Benjamin E Lauderdale

Pragmatic Social Measurement

July 18, 2022

Contents

Preface

This is a **DRAFT** book manuscript about measuring social science concepts. The current version is rife with grammatical errors, inconsistent UK and US spelling, infelicities of language and mathematical expression, missing citations, hideous regression outputs, and other rough edges. The last five chapters are missing entirely. Feedback and suggestions to (b.lauderdale@ucl.ac.uk) are very much welcome, ideally more on the content than on the writing at this stage.

This book aims to fill a niche in the range of social science methods texts that is (in the author's opinion) both important and largely vacant. That niche is defined by two important distinctions: between measurement and other kinds of inference (both population and causal), and between *pragmatic* and *representational* measurement (Hand, 1996).

The vast majority of applied statistics texts have historically been focused on population inference: making claims about populations on the basis of samples from those populations. There is a growing collection of good applied texts on causal inference written in the last fifteen years. There is not, to my knowledge, even one applied text on the practical challenges of measurement as such, even though there are lots of texts that cover statistical models that are often used for measurement. There is one recent book on some of the conceptual issues with the kind of measurement I will be exploring (Goertz, 2020) as well as some older classics (eg Blalock, 1982).

Few applied statistics textbooks cover measurement at all, they tend to assume that the origin of the numbers which will be the object of analysis is a solved problem. Others cover *representational* measurement methods—eg survey sampling design or ecological inference—where we are aiming to quantitatively measure a well-defined attribute of the world, how many units are there of a given type, with more or less adequate data. What is missing is coverage of the methodological issues involved in *pragmatic* measurement, where we are to some extent inventing—or at least conceptualising—the target concept of measurement at the same time that we conduct the measurement. A great deal of social science measurement is of this type and it presents its own distinctive challenges. These are the subject of this book.

The set of existing textbooks with which the book most closely overlaps are those that cover multivariate summary methods (eg principle components analysis) and latent variable models (eg factor analysis). These are also a subject

of this book, but are motivated differently than they are elsewhere. Here, latent variable models and related summary methods are pragmatic methods used for exploratory measurement in the absence of better data that would allow you to reliably define the target concept that you wish to measure. Ideally we would have relevant data that could more definitively link indicators to the concepts that we want to measure.

The first half of this book covers the various kinds of "supervision" that we can use to explicitly specify the connections between the target concepts we want to measure and the indicators that we have actually collected. In the second half, we turn to the "unsupervised" methods that are often used when researchers are a bit less clear about what they want to measure, or otherwise lack the data necessary to use supervised methods. Covering supervised measurement before unsupervised measurement makes it clearer that the "fancier" models involved in the latter replace substantive information about the relationships between indicators and the target concept that we wish to measure with the (often naive) hope that the target concept is the thing that maximises explained variation in the set of indicators.

Prerequisites

To make sense of this book, at a minimum you will need to be comfortable with linear regression and binary logistic regression. In particular, you need to be very comfortable thinking about linear and additive functions of variables. I will use ordinal logistic regression at several points, but if you are familiar with binary but not ordinal logistic regression you should be ok. You also need to be familiar with the basics of what I will refer to as *population inference*, which is often simply called statistical inference. That is, the logic of how you can (and cannot) make claims about a population on the basis of a sample of data from that population. At its core, such inference asks "do I have enough evidence to say something more general about the population or process from whence my data came?" This is a question that is relevant at many points here.

Some additional topics are useful, but not strictly required. Understanding causal inference is useful conceptually for thinking about both the origins of and consequences of measurement error, although it is not necessary to understand most of the material in this book.[1] If you have no familiarity with modern causal inference, Morgan and Winship (2015) or a similar book would be a good place to start. It would also be helpful to have some familiarity with machine learning concepts such as regularization and cross-validation. James et al. (2013) provides a gentle introduction to the relevant concepts, but this is by no means necessary to understand the material in this book.

The final chapters of this book, starting with Chapter 16, will assume additional outside material. You will need to know something about multilevel modelling to make sense of Chapter 16. If you actually want to design new models for specific problems, as discussed in Chapter 17, you will need some

[1] That said, causal inference is more important material for most social scientists to understand than anything that I cover in this book.

familiarity with the mechanics of maximum likelihood or Bayesian estimation, which are substantial topics in themselves. However this text is not about theories or practical details of statistical estimation: we will will focus on which models and estimators to employ, and largely ignore how that estimation actually works.

Structure

Chapter 1 concerns the question of what measurement is, and how it relates to other kinds of quantitative data analysis frequently conducted in the social sciences. I also discuss some of the sordid history of social science measurement, and the researchers who invented some of the methods covered in this book. In Chapter 2, I develop several of the key theoretical ideas that run through the rest of the book, in particular how the nature of the causal relationships between the concepts that we want to measure and indicators of those concepts inform how we go about measuring them. This is the chapter where we develop definitions of (and relationships between) *concepts*, *measures*, and *indicators*. The first two chapters are thus about defining the basic task and terminology of measurement.

Chapter 3 defines *measurement error* and develops *validity* and *reliability* as ways of understanding different patterns of measurement error that might exist for any given concept-measure pair. Chapter 4 focuses on measurement error from the perspective of individual units, and discusses questions about which kinds of measurement errors are *fair* and which are *unfair* to the units (eg individual people). Chapter 5 then provides illustrations of (some of) the consequences of measurement error for aggregate analyses of data that has been measured with error. All three chapters on measurement error are presented with a modest level of mathematical detail, an applied focus, and consequently with some hand waving towards more precise treatments elsewhere. These three chapters set the stakes for the rest of the book: what goes wrong when we do measurement poorly and why should we care?

Chapter 6 is where we start considering how to actually construct measures. In that chapter, we consider cases where we have theoretical logic that connects one or more indicators to the target concept that we want to measure. In these cases, we can sometimes derive a measure from the indicators, specify axioms that a measure ought to satisfy with respect to the indicators, or use arguments about dimensionality/units to limit the set of possible relationships between indicators and measure.

Chapter 7 considers Bradley-Terry models for comparison/competition data. These models are an introduction to several key ideas that recur in later chapters, including the use of latent variable models and the distinction between measures as summary statistics versus measures as estimated model parameters. This chapter also illustrates how creating competition data can be a useful component of a social science measurement strategy, particularly as

a means of translating qualitative expertise into a quantitative measurement strategy (an idea further developed in Chapter 9).

Chapter 8 describes the use of regression as a measurement tool. This chapter is similar to introductions to *supervised learning* as covered in machine learning texts, but with a focus on the measurement properties of using fitted values from a regression as a measure. This approach relies on the presence of a *gold standard* measure for a(n ideally random) subset of units, in order to *supervise*/*train* a model that predicts the target concept as well as possible using available indicators.

Chapter 9 considers cases where such gold standard training data are unavailable, but one nonetheless wishes to construct an *index* of indicators to measure some target concept. There are a tremendous number of indexes used across academic and non-academic applications, many of which are of dubious quality, albeit in part because these sorts of measures are difficult to construct well. This chapter discusses the properties of these (usually linear) functions of indicators, strategies for indicator selection, as well as how to specify and validate the coefficients/weighting on/of indicators. Chapter 10 discusses distinctive issues related to measurement error that arise for categorical measures, and discusses strategies for supervised measurement of categorical concepts using supervision via (logistic) regression (analogously to Chapter 8) or using coding rules (analogously to Chapter 9).

One very widely applied, but not always appropriate, strategy for combining indicators into a measure is to find a simplified representation of the indicators that maximises explained variation in those indicators. This principle, in various forms, underlies a range of summary statistics and latent variable models that are described in Chapters 11 and 12. Chapter 11 covers approaches suitable for interval-level indicators, primarily principle components analysis and exploratory factor analysis. Chapter 12 covers approaches suitable for nominal-level and ordinal-level indicators, primarily item response theory models. Following this, Chapters 13 and 14 provide analogous discussions for the measurement of categorical concepts. The former covers the most widely used techniques for unsupervised measurement of categories, including clustering algorithms and also model-based methods like gaussian mixture models for interval/ratio-level indicators, while the latter covers latent class analysis for nominal/ordinal-level indicators.

Five further chapters have not yet been written, but will proceed roughly as follows should they summon the will to exist. Chapter 15 will cover the hybrid case of class mixture models, primarily Latent Dirichlet Allocation which is used widely in the quantitative analysis of text data. The next three chapters of will cover additional advanced topics. Chapter 16 will cover the use of *hierarchical* or *multilevel* models in cases where one has multiple measurements of the same quantity that one wants to combine or one does not have enough data to form precise measures for the units of interest based on data from those units alone. Chapter 17 will cover the development of *structural* measurement

models to solve novel measurement problems by describing the statistical relationship between observed indicators and concepts of interest. Chapter 18 covers issues related to partially missing indicator data, covering frequently used strategies for working with missing data and how they apply to various methods covered throughout the book. These final three chapters should be accessible without additional background, at least to the point of being able to identify when these methods are relevant. Applying these techniques may require you to go learn some additional methods elsewhere. A short concluding Chapter 19 will wrap up.

The examples in this book cover a range of social science fields, albeit with some bias towards my own field of political science. I cannot claim to be an expert on all of these topics. I have aimed to illustrate the methods with a range of topics because they are genuinely widely applicable, and to make the book widely accessible. There is a risk in this, which is that I may have failed to do justice to the substantive considerations that should inform the application of these methods in some of the examples. Substantive expertise is absolutely required to do pragmatic social measurement well. Where the applications fall short of demonstrating such expertise, I hope that readers will recognise that there may well be ways to address these limitations, perhaps through the use of more advanced measurement techniques than I discuss here. It is often through the back and forth between substantive objections and methodological innovations that measurement techniques and substantive understandings of social science improve.

Scope

Like any textbook, this one does not cover everything that it might. What else might you want to learn to further extend your understanding of methods for measurement?

First, the scope of this book does not include survey sampling design or design of survey questions. These are clearly about "measurement", but survey sampling design is more a *representative* measurement problem rather than *pragmatic* in the sense I outline in the first chapter of the book. Probability sampling is a well examined topic, and is about population inference not measurement inference. Question wording is closer to the task here, and edges up to some of the issues that are discussed in later chapters of this text, but is not my primary area of expertise and has been covered well elsewhere.

For similar reasons, I have also not included a set of methods for situations where there is a well-defined quantity which one wishes to measure, but for reasons of data availability it cannot be directly measured. One set of such methods are those for ecological inference, where the limitation is a lack of individual-level data necessary to estimate an association in a population. Another set of such methods are small area estimation methods such as multilevel regression and post-stratification, where the limitation is a very small num-

ber of observations at the level of aggregation which is of interest. A third set of such methods are list experiments and randomised response experiments where the limitation is a (possible) inability to get truthful responses to direct survey questions. I have excluded these from this book because these problems lack the questions about conceptualisation that are the core of this book. These are methods where the thing you want to measure is in principle directly measurable, but you just do not have the data. They are all characterised by using untestable modelling assumptions to compensate for limitations of the data, and all involve risks that these assumptions will be wrong, leading to biased measurements. These are all interesting measurement problems, some of which I have personally worked on, but they are not quite in the scope of this book.

Similarly, this is not a textbook on multivariate statistics or latent variable modelling as such. Again, very good textbooks exist (Bartholomew et al., 2008) but have a somewhat different focus than this book. Here, our focus is on the problem of measuring pragmatically defined social science concepts; existing textbooks are focused on multivariate statistics and their properties. There is substantial overlap of course, and such textbooks are excellent reference materials for mathematical details that I skip in this text. In general, this book is oriented towards providing sufficient mathematical detail to apply methods, not sufficient mathematical detail to implement or derive methods. Where I skip over the underlying statistical detail, I aim to provide citations.

Structural equation models (SEMs) are not covered in this book, even though they provide a more general framework for some of the statistical models considered herein. Traditional structural equation models aim to combine measurement and causal inference in a way that is in principle very powerful, but in practice often problematic because the maintained assumptions are too complex to mentally engage with. Chapter 17 covers some related material, but focusing on modelling the relationship between indicators and target concepts rather than on introducing a general framework. If you are interested in SEMs, there are a wide variety of textbooks to consider, as well as chapters in the multivariate statistics texts like the one referenced in the preceding paragraph.

There is some overlap between the issues discussed in Chapters 8 and 9 and the problem of economic valuation and cost-benefit analysis. Such methods attempt to translate a range of quantities onto a common scale of economic value. The issues around these are well elaborated elsewhere. The UK government has official guidance on methodology, the "Green Book". For one interesting application, see the report "Measuring Economic Value in Cultural Institutions"

Finally, this book does not cover the analysis of social network data. This is again in part because of my relative lack of expertise, part because it is a well-covered area, and part because network data presents a number of unique challenges as a data structure. Nonetheless, there are methods widely used in

network analysis that are closely related to the methods covered in this book. There is a close relationship between community detection in networks and clustering of other data types; some types of exponential random graph models are very similar to factor analysis and IRT models; and the question of which network statistics are relevant to measuring different concepts is very much the sort of question that is aided by the tools covered in Chapter 6.

Acknowledgements

To be written… If you provide useful feedback on this or other drafts, your name will appear here!

1

What is Measurement?

> "For my money, the #1 neglected topic in statistics is measurement. In most statistics texts that I've seen, there's a lot on data analysis and some stuff on data collection—sampling, random assignment, and so forth—but nothing at all on measurement. Nothing on reliability and validity but, even more than that, nothing on the concept of measurement, the idea of considering the connection between the data you gather and the underlying object of your study." - Andrew Gelman, 2015

1.1 A Core and Understudied Component of Science

Measurement is a core component of making the study of humans and their interactions into something we can reasonably call "social science". Here are some examples of concepts regularly used in different social science fields:

- Psychology: Perception, Personality, Emotion
- Sociology: Social Class, Mobility
- Political Science: Ideology, Democracy
- Economics: Income, Productivity, Inflation
- Development: Poverty, Inequality, Development
- Education: Knowledge, Understanding
- Geography: Distance, Composition, Distribution

These concepts, and other like them, are used widely across the social sciences. They are how we think about and talk about what is going on. The most important thing to note about all of them is that they are not "unproblematic" or "uncontested" in their definitions. While some might look to outsiders like they are simple measurement problems, these are all concepts with varied meanings in different applications and contexts.

Social scientists often do not completely agree on how to measure the things that they are interested in studying, or indeed on how they ought to be defined in the first place. This is where social science tends to be on more uncertain ground than contemporary physical and biological sciences. Concepts like length, mass, temperature, acidity, and others are now universally agreed within the relevant fields, both in terms of their definitions and also how to

measure them. But this was not always the case. If you go and read about the history of science, a great deal of it is about developing tools for solving measurement problems. For the history of measurement in different scientific fields, "Measurement: A Very Short Introduction" by David J Hand (2016) and "Inventing Temperature" by Hansok Chang (2004) are both worth a read. If you are less interested in history and more interested in mathematical theory, multivolume treatments of the mathematics and philosophy of measurement exist as well (Suppes and Krantz, 2007; Krantz et al., 1971; Luce et al., 2007). What follows is a very slimmed down summary of some of that history and theory, as relates to the aims of this book.

The reason that some concepts—length, mass, etc—gained this status of being uncontested was (1) that they proved to be not just *useful* but *required* for scientists who wanted to describe the world to one another and (2) that scientists developed tools to measure them reliably in ways that could be understood and reproduced by other scientists. Once a concept becomes universally agreed on and the subject of reliable measure, it can become a building block on which further scientific inquiry is supported. Isaac Newton's famous statement that "If I have seen further it is by standing on the shoulders of Giants" suggests that scientists are directly supported by those that came before them. While poetic, this is a misleading metaphor for how science proceeds. Scientists are supported by the concepts and tools that other scientists build and leave behind for us to work with. We do not stand on the other scientists—they are mostly far away and dead—we build on the ideas that they create and communicate.[1]

You would have to be slightly mad to get into a fight with a physicist about the definitions of concepts like length, mass or energy. But it is also important to remember that reconceptualising these was fundamental to Albert Einstein's development of special and general relativity in the early 20th century. Even in the natural sciences, core concepts sometimes need revision long after people think they are fully understood. In the social sciences, we have to spend a lot of effort thinking about the measurement of seemingly basic concepts. We are not very good at measuring a lot of the stuff we care about.[2] Critical engagement with concepts and measurement strategies is an important part of being a social scientist.

Whether in the physical, biological, or social sciences, the best measurement techniques leverage aspects of our scientific understanding in a useful way. A classic liquid-in-glass bulb thermometer is based on the fact that (most) liquids expand in volume when they get warmer. The fact that some do so in a consistent, linear way that depends only on their temperature is both a fact that scientists figured out about how liquids behave under temperature change, and also what enabled people to make thermometers that were reasonably accurate starting with mercury-in-glass thermometers in 1714. To get to this point, scientists engaged in "promiscuous measurement" using many kinds of apparatuses and building on various intuitions that they had about the

[1] This is fortunate, because (as we will be reminded later in this chapter), many generally useful scientific ideas were developed by people who wanted to use them for abhorrent purposes.

[2] This is primarily because social science is much more difficult than biological and physical science. The systems we study are more complex, less predictable and our ability to manipulate them is more limited by practical and ethical constraints.

underlying processes involving temperature (Porter, 2020, p18). Developments in the underlying theory of temperature and the technology used to measure it developed partly in tandem and partly in parallel, not always by the same people.

Technology moves on, and now when I check my son's temperature, I instead use an infrared thermometer that relies on the fact that objects emit radiation at distinctive wavelengths in the infrared (invisible) range depending on their temperature. In 1714 no one knew about black body radiation, let alone how to measure it.[3] These measurement strategies are themselves connected to our understanding of the underlying concept of temperature and its relevance to understanding a range of phenomena in the world like the expansion of liquids and black body radiation. If the concept was irrelevant to any observable phenomena, we would struggle to measure it.

[3] In this case the technological innovation came very recently. When I was a kid way back in the 1980s, if I had a fever, my parents used a liquid-in-glass bulb thermometer on me, which was the same basic temperature measurement technology that people had been using for nearly 300 years.

1.2 Measurement versus Other Types of Inference

One way to understand measurement is that it is one of three kinds of inferences that we might want to make from data that we have observed. The terminology that I will use for these three kinds of inference is:

1. Population inference: inference from observed data to the data we would have measured if we had access to a broader population of units
2. Causal Inference: inference from observed data about the data we would have observed for the same units given counterfactual circumstances
3. Measurement inference: inference from observed data to different quantities describing the same units

Population inference is often the first kind of inference that is taught to students learning statistics, and is often simply called "statistical inference". To have reliable population inference, you need to understand the relationship between the population of interest and the sample of that population that you have observed. How did we come to observe these units, as opposed to the rest of the units that we want to characterise? Did you determine which units you observed? The ideal, *simple* case is random sampling from the population, which enables straightforward statistical inferences from sample to population, with known properties.[4] Of course in practice, simple random samples are difficult to collect in many contexts, and much applied work makes do with more or less good approximations thereof.

[4] Stratified sampling and other variations on simple random sample can be more efficient, however require more information before the sample is collected and complicate data analysis afterwards.

Causal inference is distinct from population inference because we are not making an inference from observed units to unobserved units but rather making an inference from observed potential outcomes for a set of units to unobserved potential outcomes for a set of units. To have reliable causal inference, you need to understand the relationship between data and treatment assignment. How did we come to observe the units under the circumstances that they experienced? Did you determine those circumstances? Analogously to

population inference, the ideal, simple case is a randomised experiment, which enables straightforward estimation of an unbiased estimate of an average treatment effect, with known properties.

Measurement inference[5] is distinct from either population or causal inferences because both of the former are inferences from observed values of a given quantity to some kind of unobserved values of *that same quantity*, whether for different units (population inference) or the same units under different circumstances (causal inference). In constrast, when we are engaged in measurement, we are attempting to use observed values of one or more quantities ("indicators") to make claims about a different kind of quantity: some target concept of interest. Just as for population inference and causal inference, for this to be successful we need to understand the relationships between the data we observe and the quantities about which we want to make inferences. In the case of measurement, this means we need to understand the relationship between observable indicators and the target concept that we want to measure.

[5] In the context of latent variable modelling, Bartholomew et al. (2011) refer to this kind of inference as "psychometric inference".

1.3 Representational versus Pragmatic Measurement

The "Pragmatic" in the title of this book is meant in the colloquial sense as well in a more technical sense. The colloquial sense of pragmatic applies in that we are aiming to solve problems based on practical constraints and goals. The technical sense reflects a very useful distinction, described by David Hand (Hand, 1996, 2016), between "representational measurement" and "pragmatic measurement".

1.3.1 Representational Measurement

Representational measurement theory is based on the idea that we are measuring real attributes of empirical systems in the world. There are objects in the world. Those objects interact with other objects in the world through causal processes. There are attributes of those objects that condition those causal interactions. Because there is a causal connection between the attribute of the object and these interactions—changing the attribute would change the subsequent interactions—we can learn things about the attribute by virtue of observing the interactions of the object. Representational measurement thus takes a causally *generative* view of how the attributes that we want to measure are related to the data we collect in the process of measurement. The observations that we make are causally shaped by the attributes we want to measure. The fact that an attribute of an object conditions the causal processes that the object is involved in is how you know it is a "real" attribute of the object.

For example, think about the height of a person. Your height conditions how you interact with the world. Among many other consequences, it is a consistent feature of the world that your height is a constraint on the height of openings that you can walk through without bending over. The fact that

Figure 1.1: Alice's Adventures in Wonderland by Lewis Carroll. Illustration by John Tenniel.

objects cannot generally pass through other objects is a more general causal mechanism in the world which facilitates a measurement strategy. We can measure whether an object is larger or smaller than some value by whether it can pass through an opening of that size. This is the principle behind the baggage size testers in airports: if you want to know whether something is a certain size or smaller, you can find out by trying to put it through an opening of that size. We could, if we were unkind, measure human height by having people walk under lower and lower bars until they hit their heads. It turns out there are less painful ways we can measure height too, because the physical size of an object, of which human height is just one example, causally conditions *many* physical interactions. These varied causal interactions generate the possibility of many different measurement strategies.

More generally, from the representational perspective: 1. Objects have attributes. 2. Attributes of objects causally determine the consequences of objects' interactions with other objects; if the attributes changed, so too would the interactions. 3. Therefore, we can learn about objects' attributes through a measurement procedure that involves observing *calibrated* interactions with other objects. The crucial word here is "calibrated". By calibrated, I mean that we already have a body of knowledge about how different observable interactions arise from different levels of the attribute that we are trying to measure.

Where exactly is this body of knowledge supposed to come from initially, if we lack already existing measurement procedures? The book "Inventing temperature: Measurement and scientific progress" (Chang, 2004) provides a detailed history and philosophic discussion of these issues with respect to the measurement of temperature, and is well worth reading. The way to solve this "chicken and egg" problem is in some sense the same way that real chickens and eggs came into being. Neither came first, rather they jointly evolved over time. Human understanding of temperature and the instruments for measuring it were developed together over hundreds of years, with refined understanding of the concept facilitating better measurement and better measurement facilitating a refined understanding of the concept.

1.3.2 Pragmatic Measurement

In the preceding paragraph, we made a distinction between "temperature" and "human understanding of temperature". Temperature is a defined quantity that scientists use to describe the world. However it is debatable whether "temperature" really existed prior to human definition of the concept and the scales on which it is measured. Temperature is a very convenient summary of the energy states of matter, because for many interactions that humans care about, it is sufficient to know the temperature (as defined) of materials without knowing further lower level details. This is in part due to the details of those lower level processes: the underlying physics of how individual atoms and

molecules interact with one another make temperature a useful *summary* of the system.

Humans discovered how useful this summary was from a process of conceptual refinement that very likely began with the simplest possible measurement technology: the subjective perception of temperature in the human environment combined with the invention of some ancient variation on the concepts of "hot" and "cold".[6] People found it useful to have language to *discriminate* between different subjective experiences of the world in order to communicate them to one another, the fact that there turned out to be a deep connection to thermodynamic processes need not have been the case. While subjective experience of temperature by humans is vague and qualitative, it is a starting point for conceptual refinement. If someone then observes that bodies of water freeze over only when it feels "very cold", you have a benchmark for temperature that can be usefully communicated to other people, "cold enough that the water becomes solid", and the process of conceptual refinement has begun.[7]

[6] There are many thresholds at which one might make distinctions between temperature values and different languages make different distinctions (see Koptjevskaja-Tamm, 2015).

[7] If you like this sort of conjectural history, you will enjoy the philosophy of of Jean-Jacques Rousseau.

If we take seriously that measurement is a human project, oriented towards communicating our understanding of the world with other humans, we are encouraged to take a pragmatic, or "operationalist", perspective on measurement, rather than the representational perspective described earlier. The pragmatic perspective starts, not with the idea that objects have real attributes, but rather with the idea that we have measurement procedures.

1. We can specify measurement procedures.
2. Attributes are defined as the output of a measurement procedure.

One advantage of this approach is that it is epistomologically modest. We avoid making a strong assumption about objects in the world *really* having particular attributes. We don't need to be as bothered by whether an object "really has" a temperature, so long as temperature constitutes a useful summary for some purposes. If "temperature" is a useful way to talk about world, and we have measurement procedures to facilitate that communication, that is good enough. What makes temperature useful is that it is *discriminative*: it helps us communicate distinctions between different states of the world to one another. In this instance, we further know enough physics to be able to say it also seems to be a summary of many *generative* processes, which is helpful at a pragmatic level because it facilitates robust measurement procedures. But a concept and corresponding measure can be useful even if it is not an attribute of a generative processes, if it helps us communicate differences to one another.

From this pragmatic perspective, measurements are useful summaries of things we observe in the world, reflecting attributes that we invented to help us understand and talk about the world, not attributes that necessarily existed already. This is why it generally makes sense to call them "concepts" rather than "attributes", to emphasize that someone did the conceptualisation that brought them into being. This is not an idea that everyone finds appealing,

particularly in the context of the natural sciences. "Some people find distasteful the fact that operational theory draws conclusions only about the results of measurement procedures, and leaves researchers to make an inference to an 'underlying reality' if they so wish" (Hand, 1996). However, in the context of the social sciences, it is clearer that many of the concepts that are widely used are pragmatically defined and measured.

For example, consider the concept of "democracy", measured at the level of countries.[8] This is an example we will return to in coming chapters. It would be difficult to argue that countries "really" have a binary attribute of "democracy" or "not democracy" or a continuous attribute of the extent of "democratic-ness" that itself causally determines whether they do things like hold elections or not. Rather, "democracy" is a label, a *summary*, that we use to describe how the political institutions of a country are organised (including whether they hold elections or not, the properties of those elections, etc). Scholarly arguments about which summary to use—how to measure democracy—are in fact arguments about *conceptualization*, how we operationalise the concept of "democracy".

[8] Or consider the concept of a "country", to which the same issues apply.

1.3.3 Contrasts and synthesis

One way to think about the contrast between the representational and pragmatic perspectives on measurement is that the former is a "realist" account and the latter is not. By "realist" here, I mean that representational measurement is committed to the idea that you are measuring something in the world that exists independent of you measuring it and that *causally* shapes some interactions in the world. The contrasting perspective might be called "constructed": attributes are created by the measurement procedures, they do not exist in the world until we create them. You may well have run across realism vs constructionism/constructivism debates in other contexts, this is a similar underlying contrast in perspectives.

Human height is a good example of the contrasting interpretations, if only because it seems like such a strong case for realism. You might be thinking: obviously human height is a real attribute of humans that exists whether we measure it or not. But does it, exactly? Humans are foldy in the middle and squidgy around the edges. Even if we agree on a definition of length (eg the standard metric measurement scale) we are going to get different numbers depending on whether we stand you up, lie you down, hang you upside down, put you in zero gravity, or have you walk into a giant funnel until you bump your head. Which one is your *true* height? That obviously depends on how we define the measurement procedure for "true height". So even here, with a physical measurement, some pragmatism or operationalism is difficult to avoid.

Representational measurement is, at best, aspirational for many social science applications. In practice much of what we do is pragmatic, thus the title of this book. This is not meant to be an argument for the pragmatic interpretation

against the representational interpretation in general, but rather to highlight the distinction. My own view, and the view adopted in this text, is we should aim for representational measurement where possible and be clear about when we are measuring quantities that we believe are part of a causal *generative* process. But we should also be clear about when we are being pragmatic: defining, conceptualising and measuring quantities that we create in order to be *discriminative* between different cases. Being pragmatic does not mean you are denying that there is a real world out there and that you would like to measure its real properties. No one had to articulate the concept of height or a measurement strategy for height in order for giraffes to consistently struggle to walk through low doorways. The causal processes were there regardless of whether we tried to measure them or not.

It is important to recognise, however, that if you are happy to embrace this "pragmatic realist" synthesis of the representational and pragmatic perspectives, that it implies some measurements are more or less representational versus pragmatic. Some measurement strategies will be closely connected to important and well-understood causal relations that we are very confident really exist in the world. Other measurement strategies will be less closely linked to causal relationships that we are confident exist in the world, and are simply summaries of data that we find useful. Scientists—physical, natural and social—often start with pragmatic summary measures and aim to move towards more representative measures as we improve our understanding of the underlying processes that govern what we study.[9]

This book, as clearly indicated by its title, is focused on measurement methods that are more "pragmatic" in orientation. This means that some important measurement topics are not covered in this text. An obvious "measurement" topic that I do not cover is survey sampling. Survey sampling is fundamentally about measurement: how do we collect samples of data that allow us to characterise the properties of the populations from which those samples were drawn? This is very much towards the "representational" end of the spectrum of social science measurement. There really are some number of people out there in a population, if it has been reasonably defined. If you are interested in measuring population statistics for something you can measure at the individual level, there is a causal relationship between that property of the population and the individual-level data that is generated by a well-defined sampling procedure. That causal link is encoded in the canonical statistical results on this subject, and that makes survey sampling a representational measurement exercise.

I do not think it is a coincidence that survey sampling is amply covered in many, many other books. Representational measurement tasks are often more amenable to formalisation, precisely because the causal structure that underpins the relationship between the concept being measured and the data being observed dictates the quantitative structure of the problem. In contrast, saying usefully general things about pragmatic measurement is more difficult. The researcher needs to develop the mathematical structure as part

[9] In the context of social relations of humans, measurement strategies can be as much a cause of as a reflection of the causal relationships that exist in the world. Concepts that humans invent can shape their future interactions. Sometimes things happen only because we have chosen to measure the world in a particular way.

of the process of conceptualization, and the criteria of evaluation for which mathematical structures are best tend to be more ambiguous. The fact that pragmatic measurement problems tend to provide less guidance about which maths are appropriate has a number of important consequences, among which is a tendency to apply and re-apply a set of quantitative tools with convenient mathematical properties. These tools, and the process of reasoning through their application to new problems, are the focus of this book.

1.4 Perils of Quantitative Measurement

This is a book about *quantitative* measurement of social science concepts. One articulation of why quantitative measurement is particularly valuable is given by Theodore M Porter in his book "Trust in Numbers: The Pursuit of Objectivity in Science and Public Life":

> "Since the rules for collecting and manipulation of numbers are widely shared, they can easily be transported across oceans and continents and used to coordinate activities or settle disputes. Perhaps most crucially, reliance on numbers and quantitative manipulation minimizes the need for intimate knowledge and personal trust. Quantification is well suited for communication that goes beyond the boundaries of locality and community. A highly disciplined discourse helps to produce knowledge independent of the particular people who make it." (Porter, 2020, p. xxi)

This book is not a history of quantification in the social sciences or elsewhere, nor is it centrally focused on *justifying* the project of quantification of social science concepts in general. I assume that we are embarked on the process of quantification (quantitative measurement), and focus on developing the skills and understanding necessary to do that quantification well, as opposed to doing it poorly. To do that, it is necessary to understand the criticisms of quantification, because most of them reflect real ways in which quantitative measurement can go wrong.

Within many social science fields, there are "quant-qual" divides, with scholars who do not merely use different methods in their own research, but sometimes doubt whether the methods used by others generate useful human knowledge. Different fields have vastly different balances of power and influence between those who work with quantitative data and those who do not (think economics versus anthropology). Among other critiques, quantitative research is variously accused of using simplistic and superficial data, for encouraging over-generalisation of findings, for encouraging the study of parochial topics at the expense of more important ones, and for creating an illusion of objectivity.

These criticisms of quantitative research in social science are complemented by criticisms of quantification in public policy. You may be familiar with arguments that the use of quantitative social measurement (or "metrics") can have negative consequences for society. Criticisms of educational testing are

widespread, and include criticisms that tests are too unreliable as measures of learning and too narrow in what they test. Statistical models used to make decisions in the criminal justice system regarding probation, parole and the likelihood of recidivism have been criticised for being limited, opaque and relying on information that is potentially unfair to use in evaluating an individual. More recently, Kay and King (2020) criticise mathematical modelling and "bogus quantification" in policy-making.

The most vociferous objections to quantification and ranking often come from those who are being quantified, particularly if they believe that they are themselves better equipped to make evaluations than those who would evaluate them. Rankings of universities are widely ridiculed by academics, even as they find themselves adapting their behaviour in response to rankings like that of US News and World Reports:

> It's one of the real black marks on the history of higher education that an entire industry that's supposedly populated by the best minds in the country—theoretical physicists, writers, critics—is bamboozled by a third-rate news magazine.… They do almost a parody of real research.… I joke that the next thing they'll do is rank churches. You know 'Where does God appear most frequently? How big are the pews?[10]

[10] Leon Botstein, president of Bard College, as quoted by Alice Gregory, "Pictures from an Institution", https://www.newyorker.com/magazine/2014/09/29/pictures-institution September 22, 2014

Kieran Healy argues that this impulse to reject quantification of one's own performance reflects "the loss of a profession's control over its ability to make judgments about quality and prestige in its own domain" (Healy, 2017, p513):

> Law school faculty, deans, and administrators are long past the time of their lives when their individual performance is routinely assessed in terms of As or Bs, as magna or summa cum laude. But they still exercise that capacity for judgment over others every week of the semester. They believe in it. They are still committed to the view that they know and can assess quality when they see it, and they usually think they can reliably quantify it. It is just that they would rather not be subject to that pressure themselves. Becoming a faculty member should have been a way to escape it. At the heart of an academic ranking system is the experience of having one's own knife turned back upon oneself, and finding that it still cuts like it used to. (Healy, 2017, p519)

Quantification tends to be unpopular with the quantified. Sometimes this is because evaluation is unpopular with the evaluated: the quantification per se is only part of the issue. Unpopularity is itself not much of an argument against quantification though.

What kinds of problems tend to arise from quantitative social measurement? I will discuss four general types. First, problems of narrowness: the demands of quantification might lead us to focus on some concepts at the expense of others, or narrowly defined conceptions of those concepts at the expense of richer ones. Second, quantification may create problems of fairness: the we may introduce (but also hide) biases through the process of quantification. Third, the use of quantitative measures may create unintended consequences, as people modify their behaviour in response to incentives created by the

use of these measures. Fourth, quantitative measures may enable individuals with malign intentions to exert objectionable patterns of social control. In discussing these four types of problems, I will use examples of policing, foreign aid and education as examples, as the scope and use of measurement in all of these areas have expanded over the last half century, and the potential problems associated with this have been discussed extensively by scholars and policy-makers.

Obviously it is not my view that these problems so fundamentally undermine quantitative measurement in the social sciences that we should not do it at all. Nonetheless, I want to emphasize that these are all real problems, and ones that researchers developing and using quantitative measures of social science concepts need to seriously consider in the context of their work.

1.4.1 Problems of Narrowness

Some concepts are rich, complex and/or multifaceted. Or to be less charitable, some concepts are vague in their definition. It is "easier" to maintain a rich conceptualisation when we discuss those concepts in words than when we try to translate them to quantitative measurements. When we generate measures, they are often therefore narrower, or more *minimalist*, conceptualisations. Narrower conceptualisations can be beneficial for providing clarity, but sometimes they simply exclude elements of the concept that one would have preferred to have included. The more that we then use those measures, the more that we may tend to lose track of the elements of the original concept that we did not know how to quantify. Bueno de Mesquita (2019) argues that this phenomenon is common in the evaluation of public policy, where the difficulty of measuring some consequences of public policies tends to encourage researchers to implicitly adopt a philosophy of "crass utilitarianism" where only those outcomes that can be translated into monetary terms are considered.

In the area of policing, a focus on crime statistics may reduce our focus on other criteria we might care about but which are more difficult to measure, and which may or may not vary in the same way as crime statistics (Sparrow et al., 2015). Do people feel safe going about their daily lives? Do they make decisions not to do things that they would otherwise do based on safety concerns? Do people trust police? Do people feel comfortable interacting with the police? Do the police use appropriate methods in policing? The worry is that the introduction of and focus on quantitative data on crime incidence might have the effect of displacing these concerns, although making such a causal attribution of any changes to the increased availability and use of crime rate statistics is extremely difficult one way or the other.

In the area of foreign aid, the *effective altruism* movement seeks to use evidence to maximise the amount of good that can be achieved, particularly though not exclusively in the domain of charitable donation (Singer, 2009). While few would argue against effectiveness as a goal of philanthropy, one

risk of focusing heavily on measurable quantities like cost-effectiveness is that it prioritises measurable goods at the expense of unmeasurable (or not yet measured) goods (Rubenstein, 2016). Organisations like GiveWell provide lists of "Top Charities" for which a clear quantitative evidence base can be provided, such as anti-malarial interventions, vitamin deficiency interventions, cash transfers, treatments for parasitic worm infections, and others. But does this have the effect of diverting donations from other areas where quantitative evaluation is more difficult?

In the area of education, it is common to worry that exams fail to measure many of the skills that we would like students to gain from their education. Sometimes this is obvious: a multiple choice exam will do little to test whether students can explain how they selected an answer to that question. Some of this is more subtle: if one goal of education is to develop students ability to work collaboratively, then examining them individually will never assess whether they can do so. Some of the narrowness can also be the result of the fact that exam performance depends on the things that we do not want to measure. For example, if some students are just better at taking exams, independent of their understanding of the material, they will perform better than students who are less adept at taking exams. Exam-taking ability may be a valuable skill in school, but there are very few exams once one leaves school. School is meant to be preparing one for doing well in life, not just for doing well in school.

In an essay "We Still Can't See American Slavery for What It Was", the journalist Jamelle Bouie asks "How do we wield these powerful tools for quantitative analysis without abstracting the human reality away from the story?" His concern about the potential narrowness of quantitative measurement and characterisation of the history of American slavery is closely related to the problems of narrowness described above in other domains. His conclusion is, I think, the correct one:

> "All of this is to say that with the history of slavery, the quantitative and the qualitative must inform each other. It is important to know the size and scale of the slave trade, of the way it was standardized and institutionalized, of the way it shaped the history of the entire Atlantic world. But as every historian I spoke to for this story emphasized, it is also vital that we have an intimate understanding of the people who were part of this story and specifically of the people who were forced into it."

Quantitative measurement should not blind us to the unmeasured and unquantified aspects of individuals who enter into our data analyses. Many of the failures of quantitative social measurement are such failures. At the same time, refusing to quantify is to refuse to engage with the scale or magnitude of social problems and sets us up to only care about the individuals we can see in qualitative detail. We live on a planet with billions of people. The people we can see and understand at qualitative levels of detail are disproportionately those who are close to us in space and in society, and those that various forms of media

bring to our attention for various and often unrepresentative reasons. Quantitative measurement, and quantitative methods more generally, are necessary if we want to see the world in representative ways and situate ourselves to draw conclusions and reach decisions that are responsive to the whole world, not just the parts we can see in detail. But if we want those quantitative measurements to reflect the concepts we aimed to measure in ways that do not mislead, in depth engagement with and understanding of some cases at a qualitative level is simultaneously necessary.

In sum, quantification encourages and relies on parsimonious conceptualisations. While this is often useful in providing clarity and for expanding the potential scope and scale of inquiry, it can also lead researchers to lose track of relevant aspects of the concepts that they aim to study. Whenever one is developing or using quantitative measures, it is important to spend some time thinking about how close those measures are to the concept that one actually wanted to measure. What is missing? What is there that ought not to be? The answers to these questions can sometimes motivate improved measurement, but they should always shape the conclusions that we draw from any analysis using the measures that we have.

1.4.2 Problems of Fairness

Quantitative measurement can create problems of fairness. It can also simply expose existing unfairness. There are of course many senses of "fairness" that are relevant in any social science application, most of which do not have anything to do with the quality of measurement. The "fairness in measurement" we are interested in here is fundamentally about measurement error, and is discussed in more detail in Chapter 4. The definition of fairness that I use here is that a measure is unfair when a measurement strategy systematically misrepresents the relative values of the underlying quantity that one aimed to measure when comparing different groups of units. Fairness is a comparative claim: that units which ought to be treated similarly are in fact treated differently by the measurement strategy.

Because we are most often interested in fairness to individuals, fairness criticism most obviously apply to measurements at the level of individual humans. Nonetheless, it may make sense to talk about a measurement strategy being unfair for other kinds of units as well: a measure of whether countries are democratic might be unfair to countries in one part of the world that use one set of institutions if those institutions are *incorrectly* treated as less democratic than other institutions when they *ought* to be treated similarly. Of course if those institutions are *correctly* treated as less democratic, there is no problem of fairness. This highlights the crucial point, developed further in Chapter 4, that making any claim about fairness requires you to make a clear distinction between the target concept that you want to measure and the measure you actually have. It is in the discrepancies between these—the

measurement error—that unfairness can be found.

Making a claim that a measure is unfair requires a normative commitment regarding what fair treatment is. In the context of measurement, it can be challenging to articulate the normative standard, as doing so tends to come close to requiring the development of an improved measure with which to benchmark the one being criticised. If such an improved measure were available, and everyone agreed it was better, presumably people would use that one instead. Instead, there is often disagreement about which measures better approximate the concept of interest, because there is a disagreement about the appropriate conceptualisation of that concept. Thus arguments about the fairness of measures are sometimes proxies for arguments about the appropriate conceptualisation of the concept we want to measure.

In the area of education, fairness concerns are raised with respect to both measurement of student performance and also teacher performance. With respect to student performance, we might worry that disparities in test performance between different groups (whether by race, gender, class or other descriptive categories) are not due to different understanding of the material but instead due to differential performance on the tests given the same underlying understanding of the material. For example, research by Freedle (2003) and Santelices and Wilson (2010) argued that some items on the verbal component of the SAT college entry exam in the US advantaged white test takers over black test takers, because they relied on cultural expressions that were more prevalent among white communities than black communities in the US. They observed that black students who did equally well as white students on the more difficult items on the test did worse on average on the easy ones, which tended to reflect more informal and culturally-specific uses of language.

With respect to measuring teacher performance, there are many challenges to measurement. An obvious fairness concern has to do with the fact that different teachers teach different students from different populations, and so direct comparison of student performance at the end of the academic year is unlikely to provide a fair comparison of teaching quality between different teachers. To address this, "value added modelling" is often used to adjust for students' past test scores, when assessing the gains that they have made while being taught by particular teachers. While these comparisons of gains are an improvement on naive comparisons of end-of-year performance, there are still a number of potential biases that put some teachers in a better position to excel by these measures.[11]

Quantification is no more likely to create problems of fairness than non-quantification: problems of fairness are rife throughout the social world. Quantification is useful in creating a framework for assessing the magnitude of such problems and in articulating exactly what is meant by fairness. At the same time, "naive quantification" which fails to engage with these issues can often distract people from the potential for such problems. What we even mean by "fairness" is often itself difficult: there are different conceptions of what fair-

[11] A distinct problem is the reliability of these measures for new teachers: it can take many years of students to have enough student data to confidently identify which teachers will perform better by these measures over the long-run. This distinction is developed further in Chapters 3 and 4.

ness actually requires, some of which cannot be simultaneously satisfied. This is covered in detail in Chapter 4.

1.4.3 Problems of Unintended Consequences

Measurement can distort the incentives of the people who are being measured as well as those who are doing the measuring. As discussed above, we often find it easier to measure certain aspects of a concept than others. If we go further, and attach incentives to our measurements (often called "metrics" in this context) we then will be actively focusing people on only certain aspects of the target concept, potentially to the detriment of their attention to others. Attaching incentives to measurements makes people pay less attention to the things you cannot measure, and potentially encourages them to "game" the measurement strategy by optimising the metric at the expense of the underlying concept that you aimed to measure. This will tend to undermine the measurement strategy's validity prospectively, regardless of how well it worked before one started trying to "use" the measure.

This idea is variably attributed as Campbell's Law to the American psychologist Donald Campbell or as Goodhart's Law to the British economist Charles Goodhart, as both authors first published relevant statements in 1975 (Rodamar, 2018):

> Campbell's Law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."
>
> Goodhart's Law: "Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes."

How do we know when policing is being done more or less effectively? Crime statistics have become increasingly widely used "metrics" for evaluating the performance of police over the last half century. One problem with this is that police themselves generate the statistics that are used to evaluate police performance.[12] If police are judged on the number of serious crimes, or the number of unsolved crimes, they may take advantage of available opportunities to record lesser charges or avoid recording that a crime happened at all. Already by 1976, early in a multi-decade rise in (recorded) crime in the US, this effect of increased attention to crime statistics "had as its main effect the corruption of crime-rate indicators, achieved through underrecording and downgrading the crimes to less serious classifications" (Campbell, 1976). In 2013, a whistleblower in the Metropolitan Police in London told Parliament that "that massaging statistics had become 'an ingrained part of policing culture' ".[13] Beyond simply reclassifying unsolved crimes, the incentives to have high clearance rates may induce police to focus on frequently arresting low-level criminals rather than engaging in lengthy, large scale investigations of more consequential criminal enterprises (Muller, 2018, p129).

[12] If you have ever watched *The Wire*, you will be familiar with the phrase "juking the stats".

[13] "Police fix crime statistics to meet targets, MPs told" 19 November 2013. https://www.bbc.co.uk/news/uk-25002927

Which foreign aid is worthwhile and where should aid spending be focused? As already discussed, there is a strong recent movement to try to establish which charities are more effective, but it is often very difficult to measure efficacy. We have already discussed the ways that the measures might be limited, but these various limitations of the measures beget problems of unintended consequences, as organisations refocus their attention and resources on measurable outcomes rather than potentially more important but less easily measured ones. One risk is a heavy focus on narrow conceptions of well-being involving relatively easy to measure goals like increasing monetary incomes. Another risk is a focus on short-term goals rather than long-term goals, as it is easier to demonstrate effectiveness if you do not have to wait for a long-term evaluation (Muller, 2018, p153-156).

In the area of education, the possibility that teachers will narrowly "teach to the test" is a common worry when teachers are given performance incentives that are based on student test performance.[14] The theoretical argument is detailed by (de Mesquita, 2016, p205-213). Tests can bias teachers towards expending effort inefficiently on outcomes that are more strongly reflected in test results and away from expending effort on other valuable student outcomes that are not captured by the test. If you want students to learn a set of skills, some of which are measured M by the test and some of which are not N, there is an incentive for teachers and students to focus on M rather than N. In this context, making the test "better" by improving the signal it provides about the measured components of student understanding (M), without improving coverage of the other components (N), only exacerbates the problem, pushing teachers towards even more lopsided focus on M. In one sense, this might be said to show that improved measurement is not always good; in another sense, it illustrates that we need to be careful about what we call an improvement.

[14] Mehrens and Kaminski (1989) provides an interesting discussion of the ethics of teaching to the test, from the perspective of the teacher.

One rejoinder to concerns about the incentives created by high stakes testing is that teaching to the test is good, if the tests are well designed. What is meant by a good test here is a good measurement of the concept that we want to measure, which should be students' understanding of the material, broadly rather than narrowly understood. One way that a test can be bad is if it is the sort of test which makes it possible for students to do well on the test while struggling with even a moderately different assessment of their understanding. Thus a critical question in test design is the feasibility of very narrow test *preparation*. If the test is written in such a way that one can narrowly prepare for it without a broader understanding, that creates a much larger incentive problem than if this is an ineffective strategy. The easiest tests to narrowly prepare for are those where there is a fixed set of possible questions, known in advance, and students can simply learn the answers to all these questions. Such tests create very strong incentives to simply learn all the correct answers, without learning *why* they are the correct answers, and thus failing to train students to answer any *other* related questions that they might face in the future. The problem is that "good tests" that do not have this property can be

more difficult to design than "bad tests".[15]

This point generalises to other measurement problems in the sense that some measurement strategies are based on quantitative indicators that are easy to optimise for without actually improving the underlying concept that one aimed to measure, while for others this is more difficult. Some measurement strategies are more susceptible to problems of unintended consequences than others, because some are much easier for the actors to manipulate. If police were not themselves responsible for the generation of crime data, such data might be a better tool for evaluating police performance. There would be much less enthusiasm for value-added measures in evaluating teachers if the teachers got to individually set the tests by which they were evaluated.

But even if those generating and analysing the data are not those being ranked, the incentives for manipulation can still create problems. Starting in 2003 the World Bank published an annual "Doing Business report" that ranked countries on the basis of a linear index that aimed to measure the extent to which different countries were good places to do set up and run businesses. Rankings depended on scores in 10 sub-indices covering regulatory obstacles, the availability of electricity and of credit, and the functioning of the tax and legal systems, among others. This became a widely used index in academic work and is believed to have encouraged some countries to try to ameliorate some obstacles to setting up successful businesses. However, as the prominence of the index rose, so too did the incentives to manipulate it. In 2018, the World Bank revised past reports after it became apparent that the group responsible for the index had manipulated the index values for Chile to penalise governments of the political left in the preceding decade relative to those of the political right, in a way that did not reflect any actual changes made by the governments in the relevant domains of law and regulation. Then, in early 2021, the law firm WilmerHale was employed by the World Bank to investigate and report on further improprieties related to the construction of the Doing Business reports for 2018 and 2020.

The investigation report found that in 2017, the CEO of the World Bank and senior staffers to the President of the World Bank intervened in the development of the index after the initial draft of the Doing Business report for 2018 saw China falling in the rankings, exploring methodological changes that might improve China's ranking. Ultimately this led to changes in the values of three data points for China, relating to the Bank's assessments of "the Starting a Business, Legal Rights - Getting Credit, and Paying Taxes indicators. The modifications, for example, reduced the amount of time necessary to comply with various regulations, which in turn, boosted China's score. These changes boosted China's score by nearly a point and increased its ranking by seven places to 78, the same ranking that the country had in *Doing Business 2017*."[16] The investigation report also documents similar changes made to improve Saudi Arabia's ranking relative to Jordan for the 2020 Doing Business report, as well as other irregularities. The investigation report further states that mem-

[15] Some goals of a test might also make those tests poor for other purposes: tests designed for ranking students may not be the best tests for aiding the instruction of those students or for evaluating teacher performance.

[16] https://thedocs.worldbank.org/en/doc/84a922cc9273b7b120d49ad3b9e9d3f9-0090012021/original/DB-Investigation-Findings-and-Report-to-the-Board-of-Execut pdf

bers of the Doing Business team "felt powerless to object to carrying out the data improporieties being requested by senior bank management."

In response to these findings, the World Bank chose to discontinue the production of the Doing Business report in 2021. Tim Harford, an author and commenter on the use of statistics wrote in 2021 that "I fear Doing Business was a victim of its own success. There are two types of statistics in the world: the ones that politicians ignore and the ones that politicians want to manipulate. The demands for manipulation will never go away, but the answer is not to cancel the gathering of statistics. It is to defend the independence of the statisticians."[17] But it is worth noting that the World Bank, by virtue of its other activities, was particularly at risk of this kind of pressure. If the World Bank needs the cooperation of countries like China and Saudi Arabia to complete its core mission, perhaps that means it should not also be in the business of rating those countries. It might be that a different institution, one that was itself independent, would be better able to provide measures that were not subject to manipulation. Similar concerns have been raised regarding the role of credit-rating agencies in the 2008 financial crisis, as the firms that rated mortgage-backed securities were being paid to do so by those who were creating those securities. The credit-rating agencies defence of their actions is that they were incompetent rather than corrupt in setting overly optimistic ratings for the safety of these securities.

[17] https://www.ft.com/content/c611a877-9a98-40d4-995d-e69d901df6f6

In sum, the problems of unintended consequences are real ones. They potentially undermine the usefulness of many measurement strategies and force us to think carefully about *who* should be designing and implementing measurements that will be used for high stakes purposes. There is some temptation to despair after one has read about enough such examples. If we cannot *use* measures for anything without creating bad incentives for both those entities being rated and those creating the ratings, what is the point of generating them in the first place? Can we only create measures so long as we do not try to use them for anything important? Surely it is not wrong to do *any* evaluation of police, foreign aid, and the quality of educational institutions, given the public resources that goes into funding all three?

If we could measure exactly the quantity that we wanted to measure—our preferred conceptualisation of what it means for policing or foreign aid or education to be successful—these problems would go away. The incentives would be perfectly aligned: actors optimising for the measure would be optimising for the concept, which is what we want them to do. The problem comes from the measurement error: the discrepancies between the measures and the concept of interest. This means that as we think about deploying measurement strategies, we need to think carefully about measurement error not just because we might come to the wrong conclusions about whatever it is we are measuring, but also because those measurement errors might induce behaviour that undermines the value of the measurement scheme for evaluation.

1.4.4 Problems of Malign Intentions

The preceding discussion was about the *unintended* consequences of measurement, which are sometimes bad. Sometimes the *intended* consequences of measurement are bad, or at least are considered bad by some people. In discussing these issues, it is important to make a distinction between the problem discussed in the previous section and the one discussed in this section. The former is about the way that measurement can shape which problems we are attentive to and, if poorly deployed, may therefore misalign incentives with the goals we actually have. The latter is about the way that measurement has been used to enforce social hierarchy and control: measurement can be an instrument of power, whether for good or for ill. The last section was about unintended bad consequences of measurement; this section is about cases where the intended consequences of measurement were bad.

Most of the readers of this book will be very familiar with the ways that social measurements are used intentionally to shape behaviour, because most of the readers of this book have attended or are attending university. Educational systems are, viewed in one way, a giant scheme to incentivise certain kinds of behaviour as opposed to others, through the use of social measurement.

In the UK, at the end of an undergraduate degree, students are awarded a classification: First-class honours, Upper second-class honours, Lower second-class honours, and Third-class honours, or an Ordinary degree. The category labels are confusing to outsiders, but this is simply a five-category, ordinal measurement scheme. At University College London, where I teach, there is an interval level measurement scheme underneath the ordinal categories, with each of the classifications covering a range of results on a 0-100 scale. The interval-level measures are constructed via a convoluted weighted average of 0-100 individual module/course marks, each of which conventionally does not use the full 0-100 range. Years 1, 2, and 3 have relative weights 1, 3 and 5, respectively, with further complicated rules regarding how many modules count in which years.[18]

[18] https://www.ucl.ac.uk/basc/current/degree/classification

At a basic level, the purpose of classification is to A) provide a signal to the broader world about which students understand the content of the degree more versus less well and to B) incentivise students to aim for a greater understanding of the content of their degree. Is this a problematic incentive scheme? What if the real value in a university education ought to be something besides being able to write essays that your examiners like? What if most of the learning comes from the parts of being a student that we are less good at assessing? The unintended consequence worry is that focusing on maximising your marks means that you end up with less of the things we would like you to learn. The intended consequence worry is that the whole thing is part of a malign system of social control in which we have been slowly training you to do whatever tasks you are told to do and rewarding those of you who will not cause trouble later with high degree classifications to signal to the rest of the

world that you will be well-behaved.[19]

The short answer is that, despite these negatives associated with incentivised measurement and the undeniable fact that it is in fact a social control mechanism, "we"[20] think that not doing the measurement at all would be worse. We want to make you do certain things that we think you would not otherwise do, that *is* the point. Beyond signalling to the broader world that you are the kind of person who can "do well" at stuff, classification provides some information that you did not merely do enough to get admitted before you arrived, but that you actually learned things while you were here that you might remember and be able to apply later. Marks and classifications incentivise you to actually learn the material, which we (the faculty) really do think is a valuable use of your time. We recognise that it is likely to be less immediately exciting than spending time with your friends, so we need to incentivise you by committing to tell your future employers whether you learned the material or not.

So if you feel like degree classification and course marking are all a scheme to control students, you are correct. The normative question we are not going to engage with here is whether it is a good scheme, all things and alternatives considered.

I teach at a university with a rich institutional history of measurement for purposes of (what almost everyone now agrees were) *malign* social control. One of the important figures in the early development of University College London is Sir Francis Galton (1822-1911), who donated the residue of his estate to the university in 1911. There was (until 2020) a Galton laboratory, lecture theatre, professorship, etc. Galton was not merely a donor, but also an important early statistician, with work on correlation, bivariate normal distributions and regression analysis, among other contributions.

In one of his most cited papers, Galton described a contest where people guessed what the weight of an ox would be after it was "slaughtered and dressed". He tabulated the guesses of the hundreds of participants, and reported in the journal *Nature* that the median guess was just 0.8% off of the truth of 1198 lbs. He made an argument, embedded in the article title "Vox Populi" that this was an endorsement of democratic methods, because on average the people got it right, "This result is, I think, more creditable to the trusthworthiness of a democratic judgment than might have been expected." This idea of the "wisdom of crowds" has been quite widely applied.[21]

Galton's entire career was deeply shaped by "The Origin of Species", published by (his cousin) Charles Darwin in 1859. Charles Darwin's subsequent "The Descent of Man" (1871) in turn built on Galton's work on the implications of evolution by natural selection for understanding human society. Galton invented the term "eugenics" to describe the idea that the success of human civilisations was determined in part by whether they selected reproductively for human ability. Galton's gift to UCL created a Chair in Eugenics (professorship) for Karl Pearson, for which he defined eugenics as "the study… of agencies

[19] "The Repressive, Authoritarian Soul of Thomas the Tank Engine & Friends", The New Yorker, 28 September 2017.

[20] And by "we", I mean no one in particular. That is the beauty of the thing. I am just a well-behaved cog like the rest of you.

[21] When you go back and look at the data, the 5th percentile guess was 1074 lbs and the 95th percentile guess was 1293 lbs. The attendees at the "West of England Fat Stock and Poultry Exhibition" were very good at assessing the weight of cattle by eye, but is unclear that this really tells us much about democracy. This is neither the first nor the last instance of a researcher showing a cute empirical result and then claiming it tells us something far more general than it plausibly can.

under social control that may improve the racial qualities of future generations either physically or mentally" (MacKenzie, 1981, p15). While the kinds of policies that Galton and later eugenicists were willing to endorse varied, they shared a common orientation towards this goal.

Karl Pearson held the Chair in Eugenics from 1911-1933, after which it was held by Ronald A Fisher from 1933-39. Pearson founded the statistics department at UCL and made a large number of contributions to early statistics. It is difficult to identify a statistician who has ever made greater contributions to the field than Fisher. One common thread running through all three is that Galton, Pearson and Fisher were significant figures in the development of developing social measurement for the purposes of "scientific racism". Social measurement was central to the project of scientific racism. But it wasn't just that these men developed statistical methods and then applied them to their eugenics projects, MacKenzie (1981) argues that many of the key statistical insights that these men had were in fact shaped by the kinds of eugenic arguments that they wanted to make.

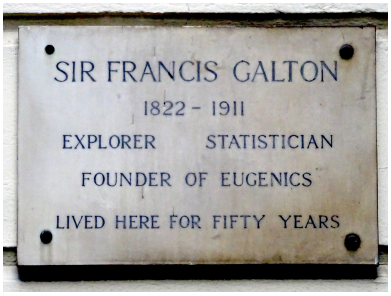


Figure 1.2: Plaque at 42 Rutland Gate, London SW7 1PD, installed in 1931.

The measurement of "human intelligence" was particularly central to this project, precisely because all the people involved viewed measurement as central to science, just as I have argued earlier. The measurement of human intelligence motivated the development of some of the core measurement methods that are covered in this book. It is important to note that not everyone who was interested in the measurement of human intelligence around the turn of the 19th to 20th century did so with an orientation towards ranking and classifying human "fitness" in either a social or evolutionary sense. As noted by Gould (1996), the creator of the original IQ test, Alfred Binet (1857-1911) was interested in developing methods to identify students who were struggling and in need of remedial help. In his writing about the test that he developed for the French government, he emphasised that intelligence was multidimensional in its forms and malleable rather than fixed by genetics. His test was meant as a rough diagnostic summary of how individuals did on tests rather than as a method of uncovering some underlying fact about the individual. In the language of this book, he was clear about the fact that he was engaged in a *pragmatic measurement* project.

However, very quickly the idea of intelligence testing was seized on as a way of providing evidence for the claim that intelligence was a real attribute of people and was largely unidimensional and therefore rankable. That is, a lot of researchers in this era assumed they were engaged in a *representational measurement* project. The UCL psychologist Charles Spearman (1863-1945) invented factor analysis, which we will cover in Chapter 11. His goal in doing so was to link intelligence test items to an underlying model of how they arose from a rankable underlying general scale of intelligence, which he called *g*. Thus both an important statistical method and its most common misuse were born together. The application of the statistical method of factor analysis does not justify a claim that you are engaged in representational measurement. The

factors you discover with such methods may or may not exist, factor analysis cannot tell you. Factor analysis is nonetheless useful as a pragmatic summary of the correlation structure in multivariate data sets, even though it cannot (alone) tell you how that correlation structure arose.

Subsequently, others went on to argue that this unitary intelligence was largely fixed from birth rather than changeable, that is, that it is hereditary.[22] Again, this view was not justified by the data, but this conclusion is appealing to those who want to justify existing social stratification and argue against any investment to close observed differences in social outcomes by class background, race/ethnicity, or gender.[23] Or indeed, if you want to argue for the eugenic ideal of improving human populations by "encouraging" some to reproduce and others not to. A unidimensional, measurable intelligence facilitates this goal: once you have a ranking the only remaining task is to decide where the cutoffs should be. The grim history of this academic tradition is detailed in *The Mismeasure of Man* by Stephen Jay Gould (1996).

This is the last that I will discuss the measurement of human intelligence in this book. The reason I will not focus on the intelligence example is that there are plenty of other, less vexatious, examples where it is easier to have a conversation about the potential for disconnects between what you want to measure and what you are actually measuring.[24] Similarly, the claims about the heredity of intelligence are really (spurious) claims about causality, and we will be using other examples to elaborate the point that even a sensible measurement strategy is not necessarily a sound causal inference strategy.[25]

Even though they are often the more evil cases, the cases where social measurement is used to justify and enforce social hierarchy and control are, I would argue, the lesser threat to the value of the project of quantitative social measurement than the previous problems that I discussed. If people have malign goals, they are likely to pursue those with or without the aid of quantitative measurement and the data analysis that it enables: the measurement is not really the problem. Indeed, racists are not the only people who use quantitative evidence to try to prove something they have already decided is true, this is done by advocates of all kinds, for causes that one would endorse as well as abhor.

The better you understand the theory and practice of measurement, the more readily you can see where people are making errors, and explain what those errors are in a precise way. The stakes are high in much the same way they are in other engineering and scientific domains. Physical mismeasurement can lead bridges to collapse; chemical mismeasurement can lead to people being poisoned; social mismeasurement can lead to profound and enduring injustice. Misuse is not intrinsic to the project of measurement itself, but the risk of misuse certainly is. Similarly, bridges only collapse if you try to build them in the first place, but we tend to treat their utility as self-evident. We usually have good reasons to try to measure the things whose mismeasurement may cause problems. The historic misuse of social measurement is a reason to

[22] Pearson and Spearman are both known to introductory students of statistics for the similar correlation coefficients associated with their names and respectively held chairs in Statistics and Psychology at UCL in the same period. Their chairs passed to Ronald A Fisher and Cyril Burt. Burt was a hack who fabricated data later in his career and also attempted to take credit for factor analysis after Spearman died. While Spearman was not particularly interested in the eugenics project, the other three men were all strongly committed to the idea that intelligence was largely hereditary, that there were class and racial differences in this innate intelligence, and therefore that efforts to remediate class and racial differences were at best pointless. (Gould, 1996, p302)

[23] But see Manski (2011) for an explanation of why even hereditary variation would not imply that no actions to reduce these differences is justified.

[24] If you are really interested in the intelligence example, read "g, a Statistical Myth" by Cosma R Shalizi.

[25] See Gould's book, and also "Yet More on the Heritability and Malleability of IQ" by Cosma R Shalizi

study and understand measurement well, not a reason to view it as something bad that should be avoided.

1.5 Conclusion

If you feel uncomfortable about the proximity of the material in this book to the intellectual history described above, that is entirely reasonable and indeed helpful. Social measurement, like social science in general, is a human enterprise and has been used by humans for good and for ill. The same is true of research in the biological and physical sciences. Humans can make antiobiotics and vaccines or they can make biological weapons; humans can produce nuclear energy or nuclear weapons. Expertise in most scientific fields can be—and has been—used to help as well as to harm.

The reason I discuss this history here, at the outset of this book, is that it reinforces the importance of understanding social measurement. Social measurement already shapes your lives in profound ways. It is not going away. It can be used to make peoples' lives better or worse. Bad measurement causes people to make bad decisions every day. As I will discuss in Chapters 3, 4, and 5, some measures are too unreliable, too biased, or otherwise unsuited to be useful for some applications. Knowing the criteria by which to evaluate measures enables you to take advantage of measures when they are suitable for an application and avoid using measures for applications in which they will mislead.

This book is focused on how to do social measurement carefully, particularly for the kinds of social science concepts that are difficult to measure. Conveying how to do measurement well requires talking about common modes of failure. In the next chapter, we will start by defining measurement error and thinking about when measurement error is likely to be consequential for analyses involving those measures.

2

Conceptualisation and Causality

Is a country a "democracy" because it has elections, or does a country have elections because it is a "democracy"? Is an individual highly "creative" because they generate a lot of new ideas, or does an individual generate a lot of new ideas because they are "creative"? Is a person "conservative" because they take on a collection of "conservative" positions or is that person taking on those positions because they are a "conservative"?

All of these are question about the conceptualisation of (i.e. how we think about) particular concepts, but they are also questions about causal relationships. Specifically, they are questions about whether these concepts are summaries of the data we observe in the world that we use to discriminate between cases or whether the concepts describe attributes of the causal processes that generate the data we observe in the world. If we say that what it means to be a democracy is that a country holds regular elections plus some other requirements, we are saying that the concept of "democracy" is a summary that we are using to discriminate between different states. If, instead, we say that countries which *are* democracies will regularly generate elections, we are saying that "democracy" is a causal attribute of states that tends to generate certain observable consequences.

This distinction between concepts that are **discriminative** between different patterns of observations versus concepts that are part of the **generative** process for those observations is far more important than is often appreciated. Our written language is often ambiguous when it comes to key concepts. As noted above, the same words often are consistent with both interpretations, and we often slip between them without being clear that we are doing so. "Professor X published six papers last year, she is incredibly productive. I wish that I were that productive." Is *productive* a summary of this output of papers that we use to discriminate between different scholars or is it a summary of some fundamental properties of Professor X that enabled her to generate more papers?

In order to properly evaluate these questions, we need to start by thinking about the relationship between concepts and measures, to introduce the idea of *indicators* as data that are used to form measures, and then to carefully

think about the causal relationships between the indicators and the concept of interest. In order to do these things in a coherent way, we have to make a critical distinction between the *estimand* μ, which is the concept that we want to measure, and the *estimate* m, which is the actual measure that we are able to construct. Why is this distinction important?

> The estimand is the object of inquiry—it is the precise quantity about which we marshal data to draw an inference. Yet, too often social scientists skip the step of defining the estimand. Instead, they leap straight to describing the data they analyze and the statistical procedures they apply. Without a statement of the estimand, it becomes impossible for the reader to know whether those procedures were appropriate. The methodological approach becomes tautological: if the thing to be estimated is defined within a statistical model, it cuts off productive consideration of a broader class of models that could accomplish the same goal. (Lundberg et al., 2021)

We cannot evaluate the quality of any measure without making a distinction between the thing we wanted to measure and the constructed measure itself. This means we need to be able to articulate clearly what it is that we wanted the measure, which is often a significant challenge in itself because it is not something for which we already have data.

2.1 *The Relationship Between Concept and Measure*

From the representational perspective, in order to measure something there must be some causal connection *from* the thing we are trying to measure μ *to* the data that we use to construct the measurement m. I will refer to such a relationship between target concept and measure as **generative**: changes in the target concept *generate* changes in the measure through a direct causal pathway.[1]

Figure 2.1 depicts this with a directed graph in which there is a causal pathway from μ to m: changes in the measure are caused by changes in the target concept. If something changes in the world (μ), then this will lead to changes in our measure. For this to be true, μ needs to capture some feature or features of the causal process that generates our measure. This is obviously the case for something like using a ruler to measure the length of an object: if the object gets shorter, that will cause you to get a smaller value of the measure when you follow the measurement procedure of holding the ruler up next to the object and reading off the difference in ruler marks at the two ends of the object.

This logic becomes less obvious when we turn to a social science concept like democracy. For the concept of democracy to be measurable from a representational perspective, it needs to be the case that there is a latent extent to which a country is a democracy, and that changing this will change whatever observable indicators (like whether we observe regular elections) form the basis of calculating the measure. We will unpack the details of where indicators fit into this causal graph later in this chapter.

[1] This terminology is borrowed from computer science language for distinguishing between different types of classifiers.

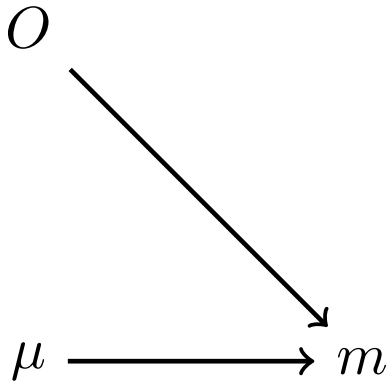


Figure 2.1: Representational measurement assumes that the target concept μ is part of the causal process that generates the measure m of that concept.

The directed graph also highlights the key threat to the quality of measurement: variation in the measurement m can be caused not only by changes in the target concept μ, but also by changes in other factors O that influence the measure. It is from these other factors O that the measurement error ϵ_M arises.

From the pragmatic measurement perspective, these causal relationships between μ and m could go either way (Figure 2.2). They might be **generative**, where μ is part of the causal process that generates m, but they might also be **discriminative**, where m is what defines μ (Figure 2.3). The defining feature of the pragmatic perspective is that it lacks the strong commitment about the direction of these causal relationships between μ to m. It might be that $\mu \rightarrow m$, as before. It might, however, be that $\mu \leftarrow m$. That is, that the concept is being defined by the measurement strategy. What it means to be a democracy might be defined by how we choose to measure it. If a country starts holding regular elections, it *becomes* (more of) a democracy. The measure discriminates between different levels of the concept, between the sorts of countries we have decided to call "democracies" and the sorts of countries that we have decided to not call "democracies".

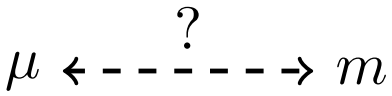


Figure 2.2: Pragmatic measurement does not specify a particular causal relationship between target concept μ and measure m.

Figure 2.3: Discriminative measurements are those where the measure m defines the target concept μ.

We have now discussed two distinctions: representational versus pragmatic and generative versus discriminative. Representational measurement aims to create generative measures: if the goal is to describe quantities that in some sense already exist in the world, those quantities need to be part of the causal processes that affect how entities in the world interact and generate observable quantities. Pragmatic measurement may create generative or discriminative measures. We might hope to describe quantities that in some sense already exist in the world, affecting causal processes, but we might also want to define a new concept in the world in order to discriminate between different cases.

2.2 Example: Flipping a Biased Coin

The example of flipping a potentially biased coin is ubiquitous in the teaching of statistics, but it is very difficult to construct a coin that has the shape of a coin and which is substantially biased with respect to flipping (Gelman and Nolan, 2002). Serious and non-subtle interventions are required to yield a substantial bias. Figure 2.4 shows a US 1 cent coin which has had an unfortunate accident.

Figure 2.4: A bent penny (heads up).

In a short trial, when flipped onto a hard surface on which it bounced until it came to rest, this coin came up heads 40 times and tails 60 times in 100 trials (see Figure 2.5). Thus, we can define a measure $m = \frac{\text{heads}}{\text{trials}} = \frac{40}{100} = 0.4$ that is the proportion of heads observed in some number of trials. We can also define a target concept μ that is the proportion of heads we would observe in the long run, in a very large number of trials.

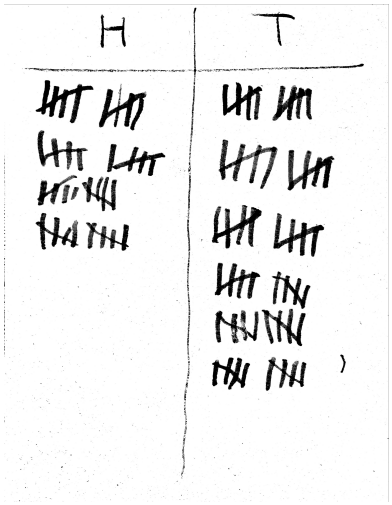


Figure 2.5: A tally of 100 flips.

In this instance, it is easy to see the conceptual distinction between m and μ. The former is the proportion we calculate in a finite set of trials, the latter is what we would calculate if we could keep doing trials forever. In this case,

recalling our knowledge of basic statistics, we can see that m will have a lot of good properties as an estimate of μ. It is unbiased ($E[m] = \mu$) and consistent ($m \xrightarrow{p} \mu$).

The potential for a numerical discrepancy between m and μ when m is calculated based on a small number of trials is familiar. As you will know from introductory statistics, if you do 100 trials of a "heads vs tails" Bernoulli process like this one, the proportion of heads will vary substantially. As a consequence, 100 trials does not provide very strong evidence that the probability of heads is actually below 0.5 for this coin (the p-value for a two sided test of the null that $\mu = 0.5$ is 0.06) but my research assistant got bored and did not want to do more trials. Previous research suggests that bending a coin in this way will reduce the heads probability by about this amount (Izbicki, 2011).

The tricky question for us here is what sort of concept we think μ represents. Is μ a 'real' property of the coin that we have sought to measure through the tedious procedure of repeatedly flipping it? Or did we collect some data, form it into a measure (proportion of heads in some number of trials) and then invent a concept (the long run heads probability of this particular object) that is actually defined by our measurement procedure?

It is clear that, if you want to say that μ is a real property of the coin, we have to recognise that this property is actually a summary of other more fundamental features of the coin (eg size, mass, and shape) as well as the context of the trial (eg the height from which it is flipped, the initial distribution of momentum and angular momentum, the properties of the surface on which it will land). If you had all of this information and a sufficiently good computer program for simulating the physics of the flips, you could perhaps simulate the coin flips and estimate the probability simply from the geometry of the coin. The quantity μ is a summary of properties of the coin and of the trial that is relevant to this particular type of trial, but it is not actually a property of the coin alone.

So, if we seek to estimate μ, whether by tedious physical coin flipping or elaborate computer simulated coin flipping, are we engaged in representational or pragmatic measurement? Are we seeking a measure that of a generative concept or a discriminative concept? In favour of the representational perspective, there is a physical process that is being characterised: the coin clearly exists in the world and the flipping process is a physical process (albeit one that is not all that precisely defined). However, in favour of the pragmatic perspective, it is clear that this quantity μ is defined in terms of the measurement strategy m: it is just the long-run limit of running these trials indefinitely. These are not mutually exclusive perspectives in general or in this instance.

In contrast, with respect to whether the concept is generative or discriminative, the answer is clearer. It is correct to say that μ is a generative concept, it is clearly possible to change μ by reshaping the coin in a way that will change the measure m (at least if we do enough trials). Causality runs from μ to m, via the observed values of the individual coin flips. Those coin flips, the data we collect

to construct the measurement m, are an example of an *indicator*, which is the topic of the next section of this chapter.

Measuring the long run probability of heads for a bent coin is a toy example, but it highlights some of the challenges with thinking about measurement. The good news is that we typically do not need to make a strong commitment about whether we are doing representational or pragmatic measurement, as the latter is flexible and we can always aspire to the former. We do, however, need to be clear on whether we are engaged in measuring a concept that describes part of the causal process that generates the observable data we are using to measure that concept, or whether we are defining a concept as a discriminative summary of some observable data that are interested in. This distinction, between generative and discriminative measures, is consequential for how we think about different measures and the appropriate methods to construct them.

2.3 Indicators

Sometimes measures arise directly, but more often they are constructed as functions of one or more observable *indicators*. What qualifies as an indicator, exactly? An *indicator* is an already measured quantity that provides evidence regarding the concept that we aim to measure.

Critically, indicators are partial constituents of or noisy manifestations of the underlying concept, not the concept itself. If you have ever watched sports and thought that the better individual/team did not win, you recognise this distinction between a concept of interest and indicators of that concept. Winning a particular match is an indicator of being a better individual or team, but it is not the same as being a better individual or team.

If I am giving an examination in a course, with the aim of measuring *how well students understand about the content of the course*, the quality of students' answers on each question are indicators of that target concept. If we take a representational perspective on the exam, students really have some level of *understanding of the course material*, and this causally influences whether they answer the questions in ways that get high marks. Variation in the indicators is caused by variation in the target concept: if a student improves their understanding of the material, they will (at least in expectation) end up with a higher mark on the exam.

If we instead take a pragmatic perspective, one possibility is that what it means to have a high "understanding of the course material" is just that you get high marks on the exam. The concept then is simply a summary of the indicators. If you followed this definition of μ, it would become impossibly to define what it means for an exam to be a poorly designed measurement scheme for "understanding of the course material", $m = \mu$ by definition. In practice though, nothing stops you from taking a pragmatic perspective on measurement and still maintaining a definition of μ that is distinct from the procedure that generates m. In this example, you can have an ideal of what "understanding of the

course material" entails, and recognise that the measurement procedure for m is at best approximating that ideal. This is true even if you recognise that your ideal of understanding of the material μ is a summary of students' ability to do a set of tasks well (perhaps a larger set than can be examined) rather than a generative attribute of students that causally generates better or worse exam answers under some causal process. From this perspective an exam m might be poorly designed for measuring μ because it includes tasks that are unrepresentative of the larger set or because it includes tasks outside that larger set. Neither taking a discriminative perspective on the causal relationship between concept and measure nor taking a pragmatic perspective on the definitional relationship between concept and measure requires you to give up on the possibility of evaluating the quality of a measurement strategy. This is a point that we will return to below and in Chapter 3.

The commonly used terms *proxy variable* and *surrogate variable* are examples of indicators that are typically used by themselves, as a measure of a concept. Upton and Cook (2014) define a *proxy variable* as "a measurable variable that is used in place of a variable that cannot be measured" and a *surrogate variable* as "a variable that can be measured (or is easy to measure) that is used in place of one that cannot be measured (or is difficult to measure). For example, whereas it may be difficult to assess the wealth of a household, it is relatively easy to assess the value of a house."

More generally, when we talk about indicators—for example, whether countries hold elections as an indicator of whether they are democracies, whether people say they are happy as an indicator of whether they are happy, whether the prices of particular goods have increased since last year as an indicator of inflation—there is some translation needed to get from these measurable quantities to the concept that we are actually interested in. Proxy/surrogate variables are often *uncalibrated*, which is to say that while they are indications of the presence/absence or level of the target concept, they are unlikely to be on the same scale as that target concept. You might know the assessed value of someone's home, and that people with greater wealth will tend to live in homes that are more valuable, but clearly you cannot simply use the numerical value of the home as the numerical value of wealth, the latter is likely to be some multiple of the former. Figuring out what that multiple typically is, and thus how to translate home values into a(n imperfect) measure of wealth, is the task of *calibration* (this is discussed further in Chapters 6.2 and 8).

Whereas proxy/surrogate variables are single indicators of a target concept, in cases like an exam with multiple questions used for measuring student understanding or a basket of goods used for measuring inflation, calibration involves combining information from multiple indicators. In order to do this, we need some information about the nature of the *collective* relationship between these indicators and the target concept.

Figure 2.6 shows the causal relationships that might exist between concept, indicators and measure, from a representational and generative perspective.

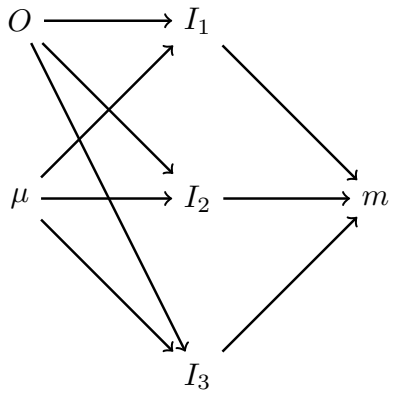


Figure 2.6: Representational measurement perspective on the causal relationships between target concept μ, indicators I, measure m and other factors O.

There is some causal relationship from the target concept μ to the indicators I_1, I_2, etc, but there are also other (potentially unknown or unobserved) factors O that are influencing the indicators as well. The measure m is a function of the indicators. The graph highlights the key threat to the quality of measurement, discussed more later in this chapter. The measurement is potentially shaped not only by the target concept, but also by any other factors that influence the measure via its constituent indicators.

From a pragmatic perspective, illustrated in Figure 2.7, our measure is a function of the indicators, but the nature of the causal relationship between the indicators and the concept of interest is unclear. Maybe we are measuring something that exists in the world and that is shaping the indicators, as in Figure 2.6, but maybe we are instead engaging in a discriminative exercise summarising differences between sets of indicator values in a way that we define that is distinct from whatever causal processes created those indicator values. The measure is definitely providing a summary of the indicators, but it may or may not reflect some component of the generative process for the indicators.

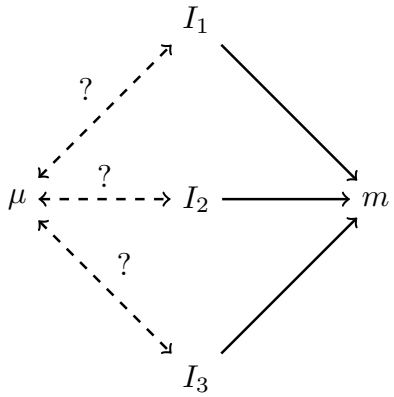


Figure 2.7: Causal relationships between target concept μ, indicators I, measure m and other factors O in pragmatic measurement.

Looking at these graphs, we can see two immediate things we are going to need to understand in order to do measurement, regardless of whether we take a representational or pragmatic perspective. First, we need to figure out what qualifies as an indicator and which indicators we should use. Second, we need to figure out how we put those indicators together in order to form a measure. In cases where there is a generative process which we want to measure some aspect of, we want to do this in a way that closely reflects the causal connections between the target concept and the indicators, so that our measure approximates the target concept as well as possible. In cases where we are engaged in a discriminative exercise, we are going to need to find some other criteria that we can use to decide how to put the indicators together, because we cannot rely on the causal processes that generate those indicator values to provide guidance.

2.4 Supervised versus Unsupervised Measurement

Figures 2.6 and 2.7 highlight an important practical question that arises once we contemplate measurement using multiple indicators. How do we go from I_1, I_2, etc to m? Where does our information about the nature of the relationship between indicators and the target concept come from? How do we figure out how to combine multiple indicators into a single measure that reflects those relationships to the target concept? To a large extent, different possible answers to this question are the content of the rest of this book. There are three kinds of general approaches that we will see:

1. Theory: we know the relationship between the indicators and the target concept because that relationship is dictated by theoretical arguments

and/or known features of the relationship between the target concept μ and the indicators I_j.
2. Supervised: we have some kind of data set that enables us to estimate the relationships between the indicators and the target concept.
3. Unsupervised: we use our data set of indicators in order to learn which relationships between indicators and the target concept would enable us to most effectively explain variation in the indicators.

Theoretical approaches are discussed in Chapter 6 and again briefly in Chapter 10, supervised approaches in Chapters 7, 8, 9 & 10, and unsupervised approaches in Chapters 11, 12, 13, 14 & 15.

These different approaches do not map straightforwardly onto whether we are engaging in *generative* versus *discriminative* measurement. In cases where we have a generative process that connects the concept of interest to the observed indicators there is some prospect of recovering it using unsupervised methods. However, we can also apply unsupervised methods for purposes of *discrimination*, albeit with the limitation that we can only claim to be summarising (co-)variation in the indicators. In cases where we have a *generative* process, we may also apply supervised/theoretical approaches to mapping indicators onto measures, leveraging existing knowledge and information to improve the quality of measurement. However these approaches are particularly valuable when we are engaging in *discriminative* measurement, because in such cases we cannot rely on the process that generates the indicators to provide any information about how to map them onto our concept(s) of interest.

In general, stronger theory and stronger supervision are good things, reflecting greater pre-existing knowledge of the relevant relationships between indicators and the target concepts that we wish to measure. Using these methods makes it more likely that you are measuring what you wanted to measure. Unsupervised methods excel at summarising (co-)variation in sets of indicators, but can provide no guarantees that the measures recovered measure anything in particular. Sometimes this is fine and useful, but it comes with a number of risks that we will discuss in the later chapters of this book.

2.5 Conclusion

This chapter has elaborated several theoretical distinctions that are relevant to understanding different types of measurement problems. The first of these, representational versus pragmatic measurement, is one related to the definition of a concept in relation to a measure. The second of these, generative versus discriminative measurement, is related to the causal processes that connect a concept and a measure. These are related in that representational measurement is necessarily generative in its perspective on causality, but they are not the same distinction in that pragmatic measurement is compatible with concepts that are either generative or discriminative.

Finally, we have introduced the idea of an indicator, and then have briefly introduced the distinction between unsupervised, supervised and theoretical approaches to determining how indicators are combined in order to construct a measure. In the next chapter, we proceed to defining levels of measurement and measurement error.

3
Measurement Error

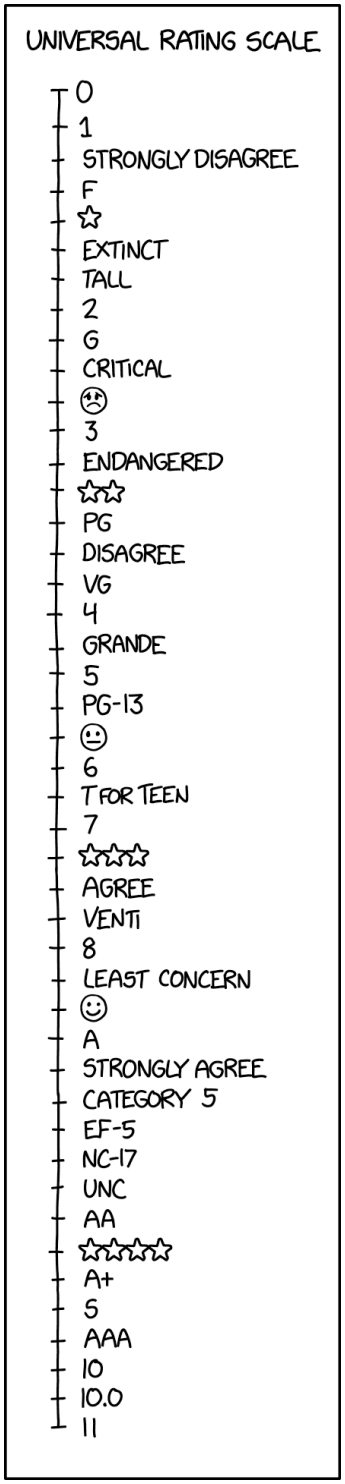


Figure 3.1: Universal Rating Scale
https://xkcd.com/2329/

In order to have sensible discussions about different measurement strategies that we might want to use later on in the course, we need to have a language in which to have those conversations. What is the thing we are trying to measure? What constitutes a measurement? What are the ways in which a measurement can be good or bad? These are questions we need to think about before we can do much else.

We already have begun to build a language for considering these questions. We have defined the idea of a target concept that we are aiming to measure μ, and noted that an actually constructed measure m may not be equal to that. The core problem of measurement is constructing m such that it is as similar to μ as possible. Ideally, $m = \mu$, however entirely eliminating all discrepancies between the two, *measurement errors*, is seldom possible. In order to be precise about what we mean by the concept of measurement error, it is necessary to start by defining the *levels of measurement* that might be appropriate to different target concepts μ.

3.1 Levels of Measurement

The mathematical details of a measurement procedure as well as the way we talk about measurement error usually differ according to the *level of measurement* that is appropriate to the target concept μ. You may already be familiar with the classic distinction between different levels of measurement (Stevens et al., 1946):

- Nominal: Numeric values convey neither ordering nor distance
- Ordinal: Numeric values convey ordering but not distance
- Interval: Numeric values convey ordering and distance
- Ratio: Numeric values convey ordering, distance, and have a meaningful zero point

The only operations that can be used on nominal measures are tests of equality = and inequality ≠, nothing else. Ordinal measures allow one to also use operations involved in sorting, such as ordering tests like greater than >

and less than <. Interval level measures can further be used to assess distance and difference, allowing meaningful use of addition + and subtraction −. Finally, ratio-level measures additionally allow meaningful use of multiplication × and division /. The four levels of measurement are themselves ordered, all operations that can be applied to a given level of measurement can also be applied to those further down the list.

Nominal measurements typically involve typologies, categorizations and classifications. These can be anything where there is no necessary logical ordering between categories, such as types of transit (1 = foot, 2 = bike, 3 = car, 4 = bus, 5 = train, 6 = plane), type of electoral system (first-past-the-post, two-round, party-list proportional, open-list, etc), and many others. There may be ways that you could typically sort the levels (eg by speed or number of passengers for the transit case, by majoritarian vs proportionality for electoral systems) but these are not intrinsic to the categorization.

Ordinal measurements typically involve classification that has a qualitative link to relative levels of an underlying concept. This includes grade classifications (A/B/C/D/F or 1st/2:1/2:2/3rd), Likert scales in surveys (strongly agree / agree / neither agree nor disagree / disagree / strongly disagree) and many others.

Interval measurements (that are not ratio-level) typically arise in cases where difference and distance are more relevant than absolute levels, or the latter is difficult to define. Map coordinates on the Earth's surface are often interval scales because there is no (non-arbitrary) zero point. The longitude of Copenhagen is about 12.5 degrees east and the longitude of Helsinki is about 25 degrees E. Nonetheless, Helsinki does not have twice as much longitude as Copenhagen, because while the difference between these is meaningful, the ratio is not. This ratio would only be meaningful if you wanted to know the east-west distance from the Royal Observatory in Greenwich, which is arbitrarily set as the 0 point for longitude. Arbitrary specifications of 0—Greenwich, England for Longitude; the freezing point of water for degrees Celsius—are usually hints that a scale is interval-level.

Ratio measurements are distinguished from interval level measurements by the fact that the zero point is meaningful.[1] Some of these are quantities that cannot be negative, like measures of physical length (eg in meters) or a number of votes. In other cases, like the net budget balance of a government, the 0 point is special (balanced budget) but both positive and negative values are possible. These are often the result of taking differences between two strictly positive ratio-level quantities (eg government revenue minus government expenditure).

Much of this book discusses interval-level and ratio-level measurement of *scales*. The aim of creating measures that are *actually* interval-level provides useful structure for several of the methods we examine. Many of the measurement techniques discussed in this book yield interval-level rather than ratio-level measures, but there are important exceptions discussed in Chapter 6

[1] Temperature does have a meaningful zero point: absolute zero. The Kelvin scale is a ratio-level scale, Celsius and Fahrenheit are interval-level scales. When you learn to do calculation with the ideal gas law $pV = nRT$ in an introductory chemistry class, you need to remember to use temperatures T in Kelvin because multiplication is only valid with a ratio-level measure.

and to a lesser extent Chapter 9. Even creating interval-level measures of many social science concepts is a struggle, ratio-level measures are not always feasible. Chapters 9, 10 13 & 14 discuss ordinal-level and nominal-level measurement of *classes* or categories.

Whether a given concept is amenable to measurement at these different levels is to some extent dictated by the concept and to some extent dictated by the indicators that you have to work with. For example, it is frequently the case that while a concept could theoretically be understood to exist on a ratio or interval scale, you lack data with which to measure that concept in a way that recovers reliable information about relative distances, and so your measure is explicitly or effectively ordinal.[2]

[2] This is a reminder that levels of measurement do not map perfectly onto whether variables are discrete or continuous. So it is possible to have continuous, ordinal-level measures. It is also possible to have discrete, interval-level or ratio-level measures (eg counts).

3.2 Defining Measurement Error

There are different definitions of measurement error appropriate to different levels of measurement. To define measurement error, the discrepancy between μ and m, you need to use mathematical operations like $\neq$, $-$ and / that are not all meaningful for all levels of measurement.

For nominal measurements, those with no necessary logical ordering between categories (eg foot, bike, car, bus, train, plane), the only allowed operations are tests of equality. This means that the only definition of measurement error for such measures is binary: either you have a measurement error $m \neq \mu$ or you do not $m = \mu$. Formally, we can define the measurement error $\epsilon_m = 1$ if $m \neq \mu$ and $\epsilon_m = 0$ if $m = \mu$.

For ordinal measurements, which add logical ordering between categories (eg strongly agree / agree / neither agree nor disagree / disagree / strongly disagree), we can use tests of equality as well as tests of ordering. This means we can define three levels of measurement error: $m < \mu$, $m = \mu$, and $m > \mu$. We can also begin to say something about the magnitude of errors, although not in a way that is typically very useful. If our ordered scale runs A, B, C, D, then if $\mu = C$, a measure $m_1 = A$ has a larger error than a measure $m_2 = B$. However the possible rankings of measurement error magnitudes are only partial, as we cannot say whether these errors are larger or smaller than the error in the case where $\mu = D$ and $m = C$.

When we move to interval-level measurements, which have meaningful distances and differences, the use of addition + and subtraction $-$ allows us to define measurement error ϵ_m as the difference between the measure you have m and the thing you wanted to measure μ:

$$m = \mu + \epsilon_m \tag{3.1}$$

$$\epsilon_m = m - \mu \tag{3.2}$$

This is the most frequently used definition of measurement error. For continuous, interval-level quantities, where exact matches between m and μ may never occur, the measurement error definitions that apply to nominal and ordinal

measurements are not especially useful because the magnitude of discrepencies between m and μ is the relevant question, not the mere fact of them.

With a ratio-level measurement, we can continue to apply the definition based on the difference, but it will occasionally make sense to apply a definition of measurement error based on the ratio between m and μ:

$$m = \mu \cdot \xi \tag{3.3}$$

$$\xi = \frac{m}{\mu} \tag{3.4}$$

$$\log(\xi) = \log(m) - \log(\mu) \tag{3.5}$$

Defining the measurement in this was is equivalent to defining measurement error as a difference on a log scale.

Why would one want to do this? Imagine that you are trying to measure the GDP of countries. A poor country might have a true per capita GDP of $500 and a measured per capita GDP of $550, while a middle income country might have a true per capita GDP of $5000 and a measured per capita income of $5100. If you calculate $\epsilon_m = m - \mu$, the measurement error for the poor country is $50 and for the middle income country it is $100, the latter being larger. However, if you calculate $\frac{m}{\mu}$, you note that the ratio for the poor country is $\frac{m}{\mu} = 1.1$ or 10% of the country's true GDP, while for the middle income country it is $\frac{m}{\mu} = 1.02$ or 2% of the country's true GDP. The measurement error in the poor country's GDP, while smaller in dollar terms, is larger in percentage terms, and may therefore mischaracterize the country's economic output to a greater extent. Depending on the application, it may make sense to think about measurement error in these terms.

3.3 Validity, Reliability; Accuracy, Precision; Bias, Variance

When we think about measurement error, it is important to keep in mind the distinction between the errors of individual measurements and the properties of a set of measurements for many units. If we want to think about the properties of many such measurements, we need to put i subscripts on m and μ, where $i = 1, 2, \ldots$ indexes each of a set of measurements of the same target concept for a set of measured units.

In the case of an nominal measure, $\epsilon_{m_i} = 1$ if $m_i \neq \mu_i$ and $\epsilon_{m_i} = 0$ if $m_i = \mu_i$. In the case of an interval level measure:

$$\epsilon_{m_i} = m_i - \mu_i \tag{3.6}$$

While the measurement error for a single measurement is a single number, there are multiple ways that a set of measures m might "go wrong" as a representation of the target concept μ across a set of units. In the following discussion I will focus on how these apply to the nominal-level and interval-level cases, as these are the most frequent cases in application (and can often be used for ordinal-level and ratio-level measurements, respectively).

There are a variety of terms have been developed to talk about ways that measures can be wrong across a set of units. For example, two different ways that measurements can go wrong are illustrated by the contrast between the first two panels in Figure 3.2. Panel a shows measurements scattered widely around a target, but on average they are centred on the target. Panel b shows points tightly clustered together, but some distance from the target. Think of each point as a measurement and the centre of the target as the "true" value that we are aiming to measure. Both plots show measurements that have errors, but they have different patterns of errors and the distinction is very important.

There are a variety of terms that are widely used to correspond to the characterise the patterns of errors in Figure 3.2.

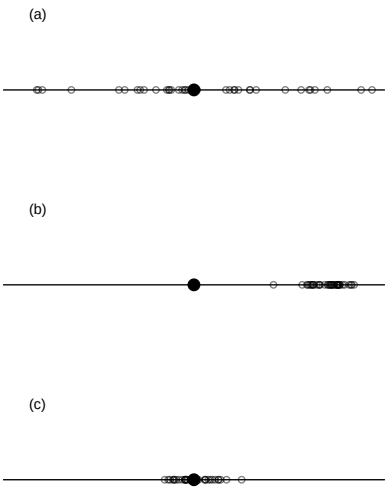


Figure 3.2: Where the large black dot is the target μ, each panel shows a set of measurements m_i with different properties.

	panel a	panel b	panel c
validity	low	low	high
reliability	low	high	high
accuracy	low	low	high
precision	low	high	high
bias	low	high	low
variance	high	low	low

These terms are not entirely interchangeable, and not only because some describe the things you want (validity, reliability, accuracy, precision) while others describe the things you do not (bias, variance).[3] Some have different statistical quantities associated with them and some decompose variation in measurement error in different ways. Note that "validity" and "accuracy" typically are used to encompass all kinds of errors, whereas the terms "reliability", "precision", "bias" and "variance" focus on particular types of errors.

There are a variety of names that have been given to ways one might assess validity, some of which we will return to below. There are similarly several common variations on the idea of reliability, which correspond to different senses in which you might measure a quantity multiple times. If the multiple measurements are conducted by different individuals (as in a manual coding exercise), it is typical to refer to "inter-rater reliability" or "inter-coder reliability". If the multiple measurements are conducted at different points in time, it is typical to refer to "test-retest reliability". If the multiple measurements are constructed using different methods of measurement, is is sometimes called "inter-method reliability", but note that failures of "inter-method reliability" can arise from failures of validity in the individual methods. [4] The terms validity and reliability do not have specific statistics associated with them, but a number of statistics can be used to capture these concepts in different contexts. These need to be considered separately for interval-level and nominal-level measures.

[3] The term "accuracy" is not always used consistently, but here I go with the ISO 5725 definition that requires both lack of bias and also precision, as opposed to only the former. Note that ISO uses yet another term, "trueness" to mean the opposite of bias, which I eschew here because it is lame.

[4] There are further senses of reliability. Measures of internal consistency among items in a multi-item test or index, such as Chronbach's Alpha (Cronbach, 1951), are also called measures of reliability. These are covered in Chapter 9.

3.3.1 Interval-Level Measures

The terms "validity" and "accuracy" are usually associated with measures like mean absolute error (MAE), mean square error (MSE) or root mean square error (RMSE).

$$MAE\,(m|\mu) = \frac{1}{n}\sum_{i=1}^{n}|m_i - \mu_i| \tag{3.7}$$

$$MSE\,(m|\mu) = \frac{1}{n}\sum_{i=1}^{n}(m_i - \mu_i)^2 \tag{3.8}$$

$$RMSE\,(m|\mu) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \mu_i)^2} \tag{3.9}$$

Note that these are measures of "invalidity" or "inaccuracy" because higher values of MAE, MSE or RMSE correspond to lower validity and lower accuracy. These can all also be viewed as measures of (in)accuracy or of (in)validity because they capture the idea of how well you have measured the thing you meant to measure (μ). Note that since we usually do not actually know μ, we may not be able to calculate these in a given application, but they are nonetheless the quantity we would like to know in order to assess accuracy/validity. Note also that they are on (or derived from, in the case of MSE) the scale of μ, so the numerical values need to be interpreted in the context of the actual levels of μ and their variation.

Bias and variance are widely applied statistical concepts and provide a useful decomposition of mean square error into two components. If we start with the formula for the mean square error, we can show that this equal to the sum of the bias (squared) and the variance of the measurement.[5]

$$MSE\,(m|\mu) = \frac{1}{n}\sum_{i=1}^{n}(m_i - \mu_i)^2 \tag{3.10}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m}_i + \bar{m}_i - \mu_i)^2 \tag{3.11}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m}_i)^2 + 2(m_i - \bar{m}_i)(\bar{m}_i - \mu_i) + (\bar{m}_i - \mu_i)^2 \tag{3.12}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m}_i)^2 + (\bar{m}_i - \mu_i)^2 \tag{3.13}$$

$$= \frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m}_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(\bar{m}_i - \mu_i)^2 \tag{3.14}$$

$$= Var\,(m_i) + Bias\,(m_i|\mu_i)^2 \tag{3.15}$$

Here, the variance of the measure m is with respect to its mean $\bar{m}$ and the bias is the difference between that mean and the target concept μ. The bias-variance decomposition of MSE is particularly useful when thinking about

[5] The key step in the derivation is to observe that the quantity $\frac{1}{n}\sum_{i=1}^{n}(2(m_i - \bar{m}_i)(\bar{m}_i - \mu_i))$ is zero because the mean value of the measure m_i is defined such that the mean deviation

$$\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m}_i) = 0$$

.

measurement error because measurement variance is one way of representing the concept of *reliability* mathematically.[6] One can think of inverse validity (MSE) was being the sum of (squared) bias and reliability (variance). The final one of the six terms in the table earlier, *precision*, is statistically defined as the inverse of variance, so those terms already map directly on to each other.

Thus, with the statistical quantities of mean square error (MSE), bias, and variance, we can quantify all six of the terms–validity, reliability; accuracy, precision; bias, variance–that are widely used to describe the proximity of a measure m to a target quantity μ. Validity and accuracy can be quantified by mean square error with respect to the target, reliability and variance using the variance of the measure with respect to its mean, precision with the inverse of that, and bias with the expected value of the measure relative to the target. While other statistical quantities are used in various contexts, in this text I will largely stick to these quantities.

[6] But note that this conflates the cross-unit and within-unit variability of the measurement strategy, the former of which may sometimes be termed as a validity problem rather than a reliability problem. There are unit-specific biases which have some variability across the set of units (and which have mean zero) as well as variation across repeated measurements of each individual unit (which also have mean zero). Unless you have repeated measures of individual units, the unit-specific biases will be indistinguishable from the repeated measure variability.

3.3.2 Nominal-Level Measures

For nominal level measures, measured values m_i either match the target concept μ_i (which we define as the $\epsilon_{\mu_i} = 0$ case) or they do not (the $\epsilon_{\mu_i} = 1$ case). Given this, the most obvious measure of overall validity is the proportion correctly classified (*PCC*):

$$PCC\,(m|\mu) \quad = \quad 1 - \frac{1}{n}\sum_{i=1}^{n} \epsilon_{\mu_i} \tag{3.16}$$

$$\tag{3.17}$$

As in the interval-level case described in the last section, this is necessarily defined with respect to a given population of units. A measurement strategy that has high validity on one set of units may not have high validity on other sets of units.

Once again, where validity is a question of whether our measurement strategy will return the *right* answer, reliability simply asks if our measurement strategy will consistently return the *same* answer. To think about reliability, we have to think about repeatedly trying to measure the same quantity using the same measurement strategy. Imagine that we are able to apply the same measurement strategy to each unit twice, yielding m_{ij} for $j \in 1, 2,$. Now, instead of asking the proportion of m correctly classified versus μ, we are interested in the proportion of m consistently classified across the two measurements. Again, the simplest and most obvious statistic is the proportion of cases in which $m_{i1} = m_{i2}$. If $j = 1$ and $j = 2$ are different coders of the data, this is called *inter-coder reliability*. If $j = 1$ and $j = 2$ are measurements at two different moments in time, this is called *test-retest reliability*. In either instance, this is again defined as an average across a set of measured units.

There are a variety of more complicated statistics for assessing inter-rater reliability, such as Cohen's kappa and Krippendorf's alpha, which I do not

cover in detail here. These statistics are all based on the observation that the basic agreement of two measures, whether by two coders or at two moments in time, is quite sensitive to the number of categories of the nominal-level variable and the relative frequencies of those categories. Thus, such statistics attempt to adjust for this baseline level of agreement "by chance" in order to make comparisons across different variables with different distributions meaningful.

3.4 Information and Calibration

In addition to these properties of the unconditional distribution of ϵ_m, it is sometimes important to consider the conditional relationships of ϵ_m to μ. How does ϵ_m vary across different levels of μ? There are two important questions about the quality of a measure that are related to this. First, are the measurement errors systematically different at different levels of the target concept? That is, how well well *calibrated* is m as a measure of μ? Second, how big are the measurement errors relative to the variation in the target concept? That is, how much *information* about the variation in μ is contained in m? Once again, these questions apply somewhat differently to interval-level (and ratio-level) measures as opposed to nominal-level (or ordinal-level), so I will take the different levels of measurement in turn.

3.4.1 Interval-Level Measures

Miscalibration of a measure occurs when a one unit change in m is not associated with a one unit change in μ. Figure 3.3 illustrates what miscalibration looks like. Panel a shows a case where the measure varies over a smaller range that the underlying concept, and therefore the measurement error in panel b is correlated with the target concept μ. Note that the measurement error has an average of about zero in this case, so this measure is not biased overall, but it is biased conditional on most values of μ because for most values of μ the average value of m is higher or lower than μ. Note that some pragmatically defined measures do not have well defined numerical scales absent the measure itself, and therefore it can make more sense to think of them as *uncalibrated* than *miscalibrated*.

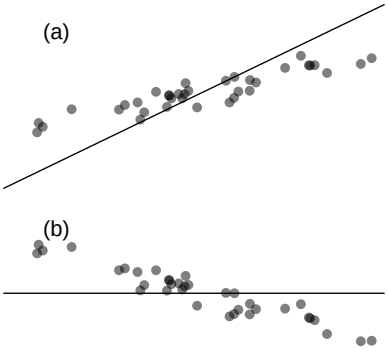


Figure 3.3: Panel (a) shows a set of measurements m_i (y-axis) that are miscalibrated with respect to the underlying target quantity μ_i (x-axis), Panel (b) shows the measurement errors ϵ_m as a function of the target quantity for the same measurements.

If we want to assess whether an interval-level measure is miscalibrated (and we have some known values of μ for at least some units) is to run the simple linear regression of m on μ and test whether the coefficient $\beta = 1$. Are the data consistent with the hypothesis that, in the broader population of applications of this measurement strategy to units like those observed, a one unit change in μ is associated with a one unit change in m? Note that this does not necessarily mean the measure is unbiased or that it has low variance. It could be that m is on average greater or less than μ, which shift α in the regression. It could also be that the standard deviation of the regression residuals σ is large or small,

which would be associated with the variability of the measurement strategy.

The other question we often want to ask about the quality of a measure is how much *information* is contained in the measure itself about the underlying quantity that we wanted to measure. For interval-level quantities, this can be captured in various ways, the simplest of which is the (hopefully familiar) Pearson correlation coefficient:

$$\rho = cor(m, \mu) = \frac{cov(m, \mu)}{sd(m)sd(\mu)}$$

.

Higher correlations imply stronger relationships between m and μ.[7]

The correlation coefficient is closely related to the familiar $R^2 = \rho^2$ statistic which describes the proportion of variation in m explained by variation in μ (and vice versa). This proportion of explained variation is a particularly useful quantity for thinking about the quality of a measure for understanding variation in the underlying concept. For example, if $\rho = \sqrt{0.5} = 0.707$, then $R^2 = 0.5$, which is to say that *half* the variation in the target concept is explained/captured/predicted by the measure.

What qualifies as a good correlation r and proportion of variance explained R^2 between the true value of μ and the measure m? Chapters 4 and 5 further develop some of the possible consequences of measurement error, but it is difficult to give a generic answer to how much measurement error is ok (and thus, how low $\rho(m, \mu)$ can be). An alternative non-parametric correlation statistic, Kendall's tau statistic τ is perhaps more useful than ρ for providing intuition in this case. Kendall's tau is (in this application) the proportion of pairwise comparisons of m_i with $m_{i'}$ for two units i and i' which are in the same direction as the pairwise comparisons of the underlying μ_i with $\mu_{i'}$.

[7] Note that it is possible to have a strong correlation of m and μ in cases where m has a significant bias with respect to μ. To take an extreme example, imagine that $m_i = \mu_i + a$ for some constant value a for all units i. This means that the measure is always "too high" by a, but the correlation between the measures will be 1.

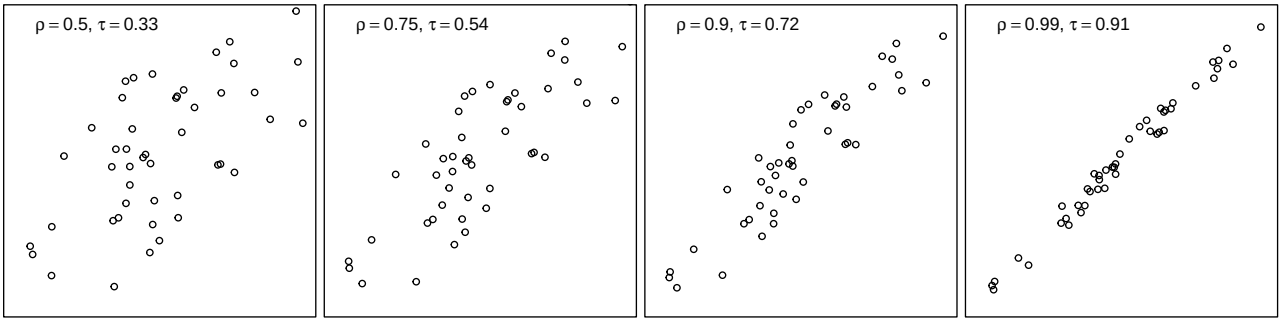


Figure 3.4: In the context of measurement, only numerically high correlations ρ between the measure (y-axis) and the target concept (x-axis) yield a high probability τ that pairs of units are correctly ordered.

Figure 3.4 shows a set of examples of ρ and τ, given a multivariate normal distribution of m and μ. This figure highlights that very substantial Pearson correlation coefficients are required to achieve high proportions of correct pairwise comparisons with the respect to the underlying concept of interest μ. Correctly ordering the units, being able to reliably determine which has more or less of the target concept of interest, is of course only one application of a

measure. However, along with R^2, correct ordering is one of the few that can be assessed without reference to any other variables. To say anything more generally about what is a large and what is small amount of measurement error, we need to think more about the kinds analyses that we might conduct *using* variables measured with error, and the relationship of that error to the other variables X that enter into those analyses.

Both the correlation coefficient and the linear regression coefficient are useful tools for thinking about measurement error in interval-level measures. Of course the relationship between m and μ need not actually be linear. Sometimes measurement error has more complicated relationships with the value of the target concept μ, as we will see in the example in Chapter 5.5.

3.4.2 Nominal-Level Measures

For nominal-level measures, assessing the conditional relationships of ϵ_m to μ requires focusing on the particular possible values of μ. This requires tools that are appropriately adapted to the structure of errors that are possible when a variable can only take on a limited set of values. With binary quantities $\mu \in 0, 1$, for either true value of μ there is one correct value of m and one incorrect value of m. If $\mu_i = 0$, then $m_i = 1$ is a *false positive*; if $\mu_i = 1$, then $m_i = 0$ is a *false negative*. Most of my focus here will be on the binary case, but the key ideas can be generalised to nominal variables with more than two categories.

Despite the apparently simplicity of having only two possible kinds of errors, there are a very large number of ways to describe errors in binary variables. In many applications, the rate of errors when $\mu = 0$ is very different than the rate of errors when $\mu = 1$, because one of these is far more common than the other. One of these errors may be far more important than the other substantively as well, eg we might be more worried about failing to detect a disease than about falsely diagnosing one, or we might be more worried about imprisoning the innocent than about failing to imprison the guilty.

Given that these kinds of asymmetries are quite common when we want to measure binary quantities, there is a terminology for distinguishing the rate of false positives separately from the rate of false negatives. First, we define the types of errors and non-errors that are possible in a binary variable:

	$m = 0$	$m = 1$
$\mu = 0$	true negative	false positive
$\mu = 1$	false negative	true positive

The above 2x2 matrix of possibilities is often called the *confusion matrix*. The *accuracy* of a binary measure is simply the proportion of cases in which $m = \mu$:

$$\frac{p(\text{true positive}) + p(\text{true negative})}{p(\text{true positive}) + p(\text{true negative}) + p(\text{false positive}) + p(\text{false negative})}$$

The quantities *sensitivity* (Se) and *specificity* (Sp) are defined as follows:

- Sensitivity is the *true positive rate*, the proportion of the cases where $\mu = 1$ for which $m = 1$:

$$\frac{p(\text{true positive})}{p(\text{true positive}) + p(\text{false negative})}$$

- Specificity is the *true negative rate*, the proportion of the cases where $\mu = 0$ for which $m = 0$:

$$\frac{p(\text{true negative})}{p(\text{true negative}) + p(\text{false positive})}$$

The best possible value for both of these is 1. Note that you can always achieve this for either sensitivity or specificity at the expense of the other, by setting $m = 1$ for all units (achieving perfect sensitivity) or $m = 0$ for all units (achieving perfect specificity). This obviously does not make for a good measure though, as such a measure carries no information at all about μ!

Both of these quantities *condition* on the true value μ: they ask questions about the proportion of correctly measured values m among units with a given true value μ. There are further statistics that we can calculate which condition on m instead of μ. The corresponding quantities are usually called *positive predictive value* (PPV) and *negative predictive value* (NPV):

- Positive predictive value is the proportion of the cases where $m = 1$ for which $\mu = 1$:

$$\frac{p(\text{true positive})}{p(\text{true positive}) + p(\text{false positive})}$$

- Negative predictive value is the proportion of the cases where $m = 0$ for which $\mu = 0$:

$$\frac{p(\text{true negative})}{p(\text{true negative}) + p(\text{false negative})}$$

Whereas sensitivity and specificity describe the rate of (non-)errors among units with a common true value, positive and negative predictive value describe the rate of (non-)errors among units with a common measured value. Positive predictive value asks "of the units I measured as $m = 1$, how many of them really have $\mu = 1$?", as opposed to sensitivity, which asks "of the units that really have $\mu = 1$, how many of them will I measure as $m = 1$?" Both of these types of questions are potentially relevant to thinking about a given measurement problem, a point we will develop further in Chapter 4.

There are further statistics that aim to describe the overall information content of a binary measure, which is to say the strength of the relationship between the measure and the target. You can think of these as being somewhat

analogous to the correlation coefficient in the interval level case we considered in the previous section. One of the simpler ones to understand is the "Diagnostic Odds Ratio", which is defined as:

$$DOR = \frac{\frac{p(\text{true positive})}{p(\text{false positive})}}{\frac{p(\text{false negative})}{p(\text{true negative})}}$$

This quantity can be understood as telling you how much more you should believe $\mu = 1$ after observing a measured value $m = 1$ by comparison to what you would have believed had you observed $m = 0$. If the value of the $DOR = 1$, there is no information content to the measure; if observing the measure perfectly resolves (at least one of) whether $\mu = 0$ or $\mu = 1$, then the $DOR = \infty$. If the DOR is less than 1, that indicates that the measure provides a negative signal of μ, where observing $m = 1$ makes it more likely that $\mu = 0$ and observing $m = 0$ makes it more likely that $\mu = 1$. This is obviously a very bad property for a measure to have, it is equivalent to the continuous measure case above where the correlation between m and μ is negative. As in that case, such situations are often easily fixed by redefining the measurement strategy in a simple way that reverses its sign.

3.5 Assessing Validity

Which approaches to assessing validity are possible varies from application to application. There is a bewildering menagerie of overlapping terminology around the concept of validity with respect to measurement. These terms have largely been developed in the field of psychology/psychometrics, which has an especially rich tradition of developing new concepts and then trying to measure them through survey-based batteries of questions. This then leads to a number of overlapping approaches to assessing validity. I have made an effort here to translate these each into the terminology and notation we are using here.

- Concurrent validity: a measure m_1 correlates well with a previously validated measure m_2 for the same target concept μ.
- Construct validity: an all-things-considered assessment of the adequacy of m as a measure of the target concept μ.
- Content validity: the extent to which a measure m reflects "all aspects" of the target concept μ.
- Convergent validity: the extent to which m_1 measuring μ_1 is correlated with m_2 measuring μ_2 in a case where we expect μ_1 and μ_2 to be highly correlated.
- Criterion validity: the extent to which m is correlated with some quantity x (which may or may not be another measure of the target concept) that we expect is correlated with μ.

- Discriminant validity: the extent to which m_1 measuring μ_1 is *not* correlated with m_2 measuring μ_2 in a case where we expect μ_1 and μ_2 to *not be* correlated.
- Face validity: the extent to which m is subjectively perceived to be measuring μ by an observer or by an expert.
- Predictive validity: a measure m correlates well with some x (which is not another measure of the target concept) that we expect is correlated with μ.

If you read these carefully, you will note that some of these are more precisely defined and some correspond to specific approaches to assessing validity while others are stated at the level of principles. These do not necessarily go together though: face validity may appear to be the least precisely defined, but in Chapter 9.2.5 we will see a strategy for using pairwise comparisons of units to not merely validate, but actually construct an index measure. Subjective perceptions may vary, but they are themselves measurable: you just need to find someone (ideally a relevant expert) to make some evaluations of units with respect to a concept of interest. Many of these terms overlap: criterion validity includes concurrent, convergent and predictive validity. I will generally limit my use of these terms, because I think they are easily confused and I would rather just talk about ms and μs and the structure of the errors ϵ_m between them. All approaches to assessing validity are about trying to understand this relationship.

From the perspective of the analyst, the best, non-trivial,[8] cases for validation are those where you know the right answer μ exactly for a representative subset of the units you ultimately want to measure. In practice, the benchmarks for validation that are available are often far more limited than this. There are two broad cases. First, there are cases where there is some benchmarking data available, either μ itself for some units, or another measure m for some units, where those units are or are not representative of the units you ultimately wish to measure. Second, there are cases where there is no benchmarking data available, at least not on the scale of the quantity μ that you are trying to measure.

[8] The trivial case is that you know μ for *all* the units you want to measure, in which case you have already solved the measurement problem.

3.5.1 With representative benchmark data

The best case for testing validity is that you know μ for a representative sample of units. In these cases, you can simply calculate the kinds of quantities described in the preceding sections of this chapter, within your benchmark sample, and have a reasonable expectation that these are representative of the broader population to which you will apply the measurement strategy. In the continuous cases you can calculate the RMSE of m versus μ, the correlation of m with μ, and the unconditional $E[m] - E[\mu]$ and conditional biases $E[m|\mu]$. In binary/categorical cases, you can directly calculate true negative, false negative, true positive and false positive rates, as well as the quantities that derive from them. These will all be sample estimates, whose precision depends on the size of the benchmark sample, but they will in expectation give you an accu-

rate picture of how the measurement strategy will perform across the entire population.

Or rather these calculations will give you an unbiased assessment of the measurement strategy if the benchmark data was *not* involved in any way in selecting or optimising the measurement strategy. If, however, you have used the benchmark data to choose between multiple measurement strategies, or to calibrate, optimise or estimate aspects of the measurement strategy (see, eg, Chapters 8 and 10), they you need to worry about *over-fitting*. That is, if you chose the measurement strategy that you use on the basis of it having a low measurement error versus alternatives (selection) in the benchmark data, or you chose features of the measurement strategy itself to minimise measurement error in the benchmark data (calibration/optimisation/estimation), then the magnitude of measurement error in the benchmark data sample will generally be an underestimate of the measurement error in the population to which you want to apply the measurement strategy.

The principle strategies for accounting for over-fitting in other applications apply to measurement applications as well. If you want a realistic assessment of your likely patterns and magnitudes of measurement error "out-of-sample", that is to say in the population as opposed to the sample for which you have the benchmark data on μ, you need to do any measurement strategy selection and calibration/optimisation/estimation using subsets of the benchmark data and then test its performance on the residual subset that was not used. There are a number of strategies for doing this, by dividing the benchmark data into a training set and a test set or by applying cross-validation. The mathematical details and practicalities of these are covered in detail in other sources, the important thing to note here is where the need to apply these arises in measurement applications.

The core question you need to ask is whether the benchmark data you are using in any way informed the selection of a measurement strategy or was used to optimise/calibrate/estimate aspects of the measurement strategy. If it was, measurement error in the benchmark data ceases to be representative of measurement error in the broader population of units, because you have selected for a measurement strategy that makes the measurement errors in the benchmark data small. The hope, of course, is that this will make measurement errors in the broader population you wish to study small. But there is a risk that it will simply be over-fitting the idiosyncrasies of the benchmark data, particularly when the benchmark data set is small and/or the measurement strategy has many degrees of freedom / parameters over which you have optimised performance for that benchmark data.

3.5.2 With unrepresentative benchmark data

Additional considerations arise when you are attempting to assess validity with benchmark data which is unrepresentative of the population to which you

want to apply a measurement strategy. Are the errors larger for certain kinds of units than for others? What does this imply about the measurement errors when applied to the population?

One way to assess the potential for non-homogeneous measurement error is to model the error structure as a function of observable characteristics of units among the benchmark data that you do have. Note that what exactly it would make sense to do here will depend on the problem and the ultimate application, so it is difficult to provide general advice. In most applications, one would be primarily looking for evidence of a gross mismatch between the errors in the kinds of units that are over-represented in the benchmark set versus those that are under-represented. For example, you could have a situation where the measurement error ϵ_m has a different mean value in the over-represented types of units by comparison to the under-represented types of units.

Of course, as with any exercise in re-weighting, you can only assess the range of units where under-representation in the benchmark sample is finite. If there are types of units which are entirely missing from your benchmark data, and you lack credible evidence that the error patterns observed in the benchmark data carry over to those types, for those unrepresented unit types you are effectively in the situation discussed below where you lack benchmark data at all.

3.5.3 With imperfect benchmark data

Another common case is that your benchmark data itself has some (unknown) error with respect to the concept of interest. That is, you are attempting to validate m using m^* rather than using μ. As should be obvious, how useful m^* is for validating m depends on whether the measurement errors in the benchmark ϵ_{m^*} tend to be larger or smaller than the measurement errors ϵ_m in the new measure you are trying to validate. You will generally not have direct data on this question, but may have some reasonable expectations. For benchmarking against m^* to be useful, you generally need the ϵ_{m^*} to be smaller than the ϵ_m, and also to not have problematic biases for your particular problem (eg correlation with your other quantities of interest).

3.5.4 Without benchmark data

For many measurement projects, there simply is not an existing measure for the concept of interest. Indeed, this is necessarily the case the first time someone tries to measure something new. Part of the reason why there has been the proliferation of validity concepts listed above is that there are a number of possible ways one can begin to validate in the absence of benchmark data on the target concept. "Criterion", "concurrent", "convergent" and "predictive" validity are all variations on the idea that you might look at whether m is correlated with other variables that you expect to be correlated with μ. "Face" validity is

the idea that you might just look at some of the values produced by the measurement strategy (in a more or less structured way) and decide if they make sense. If you do not have benchmark data for μ, you need to either look to other variables you do have that you think might be a reasonable proxy, or you need to trust your intuition as to what "looks right". Since one does not know what the right magnitudes of correlations with other variables are ex ante, even the criterion/concurrent/convergent/predictive validation methods ultimately come to the same "does it look right?" standard as face validation methods.

This all sounds very nebulous and unscientific. While it is certainly the former, science often begins with nebulous hunches rather than following some strict procedure that guarantees sound results. There is no general method for validating a new measure of a previously unmeasured concept.

Face validation can be given a formal structure that mitigates some of these issues. For a continuous (interval or ratio-level) concept, you randomly select pairs of units for which a new measure has been generated, and present the units (and potentially relevant information about them) to a subject area expert who is ideally not the person generating the measure itself.[^classificationvalidation] For each pair of units, you ask the expert for the relative ranking of the two units with respect to the concept of interest. The sign of the difference between the measures for these units can then be compared to this expert subjective assessment benchmark. If the measurement has high validity for the concept of interest, the differences in the measures will generally have the same sign as the expert's subjective assessments. Of course one cannot expect to achieve a perfect match here, and the structure of the task may change the results radically. If you are measuring properties of countries, you might assume the experts already know the units and provide no further information: "does country X or country Y have a higher level of this concept?" But if you are instead trying to measure properties of large numbers of individuals who the expert will not already know, you need to provide information about the units, including at least the underlying indicator data that went into the measurement strategy itself.[9]

For a categorical (nominal or ordinal) concept, the experts can often provide direct codings of individual units. If these are randomly selected from the population of interest, one can use the expert codings of a random subset to calculate the various categorical variable validity statistics (true positives, etc) described earlier in this chapter. As with the pairwise comparison approach for continuous concepts, the limitation of this approach is how much measurement error there is in the experts' evaluations, which is itself not typically known quantitatively. That is, the expert codings are just another m^*, they are not μ. Experts will invariably disagree with one another and have different conceptions of the concept μ. Nonetheless, validating against a m^* is potentially better than doing no validation at all, and since the concepts that social scientists are interested in measuring are often invented by social scientists themselves, validating against expert understanding of those concepts is at least

[9] For these latter cases, in Chapter @**??**floridi-lauderdale), I describe a method for structuring an exercise like this to generate a measure from indicators using experts to calibrate the contribution of the different indicators. This switches the role of the experts from providing validation for a measure to providing evidence that forms part of the construction of the measure, even though the pairwise comparison task can be structured in the same way.

a check for consistency.

3.6 Assessing Reliability

Recall that the reliability of a measurement is the extent to which repeating a measurement procedure generates stable, as opposed to unstable, measured values for a given unit. As noted earlier, some usages of the term "reliability"—particularly those that compare measures across different measurement methods for the same quantity of interest or across measures constructed from difference subsets of indicators in a multi-indicator measurement strategy—are probably better understood as indirect tests of validity. Here, I will focus on practical strategies for assessing stability across repeated applications of the same procedure.

Perhaps obviously, there are two relevant cases to consider: those where it is possible to repeat the measurement procedure and those where it is not possible to repeat the measurement procedure. Measures that involve any sampling procedures are unreliable in the sense that if you resampled, you would get different measures for the quantities of interest. These include be area measures constructed from surveys of individuals, inflation measures constructed from surveys of goods, and measures of a wide variety of social science concepts constructed from expert coding of cases. In contrast, measures that are constructed from fixed indicators like official national statistics can be perfectly reliable in the sense that if you followed the procedure you would end up with the same number exactly.

To assess reliability in cases where repeated measures can be constructed, one can look at the same sorts of statistics that are used to assess validity, but comparing m_1 and m_2 rather than m and μ. In cases where repeated measures are not available, there is an important conceptual distinction to be drawn between instances where the repeat measures are practically unavailable, but would in fact be different from the observed measures if you applied the same procedure again, versus instances where applying the same procedure would in fact generate the same numerical values because all the inputs are fixed.

Perfect reliability isn't necessarily a good thing, unless it comes with high levels of validity in the sense of proximity to the target concept μ. Holding constant the overall level of validity of a single measure (eg RMSE of m versus μ), it is better if the source of the lack of validity is a lack of reliability, as you then have the prospect of generating more valid measure simply by collecting more data (eg some sort of larger sample). If you literally have multiple measures, each constructed in the same way, pooling multiple realisations of an unrealiable measurement strategy will typically be an improvement on single realisations of the measurement strategy. For continuous measures, this pooling typically involves taking the mean; for categorical measures this may involve taking the modal categorisation.[10] Often there is a feasibility or expense constraint on pooling evidence from repeated measurements into a

[10] See Chapter 10 for a more extensive discussion of using point classifications versus probabilistic classifications in subsequent analyses.

better synthetic measure, but where a measurement strategy generates different values on repetition, the long run average of these is nearly guaranteed to have better properties than a single realisation, and so the closer you can get to that long run average, the better.

A lack of reliability is a lesser problem by comparison to a lack of validity. Or rather, the thing you care about is the relationship between m and μ, which is to say ϵ_i, and how it covaries with the true values of μ and the other quantities relevant to your application. Whether you could obtain different values of m for the same units may be relevant (or even useful) in particular problems, but this is only important for subsequent analyses using m if you are in fact going to conduct multiple measurements. If you have a single measurement that you are going to use in an analysis, all that matters is the validity of that measure, which is to say the properties of the ϵ_i in the measures you actually have.

3.7 When is a Measure Not Good Enough to Use?

> "For we dare not make ourselves of the number, or compare ourselves with some that commend themselves: but they measuring themselves by themselves, and comparing themselves among themselves, are not wise." 2 Corinthians 10:12, King James Bible

> "The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data." (Tukey, 1986, p74-5)

One sometimes hears the suggestion that there are some things that we should not try to measure, or more specifically that there are some things we should not try to *quantify*, because we cannot do so well enough. This is often a difficult claim to produce evidence for or against. Sometimes it is based on concerns about unintended consequences of the use of the measure, as discussed in Chapter 1. However sometimes it is simply a statement that we will generate misleading claims about reality if we do so. There are clearly some measures that should not be used for some purposes. We can trivially generate examples, such as using the average income of a local area as a measure of the average education level of that area or quality of life for those who live there. Income is probably correlated with these things, but the measurement errors are very likely to have associations that lead us to errant conclusions about the target concepts (education levels, quality of life). This might seem obvious in this instance, but many a research project has used measures that would probably have worse properties for measuring the concepts they purported to measure than the above examples. The question is how you identify when there is a problem in applying a given measurement strategy to a given prospective problem.

One way to come at this problem of when measures are not good enough to use, is to ask what are the ways that measures might fall short of support-

ing different applications. I have identified three general ways that we might overclaim regarding the quality of measurements, all of which we should aim to avoid:

1. We should not overclaim regarding levels of measurement. We should not present measures that are ordinal (providing information about ranking) as though they provide interval- or ratio-level information about magnitudes of difference.
2. We should not overclaim regarding the validity of measures. We should present information about the known magnitude and patterns of measurement error wherever available, and where this information cannot be constructed we should clearly describe the potential risks of using measures with patterns of error that we cannot quantify.
3. We should not overclaim the domain over which our measures are useful. When describing measures we should describe them in terms close to the measure we constructed rather than the most general concept that we could plausibly associate our measure with.

These each correspond to an essential concept that we have considered: level of measurement of m, the validity of m as a measure of μ, and the definition of the concept μ for a given measure m. You need to make sure that the actual level of measurement you have supports the kind of claim you want to make; you need to make sure the actual measurement you has a relationship with the target concept that is sufficiently good to support the kind of claims you want to make; you need to make sure that the language you use to describe the target concept does not inflate the distance of that concept from the measure you actually have.

In the next two chapters, we will explore ways that measurement error can create problems in applications. These problems arise from working with measures that do not have perfect validity, where $m \neq \mu$. These are roughly divided into what we might call "normative" problems in Chapter 4 and what we might call "empirical" problems in Chapter 5.

4
Fairness in Measurement

If a concept is measured with error, some units will usually be measured better than others. Usually some will have larger errors and others smaller errors; usually some will have positive errors and others negative errors. Were we measuring the diameters of ball bearings we would not worry about whether we were being fair to the ball bearings for which the measurement error was negative versus the ball bearings for which the measurement error was positive. However, given that we are engaged in social measurement, and the units we are measuring are people, groups, countries, etc, we do have to think about fairness in how they are being treated by any given measurement procedure.

The question of how measurement error $\epsilon_m = m - \mu$ relates to other variables X is particularly acute where those other variables specify membership in legally protected groups or in otherwise sensitive groups of units. There are a very large number of fairness criteria that scholars have applied in various contexts, most of which are statistically equivalent, tabulated by Barocas et al. (2020).[1] It is intuitive that if measurement error in an individual-level measure is correlated with some X like race, ethnicity, sex or gender, the use of that measure might lead to biased decisions with respect to individuals of different races, ethnicities, sexes or genders. It turns out that the challenges are more fundamental than this: the very existence of any measurement error at all creates basic challenges for the fair treatment of units, if we are to use their measured values for any purpose.

[1] Chapter 2 of Barocas et al. (2020) provides a more detailed mathematical treatment of the issues covered in this section. See also Kleinberg et al. (2016).

4.1 Separation and Sufficiency

When we look at a measure m, and we see differences as a function of some other variables X, we do not know if those differences are due to "real" differences in μ between those groups or different measurement errors for those groups. For this reason, looking at group differences in the distribution $p\left(m|X\right)$ of measures m cannot tell us much about whether the measurements treat different groups X fairly. If we see that one group has higher values of a measure than another, it could reflect a fairness problem with the measurement (a difference in the distribution of ϵ_m for different groups) but it could

also reflect a fairness problem with reality (a difference in the distribution of μ for different groups). In order to meaningfully define what fairness means with respect to *measurement*, we need to simultaneously think about the measure m, the underlying concept that we aimed to measure μ, and the groupings X with respect to which we have potential concerns about fairness.

Ideally we would like any measurement to be independent of X, conditional on the true value of the concept of interest μ:

$$p\left(m|\mu, X\right) = p\left(m|\mu\right)$$

In the machine learning literature, this conditional independence condition is called "separation". This criterion says that, among units with the same true value of the target concept μ, we want the distribution of the measurement m (and thus the measurement error ϵ_m) to be identical for all levels of X. So, for example, among students with the same understanding of the material in a course, we want their distribution of final marks (the measurement generated by the assessment procedure) to not depend on whether they are men or women, or black or white, or a variety of other attributes X for which we are concerned about bias.[2]

There is another condition that we might want to satisfy, called "sufficiency". Sufficiency is (like separation) a conditional independence condition, but conditioning on the measure rather than the target concept. Sufficiency is met when the distribution of the true value of the concept of interest μ is independent of X, conditional on the measured value m:

$$p\left(\mu|m, X\right) = p\left(\mu|m\right)$$

Why is sufficiency a criterion that we would like a measure to satisfy? The idea is that, if we are going to use a measure m for some application or decision, we would like it to be the case that the distribution of the true values of the underlying concept that we wanted to measure are the same for individuals in different groups who have the same value of the measure. We would like it to be *sufficient* to know the measure m. Knowing X should convey no further information about the likely value of μ for a given unit once you know m. So, for example, if we are going to award First class degrees to all students who achieve an average of 70 across their modules, sufficiency says that the men and women who get a 70 (m) should have the same distribution of understanding of the course material (μ).

Thus, separation says that units (individuals) with the same value of the thing you wanted to measure have the same distribution of measures, while sufficiency says that units with the same value of the measures have the same distribution of the thing that you wanted to measure. These both seem like properties that we would want to satisfy, in order that a measurement strategy for generating m is fair to units in different groups X.

The bad news is that separation and sufficiency are in fundamental conflict. Separation and sufficiency can only both be satisfied if both the measure m and

[2] We are likely to be particularly concerned with the mean of the distribution ($E\left[m|\mu, X\right] = E\left[m|\mu\right]$). For example, if we are constructing an educational assessment, we do not want a situation where men tend to get higher average marks m than women who have the same underlying understanding of the material μ, or vice versa.

the concept of interest μ are independent of group X. Stated differently, if there is any difference in the distribution of the the true values μ of the concept that you wanted to measure between the groups defined by X and there is any measurement error in m, you cannot achieve both separation and sufficiency. The proof of this is beyond our scope here, but it is a straightforward implication of the laws of probability and the definitions of separation and sufficiency given above (Barocas et al., 2020).[3]

Given that this tension exists, any application must consider the question: if you can only satisfy one, do you want a measure that satisfies separation or one that satisfies sufficiency? Do you want measures to have the same distribution given equality of the underlying quantity of interest or do you want the same distribution of the underlying quantity of interest given measured equality? Does it matter who you are or what the circumstances are? In order to properly consider these questions, it is helpful to turn to some examples.

[3] The brief sketch of the proof fits in a footnote. If $p(\mu|m,X) = p(\mu|m)$ and $p(m|\mu,X) = p(m|\mu)$ then it is implied that $p(m,\mu|X) = p(m,\mu)$. That is, if both sufficiency and separation are satisfied, then the joint distribution of m and μ must not depend on X. Therefore, if the distribution of m and/or μ does depend on X, at least one of sufficiency or separation must not be satisfied.

4.2 Application - Predicting Recidivism

In 2016, there was a public controversy over a risk assessment tool used in the US to predict the risk that defendants being sentenced for criminal convictions would commit further crimes in the future. This tool for assessing "recidivism risk" is called "Correctional Offender Management Profiling for Alternative Sanctions" but is generally referred to as COMPAS. The COMPAS score, on a 1-10 scale, is based on age, sex and criminal history, but not race/ethnicity.

ProPublica, a nonprofit organisation which conducts investigative journalism "in the public interest", published a long article arguing that COMPAS was biased against black defendants because black defendants who ultimately did not reoffend ($\mu = 0$) had substantially higher risk scores on average than did white defendants who ultimately did not re-offend. Northpointe, the company that developed the software, argued in response that COMPAS was not biased against black defendants because, given the same risk score m by the software, black and white defendants were equally likely to re-offend. This is precisely the tension of fairness in measurement that was discussed earlier in this Chapter. ProPublica observed that COMPAS failed the separation test: given the true target concept μ (whether someone would ultimately re-offend), black and white defendants (X) had different distributions of the measure m. In response, Northpointe argued that COMPAS passed the sufficiency test: given the risk assessment measure m, black and white defendants had the same distribution of re-offending.

There was a particularly accessible exposition of the fundamental tension between these notions of fairness published by Corbett-Davies, Pierson, Feller and Goel in the Washington Post which my discussion below follows closely. The ProPublica article was based on a public records request in Broward County, Florida, and the data we will be analysing includes race, COMPAS scores and whether the defendant committed another offense within the fol-

lowing two years for 3175 black and 2103 white defendants.

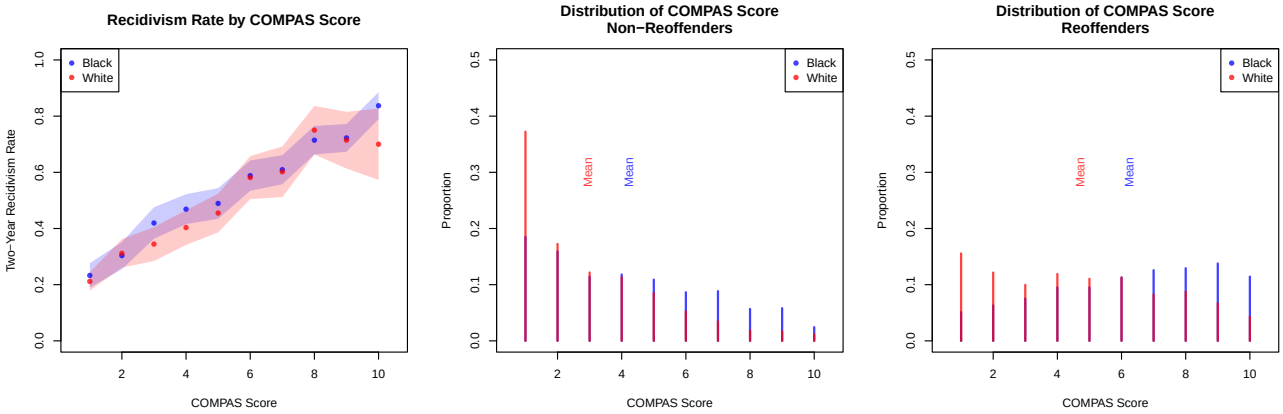


Figure 4.1: Recidivism by COMPAS Score (left), distribution of COMPAS Score among non-re-offenders (center), distribution of COMPAS Score among re-offenders (right).

We can see the relevant patterns in the data very easily. Figure 4.1 shows the arguments on both sides. The left panel shows that the two-year recidivism rate for black and white defendants is essentially the same, for each COMPAS score. There are a few minor deviations, but they are small and are mostly of the magnitude we would expect to see by chance given the number of observed defendants. We also can see from this plot that the COMPAS score is highly predictive of recidivism within two years: the recidivism rate rises roughly linearly from 22% at a COMPAS score of 1 to 81% at a COMPAS score of 10. This is Northpointe's case for the COMPAS score: a given score "means the same thing" for white and black defendants in terms of their future recidivism and it is highly predictive of their future recidivism.

ProPublica's case against the COMPAS score is in the center and right plots of Figure 4.1. If we look at just the defendants who did not re-offend, black defendants had higher COMPAS scores (center). The average COMPAS score among black non-re-offenders is 4.22 while among white non-re-offenders it is 2.94. It is also the case that among those who did re-offend, black defendants had higher COMPAS scores (right). The average COMPAS score among black re-offenders is 6.24 while among white re-offenders it is 4.72. It is indeed striking that the average COMPAS score for black non-re-offenders is closer to the average for white re-offenders than it is to that for white non-re-offenders. The average black non-reoffender is treated with greater suspicion of re-offending by the system than the average white non-reoffender.

These results are internally consistent because the rate of recidivism is substantially higher among the black defendants in these data. Black defendants had an overall recidivism rate of 0.52, versus 0.39 for white defendants. The average COMPAS scores for black defendants were higher as well 5.28, versus 3.64 The combination of the fact that the recidivism rate is different for the two groups with the fact that the COMPAS scores satisfy the sufficiency requirement of having equal recidivism rates given the COMPAS score guarantees

that the distribution of the scores must be different for black versus white non-offenders and/or for black versus white offenders.

When we think about fairness, we have to ask "fairness with respect to whom?" Fairness is a statement about equal treatment. People have a claim to be treated equally with those who are similarly situated with respect to relevant criteria (eg past and future criminal behaviour), in a way that does not depend on protected criteria (race, gender, etc). The risk score is meant to encode the information from the relevant criteria, at least the ones that are known at the time. If we are thinking about someone who is black, do we want to be fair by treating them the same as a white person with the same risk score or as a white person with the same ultimate recidivism? To what extent does the fact that the risk score is known at the point of decision, while the recidivism is not, push us towards one answer? To what extent does the fact that the recidivism is a real thing that the person will or will not do, while the score is a mere prediction, push us towards the other answer?

Corbett-Davies, Pierson, Feller and Goel write:

> It's hard to call a rule equitable if it does not meet Northpointe's notion of fairness. A risk score of seven for black defendants should mean the same thing as a score of seven for white defendants. Imagine if that were not so, and we systematically assigned whites higher risk scores than equally risky black defendants with the goal of mitigating ProPublica's criticism. We would consider that a violation of the fundamental tenet of equal treatment.
>
> But we should not disregard ProPublica's findings as an unfortunate but inevitable outcome. To the contrary, since classification errors here disproportionately affect black defendants, we have an obligation to explore alternative policies. For example, rather than using risk scores to determine which defendants must pay money bail, jurisdictions might consider ending bail requirements altogether — shifting to, say, electronic monitoring so that no one is unnecessarily jailed.

Lurking in the background of this case are several other normative considerations that make it difficult to focus solely on the measurement fairness question. One could take the view that predictions about future re-offending should have no role in sentencing, although they might be put to other uses.[4] One could take the view that because of systematic racism in US criminal justice, blacks are more likely to be charged with crimes in the future by comparison to whites with similar criminal behaviours, and so satisfying the sufficiency test is actually just reinforcing the systematic biases elsewhere in the system. While this is an unusually clear illustration of the tension between two notions of fairness when treated narrowly, there are clearly many issues of fairness and justice in its vicinity that might shape how we make the choice between different notions of fairness. Of course this is not unique to this particular example: fairness concerns regarding social measurement are particularly likely to arise in the context of broader questions of fairness and justice with respect to different groups.

[4] In a system more interested in rehabilitation, such scores might be used to target support to those more likely to otherwise re-offend, which would put these disparities in a different light.

4.3 Application - Predicted A-Level Grades

In England, Wales and Northern Ireland, the primary qualification for university entry are A Level exams. Students typically prepare for these subject-based exams over a two year period, with most students sitting the exams in May - June, receiving results in August, and entering university around the beginning of October. Because of this tight timing, university admissions has historically relied on *predicted grades* for each student, given by their teachers based on the coursework they have done in preparing for their exams. Universities then make *conditional offers* to students based on these predictions and other application materials. These offers are subject to the student achieving a set of A-level grades. These A level grade *conditions* are typically slightly lower than the predictions, for reasons that will become clear below. If the student achieves that grade level, they are admitted; if they do not, are sometimes still able to attend that university, sometimes have an offer from another university for which they have met the conditions, or may use the "Clearing" process in August to secure an available place at another university with lower entry requirements.

The pass grades for A Levels are A*, A, B, C, D and E. A typical student completes three A Level exams, and most university entry requirements are based on three A level results. For our purposes here, and roughly approximating how the exam scores are used, we will conduct our analysis in terms of *A level grade points*, which are calculated A* = 6, A = 5, B = 4, C = 3, D = 2, E = 1, non-passing grades = 0 and adding over the three exams to yield a total number of points between 0 and 18.

One way to think about the *predicted grades* set by school teachers for each student is that they are a *measure* of the student's ultimate *achieved grades* on their A level exams. Teachers are not able to perfectly predict A-level exam grades for their students, so these measures have measurement error. Analysis by Murphy and Wyness (2020) shows that there are several important systematic differences between the A level grade points the students achieve and the predictions that their teachers set. In the years 2013-15, 16% of students achieved exactly their predicted grade points, 9% achieved more than their prediction, and 75% under-achieved their predictions. On average, predictions are 1.7 grade points higher than the results that students achieve (Murphy and Wyness, 2020).

Our question here is whether these under-predictions and over-predictions are systematically associated with other attributes of students. If teachers are, on average, too optimistic about student performance that is not a fairness problem, even though it is a bias in the predicted grades as a measure of achieved grades.[5] A constant bias that applies to everyone is not a fairness problem. What could be a fairness problem though, is if teachers are over-predicting grades for some kinds of students more than for other kinds of students. Concerns about these kinds of biases are part of an ongoing discus-

[5] In a world where most teachers make predictions that are too high, there is little incentive for a teacher to recalibrate their predicted grades to be more realistic as it will only disadvantage their students in admissions.

sion in the UK about whether predicted grades should be used in admissions, whether some schools have policies of intentionally overstating their predictions, and whether already advantaged students are able to haggle with their teachers for better predictions.

In 2020, the coronavirus pandemic led to the cancellation of all A level exams. The UK government initially decided that the exams regulators[6] should award "exam" grades based on teacher rankings of their students, mapped onto to the past performance distribution at that school. Because of the persistent discrepancy between predicted grades and achieved grades, this meant that a very large proportion of students received grades that were lower than the predictions they had received for purposes of university admissions. This meant that they failed to meet the conditions of their offers and lost their places at their preferred universities. This is what happens in a normal year, and to a similar number of students, but in 2020 the students had not actually had exams, and so they had no agency in their "failure". This algorithm (the 2020 measurement procedure) clearly lacked the public legitimacy that the exams (the pre-2020 measurement procedure) had for assessing student achievement and preparation for university. The political fury was such that the government reversed course four days later and instructed the exams regulator to give all students the better of these simulated grades and their predicted grades. This meant that many more students met their conditional offers than in a normal year, and some programmes ended up with far more first year students than expected.[7]

[6] This process happened in parallel, with four different exams regulators, in England, Scotland, Wales and Northern Ireland, with slightly different details.

[7] My department at UCL ended up with about an extra 100 first year undergraduate students versus what we would have under either the original algorithm or a typical year's pattern of performance.

Our focus here will be on the predicted grades and how they relate to achieved grades in a typical year, as this is both interesting in itself and also structured why the 2020 exam debacle took the form that it did. Some groups of students received greater downgrades between their predicted grades and the grades awarded by the algorithm because those groups had tended to receive overstated predictions to a greater extent in previous years. While the algorithm did not explicitly use factors like race/ethnicity or school type in its calculations, the fact that it was conducted at the school level meant that typical patterns of overstatement with respect to these factors were apparent in which students tended to receive downgrades versus their predictions.

Previous academic research has raised measurement fairness concerns with predicted grades of the type that we have discussed earlier in this chapter. Murphy and Wyness (2020) write:

> "In addition to final achievement, we find that the Socio-Economic Status (SES) of the student and the type of school attended are associated with accuracy. We find among students who are equally high achieving, low SES students receive predictions that are lower than those from high SES backgrounds, by around 0.059 grade points (where 1 point is equivalent to a full A-level at the lowest grade). Moreover, high achieving students from state schools also receive lower predictions than those from private schools."

We know, from the theoretical discussion earlier in this chapter, that this

statement is describing failures of *separation*. It is a fairness claim that involves comparing students with the same *achievement*, which is to say the same value of the target μ. The claim is that, among those with the same level of achievement, those with low SES backgrounds and those from state schools had lower predicted grades than students with high SES backgrounds from private schools. But we also know, from earlier in this chapter, that there is a potential tension with *sufficiency*. If you compare students with with the same predicted grades from different backgrounds and schools, does one group tend to outperform the other in their exams?

In order to assess this question, we examine data on predicted grades, achieved grades, sex, ethnicity, local average educational attainment in the area that a student lives (which Murphy and Wyness refer to as SES), and school type. The data that I use here are from the Universities and Colleges Admissions Service (UCAS) and include all students in the UK who took A levels and applied to university in the 2017-2019 admissions cycles.[8]

Figure 4.2 shows both how achieved grades vary by predicted grades for relevant groups (the quantity relevant to evaluating sufficiency) and how predicted grades vary by achieved grades (the quantity relevant to evaluating separation). The fitted curves are from LOESS non-parametric regressions in order to illustrate the non-linearities in the relationships between predicted and achieved grades. In the top row, we see the best case for fairness. Men and women have the same average predicted points, the same average achieved points, and essentially identical relationships between achieved and predicted points. This is the best case because both sufficiency and separation are only possible when the underlying distributions of the target μ are the same.

The remaining pairs of panels illustrate different manifestations of the tensions between the two notions of fairness. The second and third rows show the conditional relationships between predicted and achieved grades for the two variables highlighted by Murphy and Wyness (2020): school type and local area educational attainment. The right plots in these rows show that the discrepancies quoted above based on data from 2013-15 remained in the data for 2017-19. Among students with the same level of achievement, independent school students and those from the highest quintile of local area education attainment both had higher predicted points than those from state schools and those from the lowest quintile of local area education attainment. This seems unfair because predicted grades determine which students get university offers at which universities. State school students and those from areas with low local area educational attainment are broadly understood to be already disadvantaged compared to students going to independent (which is to say private) schools and those living in areas with high educational attainment. These disadvantaged groups are both predicted to score lower on average on the exams and do in fact score lower on average on the exams, but this failure of *separation* means that those from the disadvantaged groups had the further apparent disadvantage of having had worse predictions fed into

[8] These data are available for purchase from UCAS with strong anonymity controls and under a restrictive publication license. The terms of the license restrict publication to "100 data points", which include numerical quantities such as averages and regression coefficients calculated from the data. Since the analysis below uses non-parametric regression, the equivalent quantity is the number of degrees of freedom (calculated as the trace of the projection matrix H), which may not have an integer value. The total number of effective data points revealed in in this section is 95.3.

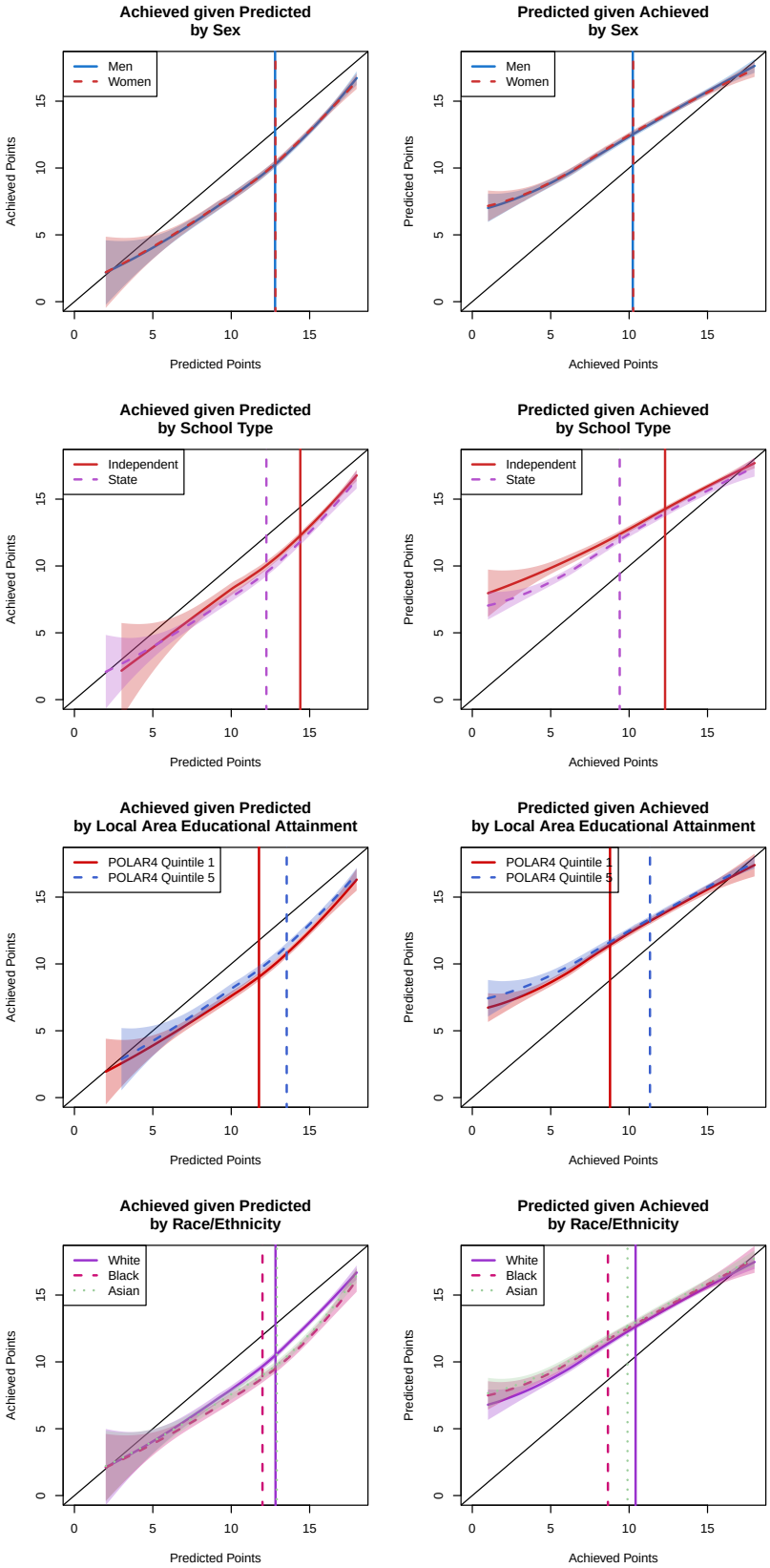


Figure 4.2: Achieved A level grade points given predicted A level grade points (left column) and predicted A level grade points given achieved A level grade points (right column), by sex, school type, race/ethnicity and local educational attainment. The mean values of the relevant groups are shown for predicted (left column) and achieved (right column) A level grade points as a vertical line on each plot.

their university admissions applications even where they ultimately had equal achievement on the exams.

I say apparent disadvantage because the corresponding plots in the left column seem to tell a different story. Here, we observe that students from the disadvantaged groups (state school and low local area education attainment) achieve lower A level grade points than those from the advantaged groups (independent school and high local area education attainment) given the same predicted A level grade points. Thus, a university admissions committee could expect the students from the advantaged backgrounds to be more likely to make their predicted grades than those from the disadvantaged backgrounds. Whereas the previous analysis seemed to suggest that that the disadvantaged students were getting further disadvantaged by the measurement error in predicted grades, this *sufficiency* analysis seems to suggest that the measurement error works in their favour.

The counter-intuitive part of the preceding discussion is that it can all be true at the same time. In these instances, there are failures of *both* separation and sufficiency, but they seem to point in opposite directions as to which group the measurement error is "helping" and which group the measurement error is "hurting". This is possible because of the large gaps in average A level grade points between the groups in question (marked with vertical lines in the plots), which creates more tension between the two notions of fairness.

The final row of Figure 4.2 shows the same relationships by race/ethnicity, categorised using the top-level UK census categories of White, Black and Asian. Predicted grades and achieved grades are somewhat higher for White and Asian students relative to Black students, but the differences are not quite as large as the differences by school type and local area educational attainment. Here we see a clear failure of sufficiency (left plot) with white students achieving higher A level grade points than both Black and Asian students at the same level of predicted grades. At the same time, the data come very close to satisfying separation (right plot), with students of all three groups holding nearly identical predicted grades at levels of achievement other than the very low end of the range.

So, once again, how we ought to evaluate these patterns depends on which standard of fairness we want to adopt. There are arguments in favour of both possibilities. The argument in favour of focusing our attention on *sufficiency* (and thus the plots of the left) is that predictions determine admissions, and we want to be in a world where knowing someone's school type or local area or race/ethnicity does not tell you anything about whether they are likely to overperform or underperform their predicted grades. If sufficiency fails as we see it does, university admissions have reason to "adjust" the predicted grades for some groups relative to others, which we might find troubling. The argument in favour of focusing our attention on *separation* is that the whole system is based on the idea that the achieved exam scores are the canonical truth, the predictions are just predictions. Thus it seems compelling that we

ought to want to treat students with the same ultimate achievement fairly in the admissions processes, and thus it is important that those students with the same ultimate achievement receive the same predicted grades on average regardless of their school type or local area or race/ethnicity.

There is a final subtlety here that makes reasoning about this case even more difficult. As just noted in the last paragraph, we have been discussing the predicted grades as a noisy measure of the achieved grades, and the latter as the canonical, gold-standard truth that the predications are meant to recover. But exams themselves are not wholly reliable measures of students' understanding of the body of material that the exams are meant to be assessing. Anyone who has ever sat an exam will know that sometimes the questions are the ones you are better prepared for and sometimes they are the ones you are less well prepared for, sometimes you get a good night's sleep and sometimes you do not, sometimes you are not feeling well and sometimes you are. All of these contribute to measurement error in the achieved grades, with respect to the underlying concept that the exams exist to measure: students' understanding of a body of material.

Once we consider the possibility that achieved grades are themselves only a noisy measure of the thing that we really wish university admissions could be based on, this is a further complication in thinking about what fairness actually requires in this situation. What if students' exam performance has *low reliability / high variance* in the sense that it would vary a lot from one sitting to the next? What if students' predicted grades, based as they are on a teacher's experience of the student over a longer period of time, are *high reliability / low variance*? If this is the case, then it is entirely possible that the predicted grades are actually a *more* reliable indicator of how a student would do on average across many sittings of the A level exams than is that same student's results on a single sitting of the exam. To my knowledge there is no good data on the "test-retest" reliability of A level exams, and so this is merely a speculative possibility to further contemplate how we think about fairness in this context.

4.4 Conclusion

The preceding discussion fails to take a view on whether sufficiency or separation is the appropriate criterion for evaluating measurement fairness in the examples we have considered. There may be a clear argument in one case or the other, but I have not been convinced one way or the other.[9] Of course a further possibility is that a compromise is appropriate, and one ought to aim to keep violations of both criteria small. Regardless of where you come down on this, I hope that you take away from this chapter an appreciation for the difficulty of defining fairness in measurement and some of the tensions that exist.

In the next chapter, we will start to unpack how measures are constructed. We will continue to focus on the problem of measurement error, but thinking

[9] Perhaps in a future edition of this book I will take a view!

more about its consequences for subsequent data analysis using the measures as opposed to the consequences for specific units being measured.

5
Consequences of Mismeasurement

In the last two chapter, we introduced the definition of measurement error, discussed basic properties of measurements like reliability and validity, and considered the challenges in assessing questions of fairness in measurement in the presence of measurement error. That latter discussion was about the consequences of measurement errors for the measured units themselves, and thus also for society. In this chapter, we consider the consequences of measurement errors for *social science*. What happens when the measures that we are using for social scientific analyses are measured with error? What are the potential problems that can arise, when will they arise, and will we be able to easily assess whether we have these problems?

Measurement error can potentially lead to mistaken conclusions in subsequent analyses that employ the measures. By mistaken conclusions, I specifically mean mistaken claims about relationships involving the concepts μ that those measures m claim to represent. These mistaken claims arise because subsequent analyses cannot distinguish between the component of the measure m that reflects the concept of interest μ and the component of the measure that does not (the measurement error ϵ_m). There are three basic cases that are important to consider, corresponding to the three most important ways that a (potentially mis)measured variable might enter a subsequent analysis.

1. Our mismeasured variable might be the outcome variable of our analysis.
2. The mismeasured variable might be the primary explanatory or treatment variable.
3. The mismeasured variable might be a control or conditioning variable, that we are using to ensure that we are making fair comparisons units with different values of the primary explanatory or treatment variable.

We consider each of these cases in turn.

5.1 Error in the Outcome Variable

Figure 5.1 illustrates the case where the mismeasured variable μ is the outcome variable, and we want to estimate the causal effect of some treatment variable

T on μ. Here, we assume that our measurement m has the sort of *generative* relationship to μ that is assumed by the representative perspective on measurement, an assumption that we will relax below. If in fact we can only estimate the causal effect of T on m, what might go wrong?

If T has a causal effect on μ, we can see that this will in turn have an indirect effect on the indicators I_1, I_2, etc that will in turn cause the value of the measure to change. So, if there is a causal effect of T on μ, there will also be a causal effect of T on m. This is the good news.

The bad news is that this is not all that might happen, there is also the potential that the treatment will have an effect on *any* of the other quantities O that influence the indicators I_1, I_2, etc and therefore the measure m. Thus, a treatment T might have an effect on O that generates an effect on m, *without* having any effect on μ. Given that μ is the thing that m ostensibly measures, this is bad, and could lead to misattribution of the causal effect. There might be no causal effect of T on μ, but we might think there is because we observe a causal effect of T on m, which is supposed to measure μ. The causal effect of the treatment is on the measurement error ϵ_M, rather than on the quantity we wished to measure μ.

Can we tell the difference between these two scenarios? Unfortunately, the answer is generally no. In some cases we may be able to do so if we have additional information about the other factors O or the measurement errors ϵ_M that result from them. But this is not usually the case because if we had such information we would typically just use it to improve our measure to eliminate those errors.

What if our measurement is not *generative*, as is assumed by the representative measurement perspective, but rather *summary*, as is allowed by the pragmatic measurement perspective? Remember, the implication of this is that there is no necessary causal relationship from the concept we aim to measure μ to the indicators I and the measure m. Figure 5.2 shows that this generates an additional problem beyond the ones described above (which still apply, but are omitted from the figure for clarity). The additional problem is that there is now no guarantee that a treatment effect of T on μ will actually be conveyed through to make an observable effect on m.

Thus, from looking at these cases, we can identify two general concerns that apply in the case where the outcome variable is potentially subject to measurement error. First, the causal effect of the treatment may be on the measurement error rather than on the concept of interest. Second, if the measurement has a *summary* relationship to the concept of interest rather than a *generative* one, the causal effect of the treatment T on the concept of interest μ may fail to have any effect on the measure m. Thus, in practice, we need to think carefully about both spurious effects in the first instance and spurious non-effects in the second.

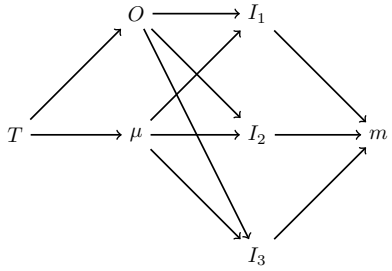


Figure 5.1: Measurement error in outcome μ, with representative measurement m for μ. Causal relationships between treatment variable T, and target concept μ, indicators I, measure m and other factors O.

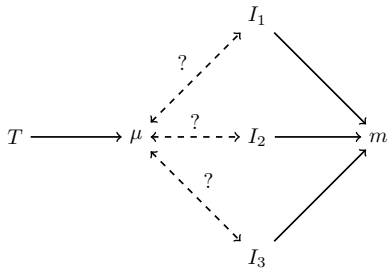


Figure 5.2: Measurement error in outcome μ, with pragmatic measurement m for μ. Causal relationships between treatment variable T, and target concept μ, indicators I, and measure m.

5.2 Error in the Treatment Variable

Are things any better if our (potentially mis)measured variable is the primary explanatory or treatment variable? The answer, unfortunately, is no; the set of potential problems are very similar. Figure 5.3 shows the causal graph for this situation, and again there is a crucial ambiguity. If we observe changes in m, these might reflect changes in μ or they might reflect changes in O. It might be that there is a causal effect of changing μ on Y, or a causal effect of changing any of the other factors O that influence our measure via its indicators. If all we observe is m and Y, we have no way of distinguishing these two scenarios.

As was the case when we considered mismeasured outcome variables in the previous section, an additional inferential problem is introduced if we cannot be confident that our measure has a generative relationship to the concept of interest. Figure 5.4 shows that in this situation, there is no longer any necessary implication about the effect of μ on Y if we have observed an effect of m to Y. As with the previous case, losing the directed causal path from μ to m undermines our ability to make causal claims involving μ. For representatively measured variables we need to worry about alternative causal pathways connecting μ via the measurement error; for pragmatically measured variables we also need to worry about the possibility that changes in the measured variable will fail to reflect, or fail to be generated by, changes in the quantity we wanted to measure.

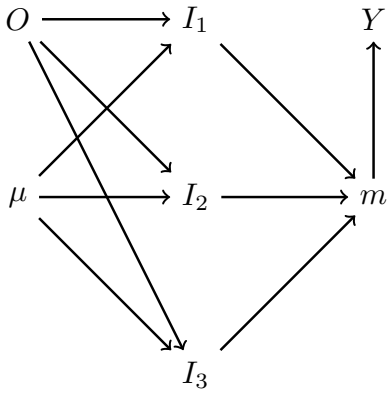


Figure 5.3: Measurement error in treatment μ. Causal relationships between outcome variable Y, and target concept μ, indicators I, measure m and other factors O.

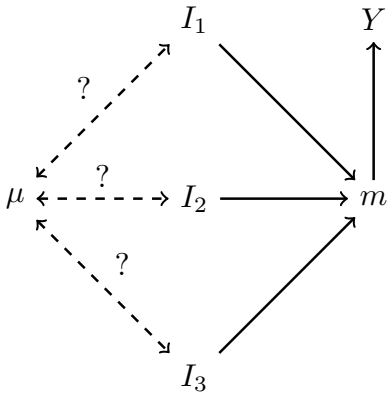


Figure 5.4: Measurement error in treatment μ, with pragmatic measurement m for μ. Causal relationships between outcome variable Y, and target concept μ, indicators I and measure m.

5.3 Error in the Control Variables

The third important case to consider is measurement error in control variables. We include control variables to enable fairer comparisons between units with different values of the primary explanatory / treatment variable, in cases where that variable is not randomly assigned. But what if those control variables are measured with error? It is common for researchers to include many controls that they think might be associated with likely confounding, but without strong claims that the exact variables which they include are the specific causal source of the confounding. Is this strategy likely to eliminate omitted variable bias, where it exists?

The answer, once again, is unfortunately no. Figure 5.5 illustrates the situation. If the outcome variable Y is causally influenced by both the true value of the control variable μ and also the treatment variable T of primary interest, conditioning on m instead of μ will fail to appropriately control for μ because m is contaminated with the other factors O.

To illustrate the consequences of this in a simple numerical example, imagine that we are interested in the effect of a variable x on an outcome y. In reality, the relationship between x and y follows this equation:

$$y = \beta x + \gamma w + \epsilon$$

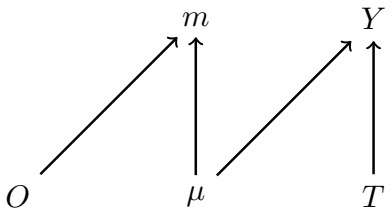


Figure 5.5: Measurement error in regression control variable μ.

That is, y may depend on x, but also on w. We assume that ϵ is uncorrelated with x and w, so this is a well-behaved regression problem, with no consequential omitted variables. Or it would be, if we have measured x and w precisely.

We do observe x exactly, there is no measurement error in our primary variable of interest. But there is measurement error in the "control" variable w. We do not observe w exactly, instead we observe $z = w + \nu$, which has measurement error ν with mean 0 and standard deviation σ_ν. But remember, we are interested in β, not γ. Does it matter that we have mismeasured the control variable w? We make the assumption that the measurement error $\nu = z - w$ is uncorrelated with x,[1] so the fact that x is correlated with z also implies that it is correlated with w.

[1] If the measurement error is correlated with x that creates an additional bias term. If you were fantastically lucky, this might cancel out the bias that we focus on here, but you are not that lucky.

$$\hat{\beta} \rightarrow_p \beta + \frac{\gamma \sigma_\nu^2 \sigma_x \sigma_w \rho_{xw}}{\sigma_x^2 \left(\sigma_w^2 + \sigma_\nu^2\right) - \sigma_x^2 \sigma_w^2 \rho_{xw}^2} \tag{5.1}$$

This regression will be unbiased only if the second term is zero, which is true if any of the following four conditions are met:

- If the coefficient on the control variable is zero ($\gamma = 0$), then the coefficient of interest is unbiased.
- If there is no variation in the (true) control variable ($\sigma_w = 0$), then the coefficient of interest is unbiased.
- If the control variable is uncorrelated with the explanatory variable of interest ($\rho_{xw} = 0$), then the coefficient of interest is unbiased.
- If the measurement error of the control variable is zero ($\sigma_\nu = 0$), then the coefficient of interest is unbiased.

Unfortunately, *none* of these are interesting or useful cases. If the coefficient on the control variable is zero, the control variable was unneeded. If there is no variation in the true control variable in the data, the control variable was unneeded. If the control variable was uncorrelated with the explanatory variable of interest, the control variable was unneeded. If the measurement error of the control variable is zero, we are back to the standard regression problem without the measurement error that is our interest here. These four conditions cover no interesting cases: if there is measurement error in a control variable, and you actually needed the control variable in the first place, the estimate of the coefficient on the explanatory variable of interest will be biased.

5.4 *Application - Behavioural vs Self-Reported Information Seeking*

Clifford and Jerit (2018) report on two experiments that aimed to assess whether and how emotional responses impede learning. We will focus on one component of Study 2, which involves presenting information about dengue fever to 748 students at the University of Houston.

> "The treatment described how the climate in Houston is ideal for the spread of dengue, how the disease can spread un-noticed, and that there is currently no

> vaccine.… Our primary manipulation consisted of the presence (or absence) of three disgusting images, a common method for inducing disgust. The images displayed symptoms of dengue fever; however, this information also was included in the text. This design feature increased our confidence that any effects of the images were produced by manipulated disgust rather than by some other mechanism."

Figure 5.6: Aedes aegypti mosquito, the primary vector for dengue fever.

The authors then "gauged the motivation to seek new information about dengue, using two behavioral measures and two self-reports."

Label	Scale	Type
B1	Binary	Behavioural
B2	Binary	Behavioural
S1	Ordinal (0-4)	Self-Report
S2	Ordinal (0-4)	Self-Report

The treatment effects on these four outcome measures are shown in the table below.

Measure	Control	Treatment	Difference	Interval
B1	0.195	0.135	-0.060	0.01-0.11
B2	0.223	0.157	-0.066	0.01-0.12
S1	1.252	1.281	0.029	-0.18-0.12
S2	1.244	1.237	-0.007	-0.15-0.16

Both behavioural measures exhibit significant differences between treatment and control conditions: respondents exposed to the disgusting treatment are less likely to want to be informed about a further information session and less likely to want to receive more information about dengue. In contrast, there is negligible difference between the treatment and control groups with respect to the self-reported measures. Respondents were similarly likely to say they would look up more information or discuss dengue with their family or friends in the next week, regardless of whether they saw the disgusting images or not.

One way to read these results is to simply focus on these as different outcomes. The treatment had an effect on some outcomes and not on others. The treatment made people less likely to immediately sign up for more information, but did not change their stated intention to further research and discuss dengue fever in the future. These are easy enough to reconcile, though the discrepancy is certainly interesting.

However, the authors' study aims to make more general statements about the concept of "information seeking". If we view these measures as all reflecting information seeking, there discrepancy becomes more relevant. Which of these are the "right" measures of "motivation for information seeking"? Are there biases that are likely to affect only the self-reported measures and not

the behavioural measures, or vice versa? There are (at least) five questions that we need to consider, in order to decide what conclusions to draw from these results.

1. Are the behavioural measures good measures of "information seeking"?
2. Are the self-reported measures good measures of "information seeking"?
3. Is there a treatment effect on "information seeking"?
4. Is there a treatment effect on the measurement error in the behavioural measures?
5. Is there a treatment effect on the measurement error in the self-reported measures?

To get an initial sense for the answers to the first two questions about the quality of the measures, we might look at the extent to which the different measures are correlated with one another.

Table 5.3: Pairwise (pearson) correlations for the four measures of information seeking.

	B1	B2	S1	S2
B1	1.00	0.29	0.31	0.28
B2	0.29	1.00	0.30	0.23
S1	0.31	0.30	1.00	0.63
S2	0.28	0.23	0.63	1.00

All four measures are positively correlated with one another, although the correlations are not very strong with the exception of the two self-reported measures, which have a correlation coefficient of 0.63. The fact that all the measures are positively correlated with one another provides some evidence there is a common element to all four measures.[2] People who tend to seek information in one way tend to also do so in others.

So we have some reason to believe that these four measures reflect some more general concept that we might call "information seeking". Nonetheless, the low correlations strongly imply that they are all, at best, "weak" measures. There must be substantial measurement error in these measures, or they would be more highly correlated with one another. One plausible interpretation is that they are all weak measures of the target concept of "information seeking", and the relatively strong correlation between the two self-reported measures is not because they are better measures of the target concept, but because they are extremely similar survey prompts or because they are measured on 5 point ordinal scales rather than as binary yes/no responses. It is also possible, but perhaps less substantively plausible, that it is the two self-reported measures that are high quality and the behavioural measures lower quality. Most researchers would assume that a behavioural measure is likely to be

[2] This kind of pattern of positive pairwise correlations is discussed in more detail in Chapters 9 and 11. Applying the one of the methods discussed in the latter of these, it is relevant to note here that the first principle component for the set of (standardized) measures has similar, positive coefficients for all four.

better than a self-reported measure, not worse, since it involves a costly action and not just cheap talk. Finally, these correlations are consistent with the possibility that one specific measure is in fact a very good measure while the rest are very poor, but this seems implausible given what we know about how the measures were generated.

Let's say we accept the idea that these four measures all reflect the target concept to similar degrees, albeit with substantial measurement error. The answers to the three questions (Q3, Q4 and Q5) posed earlier about treatment effects have possible answers that are interlinked.

One way to reconcile the results is that there *is no* treatment effect on information seeking, rather there is a negative treatment effect on the measurement error in the behavioural measures but not on the self-reported measures. Is this plausible? Perhaps the presence/absence of disgusting images had no effect on respondents interest in learning about dengue fever, but simply made them less inclined to receive information specifically from the people running the experiment. The behavioural measures were linked to getting information from the specific people running the experiment (S2) and from an information session with a local organization with some relationship to the experimenters (S1). Perhaps the disgust response was not really on general interest about learning more about dengue, but more focused on the experimenters and those that they work with?

Another way to reconcile the results is that there *is* a negative treatment effect of the disgust treatment on information seeking, there is no treatment effect on the measurement error in the behavioural measures, and there is a positive treatment effect on the measurement error in the self-reported measures that cancels out the treatment effect on information seeking. This is the authors' own interpretation of the results. Note that this is a bit more complicated, as it involves countervailing treatment effects on the target concept and the measurement error. Nonetheless, this is not implausible. The disgust treatment may make people less inclined to seek information, but the fact that it is emotionally engaging might also make respondents feel they *ought* to be seeking out more information. This might increase social desirability biases in responses, which would manifest primarily in the self-reported measures because they are cheap talk and do not require immediately engaging with more disgusting information about the effects of dengue.

There are further possible interpretations that are more complicated, involving more complex cancelling out of positive and negative treatment effects. The point of this discussion was not to be exhaustive, but rather to highlight how the potential for measurement error in an outcome variable complicates the interpretation of an experiment. This is especially true where the researcher is not interested in the outcome variables in their own right, but rather in underlying concepts that those outcome variables are meant to measure.

5.5 *Application - Objective vs Subjective Sleep Hours*

In an article "Sleep duration and health among older adults: associations vary by how sleep is measured", Lauderdale et al. (2016) report on a study that compares subjective (self-reported) sleep hours and objective (measured with wrist monitors) sleep hours as predictors of various health outcomes.

> "Cohort studies have found that short and long sleep are both associated with worse outcomes, compared with intermediate sleep times. While demonstrated biological mechanisms could explain health effects for short sleep, long-sleep risk is puzzling. Most studies reporting the U shape use a single question about sleep duration, a measurement method that does not correlate highly with objectively measured sleep. We hypothesised that the U shape, especially the poor outcomes for long sleepers, may be an artefact of how sleep is measured."

The data for this study come from a national probability sample of older US adults, mostly aged 60-90 at the time of the study. Two survey measures of typical sleep duration were collected, one based on a single question about typical sleep duration ("Subjective 1") and another calculated from reported times when respondents went to bed and woke up ("Subjective 2"). Both of these are typical survey measures used in epidemiologic studies of sleep. In addition, "Objective" sleep duration was measured using an "actigraph" wrist monitor that measures patterns of physical movement. Finer details of each of these measures can be found in the original source article.

The first question that we might ask is how strongly correlated these different measures are? To what extent do these different measures capture the same cross-sectional variation across respondents in sleep duration? Figure 5.7 shows that the correlations between these measures are weak, but nonetheless positive. We will assume for the purposes of discussion here that the objective measure has minimal measurement error with respect to the target concept of actual biological sleep duration, and therefore that all error is attributable to respondents' misperception/misreporting of their own sleep duration. The modest correlations here suggest that variation in how long people report they sleep has only a weak relationship to how long they are actually sleeping.

How are these measures related to other health quantities of interest? The authors of the original study examine a standard self-reported health question: "Would you say your health is—excellent, very good, good, fair, or poor?" Following the authors of the original study, we dichotomise this into fair/poor versus excellent/very good/good. Note that while this sort of self-reported health question is obviously subjective, it has been validated in the epidemiologic literature and does predict mortality and future use of medical services (Miilunpalo et al., 1997).

Figure 5.8 shows the partial association of fair/poor self-reported health with all three measures of of sleep duration, controlling for age, gender and race/ethnicity using a generalised additive model.[3] The difference between the three plots is striking: there is a U-shaped relationship between the two sub-

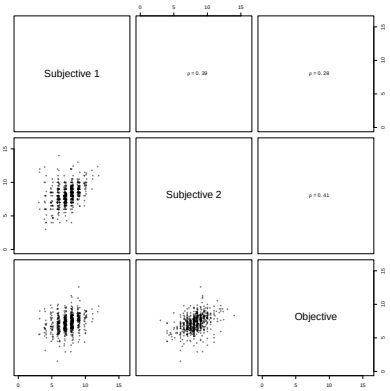


Figure 5.7: Correlation between two subjective measures of sleep duration and one objective measure of sleep duration, all measured in hours.

[3] Generalised additive models are regression models that allow for flexible, non-linear partial associations with continuous explanatory variables. See Simon Wood, "Generalized Additive Models: An Introduction with R" (2017) for a good introduction. A similar plot could have been generated by fitting a multiple regression with polynomial terms for the displayed variable.

jective sleep measures and fair/poor self-reported health. These relationships are strong, with those at the extremes of subjective sleep duration 20-40 percentage points more likely to say they have fair/poor self-reported health. This relationship is absent in the objective data, with only a very weak (and statisticaly insignificant) negative relationship between sleep duration and fair/poor self-reported health.

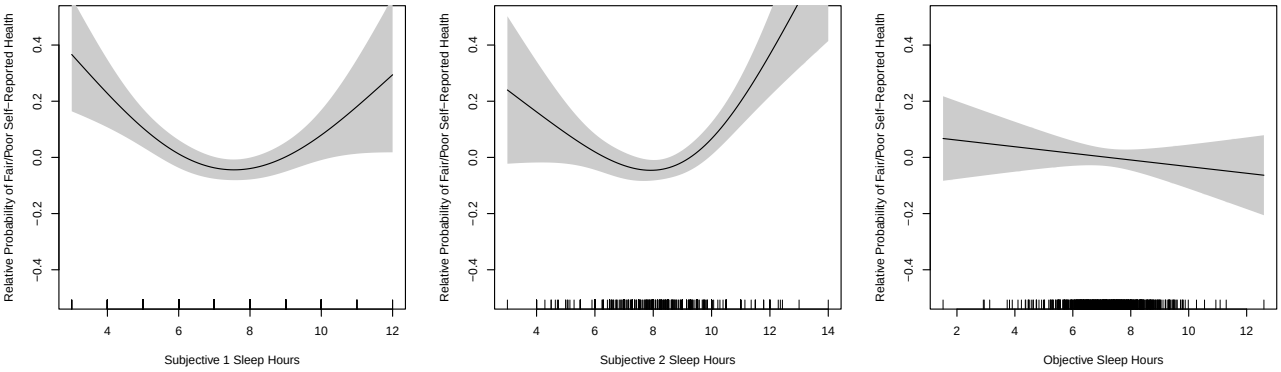


Figure 5.8: Partial association of self-reported health with two subjective and one objective measure of sleep duration, controlling for age, gender and race/ethnicity.

If we take the objective measure to have minimal measurement error, we can calculate the measurement error for each of the subjective measures as the difference between the subjective measure and the objective measure for each person. Figure 5.9 shows that this measurement error for the first subjective measure has the same U shaped relationship with self-reported health as above. The association of subjective/self-reported health with this measure of subjective/self-reported sleep appears to be *entirely* associated with the measurement error in subjective/self-reported sleep, rather than with the objectively measurable variation in sleep duration.

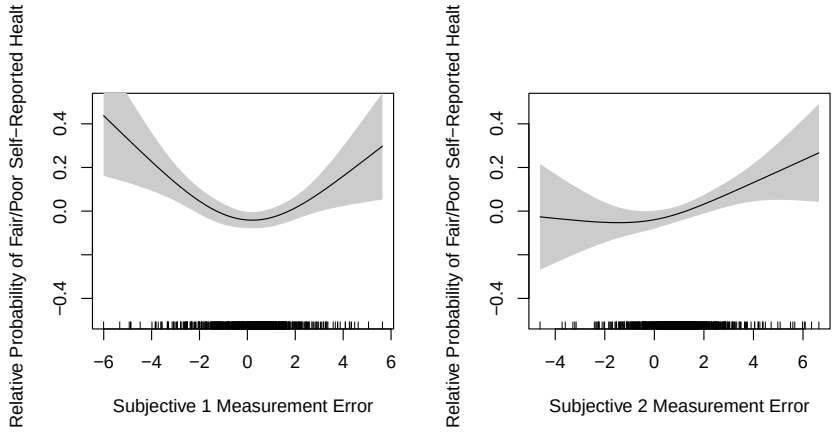


Figure 5.9: Association of self-reported health with measurement error in subjective (self-reported) sleep duration.

A final, and striking, feature of the left plot in Figure 5.9 is that the minimum is located very close to a measurement error of 0. The people with the best self-

reported health are the people who most accurately report how many hours they sleep. Perhaps measurement error is bad for your health?[4]

[4] Measurement error is definitely not randomly assigned, so the implicit causal inference in this statement is unjustified. A more careful statement is that those 60-90 year old Americans who accurately report their sleep hours are more likely to feel healthy than those who inaccurately report their sleep hours. There are a number of possible mechanisms for this, which are beyond the scope of the discussion here.

5.6 Illustration - Measurement Error in a Control Variable

In section 5.3 above, equation 5.1 shows that a regression coefficient on an explanatory variable of interest x will be biased due to needed control variables w being measured with error, even if the explanatory variable of interest is measured without error. The bias for $\hat{\beta}$ is given by the following expression, in terms of the true coefficient on the control variable γ, the standard deviation of the measurement error σ_v, the standard deviation of the control variable σ_w, and the correlation of the explanatory variables x and w:

$$\hat{\beta} \rightarrow_p \beta + \frac{\gamma \sigma_v^2 \sigma_x \sigma_w \rho_{xw}}{\sigma_x^2 \left(\sigma_w^2 + \sigma_v^2\right) - \sigma_x^2 \sigma_w^2 \rho_{xw}^2}$$

How substantial are these biases? To get a sense, we consider a simple case. Let's consider the case where the true coefficient of interest $\beta = 0$, the case where there is no true effect of the primary explanatory variable of interest. The question, in these cases, is the extent to which our inferences about β are biased by having a mismeasured control variable such that we erroneously conclude that there is an association with the explanatory variable. For purposes of illustration, we set the coefficient on the control variable $\gamma = 1$ and assume that the control variable w is correlated with the primary explanatory variable with a correlation of $\rho_{xw} = 0.5$.

Figure 5.10 shows how the bias in β varies as we increase the standard deviation of the measurement error σ_v, holding everything else constant. Bias increases as the degree of measurement error σ_v relative to the true variation σ_w in the control variable increases. As the measurement error comes to dominate, $\sigma_v \gg \sigma_w$, the bias approaches the degree of bias that would result from omitting the control variable entirely. In these plots, $\sigma_v = 1$ is the case where the measurement error has the same magnitude of variation as the true control variable $\sigma_w = 1$, which corresponds to the case where the measure is correlated with the true measure at $\rho_{xw} = 0.71$, or $R^2 = 0.5$. At this point on these curves, the bias is already a substantial fraction of the omitted variable bias that would result from omitting the control entirely.

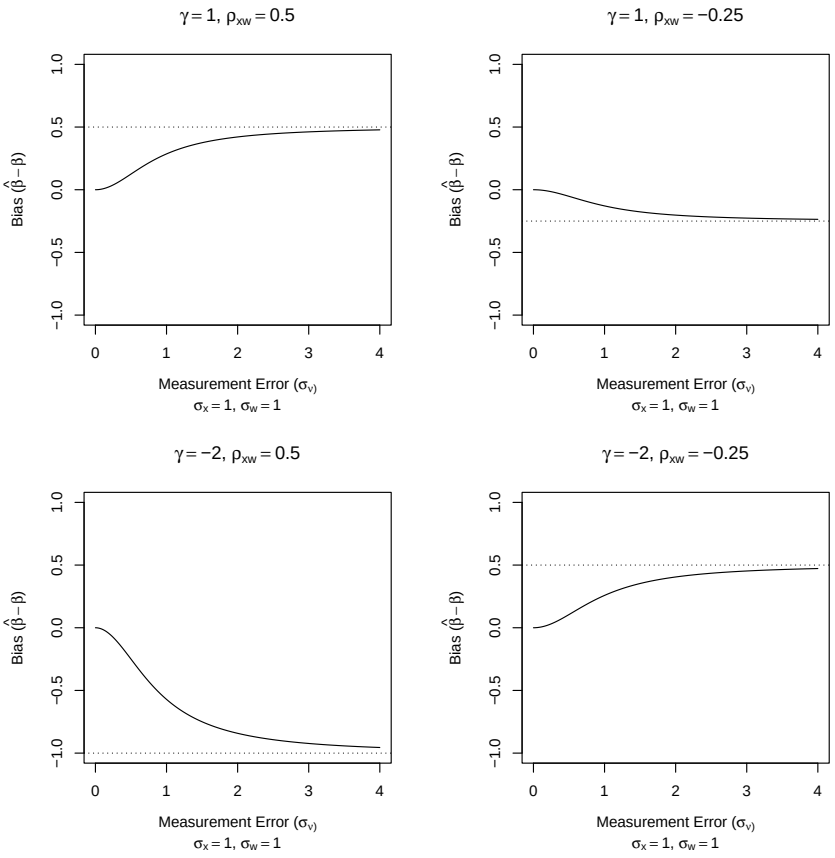


Figure 5.10: Bias in coefficient β on primary explanatory variable as measurement error in control variable increases. Dotted line shows the bias that would result from omitting the control variable from the model.

The stronger the correlation ρ_{xw} between the primary explanatory variable x and the true control variable w, the more rapidly measurement error becomes a problem. To illustrate this, we continue to examine the $\sigma_v = 1$ case where half the variance of the control variable is measurement error and half is the true control variable, but now varying ρ_{xw}. The two panels of 5.11 show that while the mismeasured control variable reduces the the bias versus no control by half when the two explanatory variables are weakly correlated, the proportional bias reduction declines as the control variable becomes more strongly correlated with the explanatory variable of interest. The more that you need the control variable, because it strongly confounds the variable of interest, the more bias results from a given amount of measurement error.

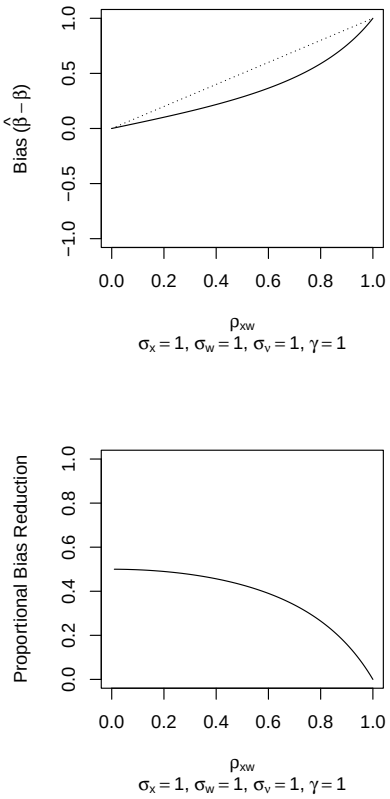


Figure 5.11: Top panel: bias in coefficient β on primary explanatory variable as correlation ρ_{xw} between control variable and primary explanatory variable varies; dotted line shows the bias that would result from omitting the control variable from the model. Bottom panel: proportional reduction in bias from including the control variable in the model.

5.7 Conclusion

A final, overarching point, is that measurement requires you to be attentive to the causal relationships that generated your measure. A good measure is one that makes the differences between the measure and the target concept as small as possible, which makes it sound like a purely predictive problem. And it could be, so long as you never used your measures for anything.

However, in order to use your measures in applications where you care about the underlying concepts, rather than the measures as such, you do need to be attentive to the causal relationships between your target concept and your measures (typically via your indicators). It is not just that you want your measurement error to be as small as possible, but also that you want it to be as idiosyncratic as possible *with respect to the other variables of interest*. As several of the examples above show, cases where measurement errors in one variable have relationships to other variables of interest are particularly likely to lead to spurious conclusions.

6
Deriving Scales using Theory

This chapter discusses general strategies for deriving scale measures from theoretical arguments, with a large number of worked examples. Because theoretical arguments are necessarily specific to particular applications, most of the chapter consists of the applications. The examples we will be considering all involve concepts that are in some sense "close" to the data: we have data that is clearly relevant to the concept we are interested in, but there is more than one way that we might translate the data into a measure and so we need to think about which one to choose.[1]

[1] Even apparently simple measurement problems can have myriad solutions: Choi et al. (2010) compile a survey of 76 published measures of similarity and distance between two vectors of binary quantities.

6.1 Axiomatic Analysis

One strategy for deriving a measure—or for checking whether a measure you have otherwise derived makes sense—involves setting out axioms that the measure should satisfy. Listing these is a very useful way of figuring out the connection between the concept that you are interested in and the data that you have to work with. Here are five criteria that frequently form the basis of useful axioms.

1. Special/extreme/limiting cases. What should happen to the measure when the observable data go to their limits? Depending on your application, the limits might be $-\infty$, 0, 1, ∞ or some other value that is relevant.
2. Equal cases. Which distinct profiles of the indicator data correspond to the same values of the concept of interest?
3. Derivative conditions. What are the directions of the associations between the data/indicators and the concept of interest? That is, when an indicator goes up/down, what should happen to the measure?
4. Continuity and smoothness conditions. Is the relationship between the data/indicators and the concept continuous or does it have discontinuities? Is the relationship smooth or does it have kinks (discontinuities in the first derivatives)?
5. Functional form restrictions. Is there a reason to restrict the possible relationships between the data/indicators and the measure to a simple family of possible functions?

Unlike the first four criteria, functional form choices tend to be arbitrary, but are useful in getting to a specific measure that is not too complicated to work with once you have satisfied all the other conditions. You can think of these as applying something like Occam's Razor: "everything should be made as simple as possible, but no simpler."

Note that one need not have all of these in a list of axioms, these are just examples of the kinds of axioms that are typically useful to specify, and that you should therefore consider.

6.2 Dimensional Analysis

Any time you are making a claim to have measured something, it is important to assess whether the units of the measure are internally consistent. This process of *dimensional analysis* is widely used when solving problems in the physical sciences, particularly for checking the plausibility of a final calculation. It is more rarely taught in the social sciences (particularly outside of economics). At its core, dimensional analysis is nothing more than ensuring that you never make statements like "my height is 70 kilograms".

To do dimensional analysis, you need to work out the *dimensions* of all the quantities that you are working with. These could be time, money, people, etc. Mathematically, each of these dimensions is described in terms of some *units*, for which there might be several choices. Time might be measured in units of days or years; money might be measured in units of $s or £s, people might be measured in units of people or in thousands or millions of people.

One of the basic mathematical operations that one learns to do as a child is a unit conversion:

$$1 \text{ day} \times \frac{1 \text{ years}}{365.25 \text{ days}} = \frac{1}{365.25} \text{ years}$$

$$1 \text{ £} \times \frac{1 \text{ \$}}{0.78 \text{ £}} = 1.28 \text{ \$}$$

Unit conversations rely on the basic mathematical fact that you can always multiply a quantity by 1 without changing that quantity. The unit conversion ratios used in the expressions above, eg $\frac{1 \text{ years}}{365.25 \text{ days}}$, are themselves equal to 1. The *dimension* of the numerator and the denominator are the same (time) but the *units* (years versus days) are not.

$$\frac{1 \text{ years}}{365.25 \text{ days}} = \frac{365.25 \text{ days}}{365.25 \text{ days}} = 1$$

This means that not only are these ratios equal to 1, they also are also dimensionless overall, so multiplying by a unit conversion ratio does not turn a unit of time into a unit of something else.

There are many quantities that we are interested in that have compound dimensions/units. For example, per capita Gross Domestic Product (pcGDP) is a widely used measure of the economic output of countries. It has dimensions of money per person per time period, typically US$ per person per year

($\frac{\{\$\}}{\{\text{person}\}\{\text{year}\}} \times \{\text{persons}\}$). These units help indicate which kinds of mathematical operations make sense. For example, if we wanted to calculate total GDP for a country, we would multiply per capita GDP by the number of people in that country, which cancels out the units of {persons} from the denominator of pcGDP:

$$pcGDP \times Population = GDP$$
$$\frac{\{\$\}}{\{\text{person}\}\{\text{year}\}} \times \{\text{persons}\} = \frac{\{\$\}}{\{\text{year}\}}$$

In general, if you are doing a sensible mathematical operation, it will obey a few key rules:

1. If you want to add (+), subtract (-) or compare (=,<,>) two numbers a and b, they must have the same units $\{a\} = \{b\}$. The resulting units after addition or subtraction remain the same.
2. You can multiply ($\cdot$) and divide (/) numbers with different units. If a has units $\{a\}$ and b has units $\{b\}$, then $a \cdot b$ has units $\{a\} \cdot \{b\}$ and a/b has units $\{a\}/\{b\}$. By implication, if you raise a quantity a to the power p, $\{a^p\} = \{a\}^p$.
3. Summation ($\sum$) and integration ($\int$) across the entire set of units multiplies the units of the summand/integrand by the units of the summation/integration limits. Thus, $\left\{\sum_{i=1}^{n} a\right\} = \{n\} \cdot \{a\}$ and $\left\{\int_{x_0}^{x_1} a \cdot dx\right\} = \{x\} \cdot \{a\}$.

It can sometimes be slightly tricky when doing a sum ($\sum$) to figure out whether the units remain the same (following rule 1) or are multiplied by the dimension that you are summing over (following rule 3). The reason that this can be confusing is that summation and integration are really surreptitious multiplication. The distinction usually turns on whether you are simply adding two quantities (in which case, rule 1) or whether you are cumulating the quantities (in which case, rule 3), but this distinction can sometimes be tricky to figure out in particular problems.

6.2.1 Standardization

If we want to find the distance in two spatial dimensions between two points, x_1, y_1 and x_2, y_2, we can see that Euclidean distance D_2 satisfies dimensional analysis:

$$D_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{6.1}$$
$$\{length\} = \sqrt{\{length\}^2 + \{length\}^2} \tag{6.2}$$

The same is true of "city-block" distance D_1:

$$D_1 = |x_2 - x_1| + |y_2 - y_1| \tag{6.3}$$
$$\{length\} = \{length\} + \{length\} \tag{6.4}$$

If you were instead constructing a time-based measure of distance between two locations, that would involve decomposing travel between those locations into a sequence of steps, each of which are denominated in time, and then adding them up.

Distance (or dissimilarity) is a useful concept in a variety of areas, but sometimes runs into problems with units. Can you define distances without common units? Is it meaningful to talk about whether a 30 year-old male is closer to a 25 year-old woman than to a 50 year-old man? Dimensional analysis is useful here for illustrating what you would need to assume in order to make such a claim.

In this example, we have two component distances: a distance in units of years of age and a distance in units of gender. In order to *add* these together to form a "city-block" distance D_1, or to apply the Euclidean distance formula D_2, we need to transform these two distances to have the same units. But how can we possibly do that? Well, if we have something with units of years, and we want to "convert" it to some other units, the simplest way to do that is to multiply it by something with units of {desired units}/{years} or divide it by something with units of {years}/{desired units}.

One common way to do this is called "standardizing" a variable x, which involves calculating its mean $\bar{x}$ and standard deviation $sd(x)$, and defining a new variable $x_s = \frac{x-\bar{x}}{sd(x)}$. What are the units of this?

$$x_s = \frac{x - \bar{x}}{sd(x)} \tag{6.5}$$

$$\{x_s\} = \frac{\{x\} - \{\bar{x}\}}{\{sd(x)\}} \tag{6.6}$$

$$\{x_s\} = \frac{\{x\} - \{x\}}{\{\{x\}/\{sd\}\}} \tag{6.7}$$

$$\{x_s\} = \{sd\} \tag{6.8}$$

You might reasonably look at the above and say: "what do you mean by a unit of standard deviation?" But that is the key idea of standardization: you are redefining the variation in x to be measured in standard deviations (of x). Of course this means that you have not entirely fixed the problem. If you apply this transformation to a series of variables in different units, you will have distances that are measured in the standard deviations of those variables. If you then add those up, you are treating the standard deviations of the variables in your data as comparable in the sense that you can talk about two units being closer together on variable x_1 than on x_2 if they are different by 0.5 standard deviations on the former variable and 1.3 standard deviations on the latter.

Standardization makes the units numerically comparable (mostly falling between -2 and +2) and you can talk about the differences in distances on the different dimensions more comfortably. This does not mean that it is necessarily a good idea to do so! If you went around saying that 30 year-old men are more like 50 year-old men than like 25 year-old women, people would

rightly look at you as though you were a bit daft. "What do you mean 'more like'? 'More like' in what sense?" The only answer justified by standardization is "in terms of their distributions in the adult population", which may not be very convincing.

The standard deviation of a binary 0, 1 variable for gender is approximately 0.5; in an adult population the standard deviation of age is about 20 years. So a man and a woman are 2 standard deviations apart on gender. The 5 year age gap is 0.25 standard deviations while the 20 year age gap is 1 standard deviation. As a result, regardless of whether you use Euclidean or city-block distance, you would conclude that the 30 year-old man is more like a 50 year-old man ($D_1 = 1, D_2 = 1$) than like a 25 year-old woman ($D_1 = 2.25, D_2 = 2.02$). You can do this, and people do it implicitly every time they make comparisons across standardized variables, but please note that there is a bit of a cheat involved and you are relying on the population/sample variation in all the component distances being equally meaningful in your application. *It may not be*. Note that if, in a particular application, you had some other way of translating the x values onto a common scale that was more meaningful than using the sample/population variation, you could use that instead.

A multivariate generalization of standardization is the Mahalonobis distance,[2] which rescales any number of interval-level variables (plus binary variables, as with gender above) into a distance that is calculated in terms of the covariance matrix of those variables. Where x_1 and x_2 are vectors corresponding to measurements of any number of such variables for two observations, and the covariance matrix of those variables in the sample or population is S:

$$D_M(x_1, x_2) = \sqrt{(x_1 - x_2)' S^{-1} (x_1 - x_2)} \tag{6.9}$$

For uncorrelated variables this reduces to the Euclidean distance of the variables when standardized as above. For correlated variables, it makes a sometimes useful distinction between units that differ across multiple variables in ways that are typical given the correlations of those variables (smaller distance) versus ways that are atypical given the correlations of those variables (larger distance), as depicted in Figure 6.1. The Mahalonobis distance is used as a metric for clustering/classifying units that have measurements across many dimensions, for matching treated and untreated units in causal inference and for identifying units that are outliers in a multivariate rather than a univariate sense.

[2] Prasanta Chandra Mahalanobis (1893-1972) developed this measure with the goal of taking a large number of different kinds of measurements of humans in order to associate patterns of physical differences with Indian castes. Like so many of the methods covered in this book, this measure was developed for dubious classification of human populations.

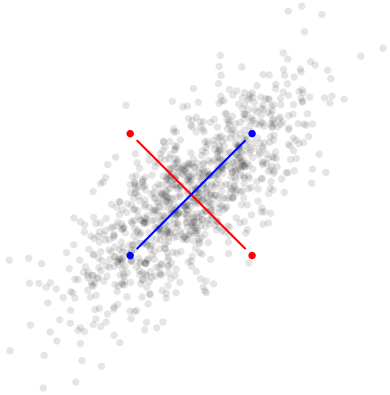


Figure 6.1: By Mahalonobis distance, the red points are more distant from one another than the blue points, even though the Euclidean distance is identical.

6.2.2 Regression

You may not have thought about it, but regression models face precisely the same unit comparability problem that we have just discussed. How is it that we can fit a regression model that adds up effects of variables x_1, x_2, etc, which are measured on all sorts of different scales and relate those to a variable y measured on its own scale? This is a question that more attentive students are

sometimes bothered by when they first see a multiple regression model.

The answer is that the β coefficients fix the unit problem:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \epsilon \tag{6.10}$$

$$\{y\} = \{y\} + \frac{\{y\}}{\{x1\}}\{x1\} + \frac{\{y\}}{\{x2\}}\{x2\} + \cdots + \{y\} \tag{6.11}$$

$$\{y\} = \{y\} + \{y\} + \{y\} + \cdots \tag{6.12}$$

You already know this implicitly because you know how to interpret a regression. The units of α, the intercept, are the same as the units of y, because α is the expected value of y when all x equal 0. The units of β_1 are $\frac{\{y\}}{\{x_1\}}$ because the interpretation of that coefficient is the change in y for each one unit change in x_1.

Note that this also explains why, if you want to compare the magnitudes of different βs to one another, you need to first standardize your X variables as we discussed in the previous section. This puts the βs on a common(ish) scale: units of y per standard deviation of x. This relates everything to the distribution of x in your data. Again, as noted above, this may or may not be desirable, but you do have to make some kind of assumption like this in order to make comparisons across variables on different scales.

6.3 Application - The Debt-GDP Ratio of Countries

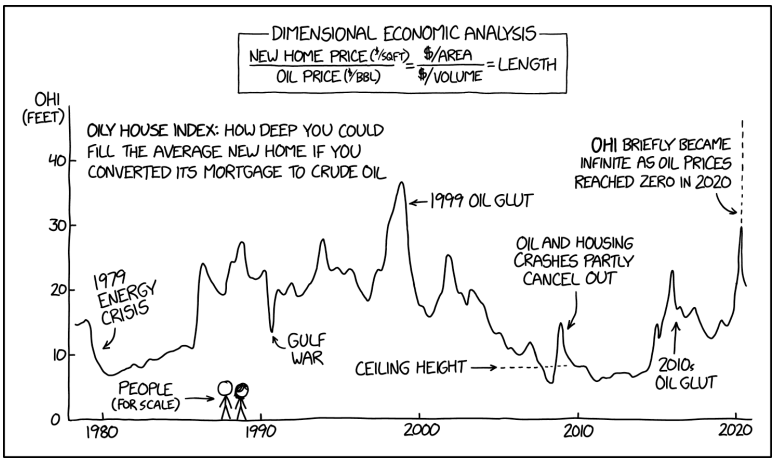


Figure 6.2: Oily House Index
https://xkcd.com/2327/

For our first example, we focus on a very simple case of thinking about how units can clarify what operations and comparisons are sensible. During the European sovereign debt crisis that started in 2009, there was extensive use of the *debt-GDP* ratio as a measure of the sustainability of countries borrowing. The numerator of this measure of debt is the total amount of money that the government was currently borrowing from bond holders. The denominator is the current gross domestic product (GDP) of the same country.

In 2011, for example, Greece was borrowing about €360 billion from various creditors. Greek GDP in 2011 was €207 billion per year. The debt-GDP ratio of Greece was therefore $360/207 = 1.74$.

But what are the units of this ratio? Debt has units of {billion €}, GDP has units of {billion €}/{years}, therefore the units of the dept-GDP ratio are:

$$\{\text{debt-GDP ratio}\} = \frac{\{\text{billion €}\}}{\frac{\{\text{billion €}\}}{\{\text{years}\}}} = \{\text{years}\} \tag{6.13}$$

Why does the debt-GDP ratio have units of years? Debt is a quantity of money. GDP is a quantity of money produced per year. We are comparing a *level* (debt) to a *flow* (GDP). The ratio tells us how many years worth of the gross domestic product of the country are required to equal the total outstanding debt. In this case, Greece's outstanding debt load was equal to 1.73 years of the entire economic product of the country, which was viewed as an exceptionally high level of debt at the time.[3] That is a coherent quantity to calculate and is relevant for benchmarking debt against the ability to pay off that debt. Of course a country cannot simply devote its entire national product to paying off its creditors as everyone would starve to death. In order to calculate the time necessary to realistically pay down a debt, one would need to further calculate the proportion of GDP that could feasibly be devoted to that purpose.[4]

[3] According to the IMF, as 2020, the Greek debt-GDP ratio is 2.01, the second highest in the world. Japan has a debt-GDP ratio of 2.52. Japan's debt-GDP ratio has been the highest in the world for decades.

[4] Similar ratios of debt to income are used when people apply for mortgages to buy a home. Even though home mortgages have higher interest rates than governments typically pay, much higher debt ratios are supportable for home purchasers than for governments because individuals can devote far more of their personal income to paying off that debt than a government can feasibly tax and then devote to debt service. In the UK, as of 2020, most mortgage lenders will lend up to 4.5 times annual income.

6.4 Application - Measuring Inequality

There are many kinds of social inequality that we might want to assess, with respect to a variety of "goods" such as wealth, income, life expectency, and others. How do we numerically measure the extent of inequality? One useful starting point for defining a measure of inequality is the distribution of the relevant good. We assume, for our purposes here, that we are looking at a single good y_i measured for n individuals (or households or some other relevant unit). Any translation of multiple kinds of goods into this single metric (eg monetary value) has already occurred. We also assume that we have solved any difficulties in measuring the y_i, how much of the good each individual has.

So, if we know how much of a good y_i each of the individuals i in our population has, how do we translate these into a single number describing the level of inequality among that set of individuals? There are many functions $I(y_1, y_2, \ldots y_n)$ we could calculate given a set of y_i, which ones best map onto the concept of inequality? Before we start proposing statistics, it might be wise to try to specify some axioms. Here are some of the axioms found in the economics literature on measuring inequality:

- Axiom 1 (symmetry): It does not matter which individuals have more or less of the goods, all that matters is the distribution $f(y_i)$. So if we swap the total goods held by any two individuals y_i and $y_{i'}$, the measure of inequality should not change.

- Axiom 2 (homogeneity): It does not matter how large/small the total quantity of goods are, just their relative value. So if we multiple all y_i by a common factor a, the measure of inequality should not change.[5]
- Axiom 3 (population independence): It does not matter how many individuals there are in the population. If we calculate the level of inequality based on $y_1, y_2, y_3, \ldots, y_n$ the measure of inequality should be the same as if we calculated it for $y_1, y_1, y_2, y_2, y_3, y_3, \ldots, y_n, y_n$.
- Axiom 4 (transfer): If we take a small amount of the good from an individual i with a larger than average y_i and give it to the individual i' with a smaller than average $y_{i'}$, the level of inequality should decrease.
- Axiom 5 (minimum): The lowest level of our measure of inequality is found if (and only if) all individuals have the same amount of the good (perfect equality), $y_1 = y_2 = \ldots = y_n$.
- Axiom 6 (maximum): The highest level of our measure of inequality is found if (and only if) one individual has all of the goods.

[5] Note that this axiom ensures that we get the same measure of inequality regardless of which units we measure y in, so that the level of inequality in a society does not change if we convert from £s to $s for example.

These all seem like sensible axioms. Note that the first three do not really say much about inequality per say, but the last three do. In particular, axiom 4, the transfer axiom, is vital to enforcing the concept of inequality. Note that there are many different ways that this could be stated, more or less precisely. Note also that the axiom, as stated, implies that the reverse transfer from someone with less than average y to someone with more than average y will increase inequality, so we do not need to state that as a separate axiom.

It turns out that there are lots of measures of inequality that satisfy all (or most) of these axioms.[6] Some of these are simple ratios of how much "the rich" have to how much "the poor" have:

[6] There are further axioms that one can define, such as having a measure that can be mathematically decomposed into contributions from subgroups, in order to more narrowly define the set of measures that meet all axioms.

- The 20:20 ratio is the ratio of the total goods held by the top 20% of individuals to the total goods held by the bottom 20% of individuals.
- The Palma ratio is the ratio of the total goods held by the top 10% of individuals to the total goods held by the bottom 40% of individuals.

These are dimensionless quantities, as they are a quantity of goods divided by another quantity of goods. Note that somewhat arbitrary choices are required to define these ratios. Why 10%, 20% or 40%? The idea of a ratio of how much more the rich have than the poor have is attractive, but leaves open the question of who qualifies as rich and who qualifies as poor. Another feature of these ratios is that they ignore the incomes of the middle of the distribution entirely, which may or may not be a desirable feature (Cobham et al., 2013). Shifting a small amount of income from someone a bit above the average to someone a bit below the average will (for most distributions of income) have no effect on these measures (Axiom 4 above is weakly violated).

Several measures exist that act on the entire distribution of y.

- *The Hoover Index/Robin Hood index* The (minimum) proportion of the total good $\sum_i y_i$ that would need to be transferred from the individual that

currently has it to another individual in order to achieve perfect equality.

$$I_{Hoover} = \frac{1}{2n} \sum_i \frac{|y_i - \bar{y}|}{\bar{y}}$$

- *The Gini coefficient* The Gini index can be described in several ways. One way is as the average difference between the goods y_i held by all pairs of individuals, relative to the average goods $\bar{y}$.

$$I_{Gini} = \frac{1}{2n^2} \sum_i \sum_{i'} \frac{|y_i - y_{i'}|}{\bar{y}}$$

- *The Coefficient of Variation* The coefficient of variation is the standard deviation of goods $sd(y)$ divided by the mean level of goods $\bar{y}$. This can be written similarly to the two measures above.

$$I_{CV} = \sqrt{\frac{1}{n-1} \sum_i \frac{(y_i - \bar{y})^2}{\bar{y}^2}}$$

- *The Theil Index* The Theil Index does not have a simple interpretation in terms of differences in income, because it is derived from information theory.

$$I_{Theil} = \frac{1}{n} \sum_i \frac{y_i}{\bar{y}} \log\left(\frac{y_i}{\bar{y}}\right)$$

Like the ratio statistics, these are also all dimensionless quantities. If we look at the Hoover/Robin Hood Index, we can easily demonstrate this:

$$\begin{aligned}
I_{Hoover} &= \frac{1}{2n} \sum_i^n \frac{|y_i - \bar{y}|}{\bar{y}} \\
\{I_{Hoover}\} &= \frac{\{1\}}{\{2\}\{persons\}} \sum^{\{persons\}} \frac{|\{goods\} - \{goods\}|}{\{goods\}} \\
\{I_{Hoover}\} &= \frac{\{1\}}{\{persons\}} \sum^{\{persons\}} \frac{\{goods\}}{\{goods\}} \\
\{I_{Hoover}\} &= \frac{\{1\}}{\{persons\}} \sum^{\{persons\}} \{1\} \\
\{I_{Hoover}\} &= \frac{\{1\}}{\{persons\}} \{persons\} \\
\{I_{Hoover}\} &= \{1\}
\end{aligned}$$

The fact that these are all dimensionless quantities is not an accident; if you try to use a measure that is not dimensionless you will quickly discover that it inevitably violates axiom 2. For example, what if we just used the standard deviation of the income distribution as our measure of inequality? The standard deviation is a measure of dispersion—how spread out a distribution is—and surely that captures the concept of inequality? But the standard deviation has

dimensions of $\{goods\}$:

$$
\begin{aligned}
sd(y) &= \sqrt{\frac{1}{n-1}\sum_{i}^{n}(y_i - \bar{y})^2} \\
\{sd(y)\} &= \sqrt{\frac{1}{\{persons\}-1}\sum^{\{persons\}}(\{goods\} - \{goods\})^2} \\
\{sd(y)\} &= \sqrt{\{goods\}^2} \\
\{sd(y)\} &= \{goods\}
\end{aligned}
$$

This will get bigger as the total quantity of goods increases for a given set of individuals, violating axiom 2.

One of the important features of this example is that there are competing indices of inequality. There is not just one single measure that captures the concept of inequality from a collection of y_i, even once you have settled on a single good y that is measured at the individual level. There are many coherent ways to map a set of $y = y_1, y_2, \ldots, y_n$ into a single number that seem consistent with the target concept of "inequality", as specified by the axioms above.

Given that they all satisfy the axioms, how do these different indices differ? One way that they differ is in the range that they cover. The ratio statistics—the 20:20 and Palma—have a minimum inequality of 1 and a maximum inequality that is unbounded. The Hoover and Gini statistics have a minimum of 0 and a maximum of 1 (or more precisely, $1 - 1/n$). The coefficient of variation has a minimum of 0, and a maximum that increases as $\sqrt{(}n)$ with the size of the population. The Theil Index has a minimum of 0 and a maximum that increases as $\log(n)$ with the size of the population. It does not really matter whether the minimum of the measure is at 0 or 1; however the question of whether maximum inequality should depend on the population size is a more interesting question. Is a society with 10 people, only one of whom has all the goods, less unequal or just as unequal as a society with 100 people, only one of whom has all the goods? The 20:20 and Palma ratios give the same answers (∞) in both cases, the Hoover and Gini give very nearly the same answers ($\approx$ 1) in both cases, but the coefficient of variation and the Theil index both indicate that the 100 person society is substantially more unequal than the 10 person society. Different choices here could be the basis of another axiom, or a refinement of the existing axiom 6.

How strongly associated are these different indices with one another? The answer to this depends on which range of distributions of y you are interested in. In the extreme edge cases near perfect inequality, they do tend to make very different relative statements. However, most societies are not very close to perfect inequality!

Figure 6.4 shows pairwise comparisons of these six inequality measures for 1000 simulated 100 person societies, generated from symmetric Dirichlet distributions with $\alpha = 1$. This is a somewhat arbitrary choice, but it does

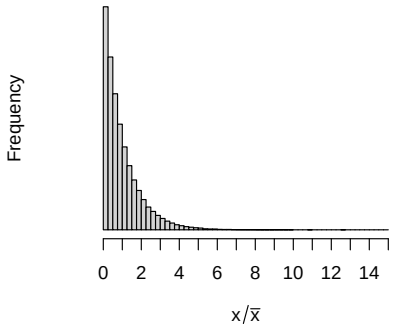


Figure 6.3: Large sample income distribution used for simulations.

generate societies with varying degrees of inequality over a moderate range of inequality, albeit without any very equal or very unequal societies.

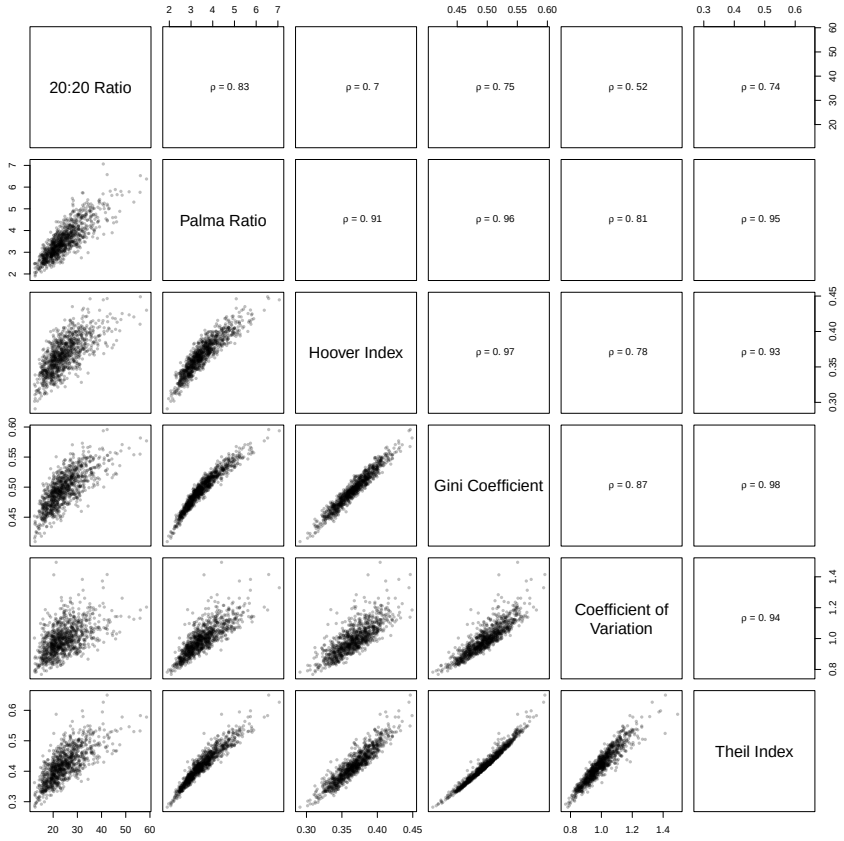


Figure 6.4: Pairwise comparisons of inequality measures for 1000 simulated societies.

For moderate ranges of inequality, the different measures are generally highly correlated with one another (although the 20:20 ratio less so than the others). This is at least a plausibility check that they are measuring similar conceptualisations of the concept of inequality, although this does not tell us that they are good conceptualisations of that concept. The case for these measures capturing the concept of inequality well has to come from the axioms, not from the mere fact that they are correlated with one another in a simulation like this.

6.5 Application - Measuring Poverty

The problem of measuring *poverty* is clearly related to the problem of measuring *inequality*, but these are distinct concepts. The most crucial difference is that the concept of poverty introduces the idea of a threshold: the idea that individuals are either poor or not poor. While one can talk about the depth or intensity of poverty (eg the poor vs the very poor), the concept aims to focus our attention on a binary distinction and an absolute threshold in a way that inequality does not. Note that the axiom 2 that we considered above for

inequality *should not* apply to *poverty*: if we make everyone twice as well off, poverty should go down, not stay the same.

The most widely used measures of poverty at a (sub-)population level are based on the idea that there is a poverty threshold z that we define ex ante. The question, then, is how we translate a set of individual incomes y_i into an aggregate measure of poverty for a set of people $i \in 1, 2, \ldots, n$ (a population). Foster et al. (1984) describe a family of poverty indices with the following functional form:

$$P_\alpha(y, z) = \frac{1}{n} \sum_i^q \left(\frac{z - y_i}{z} \right)^\alpha \tag{6.14}$$

The sum is not over all n individuals in the population, but rather over the q individuals who are below the poverty threshold, for whom $z - y_i$ is positive. The incomes of the individuals above the threshold ($y_i > z$) make no contribution to the calculation, except through increasing the population n.

The authors write that "[t]he parameter α can be viewed as a measure of poverty aversion: a larger α gives greater emphasis to the poorest poor" (p763). While α can in principle be set to any non-negative real number, the most widely used measures are those for $\alpha = 0$, 1 and 2:

$$\begin{aligned} P_0(y, z) &= \frac{1}{n} \sum_i^q \left(\frac{y_i - z}{z} \right)^0 = \frac{q}{n} \\ P_1(y, z) &= \frac{1}{n} \sum_i^q \left(\frac{y_i - z}{z} \right)^1 \\ P_2(y, z) &= \frac{1}{n} \sum_i^q \left(\frac{y_i - z}{z} \right)^2 \end{aligned}$$

Note that, as with the inequality measures, these are all dimensionless quantities. The core fraction $\left(\frac{z - y_i}{z}\right)$ is dimensionless, and so any exponent α thereof is as well. Similarly, the averaging $\frac{1}{n} \sum_i^q$ is also dimensionless.

The measure $P_0(y, z)$ is called the "headcount ratio" because it is simply the ratio of the number of poor persons to the population: the fraction of poor people in the population. This measure is insensitive to how far below the poverty line people are, the poorest poor count the same as the narrowly poor.

The measure $P_1(y, z)$ is called an "income-gap measure" because it indicates, on average, how far poor people are below the poverty line. This measure is sensitive to how far below the poverty line people are on average, but makes no distinction between a case where $y_1 = 0.1z$ and $y_2 = 0.9z$ and a case where $y_1 = y_2 = 0.5z$. They are both cases where $P_1(y, z) = 0.5$.

The measure $P_2(y, z)$ is a measure that particularly emphasises the presence of the poorest poor, as taking the square of the income gap $\left(\frac{y_i - z}{z}\right)^2$ treats incomes far below the poverty threshold as much worse that those just below. Figure 6.5 illustrates the relative contributions of different values of y_i to each of the three measures.

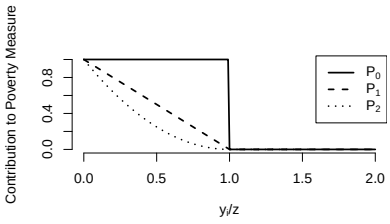


Figure 6.5: Contribution to poverty measure for different individual values, relative to the poverty line.

Foster et al. (1984) describe different axioms that these measures do (and do not) satisfy, noting that the P_2 measure satisfies additional desirable axioms that P_0 and P_1 do not. Here, we can simply note that they place different emphasis on incomes just versus far below the poverty threshold. The headcount measure P_0 is surely the most intuitive "poverty rate", but Figure 6.5 highlights that it has some potential pathologies. It puts a lot of weight on being just above vs just below z, and makes no distinctions among different levels of poverty below z. This potentially makes it more sensitive to the choice of threshold value as well, whereas P_1 and P_2 treat individuals just above versus just below the threshold very similarly, reflecting their similar incomes and experiences. In some applications the simplicity of P_0 may be sufficiently advantageous to overcome these limitations, but in others it may make more sense to use measures that capture the depth of poverty in some way.

6.6 Application - Effective Party Count

Political scientists interested in how party systems work in different countries often would like to characterize how many parties there are competing in different political systems. Some electoral systems encourage lots of parties to form while others encourage consolidation into a small number of major parties.

The obvious way to measure the number of parties competing would be to count all the parties who receive votes in an election, but it becomes immediately clear when you look at data on election results that this will not work because most countries have large numbers of very small parties which have very little consequence for politics. If you are interested in how many parties are *competing* in a meaningful sense, rather than simply how many exist, you need a measure that is designed to get at that concept, rather than how many parties/candidates are simply standing for election.[7] Otherwise your measure of the number of parties in a political system will be largely determined by how easy it is for crackpots to get on the ballot and how write-in votes are or are not reported in election results.[8]

To give an illustration of the problem, consider the UK 2017 general election. For the moment, let us consider just the parliamentary constituency of Maidenhead, which is west of London and which was the constituency of the sitting Prime Minister and leader of the Conservative Party, Theresa May. She easily won re-election, with 65% of the vote in her constituency (37,718 votes). In the UK, there is a rich tradition of eccentric minor candidates choosing the stand against the Prime Minister in his/her constituency. For the modest price of the £500 deposit that they lose when they fail to get 5% of the vote, they get to appear on television next to the Prime Minister as the results are announced on election night. Theresa May therefore had 12 opponents in her constituency, including Bobby Smith of the Give Me Back My Elmo party (dressed in an Elmo suit), Howling Laud Hope of the Monster Raving Loony party, and Lord

[7] Note for US readers: in the US candidates "run" for office, in the UK candidates "stand" for election. Whatever joke might occur to you here is unoriginal.

[8] Note the similarity to the poverty measure example. Making simple binary distinctions between poor and not poor or party receiving votes vs not receiving votes might seem superficially attractive, but it often does not yield a measure with desirable properties.

Buckethead of the Gremloids party.

Figure 6.6: Theresa May speaks upon her 2017 re-election as MP for Maidenhead constituency.

So, should we think of Maidenhead constituency as having had a 13-way competition? Or should we think of it as barely having any competition at all, given that the Conservative candidate won 65% of the vote, with the "most" competitive Labour and Liberal Democratic candidates far behind on 19% and 11% respectively?

In order to translate election results like this into a meaningful measure of competitiveness, political scientists have developed a variety of measures of "effective party count". The most widely used such measure is one described by Laakso and Taagepera (1979) which is a function of the share p_i of votes or seats (depending on application) secured by each party i:

$$N_{\text{effective}} = \frac{1}{\sum_{i=1}^{n} p_i^2} \tag{6.15}$$

This is not the only such measure that one might use, and there have been a series of subsequent papers in political science promoting alternative measures (Molinar, 1991; Dunleavy and Boucek, 2003; Golosov, 2010). We will focus here on the measure by Laakso and Taagepera, and return to some of the proposed alternatives later.

Table 6.1: Results of 2017 UK General Election in the constituency of Maidenhead.

Candidate	Party	Votes	Percent
Theresa May	Con	37718	64.8
Pat McDonald	Lab	11261	19.3
Tony Hill	LD	6540	11.2

Candidate	Party	Votes	Percent
Derek Wall	Green	907	1.6
Gerard Batten	UKIP	871	1.5
Andrew Knight	Animal Welfare Party	282	0.5
Lord Buckethead	Ind	249	0.4
Grant Smith	Ind	152	0.3
Howling 'Laud' Hope	MRLP	119	0.2
Edmonds Victor	CPA	69	0.1
Julian Reid	The Just Political Party	52	0.1
Yemi Hailemariam	Ind	16	0.0
Bobby Smith	Ind	3	0.0

If we work out the value of this measure for Maidenhead constituency, it gives an effective party count of 2.13. This is mostly driven by the top three parties (see Table 6.1). If we grouped all candidates and votes for remaining parties into a single category of "Other", the effective party count would be 2.12. The fact that there are 10 candidates sharing the last 4.7% of the vote rather than just one makes almost no difference to this measure.

The measure successfully focuses our attention on how many "serious" parties/candidates there are. In this case, very roughly speaking, there is one dominant party and two weak competitors from the Labour and Liberal Democrat parties, and the value of the measure (about 2) is meant to reflect that numerically. Arguably there really is only one competitive party in this constituency, and some of the debate about this measurement problem in the political science literature is about whether examples like this should even be given effective party counts as high as 2.13, or whether the appropriate value should be closer to 1.

Zooming out to the UK as a whole, we can do the same calculation. There were 73 registered parties[9] fielding candidates in the election, but the effective number of parties according to the Laakso and Taagepera measure is still just 2.89. The reason that it is not higher than this is that just two of those parties secured 82% of the vote (see Table 6.2). The fact that the remaining vote was highly fragmented gets very little weight in this measure.

[9] While the Monster Raving Loony party is registered party with 12 candidates in different UK constituencies, Lord Buckethead is bundled in with other independents in these calculations.

Table 6.2: Results of 2017 UK General Election by party.

Party	Votes	Percent
Conservative	13636684	42.3
Labour	12877918	40.0
Liberal Democrat	2371861	7.4
Scottish National Party	977568	3.0
UK Independence Party	594068	1.8
Green	525665	1.6

Party	Votes	Percent
Democratic Unionist Party	292316	0.9
Sinn Fein	238915	0.7
Plaid Cymru	164466	0.5
Independent	151471	0.5
Social Democratic and Labour Party	95419	0.3
Ulster Unionist Party	83280	0.3
Alliance	64553	0.2
Speaker	34299	0.1
The Yorkshire Party	20958	0.1
National Health Action Party	16119	0.1
Christian Peoples Alliance Party	5869	0.0
People Before Profit Alliance	5509	0.0
Ashfield Independents	4612	0.0
British National Party	4580	0.0
Monster Raving Loony Party	3890	0.0
Liberal	3672	0.0
Women's Equality Party	3580	0.0
Traditional Unionist Voice	3282	0.0
The North East Party	2355	0.0
Pirate Party	2321	0.0
English Democrats	1913	0.0
Christian Party, Proclaiming Christ's Lordship	1720	0.0
Independent Save Withybush Save Lives	1209	0.0
Socialist Labour Party	1154	0.0
Animal Welfare Party	955	0.0
Justice and Anti-Corruption Party	842	0.0
Southampton Independents	816	0.0
Workers Revolutionary Party	771	0.0
Workers Party	708	0.0
Something New	552	0.0
Demos Direct Initiative Party	551	0.0
Libertarian Party	524	0.0
Social Democratic Party	469	0.0
The Peace Party	468	0.0

In their paper on effective party count measures, Laakso and Taagepera (1979) list 6 axioms that they think an effective party count measure ought to satisfy. Here they are, slightly rewritten:

1. If all components have equal vote/seat shares, then the effective number must be the same as the actual number of parties: $N_{\text{effective}} = n$.
2. If all components except one have zero vote/seat shares, there is only one effective party: $N_{\text{effective}} = 1$.

3. Adding zero-share parties should not change $N_{\text{effective}}$.
4. Small changes in component shares must lead to small changes in $N_{\text{effective}}$.
5. Relabeling which parties get which indices should not change $N_{\text{effective}}$
6. Vote shares must be transformed in a consistent way $f(p_i)$ and cumulated additively such that the formula includes the expression $\sum_{i=1}^{n} f(p_i)$.

Axioms 1, 2 and 3 each identify special/limiting cases where we know what the answer should be on theoretical grounds, given the target concept. Axiom 4 is a continuity condition. Axiom 5 is an example of specifying different cases (permutations of the indices) that should yield the same measure. Axiom 6 is a functional form restriction, intended to limit the range of mathematical possibilities to relatively simple functional forms. Axiom 6 is not dictated by the concept of "effective parties", it is included by the authors for convenience.

Does the formula that we saw earlier, $N_{\text{effective}} = \frac{1}{\sum_{i=1}^{n} p_i^2}$ satisfy all these conditions?

1. If all components have equal vote/seat shares, $p_i = 1/N$ and therefore

$$N_{\text{effective}} = \frac{1}{\sum_{i=1}^{n} p_i^2} = \frac{1}{\sum_{i=1}^{n} 1/n^2} = \frac{1}{n/n^2} = n \tag{6.16}$$

2. If all components except one have zero vote/seat shares, then $p_1 = 1$ and $p_i = 0$ for all $i > 1$, and therefore

$$N_{\text{effective}} = \frac{1}{\sum_{i=1}^{n} p_i^2} = \frac{1}{1^2 + 0^2 + 0^2 + \cdots} = 1 \tag{6.17}$$

3. Adding zero-share parties does not change $N_{\text{effective}}$ because this involves adding $0^2 = 0$ to the (non-zero) denominator.
4. A small change in the component shares involves moving a small proportion of the vote/seat δ from one component to another. $\frac{1}{(p_1+\delta)^2+(p_2-\delta)^2+\cdots} = \frac{1}{(p_1^2+2\delta p_1+\delta^2+p_2^2-2\delta p_2+\delta^2+\cdots}$. As $\delta \to 0$, this $\to \frac{1}{p_1^2+p_2^2+\cdots}$.
5. Relabeling which parties get which indices does not change $N_{\text{effective}}$ because the order in which terms are added in the denominator has no consequence.
6. Vote shares are transformed in a consistent way $f(p_i) = p_i^2$ and cumulated additively such that the formula includes the expression $\sum_{i=1}^{n} p_i^2$.

Dunleavy and Boucek (2003) argue that the Laakso and Taagepera measure fails to respect an additional criteria that they (Dunleavy and Boucek) think a measure of effective parties ought to respect. Figure 6.7 shows how there is a "kink" in the Laakso and Taagepera effective party count measure in the "minimum fragmentation conditions" as the largest party's vote share falls below 50% and an additional party is added to the system. A kink is not a discontinuity where the level of the measure jumps, but is a discontinuity in the first derivative (rate of change) of the measure. Note that "minimum fragmentation conditions" implies two parties if the largest party is at or above 50%, and three parties if the largest party is below 50%. The kink in this plot

occurs around the point where there are two parties that each have 50% of the vote, where $N_{\text{effective}} = 2$. The authors are showing that if you move to a 51-49 split between these two parties, the effective parties $N_{\text{effective}} = 1.999$ barely declines, but if you instead add a very small third party to create a 49-49-2 split, the index increases much more substantially to $N_{\text{effective}} = 2.08$. Essentially, Dunleavy and Boucek are making an argument for an additional axiom. That axiom would say that not only should a measure of effective parties vary continuously as you make small changes to vote shares (which is what Laakso and Taagepera axiom 4 already specifies) but it should also vary smoothly.

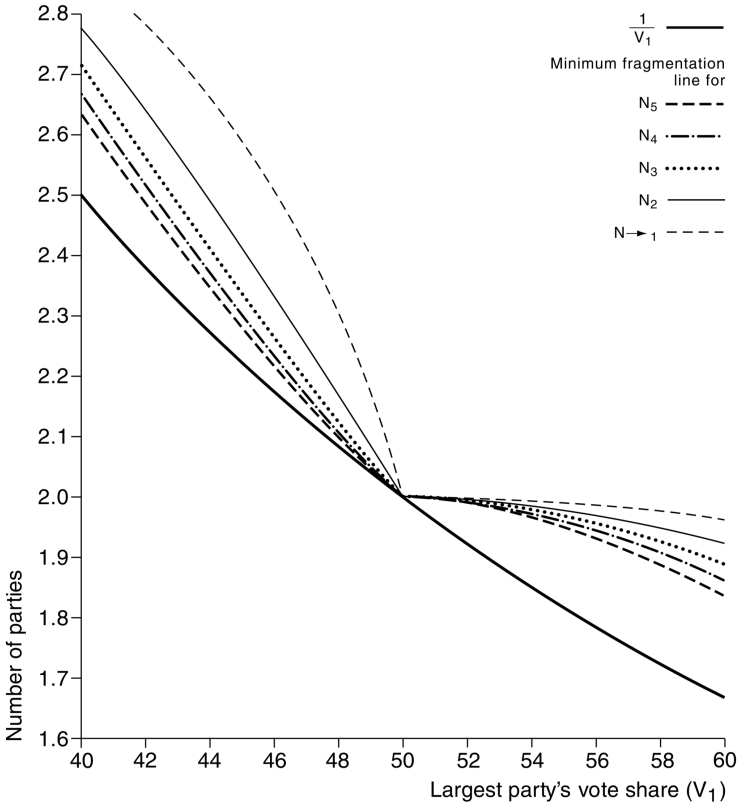


Figure 4. The behaviour of effective number of parties indices around the 50 percent anchor point, under minimum fragmentation conditions

Figure 6.7: Figure from Dunleavy and Boucek (2003), N_2 is the Laakso and Taagepera effective party measure.

My view is that Dunleavy and Boucek's argument is not very convincing because the proposed axiom is not very compelling. I do not think that an effective party count measure needs to vary smoothly across the introduction of an additional infinitesimal party into the system; the important thing is that there is no discontinuity in the measure when such a tiny party is introduced (Laakso and Taagepera axiom 3). But whether or not you think that smoothness is an appropriate axiom in this case, it is the right *kind* of criticism to make of an existing measure. Having an argument about which axioms we want to

satisfy is a good way to specify, assess and refine a measure like this one.

Let's return to the Laakso and Taagepera measure of the number of effective parties in a political system from the perspective of dimensional analysis. If you claim your measure is of an "effective number of parties", it should *not* be dimensionless. It should be in units of "number of parties". How can we check this? Let's look back at the effective party formula:

$$N_{\text{effective}} = \frac{1}{\sum_{i=1}^{n} p_i^2} \tag{6.18}$$

So how do we work out the units? First question, what are the units of a party vote share p_i? If we think about how we calculate a party vote share, it is the votes for the party divided by the total votes for all parties:

$$\{p_i\} = \frac{\{\text{votes}\}/\{\text{parties}\}}{\sum^{\{\text{parties}\}}\{\text{votes}\}/\{\text{parties}\}} = \frac{1}{\{\text{parties}\}} \tag{6.19}$$

So vote shares have units of inverse parties. We can therefore rewrite the overall formula as:

$$\{N_{\text{effective}}\} = \frac{1}{\{\sum^{parties}\}\left(\frac{1}{\{\text{parties}\}}\right)^2} = \{\text{parties}\} \tag{6.20}$$

which yields the conclusion that we were hoping for: that the Laakso-Taagepera index of effective parties indeed has units of parties.

The original paper by Laakso and Taagepera does not list having correct dimensions as a desirable criteria for their measure, nor do they ever explicitly demonstrate that their measure does have the correct units. Nonetheless, their measure satisfies the criteria.[10]

Similarly to Foster et al. (1984), Laakso and Taagepera (1979) suggest a whole family of further indices that satisfy the axioms that they have specified. They note that for any positive value of a, one could define a defensible effective party measure of the following form:[11]

$$N_{\text{effective,a}} = \left[\sum_{i=1}^{n} p_i^a\right]^{\frac{1}{1-a}} \tag{6.21}$$

For $a = 2$, this yields the effective party count measure that we have looked at above. Do all values of a and therefore all of the possible indices of effective party count yield the correct units?

$$\begin{aligned}
\left\{N_{\text{effective,a}}\right\} &= \left[\sum^{\{\text{parties}\}}\left(\frac{1}{\{\text{parties}\}}\right)^a\right]^{\frac{1}{1-a}} && (6.22)\\
&= \left[\left(\frac{1}{\{\text{parties}\}^{a-1}}\right)\right]^{\frac{1}{1-a}} && (6.23)\\
&= \left[\left(\{\text{parties}\}^{1-a}\right)\right]^{\frac{1}{1-a}} && (6.24)\\
&= \{\text{parties}\} && (6.25)
\end{aligned}$$

[10] The way that Laakso and Taagepera define their axioms guarantees that their measures will have the correct units.

[11] Laakso and Taagepera note that the $a = 0$ case yields the actual number of parties. The calculation for the $a = 0$ special case is $N_{\text{effective,o}} = \left[\sum_{i=1}^{n} p_i^0\right]^{\frac{1}{1-0}} = \left[\sum_{i=1}^{n} 1\right] = n$. The trickier special case is $a \to 1$, which has to be solved via limiting arguments because plugging in the value $a = 1$ yields an indeterminant expression. This yields an effective party number measure $N_{\text{effective,1}} = exp\left(-\sum_{i=1}^{n} p_i \log p_i\right)$ which had previously been used in the literature and which has analogies to Shannon's H measure of entropy which was mentioned above (Laakso and Taagepera, 1979, p.5).

Logically consistent units are not *sufficient* to guarantee that you have a sensible measure, but they are *necessary*. There are published measures that fail this test, including the effective party count measure proposed by Golosov (2010). Where p_1 is the vote share of the largest party, he suggests:

$$N_{\mathrm{G}} = \sum_{i=1}^{n} \frac{p_i}{p_i + p_1^2 - p_i^2} \tag{6.26}$$

The denominator of this expression appears to violate dimensional analysis because it involves adding/subtracting p_i, which has units $\frac{1}{\{\text{parties}\}}$ to p_1^2 and p_i^2, which have units $\frac{1}{\{\text{parties}\}^2}$. However, note that you could rewrite the above as:

$$N_{\mathrm{G}} = \sum_{i=1}^{n} \frac{p_i}{p_1^2 + p_i \cdot (1 - p_i)} \tag{6.27}$$

The denominator is actually ok because vote shares add up to 1 and so $(1 - p_i)$ can be rewritten as the sum of all parties other than i: $(1 - p_i) = \sum_{i' \neq i} p_{i'}$. Thus, despite initial appearances, the denominator is consistent in terms of dimensions, as it involves adding two quantities, each of which has units of $\frac{1}{\{\text{parties}\}^2}$.

However, if we analyse the units overall, we find that the measure has units of $\{\text{parties}\}^2$ rather than having units of $\{\text{parties}\}$:

$$N_{\mathrm{G}} = \sum_{i=1}^{n} \frac{p_i}{p_1^2 + p_i \cdot (1 - p_i)} \tag{6.28}$$

$$\{N_{\mathrm{G}}\} = \sum_{i=1}^{n} \frac{\frac{1}{\{\text{parties}\}}}{\frac{1}{\{\text{parties}\}^2} + \frac{1}{\{\text{parties}\}^2}} \tag{6.29}$$

$$\{N_{\mathrm{G}}\} = \sum_{i=1}^{n} \{\text{parties}\} \tag{6.30}$$

$$\{N_{\mathrm{G}}\} = \{\text{parties}\}^2 \tag{6.31}$$

This is the wrong dimension for a measure of "effective parties".

6.7 Further Examples

There are many more examples of measures that translate data into a measure of a target concept through theoretical arguments about how the calculation ought to proceed. Here are some examples.

Measures of Diversity/Concentration try to solve the inverse problem to the one solved by effective party number measures. Instead of trying to translate a set of fractions that add up to 1 into an effective number of significant blocs, they instead try to create an index of how concentrated the blocs are. That is, they are large when there is one dominant bloc and small when there are many small blocs instead of the other way around. There are two commonly used such indices, Simpson's index / Herfindahl index (Simpson,

1949; Hirschman, 1964) and Shannon's H index (Shannon, 1948). The Simpson/Herfindahl index is: $H = \sum_{i=1}^{n} p_i^2$, which is the inverse of the Laakso and Taagepera index of effective party number discussed above. Shannon's H index is $H = -\sum_{i=1}^{n} p_i \log p_i$, is the Laakso and Taagepera index discussed above, multiplied by -1. Indeed, Laakso and Taagepera (1979) discuss the relationship of their effective party number measures to both of these indices. These measures of concentration are used in a variety of applications, and have been reinvented by numerous authors across many fields.

Power Indices attempt to characterise how much *voting power* is held by individuals/parties under majority rule voting when those individuals/parties have varying numbers of votes. Here, the idea is that having a given number of votes only makes you more powerful if it means you are more likely to hold the balance of power that determines a majority. So if you have, for example, a situation like the 2015-17 UK Parliament, where the Conservative Party held 330 of 645 seats,[12] than the simple fraction of seats that they held (a narrow majority of 51.2%) is a misleading representation of their power within the parliament. Assuming party unity, whatever the Conservative Party wants will win: they have all of the power. In contrast, after the 2017 election, the Conservative Party only had 317 of 642 seats, falling just short of a majority with 49.4%, and therefore necessarily reliant on other parties to form a majority, implying a very substantial loss of voting power despite a modest loss of seats. The Banzhaf (Penrose, 1946) and Shapley-Shubik (Shapley and Shubik, 1954) power indices are two different ways of translating distributions of votes into indices that characterise the potential for different blocs to be pivotal (Banzhaf) or to be part of a governing majority (Shapley-Shubik). These are different target conceptualizations of voting power, both of which yield calculations of voting that are not simply proportional to the size of parties / voting blocs, but which depend on the different coalitions that can form to get to a majority. Different theoretical assumptions about how voting works yield different indices, and subsequent work has continued to debate the best ones for capturing the concept of voting power (Gelman et al., 2004)

[12] Excluding the Speaker (who does not vote) and Sinn Féin MPs (who do not take their seats).

Agreement and Disagreement are concepts that we might want to measure for a set of voters across a series of votes. These are very commonly used in the study of legislatures. One of the most well known of these is the Rice Index (Rice, 1928) which aims to measure *cohesion*, and which is measured in terms of the number of majority votes m_1 and the number of minority votes m_2 as $(m_1 - m_2)\,/(m_1 + m_2)$. This a positive number between 0 and 1 which is larger the more lopsided the vote is towards the majority. The highest possible measure of cohesion is therefore achieved via unanimous votes, the lowest by minimal majorities. Just as in the examples above, there are subsequent papers examining how the measurement of the concept of cohesion might be refined (Desposato, 2005). There are further approaches to measuring agreement and disagreement that involve translating voters/legislators into an ideological space and then examining their dispersion in that space.

Disproportionality measures aim to describe the extent to which a distribution of seats in a legislature is different from the distribution of votes in the election that determined the legislature's composition. Proportional representation systems minimise these differences by design, while other electoral systems (like those used in the US and UK) may lead to large discrepencies. The most commonly used measure is the Gallagher (1991) index, which is calculated from squared differences between the vote shares v_i and the seat shares s_i:[13]

$$G = \sqrt{\frac{1}{2}\sum_{i=1}^{n}(v_i - s_i)^2} \tag{6.32}$$

[13] Typically percentages rather than shares/proportions are used, yielding a 0-100 scale rather than a 0-1 scale, but I use shares here for consistency with the measures discussed above.

This is far from the only such index however. Taagepera and Grofman (2003) evaluate 19 possible indices for this concept! Why are there so many different measures? In part it is because of there are many different conceptualizations of the underlying concept. "Working on the basis of the examples I have explored here, it seems to me that the Sainte-Laguë is getting at what I would call 'disproportionality in itself'… The Loosemore-Hanby and Gallagher indices, meanwhile, appear to be closer to getting at the impact of disproportionality-in-itself on how the country is actually governed" (Renwick, 2015).

There are many different measures that try to capture the concept of ***political representation*** of voters by legislators. Note that this is closely related to the idea of disproportionality, but potentially moves beyond legislators' identities in terms of parties to other attributes like the policies they adopt and the votes that they take. Achen (1978) outlines three different conceptualizations of political representation—proximity, centrism, and responsiveness—each of which implies different measurement strategies and which is more or less sensitive to different kinds of deviations between legislators' actions and what voters want. Of course to even engage in this kind of assessment, one must address further measurement problems associated with putting "legislators' actions" and "what voters want" on comparable scales. As with the other concepts listed above, there is a rich literature of studies proposing new measures and making arguments for and against different conceptualizations of representation (Golder and Stramski, 2010; Matsusaka, 2015).

Even in the space of voting and elections that I have been exploring in the above examples, there are yet more concepts scholars have aimed to measure. These include ***party system polarization*** (Dalton, 2008), ***electoral competitiveness*** (Cox et al., 2020), ***types of environments for multiparty government formation*** (Laver and Benoit, 2015) and many others.

6.8 Conclusion

Theoretical arguments for how you ought to construct measures from indicator data are powerful when they are available. The approaches discussed in this chapter—defining axioms and then looking for simple mathematical forms that satisfy them and dimensional analysis—can be widely used to identify

how you can measure concepts that are already reasonably "close" to the data.

In the next chapter, we consider another case of data that is close to the target concept that we want to measure, but where the measurement process involves estimating a model of a particular form rather than simply specifying a fixed relationship between the indicator data to the concept of interest. In chapters after that, we will explore cases where there is not such clearly relevant indicator data for the target concept.

7
Supervised Scale Measurement using Comparison Data

This chapter discusses measurement using comparison data. The measurement problem is assessing the relative degree to which units have some concept of interest that contributes to determining the result of the competitions. The kind of data we will focus on are data which constitute direct comparisons between the units. We will call the measurement problem *scoring/ranking* and the type of data *comparison/competition data*.

It is almost impossible to avoid sports examples for this chapter, because this is the situation and kind of data that is at the core of nearly every sports competition. We want to know which individual/team is the best at and we have a bunch of data on head-to-head competitions between the individuals/teams. The problem of assessing the underlying concept of individual/team strength on the basis of competition data is lurking under the surface of the Premier League table, Association of Tennis Professionals (ATP) rankings, Fédération Internationale des Échecs (FIDE) rankings of chess players, and many other sports examples. I will mention each of these systems below at various points because they reflect familiar and potentially useful approaches to solving the underlying measurement problem.

While sport is a social enterprise, and thus plausibly in the remit of this book, competition data also arise in many other contexts. As social scientists, sometimes we can *create* competitions to help solve measurement problems where competition data do not already exist.

7.1 Wins and Losses

In many competitions, the winner is determined simply by the number of wins accumulated by each side. There are some core assumptions required in order for this model of "scoring" to make sense as a means of measuring which individuals/teams/units have more of some underlying quality or concept of interest.

Let's call this underlying quality or concept α_j, where j indexes individuals/teams/units. Let $W_{jk} = 1$ if j defeats their opponent k, and 0 if k defeats j. Then the total number of wins we would expect side j to receive in a series of

competitions is:

$$E\left[\text{Wins}_j\right] = \sum_{k=1}^{n_j} p\left(W_{jk}|\alpha_j, \alpha_k\right)$$

If we want Wins_j to be a measure of α_j, we need it to be the case that, in expectation, better individuals/teams/units (those with higher α_j) receive more wins over a series of competitions. We can guarantee that $E\left[\text{Wins}_j\right]$ is increasing in α_j via the following requirements on the structure of the competitions:

1. Individuals/teams/units that have more of the concept of interest will be more likely to succeed in the pairwise competitions: $\frac{\partial E\left[W_{jk}\right]}{\partial \alpha_j} > 0$ and $\frac{\partial E\left[W_{jk}\right]}{\partial \alpha_k} < 0$.[1]
2. Every individual/team/unit has the same number of matches n_j.
3. Every individual/team/unit has opponents with the same distribution of strengths $f(\alpha_k)$.

[1] Note that $E\left[W_{jk}\right] = p(W_{jk} = 1)$. The expected number of wins from a single competition is just the probability of winning that competition.

The first requirement is a reminder that we cannot measure just anything we want from competition data. If your want to measure which individuals/teams/units have a higher propensity to win competitions, competition data is obviously a sensible kind of data to use. If, however, you want to measure some other concept that is not determining the outcome (eg which individuals/teams/units are the most sportsmanlike, or something like that) this is not a good measurement strategy for that concept.

The second assumption again seems kind of obvious. If one individual/team/unit had only had five competitions, and another had fifteen, it would hardly be fair to assess their strength based on which one won more matches![2] The underlying motivation for the second assumption is that all individuals/teams/units should have similar oppurtunities to succeed.

[2] One could use the proportion of, rather than count of, wins to partially address this issue, but that would give a winner of a single match an insurmountable record.

The third assumption listed earlier is the most interesting, and is the one that some competitions used in professional sport actually fail. Sports league competitions are sometimes structured with balanced schedules so that the competition is *strictly* fair. For example, the English Premier League (an example used later in this chapter) involves every team facing every other team twice, once at home and once away. Indeed the size of the division is essentially determined by the number of matches that can be scheduled in a season. The logic of this is that it gives all of the units/teams/individuals an equally difficult set of competitions, because they all face the same set of other units/teams/individuals under conditions that are as similar as possible.[3]

[3] Most US sports leagues have league/division structures with unbalanced schedules that undermine direct comparisons of win totals. The NFL has too many teams relative to the length of the season, making balanced schedules impossible, and necessitating a convoluted playoff qualification structure to make the competitation vaguely fair.

If all three of these conditions are met, then the win count will measure the relative strength of the different individuals/teams/units versus one another. Obviously we cannot learn about the relative strength of these individuals/teams/units versus other units that were not in the competition. This standard of strict fairness of the schedule is difficult to meet without small leagues and long seasons.

7.1.1 Adding Draws

We can use this logic to think about how to handle cases where there are draws/ties in individual match-ups. The obvious way to do this is to count draws as intermediate between a win and a loss. While you might do this by treating them as half a win for each side, in practice it is more common to define a point system so no one has to cope with fractions. If you redefine W_{jk} above as the number of points received instead of the number of wins, the listed conditions still supply a fair competition and allow you to use the points as a measure of the quality of the teams. At least if, in expectation, increasing the quality α_j increases the number of points that j receives and increasing the quality of their opposition α_k decreases the number of points that j receives.

The obvious way to achieve this is to say that a win is worth 2 points, a draw is worth 1 point, and a loss is worth 0 points, and then count these up at the end of the competition. From a measurement perspective, the logic of counting the draw as halfway between a win and a loss seems very sensible because it ensures that the same number of points are awarded for each match, and so if all the units have the same number of matches, the total number of points to be awarded is fixed in advance and does not depend on the results.

If you know anything about sports leagues with point systems, you know that they often **do not** count a draw as intermediate between a win and a loss. In domestic and international football competitions, it is standard to award 3 points for a win, 1 point for a draw, and 0 points for a loss. This means that 3 points are awarded in total for matches which end in a win for one of the teams, and only 2 points are awarded for those which end in a draw, 1 for each side. Other leagues have even more convoluted point systems. In the National Hockey League in the US and Canada, two points are awarded for a win, one point for losing in overtime or in a shootout, and zero points for a loss in regulation time. These systems make little sense from the perspective of measuring the quality of teams. They are designed to incentivise certain strategies and disincentivise others. The football system encourages teams to avoid draws. The hockey system (bizarrely) encourages teams to have draws in regular time and then to go for victory in overtime when they have nothing to lose.

These systems are still *fair* with respect to the comparison of teams across the season, in that all teams face the same incentives in every match, and have the same number of oppurtunities to gain points. They also do not undermine our ability to use points as a measure of team quality, although the incentives around encouraging/avoiding draws arguably change the definition of team quality.

This kind of "fairness in distribution" strategy does not work well for all ranking tasks. Consider the problem of determining who are the strongest tennis or chess players. These are individual sports where there are too many individuals to have everyone play everyone else regularly. If you did select

opponents randomly from a very large pool, most of the competitions would be so lopsided as to be uninteresting as competitions. Tennis competitions are designed to make the strongest individuals play each other towards the end of knock-out tournaments, because it is more interesting for the spectators to watch that way. But then naively using win counts would be a bad way of assessing which tennis players are stronger. How can we measure which individuals/teams/units are strongest if we observe a very unbalanced set of competitions?

7.2 Rating Transfer Systems (ELO)

Chess ratings use a rating system called Elo, named after their inventor Árpád Imre Élő, who was a physicist and strong amateur chess player (born in Hungary in 1903, emigrated to the US in 1913).

The core idea of this kind of system is that everyone starts with an endowment of points, which is your rating. Whenever you face an opponent in a match, depending on the pre-match ratings and the result of the match, some number of points are transfered between the two opponents. You gain more points for a better result, but the amount of points that you gain is larger if your pre-existing rating was worse relative to your opponent. This means that you can rise in the ratings rapidly by defeating highly rated opponents, but a highly ranked individual cannot gain much rating by repeatedly defeating weak competition. Two evenly rated opponents will trade points equally: if two chess grandmasters with ratings of 2400 face one another, each will stand to gain the same number of points from a victory.

One advantage of this type of system is that decentralised calculation is possible, so long as everyone is honest. If we both know our rankings going into a match with one another, we can figure out our rankings after the match, and no one needs to keep track of all the matches centrally. This means that the system works even with very large numbers of competitors.

There are some well-known disadvantages to schemes like this. They are sensitive to grade inflation over time as new entrants introduce more points to be redistributed. For this reason, initial ratings for new players are tricky to specify. If you trust everyone, you only need to know the current state of the system, but you need a full, trusted history of time-stamped matches to actually recreate the current rankings.[4] Finally, you have to make a judgement call in setting up the system regarding how many points will be at stake in each match, as this determines how sensitive the system is to what has happened very recently versus long-run performance. If too many points are at stake in each match, ratings will be excessively volatile; if too few points are at stake in each match, ratings will take a very long time to reflect changes in the performance of the competitors.

We are not going to go into any detail on such systems, as they are rarely used for social science applications, but they begin to develop the key intuition

[4] This is a rare problem for which the solution is actually is a blockchain.

behind the models we will look at for the rest of the chapter. That intuition is that not all wins are equally impressive. If you want to accurately measure strength from data that involve an imbalanced schedule of competitions, you need to explicitly model how results wins/losses/draws vary depending on the quantity that you are interested in measuring.

7.3 Bradley-Terry Models

We are now going to embed this logic in a statistical model for the outcome of competitions. We want to connect the unobserved quantity that we are interested in measuring ("strength" or "propensity to win") to the observed data (wins, losses and draws). The class of models that we will be using was first described by Bradley and Terry (1952), and are extremely simple. We assume that each team/individual/unit j has a strength in competition that is described by a single parameter α_j.

There are several ways to parameterise such a model, but we are going to focus on one with an underlying link to logistic regression models.[5] We then assume that the log-odds of the competition results are determined by the difference between the parameters for the two sides:

$$log\left(\frac{p(j \text{ defeats } j')}{p(j' \text{ defeats } j)}\right) = \alpha_j - \alpha_{j'} \tag{7.1}$$

Figure 7.1 illustrates the implications of this functional form. The more positive the difference between α_j and $\alpha_{j'}$, the greater the probability that j wins. The more negative the difference, the greater the probability than j' wins. If $\alpha_j = \alpha_{j'}$, both j and j' are equally likely to win.

The Bradley-Terry model is so simple, in fact, that it is just a special case of a logistic regression model. In order to fit a Bradley-Terry model using a logistic regression, we define $Y = 1$ to correspond to a victory of j over j', and $Y = 0$ to correspond to a victory of j' over j. We then define a set of indicator variables, one for each individual/team/unit, which equal 1 when that individual/team/unit is j and -1 when that individual/team/unit is j'.

In order to be able to fit this logistic regression model, we need to exclude one of the units, so that we are estimating the strength of all other individuals/teams/units relative to that one. The reason for this is that, if you look at Equation 7.1, we can add any constant number to all the α_j parameters without changing any of the model predictions. This means the absolute levels of the α_j parameters are arbitrary, only the differences between them matter. This is an example of a scale which is interval-level but not a ratio-level: numerical differences are meaningful, but the zero point is not.

In addition to excluding one of the units, one can also exclude or include the logistic regression intercept. In sports competitions, there is often a "home" side and an "away" side to the competition. In many sports there is a home side advantage, where there is a clear pattern that home sides are more likely

[5] The original statement of the model was linear, rather than logistic, where $p(j \text{ defeats } j') = \frac{\alpha_j}{\alpha_j + \alpha_{j'}}$, for positive α. Alternatively, in the spirit of linear probability models for binary data, one can fit $p(j \text{ defeats } j') = \alpha_0 + \alpha_j - \alpha_{j'}$, with α_0 either set to 0.5 or estimated to account for asymmetries like home side advantage. Analogously to linear probability models, this potentially yields invalid predictions for $p(j \text{ defeats } j')$ that are outside the range from 0 to 1 if the variation in the strength of the different j is large. But where the variation in the strength of j is small, this linear model has the advantage of yielding α values that are easily interpretable as differences in probability of winning.

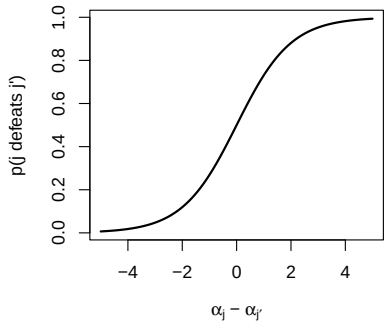


Figure 7.1: Probability that j defeats j' under a logistic Bradley-Terry Model.

to win. If you include the intercept in the logistic regression, and always code the "home" side as j and the away side as j', then the intercept will estimate the extent of home side advantage. If you exclude the intercept, you are assuming there is no home side advantage.

The fact that the Bradley-Terry model is just a logistic regression with dummy variables for each team, coded positive or negative to match the coding of the dependent variable, immediately suggests a variety of extensions to the basic model. If we want a version of the model that can cope with draws, we can use an ordinal logistic regression model:[6]

$$log\left(\frac{p(j \text{ defeats or draws} j')}{p(j' \text{ defeats } j)}\right) = \gamma_0 + \alpha_j - \alpha_{j'} \quad (7.2)$$

$$log\left(\frac{p(j \text{ defeats } j')}{p(j' \text{ defeats or draws} j)}\right) = \gamma_1 + \alpha_j - \alpha_{j'} \quad (7.3)$$

[6] This extension was proposed by Rao and Kupper (1967) but neither the original Bradley-Terry model nor this extension are explicitly related to binary and ordinal logistic regression models by their original authors because they predate the general statements of the binary logistic regression (Cox, 1969) and ordinal logistic regression (McCullagh, 1980) models.

If we want to include covariates, we can do that by adding them to the binary/ordinal logistic regression, keeping in mind that we need to code them in a way that makes sense given how we have defined the outcome. Following on the point about home side advantage above, we can include this in the model in two different ways. First, as suggested above, we could simply define the dependent variable as the home side winning, coding the indicator for the home side as +1 and the indicator for the away side as −1. In that case, the overall intercept for the logistic regression model becomes the home side advantage parameter, because it corresponds to the log-odds of the home side winning when the α parameters for both sides are equal:

$$log\left(\frac{p(\text{home defeats away})}{p(\text{away defeats home})}\right) = \beta_{advantage} + \alpha_{\text{home}} - \alpha_{\text{away}} \quad (7.4)$$

Alternatively, we could exclude the intercept, and include a +1/-1 variable home_j that takes on the value +1 if j is at home and -1 if j' is at home. This approach would be useful if some competitions are on neutral ground, so that we could instead define a three-level variable that is +1 if j is home, 0 if on neutral ground, and −1 if j' is at home.

$$log\left(\frac{p(j \text{ defeats } j')}{p(j' \text{ defeats } j)}\right) = \alpha_j - \alpha_{j'} + \beta_{advantage} \cdot \text{home}_j \quad (7.5)$$

This approach works for other covariates as well, we just need to remember to use +1/-1 or +1/0/-1 codings rather than +1/0 codings, so that the model yields the same predictions regardless of which side you chose to list as j and which you chose to list as j'. For example, imagine you wanted to estimate whether sports teams were at a disadvantage if they had played more recently than than the other side. You might code a variable that was +1 if j had played more recently than j', 0 if they had played equally recently, and −1 if j' had played more recently than j. Then, the coefficient $\beta_{\text{short rest}}$ on that variable

would indicate the advantage/disadvantage of being on shorter rest than the other side.

7.4 Application - 2018-19 English Premier League Season

The English Premier League consists of 20 teams playing home and away against every other team in the league, for a total of 38 matches per team. The Premier League champion Manchester City secured 98 points on the basis of 32 wins, 2 draws and 4 losses, while Liverpool came second on 30 wins, 7 draws and 1 loss. Note that the 3-1-0 point win-draw-loss system was pivotal, on a 2-1-0 points system Liverpool would have won the league by 67 to 66.

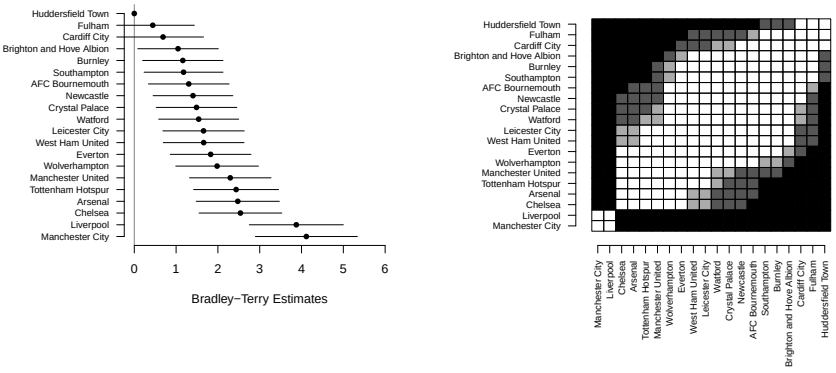


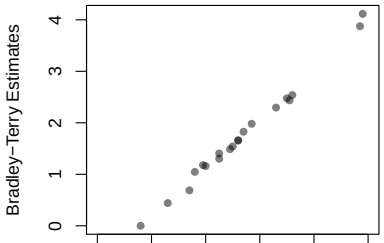
Figure 7.2: Bradley-Terry model estimates for Premier League 2018-19 season (relative to Huddersfield). Right: Hypothesis test results for pairwise comparisons of teams, where the four shades from white to black correspond to $p > 0.1$, $0.1 > p > 0.05$, $0.05 > p > 0.01$, and $0.01 > p$, respectively.

If we estimate an ordinal Bradley-Terry model on these data (which have many ties/draws) with estimated home side advantage, we get the estimates shown in the left panel of Figure 7.2. Note that all the estimates, and also the confidence intervals, are relative to the omitted side, which is the lowest ranking Huddersfield Town. Because the confidence intervals are relative to the omitted side, they are not very helpful for determining the pairwise comparisons of other teams, so the right panel of Figure 7.2 shows the results of pairwise hypothesis tests for whether teams have equal strength parameters α.

The top two sides, Manchester City and Liverpool are indistinguishable from one another, and are significantly stronger than Chelsea in third. The estimate for Chelsea, which came third, is not significantly different from the teams ranking first through eighth (Everton) at the 0.1 significance level or the tenth (Leicester City) at the 0.05 level. Leicester City, situated in mid-table, is not significantly different from the teams ranked third to seventeenth at the 0.05 level. In a 38 match season, there is a lot of room for good or bad fortune to intervene, we cannot be very confident about which teams were *really* better or worse.[7]

As the plot on the right of the figure shows, with a fully balanced schedule, the point totals from the 3-1-0 system are very highly correlated with the Bradley-Terry estimates. The correlation between the two measures is 0.996.

[7] By "really" better or worse, I mean which teams be more successful in a hypothetical infinite season in which the teams each played each other an infinite number of times, but nothing else about the teams changed.



If you have a perfectly balanced schedule of competitions, each side facing each other side the same number of times, point systems work very well. Nonetheless, there are advantages to fitting a measurement model like the Bradley-Terry model. First, we get the confidence intervals, which are relevant if you want to think about how confident we should be that the best team won the Premier League, or questions like that. Second, we can calculate predictions. Given their performance over the season, what are our predicted probabilities that Manchester City defeats Liverpool head-to-head?

To calculate this, we need to numerical values of the coefficients from the model. The key pieces are the values of α for Manchester City and for Liverpool, estimated at 4.115 and 3.876 respectively, plus the intercepts for each of the two model equations, which are estimated as -0.909 for Away Win versus Draw or Home Win and 0.142 for Away Win or Draw versus Home Win. Therefore, if we want to calculate the predicted probabilities, we need to solve the ordinal logistic regression equations:

$$log\left(\frac{p(M \text{ defeats or draws } L)}{p(L \text{ defeats } M)}\right) = -0.909 + 4.115 - 3.876 \quad (7.6)$$

$$log\left(\frac{p(M \text{ defeats } L)}{p(L \text{ defeats or draws } M)}\right) = 0.142 + 4.115 - 3.876 \quad (7.7)$$

Note that, because I fit the model with the home side as j and the away side as j', this estimates the probabilities for a match held in Manchester. If we want to calculate the probabilities for a match held in Liverpool, we have to swap the sides, which does matter because the two intercepts are not symmetric around 0:

$$log\left(\frac{p(L \text{ defeats or draws } M)}{p(M \text{ defeats } L)}\right) = -0.909 + 3.876 - 4.115 \quad (7.8)$$

$$log\left(\frac{p(L \text{ defeats } M)}{p(M \text{ defeats or draws } L)}\right) = 0.142 + 3.876 - 4.115 \quad (7.9)$$

In the former case, where Manchester City is at home, the probabilities work out to 0.52 for a Manchester City victory, 0.24 for a draw, and 0.24 for a Liverpool victory. In the latter case, where Liverpool is at home, the probabilities work out to 0.34 for a Manchester City victory, 0.26 for a draw, and 0.41 for a Liverpool victory. You can see from this that home side advantage in the Premier League is substantial.[8]

Alternatively, we can make some kind of statement about just how much better the best teams are than the worst teams. The predicted probability of Manchester City defeating Huddersfield Town, in a match where Manchester City plays at home, is 0.98. Moving the match to Huddersfield helps very little, Manchester City is still predicted to win with probability 0.96. The gulf between the top and bottom of the Premier League in 2018-19 was vast.

It is obvious how we talk about the units of wins or of points, but what are the units of the Bradley-Terry estimates? The straightforward answer is the

[8] We are estimating a general home side advantage, not one that is specific to particular teams. You could do the latter either by having intercepts that vary for each team in the model, or by estimating different strength parameters for each team depending on whether they are at home or away, essentially treating them as two distinct teams.

correct one. The coefficients of a binary or ordinal logistic regression model are log odds-ratios, so the Bradley-Terry estimates are log-odds ratios of better results versus worse results. When you fit this model, you are deciding to measure the strength of each side according to their log-odds of getting better, as opposed to worse, outcomes in competition with one another. You may not love log-odds as a unit, but they are a good unit of account for competition data for all the same reasons they are a good basis for a limited dependent variable model with binary or ordered categorical outcomes. They have a relatively simple mathematical form and they translate in a straightforward way into valid predictions. The number of wins or the number of points may have an even greater virtue of simplicity, but they do not tell you anything quantitative about who is likely to win the next match.

7.5 Application - 2019-20 English Premier League Season

The English Premier League, like nearly all football leagues, uses a balanced schedule of competition. However, the 2019-20 season was interrupted by a pandemic caused by the novel coronavirus SARS-CoV-2. This stopped all competition on 13 March 2020, with 288 of 380 scheduled matches completed (76%). All teams had completed 28 or 29 of their scheduled 38 matches, but this meant that some teams had faced weaker competition than others, simply because more of their matches against stronger teams remained uncompleted. For a while, it was unclear whether the season could be finished, and therefore which teams should be relegated to a lower league (the bottom three) or be able to compete in the European Champions League (the top four) in the next season? Given that the standings as of the suspension of the season did not even have all sides with the same number of matches, let alone the same strength of schedule, surely this would not be a fair basis on which to allocate the very substantial financial rewards of staying in the Premier League as opposed to being relegated, or entering the Champions League as opposed to not? One commentator despaired:

> "What precise form of points per game you use is a really interesting argument because once you accept that a simple points per game might not quite be adequate, and you want to start weighting it, and you say that they have six home games left and only four away and maybe we should do it home points and away points, why stop there? Why is that the dividing line? Why don't you move one further down the line and say well actually they still have to play all of the top six whereas this team had none of the top six left to play. And that then gets incredibly complex and of course you cannot at this stage make a fair assessment of that unless you get some boffins from North Korea who have no idea that the Premier League even exists and lock them away in a biosecure environment and you say to them 'What is the fairest way of doing this?' and they come out with the equivalent of the Duckworth-Lewis charts and say this is the way we do it, and whack it in the algorithm. But the problem is Duckworth-Lewis was invented looking at past performance and calculating probabilities but there was not a game going on at the time, so you didn't have an immediate knowledge of what the impact

would be. So it's an almost impossible thing." Jonathan Wilson, Guardian Football Weekly Podcast, Thursday 21 May 2020.

The Bradley-Terry model is one such algorithm. As we specified it in the preceding section, and applied it to the 2018-19 season, we can also apply it to the partial results of the 2019-20 season up to the halt of competition due to the pandemic. Figure 7.4 shows the Bradley-Terry estimates in the same way as before. While we can be pretty confident that Liverpool was by some distance the strongest team in 2019-20, the shortened season leaves a great deal of uncertainty about the relative rankings of every other team in the Premier League.

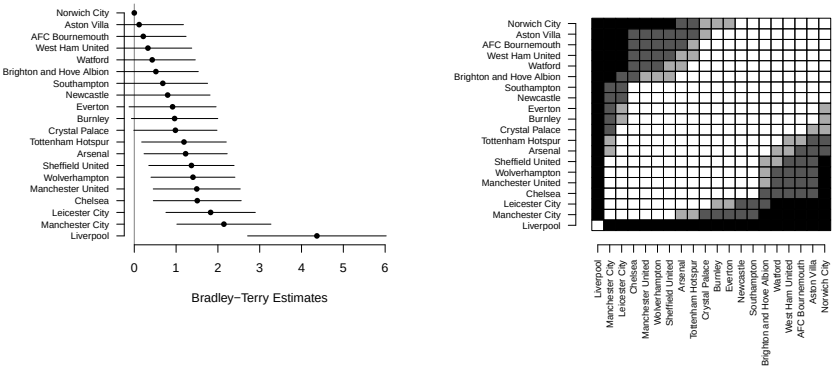


Figure 7.4: Bradley-Terry model estimates for Premier League 2019-20 season (relative to Norwich). Right: Hypothesis test results for pairwise comparisons of teams, where the four shades from white to black correspond to $p > 0.1$, $0.1 > p > 0.05$, $0.05 > p > 0.01$, and $0.01 > p$, respectively.

Because the Bradley-Terry model is a *probability model* that relates parameters α to the observable outcomes of matches, it can make predictions about the results of the unplayed matches. For each of the unplayed matches, we can construct predicted probabilities from the model, in the usual way that we do so for (ordinal) logistic regression models. From this, we can construct expected end of season point totals for every team.

For example, one of the next matches that was to be played after the pandemic halted the season on 9 March 2020 was between Manchester City (home) and Arsenal (away) on 11 March 2020. What does the model say about the likely outcome of that match? The predicted probabilities for that match are 65% chance of a Manchester City win, 22% chance of a draw, and 13% of an Arsenal win. Given these probabilities, we can calculated the *expected points* that each team would have gained from this match, which is just $3 \times p(Win) + 1 \times p(Draw)$. These work out to be 2.2 points for Manchester City and 0.6 points for Arsenal.

If we tally up the total points that all the teams had secured from the matches they had played at the time the season was suspended with their expected points for the remaining matches, we get the following league table:

Team	Matches Completed	Points	Expected Points
Liverpool	29	82	107.3
Manchester City	28	57	77.6
Leicester City	29	53	70.2
Chelsea	29	48	62.2
Manchester United	29	45	60.9
Wolverhampton	29	43	58.2
Sheffield United	28	43	57.9
Tottenham Hotspur	29	41	54.4
Arsenal	28	40	53.4
Burnley	29	39	50.4
Crystal Palace	29	39	49.4
Everton	29	37	48.3
Newcastle	29	35	46.1
Southampton	29	34	44.6
Brighton and Hove Albion	29	29	36.9
West Ham United	29	27	36.4
Watford	29	27	36.1
AFC Bournemouth	29	27	34.0
Aston Villa	28	25	32.3
Norwich City	29	21	28.6

The incomplete schedule turned out not to be very consequential for ranking the different Premier League teams. The expected order of the teams at the end of the season matched the current point totals. However, the model did break some ties between teams with the same number of points. Most consequentially, at the time that the season was suspended, AFC Bournemouth, Watford and West Ham United were tied for sixteenth to eighteenth place in the league with 27 points and 29 matches completed. Since the bottom three teams are relegated from the league under normal circumstances, which of these teams is placed eighteenth is particularly important. Based on the Bradley-Terry model at the time that the season was halted, AFC Bournemouth was the side that we should expect to secure the fewest points in their remaining matches, and thus to be relegated to a lower league.

7.5.1 Measurement and Predictive Uncertainty

While the calculations for the expected point totals for all teams give a single prediction, there was of course substantial uncertainty in how the end of the season would have actually played out in the absence of the pandemic. There are two quantifiable sources of uncertainty that we need to think about in cases like this where we want to use measurements to make predictions about future outcomes.

First, there is *predictive uncertainty*. Individual matches yield specific results: either one side gets 3 points and the other 0, or both get 1 point. Even if the estimates of the probabilities of these results are sound, football matches can go one way or another, and a side can have a good run in their remaining 9 or 10 matches relative to expectations, simply through good luck. This is quantifiable uncertainty, as we can simulate the results of each match, with the predicted probabilities, and see what the distribution of results for the rest of the season looks like if you simulate match outcomes from those probabilities.

Second, there is *measurement uncertainty*. The estimates of the relative strengths of teams are measured with substantial uncertainty. Some of the teams are *actually* stronger and some are weaker than the point estimates of α imply. Whereas predictive uncertainty reflects the potential for teams to be lucky or unlucky in future matches, measurement uncertainty reflects the fact that some teams will have been lucky or unlucky in the completed matches that we used to estimate the model. Again, this is quantifiable uncertainty, because we have fit a model that tells us the uncertainty in all of the model parameters. As discussed below, there are ways to incorporate this uncertainty in simulations of the results for the rest of the season.

In addition to these quantifiable sources of uncertainty about how the season would actually play out, there are unquantifiable sources of uncertainty related to the adequacy of the Bradley-Terry model we are using. Remember, this is a very simple model with strong simplifying assumptions. All teams have a single strength parameter that applies throughout the season, neither getting stronger nor weaker, even if they reach the end of the season with little to play for. All teams have the same home side advantage. These assumptions are very likely wrong, although perhaps not wrong enough to matter very much at all. Given the limited data in a single season, we cannot do much to relax them, but if one was embarking on a broader analysis of football results, one might wish to assess the evidence regarding how accurate these simplifying assumptions are across many seasons. We will set aside these concerns here, but it is important to acknowledge that they exist.

In order to the assess the magnitude of predictive uncertainty we can simulate the remainder of the season a large number of times, using the predicted probabilities calculated from the ordinal logistic regression coefficient point estimates. In order to assess the magnitude of predictive *plus* measurement uncertainty, we replace the coefficient point estimates with draws from a multivariate normal distribution with mean equal to those point estimates, and variance matrix equal to the estimated variance-covariance matrix of the model coefficients (King et al., 2000). The square roots of the diagonal elements of this variance-covariance matrix are the standard errors of the regression coefficients.

Figure 7.5 shows 95% intervals around the predicted point totals for each Premier League side, based on simulating the remaining matches of the 2019-20 season after the suspension of play on 13 March. The degree of predictive

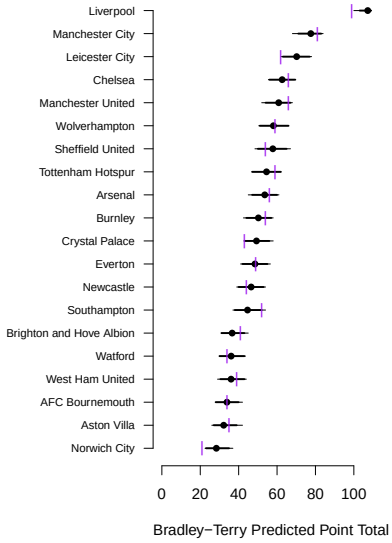


Figure 7.5: Predictions for final point totals of Premier League 2019-20 season based on Bradley-Terry model estimates. The thick error bars include only predictive uncertainty, the thin error bars include both measurement and predictive uncertainty. The final outcome of the season after restart is depicted with a purple vertical line for each team.

uncertainty is shaped by the number of matches remaining and how well the model predicts match outcomes; the estimation uncertainty is mostly shaped by the number of matches already completed. Nonetheless, incorporating measurement uncertainty (the thin error bars) in addition to the predictive uncertainty only slightly increases the width of the intervals, even in this case where we have only a modest number of past matches for each side (28 or 29). The major source of uncertainty about how the season would actually play out is that individual matches have uncertain outcomes, and teams might be lucky or unlucky in how those matches turn out.

In the end, the Premier League season was successfully concluded, in empty stadiums, between 17 June and 26 July 2020. The final point totals are included as vertical tick marks in Figure 7.5. The final totals are mostly within the 95% intervals obtained when the season was paused. We would expect one in twenty to fall outside their 95% intervals, and in fact there are three: Liverpool, Norwich and Leicester. Liverpool and Norwich had nothing to play for at the end of the season, the former because they had secured the league title, the latter relegation. As noted above, this kind of change in motivation is not something the model is designed to capture, although in principle with data from many seasons it would be possible to estimate the consequences of playing out the end of the season in such circumstances.

7.6 Designing Competition Data Collections for Measurement

Ok, enough about sport. Is this kind of model useful for social science? Yes, because sometimes the best way to measure an unobserved quantity is to setup comparisons that are responsive to that quantity. Let's say you want to figure out which political parties are further to the right and which are further to the left, across Europe. One thing you might do is ask some experts on European political parties to rank each party position as 0, 1, 2, 3, etc on a 0 - 10 left-right scale. That is a really difficult question to answer, and to answer consistently across countries.

What if we asked for pairwise comparison instead? Obviously it is a much easier question to answer whether the UK Labour party is to the left of the UK Conservative party than to put these parties on a vague 10 point scale. But it also might be easier to answer the question of whether the UK Labour party is to the left of the Irish Labour party or the German Social Democrats or other relevant comparisons of parties cross-nationally. Those are not trivial assessments, but they are a better test of whether it is possible to make meaningful cross-national comparisons of this type at all. If your experts cannot make these kinds of pairwise comparisons, their 0-10 scale scores are definitely useless. In contrast, experts might be able to make meaningful binary comparisons without being able to generate valid 0-10 scores. Making pairwise comparisons is a less demanding task.

There are other applications where asking some set of respondents to make

pairwise comparisons is useful. Loewen et al. (2012) want to assess which arguments are strongest in a Canadian political referendum. So they use a Bradley-Terry model to analyse data from a general population survey where respondents are given a comparison between two arguments randomly selected from a larger pool of possible arguments, and asked to indicate which argument is stronger. Again, you could ask people to directly assess the strength of the arguments, but it is probably easier to have them make pairwise comparisons between arguments than to evaluate single arguments against some abstract scale of argument quality.

Zucco Jr et al. (2019) use a Bradley-Terry model to assess which ministerial roles in government are valued more highly in Brazil. Rather than trying to get direct assessments of 37 different ministerial roles, they surveyed legislators and other experts on Brazilian politics, giving each randomly generated pairwise comparisons between ministerial roles, asking respondents to "choose the ministry they thought a typical politician would prefer to obtain" for his/her party in a coalition negotiation. They estimate, very plausibly, that the Finance, Health and Education ministries are all near the top, while Tourism, Culture, Sports and Fisheries are near the bottom. They also note that this kind of data collection is both engaging and quick to complete for respondents:

> Survey instruments with pairwise comparisons are impressively user-friendly. Several ABCP colleagues reported to us that the survey was "fun," which may explain why 273 of the 278 participants opted to continue on past the first eight pairwise comparisons. Median time to completion for the entire expert survey was just over 4 min.

This example highlights that pairwise comparisons are useful for extracting quantititive measures in areas where there is as yet unquantified "common knowledge" held by some relevant population of people who you can survey. This could include which legislators are more or less competent, which London tube stations are more or less pleasant to travel through, and many other applications. In any case where you are contemplating asking people to *rate* a set of units (eg on 0-5 or 0-10 scales) it is worth considering whether it would be better to have them make pairwise *comparisons* of pairs of units instead.

Hopkins and Noel (2021) use pairwise evaluations of US Senators to estimate Senator ideology as perceived by surveyed political activists. For each randomly selected pair of Senators, they asked "Which of these two politicians is more liberal?" or "Which of these two politicians is more conservative?"[9] They then ran these pairwise comparisons through a Bradley-Terry model to generate summaries of activist perceptions. Figure 7.6 shows a comparison of these scores to another set of ideology scores for US Senators, which is a summary measure of legislative votes taken in Congress. We will cover the methods for generating measures from such data later in Chapters 12 and 17.

While these two different measurement strategies do not recover exactly the same relative ideological positions—if they did, all the points would be on the diagonal line—they are highly correlated both across and within party. There

[9] The form of the question was randomised, to ensure that there was not an asymmetry between perceptions of liberalism and conservatism. The data are then recoded so that the "winners" of the comparisons are consistently one ideological label or the other, regardless of which question was asked.

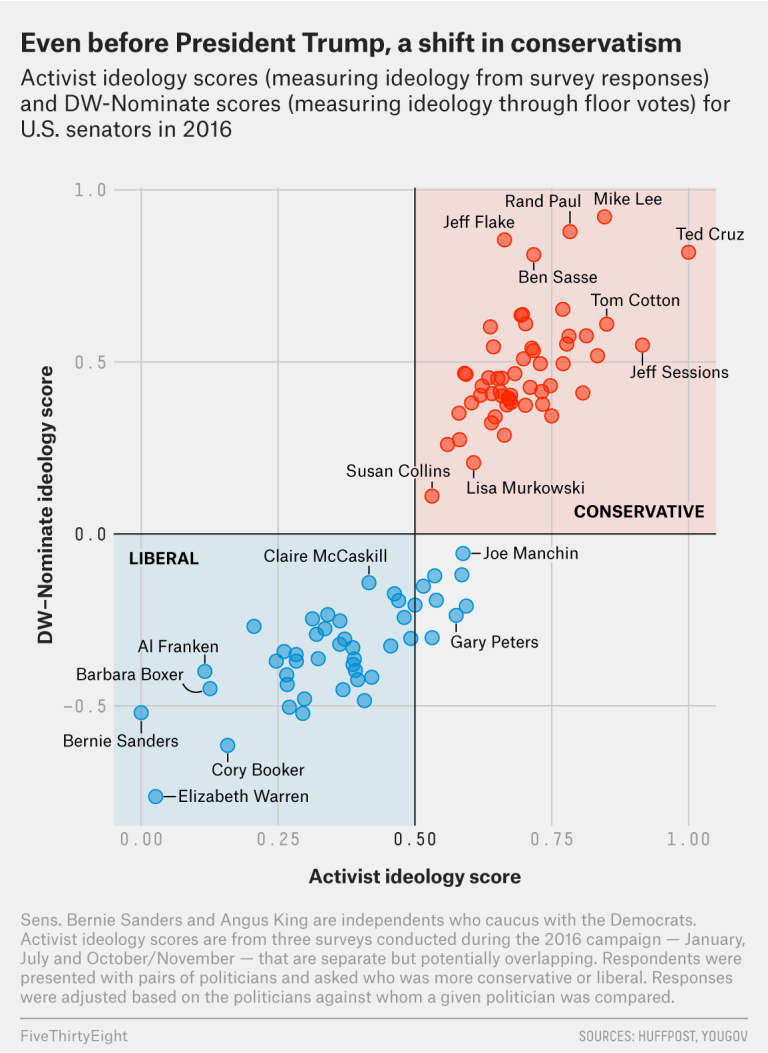


Figure 7.6: Comparison of Bradley-Terry estimates of US Senator ideology based on political activist perceptions (x-axis) and a summary of US Senator voting behaviour in Congress (y-axis). https://fivethirtyeight.com/features/how-trump-has-redefined-conservatism/

are nonetheless some differences which are interesting. While Democratic and Republican Senators overlap in ideology according to activists' evaluation—some Republicans (eg Susan Collins) are to the left of some Democrats (eg Joe Manchin)—by legislative voting there is no overlap. In their article, Hopkins and Noel (2021) discuss what we learn from some of the Senators whose "activist ideology score" and their voting behaviour score ("DW-Nominate ideology score") differ substantially. They note that these cases where the fact that these measurements are summarising different kinds of data really matters: some Senators are perceived in public differently from how they vote. Sometimes these discrepencies are likely to be intentional on the part of the Senator, sometimes not.

7.6.1 How much data do you need?

First and foremost, this depends on how "big" the differences between units are: if the "stronger" units that you are studying almost always "win", across the range of units in your data, you do not need as much data as if the stronger units only win somewhat more often. Second, you should think in terms of a number of comparisons involving each unit of interest. You might need 25 or 50 or 100 or many hundreds comparisons involving each unit, depending on how big the differences between units are. Thus, larger sets of units require larger sets of comparisons. Note that while this is fairly demanding in terms of data, the good news is that it is proportional to the number of units n and not to the number of possible pairwise comparisons $n^2 - n$, so you are not completely out of luck in medium data sets. Nonetheless, this is a measurement strategy that works best with small to medium numbers of units that you want to put on a scale relative to one another. Strategies for extending the pairwise comparison idea to larger data sets where it is not feasible to conduct enough pairwise comparisons involving each unit that you want to measure are discussed in Chapter 9.

7.6.2 How should the competitions be structured?

Balanced competitions are those where you observe all pairwise comparisons the same number of times. If there are too many possible pairwise comparisons, a balanced competition can be approximated by selecting pairs of units at random. In cases where you can just barely generate enough comparisons for the number of units, it can make sense to do adaptive testing to avoid re-running the competitions where you already know the result. You do not learn anything from having Serena Williams (2015 edition) play someone ranked 1000th in the world, because Williams will always win. Similarly, you don't learn anything from having your expert/crowd coder make really obvious comparisons—is the United Kingdom or the Democratic People's Republic of Korea more democratic?—except maybe whether they are paying attention.[10]

You cannot push the adaptive testing idea too far though, as you do need all

[10] Owing to weak labelling standards for countries, the country with "Kingdom" in the name is more democratic than the country with "Democratic" in its name.

the units you want to compare to be connected by competitions. You cannot use a Bradley-Terry model to assess the relative strengths of two groups of units that never face one another (eg two different sports leagues). And if you have only a few "bridging observations" between two groups, you will be very uncertain about their relative strengths. The good news is that the estimation will automatically account for this, you will see the problem in the standard errors (or, in the extreme case of completely disjoint competitions, the regression will fail to fit or will drop coefficients). This is one reason that using a measurement model is attractive: it tells you how much you learned from the data.

7.6.3 How can I be sure I am measuring the right thing?

If you are setting up competitions to solve a measurement problem (as opposed to working with competition data that already exist) you need to make sure that people are answering the question that you intended them to answer. This requires clarity, but it also requires relevant knowledge or competence. It might be that you expect them to already have that knowledge, as in the example of Brazilian legislators indicating which ministries they would prefer (Zucco Jr et al., 2019). It might be instead that you expect them to make an evaluation that you give to them, as in the case of comparing Canadian referendum arguments (Loewen et al., 2012). If you asked members of the general public which ministries were most attractive to legislators, many of them would have no idea. A wide variety of further survey design considerations will apply in specific applications, including limitations on how much information you can expect people to process and social desirability biases.

7.6.4 The latent variable that you just made up is not a real thing in the world

The Bradley-Terry model is our first example of the broader class of "latent variable models". We will see many more. It is a very simple latent variable model that is easy to understand, which makes it a good basis for talking about the most common conceptual error that people make when interpreting latent variable models.

What makes the Bradley-Terry model a latent variable model is that we have *hypothesized* a variable—call it "quality" or "strength" or "propensity to win competitions"—that describes each individual/team/unit. That variable is not observable directly; it is latent. But we *assume* that it predicts the wins and losses (and draws) that we do observe.

Note that it is very easy to imagine that this is meant to be a representational model. That is, that there is a *real thing* in the world that we are calling the "strength" or "quality" of the individual/team/unit, and that thing is determining the outcomes. If we are going to use these estimates for prediction of future competitions, it must be, right? If we are claiming to measure something, that something must be a real thing that already exists, right? Many

people have a strong intuition that the answer to these questions must be yes, but the answer to both is no.

Imagining that the latent variable model that you specified *represents* real causal attributes rather than merely approximating them is one of the most frequent and damaging classes of errors that people make when working with latent variable models. The Bradley-Terry model is a *pragmatic* measurement model, it should not be understood as *representative* measurement.

The Bradley-Terry model, if interpreted representationally, is an implausible monocausal[11] story for the outcomes of competitions. In some sense, the model already acknowledges this because it is probabilistic: the "stronger" unit does not always win the competition. So the model only attempts to measure one "factor", but acknowledges that other things must also matter. We will see this repeatedly in the coming chapters as we explore further latent variable models: we will often try to measure one (or two or three) "factors" that might predict an outcome, and then treat everything else as noise. It is nonetheless important to recognise that just because you hypothesize a monocausal explanation for something, that does not make it true. Even if your model predicts "pretty well", that still does not make it true.

[11] In tennis, Rafael Nadal and Roger Federer spent years fighting close matches on hard courts while Federer almost always won on grass courts and Nadal always won on clay courts. Why did this happen? Because tennis involves a multidimensional skillset (serve, forehand, backhand, pace, spin, mobility, etc) and some of those skills are more valuable on some court surfaces than others.

But if these are not representative models, in what sense have we "measured" something? What use is this very simplified and stylized description of the data? The answer is that it is useful precisely because it is simplified and stylized. If you follow sports, you will know that it is very natural to talk about and assess which sides are stronger, even though you know that strength in competition consists of much more complicated details. If you think about other comparisons we might make, they are similarly reductive. Imagine the last two restaurants you have been to. Which one is a better restaurant? That is a question that you might ask, and might be answered consistently or inconsistently over time by one person or across people, but it is clearly reducing a number of underlying attributes into a simple comparison.

Models like the Bradley-Terry model are models for pragmatic measurement. The estimates from these model are simple summaries of features of the world (the outcomes of competitions) but they do not represent something that exists already in the world. Measuring the relative "strength" of Manchester City and Liverpool football clubs in a given year is a pragmatic measurement task. In contrast, measuring the number of people who show up to a match is a representational measurement task. The former is aiming to simplify a great deal of complexity down to a couple of numbers; the latter is aiming to provide numerical representation of something that exists whether you measure it or not. These are both measurement tasks, and we can do them more or less well, but they have distinctive characteristics as problems to solve. It is important to be clear about which one you are doing.

7.6.5 Common sources of measurement error

Finally, let us think about measurement error in the context of models for competition data. This is going to be a recurring theme in the coming chapters, using the ideas that we developed in Chapter 5.

First, let's think about the variance of Bradley-Terry estimates. In this context, variance reflects how much our estimates would vary around whatever value we would measure, with the model we are using, if we were able to run an infinite number of competitions. High measurement variance comes from having insufficient data. If we do not have enough data on individuals/teams/units, we will recover imprecise estimates. Recall that high imprecision/variance means that if we went out and ran new competitions, and nothing about the competing units had changed, our measurement might still change a lot.

What does bias mean in this context? Bias is how the measurements tend to deviate from the thing we actually wanted to measure, on average. Consider some examples. Old data can cause bias. For example, performance in last year's competition is likely to be an imperfect predictor of performance this year, perhaps in systematic ways (eg by age, younger individuals/teams getting better in expectation, older ones getting worse). Data that was generated by a different process than the one we intended can cause bias. If we ask people to code which of two political candidates is more charismatic, but they actually just code which one they would vote for, we will measure something about the latter rather than the former, which is a form of bias.

One of the strengths of competition data can be a very close connection between the data and the concept. If we define the concept as "which units tend to win competitions like the ones that we observe" then observing data about who wins is really the best data we can hope for. Note that this is not true when we start to move towards using this as a measurement strategy for human/expert/crowd coded data. If we ask people to make pairwise comparisons on the basis of some concept, we will learn how they think about that concept in this application, but that doesn't necessarily make their conceptualization of that concept correct in any more general sense.

8

Supervised Scale Measurement using Regression

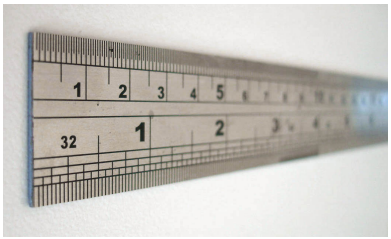
Where do the marks on a ruler come from? Why are we confident that they match the marks on other peoples' rulers? What defines the length of a (centi)meter (or a foot, or any other unit)? From 1983-2019 the definition of a meter was the length of the path travelled by light in vacuum during a time interval of $\frac{1}{299792458}$ of a second.[1] If this is the "gold standard", what gives you license to make claims about the length of things in meters on the basis of a simple straight-edged ruler made out of plastic, wood or steel?

The key feature of a ruler is that it is *calibrated* to match the gold standard. The marks on the side of a ruler tell you how to use it as an *indicator* of length. A typical ruler will not have been calibrated directly against the speed-of-light based standard, but against some more convenient approximation of that standard. However the calibration was conducted, its result is that a given number of marks on the ruler closely corresponds to a given distance in the conventional units.

Properly used, a ruler will yield a measurement that is only accurate to a certain degree of precision. There is measurement error, from imperfections in the manufacture of the ruler, from the fact that the marks on the ruler have width, and from the process of actually holding it up to what you wish to measure and attempting to see exactly which marks on the ruler correspond to the distance you are interested in. You would not expect to be able to accurately measure lengths at the micrometre (10^{-6}m), let alone the nanometre (10^{-9}m) scale. But for everyday objects and many purposes, accuracy to within a millimetre (10^{-3}m) or even a centimetre (10^{-2}m) is good enough, and the cheapness and ubiquity of a straight-edged ruler makes it an excellent measurement instrument for such tasks.

In this chapter, we consider the much messier, but nonetheless analogous, production of social science measurement instruments in the context where we have a gold standard for what we want to measure (like the international definition of the meter) and wish to use more readily available indicators (like a long piece of plastic, wood or metal) to measure that same concept. How do we put the lines on the ruler?

[1] The original meter was defined relative to the size of the earth in 1798. From 1799 to 1960, the meter was defined in terms of a series of canonical objects made out of platinum and iridium, stored in Paris. From 1960 to 1983, the meter was defined in terms of a number of wavelengths emitted by a particular electron state transmission in Krypton 86. In 2019 the definition was amended to clarify the definition of a second, which is itself defined in terms of electron state transmissions in Caesium 133. see https://en.wikipedia.org/wiki/Metre



8.1 Training with Continuous Data

Consider the case where we have a set of "gold standard" measurements m from some pre-existing measurement procedure for the concept of interest μ. These measurements should either be on the desired measurement scale, or some transformation thereof. This is often called the *training* data, which we are using here to *calibrate* a new measurement procedure. Our new measurement procedure will be based on a set of one or more indicators I (I_1, I_2, etc) that we want to use to to measure this concept. Our goal, then is to determine how to most effectively use them to approximate μ, given the indicator variables I that we have, plus the information contained in m about how they relate to μ.

If m are on the scale on which we want to measure μ, we have the most straightforward case to implement. Constructing the best approximation of m using I is straightforward: this is what regression does. If we use *linear* regression:

$$m_i = \alpha + \beta_1 I_{1i} + \beta_2 I_{2i} + \cdots + \epsilon_i \tag{8.1}$$

Recalling the definition of linear regression, a linear regression of m on I will give us the best linear approximation $\hat{m}$ of m using I, which is to say the one that minimises the sum of squared errors. Equivalently, it minimises the mean square error:

$$MSE = \frac{1}{N} \sum_i (m - \hat{m}_i)^2 \tag{8.2}$$

Note that the errors/residuals ϵ_i from this regression are the measurement error $\epsilon_m = m - \mu$, as we have defined it previously.

Our new measure of μ for a given unit i is then the fitted value $\hat{m}_i = \hat{\beta}_0 + \hat{\beta}_1 I_{1i} + \hat{\beta}_2 I_{2i} + \cdots$. Thus, through the simple application of regression, we have a measurement instrument that uses only the $\hat{\beta}$ that we estimated in this calibration exercise plus the indicator values I for the units for which we want to measure μ.

It is important to note that one need not limit the regressions considered to linear regressions using the set of indicators. Within the framework of multiple regression, one can include interactions of indicators or non-linear functions of indicators. More generally, one can use semi-parametric or non-parametric regression methods. The goal is the generic goal of all regression methods: to use the indicators to best approximate the target concept, which is to say, to minimise mean square error $\frac{1}{N} \sum_i (\hat{m}_i - m)^2$.

When thinking about how to apply this rich variety of regression methods to this kind of measurement task, the usual considerations of "supervised learning" from machine learning apply. The primary interest is in the quality of the measurement, not in the individual indicators, which is to say we care about our estimates of the concept of interest $\hat{\mu}$ rather than about the estimates of any parameters like $\hat{\beta}$ that link μ to the observable indicators I. Parametric models will lead to particularly transparent measurement procedures, where

the contribution of individual indicators are simply describable in terms of their coefficients β. Where there is a limited size training dataset, moving from linear models to more flexible models will yield limited predictive gains, but where there are more substantial training data sets the tension between transparency of how such measures are constructed and the quality of the resulting measurements may be more severe.

8.1.1 Applicability

There are two key things that must be true to make this approach useful:

1. We have a gold standard measure m[ˆgoldstandard] of the target concept μ for some units, but lack that measure for other units.
2. We have one or more indicators I that predict the target concept μ for all units.

Note that these hold in the motivating example of the straight-edged ruler. There is a gold standard definition of length (the meter) involving the speed of light, but we cannot easily apply it to all measurements we want to make. However, we can calibrate a measurement instrument (the ruler) using the gold standard and then use that measurement instrument more widely.

Even assuming that you are in the appropriate context to construct a measurement procedure in this way, the quality of the measures derived from this approach rely on three key assumptions, stated qualitatively here:

- **Training Data Quality**: It needs to be the case that the deviations of the gold standard measure m from the target concept μ are small and are not associated with quantities relevant to the intended application.
- **Representative Training Set**: It needs to be the case that differences in the relationship between the indicators and the target concept, for the units in the calibration set versus the population where you want to apply the measurement procedure, are small and are not associated with quantities relevant to the intended application.
- **Indicator Quality**: It needs to be the case that the indicators are sufficiently predictive of the gold standard measure such that the residual error of the regression is small and is not associated with quantities relevant to the intended application.

In each case, the resulting errors can be problematic either because they are large in magnitude *or* because they induce errors associated with quantities relevant to likely applications of the measure. If the errors are large in magnitude, the consequence is that the resulting measure is a very noisy approximation of the target concept, leading to imprecise unit-level assessments, and potentially leading to attenuation biases in aggregate analyses. If the errors are associated with quantities relevant to likely applications, the measurement errors will generate biases in subsequent analyses, for the reasons discussed in Chapter 5.

What, then, do we need to be particularly attentive to, with respect to the quality of the training data, the representativeness of the training set, and the quality of the indicators?

There is sometimes very little flexibility when choosing your gold standard / training data. It is rare that one has much control over the properties of the gold standard measurement procedure, or a choice among multiple such gold standard measurements. More often the problem is the lack of any training data at all. Nonetheless, it is vital to remember that your new measurement strategy can be no better than the training data that you use to develop it, because that is its only connection to the target concept. If your gold standard does not deserve that name, the even noisier measure you calibrate using it can only be worse.

Just as you will not usually have much choice in which training data to use, you will only sometimes have any choice regarding the observations that you can use to learn the relationship between indicators and the target concept. The ideal is a sufficiently large random sample from the target population to which you aim to apply the measurement procedure that you are developing. This is possible in some applications, but more typically it is not. In many applications, where the measurement is meant to be used prospectively, the assumption is that the past relationships hold in the (near) future. The worst case, which is not uncommon, is where there are systematic differences between the units in the training data that you use for calibration of the procedure and the units to which you aim to apply the measurement strategy.

Finally, and here you typically have more control, there is the question of which indicators to use and how to model their relationship to the gold standard / target concept. Here is where you can use all the tricks you know for improving prediction: interactions, non-linear models, etc. Usually the indicators you have available are the key constraint, fancier models are seldom worth *a lot* by comparison to better (ie more predictive) indicators, but they can help a bit. Some target concepts are easier to approximate with available data than others, there is not much you can do here other than look for new and better indicators (that are nonetheless available for all units). Note that if the indicator set is missing key aspects of the target concept, those aspects will be missing from the measure.

This measurement strategy assumes that you already know how to measure the concept of interest reasonably well, but for some reason you are not able to do so for all the units that you are interested in. Thus, this kind of approach is most useful in a few types of situations:

1. When the training data is costly to construct. You have a very large set of units about which you know a few things (the indicators) and a small set of units for which you can invest effort to construct better measurements
2. When the training data is only available for the past. You have a set of units for which you observed the indicators and the target concept in the past, but

you want to measure the target concept for observations where it has yet to be realised.
3. When the training data is only available for a different population of unit than the one you are interested in. You have a set of units for which you observed the indicators and the target concept, and want to use this to measure the target concept for a different type of unit.

Note that the second and third kind of application present immediate concerns about whether the training set is representative: past relationships between target and indicators may not reflect future ones, and relationships in one population of units may not reflect those in another.[2]

[2] Arguably these are the same type of situation, defining the two populations with respect to time as opposed to other criteria.

8.1.2 Validation

The information available with which to do validation of these regression-based measurements will vary by application, but there is one source of information that is always present by virtue of the setup: the training data measurements themselves. Model fit for the training data is often the best available validation standard. The familiar R^2 statistic describing proportion of variance explained in the sample of data used to fit the model is a problematic measure here for the usual reasons: it rewards overfit models that are likely to predict poorly out-of-sample. We are precisely interested in out-of-sample fit in this instance, since the core of the measurement strategy is constructing predictions for new observations from a fitted model. We do not care about the magnitude of measurement error for the training data, we care about the magnitude of measurement error in the population of units to which we will apply the measurement strategy. Adjusted R^2 provides a more suitable estimate of *population* variance explained, for the population from which the training data were drawn. One can form *population* estimates of a variety of measures of fit—eg mean absolute error (MAE) or root mean square error (RMSE)—using cross-validation.

How much model fit can tell you about measurement error in the population of units to which you will apply the measurement strategy depends on the relationship between the training data and that target population. The best case are applications of the first type described above—where the training data is simply costly to construct—because in these cases one can generate training measures for a random sample of the target population. In these cases, unbiased estimates of model fit for the gold-standard are unbiased estimates of model fit for the target population as well. In cases where the training data is available for a different time period or for a different population of units, model fit statistics calculated using the training data may or may not give you a good sense of the magnitude of measurement errors for the measurements generated on the target population.

8.2 Training with Binary/Categorical Data

There are many applications where we want to measure a continuous quantity μ with a set of indicators I, but lack training data on the scale we want to measure. In some of these situations, however, we have binary or ordered categorical data m that is closely associated with the concept we want to measure. In these cases, we can use these training data to similarly generate continuous scale measures using logistic (or other categorical response) regression models.

This is a particularly useful approach where we want to measure a concept that is something like the "propensity to have $m = 1$ as opposed to $m = 0$" for some binary training data m or the "propensity to have higher values of m rather than lower values m for ordinal data. Note that in these instances, even though our training data is categorical, we are still aiming to measure a continuous quantity. Closely related strategies for measuring binary/categorical quantities will be discussed in Chapter 10.

If we use binary logistic regression, we have the regression equation:

$$\log \frac{p(m_i = 1)}{p(m_i = 0)} = \alpha + \beta_1 I_{1i} + \beta_2 I_{2i} + \cdots \tag{8.3}$$

Our measured scale could then be either on the log-odds scale (running from $-\infty$ to ∞)

$$\hat{m^*} = \alpha + \beta_1 I_{1i} + \beta_2 I_{2i} + \cdots \tag{8.4}$$

or on the probability scale (running from 0 to 1):

$$\hat{m} = \frac{exp\left(\alpha + \beta_1 I_{1i} + \beta_2 I_{2i} + \cdots\right)}{1 + exp\left(\alpha + \beta_1 I_{1i} + \beta_2 I_{2i} + \cdots\right)} \tag{8.5}$$

Which of these is more relevant will depend on the application, and whether it is more desirable to have a measure with a probability scale interpretation or a measure that is a linear function of the indicators.

Whereas with continuous training data m, where we can use nearly any regression method, here we do need models for binary data that generate probabilistic predictions $p(m = 1)$ as opposed to merely generating binary classifications $m \in 0, 1$ for specific units. This is nonetheless a very substantial set of methods to choose from. There are trade-offs between what is possible given the size of the training data sets, the number of indicators, and so on.

8.2.1 Applicability

The potential problems associated with training using continuous gold-standard data all apply to training with binary/ordered categorical data as well. We still need to worry about the quality of the gold standard data and any potential discrepancies between those data and the target concept. We still need to worry if the relationships between the indicators and the gold standard data in the training data set is unrepresentative of the relationship in the population to which the measurement procedure will be applied. We still need

to worry about whether our indicators are sufficiently predictive of the gold standard, and whether we have chosen the best possible approximation of their relationship through indicator selection and appropriate functional form.

To take a typical application, say that you want to measure (the concept of) someone's risk of defaulting on a loan. You take a data set of people who got loans in the past, build a predictive model of who defaulted based on characteristics that you can measure before the loan was given, and calculate the fitted value for people trying to get a loan now to give each a "loan score". What are you doing here other than **measuring** the (concept of) propensity to default on a loan? This example highlights some of the potential problems that one needs to look out for in applying this approach. You are measuring propensity to default given that someone thought it was reasonable to give that person a loan at the time. If you change the criteria for giving out loans, the propensity to default as a function of indicators may not stay the same. The issue here is a potential violation of the assumption that the relationship between the indicators and the target concept is the same for the training (gold standard) data and the data for which you want to make predictions (new measurements). This is particularly likely in this example, because the selection into the training set was conditioned on expectations about the concept that we are trying to measure: default risk. A similar selection problem arises when trying to study whether tests used in university admissions predict performance in university: you only observe university performance for those students who are admitted. Those students will be unrepresentative of the broader applicant pool and the relationship between pre-admissions test performance and university performance that holds among matriculating students may not be the same as across the entire applicant pool.

The advantage of using binary or ordered categorical training data is that it potentially makes supervision possible for a much larger number of concepts. Even if there is no existing categorical training data, in some applications it is possible to create the training data by surveying relevant experts. Sometimes you do not even need experts, random members of the public can be expected to have an idea of how to map the available indicators into the presence versus absence of the concept you want to measure. This process of "crowd-sourcing" your training data is sometimes used in quantitative text analysis, where the indicators are features of a text—the presence of words or combinations thereof—and the target concept is something that a human might be able to perceive in a written text. This might be the use of emotional language, the use of populist arguments, or any of a very wide variety of concepts. With a sufficient number of human codings of whether these concepts are present or absent, it becomes possible to train a measurement model for the presence or absence of these concepts.[3]

[3] In other applications, these data are used for validation of existing measurement strategies.

8.3 Application - Election Outcomes on Alternate Geographies

The 2016 referendum on EU membership in the UK ultimately resulted in the UK leaving the EU in early 2020. In the interim, there were two UK general elections in 2017 and 2019, called "early" in response to the political complications created by the referendum result. Many academic researchers, and political analysts more generally, were interested in how vote shifts in these elections were shaped by voters' referendum preferences, but the way in which the referendum vote was tallied and reported in 2016 meant that there were no official figures for how the referendum vote was distributed at the level of UK parliamentary constituency. The 2016 referendum was reported at the level of *local authorities*, of which there are 380 in England, Scotland and Wales.[4] That same area includes 632 parliamentary constituencies, each of which sends one MP to Parliament.

[4] The referendum results in Northern Ireland were reported on parliamentary constituency boundaries.

The analysis we are going to do here to generate the missing measures is a simplified version of the analysis done by Hanretty (2017). Hanretty's estimates are the most widely used measures of EU referendum vote on parliamentary constituency boundaries. Our analysis here will ignore the geographic overlap of the different areal units, as that is beyond our scope, but see the original paper for further discussion of how to use this information to further improve the estimation strategy. Our version of the analysis will proceed as follows:

1) Fit a regression model predicting 2016 leave share in the 380 local authorities using a selection of demographic variables measured at the local authority level.
2) Construct fitted values from the regression model for all 632 constituencies in England, Scotland and Wales using those same demographic variables measured at the constituency level.

This analysis is based on calibrating the relationship between a set of indicators and the target of the measurement and then extrapolating to a new set of units. As discussed earlier, in order to evaluate how well this is likely to work, we need to consider three potential problems.

First, how good is the gold standard measurement? In this case, it is excellent: it is the official return of the election at the local authority level. There is nothing to worry about here.

Second, is the relationship to demographic variables similar in the units on which we *train/calibrate* the model (local authorities) to the relationship in the units on which we will apply that model to construct fitted values (parliamentary constituencies)? This is a more subtle question than the first two. Local authorities are somewhat larger on average than constituencies and are also much more variable in their population sizes. They reflect enduring political entities with their own local governments (ie councils) to a greater degree than constituencies, which are substantially redrawn much more frequently. Even though they describe the same land area and set of voters overall, the Modifi-

able Areal Unit Problem means that it is possible that the relationships between vote and demographic variables at the local authority level are different from those at the constituency level. If the UK had US-style gerrymandering of constituencies, we might worry about whether constituencies were designed to have particular political alignments, but the general consensus is that UK parliamentary constituencies are not aggressively gerrymandered (and in any case, these boundaries had been in place for several election cycles and were mapped in a period when leaving the EU was not viewed as a serious proposition). These potential sources of error are real possibilities, although (in the author's estimation) both are unlikely to be terribly severe in this instance.

Third, how strong is the relationship between the indicators and the target of the measurement? Is support for leaving the EU strongly predicted by available indicators for local authorities, or not? This is something we can determine by taking the indicators we have to work out, and seeing how well they predict the vote. Let's find out.

We proceed by fitting a model predicting the Leave vote percentage in each of the 380 local authorities, with the following variables from the 2011 UK Census: median age, proportion white, proportion owning rather than renting their homes, proportion with no formal educational qualifications, proportion with university degrees or equivalent (Level 4+), proportion with "Higher managerial, administrative and professional occupations" (NSSec 1), proportion "Small employers and own account workers" (NSSec 4), proportion "Lower supervisory and technical occupations" (NSSec 5), and indicator variables for Scotland and Wales. Note that there are many variables one could use at this stage, and a variety of variable selection strategies, which are discussed later in this chapter.

This particular regression has 10 explanatory variables and achieves an adjusted R^2 statistic of 0.89 and a residual standard deviation of 3.46. The R^2 statistic suggests that the model "fits well". The residual standard deviation gives us an estimate of magnitude of the measurement error, *if our application were to local authorities*. Given a ±2 standard deviation rule of thumb, this suggests that most of the measurement errors would be less than 7 percentage points in the Leave vote share.

It is often difficult to validate the magnitude of errors from this kind of regression-based measurement strategy, especially in a case like this where we are training on a different type of unit than we ultimately want to measure. However, in this instance we have some additional information: both have the estimates reported by Hanretty (2017) and also exact known results for 27 constituencies (see discussion in the cited paper).

Figure 8.1 shows comparisons of our regression estimates to the results in the few known constituencies (left), of the Hanretty estimates to those same results (left-center) and the comparison of our regression estimates and Hanretty's estimates in those few known constituencies (right-center) and all constituencies (right). In the left panel, we see that in the 27 known

	Model 1
(Intercept)	46.19***
	(3.64)
MedianAge	0.26
	(0.16)
White	−7.31**
	(2.52)
Own	25.62*
	(11.47)
No.qualifications	24.35*
	(10.65)
Level.4.qualifications.and.above	−167.21***
	(11.42)
NSSec1	183.78***
	(22.39)
NSSec4	54.80**
	(16.75)
NSSec5	214.34***
	(31.14)
ScotlandTRUE	−17.00***
	(0.82)
WalesTRUE	−3.66***
	(0.86)
R^2	0.89
Adj. R^2	0.89
Num. obs.	380

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 8.1: Statistical models

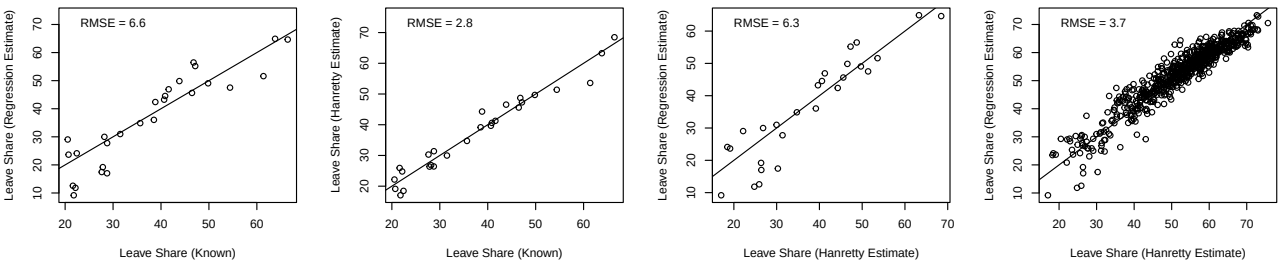


Figure 8.1: Regression estimates (left) and Hanretty estimates (left-center) of UK parliamentary constituency leave vote share in 2016 referendum, versus known parliamentary constituency results. Comparisons of regression estimates with Hanretty estimates in the subset of known parliamentary constituencies (right-center) and all constituencies (right).

constituencies, the root mean square error[5] (RMSE) is 6.6 percentage points of Leave vote share. If we think of this as an estimate of the standard deviation of the measurement errors, it illustrates that the measurement error resulting from this analysis is still non-trivial, despite the seemingly high R^2 of the regression model. Given a $\pm 2SD$ rule of thumb, this suggests that most of the measurement errors will be less than 13 percentage points.

The measurement error in the known constituencies is thus substantially larger than what the residual standard deviation estimate suggested. This could be because the latter reflect fit to local authority results and there really are substantial differences in the demographic associations with vote in parliamentary constituencies. However, there is some reason to think that the known constituency result analysis overstates the magnitude of the errors across all constituencies. The right two panels of Figure 8.1 show that the discrepancies between the regression estimates we have developed and Hanretty's estimates are atypically large in the known constituencies (which are not a random sample).

We also see from a comparison of the left two panels that Hanretty's estimates are clearly better than what we were able to achieve with this simple regression strategy. The additional information that Hanretty uses comes from the geographic overlaps between the 380 local authorities and 632 constituencies and also the strict adding up constraints for total votes. We have not used these here because they are application-specific features of this particular problem: areal interpolation is just one illustration of the method, albeit one where one can go further than our regression analysis. Nonetheless, the core of Hanretty's analysis is a similar regression strategy of estimating the relationship between local authority demographics and referendum vote and then applying that relationship using the same demographic variables measured at the parliamentary constituency level.

How important are errors of this magnitude? Here it is difficult to say anything in general. If you are interested in specific constituencies—for example, if you are a campaign deciding which seats to target in an election—you might care a great deal about an error of 10 percentage points. A seat with a Leave vote of 45% is a rather different prospect for the Conservative party than one with a Leave vote of 55%. If you are instead a social scientist, interested in patterns across all constituencies, it is less clear that this level of error is problematic. A great deal will depend on whether the patterns of measure error are uncorrelated with the other variables used in such analyses. In this situation, there is no particular reason to expect that measurement errors will be correlated with most variables of interest, as they primarily reflect deviations of real aggregate voting behaviour from a demographic model, which do not immediately suggest any important patterns of bias. This is not to say that such biases will definitely not exist.

We do not always have the benchmarks that we had in this example with which to validate regression-based measurement strategies like this one, but

[5] $RMSE = \sqrt{\frac{1}{n}\sum^{n} (\text{estimate} - \text{target})^2}$

even here we can see that validation often involves a collection of imperfect indications about the quality of the measurement. We can use predictive performance for the training data, but that does not directly apply to the target population unless the training data were a random sample. We can sometimes use other benchmarks, but they may themselves be measured with error (Hanretty's estimates) or be incomplete (the known constituency results). This means that there is a core role for qualitative reasoning about likely errors, just as we might qualitatively reason about plausible sources of sampling biases for population inference or selection biases for causal inference. We have been doing these kinds of assessments throughout the preceding discussion, but the key questions to keep in mind are those raised by the three key assumptions that we discussed at the outset of this chapter. What is the quality of the training data? How representative is the training data of the target population? How predictive are the indicators?

8.4 Application - Turnout Propensity

As discussed earlier in this chapter, one class of regression-based measurements looks a great deal like exercises in prediction. In these applications, a model is trained on past data, and used to predict future data. This kind of predictive exercise is the core of most machine learning discussions of regression, and is covered from that perspective in many excellent books (eg James et al., 2013). Here, we are going to think about the properties of the fitted values or predicted probabilities as a measurement in themselves.

One example of this kind is where we want to measure an individual characteristic of "propensity to turnout in a general election" for a large number of individuals about whom we have some indicator data (such as previous election turnout, demographic variables, etc).[6] Such measures are potentially useful to campaigns, in order to determine where to target get-out-the-vote resources (on those who are marginal to turnout rather than those very likely to do so) as well as persuasion resources (on those who are likely to vote rather than those who are not). This information is also useful to social scientists, particularly if they are interested in how politicians might represent the views of *voters* rather than *non-voters*, or other questions that turn on which kinds of people tend to vote.

[6] Indeed, it is difficult to imagine any other way to measure propensity to turnout other than using demographic indicators and past patterns of behaviour.

Here, our training data is binary, but we want a continuous measure of "propensity" to or probability of voting, not a binary classification of individuals we think will vote versus those who will not. So we will fit a logistic regression model predicting turnout in that previous election using the set of variables. Note again that we could use any predictive model here that we like, so long as it can form predicted probabilities. Let us begin by considering the same kinds of potential risks that we have discussed previously.

First, what is the quality of the training data? Unlike the previous example, where the training data was measured without error, individual voter turnout

training data has several possible sources of error. Where it is self-reported, it is at risk of errors of overstatement (Ansolabehere and Hersh, 2012; Achen and Blais, 2015); where it comes from validating against voting records, it is at risk of record linkage errors.

Second, are the relationships in the training data representative of the population of units for which the measure is to be constructed? The relationships with the indicators may change change from election to election. With this kind of predictive measurement application, we have the opportunity for validation only after the subsequent election, but this is nonetheless useful for assessing the approach in a more general sense. Of course in some sense a good or bad performance of such scores in a given election is just a single data point, one would need to repeat the exercise across many elections to assess whether the approach tends to work or not.

Third, are the available indicators adequately predictive to give useful information? If the predicted probabilities of voting only vary from 60-80% in a population with a 70% turnout rate, that is of limited use for most applications. The ideal, of course, would be if one could make very strong probability predictions that approach a binary classification: very low probabilities of voting for some units and very high probabilities for others.

To illustrate this application, I use training data from the 2015 British Election Study (Fieldhouse et al., 2016), a survey conducted with face-to-face surveys after the 2015 UK general election. This study included a vote validation exercise where voter turnout was verified against voter registration records. I then construct measures using the 2017 British Election Study (Fieldhouse et al., 2018), conducted using the same methods after the 2017 UK general election. Thus the measurement itself is the set of respondent-level turnout predictions for individuals in the 2017 study, although one could have constructed similar measures for any individual in advance of that election, given the requisite indicator data. Because this is a predictive measurement problem, I will then be able to validate the measures by comparing the predicted probabilities of voting to the validated vote data collected after the 2017 election.

As in the previous example, I will set aside the question of model and variable selection, and just proceed with a single regression specification. Once again, the appropriate criterion by which to select a regression model in this context is out-of-sample prediction, and therefore selection by cross-validation/AIC or other statistics that penalise in-sample overfitting is appropriate. The regression whose coefficients I report above reveals familiar patterns in voter turnout, many of which are robust across both different elections in the UK and also across developed countries: all else equal, turnout is higher among those who voted in the previous election, among older voters, among those with greater educational qualifications, among those who are (or have been) married, and among those who own their homes.

Figure 8.2 shows the distribution of individual-level vote propensity measures, on a probability scale. There are many individuals whom we expect

Table 8.2: Statistical models

	Model 1
(Intercept)	0.42
	(0.30)
lagged_turnout	1.44***
	(0.12)
bs(age, 4)1	−2.12***
	(0.47)
bs(age, 4)2	0.21
	(0.48)
bs(age, 4)3	−0.07
	(0.68)
bs(age, 4)4	−0.50
	(0.92)
genderFemale	0.07
	(0.11)
qualificationsLevel 1	0.32
	(0.24)
qualificationsLevel 2	0.43*
	(0.19)
qualificationsLevel 3	0.73***
	(0.20)
qualificationsLevel 4	0.44*
	(0.19)
qualificationsLevel 5 and above	0.82***
	(0.19)
qualificationsOther	−0.04
	(0.32)
maritalNever Married/Partnered	−0.50**
	(0.16)
ethnicityAsian	0.00
	(0.21)
ethnicityBlack	−0.60*
	(0.28)
ethnicityMixed/Multiple	−0.18
	(0.47)
ethnicityOther	−0.76
	(0.98)
tenureRenter	−0.71***
	(0.12)
AIC	2118.34
BIC	2224.44
Log Likelihood	−1040.17
Deviance	2073.01
Num. obs.	1967

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

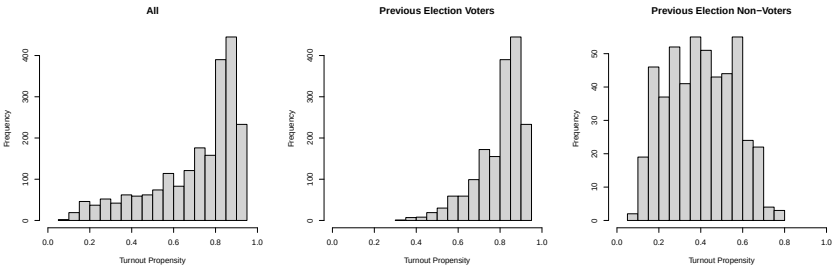


Figure 8.2: Distribution of estimated turnout probabilities for 2017 British Election Study respondents, based on demographic patterns in the 2015 British Election Study, for all voters (left), those who say they voted in the previous election (center), and those who say they did not (right).

to vote with very high probability. Much of this is due to the high predictive power of (self-reported) voting in the previous election, but the plots at center and right show that there is nonetheless a lot of variation associated with the other indicators, holding previous election voting fixed.

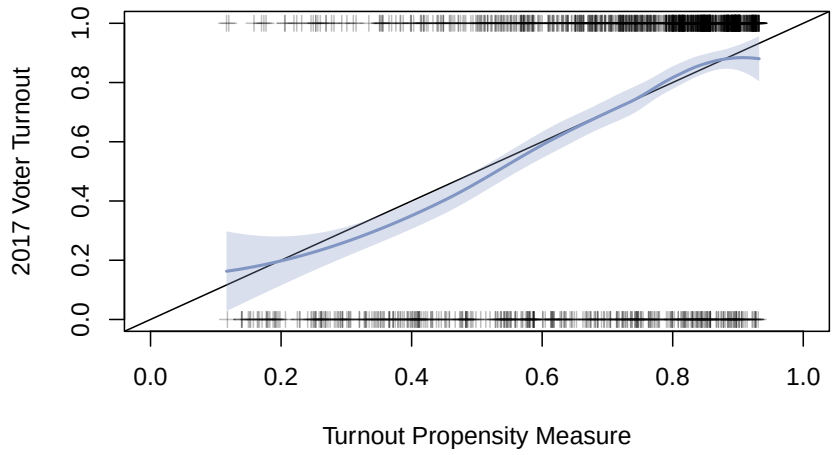


Figure 8.3: Voter turnout in 2017 for British Election Study respondents, given their estimated turnout probability based on demographic patterns in the 2015 British Election Study. Fitted curve shows that the turnout proportion in 2017 closely resembles the demographic predictions based on 2015 voting demographic patterns.

When we turn to validation, we see in Figure 8.3 that the measure of turnout propensity validates well for predicting turnout in the 2017 election. The figure shows that, as the measured probability of turning out increases, the probability of actually turning out increases, tracking very close to the diagonal line corresponding to the realised proportions equaling the predicted probabilities. This is a case where the predictive performance of the measure is very good: demographic turnout patterns in the 2017 UK general election closely tracked those from the previous election (Prosser et al., 2020), and so our measurement strategy of using past patterns of behaviour to train a measurement model describing the relationship between a set of indicators and training data from the past works well.

8.5 Application - Is this a Curry?

In 2019, an American cookbook writer and recipe columnist Alison Roman published a recipe entitled "Spiced Chickpea Stew With Coconut and Turmeric". This recipe became extremely popular and was widely shared. Very quickly, it was critised for failing to acknowledge that it is a "curry". Roman's response was "I've never made a curry, I don't come from a culture that knows about curry.… I come from no culture. I have no culture. I'm like, vaguely European." Our focus here is not on whether this makes any sense as a cultural stance, but rather on the question of what constitutes a curry, and whether we can measure "curry-ness" from a recipe.

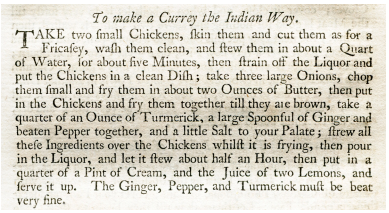
If ever there was a context in which quoting a dictionary definition seemed appropriate, this is it. According to the Oxford English Dictionary, the relevant definition of curry is "A preparation of meat, fish, fruit, or vegetables, cooked with a quantity of bruised spices and turmeric, and used as a relish or flavouring, esp. for dishes composed of or served with rice. Hence, a curry = a dish or stew (of rice, meat, etc.) flavoured with this preparation (or with curry-powder)."[7] Other definitions emphasise spicing as well, eg "a dish of meat, vegetables, etc., cooked in an Indian-style sauce of hot-tasting spices and typically served with rice".

The etymology of the term is believed to be from Tamil, from which arose the Portuguese "caril" and English and French forms of the word with various spellings. There are scattered references in sources, but the first appearance of a "currey" in a English language cookbook is in the first edition of "The art of cookery, made plain and easy" by Hannah Glasse, published in 1748. The original edition of this (best-selling and influential) cookbook included a recipe entitled "To make a Currey the Indian Way", which included the spices of coriander and black pepper. The 1758 edition revised the recipe, with the spices now including turmeric, ginger and black pepper (but no coriander). "Somewhere along the line, the word 'curry' came to be applied to a vast number of South and Southeast Asian foods… The unifying factor in all international curries is the presence of a sauce flavored by a spice mixture."[8]

If it is the spices that make something a curry, we might reasonably ask *which* spices are indicative of something being called a curry rather than a "stew" or something else entirely. How might we measure the concept of "curry-ness" based on use of spices?

Since the category of "curry" is an English language invention associated with the British Empire, we will attempt to develop a measurement strategy for whether something is a curry according to the de facto arbiter of correct British English usage: the British Broadcasting Corporation (BBC). We will use a collection of 9384 recipes scraped from the BBC recipe archive by the author in 2016 when the BBC threatened to take down the archive for budgetary reasons. I have cross-referenced the ingredient lists with a slightly modified list of herbs and spices from the Encyclopaedia Britannica in order to create a 9384

[7] "curry, n.2". OED Online. March 2020. Oxford University Press. (accessed May 26, 2020).



To make a Currey the Indian Way.

TAKE two fmall Chickens, fkin them and cut them as for a Fricafey, wafh them clean, and ftew them in about a Quart of Water, for about five Minutes, then ftrain off the Liquor and put the Chickens in a clean Diih; take three large Onions, chop them fmall and fry them in about two Ounces of Butter, then put in the Chickens and fry them together till they are brown, take a quarter of an Ounce of Turmerick, a large Spoonful of Ginger and beaten Pepper together, and a little Salt to your Palate; ftrew all thefe Ingredients over the Chickens whilft it is frying, then pour in the Liquor, and let it ftew about half an Hour, then put in a quarter of a Pint of Cream, and the Juice of two Lemons, and ferve it up. The Ginger, Pepper, and Turmerick muft be beat very fine.

Figure 8.4: Curry recipe in the 1758 edition of The art of cookery, made plain and easy by Hannah Glasse

[8] https://www.escoffieronline.com/what-makes-a-curry-a-curry/

by 72 matrix of binary variables, each indicating the presence or absence of a given herb or spice in a given recipe.

Table 8.3: Word stems for the 20 most frequently used herbs and spices in the BBC recipe archive.

	x
black_pepper	3367
garlic	2313
chilli	1218
parsley	995
coriander	844
ginger	815
thyme	785
vanilla	509
cinnamon	482
bay_lea	468
cumin	441
basil	359
mint	330
rosemary	322
chives	321
sesame	301
sage	289
turmeric	274
paprika	268
red_pepper	261

Among the 9384 recipes in the archive, 147 use the word "curry" in the title. Note that while this provides us with data that we can use as training data for what constitutes a curry, it clearly does not capture all recipes in the archive that would be commonly understood as curries. For example, there are 13 recipes that contain the character string "tikka", many of which are variants on the most popular curry in the UK, chicken tikka masala, none of which are named "curry".

Name
Chicken tikka and naan bread
Chicken tikka masala
Chicken tikka masala
Christmas barbecued coconut prawns and chicken tikka
Halibut tikka masala with basmati rice
Low-fat chicken tikka masala

Name
Monkfish tikka masala with roti
Monkfish tikka kebabs with yellow bean salad
Salmon tikka wraps
Sheek kavaab naan with malai tikka naanwich
Skewered chicken tikka with spicy lemon-scented rice
Succulent chicken tikka wraps
Tikka paneer cheese with sweet chilli dip

There are many other curries hiding under other names, so while we can be pretty confident that anything labeled as a "curry" is actually a curry, there will be false negatives associated with curries that have more specific names.[9] While we might use other metadata in the recipe archive to identify these, we are going to limit ourselves to the title text and what we can learn about the use of herbs and spices in those recipes that are explicitly labled as a "curry".

I fit a logistic regression predicting whether each of the 9384 recipes has the string "curry" in the title, using the binary variables for the presence of each of the 72 spices. There are a large number of indicator variables in this example, some of which only appear very rarely in the recipe archive. Because of this, I use a Bayesian logistic regression to regularise the coefficients.[10] The details and motivation for regularization in this kind of application are discussed in the next section, for present purposes simply note that the change of estimation procedure does not change the interpretation of the coefficients or the mathematics of constructing the fitted values / predicted probabilities as measures.

[9] Note that I am not including "curried" here, as that includes many dishes that deploy curry powder as a seasoning, many of which are not typical curries as they lack a liquid sauce, eg "Steak with curried sweet potato chips".

[10] I use the default prior scale of 2.5 on the logistic regression coefficients in the R package "arm".

Table 8.5: Estimated logistic regression coefficients for predicting the presence of the term "curry" in a recipe title, using the presence or absence of various herbs and spices.

	coef	se
(Intercept)	-5.07	0.18
asafoetida	-2.53	1.59
vanilla	-2.27	1.43
saffron	-2.18	1.47
mint	-1.94	0.83
oregano	-1.83	1.45
allspice	-1.77	1.54
sage	-1.72	1.44
rosemary	-1.68	1.44
dill	-1.25	1.50
parsley	-1.25	0.64
tarragon	-1.17	1.52

	coef	se
sesame	-1.09	0.65
thyme	-0.98	0.68
anise	-0.70	1.26
bay_lea	-0.67	0.48
caraway	-0.66	1.71
marjoram	-0.65	1.72
horseradish	-0.64	1.72
cinnamon	-0.44	0.38
paprika	-0.44	0.48
holy_basil	-0.39	1.92
wasabi	-0.29	2.00
chicory	-0.28	2.02
lavender	-0.28	2.02
sorrel	-0.26	2.04
lemon_verbena	-0.24	2.07
cassia	-0.23	1.01
mace	-0.20	0.82
black_pepper	-0.18	0.21
red_pepper	-0.14	0.47
savory	-0.11	2.26
fennel	-0.04	0.48
cayenne_pepper	0.06	0.66
black_cumin	0.09	1.12
nutmeg	0.11	0.56
star_anise	0.26	1.31
chives	0.33	0.59
basil	0.35	0.51
chervil	0.37	0.97
garlic	0.41	0.22
fenugreek	0.43	0.55
poppy_seed	0.50	1.07
ginger	0.55	0.24
cardamom	0.56	0.35
cumin	0.69	0.25
clove	0.78	0.44
chilli	0.96	0.24
turmeric	0.99	0.25
coriander	1.00	0.24
curry_lea	1.22	0.46
brown_mustard	1.37	1.04
black_mustard	1.50	0.57
lemon_grass	1.50	1.36

	coef	se
curry_powder	2.31	0.30

The coefficients indicate which herbs and spices are positively indicative of being labelled a curry (including turmeric, coriander, lemon grass, and of course curry powder) as well as which are negatively indicative (including vanilla, saffron, mint, and rosemary). Some spices, such as black pepper, provide a very weak indication as to whether a dish will be labelled as a curry or not. Note that the coefficients corresponding to rarely appearing spices have very large standard errors.

Before we return to Alison Roman's "stew", we should do some validation of the measurement strategy. First, as noted above, we have reason to expect that there are many curries that have other names, such as "tikka". What are the recipes that get the highest "curry score" which do not use the word curry in the title? Here, for the measure I just use the log-odds scale: $\alpha + \beta_1 x_1 + \ldots$, there is no reason to translate to the probability scale. These scores tend to be negative, as very few of the recipes in the archive are called curries.

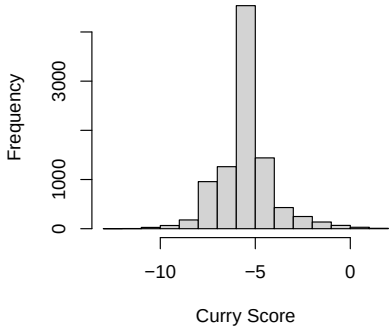


Figure 8.5: Distribution of curry scores for recipes in the BBC recipe archive.

name	score
Sambhar vada: yellow lentil soup with spiced doughnuts	1.29
Lamb madras with bombay potatoes	1.20
Beef Madras	0.98
Sweet potato bhuna masala	0.98
Cabbage with mustard seeds	0.57
Spicy haddock with stir-fried broccoli	0.52
Spicy lamb and paw paw	0.36
Indian hot-water crust pies	0.33
Kharu pork with deep-fried potatoes and salad	0.25
Masala-marinated chicken with minted yoghurt sauce	0.24
Bengal coconut dal	0.18
Coconut and prawn broth with rice, spinach and chilli	0.16
Goan-style lobster with chips	0.13
Bunny chow	0.09
Lamb dhansak	0.09
Keralan crab with Currimbhoy salad	-0.12
Quick-spiced chicken thighs with 'emergency biryani'	-0.15
Coconut and green chilli prawns (shrimp)	-0.16
Crisp tofu with chilli, garlic, spinach and soy mushrooms	-0.17
Spicy jerk chicken thighs with peppers and rice	-0.22
Masala mutton shanks with lemon rice	-0.23
Slow cooker dal	-0.25
Koli ishtew (chicken stew)	-0.27
Chicken and spinach balti	-0.47

name	score
Chicken jalfrezi	-0.47

Many of the most curry-like recipes not explicitly named curries result for the use of names for particular types of curries: "madras", "bhuna masala", "dhansak", "balti", and "jalfrezi". There are some that will be non-obvious to most readers, such as "Bunny chow", which is a "South African fast food dish consisting of a hollowed-out loaf of white bread filled with curry." The Wikipedia article on this says that this dish is believed to have been developed by migrant Indian workers to make use of widely available cheap, white bread.

There are a number of high scoring dishes that reveal the limitations of our ingredient-focused measurement strategy. We have developed a measurement model that only uses ingredients as indicators. Most definitions of "curry" emphasize a stew-like consistency involving both whole pieces of meat, vegetables, etc or a highly spiced "gravy" or "sauce" around those more solid elements. Some of the recipes that our measurement strategy identifies are soups or dal, which have a more uniform consistency. They are all "sauce". At the other extreme, there are also a number of dishes that use a wide variety of spices, but are arguably not curries because they lack a "sauce" or "gravy" component. For example, "Cabbage with mustard seeds"[11] includes black mustard seed, fenugreek, curry leaves, red chili, ginger, turmeric and chilli powder, but does not have a liquid sauce. If you view the consistency of the dish as an important component of the concept of a "curry", then our measurement strategy fails to capture that, and this could be a problematic source of measurement error for some applications.

[11] Subtitle: "If you're looking for a healthy side dish for a curry night, try adding spice to the humble cabbage."

Fortunately, this is not a problem for assessing whether Alison Roman's "Spiced Chickpea Stew With Coconut and Turmeric" is a curry or not, as it definitely has the appropriate consistency. Of the herb and spice indicators in the measurement model we just developed, the recipe includes garlic, ginger, red pepper, black pepper, turmeric, and mint. The coefficients for these in our measurement model are:

	coef	se
mint	-1.94	0.83
black_pepper	-0.18	0.21
red_pepper	-0.14	0.47
garlic	0.41	0.22
ginger	0.55	0.24
turmeric	0.99	0.25

Turmeric is a classic (sometimes definitional) curry spice, but none of the other spices in the recipe are strongly indicative of a curry, while the presence

of mint is strongly counter-indicative of a curry. When we aggregate these coefficients, we get a score of -5.39, which is actually slightly lower than the mean -5.35 and median -5.07 scores for the entire recipe archive. Only one of the 147 recipes with "curry" in the name scores lower than this on our measure, and there is a strong argument that the recipe in question, "Smoked haddock fish cake with curry mayonnaise and watercress" is a false positive in our training data rather than a real curry.

Much appears to turn on the inclusion of mint, which is included in Roman's recipe as a garnish, and which greatly diminishes the curry score, as mint appears in 3.6% of the non-curry recipes in the BBC archive, but only 0.7% of those called curries. Without the mint garnish, the recipe scores -3.45, which is higher than 93% of the recipes in the archive, and than 29% of the recipes actually named curries.

While mint is almost entirely absent from any of the recipes named curries in the BBC recipe archive, and cannot be argued to be a common curry spice, it is possible to find dishes that call themselves curries which use mint if one looks for them. According to a website site I found using Google, "In Indian cooking [mint] is widely used in chutneys, relishes, salads, sauces and teas". The inclusion of "sauces" on this list suggests that the use of mint in dishes that would be called "curries" in English is not unheard of, but the use of mint in Indian cuisine is focused elsewhere. Mint is more prevalent in other cuisines that are well represented in the BBC archive, which is part of why it ends up with such a large, negative coefficient in the model and is therefore treated as so strongly counter-indicative of curry in our measurement strategy.

In sum, our measurement strategy suggests that Alison Roman's recipe is perhaps not a curry, because it includes mint, which is atypical of curries, and is otherwise only spiced with typical curry spices to a moderate degree. You may feel that this exercise was a bit silly, and you would of course be right. Nonetheless, it illustrates the wide ranging applicability of the methodology that was the focus of this chapter. There was a public controversy about what is clearly a socially constructed concept, "curry-ness", and we were able to use a data set to quantitatively describe the patterns of herb and spice usage characteristic of something being called a "curry" in the English language.

This was not meant to be a perfect measurement strategy, it has some obvious limitations. As noted earlier, while spices are an important element of the definition of a curry, there are other elements like the presence of a gravy/sauce that are not captured by this measurement strategy. This exercise also highlights the importance of carefully considering the training data set (the BBC recipe archive is not the only one that could be used), the "gold-standard measure" of the concept (the use of the word "curry" in the title clearly leads to both false positives and false negatives with respect to a typical understanding of the concept) as well as the set of indicators (one could include ingredients other than herbs and spices, or other features of the preparation). Would coconut milk be indicative of a curry had it been included in the set of indicators?

There is room for improving this measurement strategy along all of these dimensions, if one were sufficiently motivated to do so.

8.6 Conclusions

The basic idea discussed in this chapter is that sometimes we can use an existing measure of a concept (known for some units) to learn how to use a set of indicators to generate new measures of that concept (which can be constructed for additional units). This idea can be implemented with basic regression methods, but it also can benefit from more advanced methods that are useful to be aware of for certain applications.

In the last example in the chapter, I used a Bayesian logistic regression to avoid overfitting the data in a case where there were a large number of indicators, many of which appeared only rarely. This is an example of *regularization*, a more general idea that is widely applied in machine learning and predictive applications. Regularization makes sense in applications where there are a large number of features, and one is concerned with *out-of-sample* predictive performance. There are a large number of relevant methods that implement this concept in various ways, including *lasso regression*, *ridge regression*, *elastic net regression*, *least angle regression* as well as Bayesian estimation analogues to these. The details of these are beyond the scope of this text, but the fact that they are useful in the set of cases where there is a relatively large feature set is not. For more information about these, see the very good machine learning textbooks by James et al. (2013) and Hastie et al. (2009), the former is more accessible and the latter more comprehensive.

Another set of tools that are useful for validation of these measurement models are *cross-validation* methods, which are covered in most machine learning texts. Cross-validation simulates out-of-sample predictive performance by fitting models on subsets of the data and evaluating fit using the withheld observations. This facilitates assessment of which models will predict best out-of-sample, as opposed to within-sample, where more flexible/complicated models always fit better. When considering (as we mostly failed to do in the examples here) whether it makes sense to complicate a regression model for purposes of better describing the relationship between indicators and the training data, it is important to protect against overfitting.

9
Supervised Scale Measurement using Linear Indices

In the last chapter, we considered measurement problems where we had a "training" dataset with a pre-existing measure of the target concept (or at least something close to that). Our goal was to use a set of other variables, *indicators*, to predict this quantity so that we could then do out-of-sample prediction (measurement) for units where we did not have that pre-existing measure. While a wide variety of regression models could be used to estimate the relationship between these indicators and the target concept, the most basic version of this used a linear regression to generate a measure that was a linear function of the indicators.

In this chapter, we consider problems where we again want to use a set of indicators to measure a target concept, but where we lack the pre-existing measures for some units (the training data) that would enable us to estimate how the indicators relate to the target concept. Thus, if we want to use those indicators to form a measure, we need some other way of justifying a choice of structure (additive or otherwise) and any needed coefficients. This is the problem of "index construction".

There are a very large number of indices[1] that purport to measure different social science concepts. The first year that I ran a course on the topic of measurement, one of the assignments was for students to identify a measure and critique it. Here are 15 of the country-level measures that they found, most of which are linear indices of the type discussed in this chapter.

- Fragile States Index (Fund for Peace)
- Global Liveability Index (The Economist Intelligence Unit)
- Human Capital Index (World Bank)
- World Press Freedom Index (Reporters without Borders)
- Global Gender Gap Index (World Economic Forum)
- Euro Health Consumer Index (Health Consumer Powerhouse Ltd)
- Democracy Index (The Economist Intelligence Unit)
- Freedom House Index (Freedom House)
- Polity Scores (Polity Project)
- Global Terrorism Index (Institute for Economics and Peace)

[1] "Indexes and indices are both accepted and widely used plurals of the noun index. Both appear throughout the English-speaking world, but indices prevails in varieties of English from outside North America, while indexes is more common in American and Canadian English. Meanwhile, indices is generally preferred in mathematical, financial, and technical contexts, while indexes is relatively common in general usage." https://grammarist.com/usage/indexes-indices/

- Global Peace Index (Institute for Economics and Peace)
- Global Health Security Index (Nuclear Threat Initiative, Johns Hopkins Center for Health Security & the Economist Intelligence Unit)
- Corruption Perception Index (Transparency International)
- Index of Economic Freedom (Heritage Foundation)

Where do these indices come from? How can/should the choices about how to structure them be justified? Many such indices are additive with equal coefficients/weights on different indicators, simply because their authors had no good justification for non-additivity or for any particular choice of unequal coefficients/weights. You might find either of these assumptions—additivity and equal weighting—troubling in general or in particular applications. This chapter aims to give you the tools to appropriately assess these kinds of assumptions and do better, where possible. In these cases, we have to figure out how to put the indicators together into a measure ourselves, via relevant "expertise". Thus this chapter is in large part about techniques for eliciting expertise in usable forms, whether from the analyst or from other experts.

Many of these techniques are informal "face validity" checks regarding the implications of the mathematical construction of the index, and what it implies about the equivalence of different ways of achieving the same index values through combinations of indicator values. We also discuss how it is possible to use such validation checks to create training data sets which can be used to estimate the appropriate structure of the index. This is a way of quantifying the relevant expertise held by subject matter experts, and extends ideas discussed previously in Chapters 7 and 8.

9.1 Example: Olympic Medal Tables

Every four years, when the summer Olympics are held, newspapers and media organisations around the world present medal tables that report the total gold, silver and bronze medals won by athletes of each country. Usually these are sorted to list the most successful countries at the top, but different media organisations choose different ways to sort the table. In the US it is conventional to sort by total medals (gold plus silver plus bronze), in the UK it is conventional to sort by gold medals, breaking ties with the silver medal count, and then the bronze medal count if necessary. In 2020, this meant that the US sat atop the medal table in US media throughout the Olympics, whereas in the UK, China was listed top until the final day when the US (39 gold, 41 silver, 33 bronze) edged past China (38 gold, 32 silver, 18 bronze) at the top of the table by gold medals.

These metrics on which the medal tables are being sorted are themselves examples of linear indices, the subject of this chapter. For country i, the "medal table index" m_i is defined

$$m_i = \beta_g \cdot g_i + \beta_s \cdot s_i + \beta_b \cdot b_i$$

where g_i is the number of gold medals won by country i, and s_i and b_i are the number of silver and bronze, respectively. Sorting by total medals means sorting by a measure m_i where $\beta_g = \beta_s = \beta_b$. Sorting by gold (and then using the other medals only to break ties) means sorting by a measure m_i where $\beta_g \gg \beta_s \gg \beta_b$.

The decision about how to sort the table is implicitly an act of measurement. What are we trying to measure when we sort an Olympic medal table? The concept being measured is something like "aggregate success in the Olympic competitions". Note that there are many different concepts we could choose to measure here, and by looking at "aggregate" success in terms of medal counts we are already foreclosing some quantities that we could have chosen to measure instead.[2] But even given these constraints on the problem, there are various other proposals that have been made for how to aggregate the medal counts g_i, s_i, and b_i. One of these is $\beta_g = 5$, $\beta_s = 3$, $\beta_b = 1$. The argument for this arrangement is that it respects the key axiom that gold is better than silver is better than bronze, but also that it is a bit better to win a single gold than to win a silver plus a bronze. That said, one could clearly disagree with these weights: does it make sense that two silvers are better than one gold while three bronzes are only as good as a silver? Arguing about these kinds of questions is one of the many minor traditions associated with the Olympics.

This might seem a silly example, but the UK has spent a lot of money in recent years supporting Olympic athletes. "Team GB's 67 medals won… in Brazil [2016] cost an average of £4,096,500 each in lottery and exchequer funding over the past four years." This works out to £1.09 per year per Briton. Over the period from 1996 to 2020, this money has been allocated explicitly to maximise the medal count, with money disproportionately spent on athletes preparing for events where medals are more likely to be won, and where the required investments are smaller. This strategy, including increased funding levels, has been highly successful in increasing the UK's medal count. UK Sport funding rose from £5m per year in the run up to the 1996 Olympics (1 gold, 8 silver, 6 bronze) to more than £65m per year in the run up to the 2016 Olympics (27 gold, 23 silver, 17 bronze).[3]

Recent modifications to the funding regime in 2019 have not changed this approach fundamentally, but rather have aimed to take a longer time horizon, increasing funding also for sports/athletes where success could come in the Olympics after the most immediate one. "The UK Sport chair, Dame Katherine Grainger, insisted the organisation's main focus would still be to win as many medals as possible. 'People still believe in our principal objective of success in Olympic and Paralympic Games,' the London 2012 rowing gold medallist said. 'What we heard quite loud and clear from the public is they have not had enough yet. They want more. Our aim is to pursue more medals by more medallists in more sports.' " It may be that the strategy to maximise total medals is not very different than the strategy to maximise gold medals: which athlete/team wins gold among the set of strongest competitors is difficult to

[2] There is no adjustment for the population size of the countries and all medals are equally important, whether they come in a large team sport or an individual sport, whether they come in a popular event or one of the most obscure. In 2020, San Marino easily led the Medals Per Capita table, with 3 medals (two in trap shooting, one in wrestling) and a population of 33,931, far ahead of any other country (GB averaged about one medal per million population, and the US one medal per three million population).

[3] Note that spending per medal has risen substantially, so the targeting of events where medals are more likely to be won is only part of the story. There are decreasing marginal returns in medals to increasing expenditure: there are only so many strong athletes in a given country and only so many medals to be won in the Olympics.

predict even on the day of the event. However the choice to maximise medals, as opposed to other criteria, is potentially very consequential if you are going to aggressively spend money to achieve that goal.

9.2 Defining a Linear Index

An index is a composite statistic that is formed by aggregating a set of indicators into an interval-level (but not necessarily continuous) measure.[4] Exactly how the aggregation ought to be done requires justification. In practice many indices are linear, additive functions of the following form, where the value of the index m_i for unit i as a function of various indicators j is given by:

$$m_i = \sum_j b_j \cdot I_{ij} \tag{9.1}$$

[4] Only in rare cases is it possible for an index constructed in this way be considered ratio-level.

This sort of linear index is sometimes called a "sum score": the score for unit i is the sum of scores b_j on a set of items I_{ij}.

Obviously Equation 9.2 looks like the linear regression equations we considered in the last chapter, but we are using b instead of β because we lack any training data m_i to fit a model to learn the coefficients.[5] Thus, the task of this chapter is to think about how we specify the values of b, that will predict the target concept μ as well as possible, without being able to fit a regression to estimate the coefficients that best predict known values of m because there are none.

[5] We cannot even pretend that there is a generative model here.

This type of index is sometimes described in terms of "weights" w_j rather than "coefficients" b_j, but this tends to come to the same thing mathematically. Describing these as "weights" makes sense in cases where the indicators are standardised in some way (either to have mean zero and standard deviation 1 or to range from 0 to 1) and the weights w_j are all positive ($w_j \geq 0 \forall j$) and add to 1 ($\sum_j w_j = 1$). This is a special case of Equation 9.2, where the b_j are the weights, and satisfy these constraints.

Very commonly, this form is nested, such that I_{ij} is itself a *sub-index*, constructed linearly from several further indicators. If the index is linear and the sub-indices are also linear, this is mathematically equivalent to simply defining the index in terms of the underlying indicators in the sub-indices. Working with sub-indices can nonetheless be useful in facilitating the process of conceptualization, as seen in some of the examples later in this chapter.

Indices are also sometimes described in terms of "points", particularly in cases where all the indicators are categorical. In such cases, different levels of the indicators are associated with different positive or negative point totals, and are added to yield a "score". This too is a special case of Equation 9.2, with the point totals being the b_j and the indicator variables being full sets of dummy variables for each categorical variable.

9.2.1 Requirements

A fundamental feature of this type of additive index is that it purports to be an interval-level measure. In order for an index to *actually* be an interval-level measure, equal differences in the index value must be equivalent with respect to the underlying concept. This means that the scale construction must correctly space the levels for categorical indicators and apply appropriate transformations of continuous variables. Indicators must then have appropriate coefficients such that like changes in the measure correspond to like changes in the target concept. The selection and transformation of indicators as well as how they are weighted by the b coefficients must be done carefully in order to achieve a measure that is credibly interval level.

The linear structure of the index means that changes in different indicators can be exchanged with one another. Increasing I_1 by Δ will change the index by $b_1\Delta_1$. But you can also increase the index by the same amount by increasing I_2 by $\Delta_2 = \frac{b_1\Delta_1}{b_2}$. This is most obvious if $b_1 = b_2$: if the indicators have equal coefficients/weight, you can achieve the same change to the index by changing either indicator by the same amount $\Delta_1 = \Delta_2$.

This implies that one ought to be comfortable with all such equivalences. This connects back to the discussion of unit analysis and linear regression discussed in Chapter 6.2.2. The b coefficients in the index are, by the very nature of a linear index, statements that one unit of indicator j translates to b_j units of the target concept. Checking indicator equivalence means being comfortable with these statements, when they are made explicitly. Do you think that a unit with $I_1 = I_1^* + \Delta_1$ and $I_2 = I_2^*$ has the same value of the target concept as a unit with $I_1 = I_1^*$ and $I_2 = I_2^* + \frac{b_1\Delta_1}{b_2}$? Linear scale development should involve an explicit validation strategy that involves checking whether these tradeoffs are plausible, across a range of indicator pairs and representative indicator differences.

If an index is successfully measuring the target concept, the $\hat{m}$ values for any pair of units i and i' should appear correct, in relative terms. This means that we should be able to look at units that have similar scale values, deriving from different combinations of indicator values, and feel comfortable with the claim that they ought to be viewed as similar according to the underlying concept. Similarly, we ought to be able to look at units that have different scale values, and feel comfortable with the claim that they are different in the direction that their scale values are different, even if we cannot easily assess the magnitude of that difference. Again, as with the coefficient comparisons, linear scale development should involve an explicit validation strategy that involves checking whether these comparisons are plausible, across a range of similar and dissimilar unit pairs.

These requirements create a number of opportunities to refine and validate an index, which we will discuss later. These requirements can be stated either in terms of whether the relative values of the b_j coefficients/weights on dif-

ferent indicators are correct as well as in terms of whether different units are correctly ordered.

9.2.2 Selecting Indicators

How do we choose which indicators to put into our linear index? Here it is useful to think by analogy to the regression-based analysis that we discussed in the preceding chapter. Ideally we would want to include any indicators that *would* have non-zero coefficients, *were* we able to do that sort of training exercise on some existing measure of the concept we want to measure. Non-zero coefficients imply not only that an indicator is associated with the concept of interest in the relevant population of units, but also that it is *conditionally* associated with that concept given the other included indicators. Sometimes there are multiple available indicators that are all very highly correlated,[6] which means they carry largely the same information and the consequences of choosing one versus another may be very slight.[7] Excluding an indicator that is in fact associated with the concept is thus a more serious problem to the extent that there are not other indicators which the indicator is correlated with, as such indicators can carry information about that component of the target concept in its absence.

[6] It is also the case that, even if you did have training data, you would struggle to estimate which of these highly *multicollinear* variables was most predictive of the concept via a regression.

[7] It also may not be slight, if the differences between the indicators are closely associated with your intended application.

Having established the principle of trying to identify indicators that are conditionally associated with the target concept, there are many strategies one could follow for actually picking these. One strategy, suggested by Munck and Verkuilen (2002), involves iterative decomposition of the target concept. There are three steps:

1. Conceptualization - identify the attributes that constitute the concept and how they are interrelated
2. Measurement - find measures of the constituent attributes
3. Aggregation - combine the measures of the attributes in a way implied by the conceptualization

So, for example, if you are trying the measure which countries are more or less "globalised", your initial step might be to break down the concept into sub-components of "economic globalisation", "social globalisation", and "political globalisation" (Gygli et al., 2018). Crucially, this logic can be applied recursively. At step 2 you may discover, as in this example of measuring globalisation, that there is still not an available measure for an attribute (eg economic globalisation) that constitutes one part of the target concept. You must then apply the entire process to constructing that measure of the attribute, in order to be able to include it in your overall measure of the target concept. We will see some examples of this below.

One of the useful distinctions made by Munck and Verkuilen (2002) and others is between measurement strategies that are *maximalist* and measurement strategies that are *minimalist*. A minimalist measurement strategy uses a

relatively parsimonious definition of a concept and relatively few indicators. A maximalist measurement strategy uses a more inclusive definition of a concept and relatively many indicators.

One way to think about this tradeoff is by thinking about the kinds of measurement errors that you are likely to introduce with either type of definition. Minimalist approaches particularly risk missing aspects of the concept that you want to measure, maximalist approaches particularly risk introducing unwanted elements of other concepts. If we recall the graphical representation of the measurement process from Chapter 2, we are considering a comparison of the minimalist approach on the left (one indicator) with the more maximalist approach on the right (three indicators). What is the tradeoff that we face in choosing between these?

Let's consider the case of a single indicator measure first. If you use a single indicator to measure a concept, the risk of introducing bias comes from the other factors (O) that shape the values of the indicator (I_1) besides the concept of interest (C). With multiple indicators, the risk comes from all the other factors (O) that shape the values of *any* of the indicators (I_1, I_2, I_3, etc) besides the concept of interest. Why, then, would you ever want to use additional indicators?

If you start with a single indicator, you have a limited number of potential sources of bias, but those biases are potentially very large in magnitude if some of the other factors that shape that indicator besides your target concept are strongly associated with the outcome/treatment variable you are associating your measure with. As you add more indicators, there are *more* different potential sources of bias because there are additional other factors (O) shaping the new indicators, but because there are more of them, the magnitude of the biases from each one are smaller, so long as the "other factors" that influence your various indicators are not correlated with one another.

With a sufficiently large number of indicators, it becomes difficult or impossible to reason about likely sources of correlated measurement error. A virtue of minimalist strategies is therefore that they enable you to think critically about likely sources of correlated measurement error. A virtue of maximalist strategies is that they may mitigate the magnitude of likely errors because each source of measurement error has lower weight in the measure and these sources of error cancel out in expectation, to the extent that they are not positively correlated with one another.

Sometimes adding additional indicators makes little practical difference, because the indicators are highly correlated with one another. Ogwang (1994) showed that the life expectancy indicator used in the calculation of the Human Development Index (HDI) predicted 88% of the variation in that index. Ogwang argues that, particularly in a context where a single indicator predicts most of the relevant variation, there are advantages to avoiding indices entirely, as it obviates the need to make transformation and weighting decisions like the ones that are discussed in the sections below. This is the extreme form of the

minimalism argument (which does not mean it is wrong).

9.2.3 Transforming Indicators

Once you have selected indicators, you need to consider whether they should be transformed from I to I^* in some way in order to satisfy the requirements of an interval-level measure. For continuous indicators, this may require transformation. The most commonly used transformations are log transformations,

$$I^* = \log(I)$$

standardizations,

$$I^* = \frac{I - mean(I)}{sd(I)}$$

and linear rescalings to a specified range $[min, max]$.

$$I^* = \frac{I - min}{max - min}$$

For example, the Human Development Index includes per capita gross national income (GNIpc) as one of its three components. However, this is not entered into the index directly, but rather it is both log transformed and rescaled:

$$\text{Income Index} = \frac{\log(\$\text{GNIpc}) - \log(\$100)}{\log(\$75000) - \log(\$100)}$$

This rescaling means that the index is 0 when GNIpc is $100 and 1 when GNIpc is $75000, and increases linearly in between with the *log* of income.

This transformation reflects a substantive assessment that income matters on a log scale rather than a linear scale when one is thinking about the concept of "Human Development". Thus, when GNIpc is $10,000, the Income Index ≈ 0.70, which is much closer to the value that the index achieves at $75,000 than it is to the value it achieves at $100. The log transformation is appropriate if it is the case that one thinks that multiplication of the indicator maps into addition for the concept of interest. That is, the difference in "Human Development" between an income of $2x$ and x is the same as the difference in "Human Development" between an income of $4x$ and $2x$, not $3x$ and $2x$.

For continuous indicators, one has essentially the entire universe of functions to potentially choose from, but in practice the most commonly used transformations are log transformations and linear rescaling of endpoints, both of which appear in the above example. For categorical indicators, transformation simply means assigning a numerical value, ie a number of "points", to each level of the variable.

9.2.4 Specifying Indicator Coefficients

Where transformation of indicators determines how different levels of the same indicator are associated with the target concept, specifying the indicator

coefficients determines the relative (partial) associations of different indicators with respect to that concept. Recall that our linear index looks like this:

$$m_i = \sum_j b_j \cdot I_j \tag{9.2}$$

How do we pick the b coefficients if we cannot estimate them using some training data for m?

There are a few common strategies here, all of which immediately present the challenge that it is difficult to justify *specific* values for the b_j:

1. Equal weighting ($b_1 = b_2 = \ldots$)
2. Analyst-specified weighting
3. Expert-specified weighting

The equal weighting assumption might strike you as particularly troubling. To continue the linear regression analogy, equal weights are like assuming a priori that all the betas in your regression model should be the same. This might strike you as crazy, but people have in fact made the argument for equal coefficients by assumption in regression applications where there is not enough data to estimate regression coefficients reliably (Graefe, 2015). Note that equal weighting pushes a lot of the work back to the previous step of appropriately transforming the indicators. Dimensional analysis dictates that, if you are going to have equal weighting, the indicators must all be on comparable scales, such that adding them can be a meaningful operation. This can be achieved by standardization or by transformation to something resembling the quantile/percentile in the observed distribution (that is, only either a 0-1 or 0-100 scale). Equal weighting does not require you to make any further decisions beyond whether to include indicators in the scale, and transforming them in a way that makes them numerically comparable.

In some instances, it makes sense for the analyst (the person constructing the measure) to specify non-equal weights. These need some justification. Unfortunately these arguments are very difficult to make convincingly, even for simple adjustments to weights (such as counting one indicator at twice the weight of another). Because these are judgement calls, it often makes sense to bring in a broader range of experts to make these evaluations wherever possible. This strategy has been used in some large-scale measurement projects like the Global Health Security Index as well as in one of the applications discussed later in this chapter.

As mentioned earlier in stating the requirements of a linear index, validation comes down to confirming that a unit with $I_j = I_j^* + \Delta_j$ and $I_{j'} = I_{j'}^*$ has the same value of the target concept as a unit with $I_j = I_j^*$ and $I_{j'} = I_{j'}^* + \frac{b_j \Delta_j}{b_{j'}}$, for all pairs of indicators j and j'. Thus, one strategy for developing a set of unequal coefficients is to start with equal coefficients and then interrogate the analyst's and/or experts' intuitions about whether the implied tradeoffs are appropriate. Where they are not, coefficients can be increased or decreased, and the process iterated until there are no more obvious adaptations to be made.

9.2.5 Estimation

One way of validating an existing index (for example one using equal weights) is to examine whether the relative values of the measure make sense for specific units. This sort of "face validation" process simply asks whether the comparison "looks" reasonable: should unit i really be higher/lower than unit i' in this concept? However, if we are comfortable using expert evaluations of this type as the basis for validation, we can directly generate coefficients/weights from this sort of expert evaluation of pairwise comparisons (Floridi and Lauderdale, 2018). In Chapter 7, we considered supervised learning methods for calibrating a measurement scale when gold standard data m_i measuring your target concept μ_i was available for some units. Here, we consider a strategy for calibrating a set of indicators in the absence of gold standard data m, which builds on the idea of competition data from Chapter 7.

We assume that we are able to generate data that tells us something about whether $\mu_i > \mu_{i'}$ for specific pairs of units. If we have data like this, we could feed it through a Bradley-Terry model to learn about units i and i' specifically. But this would limit us to learning about the units for which we have these comparisons. Instead, we are going to use the indicators to try to predict the results of these competitions. In the simpler case without the possibility of ties, we start with the Bradley-Terry model:

$$log\left(\frac{p(i \text{ defeats } i')}{p(i' \text{ defeats } i)}\right) = \alpha_i - \alpha_{i'} \tag{9.3}$$

and replace the individual unit "strengths" α_j with linear functions of the indicators:

$$\begin{aligned} log\left(\frac{p(i \text{ defeats } i')}{p(i' \text{ defeats } i)}\right) &= (\alpha + \beta_1 I_{1i} + \beta_2 I_{2i} + \cdots) - (\alpha + \beta_1 I_{1i'} + \beta_2 I_{2i'} + \cdots) && (9.4) \\ &= \beta_1 (I_{1i} - I_{1i'}) + \beta_2 (I_{2i} - I_{2i'}) + \cdots && (9.5) \end{aligned}$$

This is just an intercept-less logistic regression for unit i "defeating" i' with predictors equal to the differences between their respective indicator values.[8] The linear predictor from the model then provides an estimate of appropriate coefficients for the indicators:

$$\hat{\mu}_i = \hat{\beta}_1 I_{1i} + \hat{\beta}_2 I_{2i} + \cdots \tag{9.6}$$

In order for this to work, you need to have a set of pairwise comparisons that you think reflect the underlying concept you want to measure. The "strength" that determines the pairwise winners needs to be the concept you want to measure, not something else. The most obvious way to generate this kind of pairwise comparison data is to survey people who you think have some expertise for determining which indicator values correspond to greater levels of the target concept, and how to trade-off those values against one another. This approach involves conducting a *conjoint experiment* (Green and Rao, 1971;

[8] Note that you might want to include the intercept if you thought that there was some reason that the units that you are calling i are going to win more often than those you are calling i', akin to "home field advantage" in a Bradley-Terry model. This might happen if your data were generated from human responses and you worry that people will choose the first option in the survey more often than the second, for example.

Hainmueller et al., 2014) where the respondents are provided with two profiles of indicator values and asked to select which profile has a greater level of the concept of interest.[9] This relies on the ability of whoever is making the comparisons to make comparisons on the basis of the concept you are interested in.

[9] Alternatively, one could provide single profiles of indicator values and ask respondents to rate the profile according to the concept of interest, but this relies on their ability to use a consistent rating scale across many tasks, as opposed to making consistent comparisons, which is less demanding.

Floridi and Lauderdale (2018) conduct a conjoint experiment in which experts on the demographic concept of *productive aging* are asked whether hypothetical individuals who engage in varying degrees of paid work, volunteering, grandchild care and care for sick/disabled adults are more or less productive than other such hypothetical individuals. The authors then fit an ordered logistic regression (there is an "about the same" intermediate response) for predicting which hypothetical individual is selected in each expert response. The coefficients from the model are shown below, which translate directly according to Equation 9.6 into points on the scale for each level of each activity.

The Italian and Korean aging experts put broadly similar relative weights on different activities, but with some differences with respect to certain activities. Italian experts put greater value of grandparental care, Korean experts put greater weight on volunteering. When the coefficient estimates are then applied to constructing *productive aging* scores for the subjects of a Korean aging survey that includes these indicators, the scores derived from different aging experts' conjoint responses are very highly correlated with one another. They are especially highly correlated when comparisons are made among the Italian coders or among the Korean coders, with somewhat lower correlations comparing scales generated from experts from different countries. This illustrates an important point about this kind of weight calibration: not everyone will have the same weights on different activities, and this may reflect different conceptualisations of the underlying concept. If Korean and Italian aging experts perceive different activities as differently productive, then indicator weights calibrated on their assessments will be different, and so too will the measures that result. In this case the differences are not very large, but in other applications they might be.

For other concepts, it might make sense to use samples of the public to make the comparisons. For example, Hainmueller and Hopkins (2015) describe an experiment in which they randomise characteristics of immigrants, and ask respondents to indicate which immigrant they would prefer to see admitted to the United States. The analysis in that paper, as in most conjoint experiments, is primarily focused on making causal claims about which attributes of hypothetical immigrants are associated with being more or less favoured. But another way to think about what they are doing is that they are measuring the concept of "relative appeal of potential immigrants to existing US citizens". There are potential applications for such a measure. For example, one might imagine that the experience of immigrants once they arrive in the US is related to whether they are generally viewed more or less favourably by US citizens. Thus, if one took the model estimates from the conjoint experiment, and con-

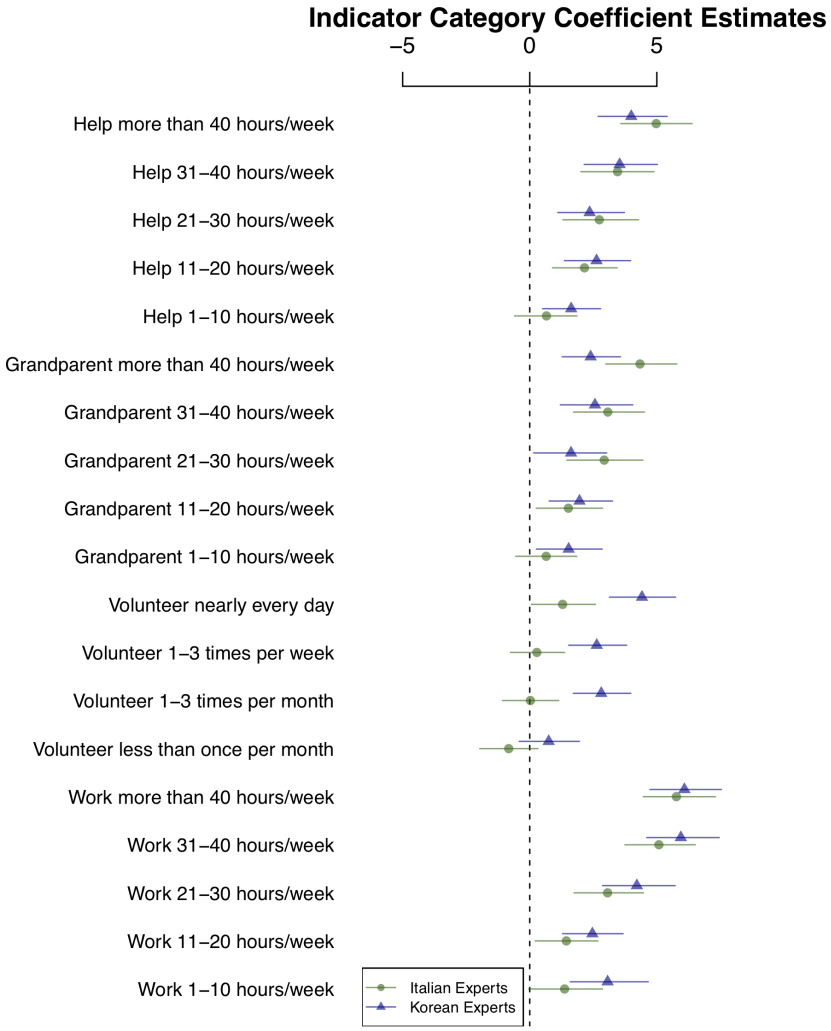


Figure 9.1: Comparison of coefficient estimates for Italian and Korean aging experts. Baseline category for each type of activity is never (for volunteering) or 0 hours per week.

structed fitted values for real individuals, this might constitute a useful measure of aspects of their experience once they arrive in the US.

This is a somewhat different perspective on conjoint experiments than you may encounter in other contexts, where the focus is on enabling valid causal inferences. The causal inference perspective on conjoint experiments is that they use randomisation to enable estimation of a well-defined causal estimand associated with changing individual attributes involved in the choice, averaged over variation in the other attributes and the sample of respondents. The measurement perspective on conjoint experiments that I am employing here is a very literal one: such experiments enable you to measure the propensity to be selected in the pairwise comparisons as a function of the randomized indicators. For measurement purposes, the randomization is useful because it allows us to explore how the measure varies across a multidimensional space of possible indicator values, rather than for the usual reason that it enables attribution of choice variation to particular indicators. Put differently, we are primarily interested in $\hat{m}$ given I, not $\hat{\beta}$.

9.3 Defining a Non-Linear Index

All of the above discussion considered linear indices, of the form:

$$m_i = \sum_j b_j \cdot I_j$$

Of course one can also aggregate the indicators in an index in non-additive ways, of which there are an infinite number to choose from. Additivity is simply mathematically convenient, and therefore a widespread default. If you think additivity is unacceptable as a baseline assumption in the absence of good theory, I have some bad news for you about how linear regression models are used in the social sciences. You should think of additive indices in a similar way to how (I hope) you think of linear regression models. They are useful "first-order" approximations to the relationship between the indicators and the target concept. If you lack any/enough theory, a linear approximation is a sensible place to start. It may or may not be a good place to stop.

There are a few common non-additive ways of aggregating a set of indicators. These are often motivated by the kinds of axiomatic arguments that we considered in Chapter 6. Such arguments say that the target concept *ought* to respond to these changes in the indicators in a way that is incompatible with additivity.

One definition of poverty used in the UK in the early 20th century required that, for people to be non-poor, they must have both a bath **and** a garden (Laderchi et al., 2003, p246). Assume for the moment that these were sensible indicators at the time. What are the implications of requiring both? Requiring both means that the poverty measure is insensitive to the difference between having neither a bath nor a garden and having just one of those. The implications of these choices can be important not just for assessment of individuals,

but also of populations. What if most people in more densely populated areas have baths but not gardens and most people in non-urban areas have gardens but not baths? Almost everyone has some of the elements of non-poverty by this definition, but almost everyone is still defined as poor. The underlying methodological question is how you aggregate the indicators. Requiring both a bath and a garden is equivalent to defining non-poverty (np) as the product of a variable for having a bath ($b = 1$ if yes, 0 if no) and having a garden ($g = 1$ is yes, 0 if no): $np = b \cdot g$. Additive aggregation would lead one to define a household poverty measure as the sum of these two indicators: $np = b + g$. This would mean that having a garden or a bath but not both would make you "halfway" between being poor and not poor. Using the product rules out this possibility.

The most widely used non-additive aggregation functions are multiplicative (products) and variants thereof like the the geometric mean. If an additive index is

$$m_i^{\Sigma} = \sum_j b_j \cdot I_j$$

, then a multiplicative index typically is constructed as

$$m_i^{\Pi} = \prod_j I_j^{b_j}$$

, with the potential for weights/coefficients to enter exponentially. This kind of expression is easiest to understand if the I_j are constrained to the range from 0 to 1, describing a proportion or share of a maximum possible, and the b_j are all 1, but this is not necessary. Multiplicative aggregation is simply additive aggregation on a log scale:

$$m_i = \prod_j I_j^{b_j}$$

$$\log(m_i) = \log\left(\prod_j I_j^{b_j}\right)$$

$$\log(m_i) = \sum_j b_j \log(I_j)$$

The important thing to recognise about multiplicative aggregation is that it understands the relationship between the indicators in a fundamentally different way than additive aggregation. Additive aggregation *requires* that the different indicators have the same dimensions/units to begin with or that the b_j themselves provide the conversion to the same units, because otherwise the addition violates dimensional analysis (recall Chapter 6). In most cases, multiplicative aggregation will mean that the quantity that you are trying to measure has dimensions that are a *product* of those of your indicators.

Put differently, additive aggregation says that the indicators are *substitutes*. If you get more of one, it can replace the lack of another. Multiplicative aggregation says that the indicators are *complements*. If you get more of one, it

makes the amount of the others that you already have contribute more. This is a major conceptual difference, and it is very important to think about which aggregation model is more appropriate to a given application.

9.4 Application - UN Human Development Index

As an initial example, we begin with a relatively simple index, albeit one that uses non-additive aggregation. We will consider the empirical consequences of not using additive aggregation, by comparing the index to the one we would recover using additive aggregation with the same indicators. We will also consider the theoretical case, the *conceptual case* for the aggregation method that is used.

The UN Human Development Index was first published in 1990 with the aim of measuring human development at the country-level. The concept of development is defined as "a process of enlarging people's choices". HDI incorporates three sub-indices, a Life Expectancy Index I_L, an Education Index I_E, and an Income Index I_I. These sub-indices are defined in very minimalist ways, with just one or two indicators (four overall). Each sub-index is defined such that the minimum score attainable is 0 and the maximum is 1. Where LE is life expectancy at birth, MYS is mean years of schooling, EYS is expected years of schooling,[10] and $GNIpc$ is per capita gross national income in purchasing power parity US dollars:

$$I_L = \frac{LE - 20}{85 - 20}$$

$$I_E = \frac{1}{2} \cdot \frac{MYS}{15} + \frac{1}{2} \cdot \frac{EYS}{18}$$

$$I_I = \frac{\log(GNIpc) - \log(100)}{\log(75000) - \log(100)}$$

These rescalings are designed to place all countries in the interval $[0, 1]$ and in the few cases where the rescaled values exceed that range, they are censored to the relevant extreme (this only occurs for the top end of the education and income indices).

The formula has changed since its creation, but as of this writing HDI aggregates these three sub-indices by geometric mean rather than additively:

$$HDI = \sqrt[3]{I_L \cdot I_E \cdot I_I}$$

Note that at its core, this is a multiplicative aggregation.[11] This means that development is understood to arise out of *jointly* possessing life expectancy, education and income. These are complementary resources which do not substitute for one another. In the extreme case that a country completely lacked any one of these, the resulting HDI would be 0 as well, regardless of the values of the other two. This mathematical structure embeds a strong substantive commitment about what meaningfully enlarges people's choices, the conceptualisation of development underlying the measure.[12]

[10] Mean years of schooling is among those aged 25 and older, expected years of schooling is a projection for those currently under 18.

[11] The cube root only makes sense here because each of the sub-indices is defined in such a way as to range from 0 to 1. Otherwise it would not make sense to ask what an *average* value of the three was, regardless of whether that average was arithmatic or geometric.

[12] Arguably this index is applying an argument about what is required for *individual* human flourishing at the *aggregate*, which could mask the consequences of inequality for which countries actually put more individuals in a position of having all three of these resources. Inequality-adjusted HDI attempts to correct for this.

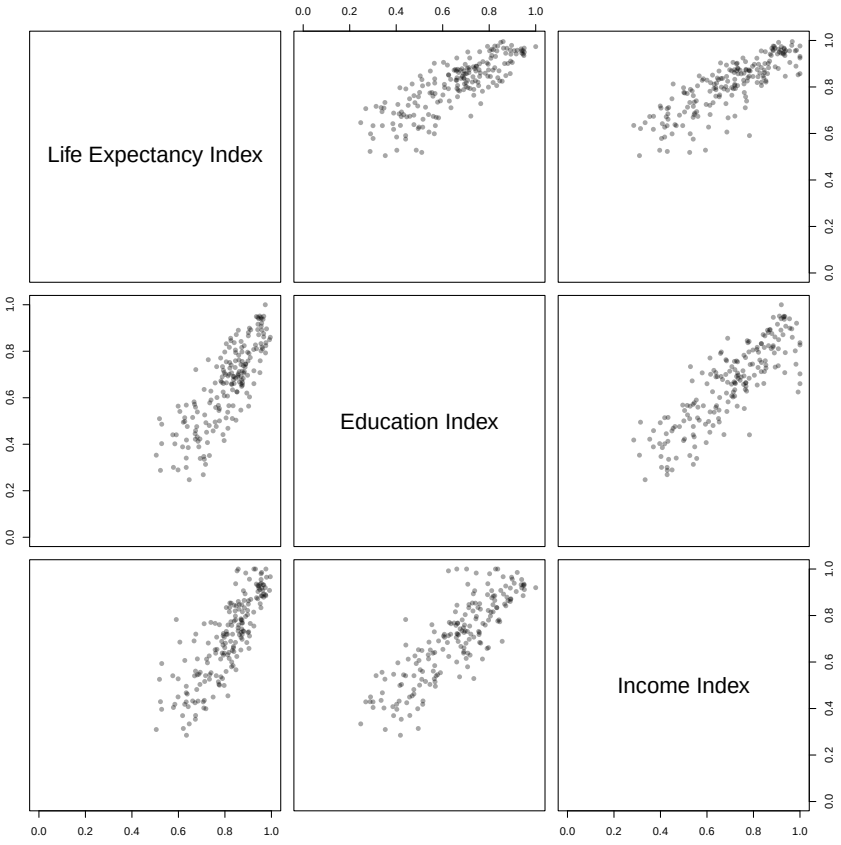


Figure 9.2: Human Development Index sub-indices.

The three sub-indices are in fact all very highly correlated with one another, in 2019 the correlation of the life expectancy index with the education index was 0.82, the correlation of the life expectancy index with the income index was 0.82, and the correlation of the education index with the income index was 0.85. Figure 9.2 shows that despite having similar pairwise correlations, the range of the life expectancy index is substantially less than the other two indices. The lowest life expectancy index value is 0.50, while the other two indices have minima of 0.25 and 0.28.

Notice that this is entirely a result of the way that the sub-indices were defined. All three sub-indices are linear rescalings of some original indicators. The purpose of these rescalings is to make the three indices comparable in scale, such that they can be aggregated meaningfully by geometric mean. Thus, the rescaling define a minimum level and a maximum level; the levels at which countries have entirely failed to achieve any development and the levels at which they have achieved "full" development. The minima have been set, by the analysts who created the index, at a life expectancy of 20 years, 0 years of education and an income of $100 per person. The maxima have been set at a life expectancy of 85 years, 15 mean years and 18 expected years of education, and an income of $75,000. The minimum observed life expectancy of any country in 2019, according to the indicators used in the HDI, was found in the Central African Republic which had a life expectancy of 52.8 years. The minimum education index value was that of Niger, which had 4.3 mean years of schooling and 7.6 expected years of schooling. The minimum income index value was that of Burundi which had a per capita GNI of $659.7 (purchasing power parity). The minimum observed values for education and income are simply much closer to the value that the index specifies as the zero point than is the minimum observed value for life expectancy.

This is not necessarily wrong, but it is a choice by those who created the index. Raising the floor value of 20 in the life expectancy rescaling to a higher value would have the effect of stretching out the variation in the life expectancy index to cover more of the range from 0 to 1. To achieve a similar range to the other two indices, a floor value around 40 would be required. But the decision of whether to do this needs to be made on substantive grounds. It could just be the case that there is less variation across countries in the contribution of life expectancy / health to their state of development, than there is for education and income. That is the appropriate grounds on which to adjudicate the scaling of the different indices.

As noted earlier, the three sub-indices are all fairly strongly correlated with one another because countries tend to have all three of these things together or lack all three. This means that, as suggested by Ogwang (1994), one could simply substitute life expectancy in some applications. But it is important to recognise that the empirical fact that one of these could largely stand in for the whole index does not really undermine the theoretical argument for conceptualising development in terms of all three component sub-indices. Indeed, perhaps the

reason that countries tend to end up with similar values for all three is that it does not make much sense to maximise any one at the expense of the others, which is in some sense the point of the conceptualisation of development as the product of all three.

9.5 Application - Immigrant Integration Index

The second index that we are going to look at in detail is a measure of the extent to which individual immigrants in a society are "integrated": the Immigration Policy Lab (IPL) Integration Index developed by researchers at Stanford and ETH Zurich. Harder et al. (2018) measure this concept via a survey instrument that asks a series of questions related to the concept of integration. The goal here is not so much to measure the average level of integration across a full population (although that is one potential application) but to measure variation in the extent of integration at the individual level so that the relationship of integration to other attributes of individuals can be studied.

9.5.1 Conceptualisation

What is immigrant integration? The authors carefully define what they mean by integration, and how it is different from assimilation:

> "In developing our measure, we defined integration as the degree to which immigrants have the knowledge and capacity to build a successful, fulfilling life in the host society"

> "Our definition distinguishes integration from assimilation, the latter of which requires immigrants to shed their home country's culture in favor of adopting the cultural practices of the host country's dominant group. In our view, immigrants need not shed their own culture to live successful and fulfilling lives in the host country. Therefore, our measure focuses exclusively on capturing the degree to which immigrants have acquired the knowledge and capacity to build successful lives rather than the degree to which they have shed their cultural heritage."

> "For example, to capture linguistic integration we measure only whether immigrants have acquired skills in the host country's or region's dominant language, but we are agnostic as to whether immigrants still use their home country's language. In contrast, a measure of assimilation would by definition take both aspects into account."

The authors identify six component dimensions of integration, which are psychological, economic, political, social, linguistic, and navigational integration. One can immediately see that that is a maximalist definition, it is intended to capture the variety of ways in which people might or might not be integrated into a society. One implication of this is that the scale already includes dimensions and indicators that one could imagine being interested in their relationship to "integration". For example, it includes measures of income and

employment, which means that you would not want to use the entire index to try to describe variation in income or employment, although you could use parts of the index to do so. Reflecting the fact that their measure is very general and includes many things, the authors include substantial analysis of the extent to which these different dimensions of integration are associated with one another.

9.5.2 Indicators

All of the indicators for this integration index are based on survey questions, which are coded or recoded onto 5 point scales. These are the items, with the first two indicators listed for each dimension constituting the short form of the survey instrument. The instrument is given in the form that the authors propose for the United States, but is designed to be cross-nationally valid with appropriate changes in the relevant proper nouns.

- Psychological
    1. *How connected do you feel with the United States?* (5pt scale)
    2. *How often do you feel like an outsider in the United States?* (5pt scale)
    3. *Thinking about your future, where do you want to live?* (5pt scale)
    4. *How often do you feel isolated from American society?* (5pt scale)
- Economic
    1. Household income (recoded to 5pt scale)
    2. Employment (recoded to 5pt scale)
    3. Ability to handle unexpected expenses (recoded to 5pt scale)
    4. *How satisfied are you with your current employment situation?* (5pt scale)
- Political
    1. *How well do you understand the important political issues facing the United States?* (5pt scale)
    2. *In the last 12 months, how often did you typically discuss major political issues facing the United States with others?* (5pt scale)
    3. 1 + number of correct answers to four political knowledge questions (5pt scale)
    4. *There are different ways of trying to improve things in the United States or help prevent things from going wrong. During the last 12 months, have you done any of the following?* (recoded to 5pt scale)
- Social
    1. *In the last 12 months, how often did you eat dinner with Americans who are not part of your family?* (5pt scale)
    2. *Please think about the Americans in your address book or your phone contacts. With how many of them did you have a conversation - either by phone, messenger chat, or text exchange - in the last 4 weeks?* (5pt scale)

3. *People sometimes participate in different kinds of groups or associations. For each group listed below, how often do you participate in a group activity?* (5pt scale)
4. *Many people help each other with everyday favors, such as getting rides, borrowing a little money, or babysitting. In the last 12 months, how often have you provided such favors to Americans?* (5pt scale)

- Linguistic

1. *I can read and understand the main points in simple newspaper articles on familiar subjects.* (5pt scale)
2. *In a conversation, I can speak about familiar topics and express personal opinions.* (5pt scale)
3. *I can write letters about my experiences, feelings, and about events.* (5pt scale)
4. *I can listen to and understand the main points in radio or TV programs about familiar subjects.* (5pt scale)

- Navigational

1. *In this country, how difficult or easy would it be for you to do each of the following? See a doctor.*
2. *In this country, how difficult or easy would it be for you to do each of the following? Search for a job.*
3. *In this country, how difficult or easy would it be for you to do each of the following? Get help with legal problems.*
4. 1 + number of correct answers to four civic knowledge questions (5pt scale)

9.5.3 Aggregation

Since each of the six dimensions has either two or four indicators, and all of the indicators are on five point scales that range from a least integrated to a most integrated option, aggregation is relatively easy mathematically. Nonetheless, it is worth pausing to think about what is assumed by adopting equal weighting, as the authors do:

> A score between 1 and 5 points is computed for each question such that there is a maximum score of 60 across all six dimensions for the IPL-12 and 120 for the IPL-24. The measure is then rescaled to range from 0 to 1 in increasing levels of integration.

Here are some response differences that all correspond to two point shifts towards integration on a single item from each of the dimensions:

- Responding that you feel a "very close" connection to the United States as opposed to a "weak connection".
- Responding that your household income is between 100% and 133% of US median income as opposed to between 33% and 66% of US median income.

- Responding that you understand the important political issues facing the United States "Well" as opposed to "Not well"
- Responding that you eat dinner with Americans who are not part of your family "once a week" as opposed to "once a year"
- Responding that you can read and understand the main points in a simple newspaper article on a familiar subject "Very well" as opposed to "Moderately well"
- Responding that seeing a doctor is "very easy" as opposed to "neither difficult nor easy"

Whether one wants to treat all these differences as identical is debatable (as the authors acknowledge) and for particular applications it may make sense to adopt different approaches to aggregating the indicators (Harder et al., 2018, p.11484). The decision to do so depends on the extent to which you are comfortable or uncomfortable with the differences above (and the many more you could similarly construct) being given equal weight in terms of the scale.

9.6 Application - National Poverty

The concept of poverty is good example of a difficult-to-measure concept, on which there is an extensive academic research literature. In 2019 the UK government adopted a new approach to publishing poverty statistics. The previous approach was based solely on income, but a Social Metrics Commission report concluded that these measures gave a misleading portrait of the "lived experience of poverty". One consequence of this, the authors of that report argue, is that published statistics substantially overstated the prevalence of poverty among those past working age. Pensioners are relatively likely to have low incomes but also are relatively likely to have substantial assets such as owning their own home and to have few regular financial obligations. Even adjusting for the number of people in the household, people on comparable incomes with children and limited assets may experience a much more precarious existence, both financially and more generally. Given this, what is the right way to measure poverty? Note that this distinction turns, critically, on what you want the concept to mean. Is poverty simply low income? Or is it something broader than that? This is a question of conceptualisation.

What indicates whether an individual is poor? Is poverty absolute or relative to other people in the same locality? Same region? Same country? In all countries? Is poverty unidimensional (just about money) or is it multidimensional (about money and other things)? Is the poverty level of a country the proportion of individual people in poverty in that country, or is it more complicated than that? We are mostly going to focus on the Multidimensional Poverty Index developed at Oxford and used by the United Nations, but with reference to other approaches as well.

9.6.1 Conceptualisation

The first step towards any kind of measure of country-level poverty is to be clear about which conceptualisation of poverty we want to employ. There are at least four approaches that have been discussed in the relevant academic literature (Laderchi et al., 2003), which I will very briefly outline here.

1. The *Monetary Approach*: Do people lack enough money?
2. The *Capability Approach*: Do people lack the range of capabilities they need to adequately function in the world?
3. The *Social Exclusion Approach*: Do people suffer exclusion from full participation in society?
4. The *Participatory Approach*: Do people think they are poor?

I have simplified each of these down to an implicit question in order to make the point that each conceptualization of poverty is based on answering a *different* question that corresponds to a *different* idea of what poverty is. Notice that moving from the concept of "poverty" to any of these conceptualizations pushes you towards different strategies for measuring the concept. For the monetary approach, you are going to need a definition of how much is enough money and some approach to measuring how much money people have/earn. For the capability approach, you are going to need to specify what capabilities are required to function "adequately" and some approach to measuring whether people have them. For all of the approaches, however you decide to measure individual-level poverty, you have to decide how that is translated to country-level poverty, both conceptually and practically. The former is likely to involve some decision about whether country-level poverty is more complicated than simply the proportion of people meeting the definition of individual-level poverty. Practically, measurement is likely to involve some kind of representative national survey. The examples we will look at are all based on assessing whether individuals meet a definition of poverty, and then using the national survey to assess what proportion of individuals meet the definition in the country overall.

You might notice that the monetary approach lends itself to more minimalist measurement strategies. The most commonly used country-level measure of poverty is that of the World Bank, which started out using the purchasing power parity equivalent of one US dollar per day as the international poverty threshold, and which is more recently $1.90 in 2011 purchasing power parity dollars. It is not quite right to say this is easy to measure—there are complications associated with conducting surveys in some countries, with putting monetary values on non-monetary transactions, implementing purchasing power parity adjustments across countries, etc—but adopting an income or consumption-based monetary threshold does make the measurement problem concrete. At the same time, this kind of minimalist measure is susceptible to criticisms that it omits important elements of the target concept. These

might include variable access to public goods as opposed to private resources (Laderchi et al., 2003).

Other researchers have developed more multidimensional measures that reflect the *capabilities* approach (more recently, the World Bank has also moved in this direction, see World Bank, 2018). The one we will look at in detail is the 2018 version of the Global Multidimensional Poverty Index (MPI)

> The original MPI was co-designed and launched in 2010 by the United Nations Development Programme (UNDP) Human Development Report Office (HDRO) and the Oxford Poverty and Human Development Initiative (OPHI) at University of Oxford. It was first published in 2010 as part of the Twentieth Anniversary of the Human Development Report (HDR). The original MPI were aligned, insofar as was then possible, with indicators used to track the Millennium Development Goals (MDGs). The global MPI has been published in every HDR subsequently, with adjustments that have been documented in the methodological reports.

In the original development of the MPI, Alkire and Santos (2010) write that "The potential dimensions that a measure of poverty might reflect are quite broad and include health, education, standard of living, empowerment, work, environment, safety from violence, social relationships, and culture among others." However, they ultimately develop a measure that only uses three of these dimensions: health, education and standard of living. Why? "…the binding constraint is whether the data exist. Due to data constraints (as well as, perhaps, interpretability) we have had to severely limit the dimensions. For example, we do not have sufficient data on work or on empowerment. Yet each of these dimensions should arguably be considered in a human development-based multidimensional poverty measure."

This kind of problem is extremely common in scale development. You have a conceptualization of the target concept that you find theoretically appealing, but you cannot identify relevant indicators. In the case of poverty measures, the social exclusion approach may be theoretically attractive for some purposes, but it turns out to be relatively difficult to operationalise (Laderchi et al., 2003). You have to make difficult choices about what you can measure, or what you can measure reliably, and there is seldom a right answer for all purposes. Often, it is trying to identify indicators that makes it clear that your conceptualization is not going to be usable as the basis of a measure.

So what are the key features of the MPI conceptualisation that we are going to focus on?

1. Poverty is a property of individuals.
2. Poverty is binary (you are poor or you are not poor) but there are also degrees of poverty among those who are poor.
3. Poverty is multidimensional, involving health, education and standard of living.
4. Country-level poverty is constructed by aggregating individual-level poverty, not by examining country-level statistics.

9.6.2 Indicators

The MPI uses 10 indicators in total, two for health, two for education, and six for standard of living. These measures are at the individual-level, based on household surveys.

Dimensions of Poverty	Indicator	Deprived if living in the household where…	Weight
Health	Nutrition	An adult under 70 years of age or a child is undernourished.	1/6
	Child mortality	Any child has died in the family in the five-year period preceding the survey.	1/6
Education	Years of schooling	No household member aged 10 years or older has completed six years of schooling.	1/6
	School attendance	Any school-aged child is not attending school up to the age at which he/she would complete class 8.	1/6
Standard of living	Cooking Fuel	The household cooks with dung, wood, charcoal or coal.	1/18
	Sanitation	The household's sanitation facility is not improved (according to SDG guidelines) or it is improved but shared with other households.	1/18
	Drinking Water	The household does not have access to improved drinking water (according to SDG guidelines) or safe drinking water is at least a 30-minute walk from home, round trip.	1/18
	Electricity	The household has no electricity.	1/18
	Housing	Housing materials for at least one of roof, walls and floor are inadequate: the floor is of natural materials and/or the roof and/or walls are of natural or rudimentary materials.	1/18
	Assets	The household does not own more than one of these assets: radio, TV, telephone, computer, animal cart, bicycle, motorbike or refrigerator, and does not own a car or truck.	1/18

Figure 9.3: Global Multidimensional Poverty Index (MPI) dimensions and indicators

Ideally one wants indicators that themselves have high measurement validity and reliability. It does not do much good if you break a single difficult measurement problem down into a large number of equally difficult measurement problems. It is important that the constituent dimensions are themselves easier to measure than the overall concept.

The measures in the table above are all reasonably straightforward factual propositions, although some are themselves tricky to assess. It is easy enough to determine what a floor is made out of if you are at someone's home, but what qualifies as "undernourished"? How far is the nearest source of safe drinking water in minutes walk, round-trip? These are non-trivial measurement problems in their own right. You can see that there are further decisions about thresholds lurking at the indicator level. Exactly which flooring materials indicate deprivation? Is a 30 minute roundtrip walk for clean water really good enough to not be deprived? Maybe it should be 20 minutes? It is often very difficult to make strong arguments for particular thresholds, but the core conceptualisation of poverty as binary at the individual level means that at some point thresholds are going to be required.

What is left off this list of indicators that could be feasibly measured? Table 4A.1 in the Annex of the 2018 World Bank Poverty Report lists indicators that are available, and their inclusion in various poverty indices. Some of the ones that are excluded from MPI (but appear in other indices) are monetary measures of standard of living (eg income below $1.90 per day), vaccination coverage and having a midwife at recent births as measures of health poverty,

and indicators like threats of crime and natural disasters which might be part of different dimensions of poverty like *security*. Note that adding additional indicators of an existing dimension like standard of living is a more modest change to the index than adding entirely new dimensions/sub-indices like security.

9.6.3 Aggregation

The approach followed by the MPI to aggregating the 10 indicators split 2-2-6 across three dimensions is a statistic M that Alkire and Foster (2011) call the "adjusted headcount measure". It may be useful here to refer back to the discussion in Chapter 6.5 of this class of poverty measures. The statistic is the weighted average proportion of deprivations, counting only deprivations among those individuals in a society who have enough deprivations in order to be considered "poor". $d_{ij} = 1$ when individual i is deprived with respect to indicator j and 0 otherwise. w_j is the weight that indicator j gets overall (where $\sum_{j=1}^{d} w_j = 1$). Given these definitions, p_i is the (weighted) proportion of dimensions on which individual i is poor, and M is the population-level statistic for the adjusted headcount measure.

$$p_i = \left(\sum_{j=1}^{d} w_j d_{ij} \right) \tag{9.7}$$

$$M = \frac{1}{n} \sum_{i=1}^{n} p_i \cdot I\left(p_i \geq k\right) \tag{9.8}$$

You might be thinking this is a bit complicated, so let's think about some simpler measures one might use instead. We might instead use a simple headcount measure that simply counts up the proportion of individuals who are poor:

$$H = \frac{1}{n} \sum_{i=1}^{n} I\left(p_i \geq k\right) \tag{9.9}$$

Alternatively, we might just calculate the average proportion of dimensions on which individuals are poor:

$$A = \frac{1}{n} \sum_{i=1}^{n} p_i \tag{9.10}$$

These are both plausible measures in their own right. H captures the "headcount" of individuals who have enough deprivations to meet the standard of being poor. This is determined by k, where the MPI uses $k = \frac{1}{3}$ such that an individual is poor if they are deprived on one third of the indicators, after they are weighted. A captures the average proportion of weighted deprivations for the entire population. H asks how many people are poor. A asks how much

poverty there is. The adjusted headcount measure M combines these two ideas, asking how much total poverty is there among the people who are poor.

Following our discussion from last time, we might think a bit about special cases for M. M is at most equal to A, because the requirement that we only count poverty for those people who qualify as poor means that any non-zero indication of poverty p_i that falls short of the cutoff k is not counted. If no one qualifies as individually poor, then M is zero, whereas A might not be.

We might also think about the units of these measures. H has units of $\frac{\text{persons}}{\text{persons}}$, it is a unitless population proportion. A has units of $\frac{\text{deprivation}}{\text{deprivation}}$, a unitless deprivation proportion. M similarly has units that cancel out. All of these are unitless measures. Note that the key to getting to these was applying thresholds for each indicator separately. By translating the presence or lack of electricity into deprivation vs no deprivation, and the number of years of schooling into deprivation vs no deprivation, the scale makes these things mathematically comparable and additive to form p_i, the weighted proportion of individual-level deprivations for individual i.

Whether the deprivations are really substantively comparable is a more difficult question. As you will have noticed above, not all of the indicators in the MPI get equal weight. The authors of that index made a decision to equally weight the three dimensions of health, education and standard of living. Having made that decision, and then concluding that they had six reliable indicators of the latter dimension and two indicators each of the first two, they made the decision to weight the indicators equally within dimensions. As a result, the nutrition and child mortality indicators get weights of $\frac{1}{2}$ within the health dimension, and since the health dimension gets weight $\frac{1}{3}$ overall, the nutrition and child mortality indicators get weights of $(\frac{1}{2})(\frac{1}{3}) = \frac{1}{6}$ with respect to the overall index. The same is true of the two education indicators, while the six indicators of standard of living each get overall index weight $(\frac{1}{6})(\frac{1}{3}) = \frac{1}{18}$.

Given the overall threshold $k = \frac{1}{3}$, this means that to achieve the overall status of being poor, an individual must either have deprivation on at least two of the four indicators in the health and education dimension, on one of those four plus at least three of the six in the standard of living dimension, or on all six of the standard of living dimensions. This means that, for example, an individual who is deprived on both of the education indicators but none of the others is poor, while an individual who is deprived on five of the six standard of living indicators, but none of the others, is not poor. If you thought this was wrong, that might constitute an argument for a somewhat different weighting or thresholding. For example, you could make both of those individuals count as poor by reducing k from $\frac{1}{3}$ to $\frac{5}{18}$, but there are many other ways you could change the weights as well.

It is often difficult to justify any particular weighting or threshold: these are very challenging aspects of this kind of scaling exercise. The authors of the MPI (Alkire and Santos, 2010) make a serious effort to discuss these issues. Regarding weights, they write:

> In the case of health indicators, it seems that malnutrition and mortality are both important deprivations and it is not clear which is the more important indicator. In the case of education, it could be argued that having one person with five or more years of schooling was the most important outcome; yet child school attendance is a time-sensitive input with long future returns, hence again we have weighted them equally. Weighting the six asset indicators equally is admittedly more difficult to justify and is also particularly important given that this is the dimension that contributes most to poverty in the poorest countries. Further research on the best comparable asset measures that can be constructed from multiple datasets would be useful in the future.

The authors of the MPI do a *sensitivity analysis* in which they examine how sensitive country-level poverty rates are to varying the relative weights on the three dimensions to 50-25-25, 25-50-25 or 25-25-50. They find that the relative ranking of countries changes only slightly (Alkire and Santos, 2010, Section 4.9). This is a generally useful strategy: if you are not sure exactly how to construct a scale, try some plausible different ways and see if it matters for anything you intend to do with the measures.

The authors also examine how country-level poverty rates vary for different levels of k, and show that their relative levels are not very sensitive to the choice of k, even as more or less stringent cutoffs increase or decrease poverty rates for all countries (Alkire and Santos, 2010, Section 4.8). The exact threshold would only matter for relative poverty rates of countries if certain countries were more likely to have lots of individuals just above or below particular thresholds. In the example under consideration that does not seem to be the case, but this sort of sensitivity analysis is valuable whenever you have to make a difficult-to-justify decision about indicator weightings or thresholds.

9.7 Application - Quality/Disability Adjusted Life Years

As a final example of the challenges associated with putting interval-level coefficients on different indicators, as well as the implications of choices about aggregation, we consider the ideas of "quality-adjusted life years" (QALYs) and "disability-adjusted life years" (DALYs) as a means for quantifying trade-offs in health economics (Zeckhauser and Shepard, 1976). The motivation of these measures is that it makes sense, from the perspective of a society, to spend more health resources where those resources will help prevent worse outcomes. The target concept of disability-adjusted life years is therefore the aggregate quality of remaining life for an individual, "worse outcomes" are those that reduce the lengths of peoples' lives as well as the quality of their lives. Since we do not know how long a given individual will live, or what their quality of life will actually be, the relevant indicators are the things that we can measure: an individual's current age and various features of their health status.

The obvious problem with developing such a measure is assessing the relative badness of different health outcomes and how they should be traded off. One way to assess this is using a person trade-off technique (Murray, 1994),

which follows the previously described logic of asking experts to specify the relative values of coefficients pairs. Experts were asked to choose between curing a certain number of individuals in one disability class versus another number in another class, to try to learn how people trade-off the "badness" of different disabilities versus death. The purpose of the disability-adjusted life years measure is to facilitate making difficult trade-offs about health resource allocation, and so identifying the relative weight to put on different kinds of bad/good health outcomes is unavoidably central to the problem. Here, we can think of different disabilities as different indicators, and the question is what weight they should have such that they become comparable for the purposes of the concept of quality/disability-adjusted life years.

Figure 9.4 highlights that explicitly facing the tradeoffs between indicators can be uncomfortable. Murray (1994) reports results from an expert panel conducted at the Centers for Disease Control (CDC) in the US which estimates that the health outcome resulting in someone needing *assistance with activities of daily living such as eating, personal hygiene, or toilet use* was assessed to have 0.92 the weight of a person dying. This means, for example, that a health intervention that results in 100 people of a given age being left with this level of disability has the same DALY value as an intervention which results in 92 of those people dying and 8 surviving without disability. The health researchers consulted in the development of DALYs did not think that one of these outcomes was clearly worse than the other.

Table 2: **Definitions of disability weighting**

	Description	Weight
Class 1	Limited ability to perform at least one activity in one of the following areas: recreation, education, procreation or occupation.	0.096
Class 2	Limited ability to perform most activities in one of the following areas: recreation, education, procreation or occupation.	0.220
Class 3	Limited ability to perform activities in two or more of the following areas: recreation, education, procreation or occupation	0.400
Class 4	Limited ability to perform most activities in all of the following areas: recreation, education, procreation or occupation	0.600
Class 5	Needs assistance with instrumental activities of daily living such as meal preparation, shopping or housework.	0.810
Class 6	Needs assistance with activities of daily living such as eating, personal hygiene or toilet use.	0.920

Figure 9.4: Disability class weights from Murray (1994), developed by person trade-off technique.

This kind of explicit numerical tradeoff involving peoples' lives often strikes people as unethical, at least on initial reading. Moreover, the structure of the QALY/DALY means that the lives of those with significant disabilities are treated as less valuable than those without disabilities. If the question is whether to save the life of someone with a disability or someone without a disability, these measures say the latter. This is a side effect of the fact that the measure treats the disabilities themselves as bad outcomes to be avoided. One might, instead, base medical decisions entirely on maximising life years, without disability adjustment. But this would mean putting no weight in decision-making on avoiding disability if that decision had any potential consequence for whether anyone lived or not. It is important to note here that the quantification is not what generates the ethical discomfort. Our ethical discomfort is guaranteed by the the motivating fact of a societal allocation of limited resources, which in turn requires choices about prioritisation, which is just another way of describing the need for a methodology to determine which outcomes are better versus worse overall.

Whether or not this is an avoidable problem, it is nonetheless an unusually stark example of the tradeoffs implicit in any linear index: you are stating that a given quantity of I_1 is equivalent to some quantity of I_2, and so on for the other indicators. If you are not comfortable with what this implies when you explicitly state the equivalence, then you need to either refine the b coefficients/weights until you are comfortable, or you need to reconsider whether

you actually want to be collapsing multiple indicators into a single concept. It may be that certain kinds of tradeoffs are unavoidable—eg you need to allocate some limited resources—but it may instead be the case that nothing is actually constraining you to find a unidimensional conceptualisation that requires making these trade-offs as part of the measurement strategy.

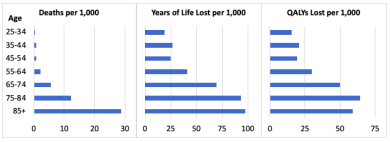


Figure 9.5: Effects of COVID on deaths, years of life, and QALYs in the US by age. https://twitter.com/JulianReif/status/1440248140427239428

9.8 Application - Global Health Security Index

The Global Health Security Index was created in order to measure the extent to which different countries were prepared for health threats, particularly from disease:

> The GHS Index is a project of the Nuclear Threat Initiative (NTI) and the Johns Hopkins Center for Health Security (JHU) and was developed with The Economist Intelligence Unit (EIU). These organizations believe that, over time, the GHS Index will spur measurable changes in national health security and improve international capability to address one of the world's most omnipresent risks: infectious disease outbreaks that can lead to international epidemics and pandemics. (Cameron et al., 2019, p5)

> The NTI, JHU, and EIU project team—with generous grants from the Open Philanthropy Project, the Bill & Melinda Gates Foundation, and the Robertson Foundation—worked with an international advisory panel of 21 experts from 13 countries to create a detailed and comprehensive framework of 140 questions, organized across 6 categories, 34 indicators, and 85 subindicators to assess a country's capability to prevent and mitigate epidemics and pandemics. (Cameron et al., 2019, p7)

The creators of this index thought that this was an important area for improving social measurement, and in light of subsequent events it would be difficult to argue otherwise. The 2020 coronavirus pandemic made exactly this kind of preparedness critically important for countries. But it also gives us a rare validation opportunity for an index of this kind. Did the scores that countries received on this index in 2019 predict their ability to cope with the pandemic that arrived soon thereafter?

At the time that I am writing this, in the summer of 2020, it is difficult to do a proper quantitative assessment of this question. Such an assessment is not straightforward given the many factors that affected which countries were seeded with cases early in the pandemic, and thus faced rising cases with the least time to prepare. With these caveats, it is nonetheless the case that the index appears to have failed at its stated goal, unless one defines "capability" in such a way that poor performance at preventing and mitigating an actual pandemic can be explained away as due to factors other than this capability.

The top ranked countries in the GHS Index (by a substantial margin) are the United States (83.5) and the United Kingdom (77.9), followed by the Netherlands (75.6), Australia (75.5), Canada (75.3), Thailand (73.2) and Sweden (72.1). This is a fascinating collection of countries given the development of the pandemic

through the summer of 2020. The United States' coronavirus response has been widely agreed to have been disastrous, and as of late July had cases and deaths trending upward. The United Kingdom locked down relatively slowly by comparison to other European countries, with the possibility that delay doubled the death total in the early stages of the pandemic. Sweden is notable as being one of very few countries that took a principled stance against the use of a "lockdown" to restrict movement and social contact, relying instead on voluntary social distancing.

As of 28 July, the United Kingdom (674 deaths per million), Sweden (563 deaths per million), and the United States (447 deaths per million) all ranked among the top ten internationally in terms of confirmed death rate, with the Netherlands (358 deaths per million) and Canada (235 deaths per million) not far behind. Of the top ranked countries in the GHS Index, only Australia (6.3 deaths per million) and Thailand (0.8 deaths per million) can be said to have performed very well at preventing fatalities. Of these, Thailand is the most notable, given that it is the only middle income country in the list. It appears to have been a genuine outlier in terms of having a highly successful pandemic response, and its presence near the top of the GHS Index is consistent with the possibility that the index was capturing something meaningful about Thailand's capacity in this area.[13]

[13] Thailand took costly action early, including banning all incoming passenger flights in April 2020, despite having an economy that is highly dependent on tourism.

Given that the GHS Index aimed to "assess a country's capability to prevent and mitigate epidemics and pandemics", the fact that so many of the worst hit countries ranked high on the index is reason for concern about the validity of the index. The examination of the countries at the top of the index is of course a very imprecise form of face validation. There is variation in coronavirus fatality rates that probably has little to do with anything that governments could do in immediate response to the emergence of the pandemic. Further, the countries that were most prepared in the ways that the GHS measured might have also been most at risk, either of pandemics in general or given the specific features of this pandemic in particular. A proper quantitative assessment would want to try to control for these other attributes of countries in order to assess whether the GHS Index provided any useful signal about the target concept that it aimed to measure. Given the observed patterns it seems unlikely that it provided a very strong signal, if any at all.

This is an example of one of the key perils of index construction. In contrast to cases where there is training data with which to calibrate the relevance of different indicators to the measurement of the target concept, the authors of the GHS Index did not have a great deal of training data to see how well countries would perform in a global pandemic. Previous epidemics in recent decades have been far more limited in scope, and have affected some countries and not others, providing little evidence regarding which attributes of countries *actually* help respond to disease outbreaks. While there are many indicators that are suggestive of capacity, and some of them really might matter for capacity, it seems that in 2020 those indicators did not turn out to be

very predictive of the ability of countries to actually "prevent and mitigate epidemics and pandemics".

9.9 Conclusion

One "feature" of scales constructed in this way is that there is seldom a natural metric for the target concept. For example, the immigrant integration scale discussed above is reported on a 0-1 scale, by linearly rescaling the 12-60 or 24-120 points to run from 0-1. But there is no right answer to what range the scale should cover.

One question you might have is whether these sorts of scale are really interval-level scales. Is a given increase really equally meaningful at all points of the scale? How credible the interval-level interpretation is depends on how the scale was constructed. To the extent that the interval-level assumptions are not met, this tends to undermines the validity of the scale overall. This is particularly true for linear indices, where the additivity of indicators requires that you believe in the interval-level interpretation. Failure of the interval-level interpretation implies you should not have been adding the indicators. Additivity is often simply a convenient mathematical structure, but if you think it is seriously wrong, you have a deeper problem than whether you can interpret your scale as interval-level rather than merely indicating ordering.

The biggest challenge in measuring the kinds of concepts we have been considering in this chapter is finding good indicators of those concepts, but indicators that are not individually already adequate as measure of the concept. Sometimes the challenge is finding the indicators at all. For example, the multidimensional poverty index that we examined was based on dimensions of health, education and standard of living. The authors of that index considered further dimensions of work, empowerment, the environment, safety from violence, social relationships, and culture, but simply could not come up with indicators of these theorised dimensions of development that could be feasibly collected for a wide range of countries (Alkire, 2013). Depending on how important you think these dimensions are relative to the three that were included, you could argue that the measure is *theoretically* missing over "half" of the target concept. *Empirically*, this might lead to measurement error because you have failed to capture important elements of what you are interested in. If those elements are correlated with the other variables in your analysis, this can lead you to erroneous conclusions. On the other hand, if all of these additional dimensions are highly correlated with the three that are included, then it might be that the measurement error is small despite the theoretically relevant omissions.

These kinds of measurement errors can be mitigated by interpreting results narrowly. The actual poverty index only measures three dimensions of poverty. So long as you remember that and state your results clearly, you won't be wrong. With all these kinds of measurements, one needs to be extremely

cautious about making causal claims where the measure is the independent/treatment variable. What would need to be the case for you to vindicate such a claim? You would need some kind of exogenous variation in the underlying concept, which manifested itself in the measure. There are circumstances where this could be plausible, but it is very likely that if you can identify them, there is an alternative definition of the treatment variable that would yield a clearer analysis. For example, you might note a natural disaster causing a shock to poverty in a given place, and then examine some downstream outcome. This might enable you to study the effects of poverty in a more causally credible way than is typically possible, but you would have difficulty disentangling those from direct effects of the hurricane that did not operate via its effect on poverty.

This is not to say that measurement scales like this have no place in causal analysis. There are situations where scale measures like these are appropriate running variables for regression discontinuity designs, but these only arise where an existing scale is used to make administrative decision. See, for example, an analysis by Lerman (2009) that uses the assignment of prisoners in California to different prison environments based on an additive scale of inmate background, crime and and sentence characteristics.

It is also reasonable to make causal claims where the kind of scale we have been discussing is the outcome variable. So long as there is a sound identification strategy for making a causal claim, these kind of scale measures can be a useful way of summarizing effects that manifest across a collection of indicators. In such situations it is reasonable to talk about causal effects of some treatment on your measure. You should be a bit more careful about making claims that there is a treatment effect on the underlying concept, because there is always the possibility that the causal effect is on the measurement error of your measurement strategy, rather than the component of the measure associated with the target concept.

10

Supervised Class Measurement

As we have followed a series of methods from principle components to factor analysis in Chapter 11 to the item response models covered in Chapter 12, we have left behind the idea of measures as summary linear functions of indicators that motivated Chapters 8 and 9. With factor analysis, and even more so with item response models, we have been developing the idea of generative models where indicators arise from latent variables. These still have linear functions at their core, but linear functions of the latent variables (the quantities to be measured) rather than of the indicators. In the case of the item response models, our indicators are then non-linear (logistic) functions of that linear relationship.

In this chapter, we will turn from tools for measuring continuous quantities to measuring categorical quantities. This chapter and the next one will re-explore some of the methodological dimensions mapped out in the last four chapters for these different quantities of interest. We will once again consider supervised and unsupervised methods, summary methods and generative models, and indicators with different levels of measurement. This chapter will cover methods based on theoretical arguments and supervised methods that deploy relevant training data, while the next chapter will cover unsupervised methods.

There is one important novel issue that arises when one is measuring a concept that is properly understood as categorical rather than continuous. Under what circumstances should we work with *point classifications* of the categorical quantity versus working with *probabilistic classifications* of that quantity? That is to say, if the indicators you have and the measurement strategy you have developed say that a unit has a 75% chance of being type A and a 25% chance of being type B, when does it make sense to say that the *measure* is "type A" and when does it make sense to say that the measure is a "0.75 type A, 0.25 type B"? When we measure continuous, interval-level quantities we do not have to worry about this issue in the same way because the expected value of the quantity we are trying to measure is also a valid level of the variable. This is not true for binary quantities.

However, before we turn to this question of what kinds of measures of cat-

egorical concepts are most appropriate to use in different applications, first we need to consider how to make the decision to conceptualise a quantity as categorical in the first instance. Once we have done that, and have considered the question of whether a categorical concept necessarily needs a categorical measure, we can proceed to the mechanics of supervised measurement. Here we will examine two approaches, appropriate for applications where we have training data and for where we do not. First, corresponding to Chapter 8 for continuous concepts, we will discuss the use of logistic regression and alternatives[1] where training data are available. Then, in a way that is analogous to the discussion of index construction in Chapter 9, we will discuss the development of "coding rules" for creating categorical variables from indicators where there is no training data. We will then consider the very large number of ways of evaluating classification error, and their relevance to describing error in categorical measurements.

[1] There are many such alternatives, as a tremendous amount of work in machine learning is on classification problems, and a very large number of tools for classification exist as a consequence.

10.1 Conceptualisation

Measuring classes (categorical variables) rather than scales (continuous variables) is often a choice of the analyst. In some contexts, one can easily imagine either continuous or categorical conceptualisations of the target concept that you want to measure. For example, consider the problem of measuring whether countries are democratic. There are continuous scale measurements of the concept of democracy—to what extent are countries democratic?—but this concept can also be conceptualised in a binary way. Is the US a democracy? Is Russia a democracy? Is Iran a democracy? The answers could just be yes or no. On the other hand, they could be "mostly" or "barely" or "a bit more so than this other country". This is ultimately a choice about how we want to talk about the underlying social science concept. Categorical conceptualisations require making sharp classification choices at the margins in a way that continuous conceptualisations do not. Sometimes this is what you want, sometimes it is not. For some purposes it is important to think in terms of classes while for other purposes it may be more sensible to acknowledge a continuum of variation via a scale, even if that scale is defined by end points that correspond to ideal types.

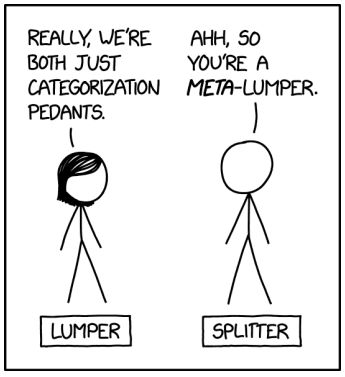


Figure 10.1: Lumpers and Splitters https://xkcd.com/2518/

Non-binary categorical measurements come in the usual types of levels of measurement—nominal or ordinal–depending on whether their levels are ordered in a single dimension. To continue the theme of measuring democracy at the country level, there is a rich history of scholars defining different conceptualisations of regime type. Plato describes five regime types—Aristocracy, Timocracy, Oligarchy, Democracy and Tyranny—which have a logical ordering in Plato's thought (Republic, Book VIII). Aristotle defines six regime types, which distinguish between the extent to which political power is concentrated versus diffuse and whether the regime is "correct" or "deviant" (Miller, 2017):

	Correct	Deviant
One Ruler	Kingship	Tyranny
Few Rulers	Aristocracy	Oligarchy
Many Rulers	Polity	Democracy

Inventing new regime *typologies* has never gone out of fashion.[2] Political scientists continue to revise and reinvent typologies of regime type. Sometimes this involves subdividing existing types, for example dividing democratic regimes into subtypes such as "liberal", "constitutional", "electoral" and "limited" (Wigell, 2008). Sometimes this involves mapping out relationships between different types that have already been discussed by previous scholars.

There are two common ways to think about the relationships between different categories in these sorts of typologies. One of these is in terms of conceptual trees, with branches defined by relevant attribute differences between types. See, for example, the regime types described by Anckar and Fredriksson (2019) depicted in Figure 10.2. The other is in terms of latent spaces, with dimensions of difference defined by different attributes. See, for example, the regime types described by Wigell (2008) in Figure 10.3.

The choice between describing a typology in terms of a tree or a latent space is related to the logical relationships between the types. Trees are more useful where the attributes distinguishing types "interact" in stronger ways, such that the relevance of one attribute depends strongly on the levels of the others. For example, in Figure 10.2, the attributes that would distinguish between "Monarchy (Constitutional)" and "Republic" in the left branch under "Democracy" may not be relevant to distinctions among different types of "Autocracy" in the right branch. Spatial metaphors are more relevant where the attributes that are relevant to categorization are more independent. Thus, in Figure 10.3, increasing "Electoralism" moves the regime type from "Authoritarian Regimes" to "Electoral-Autocratic Regimes" if "Constitutionalism" is low, and from "Constitutional-Oligarchic Regimes" to "Democratic Regimes" if "Constitutionalism" is high, and thus it is relevant in either case.

The spatial metaphor conveys ordering information in a way that the branching metaphor does not, and thus is conducive to conceptualisations involving ordered categories. For example, as one moves up the right edges of the two panels in Figure 10.3, at high "Constitutionalism", one goes from "Constitutional-Oligarchic Regimes" to "Constitutional Democracy" to "Liberal Democracy" (note that the second panel, Figure 4 from the original paper, is entirely embedded within the top right panel of the first). This kind of ordinal classification raises issues of where to set the thresholds: how much electoralism do you need to move "up" a classification? However thresholding issues are challenges with all categorical conceptualisations, in most contexts there will be difficult border cases, regardless of whether your conceptualisation uses a tree or a spatial metaphor.

[2] The Wikipedia page on forms of government lists nearly 100 types of regime that someone, somewhere, has previously defined. The page helpfully notes that "This article lists forms of government and political systems, according to a series of different ways of categorizing them. The systems listed are not mutually exclusive, and often have overlapping definitions."

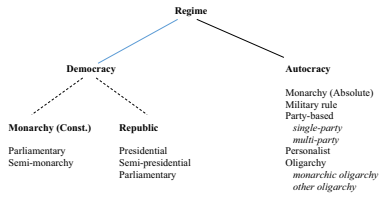
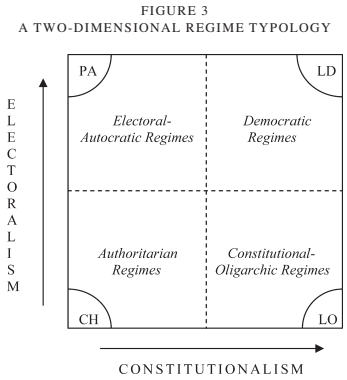


Figure 10.2: Figure from Ancker and Fredriksson (2019) depicting a typology of regime type using a conceptual tree.



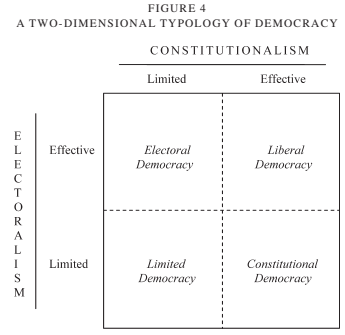


Figure 10.3: Figures from Wigell (2008) depicting how a typology of regime type relates to an underlying 2D conceptual space of electoralism and constitutionalism, both across democratic and non-democratic regime types (top) and also ithin democratic regime types (bottom).

The mechanics of how one relates the boundaries between categories to indicator data is something we will discuss in the next two sections. The first of these describes the use of training data to estimate the relationship between indicators and the classes/categories that one wants to measure. The second of these describes the use of expertise to define a set of "coding rules" to map values of indicators onto values of the classes/categories that one wants to measure. The former is the classification analogue to the kind of analysis that we saw in Chapter 8 for measuring scales, the latter is the classification analogue to the index construction techniques that we saw in Chapter 9.

10.2 Supervised measurement with training data

When we were considering the measurement of interval-level quantities, we considered examples of running regression models to predict a set of known measures using a set of indicators in a training data set. We could then use those same indicators, measured for a larger set of units, to construct measures that reflected the relationships we had observed between the indicators and the "gold standard" measures in the training data. This depended critically on the availability of pre-existing measures so that we could do the training at all, but where possible, it enabled us to find the combinations of indicators that best predicted the quantity that we wanted to measure (or at least the training data's approximation thereof).

We can use this same logic for categorical nominal-level or ordinal-level quantities by using an appropriate discrete choice / limited dependent variable regression model. For classification into two categories, we can use binary logistic regression; for classification into more than two ordered categories, an ordinal logistic regression; for classification into more than two unordered categories, a multinomial logistic regression.[3] Where we have branching logic in our conceptualisation as in Figure 10.2, we can use nested logit models to reflect this as well. The mapping between discrete choice models and the conceptualisation of the quantity to be measured is usually reasonably clear, the challenges in applying these methods are primarily identifying strong indicators and having the training data necessary to follow this approach at all.

[3] Regularized logistic regression is useful in data sets with large numbers of indicators relative to training observations.

One need not use logistic regression for these problems. There are a bewildering number of potentially applicable alternative statistical models and machine learning tools that one might use instead. These include probit regression, support vector machines, linear discriminant analysis, naive Bayes classifiers, regression trees, random forests, and neural networks. However, what they all share, is that they are tools for determining a mapping from indicator values to the classes/categories that one wants to measure, using a set of training data for which the latter has already been measured some other way. Once this mapping has been estimated, it can be applied out-of-sample as a measurement strategy for new units.

This non-exhaustive list includes some methods that provide only point

classifications and some that also provide probabilistic classifications. We can see this distinction most easily in the context of binary logistic regression. Consider the case where we have a set of "gold standard" measurements $m \in 0, 1$ from some pre-existing measurement procedure for the concept of interest $\mu \in 0, 1$. We will use this *training data* to *calibrate* a new measurement procedure. Our new measurement procedure will be based on a set of one or more indicators I—I_1, I_2, etc—that we want to use to to measure the concept of interest. Our goal is to determine how to most effectively use them to approximate μ, given the indicator variables I that we have, plus the information contained in m about how they relate to μ.

If we use *logistic* regression to estimate the relationship between I and m, we assume a model of the form:

$$\frac{p(m_i = 1)}{p(m_i = 0)} = \alpha + \beta_1 I_{1i} + \beta_2 I_{2i} + \cdots \tag{10.1}$$

The predicted probabilities of a logistic regression is given by the formula:

$$\widehat{p(m_i = 1)} = \frac{e^{\alpha + \beta_1 I_{1i} + \beta_2 I_{2i} + \cdots}}{1 + e^{\alpha + \beta_1 I_{1i} + \beta_2 I_{2i} + \cdots}} \tag{10.2}$$

We now have a choice. We could say that this predicted probability $\widehat{p(m_i = 1)}$ *is* our measure, or we could use it to make a point prediction for m_i, where $\widehat{m_i} = 0$ if $\widehat{p(m_i = 1)} < 0.5$ and $\widehat{m_i} = 1$ if $\widehat{p(m_i = 1)} \geq 0.5$. The point classification $\widehat{m_i} \in 0, 1$ has the virtue of being binary, like the target concept. The probabilistic classification $\widehat{p(m_i = 1)} \in [0, 1]$ has the virtue of incorporating uncertainty about the true classification of the unit, and turns out to have more attractive properties for many applications, as we will discuss in a later section of this chapter.

10.3 Coding rules

In many contexts, we want to develop classifications without pre-existing training data by deploying relevant theoretical arguments. As was the case when we discussed theoretically motivated measures in Chapter 6 and also index construction in Chapter 9, this requires deploying relevant expertise to determine how the indicators aggregate up to a classification.

One way you might do this would be to follow the index construction strategies described in Chapter 9, and then set thresholds in the index to translate to a(n ordered) categorical measurement. Many of the indices that we examined previously do this as well as providing index values. For example, the Global Health Security Index categorises countries into "Least Prepared", "More Prepared" and "Most Prepared" based on thresholds in an underlying index. These ordinal categorisations of index tend to use fairly arbitrary thresholds.

The more interesting uses of expertise to determine how indicators aggregate up to a classification are those that involve specifying a "coding scheme"

that directly maps the indicator levels into the classes. As an illustrative example, we will consider a binary classification of regimes into democracies and non-democracies by Alvarez et al. (1996). The authors set out four rules for classifying which countries were non-democracies in which years. Any one of these rules applying to a country-year is sufficient to classify that country-year non-democratic. - Rule 1. "Executive Selection." The Chief Executive is not elected. - Rule 2. "Legislative Selection." The Legislature is not elected - Rule 3: "Party." There is no more than one party. - Rule 4: A regime passes the previous three rules, the incumbents will have or already have had continuously held office by virtue of elections for more than two terms or without being elected for any duration, and until today, or the time when they were overthrown, they have not lost an election. The last rule is convoluted, but is carefully worded to address cases where countries have had a long period of elections without any transfer of power, and it is unclear that the incumbents would relinquish power were they to lose an election.

Regardless of the merits of this coding scheme for regime type as such, it illustrates how relevant expertise can be deployed to specify a classification scheme. In their paper, Alvarez et al. (1996) carefully explain the logic and application of each of their rules in order to create a set of indicators, as well as why they are aggregated in the way that they are. They require that a democracy have a directly or indirectly elected executive, directly elected legislature, electoral competition, and evidence of turnover in office. This aggregation rule is multiplicative (all of the above are required) rather than additive, and the authors explain why they think being a democracy requires all of these things and that we should not talk about partial democracies. The authors also do not focus on franchise requirements, explaining that they are focusing on the existence of contestation rather than any broader notion of the system being representative. These are all choices, and other ones are possible.

For our purposes here, the point is that this is the classification equivalent of index construction. By its nature, constructing an interval-level measure requires mapping different indicator levels, and combinations thereof, onto a common metric scale. In contrast, constructing an ordinal-level or nominal-level measure requires mapping different indicator levels, and combinations thereof, onto categories. In both instances, the validity of the measure relies on the quality of the expertise, in these kinds of measurement strategies there is no training data to fall back on to determine the relationships between indicators and measure.

10.4 Point classifications versus probabilistic classifications as measures

When we considered continuous, interval-level measures, we implicitly took advantage of a feature of such scales which is not present for binary/categorical measures. For interval-level quantities, $E\left[\mu|I\right]$ is itself a valid level of μ. That is to say, if the expected value of the quantity of interest that we wanted to

measure given all the indicators we have is 3.7, then our measure m can be 3.7. This is not true for categorical variables. For example, for a binary variable, $E[\mu|I]$ will almost never be 0 or 1, it will be some probability $p(\mu = 1|I) \in [0, 1]$ that is typically between 0 and 1.

As noted earlier, this leaves us with a choice. We could say that this predicted probability *is* our measure, or we could use it to make a point classification for m_i based on whether it is greater or less than 0.5. The point classification has the virtue of being binary, like the target concept. The probabilistic classification has the virtue of incorporating uncertainty about the true classification of the unit. Our task in this section is to consider some of the implications of this choice.

The idea that we should work with a binary m_i is very intuitive, surely our measure of a binary quantity should itself be binary? But in many kinds of analysis this is a poor choice, as we will now illustrate with a toy example.

Imagine that we have a single indicator $I \in 0, 1$ for a binary quantity $\mu \in 0, 1$ that we want to measure. Based on that indicator, and some measurement procedure that we have developed, we can say that the expected value $E[\mu_i|I_i = 0] = \mu_{i0}$ when $I = 0$ and $E[\mu_i|I_i = 1] = \mu_{i1}$ when $I = 1$. That is, the indicator tells us something about the quantity we want to measure, such that if $I_i = 0$ that gives us one expectation about the quantity (μ_{i0}) and if $I_i = 1$ that gives us a different expectation about that quantity (μ_{i1}).

Our ultimate application is that we are interested in comparing the mean values of some other variable Y for $\mu = 0$ and $\mu = 1$: $E[Y_i|\mu_i = 0]$ and $E[Y_i|\mu_i = 1]$. We assume that Y depends only on μ_i, not I_i, such that $Y_i = \alpha + \beta\mu_i + \epsilon_i$, where ϵ_i has an independent normal distribution with mean zero and standard deviation σ. Overall, we want to run a simple linear regression or do a comparison of means between two groups, but we have an imperfect measure of which units are in which group.

To give an immediate intuition for why we might not want to use the point classification, consider the case where $\mu_{i0} = 0.1$ and $\mu_{i1} = 0.4$. That is to say, if we observe an indicator value of 0 for unit i, our measurement strategy says that there is a 10% chance that unit i has $\mu = 1$. If we observe an indicator value of 1 for unit i, our measurement strategy says that there is a 40% chance that unit i has $\mu = 1$. These are both less than 50%, so the point classification of every unit is $m = 0$. This means **we cannot run the regression** because there is no variation in the x variable. In contrast, if we use the probabilistic classifications $m = 0.1$ for the units where $I_i = 0$ and using $m = 0.4$ for the units where $I_i = 1$, not only are we able to run the regression but the expected value of the difference in means is unbiased: $E\left[\hat{\beta}|m\right] = \beta$.

Even in cases where the indicator induces variation in m, using the point classifications still leads to bias in the subsequent analysis. If $\mu_{i0} < 0.5$ and $\mu_{i1} > 0.5$, then the measure is always equal to the indicator $m_i = I_i$ for all units i. Using these point classifications, $E\left[\hat{\beta}|m\right] = (\mu_{i1} - \mu_{i0})\,\beta$, which will be less than β unless $\mu_{i0} = 0$ and $\mu_{i1} = 1$, which is to say if there is any measurement

error at all.[4]

The key point here is that if you can quantify the measurement uncertainty in a categorical variable, you should use the probability classification rather than the point predication as your measure in subsequent analyses that look for relationships with other variables. Of course in some applications there is no measure of uncertainty about the classification, so you cannot do this. But in such instances you should still be worried that subsequent analyses will understate the true differences between mis-measured classes, particularly when one class is much more frequent than the other, and so the misclassifications are likely to be asymmetric.

10.5 Application - Predicting Clinical Diagnosis of Depression, Part 2

In Chapter 12 we looked at data from the PHQ-9 depression screening instrument, which consisted of 9 items. The standard scoring scored the four response levels on each item on a 0-3 scale, and the entire instrument as the sum of these on a 0-27 scale. As noted in the discussion there, the best evidence available suggests that using a score of 10 and above best predicts clinical diagnosis of depression, achieving a sensitivity of 0.88 (95% interval: 0.83-0.92) and a specificity of 0.85 (95% interval: 0.82-0.88) (Levis et al., 2019).

But if our goal is to predict clinical diagnosis of depression, why form a scale at all? Why not attempt to classify respondents into those who will receive clinical diagnoses of depression versus those who will not? In this section, we will use a data set recording how 656 individuals responded to a pool of 88 depression indicators in order to try to identify a simple measurement strategy for classifying those who will receive a clinical diagnosis of depression versus those who will not.[5] Since there are 88 different items, I will not list them here. They are largely variations on the kinds of questions in the PHQ-9, which we examined previously in Chapter 12, with varying wordings and numbers of categories. Of the individuals in the training set, 25% were clinically diagnosed with minor or major depression.

I begin by using the graded response model to jointly scale all 88 depression indicators, as we covered in the previous chapter. Figure 10.4 illustrates that the item response model scores from this model appear to predict clinical diagnosis of depression pretty well.

Figure 10.5 shows how we can assess the ability of the item response model scores to predict the clinical diagnosis. The plot shows how the sensitivity and the specificity vary across different thresholds that one might use in the factor score. If one uses a very high threshold, the sensitivity is very low because one fails to find the true cases of depression, but specificity is very high because there are no false positives. This corresponds to the lower right of the plot. As one lowers the threshold, sensitivity improves rapidly and specificity declines slowly because most of the clinically diagnosed cases have high scores. However, eventually the threshold falls below most of the clinically diagnosed cases,

[4] These results get messier if the measurement error in m versus μ is potentially related to the outcome Y, but we know that already from Chapter 5. It is still the case that using the probabilistic classifications is preferable in almost all circumstances to using the point classifications.

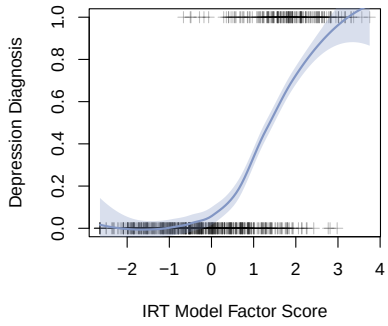


Figure 10.4: Predicting depression diagnosis with 88 item depression item response model scores.

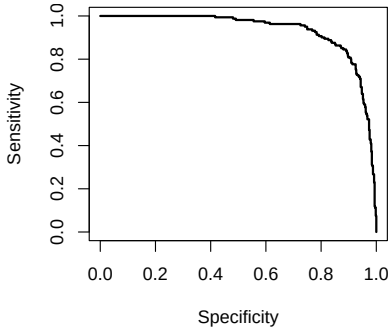


Figure 10.5: Sensitivity versus specificity for different thresholds of item response model score predicting depression diagnosis.

[5] I thank Robert Gibbons for helpful discussion and sharing these data. The data I use here are from Gibbons et al. (2013) which follows a more sophisticated version of the strategy I describe here, but see also Gibbons et al. (2012) for analogous work related to the unsupervised scaling from the previous chapter of this book.

and further reductions hurt specificity with only modest gains in sensitivity, as one approaches the top right of the curve. There are different ways to define optimal thresholds to trade off sensitivity and specificity, but those will all correspond to locations near the top right of the curve.

Recall that the preceding was an *unsupervised* analysis. We estimated the scale that best predicted the responses to the depression indicators, but in no way used the information about which individuals were clinically diagnosed. If we now turn to a supervised approach, following the strategies described in this chapter, we might be able to do better at developing a measurement strategy for identifying those who will be clinically diagnosed with depression, given the set of indicators that are available in these data.

These data have a large number of categorical indicators (88) relative to the number of units (656). If we ran a standard logistic regression, with dummy variables for each non-baseline level of each indicator, we would likely end up with a model that was substantially overfit because the number of parameters (262) would be very nearly the number of units for which we had observations. So instead, I run a logistic regression with lasso regularization. The details of how this works are beyond the scope of this text, but the key idea is that lasso regularization prevents overfitting by penalising complex models (large coefficients). Lasso regularization leads to models with only a few non-zero coefficients, even when there are a large number of predictor variables. I use this to find a relatively simple model of 88 indicator thresholds that predict clinical diagnosis well, using cross-validation to ensure that the model fits well out-of-sample.[6]

When we fit this logistic regression with lasso regularization, it estimates non-zero coefficients for 20 of the response thresholds in the data, which involve 16 of the items. Exactly which items are the most predictive ones is not really our focus here, so I have not included a table of the specific items and response thresholds here. In application, one would want to carefully consider these as a form of validity checking.[7] This means that, if this measurement model were chosen for purposes of future measurement, one would only need to survey these items, not all 88.

Figure 10.6 shows that the predicted probabilities from this model are highly predictive of depression diagnosis.[8] Figure 10.7 shows the sensitivity versus specificity tradeoff for these predicted probabilities. The improvement on the IRT model scores that we previously examined is small, but there is some. In some applications, the improvement in going from an unsupervised model to a supervised model could be quite large, but this is an instance where even the unsupervised model does quite well because the items included in the data set have been carefully curated to be closely related to the concept of interest.

These data do not include respondent level covariates for those completing the depression item battery, but it is worth noting that such variables could improve classification. For example, there could be gender differences in how respondents give answers to these questions, such that the same set of re-

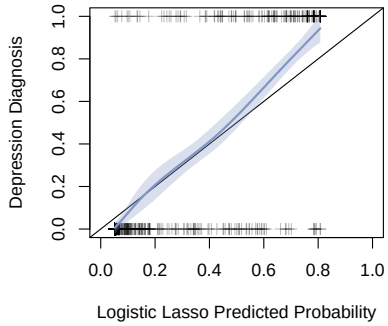


Figure 10.6: Predicting depression diagnosis with logistic lasso regression.

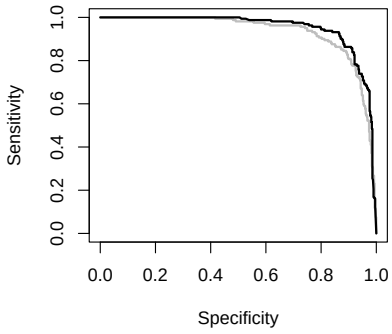


Figure 10.7: Sensitivity vs specificity curve for logistic lasso regression (black) and for the previously described IRT model scores (grey) predicting depression diagnoses.

[6] Implementation is via the cv.glmnet() function in the glmnet library for R (Simon et al., 2011).

[7] The single response threshold which has the largest coefficient in the selected model is threshold 2 for the item "In the past 2 weeks, I felt sad."

[8] Careful readers may wonder why the the proportion of depression diagnoses does not map onto the predicted probabilities one-to-one, but rather increases with a slope somewhat greater than one. This is due to the lasso penalty combined with cross-validation, which is optimising for out-of-sample prediction, rather than the in-sample prediction depicted in the figure.

sponses is more or less indicative of depression for men versus women. Or it could be the case that new parents might start responding differently to any/all of these items after the birth of their child:

- In the past 2 weeks, I had difficulty concentrating.
- In the past 2 weeks, I had difficulty sleeping.
- In the past 2 weeks, I felt sleepy all the time.
- In the past 2 weeks, have you had difficulty staying asleep?

Is this an indication that the items and the survey instrument are inapplicable to people caring for newborns? There are alternative diagnostic instruments specifically designed for postnatal depression that ask somewhat different questions. Or is it an indication that the scores should be adjusted in some way (by including whether someone is caring for a newborn in the model)? Or perhaps a lot of parents of newborns *are depressed*? Perhaps psychological interviews misdiagnose these because they dismiss the symptoms that are in the instrument? These are the sort of subtle issues that are worth considering, particularly when you are refining a measurement strategy that seems to do reasonably well overall. Making these assessments relies on having a very clear definition for the concept, which can be difficult for a concept like depression.

This discussion also highlights the possibility that a measurement strategy that works well on average for a population overall may not work well in all sub-populations. In this instance, the depression measurement model that fits well on average may not fit new parents very well, which could motivate including data on being a new parent in the training model or the use of an alternative measurement instrument for that sub-population. This is a very general problem, which has been gaining increasing attention in machine learning research in recent years due to phenomena like facial recognition systems being systematically worse at recognising the faces of people with darker skin. Supervised measurement strategies are trained to be accurate for the data with which they are trained. If those data are unrepresentative of populations, the measurements will tend to perform better on the sub-populations that are overrepresented. Even if the data are representative of populations, the measurements will tend to perform better on larger sub-population groups.

11

Unsupervised Scale Measurement with Interval-Level Indicators

In the last few chapters, we considered different ways to assemble sets of indicators into an index that measured a concept of interest. In Chapter 8 we considered cases where we had training data—pre-existing measures of the target concept—for some units that we could use to estimate the relationship between indicators and the target concept. In Chapter 9 we considered the various ways that expertise—either that of the analyst or a broader set of experts—can be used to determine that relationship.

In this chapter, we are going to consider what we might do with the set of indicators, taken by themselves. The novelty in this chapter is that we are trying to find *a* concept in the data rather than trying to find the data that measures our concept of interest. Methods that work in this direction are *unsupervised* measurement methods, as opposed to the *supervised* methods we have considered previously. Supervised methods require supervision in the sense that you, the analyst, are determining how the concept is measured from the available data. Unsupervised methods are unsupervised in the sense that the data is determining what concept is measured.

Unsupervised methods are powerful and widely used. They are also widely abused and misinterpreted. Indeed they have a rich history of abuse going back to their invention (on which more below, but also in Chapter 1). Unsupervised methods discover the dimension(s) that best explains variation in the indicators you use, in the data set you have. This is not necessarily the concept that you actually want to measure. Even from stating it in English, we can see that the criterion of "best explaining variation in the indicators" makes no reference to any specific target concept. I will show some examples where this criterion works well, and some examples where it works less well.

In this chapter we will focus on two mathematical methods for doing this translation, principal components analysis and exploratory factor analysis. These methods are conceptually very different: principal components analysis is a summary method while exploratory factor analysis is a generative model. Nonetheless, much as we saw in Chapter 7 with tallying up points for wins,

losses and draws (summary) versus a Bradley-Terry model (generative), the measures we construct with both methods tend to be very similar in practice.

11.1 Principal Components Analysis (PCA)

Principal components analysis aims to summarize the variation in a matrix of indicators I_{ij}. We have p observed variables I_{ij} $(j = 1, 2, \ldots, p)$ measured for each unit i in a sample of data. Where $var\left(I_j\right)$ is the variance of observed indicator j across all units i in the data, the **total variance** of the p variables is

$$\sum_{j=1}^{p} var\left(I_j\right) \tag{11.1}$$

This is the overall variation between the units in the data across all the indicator variables. The idea of principal components analysis is to *re-describe* the data in terms of a smaller number of *uncorrelated* new variables that capture as much as possible of the total variance. These uncorrelated new variables are the **principal components**.

Because we like things that are linear, we define the the principal components m_{ik} $(k = 1, 2, \ldots, p)$ as **linear combinations** of the original variables:

$$\begin{aligned}
m_{i1} &= a_{11}I_{i1} + a_{21}I_{i2} + \ldots + a_{p1}I_{ip} \\
m_{i2} &= a_{12}I_{i1} + a_{22}I_{i2} + \ldots + a_{p2}I_{ip} \\
&\vdots \\
m_{ip} &= a_{1p}I_{i1} + a_{2p}I_{i2} + \ldots + a_{pp}I_{ip}
\end{aligned}$$

In other words, each component is a weighted sum of the original xs, where the a_{jk} are **weights** or **coefficients**.[1] We are omitting indices for specific units to keep this from getting overly confusing, you should imagine these equations applying to each specific unit. Hopefully this is clear so far, but the key question is how we select the a_{jk} so as to achieve the goals that we specified above: (1) that the new variables m_k are uncorrelated with one another and (2) that they redescribe the total variation in the original x variables and (3) that we describe as much of the variation as possible with the first 1, 2, 3, etc principal components.

[1] Note that we are using Roman letters here because these coefficients are not part of a generative process.

To satisfy the first two conditions—that (1) our principal components are uncorrelated with one another and (2) redescribe the total variation in the original variables—we need the weights a_{jk} to satisfy the following:

$$\begin{aligned}
\sum_{i=1}^{p} a_{jk}^2 &= 1 \text{ for each } k = 1, 2, \ldots, p \\
\sum_{i=1}^{p} a_{jk}a_{jk'} &= 0 \text{ for every pair } k \neq k'
\end{aligned}$$

Together, these conditions ensure that the total variance of the PCs is the same as the total variance of the original variables $\sum_{j=1}^{p} var\left(m_j\right) = \sum_{j=1}^{p} var\left(I_j\right)$ as well as that all principal components are uncorrelated with each other $\text{corr}(m_k, m_{k'}) = 0$ for all $k \neq k'$.

This gets us most of the way to a solution, but not all the way there. There are still infinitely many sets of coefficients a_{jk} which would satisfy these constraints. We want to use the ones that explain the most variation possible with the initial principal components, so that we can explain as much variation as possible with the fewest components. In order to achieve this, the specific values of the coefficients a_{jk} are obtained from the eigenvalue decomposition of the correlation (or covariance) matrix of $I_1, \ldots, I_p$. $\left(a_{1k}, a_{2k}, \ldots, a_{pk}\right)$ is the eigenvector corresponding to the kth eigenvalue λ_k. The whole operation is equivalent to an orthogonal rotation of the p-dimensional space of the values of the p variables. For more details on the mathematics by which the coefficients/weights are calculated, see Chapter 5 of Bartholomew et al. (2008) or any other detailed treatment of PCA.

Having defined the principal components in this way, $var\left(m_k\right) = \lambda_k$, the variance of the kth PC is equal to the kth eigenvalue and the proportion of the total variance explained by the first q principal components is:

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_q}{\lambda_1 + \lambda_2 + \ldots + \lambda_p}$$

The principal components are in order of decreasing variance, $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots \geq \lambda_p$. Subject to the constraints, particularly that the total variance of the original data is preserved, each of the variances is as large as possible:

- m_1 is that linear combination of the original variables which has the largest variance across units i.
- m_2 is the linear combination uncorrelated with m_1 which explains the largest proportion of the total variance not explained by m_1 across units i.
- and so on for $m_3, \ldots, m_p$

11.1.1 Properties

Principal components analysis is sensitive to the scale of the original variables because it operates on the variances of those variables. This has several importance consequences to keep in mind. If you change the scale of all variables proportionately the principal components will not change. But if you change the scale of a single variable proportionately, the principal components will change. Because the procedure aims to redescribe the variance of the original data, with as much variance as possible explained by the initial components, if you increase the variance of one of the variables but not the others, the initial principal components will increasingly focus on explaining variance in that variable rather than the others because there is more variance in that variable as a share of the total. If $var\left(I_j\right) > var\left(I_{j'}\right)$, I_j will receive more "weight" than $I_{j'}$ in PCA.

This is particularly an issue if different variables have different units of measurement, in which case their relative variance is substantively meaningless and arbitrarily determined by the units of the difference indicators. Generally we do not want the results to be influenced by such differences and so the variables in a principal components analysis are usually standardised first so that the principal components weight the variation in each indicator equally. Recall from chapter 6 that a standardised variable has sample mean 0 and standard deviation 1 and is derived by first subtracting the sample mean from each observation of that variable and then dividing by the sample standard deviation. Therefore, the total variance (as defined above) of p standardized variables is p, the number of variables.[2] Since each variable contributes a variance of 1, PCA treats them all as having equal weight, and puts equal weight on explaining variation in each variable.

[2] Standardisation is automatically achieved by carrying out PCA on the *correlation matrix* of the original variables I_j. PCA on the *covariance matrix* amounts to using unstandardised variables.

The only exceptions, in which it is not advisable to standardise, are those where the scales of all the indicators are already substantively comparable. This means that they must have the same units in some substantively relevant sense. In these instances, differences in the variances of the indicators are substantively meaningful and it makes sense to "put more weight" on the more variable indicators.

11.1.2 How many components?

When you calculate the principal components for a data set with p variables, you will recover p principal components, in descending order of variance. Put together, they describe all the variation in the data set. But because they are in descending order of variance, the first component provides the best "single number" summary that is possible, the first two components the best "two number" summary, and so on. One common question in applications is how many components are *enough* to describe the importance variation in a data set?

- Absolute criteria: components that explain some threshold of the total variation
- Relative criteria: components that have eigenvalues λ_k of at least some threshold
- Relative criteria: components that are upwards outliers in terms of variance explained

In practice, the last of these is the most frequently used, because it focuses on the components that most efficiently explain variation in the data set. Often this is assessed visually, rather than by some strict criterion, using what is called a "screeplot". We will see examples of this in the applications discussed later in this chapter.

11.1.3 Unsupervised Measurement

PCA is based on mathematics that are very close to those we considered in the previous two chapters. Principal components are linear combinations of indicators just as the scales we developed using regression on training data and expertise were linear combinations of indicators. The difference is in how we selected the coefficients on the indicators. In Chapter 8, we used regression to learn the coefficients that best predicted a set of units where we had some pre-existing measures of the concept we wanted to measure. In Chapter 9 we considered approaches to setting the coefficients based on expertise: setting equal coefficients, setting coefficients based on theoretical/substantive arguments, or estimating them from using expert comparisons of different indicator profiles. In all of these cases, we provided *supervision* to the measurement problem to ensure that it measured what we wanted it to measure.

In this chapter, we have done something fundamentally different. We have asked the data which coefficients would best predict variation across all indicators in the data set. We might then look at the results to see if it looked like we measured what we wanted to measure, or if we had measured something unexpected. PCA is our first example of an *unsupervised* measurement method. It is *unsupervised* in the sense that we have not indicated to the data what it is that we want to measure, except indirectly through the choice of indicators. We still implicitly control what is likely to emerge from PCA through the choice of indicators that we include, but this is a very weak sort of supervision.

One consequence of this is that PCA has no idea which way is up and which way is down, with respect to any concept you might have been hoping to recover. If you look back at the way PCA is defined, the signs of the principal components are completely arbitrary:

$$\begin{aligned} m_{i1} &= a_{11}I_{i1} + a_{21}I_{i2} + \ldots + a_{p1}I_{ip} \\ m_{i2} &= a_{12}I_{i1} + a_{22}I_{i2} + \ldots + a_{p2}I_{ip} \\ &\vdots \\ m_{ip} &= a_{1p}I_{i1} + a_{2p}I_{i2} + \ldots + a_{pp}I_{ip} \end{aligned}$$

If you multiply a_{11}, a_{21}, …, and a_{p1} by -1, the sign of the first principal component m_{i1} will flip for all observations i, but nothing else will happen. All the other principal components stay the same, the variances still add up to the right total, and the principal components are still uncorrelated with one another. This means that you can always choose whichever sign is easier to talk about. It also means that which sign comes out of the computer when you calculate the principal components tells you nothing.

11.2 Exploratory Factor Analysis (EFA)

Exploratory factor analysis reverses the underlying logic of how PCA imagines the relationship between scale and indicators. Recall in Chapter 7 when we talked about two different ways to measure the strength of teams in the Premier League. We showed that official way that the league table is calculated, adding up points based on the results, yields a scale for team performance that was very similar to what we got from a Bradley-Terry model. The points based calculation derives a scale by tallying up points for wins, draws and losses to get a score. The Bradley-Terry model involved hypothesizing that there was a latent dimension of team strength, and that the probability of wins, draws and losses depended on the difference in strength between the two sides. The points strategy defined the scale as arising from the indicators; the latent variable (Bradley-Terry) model hypothesized that the indicators had arisen from the scale. Similarly, Principal Component Analysis derives scales as linear combinations of observed indicators while Exploratory Factor Analysis hypothesizes latent dimensions, with the expected values of the observed indicators depending linearly on where units are on those dimensions.

11.2.1 Mathematical Details

We will describe a factor analysis model where each observation i has q latent factors $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iq})$

We will assume that these latent factors k have a multivariate normal distribution:

$$\boldsymbol{\theta}_i \sim N(\boldsymbol{\kappa}, \boldsymbol{\Phi})$$

with mean vector $\boldsymbol{\kappa}$ ($E_i(\theta_j) = \kappa_j$ for $j = 1, \ldots, q$) and covariance matrix $\boldsymbol{\Phi}$ (variances $var_i(\theta_j) = \phi_{jj}$ for $j = 1, \ldots, q$ and covariances $cov_i(\theta_j, \theta_k) = \phi_{jj'}$ for $j, j' = 1, \ldots, q;\ j \neq j'$)

We then assume that the observed items/indicators I_{ij} (for each observation i, for each indicator from $j = 1, \ldots, p$) are related to the latent factors θ_{ik}:

$$I_{ij} = \alpha_j + \beta_{j1}\theta_{i1} + \beta_{j2}\theta_{i2} + \cdots + \beta_{jq}\theta_{iq} + \epsilon_{ij} \tag{11.2}$$

\end{equation}

Note that it is easy to get your is, js, ks, ps and qs confused here, and there are a lot of Greek letters to keep track of.[3] Consistent with our notation for PCA, we have p indicators. The number of latent factors q will now be smaller than p (in PCA, we had the same number of principal components as indicators) thus the need for a new letter. The index i refers to different units/individuals/observations, the index j refers to different indicators for those units, the index k refers to different latent factors. The θ_i are the factors—an attribute of the units i—and the α_j and β_j are parameters describing how particular indicators j relate to those factors.

[3] Notation for these models varies across reference materials. My use of I here for the observable data is idiosyncratic—it is more typical to use x or y—but maintains continuity both with our previous notation in this book and with the way we were thinking about these data as *indicators* of a target concept rather than outcomes y or explanatory variables x.

As you can see above, the model for each indicator value I_{ij} is a linear model depending on q factors. This looks a lot like PCA, but backwards. Now the indicators are a linear function of the factors instead of the principal components being a linear function of the indicators.

One other thing that is different is that there is now an "error term" ϵ_{ij} for each unit for each indicator lurking at the end. There was no error term in PCA because we had the same number of principal components as indicators, and so we could fully describe the data with the full set of components. Here, we will no longer estimate enough latent factors to perfectly predict all observations in the data, and so there is a residual. We assume that $\epsilon_{ij} \sim N(0, \sigma_j)$ for $i = 1, \ldots, p$, and that these error terms are uncorrelated with each other, and uncorrelated with all the latent factors θ as well.

Factor analysis models are estimated in much the same way as regression models. The details of estimation do not concern us here.[4] The principal is that we find the values of the latent factors $\boldsymbol{\theta}$ and the loadings $\boldsymbol{\beta}$ that make it most likely that we would have observed the data that we did in fact observe.

[4] For a more detailed treatment, see Chapter 7 of Bartholomew et al. (2008).

11.2.2 Properties

Because the items/indicators are assumed to be a function of all latent factors, when we estimate a factor analysis model, all items and all observations contribute to the scores for all factors. In influence of a given item/indicator j on the factor k is higher when the coefficient β_{jk} of that factor k on that item j is larger in magnitude. Note that in factor analysis models these coefficients are typically called "loadings".

The maximum number of factors (q) must be less than the number of indicators (p), and most factor analyses focus on one or two factors. The numerical values of the factors themselves θ are determined by further **identification assumptions** because otherwise the scale of the model is not well defined by the model specified thus far. First, we need to specify the scales of the factors. The latent factors could just as easily run from -1 to 1 or -100 to 100 or 0 to 10, they have no natural units. The most common convention is to set $\kappa_k = 0$ and $\phi_{kk} = 1$ for $k = 1, \ldots, q$ so that all factors are standardised to have mean 0 and variance 1.

However even this scale of latent factors does not fully resolve their values. Suppose we start with 2 factors θ_{i1} and θ_{i2}, and then transform them to 2 new factors with the linear combinations

$$\begin{aligned} \theta^*_{i1} &= r_{11}\theta_{i1} + r_{12}\theta_{i2} \\ \theta^*_{i2} &= r_{21}\theta_{i1} + r_{22}\theta_{i2} \end{aligned}$$

with some coefficients $r_{11}, r_{12}, r_{21}, r_{22}$. This transformation can be interpreted as a *rotation*, a change of coordinate axes in the space of the factors. A rotation changes the coefficients/loadings of the factors, and thus also the interpretation of the factors. A rotation also changes the correlations of the factors and could

make them entirely uncorrelated. Almost all rotations, with correspondingly changed loadings, can produce exactly the same model for the observed items. As a result, we can freely choose which rotation to use, but we also need to remember that the choice is arbitrary when we pick one.

Factors are easiest to interpret when their loadings have a simple structure where each factor has large magnitude loadings for some variables and small (near 0) loadings for all the rest. With two or more factors, there are infinitely many equivalent rotated solutions. With one factor, there are just two, which are mirror image reflections associated with multiplying the factors by -1.[5] If the initial solution your software finds is difficult to interpret, potentially you can find a more easily interpretable rotation of the factor space.

[5] Note the same issue here as with PCA. Neither PCA nor factor analysis can meaningfully determine which way is up and which is down with respect to the concept you want to measure.

11.3 Application: Scaling Political Attitudes with Principal Components Analysis

The face to face survey component of the 2017 British Election Study was conducted after the election, and included (among many other items) the following battery of questions about respondents' political attitudes. Most of these questions have been asked on BES surveys for decades, although a few of them are more recent additions.

- I1. Ordinary working people get their fair share of the nation's wealth
- I2. There is one law for the rich and one for the poor
- I3. Young people today don't have enough respect for traditional British values
- I4. Censorship of films and magazines is necessary to uphold moral standards
- I5. There is no need for strong trade unions to protect employees' working conditions and wages
- I6. Private enterprise is the best way to solve Britain's economic problems
- I7. Major public services and industries ought to be in state ownership
- I8. It is the government's responsibility to provide a job for everyone who wants one
- I9. People should be allowed to organise public meetings to protest against the government
- I10. People in Britain should be more tolerant of those who lead unconventional lives
- I11. For some crimes, the death penalty is the most appropriate sentence
- I12. People who break the law should be given stiffer sentences

Imagine that we want to measure something about the *political ideology* of respondents, and these questions are what we have to work with. In the previous chapter, we talked about specifying coefficients/weights for particular items. That would be difficult here because we did not design this battery of questions to measure anything in particular, and they vary substantially in

terms of topic. It is certainly not clear what the relative weights ought to be if we wanted to aggregate responses in the ways we talked about last time. We might be able to specify the *sign* of some of the coefficients/weights—based on our expectations about which positions were on the political left and which were on the political right—but it would be very difficult to determine the relative magnitudes of the coefficients. Equal weighting, in this case setting coefficients to either −1 or 1, seems inappropriate as well because it would treat all these items as equally indicative of political ideology in the UK, which seems unlikely.

So instead, it makes sense to "just ask the data" how the indicators relate to one another. Which responses tend to go together in the data that we have? Does this structure reveal anything useful about what concepts we might use these data to measure?

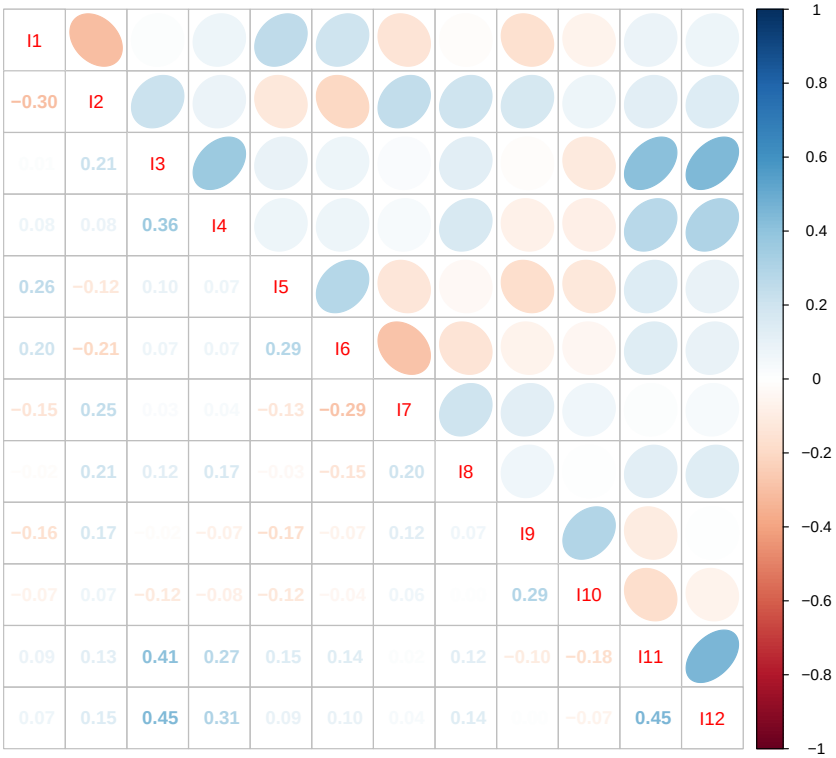


Figure 11.1: Pairwise correlations between items on the BES 2017 12 question ideology battery.

None of these items are particularly strongly correlated with one another. The strongest correlations are $r = 0.45$ between I11 and I12 and between I3 and I12. If we look at the cross-tabulation of responses to I11 and I12, we see that responses to these questions, both of which involve punitiveness of the criminal justice system, are still only moderately associated with one another.

If you look at the correlation plot carefully, you will start to notice that even though none of the correlations are especially strong, there are some indicators that seem to go together. For example, responses to I3 ("Young people today

	I12: SD	I12: D	I12: N	I12: A	I12: SA
I11: SD	35	109	155	82	47
I11: D	7	62	118	177	41
I11: N	3	14	80	128	39
I11: A	3	24	122	397	113
I11: SA	5	10	56	119	248

Table 11.1: Cross-tabulation of responses to questions I11 and I12.

don't have enough respect for traditional British values") are most positively correlated with those to I4 ("Censorship of films and magazines is necessary to uphold moral standards"), I11 ("For some crimes, the death penalty is the most appropriate sentence") and I12 ("People who break the law should be given stiffer sentences"). Those questions are also all positively correlated with one another. None of them are as highly correlated, positively or negatively, with any of the other questions. This kind of correlation pattern is indicative of a data set where there is likely to be at least one strong principal component. There is a set of questions which all seem to go together, which can therefore be described relatively well using a single common component.

We now turn to applying principal components analysis to this data set. The major pre-analysis decision for using PCA is whether to standardise the indicators. The answer is usually yes, but in this case we will not do so. Unlike many applications of PCA, here the items are already on a common scale: the strongly disagree to strongly agree scale. All the questions were asked in a bloc together. This means that the relative variance of responses on different items is probably telling us something real about the extent of variation in respondents' attitudes on the different items. If, on one of the items, everyone indicated that they agreed or strongly agreed, there is little variance in that item because people mostly share the same views. If we do not standardise, PCA will put little weight on explaining this slight variation in peoples' views. If we do standardise, PCA will put as much weight on explaining the variation in this variable as others, but that means it will put just as much weight on explaining these slight differences as the larger differences on other variables. In this instance the results of PCA look very similar regardless of whether we standarise or not, but the argument on the merits is against standardisation and so that is how I will proceed.

As you can see in Figure 11.2, the first two principal components have substantially higher variance than any of the others. Note, however, that they still only explain 23 and 17% of the variation in the original data, respectively, leaving 60 "unexplained". As with R^2 statistics, it is difficult to put a strict criteria on how much variation explained is a lot and how much is a little. With survey response scales like these, there is often a great deal of idiosyncratic variation in how individuals use the response scales, and so it is difficult to ever explain a great deal of the variation. Thus the major conclusion we can draw from the screeplot is the relative explanatory power of the different factors. We can fo-

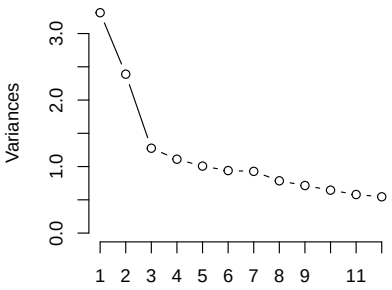


Figure 11.2: Screeplot for PCA on 12 BES questions.

PC1	PC2	Prompt
0.08	-0.39	Ordinary working people get their fair share of the nation's wealth
0.16	0.49	There is one law for the rich and one for the poor
0.44	0.06	Young people today don't have enough respect for traditional British values
0.35	0.02	Censorship of films and magazines is necessary to uphold moral standards
0.14	-0.34	There is no need for strong trade unions to protect employees' working conditions and wages
0.10	-0.38	Private enterprise is the best way to solve Britain's economic problems
0.04	0.42	Major public services and industries ought to be in state ownership
0.18	0.32	It is the government's responsibility to provide a job for everyone who wants one
-0.06	0.21	People should be allowed to organise public meetings to protest against the government
-0.11	0.13	People in Britain should be more tolerant of those who lead unconventional lives
0.67	-0.07	For some crimes, the death penalty is the most appropriate sentence
0.37	0.02	People who break the law should be given stiffer sentences

Table 11.2: Coefficients for first two principal components on 12 BES questions.

cus our attention on the content of the first two principal components in this instance.

11.3.1 Looking at the Coefficients/Loadings

Can we make sense of the first two principal components? Table 11.2 shows the coefficients for the first two principal components for each of the 12 disagree-agree items, alongside the prompts for those items. There are two things to look at for each coefficient: sign and magnitude. The sign tells you whether the principal component is positively or negatively correlated with responses to the item. The magnitude tells you whether the principal component is strongly or weakly correlated with responses to the item. Let's start by looking at the magnitudes.

Four of the items have larger magnitude coefficients on principal component 1 (PC1) than principal component 2 (PC2), while eight have larger magnitude coefficients on PC2 than PC1. The four with stronger *loadings* (larger coefficient magnitudes) on PC1 are the four we already noted were positively correlated with one another: "Young people today don't have enough respect for traditional British values", "Censorship of films and magazines is necessary to uphold moral standards", "For some crimes, the death penalty is the most appropriate sentence" and "People who break the law should be given stiffer sentences". The eight with stronger loadings on PC2 involve jobs, inequality, the economy, privatisation, and economic redistribution. Of these, there are two items, about public meetings and toleration, that load relatively weakly on both dimensions but slightly more strongly on PC2.

Can we put labels on these dimensions? Very roughly, it seems that PC1 involves traditional/authoritarian values/questions, while PC2 involves economic values/questions. We can see this more clearly by examining the signs of the coefficients. PC1 is strongly **positively** correlated with (agreement with)

"Young people today don't have enough respect for traditional British values", "Censorship of films and magazines is necessary to uphold moral standards", "For some crimes, the death penalty is the most appropriate sentence" and "People who break the law should be given stiffer sentences". These all are in the direction of tradition and authority. PC2 is strongly **positively** associated with "There is one law for the rich and one for the poor", "Major public services and industries ought to be in state ownership", "It is the government's responsibility to provide a job for everyone who wants one", which are all sentiments that we might think of as on the political **left**, rather than the political **right**. PC2 is strongly **negatively** associated with "Ordinary working people get their fair share of the nation's wealth", "There is no need for strong trade unions to protect employees' working conditions and wages", and "Private enterprise is the best way to solve Britain's economic problems", which are sentiments associated with the political **right**. Thus, to be more precise, positive scores on PC1 are more traditional/authoritarian, negative scores are less so. Positive scores on PC2 are more economically left-wing, negative scores are less so.

11.3.2 Looking at the Units/Observations

Remember our original statement of what the principal components were: they are **linear combinations** of the original variables:

$$\begin{aligned} m_{i1} &= a_{11}I_{i1} + a_{21}I_{i2} + \ldots + a_{p1}I_{ip} \\ m_{i2} &= a_{12}I_{i1} + a_{22}I_{i2} + \ldots + a_{p2}I_{ip} \\ &\vdots \\ m_{ip} &= a_{1p}I_{i1} + a_{2p}I_{i2} + \ldots + a_{pp}I_{ip} \end{aligned}$$

Thus, given the coefficients a and the observed responses in the data, we can construct the principal component values or "scores" for every observation in the data set (of which there are 2194). This is the measurement part of the exercise: we have a way of scoring individual respondents to this survey on what we now think are a traditionalism/authoritarianism scale as well as an economic (right-)left scale.

How can we assess whether these scores are sensible? We do not know the BES respondents, so doing face validity checks on individual respondents is not going to work. Instead, we might look whether the scores are associated with other relevant features of individuals in ways that we would expect. This is an assessment of *correlational* or *predictive* validity.

The first predictive validity check we can do uses another 2017 BES question which asked "In politics people sometimes talk of left and right. Where would you place yourself on the following scale?" Figure 11.3 shows that responses to this question are associated with both of the first two principal components:

positively with the traditionalism/authoritarianism PC1 and negatively with the economic (right-)left PC2. The association is somewhat stronger with PC2. These correlations are not strong overall, which tends to be the case with survey responses on questions like this because many people have only a vague idea of political left and right and many people hold heterodox combinations of political attitudes.

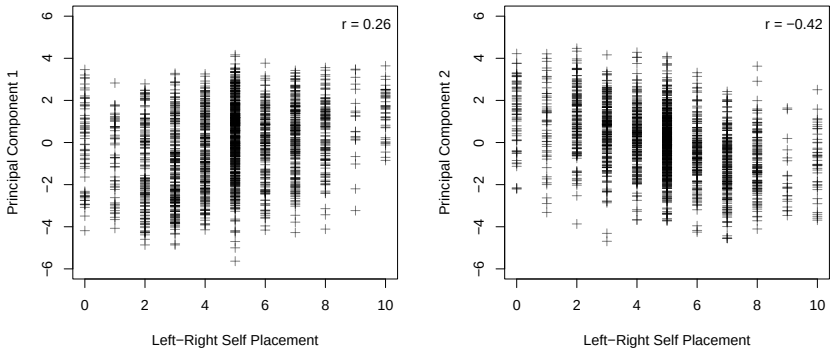


Figure 11.3: Principal Components as a function of Left-Right self-placement.

Another set of variables which we could use to assess predictive validity are the political choices made by these respondents. How are the principal components related to respondents' votes in the 2017 UK general election immediately preceding the survey? How are the principal components related to respondents' votes in the 2016 UK referendum on membership in the EU?

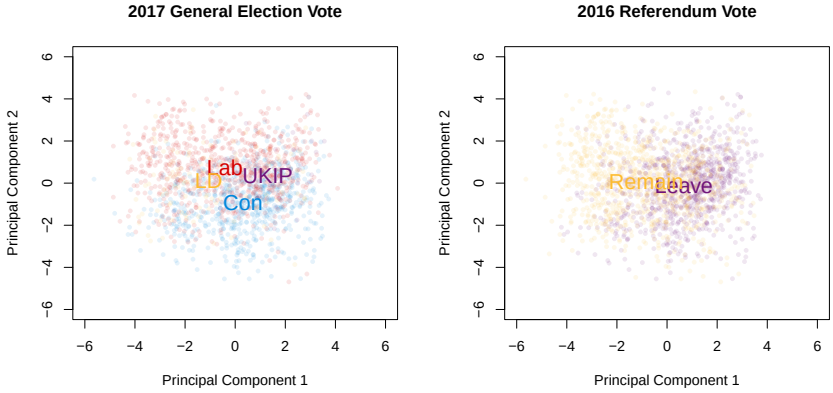


Figure 11.4: Principal Components as a function of 2017 and 2016 votes.

The left panel of Figure 11.4 shows the average position of 2017 voters for the Conservatives (Con), Labour (Lab), the Liberal Democratis (LD) and the UK Independence Party (UKIP) in the 2D space defined by the first two principal components. 2017 Conservative voters and 2017 Labour voters differ primarily with respect to Principal Component 2. We previously established that the coefficients indicate that higher scores on PC2 are associated with economically left attitudes, and we indeed see Labour voters scoring higher than Conservative voters in this "economic leftism" score. The difference on PC1 between

Labour and Conservative voters is relatively slight; the bigger contrast is between Liberal Democrat and UKIP voters who more clearly differentiate on PC1 than on PC2.

The right panel of Figure 11.4 shows the average position of 2016 voters for Leave and Remain in the same 2D space defined by the first two principal components. 2016 Leave and Remain voters are almost identical on PC2 (the economic leftism dimension), but differ to a greater degree on PC1 (the traditionalism/authoritarianism dimension).

These relationships all make sense directionally, but none are especially strong. Correlational or predictive validity always provides a weak test. Lots of things are correlated with lots of other things. If you see a moderate correlation between vote choice and your measure of someone's right-left or left-right position, it could mean that vote is actually only weakly related to that concept, or it might just mean that you did not measure it very well. In this case, we effectively guessed what concept we had measured after we did the data analysis. It is possible that the measures we have are a mixture of what we guessed and something else, or that we have mis-labeled the patterns that PCA revealed.

11.3.3 Consequences of Indicator Selection

There are many, many questions on the 2017 British Election study besides the ones I included in this example (which come from one bloc of questions in the middle of the survey). Imagine that we are only interested in economic left-right preferences of voters. We might have simply looked through the list of survey prompts, and only included 1, 2, 5, 6, 7 & 8, which are the ones that clearly have something to do with the economic organization of society. Those turned out to be the indicators that loaded most heavily on PC2 in our analysis above, but if we had only wanted to measure something specific about economic attitudes, it would have made sense to limit the analysis to those items only. We could then have either used some of the techniques from last chapter on those items (such as simple equal weighting, inverting some of the items as appropriate) or we could have applied PCA to that more limited set of items.

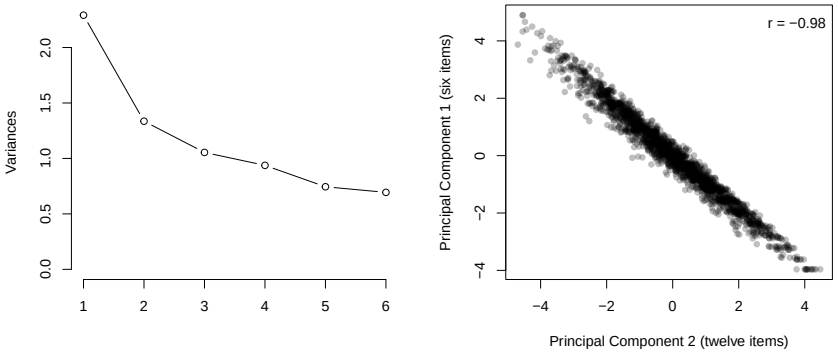


Figure 11.5: Screeplot for PCA on the six economy-related items (left). PC2 from twelve item PCA versus PC1 from PCA on the six economy-related items (right)

If we apply PCA to just those six items, we end up with a single strong principal component rather than two. That first principal component is highly correlated with our old PC2 from the initial analysis of all 12 items that we have been looking at up until now. Mathematically this is not a surprise, we have in fact selected the items that had largest magnitude coefficients on PC2, and excluded the items that had largest magnitude coefficients on PC1. Since PCA just wants to explain variance, the old PC2 becomes the new PC1. They differ slightly because the loadings for the omitted items were not exactly zero, and the loadings on the included items change a bit, but in this case the correlation between our old measure of economic right-left sentiment based on twelve items and our new one based on six items is very strong.

You will notice that the direction of the scale flips. Again, it is important to emphasise that this is entirely arbitrary. At no point did we indicate whether we wanted a left-right scale or a right-left scale. As discussed earlier, the computer just picked a direction arbitrarily. So if you do not like how we initially had PC2 with the political left as positive values and the political right as negative values, you could just multiply all the a_{j2} by -1 to have left as negative values and right as positive values.

11.4 Application: Scaling Political Attitudes with Factor Analysis

What happens if we use factor analysis instead of principal components analysis on the 12 BES ideology items? Table 11.3 reports factor loadings from R's default `factanal()` implementation of factor analysis, which uses a "varimax" rotation which tends to yield results similar to PCA. The patterns of loadings look broadly similar to the patterns of PCA coefficients we saw earlier on the same data. There are some differences; while there are again the same four issues that "load strongly" on factor 1 and five issues that load strongly on factor 2, the item about government responsibility for offering jobs now loads equally on both dimensions. But the differences from what we find with PCA are not large, and the results of using factor analysis tell the same substantive story about the correlations in the data. People tend to give correlated responses to the four traditionalism/authoritarianism items and they tend to give correlated responses to the five/six economic items.

If we look at the *factor scores*, the θ_is for each survey respondent i, we see that they are very highly correlated with the PCA scores. These two methods are measuring almost the same thing in this application. This is often true, although not always.

It is worth remembering that PCA and EFA are more conceptually different than practically different. PCA summarises variance in the indicators as efficiently as possible in terms of components that are linear functions of the indicators. FA is an effort to identify the latent factors that would have been most likely to generate the indicators, if in fact the indicators were generated by latent factors according to a specified linear model. This means that they are

F1	F2	Prompt
0.06	-0.47	Ordinary working people get their fair share of the nation's wealth
0.29	0.53	There is one law for the rich and one for the poor
0.67	0.01	Young people today don't have enough respect for traditional British values
0.48	-0.05	Censorship of films and magazines is necessary to uphold moral standards
0.15	-0.42	There is no need for strong trade unions to protect employees' working conditions and wages
0.10	-0.49	Private enterprise is the best way to solve Britain's economic problems
0.09	0.43	Major public services and industries ought to be in state ownership
0.24	0.24	It is the government's responsibility to provide a job for everyone who wants one
-0.05	0.34	People should be allowed to organise public meetings to protest against the government
-0.16	0.22	People in Britain should be more tolerant of those who lead unconventional lives
0.64	-0.13	For some crimes, the death penalty is the most appropriate sentence
0.66	-0.04	People who break the law should be given stiffer sentences

Table 11.3: Factor loadings for two factor model on 12 BES questions.

	PC1	PC2	F1	F2
PC1	1.00	0.00	0.97	-0.13
PC2	0.00	1.00	0.11	0.98
F1	0.97	0.11	1.00	-0.01
F2	-0.13	0.98	-0.01	1.00

Table 11.4: Pairwise correlations of respondent scores from Principal Components Analysis (first two components) and Exploratory Factor Analysis (two factor model, varimax rotation).

both trying to provide a "simple" summary of the correlations across variables in the data. As a consequence, when they are applied to the same data, they do tend to yield similar conclusions (at least if you pick the right factor rotation).

11.5 What are we measuring?

As noted earlier in this chapter, we have seen this kind of practical similarity between conceptually distinct methods before. Tallying up win/points in competitions (a transformation of the data like principal components analysis) gives results that are similar to fitting a Bradley-Terry model (a latent variable model like factor analysis). Before this course, you implicitly have seen this in linear regression, which can be motivated in two ways. One way is as a simple "summary" transformation of the data: the linear projection of the observed y onto the explanatory variables x that minimises the sum of square errors. The other way to motivate a linear regression is on the basis of a "generative" model for y as a linear function of x with normally distributed errors. For linear regression these two motivations give numerically identical answers; in the cases we have looked at in this course the analogous situation has yielded approximately the same answers from different methods.

More broadly, we often have a choice between an approach based on the logic of summarising a(n indicator) data set in a simple way and an approach based on estimated parameters from a model that could have generated the

observed data. When we looked at Bradley-Terry models, we already saw some of the advantages and disadvantages of the "summary measures" versus "generative measures" approach to description. Summary approaches tend to be simpler to implement, faster to compute, and clearly limited in their interpretation. Generative approaches tend to be more flexible and to provide a direct way of describing uncertainty about measures, while also being more demanding to think about, more computationally demanding, and also to risk over-interpretation.

The risk of over-interpretation is one that I highlighted back in the discussion of Bradley-Terry models, but a significant part of ugly history of the misuse of social measurement that I described in Chapter 1 involves factor analysis methods specifically. Because factor analysis models are structured around the idea that the indicators are produced by the latent factor, they often tempt people to make causal claims that would be difficult to justify if they thought carefully about the problem. Again, think back to linear regression here. If you "believe" in the linear generative model, it looks like a causal model for how y is generated by x. Surely that means that if you change x_j by 1 unit, it will change y by β_j, right? This is almost never justified by running a regression unless x was randomly assigned, but people make this mistake all the time.[6] It is far safer to view the regression model as a summary of the variation in y conditional on x, without causal assumptions. The same point holds here: you do not need to "believe" in the factor to do factor analysis and for it to be useful as a summary of the data. Nonetheless, it is almost always a mistake to adopt a causal interpretation of the factors.[7] The factors do not really exist just because you happened to fit a model with factors in it. Do not *reify* the factors!

What would reifying the factors mean in this context? It would mean believing that people are walking around with an underlying degree of traditionalism/authoritarianism and a degree of economic leftism *because* those are the labels that seem to describe the output of the factor analysis or principal component analysis. This does not follow, indeed it is wrong for the same reasons that a non-zero β_j coefficient in a linear regression model does not imply a causal effect of x_j on y. In the regression context, all you have demonstrated is that there is a partial association of x_j with y given a model with some set of other x variables in a given set of units. That may be because changes in x_j cause linear changes in y, given levels of those over variables, but it could also be for a variety of other reasons having to do with omitted $x_{j'}$ that have causal effects on y and which are correlated with x_j, causal effects of y on x_j, sample selection mechanisms that depend on x_j and y, and further more complicated mechanisms.

In the factor analysis context, the fact that a set of variables are correlated with one another does not mean that they are all the product of a single common latent factor. They may have causal relationships with one another, or they may be the product of many latent influences. Leaping to completely unjustified causal inferences is easily the most common mistake that people make

[6] Please put down this book and go read a modern causal inference textbook like Morgan and Winship (2015) if you do not know why such a claim is unjustified.

[7] If you have a credible way to argue that the factors—the latent variables—were randomly assigned to units, by all means go ahead and make the causal interpretation. I have never seen an application where this was plausible, but perhaps one exists somewhere.

in interpreting factor analysis models.

11.6 The Thomson critique

To see that the fact that a set of variables are positively correlated with one another does not mean that they are all the product of a single common latent factor, it is helpful to see an example where such patterns arise in a different way. In 1916, Godfrey Thomson wrote two papers offering a critique of factor analysis, and specifically Spearman's claims that certain patterns of correlations between items/indicators implied the existence of a "General Factor" in intelligence tests. In "A Hierarchy Without a General Factor", Thomson (1916) writes "The object of this paper is to show that the cases brought forwards by Professor Spearman in favour of the existence of General Ability are by no means 'crucial'. They are it is true not inconsistent with the existence of such a common element but neither are they inconsistent with its non-existence." The argument is a general one, and has nothing in particular to do with the intelligence testing case, and so I will illustrate the core theoretical point here using a different application.[8]

[8] This discussion was inspired by a blog post "g, a Statistical Myth" by Cosma Shalizi

Let us carry on with an example of political ideology like those considered earlier in this chapter. Let us hypothesize that, contra much writing about politics, there is no single factor that governs the left-right political ideology of individuals. Rather, let us imagine that when thinking about the structure of the economy and the role of government, people actually have a variety of intuitions about questions that are relevant to many individual items. For purposes of our example, let's say that there are six of these "intuitions" that commonly influence how people assess individual policy questions related to the economy and redistribution:

1. To what extent we are not worried by inequality of outcomes.
2. To what extent we should worry about moral hazard or the "Samaritan's dilemma".
3. To what extent we think people are motivated by personal gains as opposed to the gains of others.
4. To what extent we think people have meaningful equality of opportunity in our society right now.
5. To what extent we think that the most important goods are private rather than public goods.
6. To what extent we value individual's freedom of economic association and activity.

These are a mixture of normative commitments and factual beliefs, but they all might contribute to one's views about concrete economic policy questions. I have written them all so that positive corresponds to views that might motivate more right-leaning economic policy views and negative corresponds to views that might motivate more left-leaning economic policy views. But any given

economic policy question might only implicate a subset of these. For example, if we are thinking about the benefits of increasing baseline welfare benefits, for example via a universal basic income, this would heavily implicate intuitions 1, 2, and 4, but perhaps have little direct relationship to 3, 5 and 6. If we were evaluating a general tax increase to invest in public transit infrastructure, this might primarily implicate 2, 4, 5, and 6. If instead, we were thinking about instituting a carbon tax to address climate change, this might only implicate 5 and 6.[9]

[9] You could easily quibble with the details of these examples, but exactly which intuitions are relevant to which policies is not really the point. The important thing is that different policies might be related to different subsets of these intuitions.

Let's imagine that in the general population, these six intuitions are independently distributed. That is to say, they are uncorrelated with one another. For purposes of doing some illustrative simulations, let's say they are independently and identically distributed standard normal $N(0, 1)$.

Now imagine that you watch politics for a while, and a sequence of policy questions are raised and a subset of this population is polled regarding their views on each of these policy questions. An individual i's views on a given policy question j are given by a linear function of the relevant subset of their intuitions k for that policy, plus an idiosyncratic term, as follows. Where $\gamma_{kj} \in 0, 1$ is the irrelevance/relevance of intuition k to policy j, and $\eta_{ik} \sim N(0, 1)$ is individual i's intuition k, their view[10] I_{ij} on policy j is:

$$I_{ij} = \gamma_{1j}\eta_{i1} + \gamma_{2j}\eta_{i2} + \gamma_{3j}\eta_{i3} + \gamma_{4j}\eta_{i4} + \gamma_{5j}\eta_{i5} + \gamma_{6j}\eta_{i6} + \epsilon_{ij} \tag{11.3}$$

[10] This is the same kind of observable "indicator" or "item", the measured view of respondent i on policy j, that we had in our earlier examples in this chapter, thus the same notation I_{ij} I have used previously. What is different is the process that we are hypothesizing might have generated what we observe.

Finally, to enable us to simulate data, let's assume that each policy j implicates underlying intuitions independently and with equal probabilities of 1/2. So γ_{kj} are independently and identically Bernoulli distributed with parameter 1/2, ie coin flips. I simulate 1000 individuals from this population, expressing views across 400 issues. I then plot the screeplot and I get Figure 11.6, which very clearly indicates that there is one very strong principle component, five much less strong principle components, and very little beyond that.

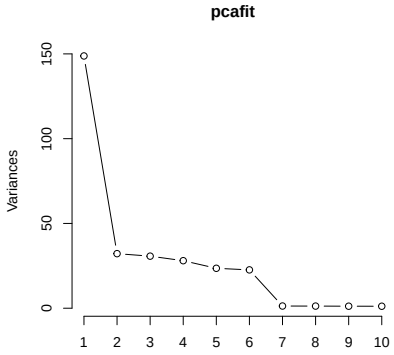


Figure 11.6: Screeplot for PCA on simulated data.

The fact that there are six components is correct, given how we simulated the data, but the fact that there is one very strong principle component is surprising and potentially misleading. The simulated data were formed from six "economic intuitions", each of which had the same average contribution to the observed data across the items and individuals in the data. Why does PCA tell us there the first principle component is massively more important than any of the following five? To understand this, it is helpful to look at the correlations between the first principle component that we recovered and the six intuitions that we used to generate the observable data.

Figure 11.7 shows that, aside from a bit of random simulation noise, the six intuitions are indeed uncorrelated with one another at the individual level. However, each of them is correlated with the first principle component at correlations of about 0.4 (again, with some variability due to the simulation). The first principle component is almost exactly equal to the average of the six intuitions $\bar{\eta}$. Principle components analysis simply aims to explain as much variation as possible with the first component, and then with each additional

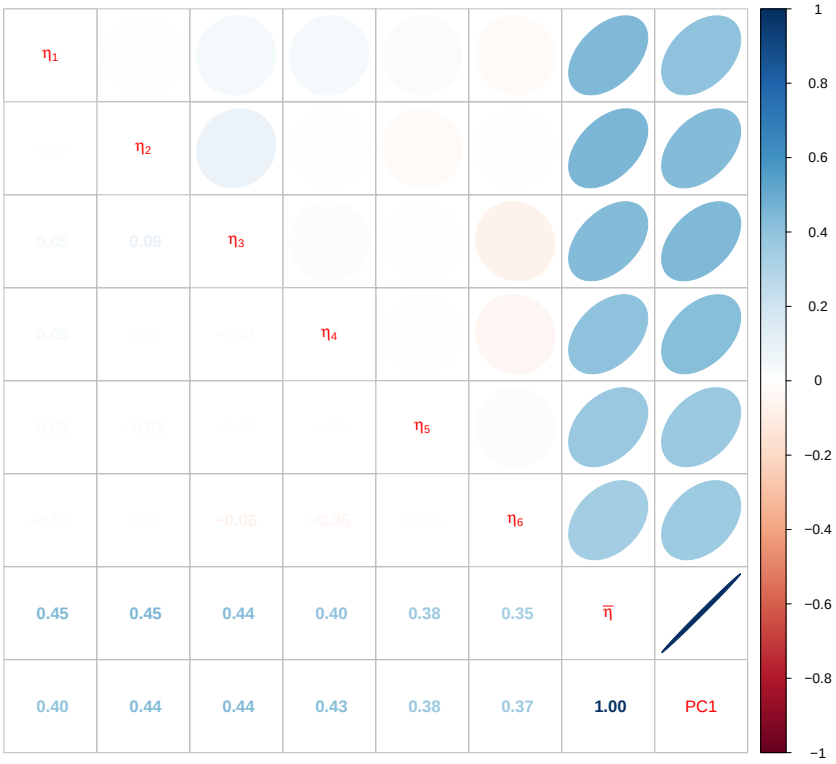


Figure 11.7: Association of each of the six economic intuitions with the average intuition and the first principle component (PC1)

factor given the previous ones. The individual-level average of the six intuitions is the best single number predictor of an individual's response given the way we have set up the problem, and it is indeed a far better predictor than any further single number predictor that you can add to it as a second (or further) principle component.[11]

What have we learned? In this example, principle components analysis will discover a strong first principle component even where, by construction, we know that there were six equally important components of the data generating process. That is, in a world where the observable policy views of individuals were equally shaped by six different, equally important and uncorrelated economic intuitions, when we run principle components analysis we still see one very strong first principle component. Thus, the single most important lesson here is that having a strong first principle component *does not demonstrate* that there is a single strong factor shaping the process that generated the data at the individual level.

If you did not recognise this, and you looked at this analysis, you might be tempted to conclude that respondents "really are" characterised by having a left-right position, as we can construct such a position (PC1) that predicts a lot of the response variation. But that is clearly not the case given how we generated the data. We can indeed summarise a good fraction of the response variation by creating a summary that we call left-right position, but that is

[11] I have made all six intuitions have equal variance and equal average coefficients for I_{ij} for purposes of exposition, but the same basic finding would persist with the first principle component as a weighted average of the intuitions if these were instead unequal.

not actually what was happening in our simulated respondents' minds as they formed positions across different policy questions.

But perhaps this is just a pathology of principle components analysis? What if we examine the same problem using factor analysis, which more closely matches the data generating process for these data. Indeed, if you compare Equation 11.3 and Equation 11.2, you will see that our six intuition model is in fact a case of the general factor analysis model. The intuitions are the factors. Do we get the right answer in this instance?

The answer turns out to be almost yes, so long as we fit a factor analysis model with a sufficient number of factors and we have enough data. Figure 11.8 shows that in a factor analysis model fit with a single factor, that factor explains nearly 40% of the variance in the responses. If we only fit the one factor model, we will tend to spuriously conclude there is a single strong factor, as we did when we used principle components analysis. But when additional factors are added (here using a "varimax" rotation of the factor space), the tendency is to have factors that explain similar amounts of variation up until one arrives at a model with at least six factors, after which adding further factors has negligible consequences. While there is still some tendency to find that some factors are stronger than others, this is due to small sample variation that makes some factors more predictive in the sample by chance. This bias goes to zero with increasingly large data sets generated in the way I have described.

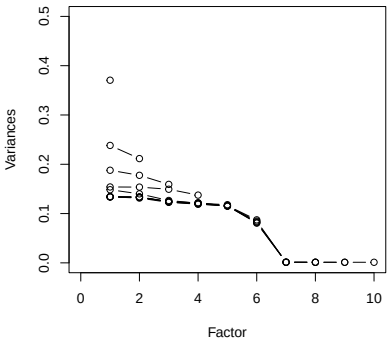


Figure 11.8: Variances explained by each factor for factor analysis models with different numbers of factors.

The single factor model is, however, spectacularly biased with respect to how much variation any single intuition actually explains, for much the same reason that the principle components analysis yields a very misleading screeplot. In this toy example, the single factor model estimates a factor that explains 37% of the variance in the response data, but in fact by construction each intuition only explains 1/7th or 14% of the variation in the population-level response data. It is really easy to convince yourself that there is a strong single factor generating the response data, even when there is not. In this instance adding more factors fixes this bias, but this is not always true.

Unfortunately this is not one of the more difficult cases for factor analysis to get the right answer, instead this is actually an unrealistically easy case. Nearly all of the variation in the data (6/7ths) is explained by a relatively small number of factors (6). We have a large number of data points (200) for each of a large number of individuals (1000). If, for example, we reduce the number of data points per individual to 20, our ability to identify that there are actually six, roughly equally important factors completely disappears. Figure 11.9 shows a typical case of what happens with what is still a fairly large number of responses per individual relative to many applications (there is a lot of random variation from sample to sample).

The tendency of factor analysis is to produce plots that look like 11.9, under a wide range of data generating processes and quantities of data, often with even steeper drops from the first to the second factor. Principle components

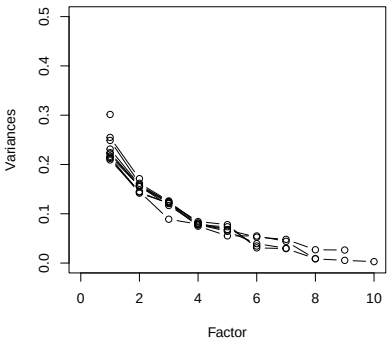


Figure 11.9: Variances explained by each factor for factor analysis models with different numbers of factors, with reduced number of responses per individual.

analysis is more or less guaranteed to generate plots that look like this. As a consequence, it is extremely important not to make inferences from the relative variance explained of principle components or factors to the number of "true" underlying factors that generated the data.

Thomson's critique observed that factor analysis tended to produce a strong first factor even when there were a relatively large number of underlying true factors that generated the data. His papers are partly focused with the mechanics of simulating such data using the technology available in 1916 (playing cards!). Sadly, this critique has been little appreciated by users of factor analysis in the century since it was made. The issue isn't specific to generating data with a large number of factors, there are many other underlying processes that will generate data that will generate a strong single factor model and first principle component.[12] But it is worth noting that the "many influences" model is both one justification for normal error terms via the central limit theorem and extremely plausible for any social/psychological process involving humans. We should expect lots of things to matter when we study complex systems, and so we need to be very careful when we apply methods that tend towards providing simple explanations not to reify those simple explanations. It may be convenient to summarise someone's views with a left-right position, but that does not mean that they really have a left-right position.

[12] "It has been known for almost as long as factor analysis has been around that positive correlations can arise in many ways which involve nothing remotely like a general factor… Thomson's ability-sampling model, with its myriad independent causes rather than a single general cause, is the oldest and most extreme counter-example, but it is far from the only one." *g, a Statistical Myth*, Cosma Shalizi

11.7 Conclusion

So, having made the point about what these techniques do not license you to say, what *are* they good for? One answer is that they are good for exploring data sets (it is called "exploratory" factor analysis, after all). We can use these methods to find *summary* measures that describe variation in a data set as simply as possible. You also might come at the question from the other direction. Why wouldn't you always do it this way? Why did we bother with the material in the previous chapter? Why ever use equal or expert-specified weights when you can "just" measure the weights from the data?

The fundamental limitation of principal components analysis and factor analysis is that we are measuring what explains the most variation in the data, not what best represents any concept we might be interested in. We have no direct control over what PCA/EFA are measuring. The dimensions/factors that explain variation best in a given set of indicators maybe the concept we wanted to measure. Or they may be a mixture of what we want to measure and something else. Or they may be other things entirely. We have indirect control over what the methods are measuring because we determined the set of indicators that went into the estimation. Nonetheless, this is a very imprecise sort of control over what is being measured.

In the preceding chapter we discussed an example where Floridi and Lauderdale (2018) used a conjoint experiment in which experts on the demographic concept of *productive aging* were asked whether hypothetical individuals who

engage in varying degrees of paid work, volunteering, grandchild care and care for sick/disabled adults are more or less productive than other such hypothetical individuals. The indicators in that example were the extent to which individuals did more or less of these activities. The desired concept, productive aging, is meant to be increasing in all these activities. The question is not whether some of these activities are productive or unproductive, but rather how much productivity to associate with the different levels of these activities.

But people have time constraints. Someone who does 40+ hours a week of paid work is going to have difficulty spending much time on the other activities, and vice versa. As a result, a lot of the variation in the data from that study is about *which* activities individuals participate in, not just how much they do *overall*. As a result, the first principal component for data on the various activities largely reflects whether individuals are in paid work versus the other kinds of activities (Floridi and Lauderdale, 2018). But the goal of the measurement exercise was to measure how productive they were overall, not to measure which types of activities people were productive in. The goal was *not* to find what explained the most variation in the data, it was to measure a particular concept of interest. This is a case where PCA or Factor Analysis give you the wrong answers because they ask the wrong question. They give you an efficient summary of variation in the data, but that summary may not be what you wanted to measure.

Both PCA and EFA can measure the wrong thing, at least if there was something in particular that you wanted to measure. Just because you want to measure a certain concept does not mean the thing that explains the most variation in your set of indicators will be that concept. This is certainly true if you did not design the indicators around that concept, and it can even be true if you did. Indeed, you might say it is vanishingly unlikely that it will be *exactly* the concept you want! Thus, the safe interpretation of PCA/EFA is that you have recovered dimension(s) that explain as much variation in the indicators in as simple a way as possible. Whether that is what you wanted to measure will depend a lot on whether you collected indicators for which the primary source of common variation was the desired concept, and not some other concept(s).

Finally, PCA and EFA present us with a first example of an issue we will have to face repeatedly in the coming chapters. How many components or factors or dimensions do we want? This is a persistently difficult question in unsupervised measurement models. In supervised models, where you set out to measure a certain thing or certain set of things, this is not an issue. But if you are asking the data for a simple measure or set of measures that explains as much of the variation as possible, determining how simple is a difficult threshold to set. The screeplot for PCA, and analogous measures of relative fit for EFA models with different numbers of factors, are statistical criteria. Statistical criteria tell you about how much variation in the data you can describe with different numbers of components/factors/dimensions. Adding components/factors/dimensions until the fit to the data stops improving very quickly

is a reasonable approach in many contexts, but it is important to keep in mind that it will not tell you whether the components/factors/dimensions that you have recovered are useful or interpretable for your purposes.

12

Unsupervised Scale Measurement with Categorical Indicators

Principal components analysis and exploratory factor analysis assume that indicator variables are measured at an interval-level. If this were not the case, assuming linear relationships between those variables and the principal components or latent factors would make little sense. In the example used in the last chapter, where the data were a scale from strongly disagree to strongly agree, this assumption was a stretch. It is not clear that respondents use such a scale in an interval-level way.

In this chapter, we will be looking at corresponding measurement methods for ordinal and nominal-level categorical response data. Thus, this chapter is the analogue of the move from linear regression to the various limited dependent variable regression models—binary logistic, ordinal logistic, etc—for factor models. In this chapter, we will focus on generative models for how the indicators might depend on latent variables. In addition to these unsupervised generative models, there are unsupervised summary methods for categorical data for example see correspondence analysis (Benzecri, 1973), but they lack the straightforward interpretation of PCA, and so we will not cover them here.[1]

[1] These methods are particularly useful in very large data sets where computation becomes a problem for methods based on generative models.

Item Response Theory is a term used to describe a broad class of factor-analysis-like models for limited dependent variables with many variants. We will focus on a couple of the most general and widely used item response models in this chapter. The canonical applications of IRT models relate to educational testing as well as indicator data that comes from *items* on questionnaires that are designed to measure attributes of survey respondents. The term *item* is simply the domain-specific version of the term "indicator" that we have been using throughout this book. Nonetheless, the models are far more general than these applications for which they were initially developed, and have been used widely across many social science fields.

Recall the linear factor model from the previous chapter. We described a factor analysis model where each observation i had q latent factors $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iq})$. Today, we are going to simplify a bit and just focus on models with a single latent factor θ_i for each observation.

Last time, for a one factor model, we assumed that the observed items/indicators I_{ij} (for each observation i, for each indicator from $j = 1, \ldots, p$) were related to the latent factors θ_i:

$$I_{ij} = \alpha_j + \beta_j \theta_i + \epsilon_{ij} \tag{12.1}$$

Recall again that the index i refers to different units/individuals/observations, the index j refers to different indicators for those units, and we have just one latent factor so we do not need a third index. As you can see above, given the value of the latent factor θ_i, the model for each indicator is just a simple linear model. Figure 12.1 shows examples of the implied relationship between $E\left[I_{ij}\right]$ and θ_i for $\alpha_j = 0$ and $\beta_j = 1$ (solid line) and for $\alpha_j = 1$ and $\beta_j = -0.5$ (dashed line).

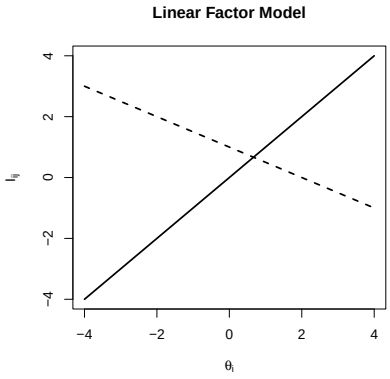


Figure 12.1: Linear item response (factor model) as a function of unit-level latent variable θ_i.

12.1 Binary Item Response Model

Assuming that the indicators will be linearly associated with the latent variable is sensible if the indicators are continuous variables, but often they are binary or categorical. In such cases, using factor analysis is analogous to applying linear regression as a "linear probability model" for a binary dependent variable. It may be a reasonable approximation, but it also can fit poorly and make invalid predictions of probabilities outside of the range between 0 and 1. Just as logistic regression and its variants were useful for better describing the likely associations between various explanatory variables and a binary/categorical dependent variable, item response models describe associations between the latent variable θ_i and the indicators I_{ij} in terms of these same logistic functional forms. We can simply take our factor model above, and replace the left hand side with the same log-odds formulation that generated a logistic regression from a linear model.

$$\log\left(\frac{p(I_{ij}=1)}{p(I_{ij}=0)}\right) = \alpha_j + \beta_j \theta_i \tag{12.2}$$

Figure 12.2 shows examples of the implied relationship between $E\left[I_{ij}\right]$ and θ_i for $\alpha_j = 0$ and $\beta_j = 1$ (solid line) and for $\alpha_j = 1$ and $\beta_j = -0.5$ (dashed line), the same parameter values as in the linear factor model depicted in Figure 12.1. The comparison illustrates the commonalities between the assumptions of factor models and item response models, as well as the difference in functional form.

By convention, these models are sometimes parametrised using the alternative form where α is the x-intercept rather than the y-intercept:

$$\log\left(\frac{p(I_{ij}=1)}{p(I_{ij}=0)}\right) = \beta_j\left(\theta_i - \alpha_j\right) \tag{12.3}$$

This form enables a more useful interpretation of α_j as the "difficulty parameter" for indicator j and β_j as the "discrimination parameter". Un-

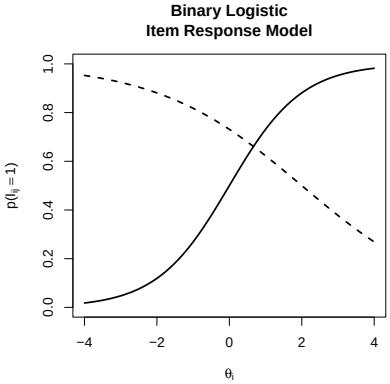


Figure 12.2: Binary logistic item response as a function of unit-level latent variable θ_i.

der this parameterization, α_j is the value of the the latent variable θ_i where $p\left(I_{ij} = 1\right) = 0.5$. Higher values of α then correspond to items with "higher difficulty", where higher values of the latent variable θ_i are required in order to make $I_{ij} = 1$ probable. Note that this interpretatation really only makes sense if all/most of the β_j are positive, which is to say that higher probabilities of $I_{ij} = 1$ are consistently associated with higher levels of the latent variable θ_i for all/most indicators j.

One application where this is true, and where the difficulty/discrimination language comes from, is standardized educational testing. One of the goals of standardized test *design* is to have test items (indicators) that cover a range of difficulties, but which all have high discrimination. That is, you want them all to test the same latent factor ("understanding of the material") but for some to be relatively easy (indicating a minimal level of understanding) and for others to be more difficult (indicating a higher level of understanding). You may also run across a simpler version of the item response model (the "Rasch model" or "one parameter logit model") which sets all $\beta_j = 1$. This assumes, a priori, equal responsiveness of all the indicators to the latent variable. There is little reason to apply this model unless you have only a small number of units. Assuming that all items respond to the latent scale equally is an assumption you can make, but it rarely makes sense to do so unless you have to. We will see below some examples where the values of the β_j are similar, but where we nonetheless learn something useful by comparing them rather than assuming they are all identical.

12.2 Ordinal Item Response Model

We can extend this model to ordinal categorical indicators in exactly the same way that binary logistic regression extends to ordinal logistic regression. This model is often called a "graded response model" (Samejima, 1969), referring to the educational testing origins of these techniques. Just as you can have a test of items that individuals get right ($I_{ij} = 1$) or wrong ($I_{ij} = 0$), you can also have "graded responses" for any number of ordered levels $I_{ij} = 1, 2, 3, \ldots$.

As with the ordinal logistic regression model, ordinal item response models or graded response models are based on a linear model for the log-odds of being above or below each of the thresholds in the ordered categorical variable. The model makes a *proportional odds assumption* that there is the same slope/discrimination parameter β_j but different intercepts α_{jk} for each threshold on a given item j:

$$\log\left(\frac{p(I_{ij} > k)}{p(I_{ij} \leq k)}\right) = \beta_j\left(\theta_i - \alpha_{jk}\right) \tag{12.4}$$

Figure 12.3 shows the implied item response curves for being above, as opposed to below, a given response level, as a function of the latent factor θ_i for two example items, one with four levels (and thus three curves) and one with

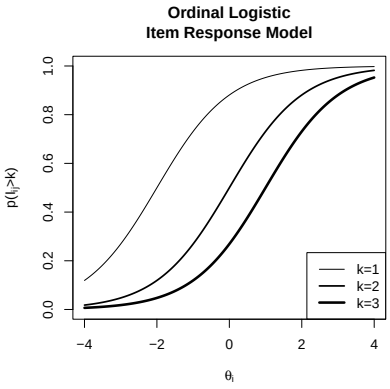


Figure 12.3: Ordinal logistic item response curves for a four level ordered response variable as a function of unit-level latent variable θ_i.

three levels (and thus two curves). The curves for a given item are "parallel" to one another, as the model assumes (the proportional odds assumption) that increasing the latent factor value is associated with increasing (solid lines) or decreasing (dashed lines) the odds of being at all higher levels relative to all lower levels. Increasing the latent factor θ_i cannot increase (or decrease) the probability of the more extreme levels of the item response at the expense of the interior levels, for example.

Notice that if there are only two levels for a given indicator, this model reduces to the binary logistic response model. Further, there is no requirement that different indicators have the same number of items. It is straightforward to extend item response models to other types of responses (unordered categorical, count, etc) in ways that are analogous to regression models. Some of these are implemented in R packages and other statistical software, while other extensions require custom modelling that is beyond the scope of this course.

12.3 Application - PHQ-9 Depression Screening

The Patient Health Questionnaire-9 (PHQ-9) is a 9 question survey instrument that was designed to provide an initial screening test for depression. The PHQ-9 is used by the NHS in the UK as well as widely in the US and elsewhere around the world. It has its own Wikipedia page. All of the questions on the instrument are of the same form:

Over the last 2 weeks, how often have you been bothered by the following problems:

- I1. little interest or pleasure in doing things?
- I2. feeling down, depressed, or hopeless?
- I3. trouble falling or staying asleep, or sleeping too much?
- I4. feeling tired or having little energy?
- I5. poor appetite or overeating?
- I6. feeling bad about yourself – or that you are a failure or have let yourself or your family down?
- I7. trouble concentrating on things, such as reading the newspaper or watching TV?
- I8. moving or speaking so slowly that other people could have noticed? Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual?
- I9. thoughts that you would be better off dead or of hurting yourself in some way?

The response options for each item are:

- Not at all (0)
- Several days (1)
- More than half the days (2)

- Nearly every day? (3)

The instrument is then used to generate a score on a 0-27 scale, by assigning 0, 1, 2, or 3 points for each of the four response options shown above respectively, for each of the 9 items, and then summing. Thus, the standard way that the data are translated into a scale is through a "sum score" of points assigned for each response. This is extremely convenient for the use of the instrument as a diagnostic instrument, in a way that fitting a model could never be. The reason that it is useful to go ahead and explore such data with item response models is that it is helpful for assessing the extent to which the responses actually are associated with a common underlying dimension of variation (a factor). We will use the point totals as a comparison for the value of θ_i that we recover from estimating the item response models.

12.3.1 NHANES Data

We are going to look at data on responses to these nine items from the 2015-16 US National Health and Nutrition Examination Survey (NHANES):

> The National Center for Health Statistics (NCHS), Division of Health and Nutrition Examination Surveys (DHANES), part of the Centers for Disease Control and Prevention (CDC), has conducted a series of health and nutrition surveys since the early 1960's. The National Health and Nutrition Examination Surveys (NHANES) were conducted on a periodic basis from 1971 to 1994. In 1999, NHANES became continuous. Every year, approximately 5,000 individuals of all ages are interviewed in their homes and complete the health examination component of the survey. The health examination is conducted in a mobile examination center (MEC); the MEC provides an ideal setting for the collection of high quality data in a standardized environment. Details of the design and content of NHANES and the public use data files are available on the NHANES website.

12.3.2 Applying the Binary Item Response Model

We begin by fitting the binary response model to the PHQ-9 data in NHANES, setting "Several days", "More than half the days" and "Nearly every day" responses equal to 1, and "Not at all" responses to 0. This means we are ignoring differences in intensities, and just focusing on whether someone is bothered at all by each of the nine problems listed in the PHQ-9 instrument. We will apply the ordinal response model to these data, using all four levels, once we fully understand the outputs of the binary response model. Table 12.1 shows the estimates for the "difficulty" (α) and "discrimination" (β) parameters describing each of the nine items.

What can we learn by looking at these coefficients directly? First, we can see that all the items have discrimination parameters β_j that are positive. This means that the latent variable / factor that we are estimating is positively associated with all of these indicators. Put differently, there are positive correlations between all of these indicators in the data set. If they were all indicators

Table 12.1: Item parameters for binary logistic item response model fit to dichotomised NHANES PHQ-9 data.

	Difficulty	Discrimination
I1	0.84	2.07
I2	0.81	3.18
I3	0.45	1.57
I4	-0.07	1.91
I5	0.95	1.56
I6	1.17	2.75
I7	1.34	1.89
I8	1.69	1.94
I9	2.20	2.48

of depression, this is what you would expect to see. It is not surprising here, since this is a measurement instrument that has been designed to assess depression.

Second, we can see that the difficulty parameters are roughly ascending. Because our model is parameterised as

$$\log\left(\frac{p(I_{ij}=1)}{p(I_{ij}=0)}\right) = \beta_j\left(\theta_i - \alpha_j\right)$$

the difficulty parameters α tell us the level of θ at which respondents were equally likely to give a response corresponding to $I_{ij} = 0$ ("Not at all") and a response corresponding to $I_{ij} = 1$ ("Several days", "More than half the days" or "Nearly every day"). Thus, these values tell you something useful about the relative frequency of different indicators of depression. The indicators with "higher difficulty" are rarer and indicate higher levels of the latent variable. Again, this is not surprising. The final items, particularly I9, are problems that are both relatively rare in the data and which are likely to be accompanied by many of the other indicators.

We can see this clearly in the raw data. For example, if we cross-tabulate I4 ("feeling tired or having little energy?") and I9 ("thoughts that you would be better off dead or of hurting yourself in some way?"), we see that I4 is far, far more common than I9:

	I9 No	I9 Yes
I4 No	2440	24
I4 Yes	2523	169

But crucially, we also see that I9 ("thoughts that you would be better off dead or of hurting yourself in some way?") is far more common among those who experience I4 ("feeling tired or having little energy?") than among those who do not. The comparison here is between the top row and the bottom row of the table. Among those who did not experience I4, only 1% experienced I9; among those who did experience I4, 6% experienced I9. I9 is rare, so neither of these numbers are high, but in the raw data we can see that the I9 is far higher among those who experienced I4, which (along with similar patterns for the other items) is why we see both items with high discrimination parameters.

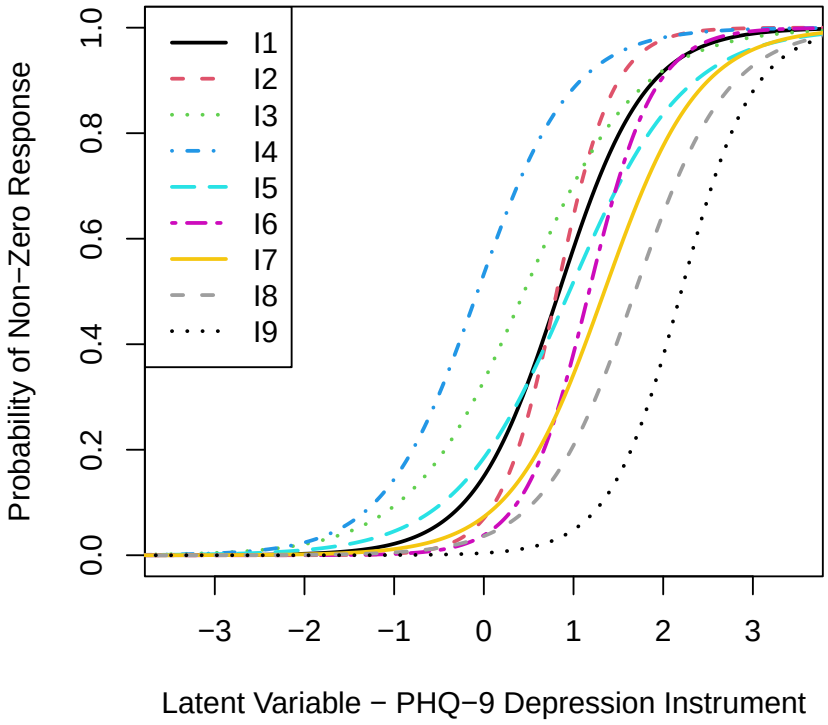


Figure 12.4: Item response curves for binary logistic item response model fit to dichotomised NHANES PHQ-9 data.

Figure 12.4 shows how the model describes these relationships. The plot illustrates the "item response curves" for all nine indicators, which are the fitted values (predicted probabilities) from the model as a function of θ_i. We can see in the plot that the indicator I4 starts to increase in prevalence at the lowest level of the latent variable, while I9 does not start to increase until much higher levels. At levels of the latent scale where nearly everyone reports I4 at least several days in the last two weeks, those in the upper half of the scale, very few people report I9. The latter is associated with much higher levels of the latent scale than the former. This means both that it is much rarer for people to give non-zero responses to this item than the others, but also that people who do are very likely to have given non-zero responses to most of the other items. We have already seen this in the raw data, the plot simply shows how this is captured by the model.

Item 2 (I2) is the most strongly associated with the scale, it has the highest discrimination parameter. This is perhaps unsurprising, because it is the item that actually uses the word *depression* itself: "feeling down, depressed, or hopeless". This is reassuring in the sense that this is the thing that the scale is meant to measure, so it is reasonable that the item that most explicitly mentions it is most strongly associated with the latent factor level. Item 5 (I5) is the least strongly associated with the scale. This is the item about whether you have experienced "poor appetite or overeating?". This too is perhaps not too surprising: while issues with diet are associated with depression, they are also

	Difficulty 0|123	Difficulty 01|23	Difficulty 012|3	Discrimination
I1	0.86	1.76	2.28	1.98
I2	0.82	1.69	2.19	3.01
I3	0.43	1.50	2.05	1.64
I4	-0.07	1.26	1.87	1.91
I5	0.92	1.92	2.49	1.64
I6	1.15	1.88	2.33	2.80
I7	1.29	2.06	2.51	2.00
I8	1.70	2.44	2.91	1.90
I9	2.21	2.80	3.22	2.46

Table 12.3: Item parameters for ordinal logistic item response model fit to dichotomised NHANES PHQ-9 data.

associated with other mental and physical health problems. It is therefore not surprising that the presence of poor appetite or overeating is less *diagnostic* for depression specifically. It is still strongly associated with the latent variable that we have estimated, just less so than the other items.

Thinking more generally, when applying these methods it is useful to look at these sorts of extreme cases of the discrimination and difficulty parameters. They indicate something useful about which items indicate high or low levels on your scale (difficulty) and which items are more strongly associated with the scale, or put differently, with the other items (discrimination). This is an example where there is not a lot of variation in discrimination, precisely because the survey items have been selected to all be useful indicators of depression. In applications with data where the indicators have not already been extensively validated, you can see much more varied discrimination parameters, with some items barely associated with the latent factor. There are also contexts in which there are a mix of positively and negatively associated items. Please note that one nice thing about these models is that they "automatically" determine which items are positively versus negatively associated with the latent variable, so it does not matter if you code all your indicators in the same direction.

12.3.3 Applying the Ordinal Item Response Model

The analysis above used a binary item response model, collapsing all non-zero response levels into a single category. We can now apply the ordinal response "graded response" model to the full data. If you compare the coefficients in Table 12.3 to those obtained earlier for the binary model in Table 12.1, you will see that the first column of difficulty parameters from the ordinal model are very similar to the column of difficulty parameters from the binary model. This is because they correspond to the same response threshold: between never having a problem and having it at least "several days" in the last two weeks. If the proportional odds assumption of the ordinal model were perfectly accurate in describing all levels of the response data (which it never will be, if only because of sampling variability) these parameters would be identical. Similarly,

the discrimination parameters are very similar across the two models, as they are defined in analogous ways. The advantage of the ordinal item response model over the binary item response model, just like the advantage of the ordinal logistic regression model over the binary logistic regression model, is simply that using more response levels yields more information and more precise estimates given the same amount of data. If the assumptions of the ordinal model are correct, and you have a lot of data, you will get the same estimates from the ordinal model as you would from dichotomising at *any* threshold and using the binary model.

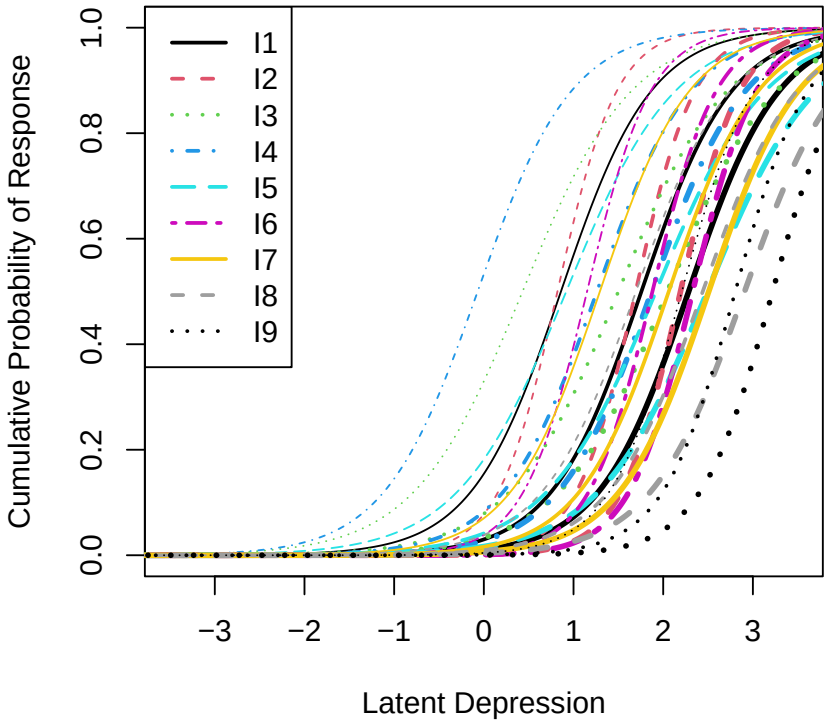
The proportional odds assumption is visible in the fact that the three item response curves for each item have the same shape/slope in the plots. If we compare, as we did before, the highest (I9) and lowest (I4) difficulty items, we see that their item response curves (which now correspond to the cumulative probabilities of giving responses above each threshold) fail to overlap. That is, saying that you are "feeling tired or having little energy" almost every day (the highest response level, 3 points on the PHQ-9) is more probable at any given level of the latent depression scale than saying that you have "thoughts that you would be better off dead or of hurting yourself in some way" at the lowest non-zero response level of several days in the last two weeks (1 point on the PHQ-9). Again, this does not seem all that surprising, given the content of the two items.

12.3.4 Comparison to PHQ-9 Points Scale

Recall that the PHQ-9 is typically scored using a 0-27 point scale, awarding 0, 1, 2 or 3 points for different responses to each item. This points scale has been extensively validated against clinical diagnoses of depression. A recent meta-analysis of studies comparing 0-27 points scores on the PHQ-9 to a gold-standard of assessments from diagnostic interviews concluded that, using a cutoff of 10 points, the instrument could achieve a sensitivity of 0.88 (95% interval: 0.83-0.92) and a specificity of 0.85 (95% interval: 0.82-0.88) (Levis et al., 2019).[2] Sensitivity, applied to this context, is the proportion of those who actually have depression (according to the gold standard) who are correctly identified as having depression according to the measurement. Specificity is the proportion of those who do not have depression (according to the gold standard) who are correctly identified as not having depression according to the measurement.

Figure 12.6 shows that if we compare the estimated factor scores θ_i from our ordinal item response model to the point totals from the standard diagnostic instrument, including only respondents who gave responses to all items, we see that there is a very strong relationship between the two. As you can see in the plot, there are some marginal cases that are classified differently by the item response model than the point system that is actually used in the PHQ-9 instrument, but not very many. The vertical dotted line indicates the optimal

[2] Note that the benchmark here is a dichotomous definition of depression—you have it or you do not—while the instrument generates an interval(ish)-level scale that is then being dichotomised. More on this point in the remainder of this chapter as well as the next one.



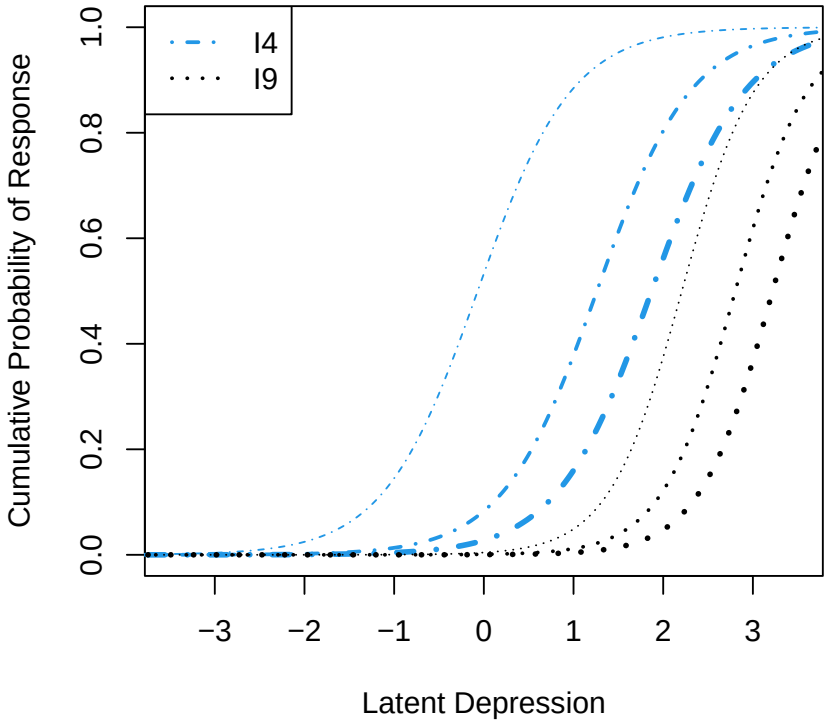


Figure 12.5: Item response curves for each level of each indicator of the PHQ-9 under a graded response ordinal IRT model with a single latent dimension. Top plot shows all thresholds for all indicators, bottom plot shows only I4 and I9. Thresholds for higher response levels use thicker lines.

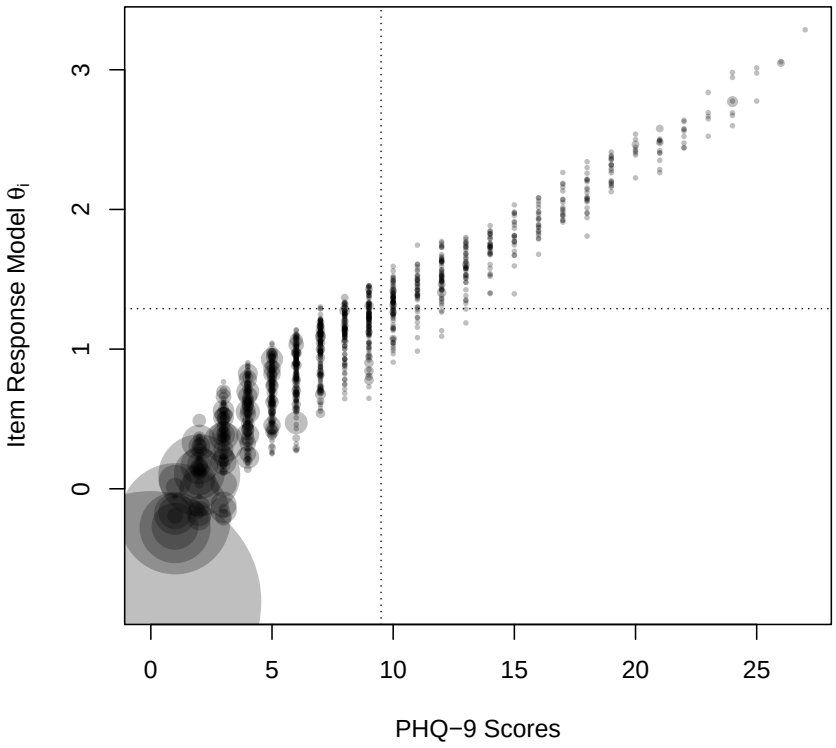


Figure 12.6: Graded response ordinal IRT scores by PHQ-9 scores for NHANES sample, point size proportional to number of individuals with each response pattern. The vertical line shows the PHQ-9 cutoff suggested by a recent meta-analysis of validation studies versus diagnostic interviews; the horizontal line is the threshold in the factor score with the same proportion of observations classified as high on the scale.

PHQ-9 cutoff from the meta-analysis mentioned earlier. According to this cutoff, 8 of NHANES respondents are classified as having responses indicative of depression. The horizontal dotted line on the plot shows the latent variable estimate that yields the same classification proportion. Only 92 of the 5134 complete observations in this data set are classified differently by the point system that is actually used and the latent variable estimated from the ordinal item response model.

This is an example where a relatively simple measurement strategy involving a scoring rule for allocating points to different responses and adding them up comes close to the same conclusions as a measurement model. The reason for this, as was the case for the previous examples where we have seen this, is that researchers developed the scale to be well-behaved in this way. All the indicators have been validated to make sure that they are similarly responsive to the concept of interest and more weakly related indicators have been discarded. The item response models are a way to evaluate the extent to which the indicators seem to reflect a common underlying concept of interest, and therefore to validate the point system. The point system has the obvious advantage that it can be quickly applied in practice, without the need to fit a larger model on a broader set of data. Nonetheless, if you are trying to develop a new measurement strategy, tools like item response models are valuable in that process.

12.4 Conclusion

Item response models are used to design and validate standardised education tests. Most tests that students take are not marked this way. No one fits a model, they just count up how many questions or items you got right. But some items may be more or less indicative of underlying understanding/ability. For example, imagine there is a weird question in the middle of your test that has nothing to do with the material of the course. One might expect that performance on that item would be weakly related to the other items. That might be a reason not to include that item on future tests. So these kinds of models are potentially useful in test development, even if they are rarely used to score tests directly.

Item response models are widely used in political science in order to model how different kinds of political responses reflect underlying political preference dimensions. Common applications include votes in legislatures (Clinton et al., 2004), decisions by judges (Martin and Quinn, 2002), and survey responses of citizens (Bafumi and Herron, 2010), to name just a few. These methods work well in some contexts, but less so in others. House of Commons voting in the UK is poorly approximated by these models because of very strong party discipline (Spirling and McLean, 2007).

Conceptually, the sources of measurement error in item response models are the same as the sources of error in factor analysis that we considered in the previous chapter. And once again, it is critical to remember that these are "unsupervised" measurement methods. Item response models discover whatever latent factors explain the most variation in your indicators. They do not necessarily measure what you want them to measure. As with principal components and factor analysis, indicator selection is crucial to ensuring that these methods measure what you want them to measure.

13
Unsupervised Class Measurement with Interval-Level Indicators

In Chapter 11, we considered principle components analysis and factor analysis, methods for generating continuous/interval-level measures from continuous/interval-level indicators. In Chapter 12, we considered item response models, which were methods for generating continuous/interval-level measures from categorical/nominal/ordinal indicators. In this chapter and the next, we fill in the two corresponding cases for measuring categorical/nominal/ordinal measures from continuous/interval-level indicators and also from categorical/nominal/ordinal indicators.

	Continuous Indicators	**Categorical Indicators**
Continuous Measure	Chapter 11	Chapter 12
Categorical Measure	Chapter 13	Chapter 14

Just as principle components analysis is an algorithm that aims to efficiently describe variation in a set of indicators in terms of a set of continuous scale components; there are methods that aim to efficiently describe variation in the set of indicators in terms of membership in groups/clusters/classes. Just as factor analysis and item response models are methods that model how indicator data could have arisen from latent continuous variables, there are methods that describe how indicator data could have arisen from latent categorical variables. I will divide up the discussion in this chapter into clustering methods (that aim to minimise within-class variation and maximise across-class variation or similar criteria) and model-based methods (that aim to describe how indicators could have been generated from underlying classes). Note that, as was the case with measuring scales, these will sometimes yield similar classifications, even though they are conceptually different ways of approaching the problem.

Measurement Models	Continuous Indicators	Categorical Indicators
Continuous Latent Variable	Factor Analysis Models	Item Response Models
Categorical Latent Variable	Gaussian Mixture Models	Latent Class Models

It is unfortunate that the standard names for these models are not more logically organised, but these are the conventional names that you will come across. These are all examples of models which describe how latent variables with different levels of measurement are related to indicator variables with different levels of measurement. All of these models can be subsumed in a general framework of Generalized Linear Latent and Mixed Models (GLLAMMs) as described by Skrondal and Rabe-Hesketh (2004).[1]

[1] Models for mixed types are also straightforward to describe in this framework. You can have a model which generates both continuous and categorical indicators from scales. You can have a model which generates both continuous and categorical indicators from classes. You can also have mixed types of latent variables, for example one binary and one continuous latent variable. This final version is rarely done, because issues surrounding identification of the latent scale and class are challenging to deal with in many applications, but it is a logical possibility.

13.1 Clustering Algorithms

There are an enormous number of clustering algorithms that have been developed, using a variety of criteria to assign units to clusters. We will consider two of the more commonly applied algorithms here, and then discuss more general conceptual issues with all such algorithms.

13.1.1 K-means clustering

One commonly used method for clustering is "k-means" clustering. The core logic of k-means clustering is that we want to partition the set of observations into k groups in the way that minimises the within-group sum of squared distances from the within-group mean. The number of groups k is chosen by the researcher. Let's say we have p observed indicators I_{ij} $(j = 1, 2, \ldots, p)$ measured for each unit i in a sample of data. We are trying to assign $G_i \in 1, \ldots, k$ such that the following is minimised:

$$\sum_{i=1}^{n} \sum_{j=1}^{p} \left(I_{ij} - \overline{I}_{jG(i)} \right)^2$$

where $\overline{I}_{jG(i)}$ is the average value of variable j for the units i assigned to the same group $G(i)$ as unit i. The algorithmic details of actually finding (or trying to find) the optimal allocation of the units into groups are beyond our scope here. It is impossible to guarantee that one has found the optimal allocation for even moderate sized data sets, but a great deal of research effort has gone into finding reasonably reliable search algorithms (Steinley and Brusco, 2008).

Like Principle Components Analysis, k-means treats variation in all variables as equally important to "explain", and so is sensitive to the scale on which the variables are defined. As a consequence, unless those variables are on comparable scales already, it is common to standardise them so that the algorithm

"tries" equally hard to explain variation on all variables. As noted when we discussed PCA and previously, this kind of equal weighting assumption may or may not be suitable to a given problem, but among many possible arbitrary choices, it is often the one people choose.

As a simple toy example to illustrate how k-means clustering works, consider the a data set with a single indicator $j = 1$, and five units with x equal to -15, -15, 5, 20 and 40. If we apply k-means with $k = 2$, it turns out that the two clusters that minimise the within-group sum of squared distances put -15, -15 and 5 in one cluster (mean -8.33, sum of squares 266.66) and 20 and 40 in the other cluster (mean 30, sum of squares 200). The total sum of squares is therefore 266.66 + 200 = 466.66.

It is perhaps slightly counter-intuitive that this is the best $k = 2$ clustering, because 5 is closer to 20 than it is to -15, but if you put 5 in the other cluster the sum of squares would increase, as the cluster -15, -15 has sum of squares 0 while the cluster 5, 20, 40 has a higher sum of squares (616.66). This toy example illustrates a key feature of k-means clustering: it will generally prefer to create clusters of about the "same size", as this tends to reduce squared error versus having a relatively spread out cluster and a relatively compact cluster.

13.1.2 Heirarchical clustering

Another commonly used strategy for algorithmic clustering is *hierarchical clustering* which creates a full series of clusterings from $k = 2$ to $k = n$, for a data set with n observations. *Agglomerative hierarchical clustering* does this by starting with the $k = n$ "clustering" where every observation is its own clustering, and then combining the two clusters that are "closest together" into a new cluster. Since each step reduces the number of clusters by 1, it yields a full array of possible clusterings. There are different possible ways to specify distances between clusters, and thus which clusters should be combined at each step. Each of these will yield different hierarchies of clusterings. The most common way is using Euclidean distances between the mean value of the units in each cluster. There are also *divisive* methods that start with all units in one cluster and sub-divide repeatedly.

Using the same toy example that we considered for k-means clustering, -15, -15, 5, 20 and 40, we can easily work out how *agglomerative hierarchical clustering* proceeds. The first pair of units that are combined into a cluster are the two with identical values of -15 and -15, as they have distance 0 and are therefore closer together than any other pair of units. This leaves clusters centered at -15 (-15, -15), 5, 20 and 40. The closest of these are 5 and 20, so these are then combined into a cluster, leaving clusters centered at -15 (-15, -15), 12.5 (5, 20), and 40. The closest of these are -15 (-15, -15) and 12.5 (5, 20), so these are then combined into a cluster, leaving clusters centered at -1.25 (-15, -15, 5, 20) and 40. Thus, the agglomerative hierarchical clustering yields a different 2 cluster solution than the k-means clustering, with all but one unit in one cluster and a

single unit in the other.

The results of this approach can look rather different from k-means, as we see both in the toy example and in the application later in the chapter. Agglomerative clustering does not have the tendency of k-means to produce similarly sized, spherical clusters, because nothing in the procedure pushes it in that direction. It simply aims, at each step, to join clusters that are as close to one another as possible (by various metrics). This can lead to very differently sized clusters as the process approaches a single cluster of all the data.

13.1.3 Alternative clustering algorithms

There are many, many alternative clustering techniques. There are a wide variety of criteria for similarity and difference that can and have been used with these, as well as other clustering algorithms not discussed here. Grimmer and King (2011) develop a set of tools for exploring and comparing a very large number of *possible* clusterings defined by different algorithms, with an eye towards identifying those clusterings that are "insightful" or "useful" for a given application. Another way to state this is that the aim is to look for clusterings that measure something of interest to the analyst.

13.2 Gaussian Mixture Models for Continuous Indicators

What factor analysis is to principle components analysis, *Gaussian mixture models* are to the K-means clustering method that we introduced earlier in this chapter. Factor analysis is a model-based method for inferring latent factors that are linearly related to observable indicator variables, whereas principle components analysis is an algorithmic decomposition of variance into linear functions of the observable indicator variables. Gaussian mixture models are a model-based method for inferring latent classes that generate multivariate normal clusters of observations, whereas K-means is an algorithmic decomposition of a data set into classes that minimise within-class variance. More generally, one can define gaussian (or, more atypically, non-gaussian) mixture models with different distributions: the key feature that connects them to measurement of a categorical latent variable is that the observed data are a mixture of data from different classes, each of which has a distinctive distribution. The unsupervised measurement task is to infer both these distinctive distributions that characterise each class and which observations belong in each class, from the observable data.

By analogy to the factor analysis and the item response models we have looked at previously, we use θ_i to correspond to the latent variable describing each unit i, only now this is a categorical variable with $\theta_i = k$ for $k \in 1, 2, \ldots, q$. As before, i indexes units; j indexes indicators, of which there are p in total; and k indexes the latent clusters/classes, of which there are q in total.

The core assumption of the Gaussian mixture model is that the indicators

are normally distributed around mean values that are characteristic of different levels of the categorical latent variable θ_i, which is to say different clusters. The clusters are sometimes specified to have independent normal distributions for each indicator:

$$I_{ij} \sim N\left(\mu_{\theta_i}, \sigma^2_{\theta_i}\right)$$

It is typical to allow the $\sigma^2_{\theta_i}$ to vary across clusters, which means that (unlike k-means) the clusters can be "wider" with respect to some indicators than others. Where possible, these models assume multivariate normal distributions across indicators:

$$I_i \sim MVN\left(\mu_{\theta_i}, \Sigma_{\theta_i}\right)$$

This is a more general specification that includes independent normal distributions as a special case, but requires more data to estimate reliably. The multivariate normal version allows for indicators to have correlated values for a given cluster, which means that the clusters can take on any ellipsoid shape rather than only spheroid shapes around the cluster mean.

Figure 13.1 is an example of the kind of data for which one might wish to use a Gaussian mixture model. There are two indicators I_1 and I_2, and the figure suggests that it might usefully be described in terms of clusters. However, the clusters do not appear to be spherical, so k-means clustering may not be suitable. Figure 13.2 shows the results of estimating a Gaussian mixture model with multivariate normal indicator distributions for three classes. We see that within each class, the two indicators are estimated to be positively correlated, and the data seem to be generally consistent with the three class model.

In fact, these data describe two physical measurements (bill length and bill depth) on a sample of penguins from three different species (Horst et al., 2020).[2] Figure 13.3 uses the species labels to colour the points, and the match to Figure 2 is extremely close with 97% of penguins correctly classified. Recall that we have used an *unsupervised* method here, we did not give the Gaussian mixture model any information about species.

If one actually had the aim of measuring penguin species on the basis of their bill length and bill depth, one might want to use one of the *supervised* methods discussed in the previous chapter. The unsupervised method works unusually well here because the Gaussian mixture model describes a generative process that closely approximates the true distribution of the data, which really is a mixture of three types/classes (species), each of which is approximately multivariate normal in these two indicators. Distributions of physical characteristics of animals are often very well approximated by multivariate normal distributions, and so this simple example is really a best case for a Gaussian mixture model yielding accurate measurements of a meaningful latent class.

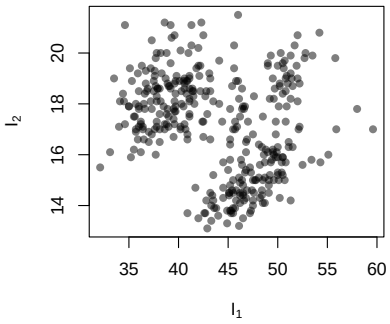


Figure 13.1: Data with two indicator variables.

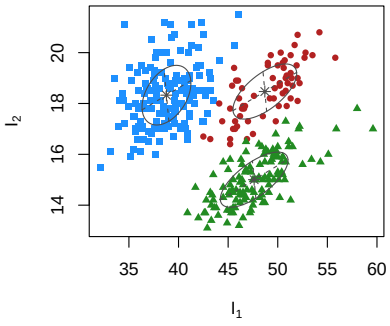


Figure 13.2: Gaussian mixture model fit with three clusters/classes. Ovals indicate the standard deviations and correlation of the indicators for each class.

[2] The idea for this example came from a tweet by Oli Hawkins

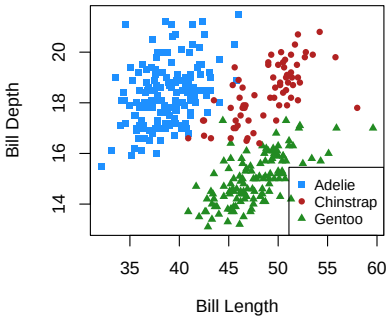


Figure 13.3: Species of penguin.

13.3 Application - Clustering Political Attitudes

13.3.1 K-means

To look at an example of k-means clustering, we will once again use the ideology questions asked in the 2017 BES, the same data set we used for our discussion of principle components analysis (PCA) in Chapter 11. This is an intentional choice. PCA aimed to summarize the variation in these data in terms of continuous summary measures (scales); clustering aims to summarize the variation in these data in terms of categorical summary measures (classes). They are both trying to achieve the same thing, and both have the same objective function of minimising squared errors / unexplained variance. They differ in that PCA provides one or more "scales" as a summary and clustering provides two or more "classes" as a summary.

As with PCA, the fact that the objective function is minimising squared errors across all indicators means that it typically makes more sense to use k-means on standardised indicators, or on variables that already have the same dimensions/scales. As you will recall from the discussion in Chapter 11, the BES ideology questions are arguably already on comparable scales, and so we will not standardise them.

First, we will compare the estimates with the PCA estimates on the same data to see how the two approaches describe the response patterns in the data in different ways. In Figure 13.4, I compare k-means with 2, 3, and 4 clusters to the first two principle components. As you can see, the k-means with $k = 2$ divides the respondents into high and low values on the first principle component. The first principle component is the continuous scale which most efficiently explains variation in the responses to this battery of questions, which means it is the scale that minimises residual variance. The k-means clustering with $k = 2$ divides the units into two groups that minimise within group variation, which is to say, residual variance. The k-means clustering with $k = 2$ will generally divide the data into "high" and "low" values on the first principle component: these are simply two different ways of describing the same variation in the data.

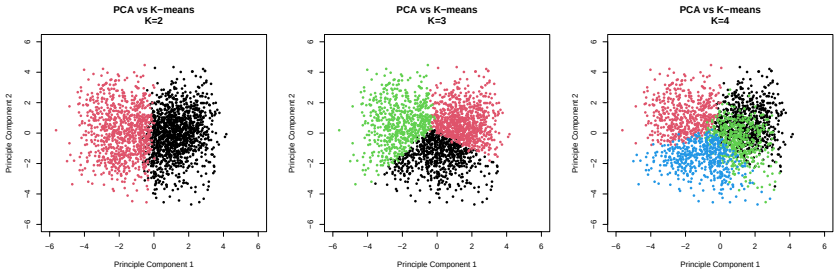


Figure 13.4: K-means cluster assignment as a function of principle components.

When we move to $k = 3$, we begin to see an important feature of k-means

clustering, which is that it tends to create clusters that are "spherical". Because within-group variation is with respect to the mean, the most efficient way to minimise that variation is to describe the data in terms of groups that are shaped like circles/spheres/hyper-spheres in terms of the underlying variables (and as a consequence, also the principle components, which are linear combinations of those variables). At $k = 3$, the three clusters correspond to three disjoint groupings in the first two principle components. Once we move to $k = 4$, the first two principle components are no longer sufficient to describe the variation in the four groups that the clustering algorithm identifies. This does not mean that either is *wrong*, it just illustrates that efficient explanation of variation begins to look increasingly different when done in terms of scales versus classes as the number of scales/classes increases. The $k = 2$ clustering will generally divide the data in the middle of the first principle component, but after that the relationships becomes more complicated.

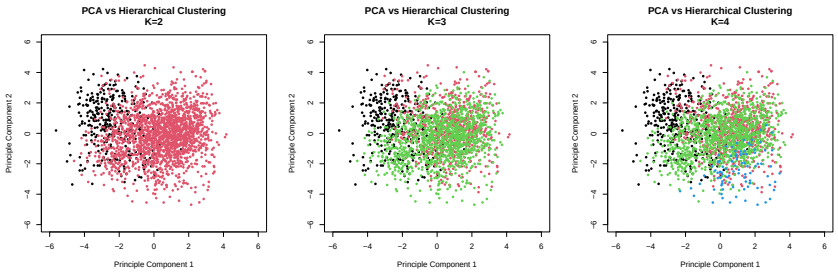


Figure 13.5: Agglomerative hierarchical cluster assignment as a function of principle components.

As we saw in the toy example earlier, agglomerative hierarchical clustering does not necessarily create equal size groups. Figure 13.5 shows that the clusters generated for these data also do not cleanly separate units based on their principle components. K-means is a clustering algorithm that is a particularly close analogue to principle components analysis, but other clustering algorithms can yield very different results.

13.3.2 Validation

You might now be wondering, which of these (or the many other) clustering algorithms should you use? There is no *right* way to do algorithmic clustering. There are many different clusterings, each as valid for a given application as the procedure that generated it is as an approximation of the researcher's goals in doing the clustering (Grimmer and King, 2011). From the perspective of measurement, these procedures each define their own target concept, all categorical in conceptualisation, but slightly different in terms of how they define what it means to be a category. None of them can be reliably expected to recover "the true" clusters in general, if such a thing were to even exist. These are *pragmatic measurement* at its most pragmatic, where it runs up against such tasks as *exploratory data analysis*, *description* and *visualisation*. Whether a given

clustering is useful for a given purpose will depend on that purpose.

One purpose we might have, with this example, is identifying "ideological groups" of voters and relating their views to political choices. Just as we examined the relationship between the components from PCA and vote choice, we can examine the relationships between the classes recovered from these clustering procedures with vote choice.

	Leave	Remain	Con	Lab	LD	SNPPCGreen	UKIP
Cluster 1	65	35	48	39	4	6	3
Cluster 2	30	70	31	50	11	7	1

With two clusters, K-means separates voting groups partially. Increasing the number of clusters does seem to identify meaningfully different groups. With three clusters, we start to distinguish between Conservative-leaning Leave voters (cluster 1) and Labour-leaning Leave voters (cluster 2).

	Leave	Remain	Con	Lab	LD	SNPPCGreen	UKIP
Cluster 1	56	44	68	20	5	5	2
Cluster 2	67	33	35	50	5	6	4
Cluster 3	22	78	21	59	12	7	1

And with four, we have groups roughly corresponding to all four combinations of Leave and Remain, Conservative and Labour, with supporters of Liberal Democrats, SNP, PC and Greens tending to end up in the cluster with Labour Remainers.

	Leave	Remain	Con	Lab	LD	SNPPCGreen	UKIP
Cluster 1	66	34	38	48	4	6	4
Cluster 2	23	77	14	65	12	8	0
Cluster 3	65	35	53	33	4	6	4
Cluster 4	44	56	63	25	7	4	1

As with when we completed similar validation for PCA on these data, none of the relationships are very strong. This reflects the fact that few citizens hold ideologically consistent views across many issues. Nonetheless, we see substantial differences in voting between the different classes, which broadly reflect the relationships that we would expect to see.

13.4 Application - Constituency Politics in the UK

K-means clustering tends to generate circular/spherical clusters in the indicator dimensions, because it weights variance in all indicator variables equally. In

constrast, the fact that a multivariate normal can have correlation across indicator dimensions means that under a Gaussian mixture model classes can have either spherical or ellipsoidal distributions of indicator values. To illustrate this useful feature of Gaussian mixture models, I will use a new example. We are going to try to cluster UK parliamentary constituencies into different classes based on five continuous variables: support for Leave in the 2016 referendum on membership in the European Union and support for the Conservative, Labour, Liberal Democratic and Scottish National parties in the 2017 UK general election. The idea is that we might be interested in describing different 'political types' of constituencies.

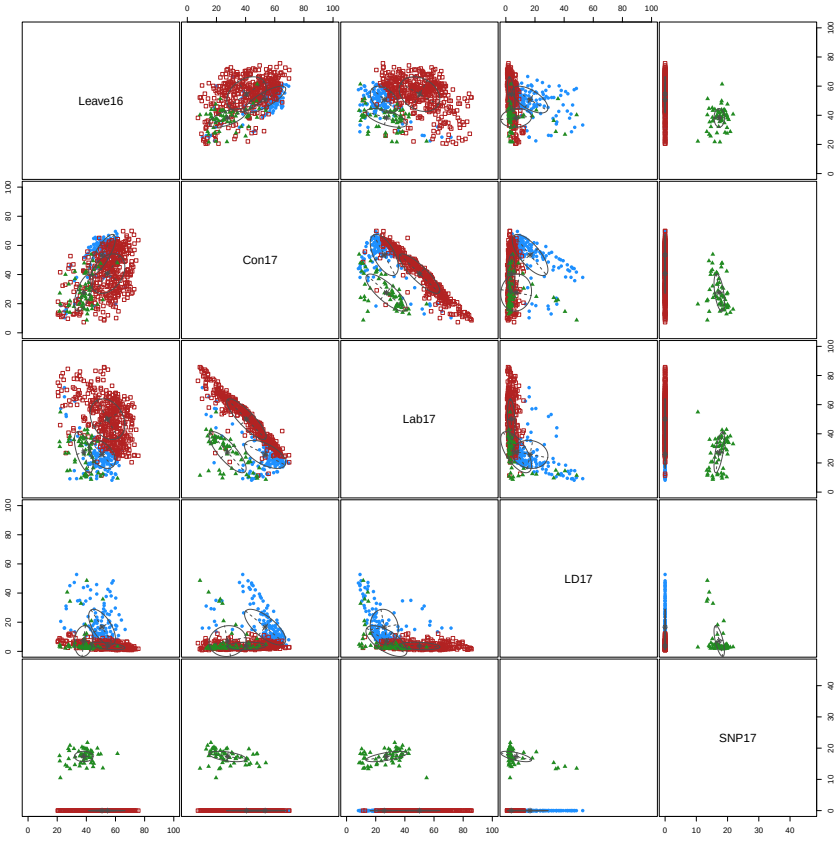


Figure 13.6: Gaussian mixture model assignment to three classes, as a function of pairs of five constituency vote variables.

In order to keep the number of figures manageable, we will focus on 3, 4 and 5 class Gaussian mixture models. Figure 13.6 illustrates how the 632 constituencies in England, Scotland and Wales are assigned to three classes as a function of pairs of the indicator variables.[3] Working out what this tells us requires examining several of the pairs plots in that figure. If you look at the plots for Con17 by Lab17, you will see one class (red squares) that is tightly clustered along the diagonal. Given that the vote shares for the two parties cannot exceed 100%, this illustrates that this class consists of "Lab-Con constituencies", those in which none of the minor parties received a significant share of the vote in

[3] I use default settings in the R package mclust (Scrucca et al., 2016).

the 2017 election. If you examine the plots where LD17 is one of the plotted variables, you will see that the blue circles are constituencies where the Liberal Democrats received a non-trivial share of the total vote. Finally, the green triangles are all seats where the Scottish national party receives a substantial share of the vote, which is to say all the seats in Scotland and none of the seats elsewhere. The three classes are thus Lab-Con seats, English and Welsh seats where the Lib Dems get a non-trivial vote, and Scottish seats.

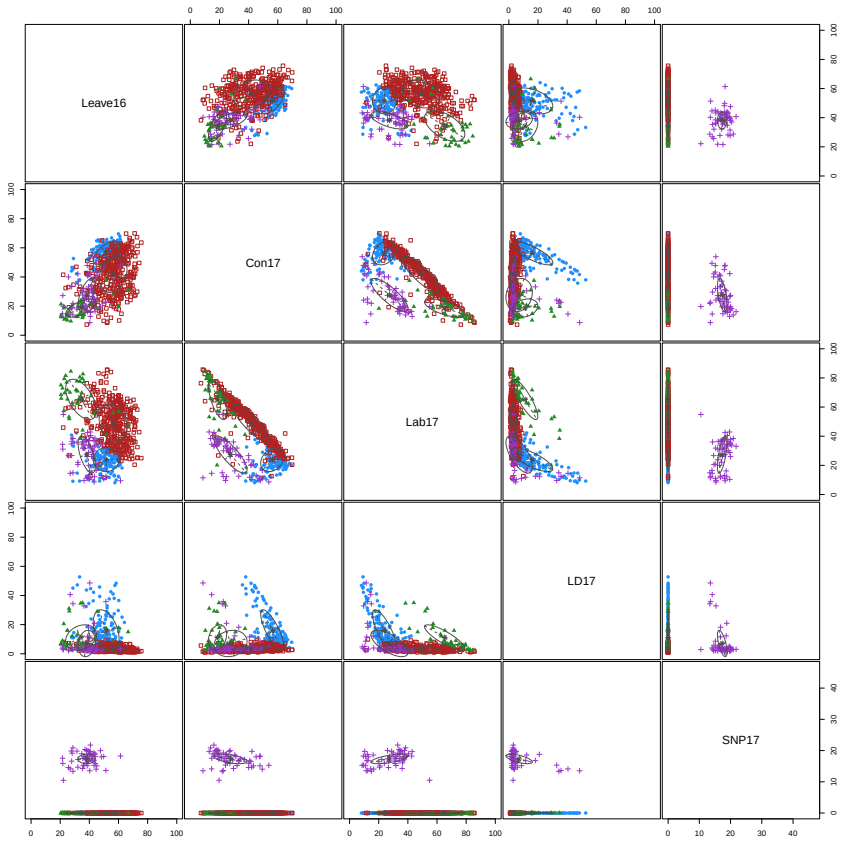


Figure 13.7: Gaussian mixture model assignment to four classes, as a function of pairs of five constituency vote variables.

This typology is reasonable, if you know anything about UK political competition, but is perhaps too parsimonious. If we move to a four class model, plotted in Figure 13.7, and look at the plots in a similar way, we see that the four classes correspond to a Lab-Con class, a Scottish class, but that there are now two classes with non-trivial Lib Dem vote, one where the Conservatives have high vote shares (blue dots) and one where Labour has high vote shares (green dots). So now the four classes are Lab-Con seats, Lab-LD seats, Con-LD seats, and Scottish seats. Talking about Lab-Con, Lab-LD and Con-LD seats as separate classes is a traditional way of talking about constituency politics in the UK.

It is important to recognise that nothing magical is happening in this *unsupervised* model, it is simply trying to describe the the major patterns in the vote

share data as efficiently as possible. It is notable that the model does not really get very "distracted" by the inclusion of the EU referendum vote share data. The classes have different average leave shares, but they overlap in distribution very substantially.

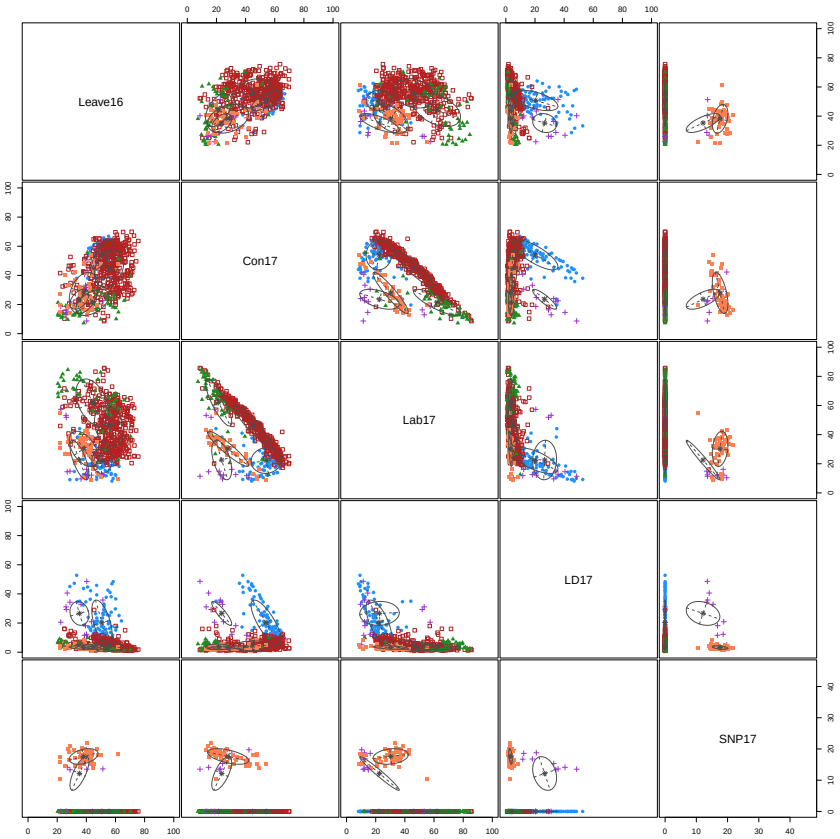


Figure 13.8: Gaussian mixture model assignment to five classes, as a function of pairs of five constituency vote variables.

Figure 13.7, illustrating a five class model, does not lead to additional clarity in this case. The classes start to be subdivided in ways that are more difficult to make sense of. You might reasonably ask, when should we stop? How many classes should we use? Here, we return to a question that first arose when we introduced *unsupervised* measurement methods with principle components analysis. How much of the variation explained is enough? The screeplots we used to visualise variation explained for principle components have analogues for gaussian mixture models. Figure 13.9 shows a similar kind of plot using the BIC fit statistic, for models with 1 to 8 classes, and a variety of different restrictions on variances and covariances of the multivariate normal distributions used to describe each class.

Figure 13.9 shows five different kinds of restrictions on the multivariate normal distributions that clusters are assumed to follow, some of which fit better than others (better fit is higher on the plot). The most restricted models assume that the variances of all indicators are equal, all clusters have the same

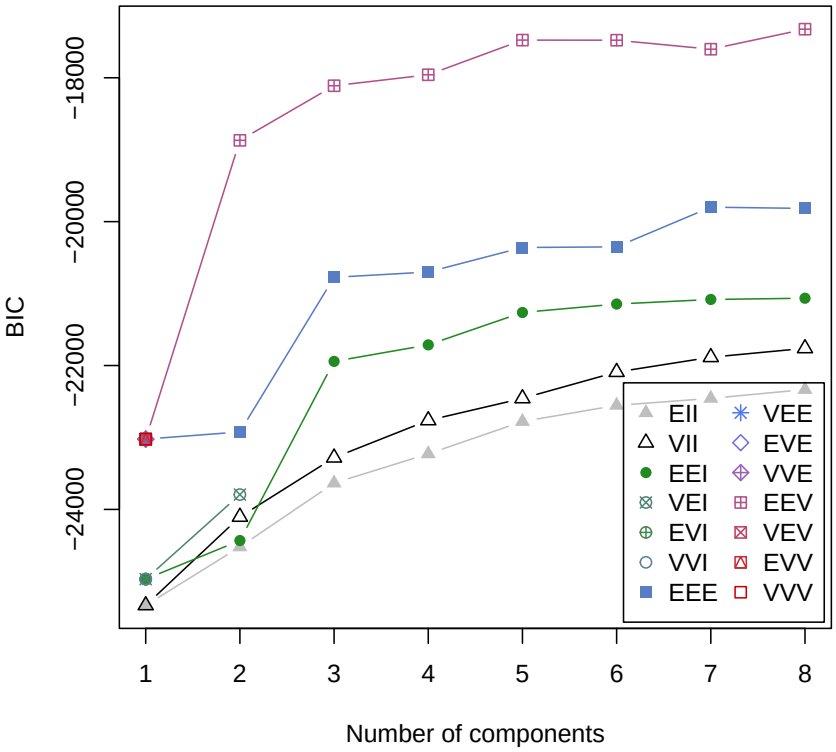


Figure 13.9: BIC fit statistic for Gaussian mixture models of 1 to 8 classes and different restrictions on the multivariate normal distributions of indicators.

variances, and that all indicators are independent (EII). This model closely resembles the underlying logic of k-means, assuming that clusters describe spheres in the space of indicators. Next, we have models where the variances of all indicators are equal, different clusters have different variances, and all indicators are independent (VII). This is the same model as before, but some clusters are allowed to have larger/smaller spheres of indicator values around them. The next model instead allows the variances of different indicators to be different, but all clusters once again have the same shape and size distribution of indicator values around their means (EEI).

The fourth set of models are those where indicators are allowed to be correlated, but all all in the same way and with the same variances (EEE). This allows clusters to have ellipsoid shaped clouds of indicator values associated with them, but the ellipses are all the same shape, orientation and size. Finally, the fifth set of models are those where the different ellipses have the same shape and size, but are allowed to have different orientations. There are many more variations that one can consider, although far more data is required to estimate models that allow all the clusters to have different shapes, sizes and relative variation in indicators.

One thing that is apparent from Figure 13.9 is that the more flexible models fit much better than the ones with stronger constraints. Unfortunately model fit does not give us a very good way of assessing which of the models will be

most useful here. Models with more components/clusters tend to fit better far beyond the level at which they give us much insight into the data (we already were struggling to interpret 5 clusters above). We have seen this already (more principle components / factors is not necessarily better for interpretation) and will see this limitation of model fit as a tool for selecting an unsupervised measurement model acutely in the next chapter as well.

14
Unsupervised Class Measurement with Categorical Indicators

In this chapter, we fill in the final cell of the table I provided in the last chapter: models for measuring categorical quantities using categorical data.

	Continuous Indicators	**Categorical Indicators**
Continuous Measure	Chapter 11	Chapter 12
Categorical Measure	Chapter 13	Chapter 14

My focus in this chapter will be on Latent Class Models / Latent Class Analysis as a method, as this is the most widely applied method of this type. There are methods like community detection algorithms for social network data which could be understood as falling into this category, but analysis of social network data is outside the scope of this book. Here, we are thinking about cases where we have collected categorical data, which are not interval level, on a number of units. We want to measure something about those units, using these data, in an unsupervised way.

14.1 Latent Class Models for Categorical Indicators

Measurement Models	**Continuous Indicators**	**Categorical Indicators**
Continuous Latent Variable	Factor Analysis Models	Item Response Models
Categorical Latent Variable	Gaussian Mixture Models	Latent Class Models

Latent class models fill in the fourth quadrant of the model-based, unsupervised measurement models grid shown earlier: categorical indicators related to

a categorical latent variable. The core idea of latent class models is that there are different latent clusters/classes, and the probability of a unit i having any given indicator value $I_{ij} = l$ depends on the latent cluster/class $\theta_i = k$ of which unit i is a member. Here, because the indicators are categorical rather than continuous, we need the additional index l to indicate the outcome level on indicator j.

The latent class model is based on the assumption that the probability that an individual i in class k has a set of indicator values I_i on indicator j is:

$$p\left(I_{ij} = l\right) = \pi_{j\theta_i l}$$

This expression looks incredibly simple, but much is buried in the indices of the parameters π. The idea is that there are distinctive probabilities π for different levels l of each indicator j that depend on which cluster $\theta_i = k$ the unit is a member of.

A simple example is helpful for getting an idea of how these models work. Imagine that you observe data on three votes taken in a legislature with 100 legislators, which for our purposes we will think of as three indicators I_1, I_2 and I_3. You observe the following pattern of votes:

Count	I_1	I_2	I_3
5	Y	Y	Y
5	Y	Y	N
40	Y	N	Y
0	N	Y	Y
0	N	N	Y
40	N	Y	N
10	Y	N	N
0	N	N	N

So, 40 legislators vote yes (Y) on votes 1 and 3, but no (N) on vote 2 (YNY), 40 legislators vote NYN across the three votes, and the remaining legislators exhibit some of the possible combinations of yes and no votes, in various numbers. Overall I_1 has 60 yes votes, I_2 has 50 yes votes, and I_3 has 45 yes votes, each out of 100 total.

A latent class model with two classes assumes that these data arise from two classes, each with distinctive probabilities of voting on each indicator I. In this instance, if we estimate a latent class model, we estimate classes with the following voting probabilities. The probability of class 1 voting yes on I_1 is 0.09, on I_2 is 1.00, and on I_3 is 0.00. This class almost all opposed I_1, all voted for I_2 and all voted against I_3. The probability of class 2 voting yes on I_1 is 1.00, on I_2 is 0.11, and on I_3 is 0.80. This class all supported I_1, mostly opposed I_2, and mostly supported I_3. This kind of pattern of voting might plausibly reflect opposition (class 1) versus government (class 2) voting patterns, if for

example votes 1 and 3 were government proposals and vote 2 was an opposition amendment.

Count	I_1	I_2	I_3	p(Class 1)	p(Class 2)	Class
5	Y	Y	Y	0	1	2
5	Y	Y	N	0.75	0.25	1
40	Y	N	Y	0	1	2
40	N	Y	N	1	0	1
10	Y	N	N	0	1	2

The table above shows the vote patterns that appear in the data alongside the latent class model estimates of which class they belong to. In this simple example, most of the units are decisively (with probability one) assigned to a particular class, however this is less likely to occur in examples with more indicators and more varied patterns of observed indicator values. The two groups of 40, the YNYs and the NYNs are assigned to different classes with probability 1, as implied by the discussion above. The group of 10 YNNs are assigned to class 2 with probability 1, because class 1 voted for I_2 with probability 1, and so these 10 legislators cannot be part of class 1. Similarly, the group of 5 YYYs cannot be in class 1 because they voted in favour of I_3, while class 1 voted for I_3 with probability 0. The only group whose assignment is probabilistic in this toy example is the 5 YYNs, who are more likely to be members of class 1 than class 2 given their pattern of voting / indicator values, but could be a member of either class.

Again, in an example with a larger set of indicators and more varied patterns of indicator values, we would not see such decisive assignments of units to classes as we do in this example. In general, latent class analysis finds clusters of units with similar patterns of indicator values, for categorical indicators. It yields probabilistic predictions for the membership of particular units in each class as well as a description of the distributions of indicator values for units in each class. Like Gaussian mixture models, it is a generative model for how the indicator values arise from class membership. Like all the methods in this chapter, it is *unsupervised*, the classes that emerge are those that best predict variation in the indicator values. This may prove to be a useful classification of the units for your measurement purposes, but it also may not.

14.2 Application - Predicting Clinical Diagnosis of Depression, Part 3

We can use a latent class model to revisit the example of the PHQ-9 depression instrument. Recall that an important goal of the PHQ-9 instrument was to aid in classification of individuals who qualify as clinically depressed. We might hope that Latent Class Analysis, with two latent classes, could help us do something like this. If we estimate a two class latent class model[1] using the same data that we previously examined, we do generate a classification that has

[1] I use default settings in the R package poLCA (Linzer and Lewis, 2011).

roughly the right properties.

Table 14.5: Estimated probability of each level for each indicator for each class, based on a two class latent class model for PHQ-9 indicator data.

Indicator	Class	Pr.1.	Pr.2.	Pr.3.	Pr.4.
I1	1	0.31	0.42	0.14	0.13
I1	2	0.90	0.07	0.02	0.01
I2	1	0.27	0.49	0.13	0.11
I2	2	0.94	0.05	0.01	0.00
I3	1	0.24	0.35	0.17	0.24
I3	2	0.76	0.19	0.03	0.02
I4	1	0.09	0.44	0.21	0.26
I4	2	0.62	0.32	0.04	0.02
I5	1	0.39	0.34	0.13	0.15
I5	2	0.87	0.11	0.02	0.01
I6	1	0.47	0.33	0.10	0.10
I6	2	0.97	0.03	0.00	0.00
I7	1	0.53	0.27	0.09	0.11
I7	2	0.95	0.04	0.00	0.00
I8	1	0.67	0.20	0.06	0.06
I8	2	0.98	0.02	0.00	0.00
I9	1	0.87	0.09	0.02	0.02
I9	2	1.00	0.00	0.00	0.00

When you estimate this model, it yields probabilities for each response level on each indicator, given membership in each class. Table 14.5 shows the probabilities for all four response levels, for each indicator, for each latent class. When we examine these, we discover that "class 1" almost always gives the lowest response option ("never") on every item, while "class 2" gives higher response options with much higher probability. In essence, individuals put in class 1 are those with very few symptoms of depression, while individuals with more than a very few symptoms are all classified into class 2.

Another way to see what the model is doing, is to look at the predicted classification for each individual, and compare these to their PHQ-9 scores (on the standard 0-27 scale). Figure 14.1 shows that the two classes are very well separated in terms of the PHQ-9 scores that they represent, with a threshold in the 4-6 point range. The fact that we find this close relationship reflects facts about these data that we already discussed in Chapter 12 when we used them to illustrate item response models. All of the indicators are positively associated with one another: an item response model describes that as reflecting positions towards one end of a latent dimension as opposed to the other while a latent class model describes that as reflecting membership in one class rather than another.

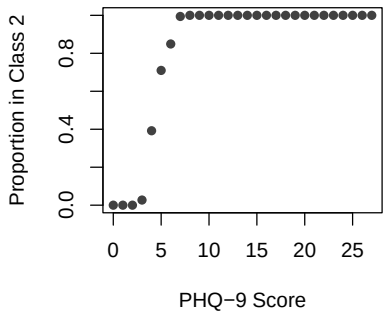


Figure 14.1: Proportion of individuals in class with more depressive symptoms, as a function of PHQ-9 score.

As noted in Chapter 12 when this data set was introduced, the best meta-analyses of PHQ-9 scores' relationship to clinical diagnosis of depression puts the threshold at 9. The threshold that the latent class model "finds" is substantially lower than the threshold validated to best predict clinical diagnosis (a score of 9). This is because the latent class model has no information about clinical diagnosis: it is an unsupervised rather than a supervised measurement strategy. Latent class models are simply trying to explain variation in the indicators. Based on the model that there are two groups in the data set, each with a different set of probabilities for each level of each indicator, the model tries to predict as much variation as possible with the most efficient set of classifications. This yields a somewhat higher number of people classified to the "depressive" class than would be medically diagnosed as such.

14.3 Application - Patterns of Survey Responses Regarding Taxation

Latent class analysis is most useful for summarising response patterns in categorical response data which have a more complex structure than that in the preceding example. In particular, this is useful in survey data where there are partially ordinal response scales, such as in cases where there is a "Don't Know" or similar option in addition to a response scale running from "Strongly Agree" to "Strongly Disagree" or similar.

One such application is presented by Lucy Barnes in a paper entitled "Taxing the rich: public preferences and public understanding" (Barnes, 2021) Figure 14.2 shows the main results of a seven class latent class analysis of six survey questions with six response categories each. The questions assess the extent to which UK respondents endorse zero-sum and positive-sum expressions of ideas around conflict for resources, the purpose of economics, and the labour market. This creates six survey questions, in pairs that are in some logical conflict with one another (zero sum versus positive sum). As with many survey batteries, the that that there are items where we might expect someone with consistent views to give "agree" responses as well as "disagree" responses is an intentional design choice.

- Zero sum: conflict - "If someone gets richer, somebody else gets poorer"
- Zero sum: purpose - "For every new person that starts working, there is one less job for the existing workforce"
- Zero sum: labour market - "Economics and economic policy are about distributing existing resources"
- Positive sum: conflict - "If someone gets richer, it means that total wealth increases"
- Positive sum: purpose - "New workers increase demand for goods and services and so create jobs for others"
- Positive sum: labour market - "Economics and economic policy are about increasing the resources available"

Figure 14.2: Latent Class Analysis of six survey responses on indicators of zero-sum and positive-sum thinking about economics. Figure from Barnes 2021.

Barnes identifies four latent classes corresponding to responders who can be characterised in terms of the substance of the research question that motivated the survey. These are "a: weak positive sum", "b: left neoliberals", "c: class conflict", and "f: strong positive sum". We will return to these below, but before we do, it is important to note that there are also three latent classes that are more about how people respond to the the survey than about what they believe about economics: "d: acquiescers", "f: non-committal", "g: high don't knows". The acquiescers (10.0%) tend to strongly agree or agree with all the six survey items, even though this is inconsistent with respect to the economic principles being queried. The non-committal respondents (8.7%) give the middle response on nearly all items. The high don't knows (6%) give the don't know response to nearly all items. These are distinct response patterns, but it isn't clear they really correspond to substantive views about the economic questions that are the target of the study. Rather, they may reflect three distinct ways that people who are not that interested in the survey quickly answer a series of questions they don't care about: agree to everything, give the intermediate response to everything, or say you don't know to everything. Collectively, these patterns characterise about 1/4 of the respondents to the survey.

The other four classes are the ones that speak to the research question about economic attitudes. These four classes are not unordered in the way the three "response types" do, they logically run from those who are more inclined towards zero sum understandings and away from positive sum understandings towards those who take the opposite views. That is, from "c: class conflict" (15.8%) on one end through "b: left neoliberals" (23.9%) and "a: weak positive sum" (28.6%) to "f: strong positive sum" (6.9%). If you carefully examine the response patterns for these groups in the figure, you will see that the typical responses shift across the response options in a consistent direction as you move across the four classes in the order that I have listed them.

14.4 When Should We Measure Categorical Quantities as Opposed to Continuous Quantities?

The fact that these four classes are ordered means that if we were instead to fit an item response model to the data for these respondents, we would estimate a continuous scale measure that would tend to array these four groups in this order. Whether we describe this variation using a latent scale or latent classes is up to us as analysts, we cannot assess which model is better on the basis of model fit. The question is what is a more useful way of summarising the patterns of response. However, what the latent scale model would typically struggle to do is to provide such an easily interpretable presentation of the three response classes that do not tidily fit into the dimension that characterises the other four, and their presence might muddy some analyses that seek to understand where different demographic and political groups fall on these questions.

This example highlights that there are some situations where it is not clear whether we should seek to measure a continuous quantity or a categorical quantity. In this application of a categorical model, latent class analysis, we see evidence of a continuum from the political left to the political right. The latent class model has no way to describe this other than by setting out a sequence of (in this case four) types from left to right, with response distributions shifting monotonically on each response question as you move across the types. In some sense, this is simply a poor approximation of an IRT model. In this application, the presence of other response types that are "off" this dimension makes the latent class approach relative attractive compared to a model with a continuous latent variable, but sometimes the application of latent class analysis will yield types that *all* seem to be embedded in a simple one or two dimensional continuous space. In such cases, should one switch from latent class analysis to an item response model with continuous latent variables?

Assume here that we are working in a situation where there is not an entirely clear answer based on theoretical grounds. Ideally the answer should be dictated by the conceptualisation of the quantity you want to measure, but sometimes one could just as easily describe a continuum or a set of types.

In the above example, patterns in the response variables do help. For example, if you really want to think of everyone as having a position on a left-right ideological scale, but you discover that some people say they do not know to everything and others just give the middle response to everything, and other always give the first response option even though sometimes that is the most left-wing view and sometimes that is the most right-wing view, you may conclude that this is not a useful conceptualisation for describing variation in the sample you are working with. Not everyone seems to have a position on a left-right scales that is driving their responses: this is true for many respondents, but not for all respondents.[2]

[2] Often, the groups whose responses seem to be driven by other considerations (eg indifference to the survey) will end up being placed in the middle of the scale, as they are clearly not consistently left-wing or right-wing. This is usually a reasonable placement, within the logic of a continuous measure, but you then need to be very careful to recognise that the centre of your scale jumbles up a variety of types of "moderates".

15
Unsupervised Mixture Measurement

To be written…

This chapter will introduce Latent Dirichlet Allocation (Blei et al., 2003) and related methods. These unsupervised measurement methods for latent variables on a simplex, where the units of interest (typically documents) are described as a mixture of categories, with weights that add to one. These models share some features with classification (measurement of a categorical latent variable) and some with scaling (measurement of a continuous latent variable), which were covered in previous chapters.

16
Multilevel Measurement Models

To be written…

This chapter will explore the use of multilevel data and multilevel models in measurement. This will include several kinds of measurement strategies. First, methods combine evidence from multiple scales or multiple raters into a consensus scale. Second, methods such as multilevel regression and post-stratification that use individual-level data to measure attributes of aggregate units (often, but not necessarily, geographic areas).

17
Structural Measurement Models

To be written…

This chapter will explore the use of structural models of behaviour or choice that connect observable data to target quantities to be measured. This will start by discussing standard methods for reducing social desirability bias such as randomised response models and list experiments. Then, the chapter will develop connections between models of choice and methods for measuring continuous variables introduced in previous chapters. Finally, the chapter will discuss general strategies and direct readers towards relevant modelling and estimation techniques that are necessary to develop custom measurement strategies for particular problems.

18
Missing Indicators and Comparability

To be written…

This chapter will introduce standard terminology for discussing missing data. The concepts of differential item functioning and measurement equivalence will be introduced and applied to examples. Strategies for managing missingness, and their limitations, will be introduced and discussed.

19
Conclusion

To be written…

This chapter will gesture vaguely at contemporary developments. In particular, I will discuss the potential of machine learning and AI methods for helping us measure things we want to measure. I will also discuss the peril these approaches present in distracting us from being clear about what it is we are trying to measure and what it would mean to measure well versus poorly.

20
Bibliography

Achen, C. H. (1978). Measuring representation. *American Journal of Political Science*, pages 475–510.

Achen, C. H. and Blais, A. (2015). Intention to vote, reported vote and validated vote 1. In *The Act of Voting*, pages 195–209. Routledge.

Alkire, S. (2013). Choosing dimensions: The capability approach and multidimensional poverty. In *The many dimensions of poverty*, pages 89–119. Springer.

Alkire, S. and Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of public economics*, 95(7-8):476–487.

Alkire, S. and Santos, M. E. (2010). Acute multidimensional poverty: A new index for developing countries. *United Nations development programme human development report office background paper*, (2010/11).

Alvarez, M., Cheibub, J. A., Limongi, F., and Przeworski, A. (1996). Classifying political regimes. *Studies in Comparative International Development*, 31(2):3–36.

Anckar, C. and Fredriksson, C. (2019). Classifying political regimes 1800–2016: a typology and a new dataset. *European Political Science*, 18(1):84–96.

Ansolabehere, S. and Hersh, E. (2012). Validation: What big data reveal about survey misreporting and the real electorate. *Political Analysis*, 20(4):437–459.

Bafumi, J. and Herron, M. C. (2010). Leapfrog representation and extremism: A study of american voters and their members in congress. *American Political Science Review*, 104(3):519–542.

Barnes, L. (2021). Taxing the rich: public preferences and public understanding. *Journal of European Public Policy*, pages 1–18.

Barocas, S., Hardt, M., and Narayanan, A. (2020). Fairness and machine learning: Limitations and oppurtunities.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*, volume 904. John Wiley & Sons.

Bartholomew, D. J., Steele, F., Galbraith, J., and Moustaki, I. (2008). *Analysis of multivariate social science data*. Chapman and Hall/CRC.

Benzecri, J. (1973). L'analyse des données volume ii. *L'Analyse des Correspondances: Paris, France Dunod-Paris, France Dunod*.

Blalock, H. M. (1982). *Conceptualization and measurement in the social sciences*. Number 04; H61, B5.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Bueno de Mesquita, E. (2019). Quantification shapes how we think about public policy—often for the worse.

Cameron, E., Nuzzo, J., and Bell, J. (2019). Global health security index: Building collective action and accountability. Technical report, Nuclear Threat Initiative and Johns Hopkins Bloomberg School of Public Health.

Campbell, D. T. (1976). Assessing the impact of planned social change. *Journal of Multidisciplinary Evaluation*, page 34.

Chang, H. (2004). *Inventing temperature: Measurement and scientific progress*. Oxford University Press.

Choi, S.-S., Cha, S.-H., and Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1):43–48.

Clifford, S. and Jerit, J. (2018). Disgust, anxiety, and political learning in the face of threat. *American Journal of Political Science*, 62(2):266–279.

Clinton, J., Jackman, S., and Rivers, D. (2004). The statistical analysis of roll call data. *American Political Science Review*, 98(2):355–370.

Cobham, A., Sumner, A., Cornia, A., Dercon, S., Engberg-pedersen, L., Evans, M., Lea, N., Lustig, N., Manning, R., Milanovic, B., et al. (2013). Putting the gini back in the bottle¿the palma'as a policy-relevant measure of inequality.

Cox, D. R. (1969). *Analysis of Binary Data*. London: Chapman and Hall.

Cox, G. W., Fiva, J. H., and Smith, D. M. (2020). Measuring the competitiveness of elections. *Political Analysis*, 28(2):168–185.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334.

Dalton, R. J. (2008). The quantity and the quality of party systems: Party system polarization, its measurement, and its consequences. *Comparative Political Studies*, 41(7):899–920.

de Mesquita, E. B. (2016). *Political economy for public policy*. Princeton University Press.

Desposato, S. W. (2005). Correcting for small group inflation of roll-call cohesion scores. *British Journal of Political Science*, 35(4):731–744.

Dunleavy, P. and Boucek, F. (2003). Constructing the number of parties. *Party Politics*, 9(3):291–315.

Fieldhouse, E., Green, J., Evans, G., Schmitt, H., van der Eijk, C., Mellon, J., and Prosser, C. (2016). British election study, 2015: Face-to-face post-election survey. *UK Data Service*.

Fieldhouse, E., Green, J., Evans, G., Schmitt, H., van der Eijk, C., Mellon, J., and Prosser, C. (2018). British election study, 2017: Face-to-face post-election survey. *UK Data Service*.

Floridi, G. and Lauderdale, B. E. (2018). Using expert judgements to measure productive ageing in italy and south korea. *Working Paper*.

Foster, J., Greer, J., and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica: journal of the econometric society*, pages 761–766.

Freedle, R. (2003). Correcting the sat's ethnic and social-class bias: A method for reestimating sat scores. *Harvard Educational Review*, 73(1):1–43.

Gallagher, M. (1991). Proportionality, disproportionality and electoral systems. *Electoral studies*, 10(1):33–51.

Gelman, A., Katz, J. N., and Bafumi, J. (2004). Standard voting power indexes do not work: an empirical analysis. *British Journal of Political Science*, 34(4):657–674.

Gelman, A. and Nolan, D. (2002). You can load a die, but you can't bias a coin. *The American Statistician*, 56(4):308–311.

Gibbons, R. D., Hooker, G., Finkelman, M. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., and Kupfer, D. J. (2013). The cad-mdd: A computerized adaptive diagnostic screening tool for depression. *The Journal of clinical psychiatry*, 74(7):669.

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., and Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *Archives of general psychiatry*, 69(11):1104–1112.

Goertz, G. (2020). *Social Science Concepts and Measurement: New and Completely Revised Edition*. Princeton University Press.

Golder, M. and Stramski, J. (2010). Ideological congruence and electoral institutions. *American Journal of Political Science*, 54(1):90–106.

Golosov, G. V. (2010). The effective number of parties: A new approach. *Party politics*, 16(2):171–192.

Gould, S. J. (1996). *The Mismeasure of Man*. W W Norton & Company.

Graefe, A. (2015). Improving forecasts using equally weighted predictors. *Journal of Business Research*, 68(8):1792–1799.

Green, P. E. and Rao, V. R. (1971). Conjoint measurement-for quantifying judgmental data. *Journal of Marketing research*, 8(3):355–363.

Grimmer, J. and King, G. (2011). General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, 108(7):2643–2650.

Gygli, S., Haelg, F., Potrafke, N., and Sturm, J.-E. (2018). The kof globalisation index–revisited. *The Review of International Organizations*, pages 1–32.

Hainmueller, J. and Hopkins, D. J. (2015). The hidden american immigration consensus: A conjoint analysis of attitudes toward immigrants. *American Journal of Political Science*, 59(3):529–548.

Hainmueller, J., Hopkins, D. J., and Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*, 22(1):1–30.

Hand, D. J. (1996). Statistics and the theory of measurement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 159(3):445–473.

Hand, D. J. (2016). *Measurement: A Very Short Introduction*. Oxford University Press.

Hanretty, C. (2017). Areal interpolation and the uk's referendum on eu membership. *Journal of Elections, Public Opinion and Parties*, 27(4):466–483.

Harder, N., Figueroa, L., Gillum, R. M., Hangartner, D., Laitin, D. D., and Hainmueller, J. (2018). Multidimensional measure of immigrant integration. *Proceedings of the National Academy of Sciences*, 115(45):11483–11488.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction, springer series in statistics.

Healy, K. (2017). Review of engines of anxiety: Academic rankings, reputation, and accountability. *European Journal of Sociology*, 58(3):512–519.

Hirschman, A. O. (1964). The paternity of an index. *The American economic review*, 54(5):761–762.

Hopkins, D. and Noel, H. (2021). How Trump Has Redefined Conservatism. *FiveThirtyEight*.

Horst, A. M., Hill, A. P., and Gorman, K. B. (2020). *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0.

Izbicki, M. (2011). "how to create an unfair coin and prove it with math".

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Kay, J. and King, M. (2020). *Radical Uncertainty: Decision-Making Beyond the Numbers*. WW Norton & Company.

King, G., Tomz, M., and Wittenberg, J. (2000). Making the most of statistical analyses: Improving interpretation and presentation. *American journal of political science*, pages 347–361.

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Koptjevskaja-Tamm, M. (2015). *The linguistics of temperature*, volume 107. John Benjamins Publishing Company.

Krantz, D., Luce, D., Suppes, P., and Tversky, A. (1971). *Foundations of measurement: Additive and polynomial representations*, volume 2. Courier Corporation.

Laakso, M. and Taagepera, R. (1979). "effective" number of parties: a measure with application to west europe. *Comparative political studies*, 12(1):3–27.

Laderchi, C. R., Saith, R., and Stewart, F. (2003). Does it matter that we do not agree on the definition of poverty? a comparison of four approaches. *Oxford development studies*, 31(3):243–274.

Lauderdale, D. S., Chen, J.-H., Kurina, L. M., Waite, L. J., and Thisted, R. A. (2016). Sleep duration and health among older adults: associations vary by how sleep is measured. *J Epidemiol Community Health*, 70(4):361–366.

Laver, M. and Benoit, K. (2015). The basic arithmetic of legislative decisions. *American Journal of Political Science*, 59(2):275–291.

Lerman, A. E. (2009). The people prisons make: Effects of incarceration on criminal psychology. *Do prisons make us safer*, pages 151–176.

Levis, B., Benedetti, A., and Thombs, B. D. (2019). Accuracy of patient health questionnaire-9 (phq-9) for screening to detect major depression: individual participant data meta-analysis. *bmj*, 365:l1476.

Linzer, D. A. and Lewis, J. B. (2011). polca: An r package for polytomous variable latent class analysis. *Journal of Statistical Software*, 42(10).

Loewen, P. J., Rubenson, D., and Spirling, A. (2012). Testing the power of arguments in referendums: A bradley–terry approach. *Electoral Studies*, 31(1):212–221.

Luce, R. D., Suppes, P., and Krantz, D. H. (2007). *Foundations of measurement: representation, axiomatization, and invariance*, volume 3. Courier Corporation.

Lundberg, I., Johnson, R., and Stewart, B. M. (2021). What is your estimand? defining the target quantity connects statistical evidence to theory. *American Sociological Review*, 86(3).

MacKenzie, D. A. (1981). *Statistics in Britain: 1865-1930; the social construction of scientific knowledge*. Edinburgh University Press.

Manski, C. F. (2011). Genes, eyeglasses, and social policy. *Journal of Economic Perspectives*, 25(4):83–94.

Martin, A. D. and Quinn, K. M. (2002). Dynamic ideal point estimation via markov chain monte carlo for the us supreme court, 1953–1999. *Political Analysis*, 10(2):134–153.

Matsusaka, J. G. (2015). 'responsiveness' as a measure of representation. *USC CLASS Research Paper No. CLASS15-19*.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.

Mehrens, W. A. and Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement: issues and practice*, 8(1):14–22.

Miilunpalo, S., Vuori, I., Oja, P., Pasanen, M., and Urponen, H. (1997). Self-rated health status as a health measure: the predictive value of self-reported health status on the use of physician services and on mortality in the working-age population. *Journal of clinical epidemiology*, 50(5):517–528.

Miller, F. (2017). Aristotle's political theory. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2017 edition.

Molinar, J. (1991). Counting the number of parties: an alternative index. *American Political Science Review*, 85(4):1383–1391.

Morgan, S. L. and Winship, C. (2015). *Counterfactuals and causal inference*. Cambridge University Press.

Muller, J. Z. (2018). *The Tyranny of Metrics*. Princeton University Press.

Munck, G. L. and Verkuilen, J. (2002). Conceptualizing and measuring democracy: Evaluating alternative indices. *Comparative political studies*, 35(1):5–34.

Murphy, R. and Wyness, G. (2020). Minority report: the impact of predicted grades on university admissions of disadvantaged groups. *Education Economics*, pages 1–18.

Murray, C. J. (1994). Quantifying the burden of disease: the technical basis for disability-adjusted life years. *Bulletin of the World health Organization*, 72(3):429.

Ogwang, T. (1994). The choice of principle variables for computing the human development index. *World Development*, 22(12):2011–2014.

Penrose, L. S. (1946). The elementary statistics of majority voting. *Journal of the Royal Statistical Society*, 109(1):53–57.

Porter, T. M. (2020). *Trust in numbers*. Princeton University Press.

Prosser, C., Fieldhouse, E., Green, J., Mellon, J., and Evans, G. (2020). Tremors but no youthquake: Measuring changes in the age and turnout gradients at the 2015 and 2017 british general elections. *Electoral Studies*, page 102129.

Rao, P. and Kupper, L. L. (1967). Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204.

Renwick, A. (2015). Electoral disproportionality: What is it and how should we measure it?

Rice, S. A. (1928). Quantitative methods in politics.

Rodamar, J. (2018). There ought to be a law! campbell versus goodhart. *Significance*, 15(6):9–9.

Rubenstein, J. C. (2016). The lessons of effective altruism. *Ethics & International Affairs*, 30(4):511–526.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*.

Santelices, M. V. and Wilson, M. (2010). Unfair treatment? the case of freedle, the sat, and the standardization approach to differential item functioning. *Harvard Educational Review*, 80(1):106–134.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Shapley, L. S. and Shubik, M. (1954). A method for evaluating the distribution of power in a committee system. *American political science review*, 48(3):787–792.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software*, 39(5):1–13.

Simpson, E. H. (1949). Measurement of diversity. *Nature*, 163(4148):688.

Singer, P. (2009). *The Life You Can Save: Acting Now to End World Poverty*. Random House.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Crc Press.

Sparrow, M. K. et al. (2015). Measuring performance in a modern police organization. *Psychosociological Issues in Human Resource Management*, 3(2):17–52.

Spirling, A. and McLean, I. (2007). Uk oc ok? interpreting optimal classification scores for the uk house of commons. *Political Analysis*, 15(1):85–96.

Steinley, D. and Brusco, M. J. (2008). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika*, 73(1):125.

Stevens, S. S. et al. (1946). On the theory of scales of measurement.

Suppes, P. and Krantz, D. H. (2007). *Foundations of measurement: Geometrical, threshold, and probabilistic representations*, volume 2. Courier Corporation.

Taagepera, R. and Grofman, B. (2003). Mapping the indices of seats–votes disproportionality and inter-election volatility. *Party Politics*, 9(6):659–677.

Thomson, G. H. (1916). A hierarchy without a general factor. *British Journal of Psychology*, 8(3):271.

Tukey, J. W. (1986). Sunset salvo. *The American Statistician*, 40(1):72–76.

Upton, G. and Cook, I. (2014). *A dictionary of statistics 3e*. Oxford university press.

Wigell, M. (2008). Mapping 'hybrid regimes': Regime types and concepts in comparative politics. *Democratisation*, 15(2):230–250.

World Bank (2018). *Poverty and shared prosperity 2018: piecing together the poverty puzzle*.

Zeckhauser, R. and Shepard, D. (1976). Where now for saving lives. *Law & Contemp. Probs.*, 40:5.

Zucco Jr, C., Batista, M., and Power, T. J. (2019). Measuring portfolio salience using the bradley–terry model: An illustration with data from brazil. *Research & Politics*, 6(1):2053168019832089.

21
Index