Week 2: Measurement Theory and Error

POLS0013 Measurement in Data Science

Dr. Julia de Romémont

Academic Year 24-25

UCL Departement of Political Science

Assessment I ('Essay') Details

- I hope you all read the information about the first assessment...
- 1500 words report on a measure that is *socially relevant* in some capacity
  - Avoid measures that are (extensivley) discussed in the lecture or book, unless you really think you have something new to add
  - Avoid the obvious ones, e.g. Consumer Price Index (CPI), Human Development Index (HDI), Global Gender Gap, unless you really think you have something new to add
- It should include the following, equally weighted, elements
  1. A description of the measure
  2. A critiqual assessment of the measure
  3. A concrete proposal on how to improve the measure

1. Description of the measure

Every measure has three core elements, which you should clearly summarise.

1. The purpose (i.e. the target concept) of the measure
    - i.e. what is it trying to measure?
2. The indicators included in the measure
    - i.e. what are the building blocks of the measure?
3. The aggregation procedure
    - i.e. how the indicators are combined into creating the measure

2. Critiqual assessment

When critiquing the measure you should think: *how good is the measure at capturing the target concept?* Specifically, this means you should discuss the following:

1. Overall, whether the measure is a ***useful* summary** of the target concept
2. What is the most important potential problem(s) of the measure, i.e. in what way might it miss its target concept?
    - What source(s) of measurement error might exist?
    - What kind of measurement error? (c.f. today's lecture)

- I strongly recommend using examples to illustrate the measurement error you identified!

3. Concrete proposal for improvement

- Your *proposal* should address the issue(s) you raised in the previous part!
  - The proposed change could be big or small, simple or complex, it needs to make sense given the most important potential problem you identified!
  - Clearly discuss how and why your proposed issue would address said issue
- Your proposal has to (only) be theoretically doable, remember it's a written task only!
- But do think carefully about and discuss the potential practicalities of implementation (e.g. costs, feasibility etc)
- Note that often addressing one issue means creating more error in another respect. Being aware of (and mentioning) these trade-offs is good!

Assessment I ('Essay') Details

- Where to find a measure to discuss?
  - Other modules, readings...
  - News, radio, podcasts...
  - Daily life...
- Remember that you can come talk to me about your idea(s) in SSF hours at any point during term!

On the use of AI in the assessment

- **Unauthorised** or **unreferenced authorised** use of ChatGPT (&others) constitutes academic misconduct.
  - Breaches of any academic misconduct rules are unacceptable and will lead to a lot of stress and unpleasentness for you
- **Under no circumstances** should you upload any course material to ChatGPT or other other GenAI tools.
- The Department has developed guidelines specific to quantitative methods courses which you should read and can find here. tl;dr: You can use AI tools in the assignment but
  - Only for certain tasks; and
  - With appropriate acknowledgment and referencing

On the use of AI in the assessment

... Only for certain tasks:

- Coding: To **correct errors** in your code and to **improve on the appearance** of tables and figures.
- Writing: To help improve your writing, including greater clarity or more accurate grammar.
- Generally: To support your efforts to resolve conceptual queries, although you should always make use of your classes, support and feedback hours, and moodle forums first.

This means you cannot use it:

- To write parts or all of an assessment;
- To write parts or all of your code;
- To generate outlines, structures and high-level arguments for essays;
- For rewriting or paraphrasing text from other sources for use in written work.

On the use of AI in the assessment

... With appropriate acknowledgement, description and referencing:

Acknowledgement

I acknowledge the use of ChatGPT (https://chat.openai.com) to improve my code and my writing.

Description

I used ChatGPT in the coding of figure XX. The prompt I entered was: "How can I add a horizontal dotted line where the Y-axis is zero?", and ChatGPT suggested the following code: `geom_hline(yintercept = 0, linetype="dotted")`.

I used ChatGPT to improve the grammar of the final version of my essay. The prompt I entered was: "Can you let me know where the grammar and clarity of writing could be improved in these paragraphs?"

References

OpenAI (2024). ChatGPT (September version). GPT-4o Large Language model, https://chat/.openai.com/chat.

Last lecture...

...we talked about Measurement

- Measurement *inference* defined as conclusions drawn from observed data to unmeasured (latent) quantities describing the same units
- This whole course will be about *how* to make such connections *credibly*
- We established some vocabulary we will use to talk about measurement this term

This lecture is about...

...Measurement Error

- Making *inferences* implies that the thing we draw conclusions *with* is not the same as the thing we draw conclusions *about*
- In other words, we need to think systematically about the difference between the thing we want to measure and the measure
- Measures can be 'wrong' in different ways. They can be
  - biased
  - unreliable
  - miscalibrated
- Deeper issues arise when measurement error is correlated with **other characteristics**! $\Rightarrow$ problem of (un)fairness

This Lecture

Defining measurement error

Measurement error and fairness

Measurement error and subsequent analyses

Defining measurement error

Notation clarification

Roman/Latin letters

We will use letters like a, b, c, etc) as denoting **data/known quantities**

Greek letters

Letters like α, β, γ, etc will represent **parameters/unknown quantities**

- Simple linear regression equation follows this convention:

  $$y = \alpha + \beta x + \epsilon$$

m and μ

- m is the measure we have
- μ is the thing that we wanted to measure (the target concept)

Describing differences between m and μ

- Let's define measure m as:

$$m = \mu + \epsilon_m$$

- Measurement error is (usually[1]) defined as the difference between the measure we have m and the thing we wanted to measure μ:

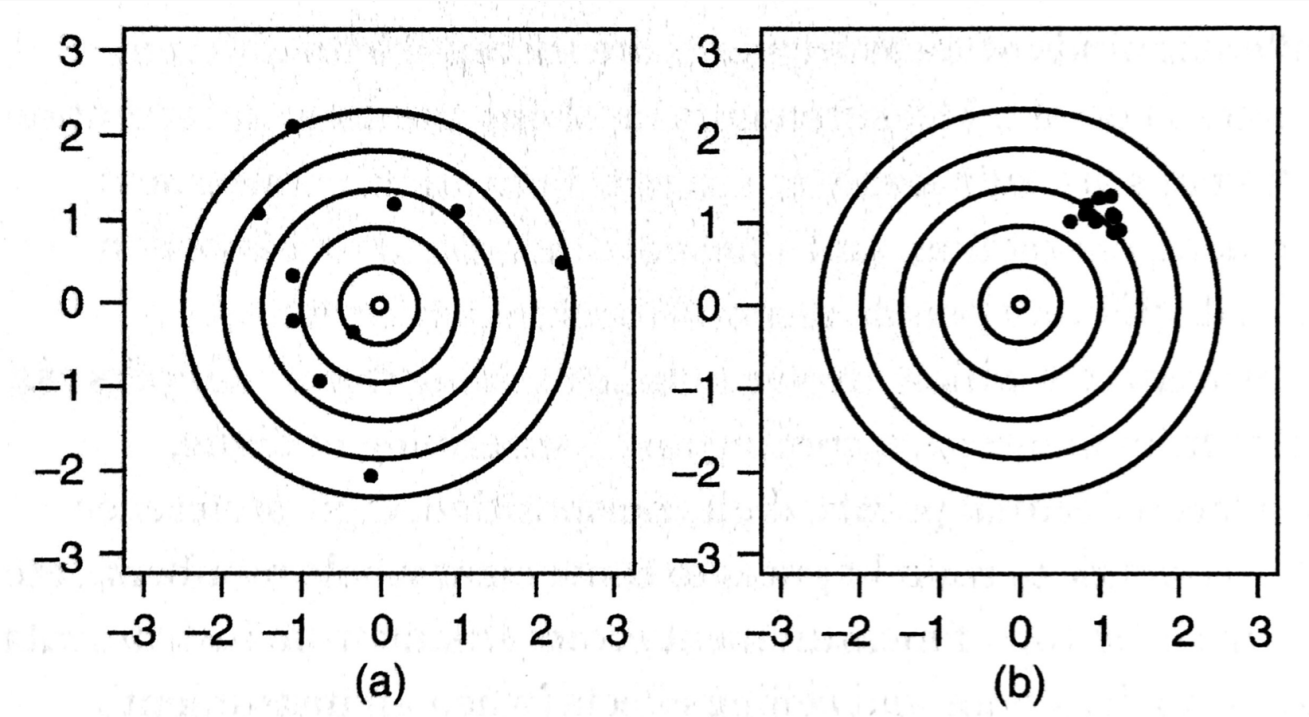
$$\epsilon_m = m - \mu$$

- Important: But we (usually) don't know what the measurement error is. What, then, makes a good measure? Or a bad measure?
  - To talk about what makes a good measurement we need to think about expectations, rather than specific values of these quantities

[1] For interval scales

Describing measurement error

- There are multiple ways that measure m might "go wrong" as a representation (or: summary) of the target concept μ

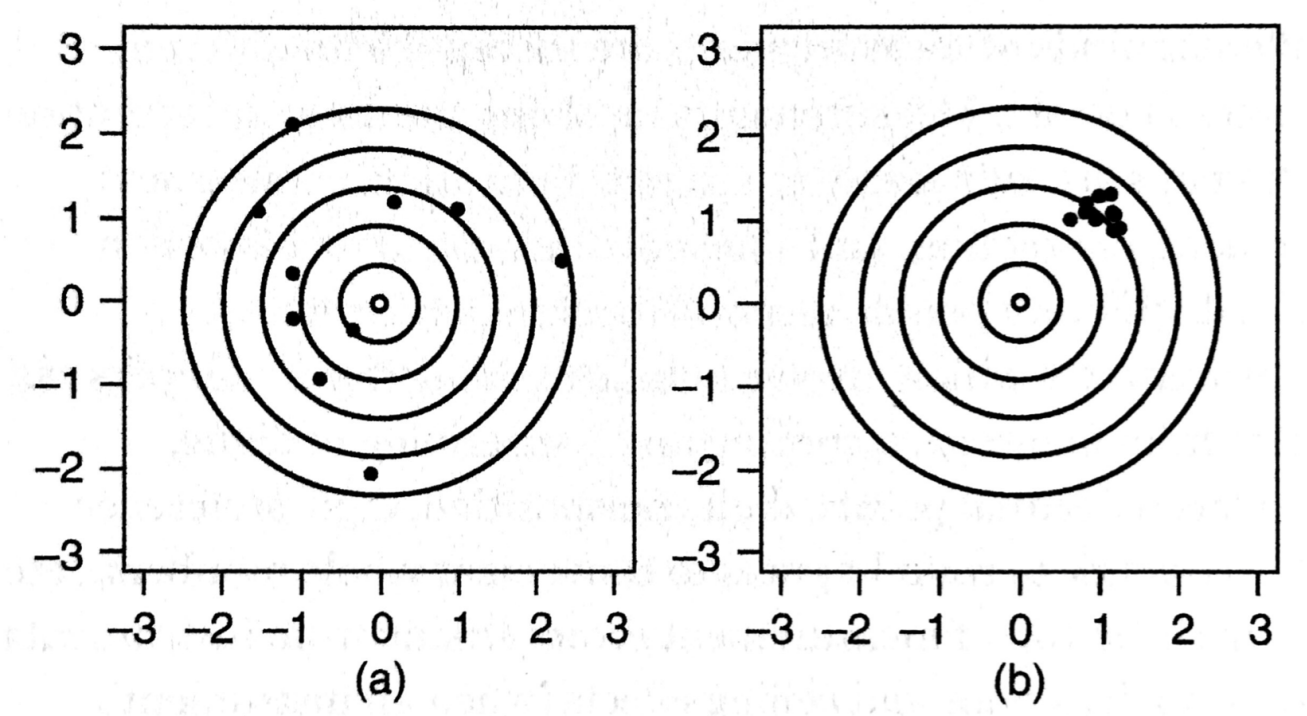


- A variety of, comparable but not necessarily interchangeable, terms exist to talk about these

Synonyms?

1. Terms that are used to describe how good your measure is at 'hitting' the target concept overall
    - Accuracy
    - Validity
2. Terms that are used to describe whether you measured the right thing, on average
    - Bias
3. Terms that are used to describe whether you would get the same answer if you measured it again
    - Variance
    - Precision
    - Reliability

Ways to (not) hit the target



Note that validity and accuracy do not actually refer to any specific type of measurement error, even though they are often (incorrectly) used as synonyms for bias (hence the "ish"'s).

Measuring measurement error

- There are different metrics that can be used to describe measurement error mathematically[2]
- Metrics *associated* with 'validity' and 'accuracy': mean absolute error (MAE), mean square error (MSE), or root mean square error (RMSE)
- Bias and variance have clear mathematical formulations and therefore specific statistics that can be used to measure them:
  - Mean error = $\frac{1}{n}\sum_{i=1}^{n}(m_i - \mu)$ = $E\left[\epsilon_m\right]$ = Bias
  - Variance of the error = $\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})$ = $var\left[\epsilon_m\right]$ = Variance

[2]Today we only look at interval-level scales. We'll discuss measurement error in nominal scales in week 7.

Example: Subjective well-being

Subjective well-being measures

- "All things considered, how satisfied are you with your life as a whole these days?"
- "Taken all together, would you say that you are very happy, pretty happy, or not too happy?"

1. Are these **reliable** measures of subjective well-being?
    - If we ask the same person twice (test-retest reliability), do we get the same answer?
    - Previous research suggests that single item measures of subjective well-being tend to be correlated at about 0.5, and multi-item scales sometimes as high as 0.8[a]

[a] Krueger & Schkade (2008)

Example: Subjective well-being

Subjective well-being measures

- "All things considered, how satisfied are you with your life as a whole these days?"
- "Taken all together, would you say that you are very happy, pretty happy, or not too happy?"

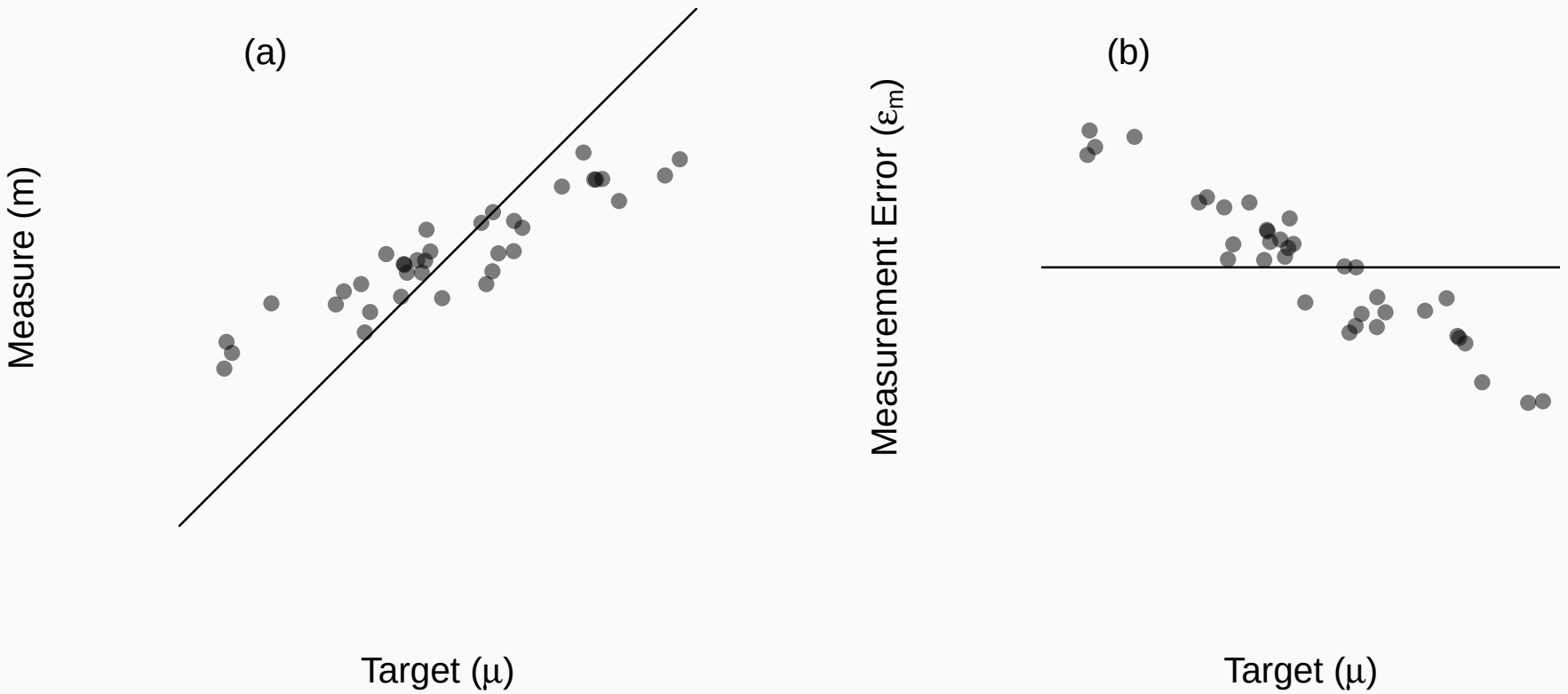
2. Are these **unbiased/valid** measures of subjective well-being?
    - Do these *actually* measure the concept of *subjective well-being*?
    - Are they *biased* by factors like the reporting context?
    - Does this concept even mean anything other than responses to this sort of question?

Bias, Variance *And* Miscalibration

Type	Quantity
Measurement Bias	$E[\epsilon_m] \neq 0$
Measurement Variance	$var[\epsilon_m] > 0$
Measurement Miscalibration	$cor[\epsilon_m, \mu] \neq 0$

- In addition to being wrong on average (bias), being noisy (variance), measures can also be *miscalibrated*
- Miscalibration means that measurement error is systematically larger at a one point of the scale compared to another

Miscalibration



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.

How much μ is in a measure m?

How can we characterise how much information about the target quantity is in a measure?

Correlation coefficient

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

Coefficient of determination

$$R^2 = \frac{cov(m, \mu)^2}{var(m)var(\mu)}$$

- These statistics ask: how much of the variance of the measure and target concept is *covariance*?
- Higher correlations imply stronger relationships between m and μ
- If $\rho = \sqrt{0.5} = 0.707$, then $R^2 = 0.5$, *half* the variation in the target concept is explained/captured/predicted by the measure

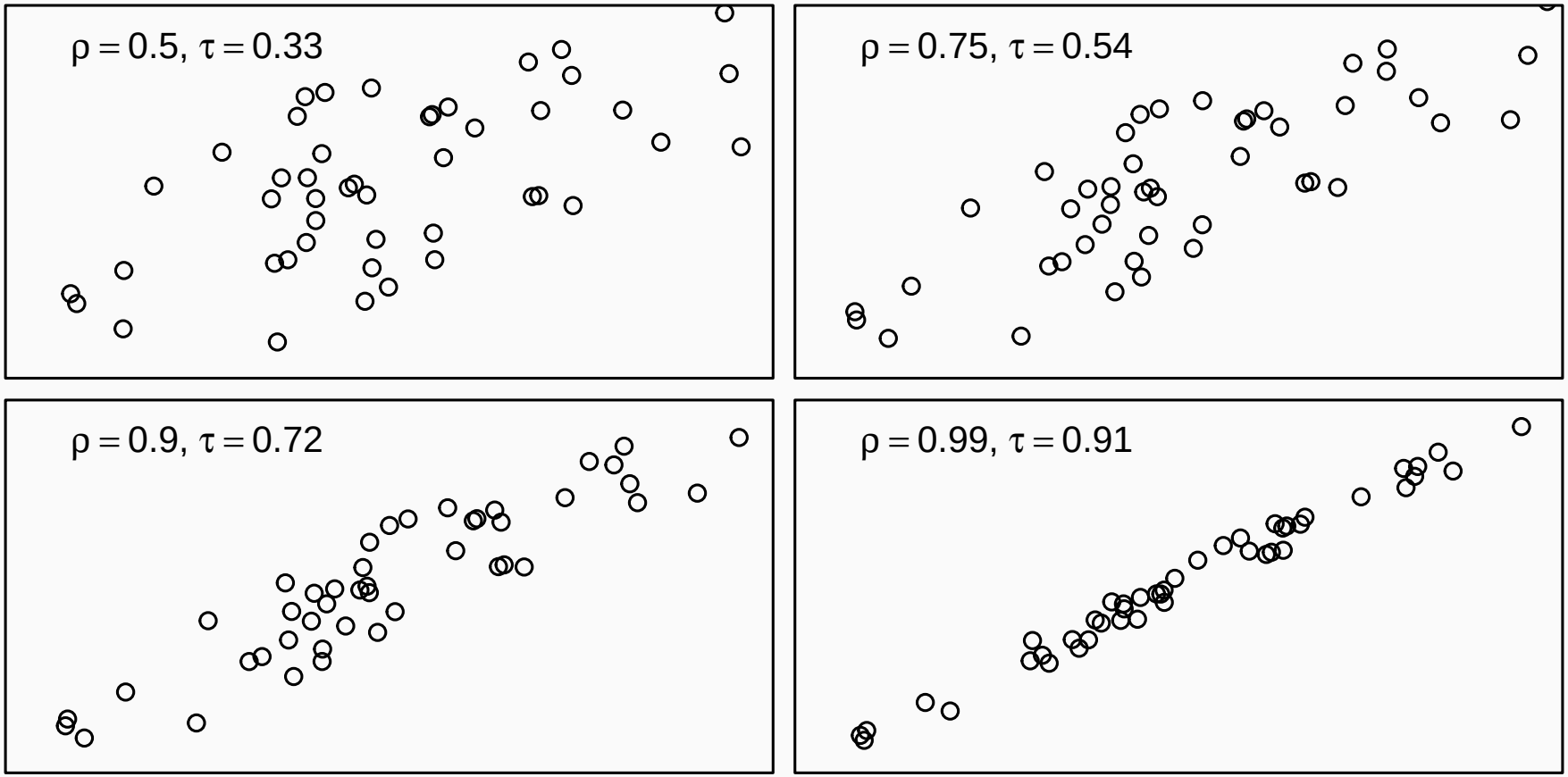
How much μ is in a measure m?

- Another correlation statistic that is sometimes used is Kendall's τ
- Kendall's τ is the proportion of pairwise comparisons between points that are ordered in the same direction

Kendall rank correlation coefficient

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \text{sgn}(x_i - x_j)\text{sgn}(y_i - y_j)$$

What qualifies as a good measure?



- Note that 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.

Correctly ordering individual units is hard

In 2020, due to the cancellation of A-level exams, UK teachers were asked to rank all their students in order, for use in a grade standardization algorithm.

- Something teachers had never been asked to do before
- Something that requires a very precise assessment of the relative strength of different students
- Potential for a lot of measurement error...
  - As long as measurement error was random/equally distributed, maybe not *too* much of a problem
  - But what if measurement error was systematically different for different types of students?

Measurement error and fairness

What makes a measure unfair?

- Simple differences in m by group do not tell us whether a measurement strategy is fair to different groups or not.
- Differences in average m for different groups X can reflect real differences in $\mu|X$, or could result from differences in measurement error $\epsilon_m|X$
- Defining fairness meaningfully requires attention to m, μ, **and the groups X.**

Desirable criterion 1: Separation

Separation

Measurement is independent of X, conditional on the true value of the concept of interest μ:

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

$=$ For units with the same true[3] value of the target concept μ, the distribution of the measurement m (and thus the measurement error ϵ_m) to be identical, *regardless* of X.

- E.g. the grades (=m) received by students with the *same* understanding of the material (μ) should not depend on (i.e. correlate with) their race/gender/background/etc.

[3] remember: but unknown!

Desirable criterion 2: Sufficiency

Sufficiency

The distribution of the true value of the concept of interest μ is independent of X, conditional on the measured value m:

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

$=$ Units with the same value of the measurement m should have the same true[4] value of the target concept μ, *regardless* of X. Knowing X should convey no further information about the likely value of μ once you know m.

- E.g. students with different race/gender/background/etc who get a 70 (m) should have the same understanding of the course material (μ).

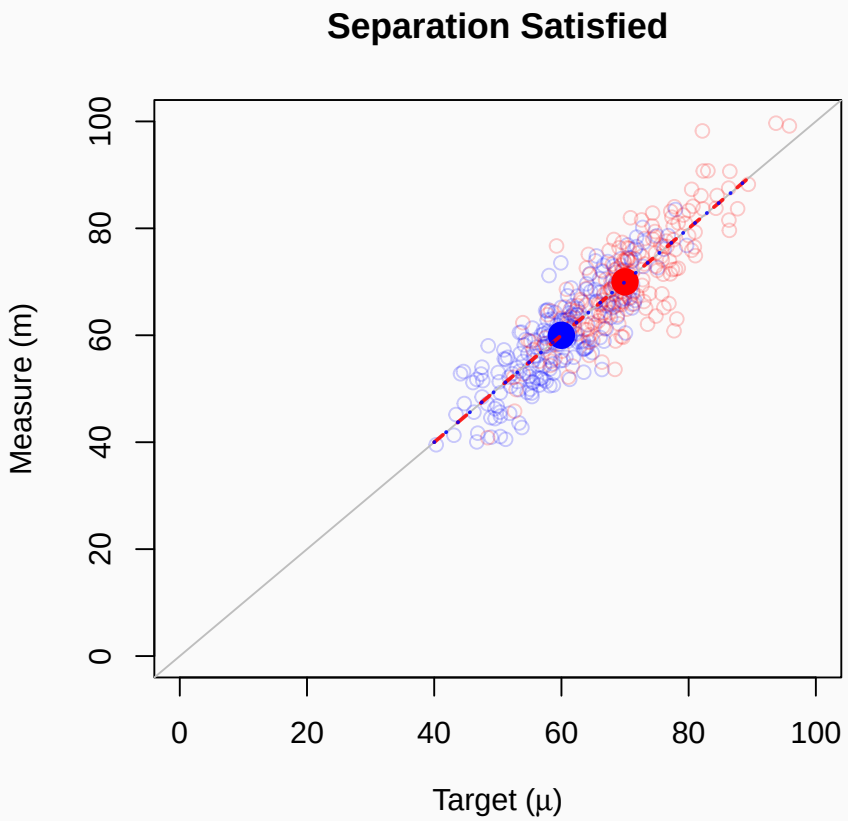
[4] remember: but unknown!

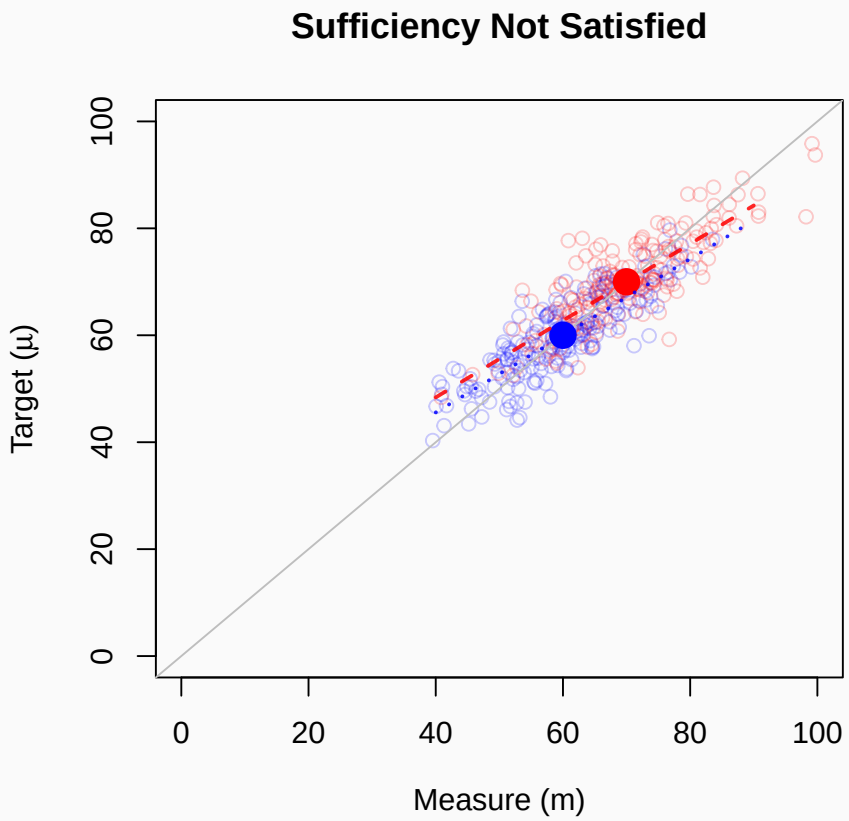
Separation vs Sufficiency

Separation amd sufficiency are usually incompatible

- If $p\left(\mu|m, X\right) = p\left(\mu|m\right)$ and $p\left(m|\mu, X\right) = p\left(m|\mu\right)$ then it is implied that $p\left(m, \mu|X\right) = p\left(m, \mu\right)$.
- If both sufficiency and separation are satisfied, then the joint distribution of m and μ can not depend on X.
- *But if there are real group differences in* μ, at least one of sufficiency or separation must not be satisfied.

Separation vs sufficiency

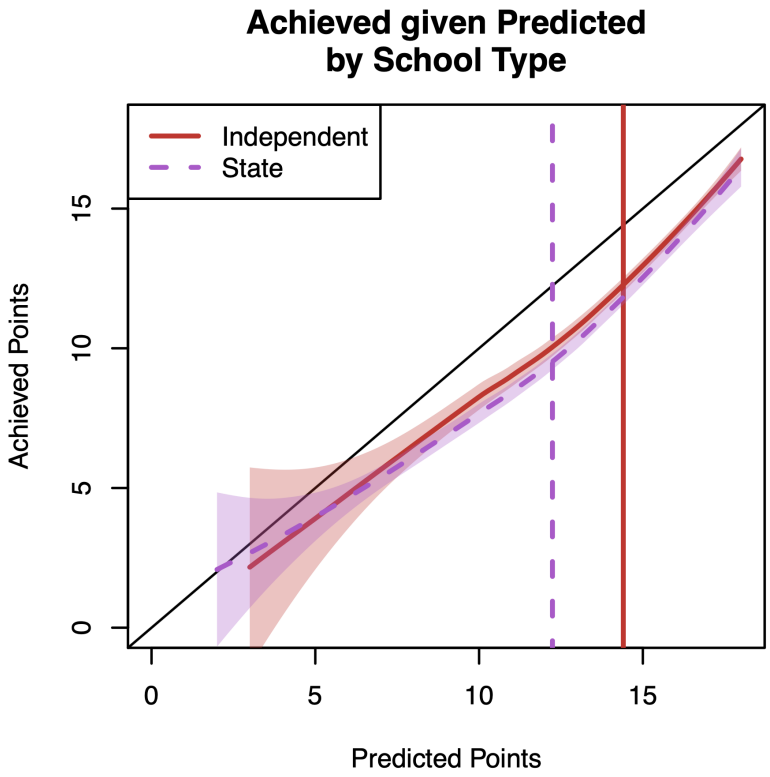


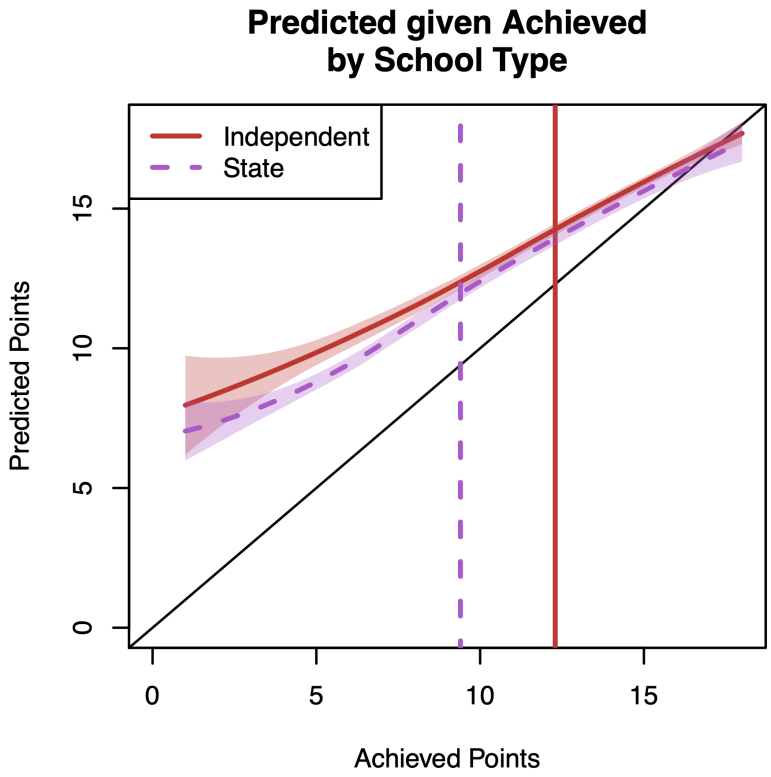


Red and Blue dots represent two values of any covariate X

- Related fact: in general, when you regress Y on X, you do not get the same line as if you regress X on Y!

Predicted and achieved A-levels by school type





- Failure of sufficiency: Given the same predicted points (m), students from state schools achieve slightly less points (μ) than those from independent schools
- But also failure of separation!: Given the same achieved points (μ), students from states schools were predicted less points (m) than those from independent schools

The correct normative standard is not obvious

In favour of *sufficiency*

- Predictions are used in admissions: we do not want group attributes to predict over/underperformance of predicted grades.
- If sufficiency does not hold, university admissions have reason to “adjust” the predicted grades for some groups relative to others.

In favour of *separation*

- The whole system is based on the idea that the achieved exam scores are the canonical truth.
- We ought to want to treat students with the same ultimate achievement equally in the admissions processes.

Deeper issues in this example

- The use of measures in both textbook examples “solve” a real problem
  - The desire to make a decision in advance of the most relevant data arriving
  - Important decisions are made on the basis of predictions/measures
  - But this introduces the potential for unfairness.
- A further issue: if we are trying to measure students’ understanding of the material, the exams are themselves measures with some error.
  - In principle, the predicted grades could perform as well as exams for measuring student understanding.

Measurement error and subsequent analyses

From individual to aggregate

- Fairness is about getting the right answer with respect to individual measured units.
- Social science is often about getting the right answer in the aggregate.
- What happens to our aggregate-level analyses when we have measurement error?

Relevant Cases of Measurement Error

- The textbook discusses:
  - Measurement error in dependent variables
  - Measurement error in independent variables
  - Measurement error in control/conditioning variables
- Measurement error can potentially lead to *mistaken conclusions* in subsequent analyses that employ the measures.
  - Claims that are accurate with respect to a independent or dependent variable measure m can be inaccurate with respect to the concept μ.
  - Controlling for m instead of μ, when you needed to control for μ, will fall short of accounting for differences in μ.

Example: Public service productivity

- The Office for National Statistics constructs a variety of productivity measures for public services.
  - Applied in domains including healthcare, education, adult social care, children's social care, social security administration, public order and safety, policing.
- In general, productivity P is defined as the ratio of outputs O per input I:

$$P = \frac{O}{I}$$

- For convenience, we are going to work on a log scale:

$$\log P = \log O - \log I$$

Measuring outputs is difficult!

- Relatively easy to measure how much governments have spent.
  - Typically expenditure in £.
- Not so easy to measure how much "public service" they have generated for that money.
  - The outputs tend to vary by domain, and include things that are difficult to measure.
- We will consider a "toy example" where we can measure I perfectly, but O imperfectly:
  - Define the "public service" output as $O = H \cdot Q$
  - H is the headcount of people served
  - Q is the (average) quality of the service provided
  - We can only measure H, not Q

Measurement error

- We want to measure:

$$\mu = \log P = \log H + \log Q - \log I$$

- For a given unit i, we are only able to calculate:

$$m_i = \widehat{\log P_i} = \log H_i - \log I_i$$

- Our measurement error for $\log P_i$ is therefore:

$$\epsilon_i = m_i - \mu_i = \widehat{\log P_i} - \log P_i = -\log Q$$

- Our measurement error is negative (log) quality.
  - We underestimate the productivity of units providing higher quality service
  - We overestimate the productivity of units providing lower quality service

Consequences of measurement error

Bad aggregate conclusions

- If we do any aggregate level assessments of which kinds of units (eg local authorities) are more productive, we will ignore quality differences in provision (obviously)
- The kinds of places that provide higher quality service will tend to look less productive than they really are.

Bad incentives

- Depending on the production function for H and Q, the levels of H and Q that maximise H may not be the same as those that maximise $O = H \cdot Q$.
- The productivity assessment will reward maximising headcount at the expense of quality, if there is any tradeoff between the two.

Summing up

- Measurement error is defined as the difference between target concept μ and measure m
- There are three ways in which a measure can be 'bad'
    1. Bias
    2. Variance
    3. Miscalibration
- Some simple indicators of measurement quality are ρ, R^2 and τ
- Measurement fairness: separation vs sufficiency
- Measurement error should be taken seriously