Week 6: Supervised Scale Measurement III: Linear Indices

POLS0013 Measurement in Data Science

Dr. Julia de Romémont

Academic Year 24-25

UCL Departement of Political Science

Big picture

Third week on measurement methods that

1. Are *supervised*
   - i.e. they are constructed by using pre-existing information (expertise or data) to connect target concept and indicator data
2. Produce a *scale*
   - i.e. the resulting measures allow us to confer *ordering* and *distance* between the observations

But where we

- *cannot* easily derive a scale from theory by coming up with a sensible mathematical aggregation formula
- *do not* have access to relevant pairwise comparisons between the units based on which we could produce a relative ordering on a scale
- *do not* have access to an already existing gold-standard measure

Last lecture...

...we considered cases where:

1. We had a set of indicators that we believed might be associated with the target concept we wanted to measure **and**
2. We had a training data set with "gold standard" pre-existing measure of the concept plus the indicators

When then could:

- Connect the already existing measure of a target concept to the set of indicators (i.e. run a regression)
- In order to learn about the relationship between the measure and the indicators (i.e. get a regression equation)
- To then calculate the scores of measure for data where we have the indicators, but not the measure (i.e. calculate fitted/predicted values)

This lecture is about...

Constructing indices

- Cases where we *lack* the necessary pre-existing measures for some units (training data) to connect indicators to the concept of interest with regression
- Combining a set of indicators related to the target concept into an index (i.e. aggregate many numbers into one)[1]
- This requires us to think carefully about
  1. What indicators to include
  2. How to combine them
  3. How to weight them

[1] Similar intuition than in lecture 3, only with usually (many) more indicators and with less sophisticated (*linear*!) mathematical aggregation formulae (the weight is more on the weights here!).

This Lecture

Motivating example

Definition of a linear index

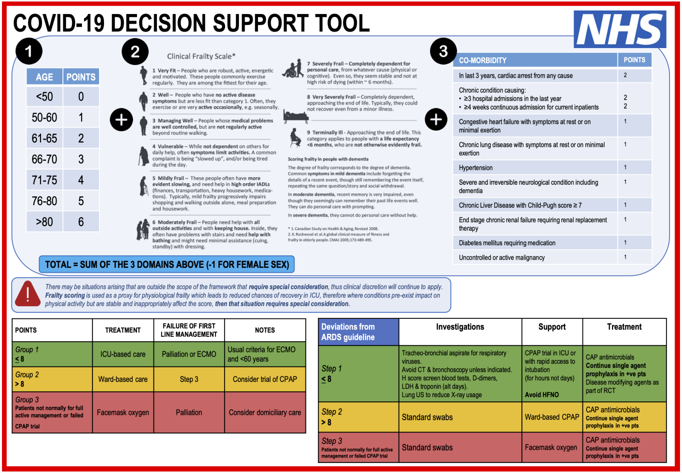
Choosing indicators

Specifying the aggregation

Further considerations

Motivating example

NHS and COVID



COVID-19 DECISION SUPPORT TOOL

NHS

1

AGE	POINTS
<50	0
50-60	1
61-65	2
66-70	3
71-75	4
76-80	5
>80	6

+

2

Clinical Frailty Scale*

1 **Very Fit** – People who are robust, active, energetic and motivated. These people commonly exercise regularly. They are among the fittest for their age.

2 **Well** – People who have **no active disease symptoms** but are less fit than category 1. Often, they exercise or are very **active occasionally**, e.g. seasonally.

3 **Managing Well** – People whose **medical problems are well controlled,** but are **not regularly active** beyond routine walking.

4 **Vulnerable** – While **not dependent** on others for daily help, often **symptoms limit activities.** A common complaint is being "slowed up", and/or being tired during the day.

5 **Mildly Frail** – These people often have **more evident slowing,** and need help in **high order IADLs** (finances, transportation, heavy housework, medications). Typically, mild frailty progressively impairs shopping and walking outside alone, meal preparation and housework.

6 **Moderately Frail** – People need help with **all outside activities** and with **keeping house.** Inside, they often have problems with stairs and need **help with bathing** and might need minimal assistance (cuing, standby) with dressing.

7 **Severely Frail – Completely dependent for personal care**, from whatever cause (physical or cognitive). Even so, they seem stable and not at high risk of dying (within ~ 6 months).

8 **Very Severely Frail** – Completely dependent, approaching the end of life. Typically, they could not recover even from a minor illness.

9 **Terminally Ill** - Approaching the end of life. This category applies to people with **a life expectancy <6 months**, who are **not otherwise evidently frail.**

Scoring frailty in people with dementia

The degree of frailty corresponds to the degree of dementia. Common **symptoms in mild dementia** include forgetting the details of a recent event, though still remembering the event itself, repeating the same question/story and social withdrawal.

In **moderate dementia**, recent memory is very impaired, even though they seemingly can remember their past life events well. They can do personal care with prompting.

In **severe dementia**, they cannot do personal care without help.

* 1. Canadian Study on Health & Aging, Revised 2008.
2. K. Rockwood et al. A global clinical measure of fitness and frailty in elderly people. CMAJ 2005;173:489-495.

+

3

CO-MORBIDITY	POINTS
In last 3 years, cardiac arrest from any cause	2
Chronic condition causing: • ≥3 hospital admissions in the last year • ≥4 weeks continuous admission for current inpatients	 2 2
Congestive heart failure with symptoms at rest or on minimal exertion	1
Chronic lung disease with symptoms at rest or on minimal exertion	1
Hypertension	1
Severe and irreversible neurological condition including dementia	1
Chronic Liver Disease with Child-Pugh score ≥ 7	1
End stage chronic renal failure requiring renal replacement therapy	1
Diabetes mellitus requiring medication	1
Uncontrolled or active malignancy	1

TOTAL = SUM OF THE 3 DOMAINS ABOVE (-1 FOR FEMALE SEX)

! *There may be situations arising that are outside the scope of the framework that **require special consideration**, thus clinical discretion will continue to apply. **Frailty scoring** is used as a proxy for physiological frailty which leads to reduced chances of recovery in ICU, therefore where conditions pre-exist impact on physical activity but are stable and inappropriately affect the score, **then that situation requires special consideration.***

POINTS	TREATMENT	FAILURE OF FIRST LINE MANAGEMENT	NOTES
Group 1 **≤ 8**	ICU-based care	Palliation or ECMO	Usual criteria for ECMO and <60 years
Group 2 **> 8**	Ward-based care	Step 3	Consider trial of CPAP
Group 3 **Patients not normally for full active management or failed CPAP trial**	Facemask oxygen	Palliation	Consider domiciliary care

Deviations from ARDS guideline	Investigations	Support	Treatment
Step 1 **≤ 8**	Tracheo-bronchial aspirate for respiratory viruses. Avoid CT & bronchoscopy unless indicated. H score screen blood tests, D-dimers, LDH & troponin (alt days). Lung US to reduce X-ray usage	CPAP trial in ICU or with rapid access to intubation (for hours not days) **Avoid HFNO**	CAP antimicrobials **Continue single agent prophylaxis in +ve pts** Disease modifying agents as part of RCT
Step 2 **> 8**	Standard swabs	Ward-based CPAP	CAP antimicrobials **Continue single agent prophylaxis in +ve pts**
Step 3 **Patients not normally for full active management or failed CPAP trial**	Standard swabs	Facemask oxygen	CAP antimicrobials **Continue single agent prophylaxis in +ve pts**

"NHS 'score' tool to decide which patients receive critical care"

– Financial Times 13 April 2020

A Linear Index

$$\text{ICU Score} = \underbrace{\text{Age Points}}_{\substack{\text{Points vary}\\ \text{from 0 to 6}}} + \underbrace{\text{Frailty Points}}_{\substack{\text{Points vary}\\ \text{from 1 to 9}}} + \underbrace{\text{Co-Morbidity Points}}_{\substack{\text{Points of 1 or 2 for a number of}\\ \text{potentially overlapping conditions}}} - \underbrace{\text{Female}}_{\substack{\text{One point reduction}\\ \text{for being female}}}$$

$\Rightarrow$ ICU care only for those for which the score sums up to 8 or fewer points[2]

[2] Note that this system was prepared, but **not** implemented.

Context

- In April 2020, the NHS was worried that ICU capacity would be exceeded
  - Needed a plan for how to ration capacity in that event
  - Score aimed to identify those unlikely to survive even with ICU care
- But where do the *numbers* come from?
  - Why do age points vary 0 to 6?
  - Why do frailty points vary 1 to 9?
  - Why do women get a point off their score with otherwise similar age, frailty and co-morbidities to men?

What did we learn last time?

What would we need to apply regression-based training to determine the appropriate points?

- Large and representative data set with individual-level outcomes (eg survival) and these indicators
- Confidence that the relationship between these was stable over time
- Definitely not possible as of 13 April 2020!
- So... just make up the numbers?

Thinking about some comparisons

- Here are two people who both get 8 points:
  1. An 85 year old (+6) woman (-1) who is "well" (+2) with hypertension (+1)
  2. A 52 year old (+1) man (0) who is "managing well" (+3) with a recent cardiac arrest (+2), congestive heart failure (+1) and hypertension (+1)
- The NHS scoring scheme says they should have equal priority for ICU care. Is that appropriate?
- Making this kind of pairwise comparison is both...
  - a potential validation strategy for the point scheme[3]
  - a potential strategy for generating a point scheme in the first place

[3] c.f. Equal cases (axiomatic analysis)

Definition of a linear index

Summarising complex concepts

- What if, for example,
  - we want to measure countries' levels of democracy or globalization or development?
  - we want to measure individuals' political ideologies or personalities or levels of human capital?
- Important concepts across various social science fields, but difficult to measure because they are difficult to define
  - How might we form an *index* of democracy (or globalization, poverty, ideology, personality etc) on the basis of *indicators*?
- Indices of the type discussed here are *extremely* widely used as social science measures
  - See links on course website for this week

Remember indicators

Indicator

⇒ An already measured quantity that provides evidence (i.e. indicates something) regarding the concept that we aim to measure.

- In other words, an indicator is one we believe *indicates* something about the presence/absence of the target concept.
- Generally, we think of indicators as *partial* and/or *noisy* signals of the target concept.

Indicators of Democracy

Example:

We are trying to measure whether a country is a “democracy”.

- Having regular elections tends to go with being a democracy
  - *an indicator*
- Regular elections may not be sufficient to be a democracy
  - *indicator is not just the concept*
- Regular elections may not be necessary to be a democracy
  - *indicator is not just the concept*
- There are *other* elements of being a democracy that are also indicators of that concept
  - *need more than one indicator to construct the index*

Indices

Index

...A composite statistic that is formed by aggregating multiple indicators.

How aggregation operates is determined by the:

1. **Functional form**: additive? multiplicative? exponential?
   - To reflect/specify *how* the indicators should be combined
   - In practice most indices are some kind of **additive** function.[4]
   - More strongly theorised measurement strategies might involve non-additive functional forms (see textbook).
2. **Weights**: equal weights? unequal weights?
   - To reflect/specify *how much* each indicator matters relative to the others

[4] In that case, indicators must be on some kind of common scale – c.f rule 1 in slides of week 3 lecture (Dimensional analysis).

Additive Indices

- Many indices are *additive* scales where the value of the index m_i for unit i is a function of various indicators j in the following form:

$$m_i = \sum_j b_j \cdot I_{ij}$$

- In some cases, this form is nested, meaning that I_j is itself a *sub-index* constructed from several further indicators
    - Note that 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 sum of underlying indicators in the sub-indices
- Sometimes, this is called a "sum score": the score for unit i is the *sum of scores* b_j on a set of items I_{ij}; Sometimes, scores are described in terms of points (e.g. NHS ICU example, Nutriscore)

Additive Indices look like regression equations

$$m_i = \sum_j b_j \cdot I_{ij}$$

- Looks like the linear regression equations we considered last time, but b_j instead of β_j **because we lack any training data m_i to fit a model to learn the coefficients.**
  - Therefore, instead of *estimating* the coefficients, we just pick some values for them.
- Often described in terms of "weights" w_j rather than "coefficients" b_j, which makes most sense when
  - indicators are standardised in some way (either to have mean zero and standard deviation 1 or to range from 0 to 1)
  - weights w_j are all positive ($w_j \geq 0 \forall j$) and add to 1 ($\sum_j w_j = 1$).

Thinking about equivalences

- This type of additive index presents itself as an interval-level measure
- This implies that **equal differences in the index value must be** (assumed to be) **equivalent** with respect to the underlying concept.
  - Increasing I_1 by Δ will change the index by $b_1\Delta_1$
  - Increasing I_2 by $\Delta_2 = \frac{b_1\Delta_1}{b_2}$ will also change the index by $b_1\Delta_1$
- For a given, well constructed index, one ought to be comfortable with all such equivalences
  - This should involve an explicit validation strategy that checks whether such trade-offs are plausible (e.g. with pairwise comparisons, as on this slide)

Choosing indicators

What should be in the Index?

- Of course, we should want to include those indicators that are **indicative** of the concept of interest
- We want to include indicators that are **conditionally associated** with the concept of interest
  - Thinking about this in terms of linear regression: *non-zero coefficients even when other relevant indicators are included*
- But remember, we don't have the data to estimate those coefficients so we have to make (defensible) decisions about
  a. which indicators to choose, especially when there are several to choose from that are highly correlated, and
  b. how these relate to the target concept, i.e. what the coefficients/weights/points are

A general strategy

Iterative decomposition of the target concept

1. **Conceptualisation**: break a concept down into constituent dimensions and identify how they are interrelated
2. **Measurement**: find measures of the indicators associated with those dimensions
3. **Aggregation**: combine the measures of the attributes in a way implied by the conceptualization

- In the NHS COVID case:
  - Outcomes expected to depend on age, frailty, co-morbidities and sex
  - Develop (hopefully comparable) points scales for each
  - Add them back up

Two extremes of a conceptualisation spectrum

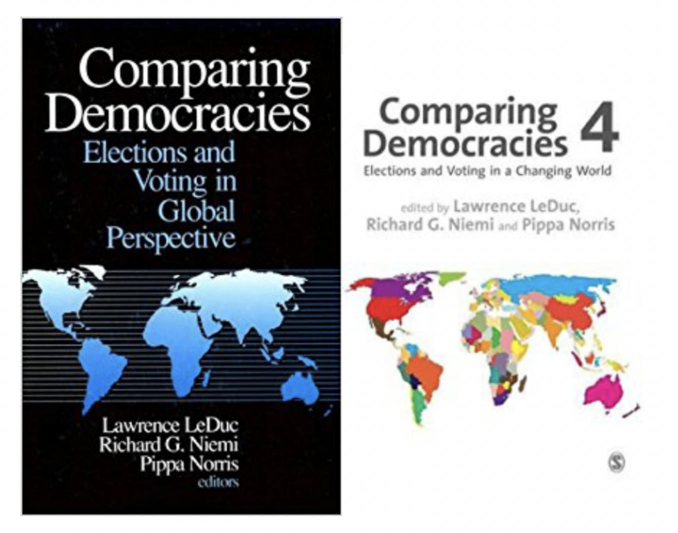
Minimalist Measurement Strategies

- Rely on relatively parsimonious (‘thin’) definitions of a concept
- Implying less constituent dimensions and therefore relatively few indicators
- Limited number of sources of bias, but possibly very large bias

Maximalist Measurement Strategies

- Rely on more inclusive (‘thick’) definitions of a concept
- This has as a consequence more constituent dimensions and hence relatively more indicators
- More potential sources of bias, but each one is smaller

Two extremes of a conceptualisation spectrum


Comparing Democracies
Elections and Voting in Global Perspective
Lawrence LeDuc
Richard G. Niemi
Pippa Norris
editors
Comparing Democracies 4
Elections and Voting in a Changing World
edited by Lawrence LeDuc,
Richard G. Niemi and Pippa Norris

Finding good indicators is difficult

- The biggest challenge in measuring concepts like these is getting the right indicators.
    - Or any indicators at all.
- The textbook discusses a multidimensional poverty index that is based on dimensions of health, education and standard of living.
    - Authors were unable to include dimensions of work, empowerment, the environment, safety from violence, social relationships, and culture.
    - Depending on how important you think these dimensions are relative to the three that were included, you could argue that the measure is missing over “half” of the target concept.

Is what you have enough?

- Depends on what you are missing out on!
- A limited set of indicators can lead to measurement error if the measure failed to capture important elements of what we are interested in
  - Especially if those missing elements are associated with other variables of interest (see week 2)
- If the missing indicators would have been highly correlated with the included ones, then the measurement error would be small because little of the target concept is being missed
- More broadly, we want to combine indicators that are all somewhat correlated with each other (they are measuring the same thing!) but not too much, as each should add information (they are not the same!)

Transforming indicators

- Before combining the indicators, you may need apply transformations to the individual indicators to satisfy the requirements of an interval-level measure (c.f. Dimensional Analysis!)
    - Common transformations (for continuous variables) include: log transformation, standardisation or linear rescaling
    - For categorical indicators, transforming the variable simply means assigning a number of 'points' to each level
- This is about determining how different levels of the **same** indicator are associated with the target concept

Specifying the aggregation

How should the indicators be added up?

- Specifying the indicator coefficients means determining the relative (partial) associations of **different** indicators with respect to the concept
- But where do the b_j/w_j/points come from?
- The textbook discusses three common strategies:
    1. Equal weighting ($b_1 = b_2 = ...$)
    2. Analyst-specified weighting (based on theory)
    3. Expert-specified weighting
- Equal weighting is very common, because it is often really difficult to justify any specific values for the b_j!

Additivity and Equal Weighting are ubiquitous

- In practice, many scales of this type are *additive* and *with equal weight* at some level because the authors have
  - no good justification for non-additivity
  - nor for any particular choice of unequal weights.
- You might find either of these assumptions - additivity and equal weighting - troubling or 'unrealistic' in general or in particular applications
  - But remember: none of the measures we are talking about are really 'realistic'! Instead we care about their relative usefulness.
  - In short: we are being *pragmatic*

Additivity as a 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
- Think of these additive scales in a similar way to how you (I hope) think of linear regression models.
  - Useful “first-order” approximations of any pattern of variation you might observe
  - 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.

Equal weighting as a default

- To continue the linear regression analogy, equal weights are like assuming *a priori* that all the β's in your regression model should be equal
  - This might strike you as crazy
  - But occasionally defensible after standardizing even for regression (Graefe 2015)
- This is like saying that
  - all the things that you measured have the same weight
  - all the things you did not measure have zero weight
- Not terribly plausible in the abstract, but does not require you to make any further decisions beyond which indicators to include or not

Examples

- The Immigrant Integration Index in the textbook is an example of equal weighting (and additivity)
  - Six equal weighted components of immigrant integration
  - Four equal weighted indicators of each component, each of 5 point scales
- The NHS ICU score is an example of unequal weighting (and additivity)

$$\text{ICU Score} = \underbrace{\text{Age Points}}_{0\leftrightarrow 6} + \underbrace{\text{Frailty Points}}_{1\leftrightarrow 9} + \underbrace{\text{Co-Morbidity Points}}_{1,2} - \underbrace{\text{Female}}_{-1,0}$$

  - Different range of possible point totals for each indicator lead to unequal weighting

Another example of equal weights

In "Targeted: The Mobilizing Effect of Perceptions of Unfair Policing Practices", Hannah Walker creates a linear index of the extent to which Americans who are in racial minority groups feel a sense of injustice regarding their interactions with police.

- Questions: "*Thinking about some things that police officers who patrol your neighborhood may or may not do, please indicate how often you think the police who patrol your neighborhood do each of the following:*"
    1. Stop people in their cars or public places without good reason
    2. Use excessive physical force or verbally abusive language
    3. Treat people like me fairly and respectfully.
- Responses: Is it very often, somewhat often, not that often, or almost never?
- Injustice index is a 0 to 9 scale, resulting from coding responses to each item 0, 1, 2 and 3 (in the appropriate direction)

Case for/against equal weights

- Why equal weights seem reasonable here
  - Three questions of the same form with the same response choices
  - Three questions designed to capture different aspects of the same concept
- Potential problems?
  - Response spacing: are very often, somewhat often, not that often, and almost never integer spaced?
  - Equal weighting: should the use of excessive physical force really get the same weight as the other two?
- On balance, it seems likely that equal weights are going to be ok in this sort of application

Arguments for unequal weights

- There are contexts in which scales include unequal weights based on substantive arguments.
  - For example, before 2010, the Human Development Index gave adult literacy twice the weight of school enrollment in forming the education sub-index
  - Most points systems are "unequal weighting", usually based on a priori integer weighting
- Such arguments for unequal weights **are difficult to make convincingly**, even for simple adjustments to weights (such as counting one indicator at twice the weight of another).
  - Is there any way we might we do better?
  - How might we try to estimate "correct" weights?
  - What sort of additional data would we need to do this successfully?

Using comparison data to learn indicator weights

- What if we have some data that we think tells us something about whether $\mu_i > \mu_{i'}$ *and* we have data on indicator values I_i and $I_{i'}$ for some pairs of units?
- Building from the Bradley-Terry model's intuition, we can replace the individual unit "strengths" α_j with linear functions of the indicators
- The coefficients from the model then provide an estimate of appropriate weights for the indicators:

$$\hat{\mu}_i = \hat{\beta}_1 I_{1i} + \hat{\beta}_2 I_{2i} + \cdots$$

- As it was the case three weeks ago, for this to work, the pairwise comparisons should reflect the underlying concept we want to measure
  - i.e. the "strength" that determines the pairwise winners needs to be the concept you want to measure, not something else.

Making the comparisons

- It is rare that these sorts of comparisons already exist, you usually need to generate them
- One way to generate these is to survey people who you think have relevant knowledge of the relationship between the indicators and the target concept
- This is an example of a *conjoint experiment* where respondents are provided with two (randomly chosen) profiles of indicator values and asked to choose between the two
- Think about applying this in the NHS COVID case:
  - How would the profiles to compare look like?
  - Who would be a relevant population to ask?
  - What would they be asked?

Further considerations

Mitigating the consequences of measurement error

- The kinds of measurement errors that can arise from missing important indicators and/or miscalibrated weights can be mitigated by **interpreting results narrowly and carefully**.
  - The poverty measure mentioned earlier might be better described as measuring “three dimensions of poverty” rather than “poverty”.
- So long as you remember that and state your results clearly, you won’t be wrong! (Or at least you’ll know you are wrong)
- More broadly, one should better discuss results of analyses involving a measure of μ in terms of associations with/effects of/effects on the *measure* of μ, and *not* just μ

Indices as ...

Independent/Treatment Variables

- Be extremely cautious about making *causal* claims where the measure is the independent/treatment variable!
- You would have to convincingly claim that *exogenous* variation in the underlying is manifested in the index for you to vindicate such a claim
    - There is often an alternative definition of the treatment variable that would yield a clearer analysis

Indices as ...

Outcomes

- Given a credible 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
- It is reasonable to talk about causal effects of some treatment on the *measure*
- We should be more careful about making claims that there is a treatment effect on the *underlying concept*, because the causal effect could be on the measurement error

Indices as ...

Running Variables for RDD

- Measurement scales like these can still be useful for causal analysis, especially when they are used to make treatment (e.g. administrative) decisions
- They can then be used as a 'running' variables for so-called regression discontinuity designs
  - E.g. Lerman (2009) leverages the assignment of prisoners in California to different prison environments based on an additive scale of inmate background, crime and and sentence characteristics
- This will make more sense when (some of) you get to Lecture 9 of the causal inference module in a few weeks!

The meaning of the numbers on the scale

- One “feature” of scales constructed in this way is that there is **seldom a natural metric** for the target concept.
    - The immigrant integration scale is reported on a 0-1 scale, by linearly rescaling the 12-60 or 24-120 points to run from 0-1.
    - There is no right answer to what range the scale should cover.
- Are these indices really interval-level scales? Is a given increase really equally meaningful at all points of the scale?
    - Note that if you don’t believe this, you don’t believe in the validity of the scale overall!
    - Additivity is often simply a convenient mathematical structure, but failure of the interval-level interpretation implies you should not have been adding the indicators.

Trying many strategies

- It is good practice to try more than one weighting scheme, if you cannot be confident in any particular one.
- For example, the Global Health Security Index assessed four different weighting schemes:
    1. Expert weights (judgements of an international panel of experts, April 2019)
    2. Neutral weights (equal by indicator category)
    3. Equal weights (by indicator)
    4. PCA weights
- What are PCA weights?
    - Our topic for in two week!

Summing up

- (Linear) indices are a very common type of measure that involve combining the values of a set of indicators, based on decisions made about 1) the functional form and 2) the weights of the aggregation
  - In practice, these often are 1) additive and 2) equal weights, if there are no good reasons to do otherwise
- Additive equal weight indices make a series of assumptions about the relationship between indicators and target concept:
  1. The relationship is linear (i.e. additive)
  2. Every indicator is equally indicative of the target concept
  3. Equal differences in the index value are equivalent with respect to the underlying concept
- As measurement error is always a threat, you should discuss results by talking about “the measure of the target concept” and not “the target concept”