Week 3: Deriving Scales from Theory

POLS0013 Measurement in Data Science

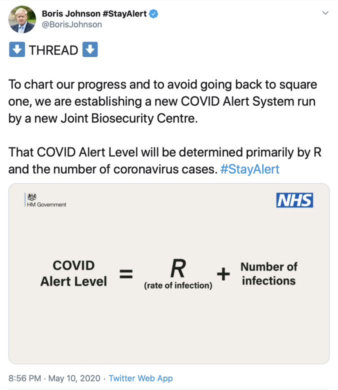
Dr. Julia de Romémont
Academic Year 24-25
UCL Departement of Political Science

Last lecture..

...we talked about Measurement Error

- Measurement error as the (distribution of the) differences between the measure m and the target concept μ
- We saw that measures can be
  - More or less wrong, on average (bias, validity etc)
  - More or less *more* wrong, on average (variance, reliability etc)
  - More or less *more* wrong, on average, at different values of μ (miscalibration)
- Measures can be considered ‘unfair’ when differences in the measure across groups are due to differences in measurement *error* instead of differences in the target concept

... why *this* is a bad tweet



... we will talk about

- Cases where we have a clear idea[a] about the connection between the target concept and the observable indicators
- In other words: when we know/can come up with the mathematical formula on how to aggregate data into the measure
- We will talk about two strategies we can follow to make sure the matematical formula makes sense
  - Dimensional analysis
  - Axiomatic analysis

[a] I.e. a theoretical intuition

This Lecture

Translating theoretical arguments

Dimensional Analysis

Axiomatic Analysis

Translating theoretical arguments

Deriving measures

- In some cases, we have a concept in mind that is very ‘close’ to the available data
- The measurement strategy then involves deciding on which way to translate the data into a measure
- This means that we can specify a fixed relationship between the indicator data to the concept of interest
- In other words: the concept is easily (mathematically) ‘translatable’ into a measure

A simple (?) case

Concept of interest
An individual's wealth

$\Downarrow$

Observable indicator
The individual's total amount of £££

$\Downarrow$

Measure
$$\text{wealth}_i = \sum £$$

Some more theoretical questions to consider:

- Do we count only disposable income?
- What about illiquid assets?
- Do we account future income, such as inheritance? etc.

A less simple case

Concept of interest

Level of competition in an industry

$\Downarrow$

Observable indicators

Companies' market shares p_j

$\Downarrow$

Measure

Herfindahl–Hirschman Index $= \sum_{j=1}^{N} p_j^2$

Have a think:

- What does the squaring of market shares do?
- What is the HHI if there is a monopoly?
- What is the HHI if all companies have equal market shares?

How to measure social utility

What decisions are we making with these two Social Welfare Functions?

$$W(.) = \sum_{i=1}^{n} u_i \qquad\qquad W(.) = \prod_{i=1}^{n} u_i$$

Moral philosophers have been discussing for a long time about the ‘correct’ way to mathematically aggregate individual utility into *social* utility.

- There even is a highly recommended[1] show about this: The Good Place
- Founding utilitarian Jeremy Bentham is an important figure in UCL’s history (and even donated his remains (‘auto-icon’) to UCL)
- How we should “weigh up” different utilities is quite a hot topic among “tech bros”... (e.g. effective altruism and longtermism)

[1] By me!

Specificities of theories

- The textbook covers more examples, e.g.:
  - Debt-GDP ratios
  - Measures of inequality
  - Measures of poverty
  - Measures of effective party count

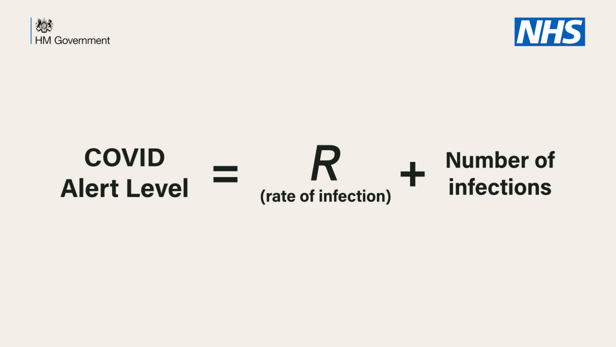
Since theoretical arguments are necessarily specific to a particular application, so are the derived measures.

- The **method** used to aggregate from data to concept is specific to the context.
- However, we can identify general **strategies** to derive a measure (that makes sense)
  1. Dimensional Analysis
  2. Axiomatic Analysis

Limitations of theoretical reasoning

- Dimensional and axiomatic analyses will only get you so far
- Most useful for concepts that are already reasonably “close” to the available indicator data
- But limited if you cannot come up with a specific aggregation formula or with a good reason for using one weighting over another
- Sometimes you will end up simply identifying a difficult data/estimation problem that you still need to solve

Example: COVID Alert Level



- "The reproduction number (R) is the average number of secondary infections produced by a single infected person."
- "If R is greater than 1 the epidemic is growing, if R is less than 1 the epidemic is shrinking. The higher R is above 1, the more people 1 infected person infects and so the faster the epidemic grows."

Taking the equation seriously

$$\text{COVID Alert Level} = \underbrace{\text{R}} + \text{Number of Infections}$$

$$\text{R} = \frac{\text{Future Infected Persons}}{\text{Current Infected Persons}}$$

$$\text{COVID Alert Level} = \frac{\text{Future Infected Persons}}{\text{Current Infected Persons}} + \text{Current Infected Persons}$$

Analogous equation

- I offer to give you £105 in one year's time for every £100 you lend me now.
- You have £1000 to lend.
- Would it make any sense to calculate the following quantity?

$$??? = \frac{\text{Future £105}}{\text{Current £100}} + \text{Current £1000}$$

- Does it make sense to add an interest rate and a current asset total?

$$1.05 + \text{£1000}$$

No!

If BJ had listened to this lecture...

Dimensional Analysis

Dimensional Analysis

Analysis of relationships between different (physical) quantities by identifying and converting their dimensions and units of measure so that inferences can made about the relations between them.

- Widely used when solving problems in the physical sciences, particularly for checking the plausibility of a final calculation.
- Helps assess whether the units of the measure are internally consistent.
- Helps understand why this is funny: https://www.youtube.com/watch?v=JYqfVE-fykk

Dimensions and units

Dimensions are the "concepts" you are measuring: time, money, people, distance, etc.

Units are the quantities in which you are measuring those units numerically (the unit of a is denoted as $\{a\}$).

- Examples
  - Dimension of time can be be measured in units of years, days, hours, minutes, seconds, etc.
  - Dimension of money can be measured in units of \$s, £s, €s etc.
  - Dimension of people can be measured in units of people or in thousands or millions of people.

Unit conversion

- Conversion between different **units of measurement** for the **same quantity**, ie. within the same dimension

$$
\begin{aligned}
1\ \cancel{\text{day}} \times \frac{1\ \text{years}}{365.25\ \cancel{\text{days}}} &= \frac{1}{365.25}\ \text{years} \\
1\cancel{£} \times \frac{1\ \$}{0.78\cancel{£}} &= 1.28\ \$
\end{aligned}
$$

- Unit conversion ratios are equal to 1, and are dimensionless overall
- You can always multiply a quantity by 1 without changing that quantity

Compound units

= Combination of different **units of measurement** across **different dimensions**

- Per capita Gross Domestic Product (pcGDP) has dimensions of money per person per time period, typically US$ per person per year
- Dimensions/units indicate which kinds of mathematical operations make sense

$$\begin{aligned} pcGDP \times Population &= GDP \\ \frac{\{\$\}}{\{\text{person}\}\{\text{year}\}} \times \{\text{persons}\} &= \frac{\{\$\}}{\{\text{year}\}} \end{aligned}$$

Rules of dimensional analysis

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 $\{a\}/\{b\}$.
    - If you raise a quantity a to the power p, $\{a^p\} = \{a\} \cdot \{a\} \cdot ... = \{a\}^p$.

Rules of dimensional analysis

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 $\left\{\int_{x_0}^{x_1} a \cdot dx\right\} = \{x\} \cdot \{a\}$.

Side note:

- Why are the units multiplied here? Why doesn't the first rule apply, namely that you can only can add numbers of the same unit?
- It's because summation/integration is secretly multiplication!
  - When you calculate 3×4, you calculate $3 + 3 + 3 + 3$
  - Similarly, when you calculate $\sum_{i=1}^{n} x_i$ you calculate $x_1 + x_2 + x_3 + ... + x_n$. The difference here is that the number over which you sum (the summand/integrand) changes.

Standardisation

- Many theoretical concepts are about the distance (or dissimilarity) between entities across more than one component
- How can we get a single measure of distance in a multi-dimensional space with widely different units of measurement?
- A common and sometimes convenient way to perform unit conversion to **standardise** the scales of the different sub-dimensions in order to then use them to perform additive comparisons
  - e.g. $x_s = \frac{x - \bar{x}}{sd(x)}$
- Two measures of distance between two observations that you can obtain through standardization are
  1. Normalised Euclidean distance
  2. Mahalanobis distance

Standardisation example

Let's look at our measures of democracy from seminar 1, and standardise them in terms of their own distribution:

```r
# on the original scales
summary(democracy$freedomhouse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    2.50    4.00    4.15    6.00    7.00    2699
```

```r
summary(democracy$polity)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## -10.0000  -7.0000  -1.0000   0.1286   8.0000  10.0000     1087
```

```r
# standardised
summary(scale(democracy$freedomhouse))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## -1.5255 -0.7992 -0.0728  0.0000  0.8957  1.3799    2699
```

```r
summary(scale(democracy$polity))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## -1.3508 -0.9507 -0.1505  0.0000  1.0498  1.3165    1087
```

Regression

- Similarly, regression models *convert* the units of the independent variables into units of the dependent variable
- The β coefficient each have unit $\frac{\{Y\}}{\{X_k\}}$, where k is the index of their respective independent variable
- Therefore, $\beta_1 X_1$ has unit $\frac{\{Y\}}{\{X_1\}}\{X_1\} = \{Y\}$
- Regression is a way to estimate unit conversion ratios, so that quantities with different units can be summed

Back to COVID Alert Levels

- You cannot meaningfully add these two quantities:

$$\begin{aligned}\text{COVID Alert Level} &= \underbrace{\frac{\text{Future Infected Persons}}{\text{Current Infected Persons}}}_{\text{a rate}} + \underbrace{\text{Current Infected Persons}}_{\text{a count}} \\ &= \frac{\{a\}}{\{b\}} + \{b\}\end{aligned}$$

$\Rightarrow$ This equation violates dimensional analysis.

- C.f. rule 1: you can only add and subtract quantities with the **same** units

A plausible alternative

- Dimensional analysis indicates that you could *multiply* them instead:

$$\begin{aligned}\text{Alert Level} &= \frac{\text{Future Infected Persons}}{\text{Current Infected Persons}} \cdot \text{Current Infected Persons} \\ &= \text{Future Infected Persons}\end{aligned}$$

- If you multiply R by the current number of infected persons, you get the number of infections at "the next generation" of the disease.
- While probably not the ideal measure of how bad the current situation is, it is at least not nonsense.

What else might we do?

- The alert level is meant to capture how bad things are. What if we had some "coefficients" to translate R and current infected persons into "badness"?

$$
\begin{aligned}
\text{Alert Level} &= \beta_R \cdot \text{R} + \beta_{\text{Infected Persons}} \cdot \text{Infected Persons} \\
&= \frac{\text{Badness}}{\text{R}} \cdot \text{R} + \frac{\text{Badness}}{\text{Infected Persons}} \cdot \text{Infected Persons} \\
&= \text{Badness} + \text{Badness}
\end{aligned}
$$

- This works dimensionally, but...
  - ...requires a linear and additive relationship between "badness" and both R and Current Infected Persons, somewhat implausible here.
  - ...we need to figure out the coefficients somehow.

Axiomatic Analysis

Axiomatic Analysis

Procedure by which a metric is generated in accordance with specified rules by logical deduction from certain basic propositions (axioms or postulates).

- What properties should a measure satisfy?
- Listing these “axioms” 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.
- Basically, axiomatic analysis means thinking through what you want your measure to look like relative to the thing you want to measure and make sure the mathematical formula achieves this.

(Five) common types of axioms

1. Special/extreme/limiting cases
   - What are special scenarios and what value should the measure have?
2. Equal cases
   - What are cases which have the same value in the target concept and therefore should have the same value in the measure?
3. Derivative conditions
   - How should the measure change for specific increases underlying indicator data?
   - Positive? Negative? By how much? Should the rate of change be the same across the range of the measure ?

(Five) common types of axioms

4. Continuity and smoothness conditions
   - Is the relationship between the data/measure continuous?
   - Are there any weird jumps/impossible numbers due to the mathematical formula that don't make sense?
5. Functional form restrictions
   - What range of possible value of the measure do we want?

Functional form restrictions

The logit link function (which transforms linear regression into a logistic regression) is an example of a functional form restriction!

- Remember, the limits of a linear regression line are $-\infty$ and $+\infty$
- But we may sometimes want our predicted values to be bounded/*restricted*, for example between 0 and 1 for probabilities!
- So we need to use link *function* to transform $\beta_0 + \beta_1 X$ into values that are meaningfully interpretable as conditional probabilities π

$$\cancel{\pi = \beta_0 + \beta_1 X}$$

$$\pi = [\ldots] = \beta_0 + \beta_1 X$$

$$\pi = logit^{-1}[\beta_0 + \beta_1 X] = \frac{e^{(\beta_0 + \beta_1 X)}}{1 + e^{(\beta_0 + \beta_1 X)}}$$

Limiting cases for the COVID alert level

Axiom If R = 0 *or* if current infections is 0, there will be no future infections, therefore the alert level should be 0.

- Both the original additive equation and the version with coefficients fail this axiom:

$$\text{Alert Level} = \beta_R \cdot \text{R} + \beta_{\text{Infected Persons}} \cdot \text{Infected Persons}$$

- The multiplicative equation is consistent with this axiom:

$$\text{Alert Level} = \text{R} \cdot \text{Infected Persons}$$

Equal cases for the COVID alert level

Axiom Different current situations which will lead to the same number of cases in the future should yield the same alert level.

- E.g., if R = 2 and our current infections are x, we will have the same number of cases "in the next generation" as if R = 1 and current infections are 2x.
- Again, additive equations fail this axiom.
- The multiplicative equation is consistent with this axiom:

$$\text{Alert Level} = R \cdot \text{Infected Persons} = 2 \cdot x = 1 \cdot 2x$$

- Note: this is not necessarily the axiom you actually want, more on this point later.

Derivative conditions for the COVID alert level

Axiom: the COVID alert level should increase if either R or the current number of infections increase.

- All the measures we have considered are consistent with this axiom.
  - The increases in Alert level for a one unit increase in current infected persons is positive
- For the additive equation with coefficients:

$$\frac{\partial \text{Alert level}}{\partial \text{Infected persons}} = \beta_{\text{Infected persons}} > 0$$

- For the multiplicative equation:

$$\frac{\partial \text{Alert level}}{\partial \text{Infected persons}} = R > 0$$

Continuity and smoothness conditions for the COVID alert level

Axiom small changes in either R or the current number of infections should lead to small changes in the COVID alert level.

- All the measures we have considered are consistent with this axiom.
  - At no point does a small change in one of the indicators lead to a sudden big increase or decrease in COVID alert level.

Axioms are useful for defining new measures

- The multiplicative measure we have considered sets the alert level at the number of infections that we would expect at "the next generation of the disease".
  - A little unclear when this is, but there is a detailed report by the Royal Society if you are interested in how R relates to growth rates in a disease over time..
- Measuring the concept of "how much we should be on alert about COVID" requires us to make some substantive/theoretical choices.
  - The axioms *are* the choices that we have made.

Summing up

- When deriving scales from theory we need to have clear **theoretical justifications** for:
  1. The chosen indicators/data
  2. How these are combined into a measure
- Two general strategies to help derive a messure from theory (or check whether a given one makes sense)
  - **Dimensional analysis** means looking at the units of the indicator data and checking whether the mathematical aggregation perfomed on them makes sense
  - **Axiomatic analysis** means looking at the mathematical aggregation formula and thinking through whether it gives results that make intuitive sense
- This approach quickly reaches its limits, as the connection between indicators and target concept becomes less clear