Introduction to Measurement

Eric Guntermann

January 8th, 2016

What you need

- R
- RStudio
- R code file
- Datasets
- You can find all of this at:
  `http://ericguntermann.com/measurement.html`

What we will learn

- Quick review of R
- What is scaling/measurement?
- Data theory
- Summated ratings scales
- Principal components analysis
- Factor analysis
- Multidimensional scaling
- Text analysis

Quick review of R

- Objects store information
- Commands/functions are performed on input objects and their output is assigned (<−) to output objects
- Commands are stored in packages

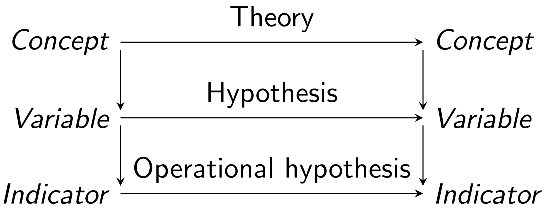
Applying a command to an input object and assigning the output to another object

```
output object <- command(input object)
```

Scaling and dimensionality

- Scaling is about optimizing information. We seek:
  - Power: explain variance
  - Parsimony: minimize number of dimensions
- Dimensionality: number of important sources of variability among set of objects
- Generally, we can present results graphically

Theories, hypotheses, and operational hypotheses



An indicator is a measure of a concept

- A concept is abstract, rarely directly observable
- An indicator is directly observable

Data theory

- Definition: study of extracting information from empirical observations
- The information we extract is our data
- Using various techniques, we produce data for analysis
- All data analysis relies on an often implicit data theory
- Knowing about data theory gives us a lot of freedom!
- Allows researchers to be creative

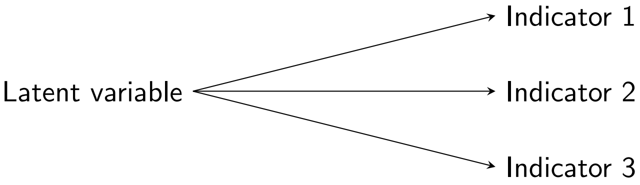
Difference between data and observations

- We observe a lot of things
- But we only retain part of these

Example: response to survey question

- Time
- Physiological reaction
- Length of response
- Answer

General Principle : Latent variable explains variability in a number of observable variables



Comparison to regression

$$X_1 = \alpha_1 + \beta_1 * \omega + \epsilon_1$$
$$X_2 = \alpha_2 + \beta_2 * \omega + \epsilon_2$$
$$X_3 = \alpha_3 + \beta_3 * \omega + \epsilon_3$$

Comparison to regression: three latent variables

$$X_1 = \alpha_1 + \beta_{1a} * \omega_1 + \beta_{2a} * \omega_2 + \beta_{3a} * \omega_3 + \epsilon_1$$
$$X_2 = \alpha_2 + \beta_{1b} * \omega_1 + \beta_{2b} * \omega_2 + \beta_{3b} * \omega_3 + \epsilon_2$$
$$X_3 = \alpha_3 + \beta_{1c} * \omega_1 + \beta_{2c} * \omega_2 + \beta_{3c} * \omega_3 + \epsilon_3$$

Other words for latent variable

- Factor
- Dimension
- Component

Coombs' Data Theory

- Two of four types of data (with their scaling methods):
  - Single stimulus data: place objects along one or more dimensions, eg. people and intelligence tests, survey respondents and left-right scale (summated ratings scale, principal components analysis, factor analysis)
  - Similarities data: proximity relation between pairs of objects from the same set, eg. distances between cities, similarity between political parties (multidimensional scaling)

Summated Rating Scales (i.e. Likert scales)

- We have scores of n units on k items
- k items are considered imperfect observations on underlying characteristic
- We assume k variables are scored in the same way
- We "collapse across the columns" (i.e. take the mean within each row)
- Major assumption: there is a dimension underlying the items (can create false dimensions)

Why?

- Give us finer resolution: one 0/1 item divides dimension into two, two 0/1 items divide dimension into three... (each item adds a cutting point)
  - k items with m categories lead to k(m-1) + 1 distinct scores
- Increase level of measurement
- Reduce measurement error. Each item consists of i's true position along dimension plus error: $V_{ij} = T_i + E_{ij}$
  - If we assume the errors cancel out (i.e. $E(E_j)) = 0$, when we add more items to the scale, it gets closer and closer to the underlying dimension
- Another assumption: Each item has a monotonic relationship to underlying dimension (i.e. Monotone homogeneity)

How do we verify our assumptions?

- We do an item analysis: make sure each item has a monotonic relationship with the underlying dimension
- Best not to use correlations:
  - Are inflated because scale contains items
  - Only measure linear relationships
- Don't rely only on Chronbach's alpha, because it measures linear relationships among items and is affected by outliers!
- Instead look at graphs showing item against the scale without the item and a loess curve (rest plot)

Chronback's alpha

$$\alpha = \frac{k\bar{r}}{1 + \bar{r}(k-1)}$$

k is the number of items

$\bar{r}$ is the mean correlation among the items

Problems with alpha

- Based on means correlation: means are strongly influenced by extreme values
- There might be clusters: items 1 and 2 are related and items 3 and 4 are correlated, but no correlation between the first two and the last two
- Only measures linear relationships
- Increases with number of items

Potential problem with summated rating scales

- Model relies on the assumption that an underlying dimension exists
- Can give false positives, especially if only use alpha. Beware of clustering!
- If you have any doubt about items, don't create summated ratings scale

Principal Components Analysis

- Get orthogonal (uncorrelated), variance-maximizing components (i.e. capture most variance)
- Each component is a linear combination of the variables: $C_k = a_{k1}X_1 + a_k X_2 + ... + a_{km}X_m$
- Atheoretical: we don't have a theory that there are one or more underlying dimensions
- Not about small number of latent variables. Just components that soak up variance
- Find one dimension that captures most variance in variables, then find a second that is uncorrelated with the first which captured the greatest amount of remaining variance, ...

Principal Components Analysis (2)

- Important to standardize data. Otherwise, variables with biggest variance will be most strongly related to first component.
- Goals: explore dimensional structure of data and possibly reduce dimensionality
- Not necessarily data reduction. Only if small number of components capture lots of variance
- Express k variables with less than k variables, which are orthogonal

Factor analysis (i.e. exploratory factor analysis)

- Goal: find factors (latent causes) that are common to two or more indicators
- Factor indeterminacy: there are infinitely many solutions
- PCA: finding underlying sources of variation
- FA: finding underlying causes. Don't try to capture all variation.
- Assumption In FA: all variables are caused by the same static source
- Factors exist in the real world. In PCA, components depend on variables.
- Usually fewer common factors than observed variables

Factor analysis (2)

- Total variance = common + specific + and random measurement error
- Communality: amount of variable's variance that is derived from common source, that it shares with other variables
- Unique variance: specific to variable
- Principle components doesn't allow for unique variance. It tries to capture all variance.
- Factor pattern matrix: factor loadings
- Factor structure matrix: correlations between factors and observed variables

Factor analysis (3)

- Unlike PCA, factors can be rotated to make them more interpretable
- We are looking for simple structure (i.e. parsimony)
  - Each factor should affect as few variables as possible
  - Each variable should be explained by as few variables as possible
- Try to get factors to run through clouds of vectors
- Varimax: orthogonal rotation
- Promax: oblique rotation
- Factor scores: estimated values of latent variable for each of our observations

Multidimensional scaling (MDS)

- Definition: family of data analysis methods all of which portray the data structure in a spatial fashion, easily asimilated by the untrained eye (Young).
- Scaling for dissimilarities data (distances among cities, differences in perceptions of parties)
- Data: matrix of dissimilarities
- Purpose (Borg, Groenen, and Mair):
  - Visualize proximity/dissimilarity data
  - Uncover dimensions of judgment
- Analogy to map: MDS starts with distances and produces a map

Multidimensional scaling (2)

- Place objects in geometric space such that rank-order of distances between objects corresponds to rank-order of dissimilarities
- Much easier to interpret small number of points than a matrix of correlations among them!
- Input data can be ordinal or interval/ratio, but the output distances are interval/ratio either way
- Metric MDS: interval/ratio input. Distances are a linear function of dissimilarities
- Non-metric MDS: interval/ratio input. Distances are a monotonic function of dissimilarities.
- Better to have large number of points. It constrains the placement of the points more.

Wordfish

- Same principle : latent variable explains the number of times each word is used
- Developed by Slapin and Proksch (2008)
- Usually used for manifestos

Wordfish Model

$$y_{ij} \sim Poisson(\lambda_{ij})$$
$$\lambda_{ij} = exp(\alpha_i + \psi_j + \beta_j * \omega_i)$$

Eiffel tower of words (Slapin et Proksch 2008)