Module Introduction: Data Reduction and Index Construction

Joseph Nathan Cohen

Fall 2019

Data reduction is the task of transforming a data set by reducing its number of variables. Such situations are more common in psychology, where a surveyor might use tens of questions to measure one respondent trait. The Minnesota Multiphasic Personality Inventory (a well-known psychopathology test) asks 567 questions. To make sense of this many variables, we have to boil them down to a smaller number. The human mind can only process so much complexity, and reduction allows people to make sense of an analysis involving many questions.

Of course, one way to reduce the number of variables in a set is to throw variables away. Alternatively, an analyst reduce the number of variables in a set by blending them. The analyst can combine multiple variables by boiling them down to one variable, synthesizing them into composite measures or "indexes". Here, ***numerical indexes*** are numerical scores that summarize the behavior of multiple metrics.

Some well-known indexes:

- The Dow Jones Industrial Average is as a metric that represents the behavior of the whole stock market by averaging the price performance of 30 stocks.
- The United Nations' Human Development Index (HDI) is a metric that measures socio-economic development based on countries' life expectancies, mean/expected years of schooling, amd gross national income per capita.
- *The Economist* publishes a Quality-of-Life index that ranks places to live by a range of material wellbeing, health, political, and social metrics.
- ESPN's MLB Relative Power Index rates the strength of Major League Baseball teams, based on a team's winning percentage, opponents' winning percentage, and opponents' opponents' average winning percentage.

Illustration

The methods that we will learn this week can be used to develop and validate efforts to consolidate and reduce variables in a data set. Figure 1 (below) presents a depiction of a hypothetical indexation operation.

In this figure, we would describe "Academic Promise" as a *latent variable*, which is a variable that we cannot observe directly. It is latent – hiding beneath the surface. In contrast, GPA, SAT, and IQ scores are *observed variables* – we see these scores, because they are in our database.

The simple way to index these scores: (1) standardize each of these metrics (in order to put them on the same scale) and (2) take the average of the three standardized scores. However, doing so assumes that each of these three metrics are truly related to a common, underlying concept (academic promise), and are all equally important in measuring it. Moreover, this simple method assumes that you know how to measure "academic promise" in the first place – what if you have no idea? Factor analysis gives you a suite of tools to solve these kinds of problems.

We assume that our observed variables are related because of their association with a common underlying latent variable. In other words, we expect correlations between GPAs, SATs, and IQs because we presume that they are all outgrowths of one diffiult-to-observe quality: academic promise.

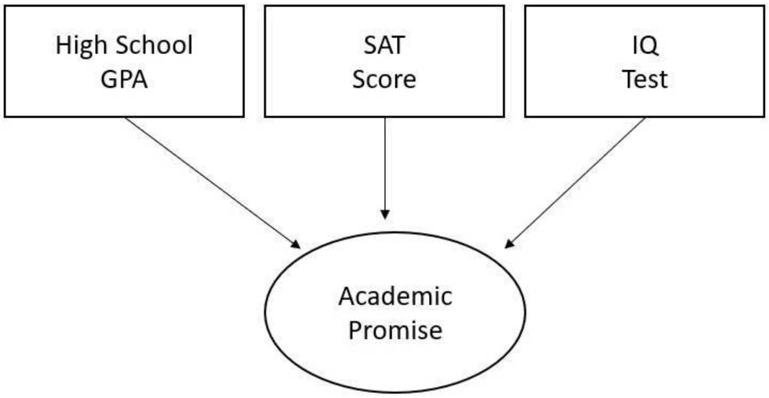

Figure 1: Depiction of a Hypothetical Data Reduction Operation

This Lesson

We discuss four methods for developing indexes:

- *Simple Indexation*, combining variables by averaging standardized variables
- *Cronbach's Alpha*, a quick-and-dirty (and commonly-used) method for justifying some combination of variables as related to an underlying latent variable.
- *Exploratory Factor Analysis*, which we use to search for common factors or latent variables in a larger number of variables.
- *Confirmatory Factor Analysis*, a more rigorous (and onerous) method for ascertaining whether or not we can infer that some latent variable underlies a collection of observed variables.

Data for this Lesson

In this module, we work two data sets. For our simple indexation lesson, we use data from the OECD *Better Life Index*, which can be downloaded here. The set measures 41 countries' overall quality-of-life using 24 different variables. You can learn more about the data set here.

For the remaining lessons, we use a generated data set that scores 400 respondents over 15 traits: happiness, optimism, sociability, anxiety, anger, jealousy, resentment, fear, boredom, tiredness, annoyance, irritability, hopefulness, friendliness, and ambition. These traits are all rated on a zero (inapplicable) to ten (fully applicable). We are interested in seeing whether we can reduce these 15 variables to a smaller number of factors. We are trying to see if we distill the major psychological states underlying these 15 variables. This set is called "Simulated Dispositional Data for EFA.xlsx", and can be downloaded here.

Both sets are in Excel format, and can be loaded using the **read_xlsx()** command in the ***readXL*** package. For example, to load the latter set:

```
library(readxl)
DATA <- read_xlsx("Simulated Emotional Disposition Data.xlsx", sheet = 1, col_names = TRUE)
```