Factor Analysis: A Short Introduction, Part 1

by Maike Rahn, PhD

Why use factor analysis?

Factor analysis is a useful tool for investigating variable relationships for complex concepts such as socioeconomic status, dietary patterns, or psychological scales.
It allows researchers to investigate concepts that are not easily measured directly by collapsing a large number of variables into a few interpretable underlying factors.

What is a factor?

The key concept of factor analysis is that multiple observed variables have similar patterns of responses because they are all associated with a latent (i.e. not directly measured) variable.

For example, people may respond similarly to questions about income, education, and occupation, which are all associated with the latent variable socioeconomic status.

In every factor analysis, there are the same number of factors as there are variables. Each factor captures a certain amount of the overall variance in the observed variables, and the factors are always listed in order of how much variation they explain.

The eigenvalue is a measure of how much of the variance of the observed variables a factor explains. Any factor

with an eigenvalue ≥1 explains more variance than a single observed variable.

So if the factor for socioeconomic status had an eigenvalue of 2.3 it would explain as much variance as 2.3 of the three variables. This factor, which captures most of the variance in those three variables, could then be used in other analyses.

The factors that explain the least amount of variance are generally discarded. Deciding how many factors are useful to retain will be the subject of another post.

What are factor loadings?

The relationship of each variable to the underlying factor is expressed by the so-called factor loading. Here is an example of the output of a simple factor analysis looking at indicators of wealth, with just six variables and two resulting factors.

Variables	Factor 1	Factor 2
Income	0.65	0.11
Education	0.59	0.25
Occupation	0.48	0.19
House value	0.38	0.60
Number of public parks in neighborhood	0.13	0.57
Number of violent crimes per year in neighborhood	0.23	0.55

The variable with the strongest association to the underlying latent variable. Factor 1, is income, with a factor loading of 0.65.

Since factor loadings can be interpreted like standardized regression coefficients, one could also say that the variable income has a correlation of 0.65 with Factor 1. This would be considered a strong association for a factor analysis in most research fields. Two other variables, education and occupation, are also associated with Factor 1. Based on the variables loading highly onto Factor 1, we could call it "Individual socioeconomic status."

House value, number of public parks, and number of violent crimes per year, however, have high factor loadings on the other factor, Factor 2. They seem to indicate the overall wealth within the neighborhood, so we may want to call Factor 2 "Neighborhood socioeconomic status."

Notice that the variable house value also is marginally important in Factor 1 (loading = 0.38). This makes sense, since the value of a person's house should be associated with his or her income.

***About the Author**: Maike Rahn is a health scientist with a strong background in data analysis. Maike has a Ph.D. in Nutrition from Cornell University.*