Data Reduction for making Comparisons: Principle Component Scores.

As published in Benchmarks RSS Matters, February 2015

http://web3.unt.edu/benchmarks/issues/2015/02/rss-matters

Jon Starkweather, PhD

Jon Starkweather, PhD
jonathan.starkweather@unt.edu
Consultant
Research and **S**tatistical **S**upport

http://www.unt.edu

http://www.unt.edu/rss

RSS hosts a number of “Short Courses”.
A list of them is available at:
http://www.unt.edu/rss/Instructional.htm

Those interested in learning more about R, or how to use it, can find information here:
http://www.unt.edu/rss/class/Jon/R_SC

Data Reduction for making Comparisons: Principle Component Scores.

Two years ago this column addressed one way of creating composite or indicator scores using Factor Analysis (Starkweather, 2012). That article approached composite score creation from a measurement modeling perspective in which each composite score represented a latent variable. The current article approaches composite score creation from a non-measurement modeling perspective.

This month's article discusses how to create composite scores from many variables using the ultimate data reduction technique: Principle Component Analysis (PCA). PCA is not measurement model based; it is a linear model based data reduction technique used to reduce the number of observed variables down to their *principle components* while maximizing the *total* amount of variance explained in the observed variables. PCA assumes linear relationships among the observed variables (i.e. it is not appropriate if curvilinear relationship are discovered among the observed variables). For a more detailed explanation of the differences between PCA and FA, please consult Starkweather (2010).

Occasionally, a data analyst is called upon to take many observed variables and combine them or reduce them to one variable or a few variables. The observed variables may, or may not, be directly related to one another and they may or may not be of the same scale. The one or a few resulting variables are weighted linear composite scores which can then be used to compare organizational units (e.g. departments within a larger organizational structure). In this situation, it is critically important to realize we are not interested in creating, assessing, or confirming a measurement model with latent variables and error. We are not assuming classical test theory model of measurement. We are solely interested in reducing many variables to one variable (or a few variables) so we can compare units. Those units may be individuals or organizations.

The Situation: *General Hospital*

Our example this month concerns a (*fictional*) General Hospital. The hospital board requested the director, Annabelle Lecter, M.D., to compare each Service Department. Each service department (Informational Services [IS], Therapeutic Services [TS], Diagnostic Services [DS], and Support Services [SS]) contains various disparate organizational structures (see pages 1 - 3 here[1]). The service departments do not initially seem comparable because each has specific tasks, budgets, number and status of personnel, degree of patient interaction, physical supply needs (weekly, monthly, yearly), and so forth. The director has access to a variety of these types of variables for each department and wants to reduce all of this information down to a single variable on which to compare the departments. Some departments have very small values on some variables by design or purpose (of the specific department) and some departments have very large values on some variables by design or purpose (of the specific department).

At first, the director thinks it might be best to transform all these variables to Z-scores (i.e. standardize them) so they are all on the same scale and then simply add or average all the Z-scores to get one number for each department. The directory quickly realizes this is not tenable because Z-scores, although used to compare individuals across two (or more) variables, are not meant to be combined. If Z-scores are averaged, the mean should be at or very near zero. Furthermore, creating a composite score using either of these two techniques (sum or mean) explicitly assume each variable is equally important and

[1] `http://www.quia.com/files/quia/users/kkacher/OrganizStHsp/Org-St-Lesson-Pln`

essentially interchangeable (with respect to the resulting composite score).

What the director really needs is a technique which creates a composite score (for each department) in such a way that each observed variable is weighted by its ability to account for variance in all the observed variables (combined). The *variance in all observed variables* is represented by the variance-covariance matrix or correlation matrix of observed variables. By submitting the observed variables' data (i.e. variance-covariance matrix or correlation matrix) to PCA and specifying the computation of Principle Component Scores (PCS) and then saving the scores of the *first* component, the director will have achieved her goal. Keep in mind, with PCA the first component is the one which accounts for the most variance and any subsequent components are accounting for variance *left over* after the variance which was accounted for by previous component has been removed. So, just to be clear; if the first component accounts for 48% of the variance of the observed variables, then that is 48% of 100% of the variance of the observed variables. If the second component accounts for 25% of the variance then that is 25% of the remaining 52% total variance of the observed variables (i.e. whatever is left after the first component has been extracted). So each subsequent component (i.e. component 3 through component *J* - 1, where *J* is the number of observed variables) is accounting for less and less of the total observed variables' variance.

Now you may be asking the question; "but what does the component score *mean*?" In order to determine that, one would evaluate the direction and magnitude of loadings of each observed variable to the first component. The variables which have the largest absolute value loadings are those most contributing to the component (i.e. accounting for the most variance in all the observed variables). Loadings are interpreted just like correlation coefficients – positive vs. negative and between -1 and +1. If more than one component is evaluated, it is very likely the observed variables will coalesce on one or the other component decisively with each component's definition (or name) becoming apparent based on which observed variables load most on a particular component. For example, say that the observed variables 1, 3, 5, 7, and 9 load most on the first component; while the observed variables 2, 4, 6, 8, and 10 load most on the second component. Then we would name the first component based on the content or meaning of observed variables 1, 3, 5, 7, and 9. Likewise, we would name the second component based on the content of the observed variables 2, 4, 6, 8, and 10.

Tutorials using PCA (with and without saving component scores) are available for each of the three most popular statistical software packages through the Research and Statistical Support instructional / tutorial websites (links provided directly below).

For users of the statistical programming language environment R, please see:
`http://www.unt.edu/rss/class/Jon/R_SC/Module7/M7_PCAandFA.R`

For users of the SAS programming suite, please see:
`http://www.unt.edu/rss/class/Jon/SAS_SC/SAS_Module7.htm`

For users of the SPSS program, please see:
`http://www.unt.edu/rss/class/Jon/SPSS_SC/Module9/M9_PCA/SPSS_M9_PCA1.htm`

References

Hospital Organizational Structure:

http://www.quia.com/files/quia/users/kkacher/OrganizStHsp/Org-St-Lesson-Pln

Starkweather, J. (2010). Principal Components Analysis vs. Factor Analysisand Appropriate Alternatives. Available in original form at Benchmarks: http://it.unt.edu/benchmarks/issues/2010/07 and available as an Adobe.pdf here: http://www.unt.edu/rss/class/Jon/Benchmarks/PCAvsFA

Starkweather, J. (2012). How to Calculate Empirically Derived Composite or Indicator Scores. Available in original form at Benchmarks: http://web3.unt.edu/benchmarks/issues/2012/02/rss-matt and available as an Adobe.pdf here: http://www.unt.edu/rss/class/Jon/Benchmarks/Composi

This article was last updated on February 16, 2015.

This document was created using LaTeX