SOME COMMENTS ON MULTIVARIATE TECHNIQUES

Path Analysis

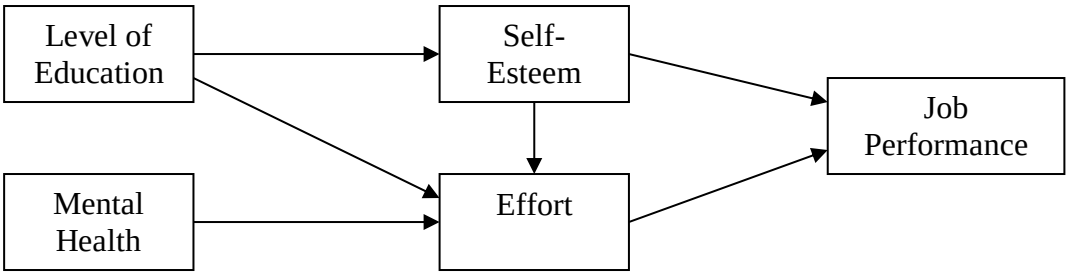
Path analysis is an extension of multiple regression analysis. A typical use of regression is to assess the relationship between one variable and a set of other variables – or to predict a single variable from one or more other variables. In the typical path analysis, more than one variable is considered dependent. In these models, the predictive ordering is of interest. In regression, we say: X explains variation in Y. In path models, a causal model is being explained where we hope to say: X causes Y, which causes Z.

Because of the demands of the multivariate estimation and simultaneous parameter estimation (there are many parameters being estimated in a typical path analysis) the data information demands are greater, requiring 300 cases or more for most models.

Recall that one structural assumption of regression is that "the right variables are in the model." This requirement is even greater in path analysis, because the results of a path analysis are highly sensitive to the presence or absence of important variables. The estimates of parameters are conditional on the correct specification of the model. One additional benefit with path analysis models is the availability of tests of model-data fit. Fit statistics are readily available based on any number of assumptions or data conditions the researcher is able to make.

The labels "dependent" and "independent" are replaced by "endogenous" and "exogenous." In path analysis, a variable can be both dependent and independent – there are multiple equations being estimated simultaneously. The values of exogenous variables are considered known – there is no need for estimation here. The endogenous variables are considered to be caused by one or more of the other variables. If a variable is being explained by another variable at any point in the model, it is considered endogenous.

Path coefficients are estimated in a method similar to multiple regression for each endogenous variable. The model is expected to explain variation in these variables. Consider the model:



There are three regressions required here because there are three endogenous variables (self-esteem, effort, and job performance), which are the dependent variables of the three equations. There are also two exogenous variables.

There is a unique distinction among path models based on the recursive nature of the model. Fully recursive models contain direct effects on all variables farther down the causal chain. The model above is not fully recursive because level of education and mental health do not have direct effects on job performance, only indirect through self-esteem and effort. In addition, mental health does not have an effect on self-esteem.

Fully recursive models always fit the observed data perfectly. In addition, recursive models contain paths (effects) that all go in one direction. Nonrecursive models contain paths that go both ways – the causation occurs in both directions.

Another important consideration is model identification, which is a prerequisite to estimation. Recall that for an equation such as

$$x + y = 10$$

there are an infinite number of solutions. We have two unknowns and one equation. We need another equation with one or both of these unknowns to complete estimation or obtain a single solution. When the parameters of a model can be uniquely determined, the model is identified.

Models with more unknowns than pieces of information are called under-identified and cannot be solved uniquely. Models with the same number of unknowns as pieces of information are called just identified models, can be solved, but cannot be tested statistically. Models with more information than unknowns are called over-identified models – these are typically called identified and can be uniquely solved and tested statistically.

A common structural equation model is confirmatory factor analysis where items are a prior hypothesized to be resulting from one or more underlying factors (latent traits). The observed variables (X) result from factors (ξ). The paths from factors to the observed variables are factor loadings (λ). Correlations and covariances (Φ) among the factors are also estimated. Covariances among the measurement errors (δ) are contained in the matrix $\theta\delta$. There exists, then, one equation for the relationship between each observed variable and its factor:

$X_1 = \lambda_{11}\xi_1 + \delta_1$ → which is analogous to the regression equation $Y = \beta_1 X_1 + e$

Since in the structural equation model, the ξs are latent variables (not directly observed), we cannot use regular regression methods to estimate the parameters of the model. Instead, a correlation or covariance matrix is used (Σ). This matrix is typically found as:

$$\Sigma = \Lambda\Phi\Lambda + \theta\delta$$

In confirmatory factor analysis, the unknowns are found in Λ, Φ, and $\theta\delta$. The information available to obtain the solution includes the elements in Σ. The number of unique elements in Σ is $p(p+1)/2$, where p is the number of observed variables. The task is then to find estimates of the unknown matrices that will best reproduce Σ. In regression we find βs to reproduce the Ys.

Principal Components

The goal of principal component analysis (PCA) is to reduce the number of variables in a data set to an essential core set of variables (typically orthogonal) that contains or explains nearly all of the variance in the original set of variables. Is there a smaller subset of variables?

$\underline{y} \sim N(\mu, \Sigma)$

A principal component closely resembles a regression equation on standardized variables – there is no constant term (intercept).

$Y = 0.2(X_1) + 0.1(X_2) + .5(X_3)$

This is a linear combination of the variables that maximizes the amount of total variance explained.

$z_1 = a_{11}y_1 + \ldots + a_{p1}y_p = \underline{a}'\underline{y}$

$s^2_{z1} = \underline{a}'\, V(\underline{y})\, \underline{a} = \underline{a}'\, \Sigma\, \underline{a} \qquad \rightarrow (1\times p)(p\times p)(p\times 1)$

Consider the geometry involved here.

The principal component is a vector that passes through the observed data points in a multidimensional space of p dimensions (number of variables). Each observation lies at a given distance perpendicularly from the vector (1st principal component). This distance is the error score for each observation. The 1st principal component minimizes the sum of squared errors.

This vector, the 1st principal component, is also an eigenvector – the amount of variance explained by each eigenvector is the eigenvalue.

If there were 10 variables, the total variance would be equal to 10, since the variables would be normalized with variance equal to one. If the variance explained by the 1st principal component is 5, this would indicate that the eigenvector accounts for the same amount of variance as five of the original variables. This is 50% of the variance.

A second principal component could be found that again maximizes the amount of remaining variance. This principal component would be perpendicular, orthogonal, to the first component. So the proportion of variance explained is additive.

The number of components required to explain 100% of the variance is the rank (true dimensionality) of the correlation matrix used to summarize the interrelationships of the variables. Usually, the number of components extracted includes all of those with eigenvalues equal to one or more – one is the minimum because it explains as much variance as a single variable so no gain is obtained by employing components with less variance than a single variable– the point is to reduce the number of variables. This is typically done until 75% or more of the variance has been explained.

It is typical to use the correlation matrix to extract components – particularly when the variables employed vary greatly in terms of their scale – where the correlation matrix standardizes variables in terms of scale. Statistically, it is preferred to use the covariance matrix, but only when the scales are similar. The results are not the same.

Typically, a sample size of 100 is a good basic minimum. However, beyond this, the total sample size should be 10 times the number of variables. So if you have 20 variables, a sample size of 200 should secure stable estimates. This is qualified by the degree of variability in the sample as a whole. As in all other linear models, the value of the information from the sample is directly related to the variability in the sample. With greater variability, fewer observations will still yield stable estimates.

Often, PCA is confused with factor analysis. Principal components analysis is typically conducted with a correlation matrix where ones are on the main diagonal or the diagonal contains variances when a variance-covariance matrix is employed. In factor analysis, values that are less than one are placed on the diagonal – these are communalities, the squared multiple correlation of each variable with all others is one form of a communality.

In practice, with enough subjects and variables (20 or more), there is little difference in the results of any method of extraction or what is placed in the diagonal of the matrix for extraction.

General Results

Trace (S) = $\lambda_1 + \lambda_2 + \ldots + \lambda_p$ → the sum of variances

Where A is nonsingular, square and full-rank, Rank(A) = p such that $|A| \neq 0$.

$$\text{Det}(A) = \prod_{i=1}^{P} \lambda_i \qquad \text{Trace}(A) = \sum_{i=1}^{P} \lambda_i$$

Thinking about the Determinant as the product of the eigenvalues highlights the idea that the determinant informs us about the degree to which a set of variables provides relatively unique nonredundant information (a large determinant, indicating relative orthogonality) or a set of variables that are largely redundant (small determinant near zero, indicating multicolinearity). Thinking about the Trace as the sum of the eigenvalues highlights the idea that the trace informs us regarding the total amount of variance available to be analyzed in a set of variables – particularly useful when the variables are largely orthogonal.

These considerations work because an eigenvalue can be thought of as the variance of a linear combination of variables. The specific amount of variance taken from each variable is represented by the eigenvector weight. Since the Eigen values sum to the total variance, eigenvalues can also be thought of as redistributions of the variances of the set of variables.

Multivariate Hypothesis Tests

Pillai: $\sum_{i}^{s} \frac{\lambda_i}{1+\lambda_i}$ Hotelling: $\sum_{i}^{s} \lambda_i$

Wilks λ: $\prod_{i}^{s} \frac{1}{1+\lambda_i}$ Roy's Largest Root: $\theta = \frac{1}{1+\lambda_i}$

Multivariate Effect Sizes

MANOVA:, the multivariate analysis of variance (multiple dependent variables)

One effect size for MANOVA is eta squared: η^2, which represents the proportion of variance in the resulting linear combination of the dependent variables that is explained by the groups represented in the independent variables.

This can be computed from Wilks Lambda: $(1 - \lambda)$

Cohen provided guidelines for multivariate effect sizes, such that .02 is a small effect, .13 is a medium effect, and .26 is a large effect.

*Discriminant function analysis is appropriate as a follow-up to MANOVA. This would identify the continuous variables that were discriminating among the groups most strongly.

Discriminant Function Analysis

An index of effect size that would be appropriate for Discrimination Functions would be the squared discriminant loadings as the proportion of shared variance between a variable and the underlying latent variable or discriminant function.

For more information on effect sizes and power related issues, see; Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155-159.

Explaining the variance-covariance structure of a set of variables through linear combinations of the variables is the goal of principal components analysis. The two primary purposes of this technique are (a) data reduction and (b) interpretation. This procedure is frequently used because there are (a) too many independent variables relative to the number of observations and (b) some variables are highly correlated, producing unstable estimates.

P components are required to reproduce all of the variability in the system; however, most of the variability can usually be accounted for by fewer than *p* components. If there are *k* components that account for most of the variability of the *p* variables, then we say that the data set has been reduced to *k* principal components.

Principal components consist of linear combinations of the *p* variables in the data set. Geometrically, the linear combinations represent a new coordinate system that was obtained by rotating the original system to maximize the variability of a simpler representation of the *p* coordinate axes (from the original *p* variables in the system).

Let $\underline{x}' = (x_1, x_2, \ldots, x_3)$ with the variance-covariance matrix $\boldsymbol{\Sigma}$ (sigma) and eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$.

Consider linear combinations of the $\underline{x}$s:

$$
\begin{aligned}
Y_1 &= \underline{a}'_1 \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1p}X_p \\
Y_2 &= \underline{a}'_2 \mathbf{X} = a_{21}X_1 + a_{22}X_2 + \ldots + a_{2p}X_p \\
&\vdots \\
Y_p &= \underline{a}'_p \mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \ldots + a_{pp}X_p
\end{aligned}
$$

Then we can obtain

$$\text{Var}(Y_i) = \underline{a}'_i \boldsymbol{\Sigma} \underline{a}_i \text{ and } \text{Cov}(Y_i, Y_k) = \underline{a}'_i \boldsymbol{\Sigma} \underline{a}_k$$

The principal components are the uncorrelated linear combinations $Y_1, Y_2, \ldots, Y_p$ with variances as large as possible.

The first principal component is the linear combination with the maximum variance; it maximizes $\text{Var}(Y_1) = \underline{a}'_1 \boldsymbol{\Sigma} \underline{a}_1$. Of course, the variance can be increased by multiplying any $\underline{a}$ by a constant. To avoid this indeterminacy, coefficient vectors are constrained to be unit length.

First principal component = linear combination $\underline{a}'_1 \mathbf{X}$ that maximizes $\text{Var}(\underline{a}_1 \mathbf{X})$ so that $\underline{a}'_1 \underline{a}_1 = 1$.
Second principal component = linear comb. $\underline{a}'_2 \mathbf{X}$ that maximizes $\text{Var}(\underline{a}_2 \mathbf{X})$ so that $\underline{a}'_2 \underline{a}_2 = 1$ and
$\text{Cov}(\underline{a}'_1 \mathbf{X}, \underline{a}'_2 \mathbf{X}) = 0$.

Etc…