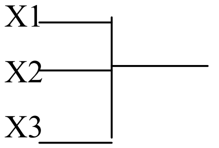
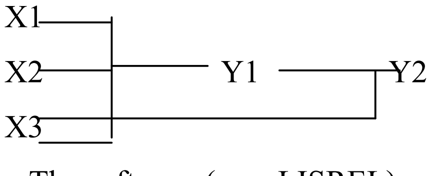
Path analysis is a data analysis technique that quantifies the relative contributions of variables ("path coefficients") to the variation in a focal variable once a certain network of interrelated variables has been specified (Lynch & Walsh 1998, 823). Some of these contributions are direct and some mediated through other variables, i.e., indirect. Although some researchers interpret "contribution" in causal terms (e.g., Pearl 2000, 135 & 344-5), others criticize such an interpretation (e.g., Freedman 2005). Here, contribution refers neutrally to the term of an additive model fitted to data.

The conceptual starting point for path analysis is an additive regression model that associates the focal ("dependent") variable with several other measured ("independent" or "exogenous") variables.



Technically, the additive model is transformed by subtracting the mean from every term, squaring the expression (so it is an equation for the variance), and dividing by the variance of the focal ("dependent") variable. The result is the "equation of complete determination," with the regression coefficients being multiplied by the SD of the other "independent" variables and divided by the SD of the focal variable to arrive at the path coefficient.

The next step is to consider more than one focal, "endogenous" variable and networks of exogenous and endogenous variables that you have reason to think are associated with one another. Indeed, the focal variable of one regression may be among the variables associated with a second focal variable and so on. In the figure below X3 has a direct link with Y2 and an indirect one through Y1.



The software (e.g., LISREL) can solve these linked regression equations, but it is up to you to compare the results using the network you specify with plausible (theoretically-justified) alternatives that may link exogenous, independent variables and endogenous variables differently. Unlike multiple regression, we do not arrive at our idea of what should be in the regression by adding or subtracting variables in some stepwise procedure.

Structural equation modeling extends path analysis to include latent (a.k.a. unmeasured) variables or “constructs.” These latent variables are sometimes the presumed real underlying variable of which the measured one is an imperfect marker. For example, birth weight at full term and the neonate APGAR scores* might be the measured variables but the model might include degree of fetal under-nutrition as a latent variable. Latent variables can also be constructed by the software in the same way that they are in factor analyses, namely, as economical (dimension-reducing) linear combinations of measured variables. Calling the networks of linked variables “structural” is meant to suggest that we can give the pathways causal interpretations, but SEM and path analysis has no trick that overcomes the problems that regression and factor analyses have in exposing causes.

This section is not needed for understanding the papers for this week. However, looking ahead to studies of heritability (part of week 12), a field in which path analysis originated, there are no measured variables except the observed focal variable (e.g., height). Path analysis can still be used if we convert the additive model on which any given Analysis of Variance (AOV or ANOVA) is based into an additive model of constructed variables that take the values of the contributions fitted to the first model. For example, in an agricultural evaluation trial of many varieties raised in many locations, the AOV model is

$$y_{ijk} = m \quad +v_i \quad +l_j \quad +vl_{ij} \quad +e_{ijk} \qquad (1)$$

where y_{ijk} denotes the measured trait y for the i^{th} variety in the j^{th} location and k^{th} replication;

m is a base level for the trait;

v_i is the contribution of the i^{th} variety;

l_j is the contribution of the j^{th} location;

vl_{ij} is an additional contribution from the i,j^{th} variety-location combination—in statistical terms, the “variety-location-interaction” contribution; and

e_{ijk} is a noise contribution adding to the trait measurement.

The path model equivalent to equation 1 is

$$y_x = m \quad +z_{1x} \quad +z_{2x} \quad +z_{3x} \quad +e_x \qquad (2)$$

where

y is the measured trait as before and x denotes the replicates

$z_{1x} = v_i$ if x if a replicate of variety i, or 0 otherwise

$z_{2x} = l_j$ if x if a replicate in location j, or 0 otherwise

$z_{3x} = vl_{ij}$ if x if a replicate of variety i in location j, or 0 otherwise

$e_x = e_{ijk}$ where x is replicate k of variety i in location j

The path coefficients are then set to equal the square root of the ratio of the variance of the contribution (v_i, etc.) to the total variance for the trait (Y). The equation of complete determination becomes

$$1 = \Sigma \text{ variance } (z_w) / Y \qquad (3)$$

where w denotes the different contributions in the Analysis of Variance model.

For the agricultural trial this equation might be written

$$1 = (V + L + VL + E) / Y \qquad (4)$$

where V = variance of the v_i terms, etc.

In human studies the VL is ignored and this is expressed as

$$1 = \text{heritability} + \text{shared environmental effect} + \text{non-shared environmental effect} \qquad (5)$$

When the same trait is observed in two relatives, their separate path analyses can be linked in one network and the correlation between the relatives calculated (Lynch & Walsh 1998, 826)—provided it is assumed that the contributions (and path coefficients) apply to both and that the noise contributions are uncorrelated. If we have data on correlations for different kinds of relatives (e.g., identical vs. fraternal twins), we can estimate the relative size of the contributions in equations such as 4 and 5. That's the crux of heritability studies.

References

Freedman, D. A. (2005). Linear statistical models for causation: A critical review. Encyclopedia of Statistics in the Behavioral Sciences. B. Everitt and D. Howell. Chichester, Wiley.

Lynch, M. and B. Walsh (1998). Genetics and Analysis of Quantitative Traits. Sunderland, MA, Sinauer.

Pearl, J. (2000). Causality: Models, Reasoning, and Inference. Cambridge, Cambridge U. Press.

* http://en.wikipedia.org/wiki/Apgar_score