Structural Equation Modeling (Pizza and Pilots Presentation)

Dr. Richard Charnigo
Professor of Statistics and Biostatistics
RJCharn2@aol.com
07 December 2015

Motivation

Structural equation modeling is useful for data analysis entailing one or both of two features that render ordinary regression modeling inapplicable:

1. There is not a clear dichotomy of "independent variables" versus "dependent variables" because some variables both predict and are predicted by other variables in the modeling.

Motivation

This may happen in a longitudinal study if the value of a variable at an intermediate time point both predicts its value at a later time point and is predicted by its value at an earlier time point.

This may happen in a cross-sectional study or a longitudinal study if a variable is a "mediator". For instance, rumination may be a mediator between stress and depression.

So, rather than using "independent variable" and "dependent variable", we refer to variables predicted by other variables in the modeling as "endogenous" and to variables not predicted by other variables in the modeling as "exogenous".

Motivation

2. What we are mainly interested in is not directly observable but is measured, imperfectly, using one or more instruments. Such instruments are often scales obtained by summations of Likert items.

   In this context, that which is not directly observable is often called a "latent construct", while a corresponding instrument is often referred to as an "observable indicator".

A Hypothetical Example

Consider a longitudinal study of college students who have a history of drinking in high school. The goal of the study is to ascertain whether sensation seeking may influence drinking, whether drinking may influence sensation seeking, or both.

Let SS_1 denote sensation seeking at the beginning of college and DR_1 drinking at the beginning of college.

These are exogenous variables in that they will not be predicted by any other variables that we will include in our structural equation model. They are also latent variables because we do not observe them.

A Hypothetical Example

Let SS_2 denote sensation seeking after one year of college and DR_2 drinking after one year of college. Let SS_3 denote sensation seeking after two years of college and DR_3 drinking after two years of college.

These are endogenous variables in that they will be predicted by other variables that we will include in our structural equation model. They are also latent variables because we do not observe them.

A Hypothetical Example

Let NS_1 be the score observed on a scale of novelty seeking;

IM_1 the score observed on a scale of impulsivity;

DF_1 the score observed on a scale of drinking frequency; and

DI_1 the score observed on a scale of drinking intensity, with all of these scales administered at the beginning of college.

Thus, NS_1 and IM_1 are observable indicators of the latent construct SS_1, while DF_1 and DI_1 are observable indicators of the latent construct DR_1.

A Hypothetical Example

Let NS_2 be the score observed on a scale of novelty seeking;
IM_2 the score observed on a scale of impulsivity;
DF_2 the score observed on a scale of drinking frequency; and
DI_2 the score observed on a scale of drinking intensity, with all of these scales administered after one year of college.

Let NS_3 be the score observed on a scale of novelty seeking;
IM_3 the score observed on a scale of impulsivity;
DF_3 the score observed on a scale of drinking frequency; and
DI_3 the score observed on a scale of drinking intensity, with all of these scales administered after two years of college.

A Hypothetical Example

The next slide depicts relationships among these quantities. Those labeled with "e" or "d" symbols represent random deviations from what is predicted and may be called, for lack of a better word, "errors".

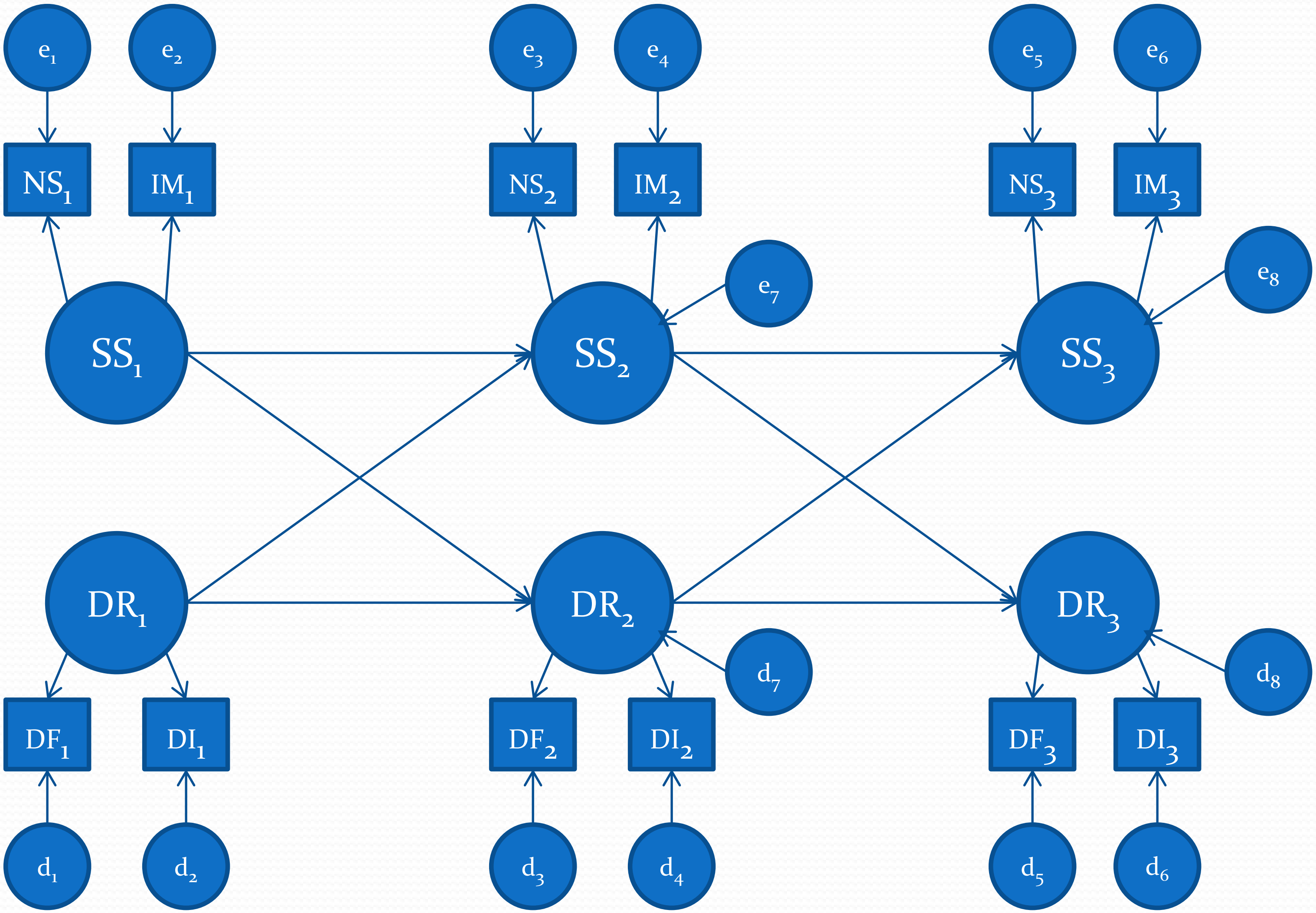
Generally speaking, a given quantity is expressed as a linear combination of those quantities which contribute to it, not unlike what is seen in regression. Thus, for example,

$$SS_2 = b_0 + b_1\,SS_1 + b_2\,DR_1 + e_7$$

and

$$DF_3 = a_0 + a_1\,DR_3 + d_5.$$

A Hypothetical Example



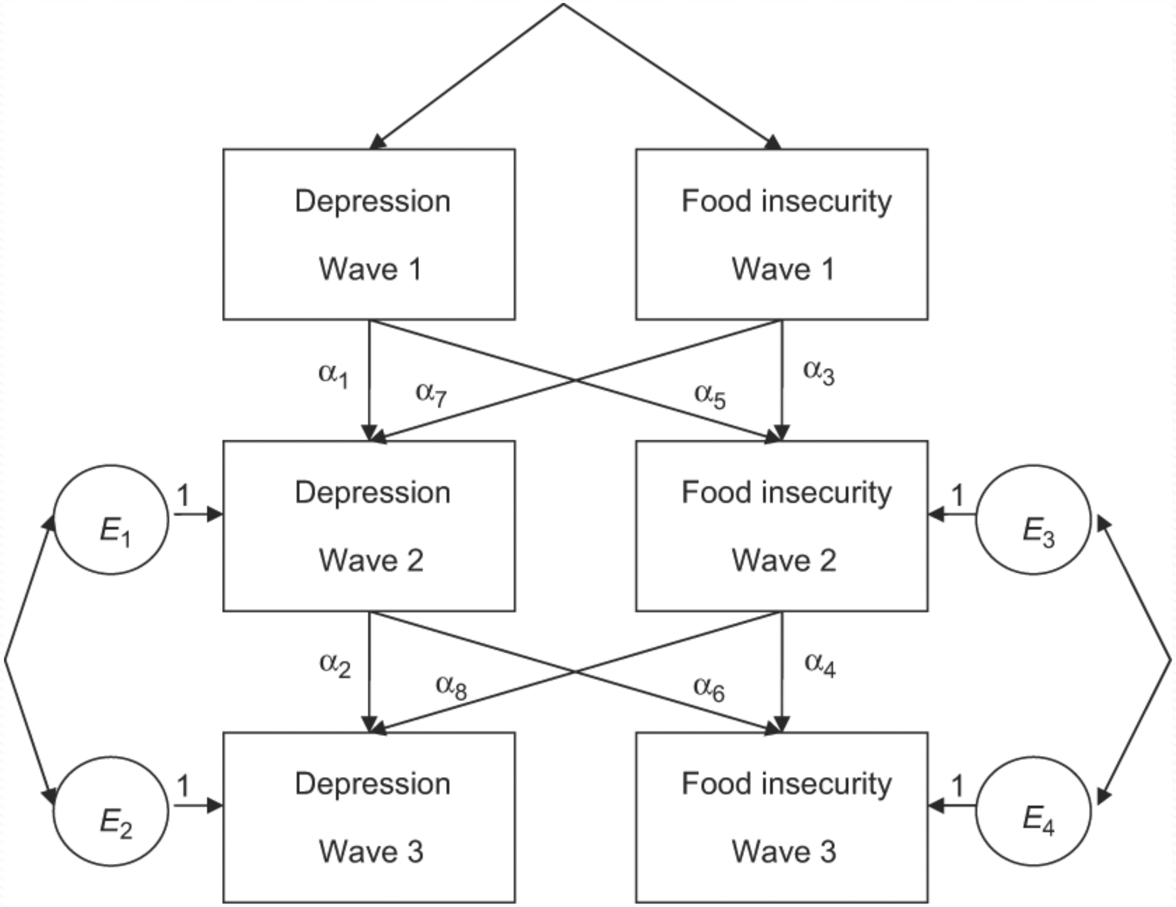
An Actual Example

Sometimes a structural equation model is not as complicated as indicated on the preceding page. For example, if there is only one observable indicator for each latent construct, then the observable indicator may be identified with its latent construct.

An example of this occurs in my 2009 *Public Health Nutrition* paper with Huddleston-Casas and Simmons. We were trying to understand whether food insecurity predicted depression, or vice versa, or both, using three waves of survey data from low-income rural mothers.

The diagram, derived from Figure 1 of the paper, appears next.

An Actual Example



An Actual Example

Some results, from Table 2 of the paper, are as shown below. Also, testing the null hypothesis that $\alpha_5 = \alpha_6 = 0$ yields P=0.034 in the complete case analysis (P<0.001 in the imputed analysis), and we have P=0.003 (P<0.001) in testing $\alpha_7 = \alpha_8 = 0$.

	Complete case analysis				Analysis with missing values imputed			
Coefficient	Estimate	SE	Standardized estimate	*P*	Estimate	SE	Standardized estimate	*P*
α_1	0·426	0·061	0·442	<0·001	0·468	0·036	0·528	<0·001
α_2	1·015	0·146	0·858	<0·001	0·870	0·064	0·853	<0·001
α_3	0·486	0·060	0·505	<0·001	0·500	0·036	0·565	<0·001
α_4	1·034	0·144	1·053	<0·001	0·911	0·073	0·962	<0·001
α_5	0·006	0·004	0·077	0·132	0·006	0·003	0·082	0·023
α_6	0·008	0·005	0·103	0·102	0·011	0·003	0·139	<0·001
α_7	2·436	0·708	0·212	<0·001	2·103	0·437	0·193	<0·001
α_8	−0·826	0·885	−0·058	0·350	−0·379	0·477	−0·030	0·427

An Actual Example

The conclusion is that the relationship between food insecurity and depression among low-income rural mothers may be bidirectional.

And, just as linear regression models have R^2 and other tools to assess goodness of fit, there are various indices of fit for structural equation models. Three of them are χ^2 / df, RMSEA, and CFI.

In the complete case analysis from our 2009 paper, we had χ^2 / df = 1.835, RMSEA = 0.068, and CFI = 0.989. These are favourable values for the indices.

Remarks

Goodness of fit is often substantially affected by the assumptions one makes regarding which error terms are correlated. Since that is difficult to ascertain from subject matter theory, some iterative process (not unlike stepwise selection in linear regression modeling) is often helpful.

Typically we assume that quantities involved are (close to) normally distributed. However, a categorical variable can be accommodated in the sense that model parameters may be estimated for and compared across multiple groups defined by the categorical variable. We did this, too, in our 2009 paper.

References

Bollen and Long (1993). *Testing Structural Equation Models*. SAGE, Newbury Park CA.

Hoyle (1995). *Structural Equation Modeling: Concepts, Issues, and Applications*. SAGE, Thousand Oaks CA.

Huddleston-Casas, Charnigo, and Simmons (2009). Food insecurity and maternal depression in rural, low-income families: a longitudinal investigation. *Public Health Nutrition*, **12**, 1133-1140.

Kline (2010). *Principles and Practice of Structural Equation Modeling, 3rd Edition*. Guilford, New York NY.

Mueller (1996). *Basic Principles of Structural Equation Modeling*. Springer-Verlag, New York NY.