# Notes on Latent Variable Modeling

compiled by D.Gueorguiev 3/31/2024

## Introductory Notes

The Ideal experiment: a single independent variable is manipulated, and the consequences are observed in a single dependent variable.

Experiments in reality : the variables which are observed are typically not the one of real theoretical interest but are merely some convenient variables acting as proxies. A full analysis will turn out to be multivariate, with a number of alternative experimental manipulators on the one side, and a number of alternative response measures on the other.

There is a variety of statistical techniques for dealing with situations in which multiple variables, some of which unobserved, are involved. In [1] are discussed a variety of methods with the following common features:

( a ) multiple variables – three or more - are involved

( b ) one or more of these variables are unobserved i.e. *latent*

*Latent variable analysis*, discussed in [1], encompasses specific methods such as factor analysis, path analysis and structural equation modeling applied to (a) and (b).

### Path Models in Factor, Path, and Structural Equation Analysis

#### Path Diagrams

*Path diagram* is a representation of the relationships among a number of variables. We use capital letters to denote variables in such diagram. The connection among variables are represented in path diagrams by two kinds of arrows : a straight, directed arrow represents *a causal relationship* between two variables, while a curved bidirectional arrow represents *a correlation* between the variables which it connects.

Figure 1: example of a path diagram, Modeling intelligence in single child family

Variables , , and are all assumed to have causal effects on . Variables and are assumed to be correlated with each other. Variable is assumed to affect but to be uncorrelated with either or .

What the diagram on Figure 1 might model?

Possible representation of the diagram on Figure 1: Modeling intelligence in single child family

would represent the child intelligence, and would represent father’s and mother’s intelligence which are assumed to have causal influence on child intelligence. The intelligence of the mother and father is correlated. represents the other variables, independent on the father and mother intelligence, which influence child’s intelligence.

Figure 2 shows another path diagram. Here is assumed to affect both and , and each of the latter variables is also affected by an additional variable – U and V respectively.

Possible representation of the diagram on Figure 2: reliability of psychometric test

and would represent scores on two alternate forms of a test. would represent the unobserved true score on the trait being measured, which is assumed to affect the observed scores on both forms of the test. U and V would represent factors specific a) to each form of the test or b) to the occasions on which it was administered. These additional factors would affect any given performance, but it would be unrelated to the true trait.

*Note*: the variance in and resulting from the influence of would be called *true score variance* and that caused by or would be called *error variance*. The proportion of the variance of or due to would be called the *reliability of the test*.

Figure 2: Reliability of psychometric test

Figure 3 shows a path representation of events over time. In this case and denote two separate variables with subscript denoting a time sequence in which they are measured. Both and are measured in time 1, is measured again in time 2 and at time 3. In this case the diagram indicates that both and are assumed to affect , but that the effect of on at time 3 is entirely via as there is no direct arrow from to . It is assumed that and are correlated, and that and are subject to additional influences independent of and . These additional influences could be labeled (say and ), but often are left unlabeled as in Figure 3 to indicate that they refer to other , unspecified influences of the variables to which they point. Such arrows are called *residual arrows* to indicate that they represent causes residual to those explicitly identified in the diagram.

Figure 3: path diagram involving events over time

##### The meaning of “cause” in a path diagram

The essential assumption made when we use causal arrow in a path diagram is that a change in the variable at the tail of the arrow will result in a change in the variable at the head of the arrow , all else being equal. This influence is one way only – imposing a change on the variable at the head of the arrow does not bring a change in the tail variable.

##### Completeness of a path diagram

Variables in a path diagram may be grouped in two classes: those that a) do not receive causal inputs from any other variable in the path diagram, and b) those that receive one or more such causal inputs. Variables in the first of these two classes are referred to as *exogenous*, *independent* or *source variables*, as those variables represent causal sources. *Endogenous variables* have at least some causal sources that lie within the path diagram; these variables are *causally dependent* on other variables as one or more straight arrows lead into them. Such variables lie *causally downstream* from source variables.

In a *proper and complete* path diagram, all the source variables are interconnected by curved arrows, to indicate that they may be inter-correlated unless it is explicitly assumed that they are uncorrelated.

Downstream variables, on the other hand, are never connected by curved arrows in path diagrams. Actually, sometimes downstream curved arrows are used as shorthand to indicate correlations among downstream variables caused by other variables than those included in the diagram. We use correlations between residual arrows for this purpose, which is consistent with our convention because the latter are source variables.

Residual arrows point at downstream variables and never at source variables. *Completeness of a path diagram* requires that a residual arrow be attached to every downstream variable unless it is explicitly assumed that all the causes of variation of that variable are included among the variables upstream from it in the diagram.

Note: The last rule is not universally adhered to. Occasionally, path diagrams are published with the notation “residual arrows omitted” which leads to ambiguity.

Figure 4: Path diagrams illustrating the implication of an omitted residual arow

(a)

(b)

Figure 4 show an example in which the presence or absence of a residual arrow makes a difference. The source variables and refer to the genetic and environmental influences on a trait . The downstream variable in Figure 4 (a) has no residual arrow. That represents the assumption that the variation of is completely explained by the genetic and environmental influences upon it. Figure 4 (b) represents the counter-assumption that genetic and environmental influences are not sufficient to explain the variation of – some additional factor or factors, perhaps measurement error or gene-environment interaction – may need to be taken into account in explaining .

Finally, all significant direct causal connections between source and downstream variables, or between one downstream variable and another, should be included as straight arrows in the diagram. Omission of an arrow between and in Figure 3 carries information: is assumed to affect only via .

Note: most path diagrams could be extended indefinitely past their source variables – these could be taken as downstream variables in an extended path diagram, and the correlations among them explained by the linkages among their own causes. Thus the parents in Figure 1 could be taken as children in their own families and the correlation between them explained by a model of the psychological and sociological mechanisms that result in mates having similar IQs. In Figure 3 one could have measured and at a preceding time zero, resulting in a diagram in which the correlation between and is replaced by a superstructure of causal arrows from and , themselves probably correlated.

##### Other assumptions in path diagrams

It is assumed in path diagrams that causes are unitary, that is, in a case such as Figure 2 that it is meaningful to think of a single variable that is the cause of and , and not (say) two separate and distinct aspects of a phenomenon , one of which causes and one . In the latter case, a better representation would be to replace by two different (possibly correlated) variables.

An exception to the rule of unitary causes is residual variables, which typically represent multiple causes of a variable that are external to the path diagram.

*Note*: for this reason, we do not always solve for the path coefficients associated with the residual arrows in their diagrams. It is , however, a good practice

## References

[1] [Latent Variable Models: Introduction to Factor Analysis and Structural Equation Analysis, John C. Loehlin, 2004](https://github.com/dimitarpg13/information_theory_and_statistical_mechanics/blob/main/literature/books/Latent_Variable_Models_Loehlin_2004.pdf)