Structural Equation Modelling (SEM)

Aims and Objectives

By the end of this seminar you should:

- Have a working knowledge of the principles behind causality.
- Understand the basic steps to building a Model of the phenomenon of interest.
- Be able to construct/Interpret path diagrams.
- Understand the basic principles of how models are tested using SEM.

Historical Background: Cause and Effect

Philosophers have had a great deal to say about the conditions necessary to infer causality. Early theories such as Hume's stressed the importance of observed temporal regularities. Cause and effect should occur close together in time, cause should occur before an effect is observed, and the cause should never occur without the presence of the effect. In short, cause is inferred through corroboration. However, the problem with this is that we know (from statistics) that covariation between variables is insufficient to infer causality (see Field, 2000, Chapter 3) for two reasons: (a) the *Third Variable* problem, and (b) *inferring direction of causality*. As an example, we might well observe a strong correlation between having dreadlocks and supporting environmental issues. However, it is unlikely that having dreadlocks causes an interest in the environment. There is, presumably, an external factor (or factors) that causes both.

John Stuart Mill (1865) described three conditions necessary to infer cause:

- Cause has to precede effect
- Cause and effect must be related
- All other explanations of the cause-effect relationship must be ruled out.

In many senses Mill owes a huge debt to Hume in that his first two conditions mirror those of Hume (correlation between cause and effect and temporal precedence). Mill's main contribution was to add the third condition, which suggests that the effects of a 'third variable' might be ruled-out.

To verify the third criterion, Mill proposed the **method of agreement** which states that an effect is present when the cause is present; **the method of difference** which states that when the cause is absent the effect will be absent also and; the **method of concomitant variation** which states that when the above relationships are observed, causal inference will be made stronger because most other interpretations of the cause-effect relationship will have been ruled out.

Example:

If we wanted to say that me talking about causality causes boredom, we would have to satisfy the following conditions:

(1) I talk about causality before boredom occurs.

(2) Whenever I talk about causality, boredom occurs shortly afterwards.

(3) The correlation between boredom and my talking about causality must be strong (e.g. 4 out of 4 occasions when I talk about causality boredom is observed).

(4) When cause is absent effect is absent: when I don't talk about causality no boredom is observed.

(5) **The manipulation of cause leads to an associated change in effect**. So, if we manipulated whether someone is listening to me talking about causality or to my cat is mewing, the effect elicited should change according to the manipulation.

This final manipulation serves to rule out external variables that might affect the cause-effect relationship.

Historically, Psychology has two traditions: (a) experimental (variables are manipulated systematically to infer causality), and (b) situations are observed and variables correlated (see Field, 2000, Chapter 7; Cronbach, 1957). The first approach allows inferences about causality because possible causes are being compared systematically (i.e. a situation in which cause is observed is compared against a situation in which cause is absent). In the later approach no such manipulation is made, therefore cause cannot be inferred. For example, we can observe that night always follows day, but this doesn't necessarily mean that day causes night (Hume).

So, in situations in which cause cannot be manipulated we cannot make causal attributions about our variables. Statistically speaking, this means that when we analyze data from non-experimental situations we cannot conclude anything about cause an effect. Structural Equation Modeling (SEM) is an attempt to provide a flexible framework within which causal models can be built.

A Simple SEM

SEM is an attempt to model causal relations between variables by including all variables that are known to have some involvement in the process of interest. As a simple example, we could test the effect of a drug on some psychological disorder (e.g. obsessive compulsive disorder, OCD). Obsessive checking behaviour sometimes characterizes this disorder. So, we could take a baseline of checking behaviour, administer the drug and then re-measure checking to ascertain the reduction in this behaviour. We might find a pretty strong relation (correlation) between drug dose and reduction in checking across patients. However, there are several problems. First, this is a non-directional relationship because although drug dose could cause the reduction in checking behaviour, the drug dose could be caused by the behaviour (because higher doses are likely to be given to the more severe cases). Second, drug dose might not have a direct effect on behaviour; instead, the drug might inhibit some cognitive process (such as worrying that a task hasn't been done) that in turn prevents behaviour. So, although the effect might be causal it is part of a chain of events. Finally, the drug might have no effect and both drug dose and behaviour are moderated by a third variable. For example, the trait of 'restraint' might (a) reduce the number of times that a obsessive thoughts lead to an actually behaviour, and (b) affect the amount of symptoms reported to the psychiatrist (and hence affects the amount of drug prescribed.

This example should hopefully illustrate several points about building a causal pathway model about the effect of a drug on behaviour.

1. **Third Variables**: In non-experimental research, all variables that may cause a spurious relationship between the hypothesized 'cause and effect' must be collected and included in the data. In this example, 'restraint' might cause a spurious relationship between the drug and checking behaviour. Therefore, we need to collect data about patient's levels of cognitive restraint and include this variable in the model. This is similar to what we do with partial correlations (see Field, 2000, Chapter3). We collect information about possible 'third variables' and include them in the model.
2. **Disturbance Terms**: There may be variables that have an influence over either the cause of effect variables, but not both. For example, 'worry' (as a trait) may affect checking behaviour

but will not affect the dose of drug. Similarly, the psychiatrist prescribing the dose is likely to directly affect the dose (because doctors will vary in their preferred dose for a given level of symptoms) but not the outcome (the checking behaviour). These variables need not be included in the model because they affect only the cause or effect, but not both. These are known as disturbance terms and are sometimes notated by the letter ζ_i for the i^{th} variable.

Model Specification

The first step in SEM is to specify a model. A model is simply a statement (or set of statements) about the relations between variables. Although you may not realise it you will have already specified many models during your statistics course. For example, a correlation implies a non-directional single relation between two variables, whereas ANOVA and Regression imply directional relations between variables (however, neither technique can statistically test the directionality of these relations). More complex models (in ANOVA for example) involve the use of planned comparisons to specify particular relations between levels of a variable. In all of these techniques, an implicit model underlies the analysis. In SEM, the model is much more central to the process, so rather than relying on a fixed framework based on an implicit model (such as Regression), in SEM we specify a model first, and then use SEM to test the model specified.

Specification involves several processes. First *parameters* need to be specified. Parameters are a set of values (constants) that indicate the nature of the relation between variables. So, these are similar to correlation coefficients, or factor loadings in which the magnitude and direction of the parameter indicate something about the strength and direction of the relationship. There are two types of parameter:

- **Fixed Parameters**: These are parameters not estimated from the data collected. Typically, their value is fixed at zero. These parameters indicate no relation between variables.
- **Free Parameters**: These are parameters that are estimated from the data collected and are believed, by the investigator, to be different from zero. The parameters indicate the presence of a relationship between variables.

Variables: There are two types of variables that are used in SEM. The first type is ***measured variables***, which are variables that can be directly measured in some way. So, these variables consist of data that have actually been collected that map directly to the construct of interest (we could measure heart rate directly though a pulse monitor and the data collected maps directly onto the variable of interest). A second type of variable is a ***latent variable***, which is a variable that cannot be measured directly but is implied by the covariances between two or more indicator variables. A latent variable is, therefore, just another name for a factor (in factor analysis) or component (in principal components). So, it is just a construct that cannot be measured directly (e.g. depression) but can be accessed through measuring other related variables (such as questions on the BDI). Latent variables are free of the uniqueness and measurement error associated with the indicators. Within a model, measured variables can either be included as indicators of a latent variable, or as stand-alone variables.

Model Components: Within a model there are two components: the *measurement model* and the *structural model*. The measurement model is basically the part of the general model in which latent variables are prescribed. Therefore, we determine which measured variables are indicators of a latent variable (or factor). The structural model is the part of the model in which we define the relationship between latent variables and other measured variables that are not indicators of some other latent variable. These two parts of the model are combined to make a whole model that comprehensively describes the relationships between variables that are free of measurement error (i.e. latent variables).

Types of Relations between Variables: So far we have learnt about the different types of variables we can use in SEM, and the basic types of models that we can use. The next important feature is how we specify relations between types of variables. There are basically three types of relations that can be specified, two of which you should have come across during basic statistics courses.

1. Association: This is a nondirectional relation between variables. This is what is measured by the correlation coefficient with which you should not be familiar. The parameters representing these effects are simply the covariances between variables (or the correlation coefficient if variables are standardized).
2. *Direct Effect*: This is a directional relation between two variables. You have come across this sort of relation in ANOVA and regression in which one variable predicts another. Just like multiple regression and factorial ANOVA it is possible to have a dependent variable related to several independent variables, and as in MANOVA and discriminant analysis it is possible to have one independent variable related to several dependent variables. The parameters representing these effects are values of the regression coefficients (i.e. weights in the regression equation (see Field, 2000, Chapter4).
3. *Indirect Effect*: This is an indirect relation between variables and stems from the possibility that a variable can be both an independent and a dependent variable within SEM. In simple terms, a given variable can be the cause of some change in another variable whilst itself is influenced by some different variable. As a simple example, if we know that 'practice makes perfect' (i.e. practise leads to a perfect performance) and we know that the trait 'dedication' (a latent variable) leads to more practise, then we know that dedication must have an indirect effect of perfection. In this example, practise is both an independent variable (because variation in practise leads to variation in perfect performance) and a dependent variable (variation in practise is affected by variation in dedication).

If the direct and indirect effects of an independent variable are summed, we gain the *total effect* of that variable. Variables that receive a directional influence from some other variable in the system are known as ***endogenous variables***. These variables may also exert their own directional influence on other variables in the system. Variables that do not receive any directional influence from another variable are known as ***exogenous variables***. Exogenous variables are usually associated with each other by nondirectional relations (although it is possible that they are not associated with any other variables in the model.

The variance in endogenous variables is not completely explained by other variables in the system. Instead it is assumed that they are also influenced by an error term which consists of both random error and systematic variable that could, for example, represent the effect of some other variable that has not been measured or included in the present model.

Additional Considerations

In SEM we are hoping to model causal relationships between variables. To do this several conditions should be met.

1. The cause and effect variables of interest must be included.
2. All variables that affect both the cause and effect variables must be included. Therefore, any variable that could produce a spurious relation between the cause and effect variable must be considered within the model.
3. The direct effects between variables must be included to indicate the causal relations of interest.

4. Relations between exogenous variables (non-direct effects) must be included.
5. Disturbance terms should be included that represent any variables that need not be explicitly represented in the model.

Schematic Representation of the Model: Path Diagrams

So far we have learnt about the different types of variables that exist with a model and the types of relations that can exist between variables. A popular way to conceptualise a model is using a path diagram, which is a schematic drawing of the system (model) to be estimated. There are a few simple rules that assist in creating these diagrams:

1. Circles or ellipses represent latent variables.
2. Measured variables are represented by squares or rectangles.
3. Directional effects are indicated using a *single-headed* arrow.
4. Non-directional relations are indicated using a *double-headed* arrow.
5. It is useful to represent the variance of a variable using a double-headed arrow that connects a variable to itself.
6. Each arrow represents either a free or a fixed parameter. Fixed parameters should be indicated by their value, free parameters should be indicated using an appropriate letter or asterisks.

So, looking back to our earlier example of modelling the effect of a drug on checking behaviour, we have identified several variables (from the literature) and have some idea of their impact on each other. We have a latent variable 'restraint', which cannot be directly measured and so is implied by measuring several other variables (for example, through a questionnaire). The resulting path diagram might look like this:

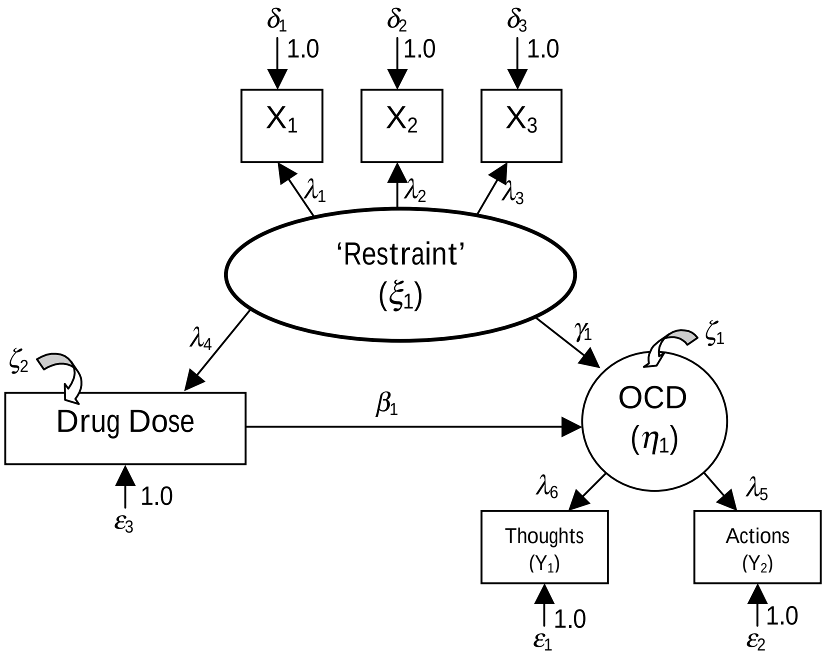

Figure 1.

This diagram looks very complex even though it describes a relatively simply set of relations between variables. Primarily we are interested in the directional relation between drug dose and OCD. Drug dose can be directly measured and so is a measured variable (hence represented by a rectangle). We are interested in the effect that drug dose has on OCD. Now, in this model OCD is a latent variable in that it cannot be measured directly. To keep things simple we assume that we can measure OCD symptoms in terms of the frequency of obsessive thoughts and the frequency of obsession-related actions. As such, OCD is represented by a circle (because it is a latent variable) that is measured by two measured variables (thoughts and actions) both represented by rectangles. The latent variable representing OCD has directional arrows coming into it and so is said to be endogenous (i.e. dependent on other variables within the model). The final variable in the model is the trait of 'restraint', which cannot be measured directly. This latent variable is represented by an ellipse and is measured through three measured variables (X_1, X_2, and X_3). Restraint has a direct effect on both drug dose and OCD symptoms and so is connected to them both with a single arrow. Restraint does not have any directional arrows coming into it and so is said to be exogenous (i.e. doesn't depend on other variables within the model).

All of the measured variables (the squares) have associated error terms, but the latent variables do not (because they are assumed to be free from measurement error). However, both OCD and drug dose have disturbance terms attached to them. In the case of OCD this might be some kind of comorbidity factor such as worry or anxiety (that affect only OCD symptoms but not drug dose). For Drug dose this disturbance term might be the psychiatrist (because different psychiatrists will have different dose preferences for a given level of reported symptoms).

The letters in the path diagram follow usual conventions, which are listed as follows (from Pedhazer & Schmelkin, 1991):

- $X \Rightarrow$ measured variable that indicates an exogenous latent variable.
- $\delta \Rightarrow$ error term associated with X.
- $Y \Rightarrow$ measured variable that indicates an endogenous latent variable.
- $\varepsilon \Rightarrow$ error term associated with Y.
- $\xi \Rightarrow$ Exogenous latent variable.
- $\eta \Rightarrow$ Endogenous latent variable.
- $\lambda \Rightarrow$ Parameter of a directional relation between a latent variable and its indicators.
- $\gamma \Rightarrow$ Parameter of a directional relation between an endogenous and exogenous latent variable.
- $\beta \Rightarrow$ Parameter of a directional relation between two endogenous latent variables.
- $\phi \Rightarrow$ Parameter of a non-directional relation between two latent variables (i.e. the covariance).
- $\zeta \Rightarrow$ disturbance terms (or error in an endogenous latent variable).

Initial Constraints on the Model

Latent variables cannot be directly measured and so do not have a scale of measurement. Therefore, once the model has been created, it is necessary to fix certain parameters such that a scale of measurement is established for all of the latent variables in the model. This objective can be achieved in one of two ways:

1. *Fix the variance of each latent variable*: Fixing the variance of each latent variable has the effect of standardizing these variables. Typically, the variance of each latent variable is set at 1 so that any resulting parameter estimates can be interpreted like standardized regression coefficients (directional relations) or correlations (non-directional relations). If measured variables are also standardized (converted to Z-scores) then relations between latent variables and measured variables can also be interpreted in this way.
2. *Fix the value of one directional relation emitted by a given latent variable*: In effect this means setting the value of the relation for each error term in the model at 1.0. So, we assign a value of 1 to the influence of each error term on its associated endogenous variable. Remember that error terms are typically exogenous and so their variances are estimated from the model. By fixing the value of their influence on an endogenous variable, it is possible to estimate this error variance, which in turn tells us how much variance in the endogenous variable cannot be explained other variables within the model.

You'll notice in the path diagram that the influence of all error terms has been set at 1. Also, although not explicitly drawn on the diagram, the variance of restraint would also be set at 1.

The question remains of how to establish a scale for endogenous latent variables because the variance of these variables is not estimated by the model but implied by other variables within the model. One way is to fix the value of one influence of an endogenous latent variable on another variable at 1. Typically, we fix the value of the influence of the latent variable on one of its indicators. An alternative is to set the implied variance of these variables at 1.0, which has the desirable effect of simplifying the interpretation of the resulting parameter estimates. However, current software typically uses the former approach.

Estimation

Once the model is constructed estimates of the free parameters must be obtained from a set of observed data. The simplest way to get these estimates is to use simple regression techniques to estimate the values of relations between variables in the model. However, computer software packages use iterative procedures in which a series of solutions are generated and then compared to the actual data. Typically, the process involves generating a series of estimates of the free parameters that imply a covariance matrix like the observed one. The implied covariance matrix is the covariance matrix that would result if the values of the fixed parameters and current estimates of the free parameters were substitutive into the structural equations that make up the defined model. So, the program begins with a set of start values that are in some cases estimated from the data (LISREL) or can be a set of default values (EQS), and the implied covariance matrix is estimated and compared to the actual covariance matrix. The differences between these matrices are stored in a residual matrix. This matrix represents the error between the model and the real data. Estimates are changed to try to minimize the values in the residual matrix (so that values in the implied covariance matrix and actual covariance matrix converge). When the residuals cannot be reduced any further the iterations finish.

The residuals are expressed as a single value that represents the degree of correspondence between the implied covariance matrix and the actual covariance matrix. This value is sometimes called the *value of the fitting function* and in a perfect model would be zero (indicating no difference between the implied and actual covariance matrices).

Goodness-of-Fit

As with even the simplest models, it is essential to establish how well the model fits the observed data (see Field, 2000, Chapter 1). The simplest gauge of how well the model fits the data would be to inspect the residual matrix. A good model should have a high number of residuals close to zero (implying high correspondence between elements of the implied covariance matrix and the actual covariance matrix). The fitting function already provides a summary of the information in this matrix and so, in itself, is a measure of how well the model fits the data. Indeed, if the fitting function is multiplied by the sample size minus 1 we get statistic that is distributed with a χ^2 distribution if the data are multivariate normal and the specified model is the correct one (the χ^2 goodness-of-fit test).

This later assumption is unlikely to be true which has led to considerable debate about the appropriateness of this test. Alternatives are the *adjunct fit indexes* pioneered by Bentler and Bonett (1980). These work on the principle of comparing the fitted model against a null model (and so are rather like hypothesis testing). The null model (sometimes called the *independence* model) used is one in which no relations between variable are specified and instead only variable variances are estimated. As such these statistics represent the improvement in fit of the actual model over a model in which all structural parameters are set at zero. Unlike the goodness-of-fit test, these indexes cannot be statistically tested and instead are treated as global indexes of model adequacy. Bentler and Bonett indexes vary between zero and 1 and an acceptable critical value is 0.9. Indexes less than this value reflect inadequate models. However, not all adjunct fit indexes follow these rules.

The goodness-of-fit and adjunct fit indexes are in many ways opposite. The goodness-of-fit gives us an estimate of the error in the model and as such large values represent large amounts of error. Therefore, it is more accurate to call these measures badness-of-fit indexes because large values represent a bad fit. When we put these values to statistical tests we, are, therefore looking fir a null result (i.e. the χ^2 is not significantly different from zero). The adjunct fit index represents the goodness-of-fit more literally in that high values demonstrate a better fit, however, they cannot be put to a statistical test.

What happens if the Model is a Bad Fit?

If the fit indicators suggest that the model is a bad fit it is possible to modify the model by either freeing some of the fixed parameters or by fixing some of the free parameters. This practise is extremely controversial. Although statisticians are not against the notion of model modification per se, there is worry over the way in which it is done. There are several options including the inspection of the residual matrix (for parameters that stand out as not fitting the data), or using statistical searches to find adjustments that will provide a better fit (rather like the DF Beta statistics in multiple regression – see Field, 2000, Section 4.2.4.2.). One example is the *modification index* used by LISREL and the *Lagrange multiplier test* provided by EQS. These statistics provide an estimate of the change in the χ^2 goodness-of-fit statistic resulting from changing a given parameter from being fixed, to being free. However, none of these is a good substitute for substantive theory on which to base modifications, and modifications devoid of theoretical basis are ill advised.

Interpretation of the Model

Assuming the model is a good fit of the data the next step is interpretation. As mentioned earlier, if latent variables are standardizes appropriately then parameters can be interpreted like standardized regression coefficients (see Field, 2000, section 4.4.1.3.) when they are directional, and like correlation coefficients (see Field, 2000, section 3.2.3.1.) when they are non-directional. The ratio of each estimate to it's standard error is also a z-score (so distributed with a z-distribution), therefore values of this ratio greater than 1.96 indicate an estimate that is reliably different from zero.

The bigger issue is that of inferring causality from the structural pathways. A common misconception with SEM is that it provides statistical evidence of a causal link between variables. This is simply not true. The estimates obtained are no different from those obtained from regression, ANOVA or a simple correlation. They tell us nothing about causality. The difference between SEM and these other techniques is in the flexibility with which causal models can be built. In ANOVA models, causality is inferred because one variable is systematically manipulated to see the effect on another. Likewise, in SEM, causality can be inferred, but only from the model originally constructed (and not from the statistical test of that model). The benefit of SEM over other approaches such as ANOVA or regression is simply the flexibility with which models can be built.

Further Reading

Hoyle, R. H. (ed.)(1995). *Structural Equation Modelling: concepts, issues and applications*. Sage: Thousand Oaks, CA.

Pedhazur, E. & Schmelkin, L. (1991). *Measurement, design and analysis*. Hillsdale, NJ: Erlbaum. Chapters 23 & 24.

Saris, W. E. (1999). Structural Equation Modelling. In H. Adèr & G. J. Mellenbergh, (eds.), *Research methodology in the social, behavioural and life sciences* (pp. 220-239). London: Sage.

References

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness-of-fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588-606.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *The American Psychologist*, 12, 671–684.

Field, A. P. (2000). Discovering statistics using SPSS for windows: advanced techniques for the beginner. London: Sage.