Chapter -1

Overview

Structural equation models may be viewed as extensions of multiple regression. They generalize multiple regression in three main ways. First, there is usually more than one equation. Second, a response variable in one equation can be an explanatory variable in another equation. Third, structural equation models can include latent variables.

Multiple equations Structural equation models are usually based upon more than one regression-like equation. Having more than one equation is not really unique; multivariate regression already does that. But structural equation models are more flexible than the usual multivariate linear model.

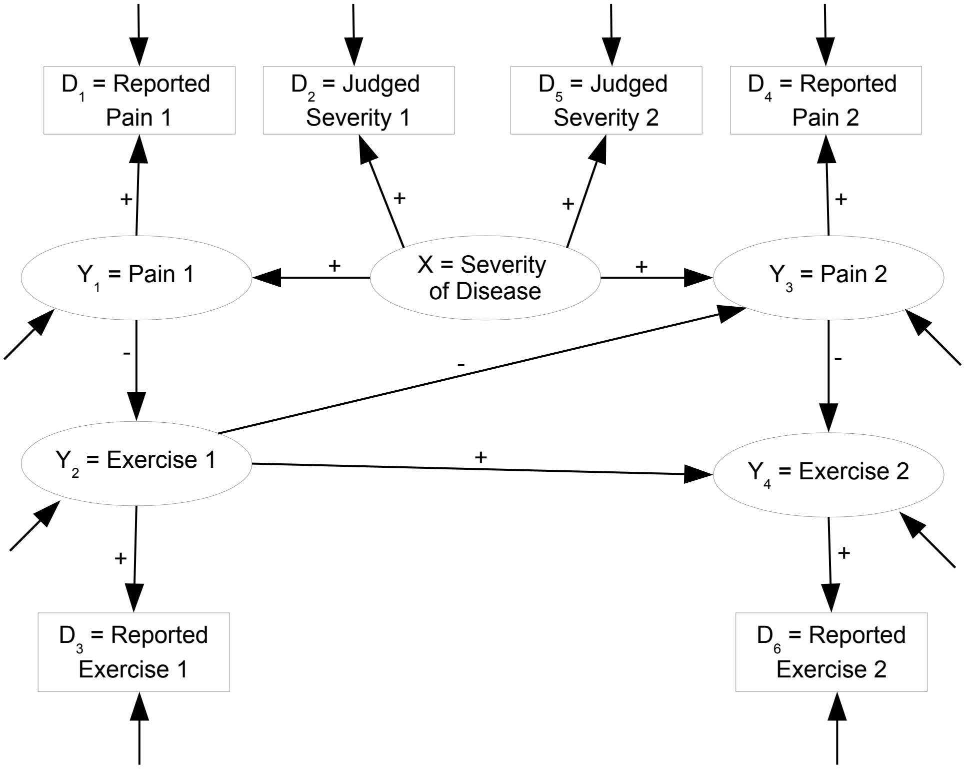
Variables can be both explanatory and response This is an attractive feature. Consider a political science study in which favourable information about a political party contributes to a favourable impression among potential voters at time one. But people often seek out information that supports their viewpoints, so that a favourable impression at time one contributes to exposure to favourable information at time two, which in turn contriutes to a favourable opinion at time two. Thus, opinion at time two is both a response variable and a response variable. Structural equation models are also capable of representing the back-and-forth nature of supply and demand in Economics. There are many other examples.

Latent variables To a degree that is often not acknowledged, the data you can see and record are not what you really are interested in. A *latent variable* is a random variable whose values cannot be directly observed – for example, true family income last year. Contrast this with an *observable variable* – for example, reported family income last year. Usually, interest is in relationships between latent variables, but the data set by definition includes only observable variables. Structural equation models may include latent as well as observable random variables, along with the connections between them. This capability (combined with relative simplicity) is their biggest advantage. It allows the statistican to admit that measurement error exists, and to incorporate it directly into the statistical model.

There are some ways that structural equation models are different from ordinary linear regression. These include random (rather than fixed) explanatory variable values, a bit of specialized vocabulary, and some modest changes in notation. Also, while structural equation models are definitely statistical models, they are also simple *scientific* models of the phenomena being investigated.

This last point is best conveyed by an example. Consider a study of arthritis patients, in which joint pain and exercise are assessed at more than one time point. Figure 1 is a path diagram that represents a structural equation model for the data.

Figure 1: Arthritis Pain



The notation is standard. Latent variables are in ovals, while observable variables are in boxes. Error terms seem to come from nowhere; in many path diagrams they are not shown at all. There is real modeling here, and plenty of theoretical input is required. The plusses and minuses on some of the straight arrows are a bit non-standard. The represent hypothesized positive and negative relationships.

As the directional arrows suggest, structural equation models are usually interpreted as *causal* models. That is, they are models of influence. $A \rightarrow B$ means A has an influence on B. In the path diagram, reported pain at time one is influenced by true pain at time one. There are other influences on reported pain, including the patient's reading level, interpretation of the questions on the questionnaire, self-presentation strategy, and so on. These unmeasured influences are represented by an error term. The error term is not shown explicitly, but the arrow that seems to come from nowhere is coming from the error term.

Structural equation models are causal models [9], but the data are usually observational. That is, explanatory variables are typically not manipulated or randomly assigned by the investigator, as they would be in an experimental study. Instead, they are simply measured or assessed. This brings up the classic *correlation versus causation* issue. The point is often summarized by saying "correlation does not imply causation." That is, if the variables X and Y are related to one another (not independent), it could be that X is influencing Y, or that Y is influencing X, or that a third variable, Z is influencing both X and Y. In the absence of other information, it's wise to be cautious. Practitioners of applied regression are often warned not to claim that the x variables influence the y variable unless the values of the x variables are randomly assigned.

Structural equation modeling adresses the correlation-causation problem by constructing a model that is simultaneously a statistical model and a substantive theory of the data. In this way, a great many details are decided on theoretical or at least common-sense grounds, and the rest are left to statistical estimation and testing. In Figure 1, for example, it is obvious that the arrows should run from Time One to Time Two and not the other way around. Notice that in the path diagram, the severity of the disease is essentially the same at Time One and Time Two. This is a theoretical assertion based on the nature of the disease and the length of time involved. All such assertions are open to healthy debate.

Not everybody likes this. Some statisticians, particularly students, don't feel comfortable with theory construction in a scientific discipline outside their field. This is less a problem than it seems. While it's true that the ideal case is for the same person to be expert in both the statistics and the subject matter (as in econometrics), frequently the statistician works together with a scientist who wants to apply structural equation models to his or her data. Most scientists get the idea of path diagrams very fast, and the collaboration can go smoothly.

It must be admitted, though, that some scientists are uncomfortable with making theoretical commitments and incorporating them into the statistical analysis. To them, data analysis is where evidence is assessed and weighed. Building theory into the statistical model seems biased, like putting a finger on the scale[1]. One response to this is that the generic statistical models in common use also carry assumptions with theoretical implications. Getting involved in the assembly of the statistical model just serves to make the black box less mysterious, and that can only be a good thing.

Path diagrams correspond to systems of regression-like equations. Here are the equa-

[1] There is a distinctly Bayesian feel to the way structural equation models depend on prior information. The objection of bias is also raised against Bayesian methods, for exactly the same reason. It is possible to do structural equation modeling in a fully Bayesian way, but the approach in this book is strictly frequentist.

tions corresponding to Figure 1. Independently for $i = 1, \ldots, n$,

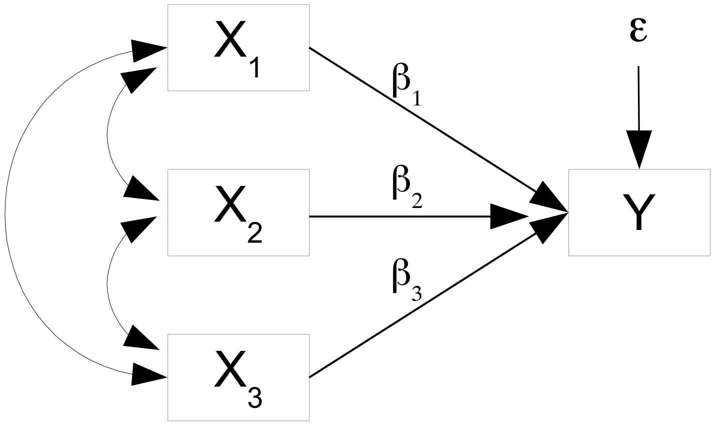
$$\begin{aligned}
Y_{i,1} &= \beta_{0,1} + \beta_1 X_i + \epsilon_{i,1} \\
Y_{i,2} &= \beta_{0,2} + \beta_2 Y_{i,1} + \epsilon_{i,2} \\
Y_{i,3} &= \beta_{0,3} + \beta_3 X_i + \beta_4 Y_{i,2} + \epsilon_{i,3} \\
Y_{i,4} &= \beta_{0,4} + \beta_5 Y_{i,2} + \beta_6 Y_{i,3} + \epsilon_{i,4} \\
D_{i,1} &= \lambda_{0,1} + \lambda_1 Y_{i,1} + e_{i,1} \\
D_{i,2} &= \lambda_{0,2} + \lambda_2 X_i + e_{i,2} \\
D_{i,3} &= \lambda_{0,3} + \lambda_3 Y_{i,2} + e_{i,3} \\
D_{i,4} &= \lambda_{0,4} + \lambda_4 Y_{i,3} + e_{i,4} \\
D_{i,5} &= \lambda_{0,5} + \lambda_2 X_i + e_{i,5} \\
D_{i,6} &= \lambda_{0,6} + \lambda_5 Y_{i,4} + e_{i,6}
\end{aligned} \tag{1}$$

Every variable that appears on the left side of an equation has at least one arrow pointing to it, and the arrows pointing to the left-side variable originate from the variables on the right side.

The path diagram contains some additional information. Note that there are no direct connections between the error terms, or between the error terms and underlying disease severity X_i. This represents an assertion that these quantities are independent. If they were not independent, covariances would be represented by curved, double-headed arrows. An example is given in Figure 2. Notice that all the variables are observable, the error term is shown this time, and the straight arrows from x to y are labelled with the regression coefficients. This is all within the range of standard notation for path diagrams.

Figure 2: Regression with Observable variables

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \epsilon_i$$



Returning to the example of Figure 1, the model as given is still not fully specified. It is common to assume that everything is normal. In most software, the default method of estimation is numerical maximum likelihood based on a multivariate normal distribution for the observable data. There is considerable robustness to this assumption so it does little harm. With the normal assumption and letting the expected values of the error terms equal zero, we have 12 more model parameters, including the expected value and variance of X_i, underlying disease severity. As usual in Statistics, the objective is to estimate and draw inferences about the unknown parameters, with the goal of casting light on the phenomena that gave rise to the data.

Parameter identifiability It is an uncomfortable truth that for the model given here, maximum likelihood estimation will fail. The maximum of the likelihood function would not be unique. Instead, infinitely many sets of parameter values would yield the same maximum. Geometrically, the likelihood function would have a flat surface at the top.

Here's why. Let $\boldsymbol{\theta}$ denote the vetor of parameters we are trying to estimate. $\boldsymbol{\theta}$ contains all the Greek-letter parameters in the model equations (1), plus ten error variances, and also the expected value and variance of X_i. Thus, $\boldsymbol{\theta}$ has 34 elements.

Assume that the model is completely correct, and that disease severity and all the error terms are normally distributed. This means the vector of six observable variables (there are six boxes in the path diagram) have a joint distribution that is multivariate normal — independently for $i = 1, \ldots, n$, of course. All one can ever learn from a data set is the joint distribution of the observable data, and a multivariate normal is completely characterized by its mean vector and variance covariance matrix. Thus, with increasing sample sizes, all you can ever know is a closer and closer approximations of the six expected values (call them $\mu_1, \ldots, \mu_6$) and the 21 unique values of the 6×6 covariance matrix (call them $\sigma_{ij}, i \leq j$). Suppose you knew the μ_j and σ_{ij} values exactly (conceptually letting $n \rightarrow \infty$, if that is an idea that helps). Would this tell you the values of all the model parameters in $\boldsymbol{\theta}$?

The μ_j and σ_{ij} are definitely functions of $\boldsymbol{\theta}$, and those functions may be obtained by direct calculation of the expected values, variances and covariances using the model equations (1). This yields 27 equations. To ask whether the 34 model parameters can be recovered from the μ_j and σ_{ij} is to ask whether it's possible to solve the 27 equations for 34 unknowns. As one might expect, the answer is no. More precisely, it is impossible to solve uniquely. There are infinitely many solutions, so that infinitely many sets of parameter values are equally compatible with any data set. This corresponds to the flat place on the top of the likelihood surface.

In general, model parameters are said to be *identifiable* if their values can be recovered from the probability distribution of the observable data. In structural equation modeling, it is very easy to come up with reasonable models whose parameters are not identifiable — like the arthritis pain and exercise example we are considering. When parameters are not identifiable, estimation and inference can be a challenge, though in some cases the problems can be overcome. In structural equation modeling, almost everything is connected to the the issue of parameter identifiability, and on a technical level, this is what sets structural equation modeling apart from other applied statistical methods based on large-sample maximum likelihood. One of the most important tools in the structural equation modeling toolkit is a set of rules (based on theorems about solving systems of equations) that often allow the identifiability of a model to be determined based on visual inspection of a path diagram, without any calculations. The story begins with an important special case: regression with measurement error.