DCONJOINT

A Mu Sigma University Publication, August 2014

Mu Sigma

In This Issue...

- Philosophy of Statistics
- Unraveling Structural Equation Modeling
- Statistical Nuggets: Rare Events in Logistic Regression
- Python in 15 Minutes

Abstract

Traditional regression modeling techniques explore relationships between a single observed variable and a set of independent observed variables. The drawback of this approach is that we do not get to study the relationship among the independent variables, i.e. how they are interconnected. Path Analysis or SEM allows us to deal with more complex relationships among the observed variables and latent variable. It allows us to have more than one dependent variable and study the interconnected relationships. In this article, we will try to unravel Structural Equation modeling in a very simple and non-mathematical way.

Unraveling Structural Equation Modeling

- By Indrajit Sen Gupta

Introduction

Structural Equation Modeling or SEM in short is a multivariate statistical technique which tries to model a causal process through a series of structural equations. As an analyst, whenever we think of building a statistical model, we tend to think of modeling techniques like Linear or Logistic regressions. Regression models are very useful when we have one outcome variable Y and a set of independent variables X_i's. We try to fit a statistical model between this outcome variable Y and the X_i's. But therein we face a limitation. In real world, seldom we have a set of independent variables in any system that we study. On the right hand side of your regression model, all the independent variables are taken together and regressed on one outcome variable. This prevents us from studying the relationships that may exist among the independent variables.

Enter Structural Equation Modeling which takes a confirmatory approach rather than an exploratory approach to data analysis, i.e., it is hypothses driven. A researcher wanting to use SEM, will need to specify the pattern of intervariable relations beforehand. In contrast, traditional approaches are more descriptive in nature. Moreover in SEM, the models may include variables that are unobserved or latent. For this reason SEM is often referred to as latent variable modeling. Another advantage of SEM over traditional multivariate procedures is that, the latter are incapable of correcting for measurement error, whereas SEM provides explicit estimates of these error variances. Given these characteristics, structural equation modeling has become a very popular technique in non experimental research.

In this article, I will introduce the fundamental concepts behind SEM, the steps that are required to implement these family of methods, one real life example of application of SEM in social sciences and the software that is currently available in the market.

The Basics

One of the most important requirements in application of SEM is your knowledge and familiarty with the business or the area of research. Everything from initial model specification to analysis and interpretation of results will be based on your domain expertise. The final outcome of your SEM will depend on how well informed you are about your research area and correct are your hypotheses.

To understand SEM, we need to distinguish between observed and latent variables. Observed variables is something we are familiar with - the ones that represent the data. They can be categorical or continuous. Latent variables, on the other hand are those which are not directly observed or measured. They are usually inferred

DCONJOINT

A Mu Sigma University Publication, August 2014

Mu Sigma

In This Issue...

- Philosophy of Statistics
- Unraveling Structural Equation Modeling
- Statistical Nuggets: Rare Events in Logistic Regression
- Python in 15 Minutes

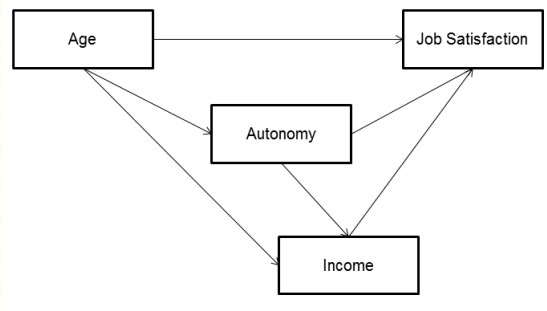
from a set of observed variables. Consider a simple question - "How are you feeling today?". A response to such a question is something that we cannot directly measure. A statement like - "student A is very intelligent" is again something we can only make based on test scores. What this illustrates is that we humans have an innate tendency to categorize our thoughts and beliefs in a way that is not directly observable. Latent variables are thus ones which are somewhat abstract and are usually linked to some observed variables to be measured indirectly.

Both latent and observed variables can be defined as either independent variables or dependent variables. An independent variable is one that is not influenced by any other variable. A dependent variable is a variable that is influenced by other variables in the model. Both of these types of variables can be associated with a residual or error terms.

In any statistical modeling exercise, a researcher usually hypothesizes a statistical relationship based on his or her knowledge and experience. Once the mathematical form is specified, its accuracy or goodness-of-fit is tested against available data. From the results of this test, the researcher either rejects or fails to reject the model. If on the basis of a poor fit, the model gets rejected the researcher will then try to locate the source of misfit, respecify and reestimate the model. This process continues till a model is identified which is meaningful and fits the data well. In the case of SEM, the process is similar, the only difference being, the researcher might have identified alternative models based on theory. If such is the case, then each of the alternative models may be tested till the best one is found.

The Structural Equation

A structural equation model is not one, but an integration of two multivariate techniques - Path Analysis and Factor Analysis. Path models were developed by a biologist named Sewell Wright in the year 1918. These models basically use correlation and regression analysis to model complex relationships among the observed variables. Though developed in the early 20th century, this technique gained prominence only in around 1960's. To illustrate path analysis, let me take a simple example from Bryman and Cramer (1990).



Path Diagram of Causal Relationships in Job Survey

DCONJOINT

A Mu Sigma University Publication, August 2014

Mu Sigma

In This Issue...

- Philosophy of Statistics
- Unraveling Structural Equation Modeling
- Statistical Nuggets: Rare Events in Logistic Regression
- Python in 15 Minutes

In the above diagram, we have a path model given by Bryman and Cramer using four variables from a job survey: age, income, autonomy and job satisfaction. They proposed that age not only had a direct effect on job satisfaction, but also an indirect effect through autonomy and income. The above path diagram gives rise to three equations with standardized beta coefficients as:

$$\text{Job Satisfaction} = \beta_1 \text{Age} + \beta_2 \text{Autonomy} + \beta_3 \text{Income} + \epsilon_1 \quad (1)$$

$$\text{Income} = \beta_4 \text{Age} + \beta_5 \text{Autonomy} + \epsilon_2 \quad (2)$$

$$\text{Autonomy} = \beta_6 \text{Age} + \epsilon_3 \quad (3)$$

In the above three equations, ϵ_i's represent error terms. The beta coefficients can be estimated by solving the above three regression equations. Thus path models are very useful when a researcher has clear understanding of the causal hypotheses.

Factor analysis, which is the second multivariate statistical technique used in SEM, is useful for exploring relationships between an observed set of variables and latent variables. There are two types of factor analysis: exploratory and confirmatory. Exploratory factor analysis is useful for situations where the relations between observed and latent variables are not completely known, whereas confirmatory factor analysis is used when the researcher has prior knowledge regarding the relationships.

To understand how factor analysis works, consider a simple situation where we have two continuous variables X_1 and X_2 observed for n different entities. Now if they have a high degree of correlation, we can imagine that both the variables possess similar kind of information. If we plot these two variables on a 2 - dimensional Euclidean plane, then it will be easy to see how a single regression line can be used to represent the information. Now if we define a single variable say F_1 to approximate the regression line, then this variable will be able to capture the information contained in X_1 and X_2. This concept can be generalized to multiple variables and with the basic idea that if a set of variables are correlated, then these variables could be reduced to a smaller set of variables containing similar information. This statistical technique was first introduced by L. L. Thurstone in 1931 and has been one of the most popular variable reduction techniques for researchers.

SEM thus combines path models and confirmatory factor models to form a structural equation for both latent and observed variables. In the next section we will briefly walk through the basic steps of building a structural equation model.

Steps in Building SEM

There are four basic steps in building a structural equation model:

1. **Model Specification** In this step, the researcher specifies the model by drawing a path diagram. This is the most difficult and important of all the steps since it requires extensive knowledge about the field of study. SEM path diagrams are usually portrayed using the four geometric symbols. For observed variables a rectangle (▭) is used. Ellipses (⬭) are used to represent latent variables. Single headed arrows (⟶) are used to denote the impact of one variable on another and double headed arrows (⟷) are used to represent inter-relationships.

DCONJOINT

A Mu Sigma University Publication, August 2014

Mu Sigma

In This Issue...

- Philosophy of Statistics
- Unraveling Structural Equation Modeling
- Statistical Nuggets: Rare Events in Logistic Regression
- Python in 15 Minutes

2. **Model Identification** This is a crucial step where the researcher tries to check if the specified model can actually be estimated by the SEM software. If the model is unidentifiable(the parameter values are not unique) then it is not a plausible model. The complexity of the model depends on the number of parameters to be estimated. The number of parameters that can be estimated in an SEM is limited by the number of unique elements (variances + covariances) of the covariance matrix. There is an easy way to check this. Suppose a model has p variables in total. Then the total number of unique elements (k) in the covariance matrix is given by the formula $k = p\,(p+1)\,/2$. This is the upper bound to the number of parameters that can be estimated in an SEM.
3. **Data Preparation** This step is identical to the EDA steps required for other modeling techniques. Some fundamental data related checks need to be in place before running a SEM model. They are
   - The data matrix must be non - singular, i.e., it must be invertible
   - Extreme multicollinearity needs to be removed as this can again make the data matrix singular
   - The data needs to be checked for outliers
   - Missing values need to be removed or imputed
   - SEM assumes data follows conditional multivariate normality for the continuous variables, i.e, each variable is normally distributed and any given pair of variables follow a bivariate normal distribution. This is often not practical in real life situations, hence variable transformations may be made to achieve normality.
4. **Model Estimation** Once the data is cleaned and checked as per the previous step, it is fed into a SEM software for estimation. Based on the output, the model is tested for goodness of fit. Most often the initial models fail, hence the researcher might need to respecify the model and rerun the analysis. If the model fit is satisfactory, the next step is to interpret and report the parameter estimates which is again very crucial for the success of the analysis. The path coefficients are interpreted in the same way the regression coefficients are interpreted in a linear regression model.

Methods of Estimation

There are a bunch of different techniques for estimation of parameters. Most SEM software will use the popular **maximum likelihood (ML)** approach. This method expects all the statistical assumptions of normality to be met. The next popular technique is the **two - stage least squares (2SLS)** method. This is similar to the OLS method used in regression. This method has some advantages over the conventional maximum likelihood method. 2SLS does not require any distributional assumptions for the independent variables. They can be non - normal, binary, etc. This method is computationally simple and does not require any optimization algorithm. For small samples 2SLS may perform better than ML approach.

There is a disadvantage in using 2SLS over maximum likelihood though. The ML estimator is more efficient compared to 2SLS and most software which come with 2SLS algorithm do not have a graphical interface for drawing path diagrams.

DCONJOINT

A Mu Sigma University Publication, August 2014

Mu Sigma

In This Issue...

- Philosophy of Statistics
- Unraveling Structural Equation Modeling
- Statistical Nuggets: Rare Events in Logistic Regression
- Python in 15 Minutes

Other methods of estimation are **generalized least squares (GLS)** and **unweighted least squares (ULS)**. The GLS technique is based on the assumption of multivariate normality but requires less computation time and memory when compared to the maximum likelihood approach. The ULS technique is again a type of OLS estimation which though not as efficient as maximum likelihood has one distinct advantage that it does not require the covariance matrix to be positive definite or non - singular.

Software for SEM

This article would not be complete without giving an overveiw of the software available in the market. The software market is flooded with tools specific to structural equation modeling. Commercial software include AMOS from SPSS, CALIS procedure from SAS, SEPATH from Statistica, MPlus, LISREL and EQS. Among the free software we have OpenMx and Onyx. R software also comes with various packages like sem, lavaan, plspm, plsdepot, pathdiagram and OpenMx packages.

Each of the above-mentioned software has its own set of advantages and disadvantages. Some have graphical user interface and some don't. Use the one that you are most comfortable with. At the end of the day the focus should be on the domain knowledge and not the tool since a computer will not be able to validate the correctness of your hypotheses.

References

[1] Randall E. Schumacker & Richard G. Lomax (2010) *A Beginner's Guide to Structural Equation Modeling*, Routledge Taylor & Francis Group, New York.

[2] Rick H. Hoyle (2012) *Handbook of Structural Equation Modeling*, The Guilford Press, New York.

[3] Kenneth A. Bollen (2002) *Latent Variables in Psychology and the Social Sciences*, Annu. Rev. Psychol. 2002. 53:605-34.

[4] Yves Rosseel (2012) *lavaan: an R package for structural equation modeling and more*, Dept. of Data Analysis, Ghent University (Belgium).

[5] Rex B. Kline (2011) *Principles and Practice of Structural Equation Modeling*, The Guilford Press, New York.

[6] Barbara M. Byrne (2010) *Structural Equation Modeling with AMOS: Basic Concepts, Applications, and Programming (2nd ed.)*, Routledge Taylor & Francis Group, New York.

[7] John Fox (2006) *Structural Equation Modeling With the sem Package in R*, Lawrence Erlbaum Associates, Inc.

[8] Diana Suhr (2006) *The Basics of Structural Equation Modeling*, Technical Report, University of Northern Colorado.