Structural Equation Modeling

I. Introduction
   a. Things you should know before attempting SEM
   b. Kinds of research questions addressed by SEM
   c. Limitations and assumptions relating to SEM
   d. Preparing your data for analysis
II. Developing the Model Hypothesis and Model Specification
   a. Drawing your hypothesized model: procedures and notation
   b. Specifying equation sets
   c. Specifying covariances, residuals, and variances
   d. Estimating, fixing, and constraining variables
      i. Estimation methods
      ii. Determining the number of data points for testing
   e. Overidentification
   f. Preparing a program to analyze your model
   g. Other issues
III. Assessing Model Fit
   a. Comparative fit indices
   b. Variance-accounting indices
   c. Parameter-based indices
IV. Model Modification
   a. Modifying a model to improve fit
   b. The Chi-square difference test
   c. The Lagrange-multiplier test
   d. The Wald test
V. Examples

Introduction

A. Things You Should Know Before Using Structural Equation Modeling.

Structural equation modeling (SEM) is a series of statistical methods that allow complex relationships between one or more independent variables and one or more dependent variables. Though there are many ways to describe SEM, it is most commonly thought of as a hybrid between some form of analysis of variance (ANOVA)/regression and some form of factor analysis. In general, it can be remarked that SEM allows one to perform some type of multilevel regression/ANOVA on factors. You should therefore be quite familiar with univariate and multivariate regression/ANOVA as well as the basics of factor analysis to implement SEM for your data.

Some preliminary terminology will also be useful. The following definitions regarding the types of variables that occur in SEM allow for a more clear explanation of the procedure:

a. Variables that are not influenced by another other variables in a model are called **exogenous** variables. As an example, suppose we have two factors that cause changes in GPA, hours studying per week and IQ. Suppose there is no causal relationship between hours studying and IQ. Then both IQ and hours studying would be exogenous variables in the model.

b. Variables that are influenced by other variables in a model are called **endogenous** variables. GPA would be a endogenous variable in the previous example in (a).

c. A variable that is directly observed and measured is called a **manifest** variable (it is also called an indicator variable in some circles). In the example in (a), all variables can be directly observed and thus qualify as manifest variables. There is a special name for a structural equation model which examines only manifest variables, called *path analysis*.

d. A variable that is not directly measured is a **latent** variable. The "factors" in a factor analysis are latent variables. For example, suppose we were additionally interested in the impact of motivation on GPA. Motivation, as it is an internal, non-observable state, is indirectly assessed by a student's response on a questionnaire, and thus it is a latent variable. Latent variables increase the complexity of a structural equation model because one needs to take into account all of the questionnaire items and measured responses that are used to quantify the "factor" or latent variable. In this instance, each item on the questionnaire would be a single variable that would either be significantly or insignificantly involved in a linear combination of variables that influence the variation in the latent factor of motivation

e. For the purposes of SEM, specifically, **moderation** refers to a situation that includes three or more variables, such that the presence of one of those variables changes the relationship between the other two. In other words, moderation exists when the association between two variables *is not the same* at all levels of a third variable. One way to think of moderation is when you observe an interaction between two variables in an ANOVA. For example, stress and psychological adjustment may differ at different levels of social support (i.e., this is the definition of an interaction). In other words, stress may adversely affect adjustment more under conditions of low social support compared to conditions of high social support. This would imply a two-way interaction between stress and psychological support if an ANOVA were to be performed. Figure 1 shows a conceptual diagram of moderation. This diagram shows that there are three direct effects that are hypothesized to cause changes in psychological adjustment – a main effect of stress, a main effect of social support, and an interaction effect of stress and social support.

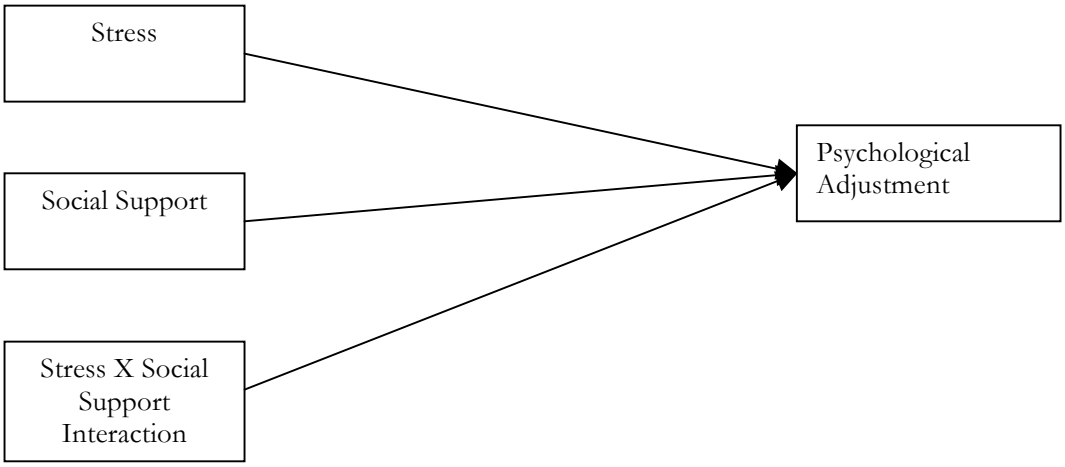


Figure 1. A Diagrammed Example of Moderation. Note that each individual effect of stress, social support, and the interaction of stress and social support can be separated and is said to be related to psychological adjustment. Also note that there are no causal connections between stress and social support.

f. For the purposes of SEM, specifically, **mediation** refers to a situation that includes three or more variables, such that there is a causal process between all three variables. Note that this is distinct from moderation. In the previous example in (e), we can say there are three separate things in the model that cause a change in psychological adjustment: stress, social support, and the combined effect of stress and social support that is not accounted for by each individual variable. Mediation describes a much different relationship that is generally more complex. In a mediation relationship, there is a *direct effect* between an independent variable and a dependent variable. There are also *indirect effects* between an independent variable and a mediator variable, and between a mediator variable and a dependent variable. The example in (e) above can be re-specified into a mediating process, as shown in Figure 2 below. The main difference from the moderation model is that we now allow for causal relationships between stress and social support and social support and psychological adjustment to be expressed. Imagine that social support was not included in the model – we just wanted to see the direct effect of stress and psychological adjustment. We would get a measure of the direct effect by using regression or ANOVA. When we include social support as a mediator, that direct effect will change as a result of decomposing the causal process into indirect effects of stress on social support and social support on psychological adjustment. The degree to which the direct effect changes as a result of including the mediating variable of social support is referred to as the *mediational effect*. Testing for mediation involves running a series of regression analyses for all of the causal pathways and some method of estimating a change in direct effect. This technique is actually involved in structural equation models that include mediator variables and will be discussed in the next section of this document.

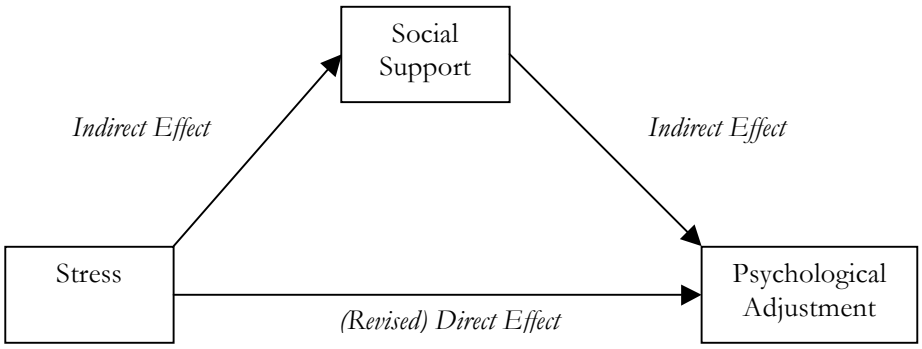


Figure 2. A Diagram of a Mediation Model. Including indirect effects in a mediation model may change the direct effect of a single independent variable on a dependent variable.

In many respects moderation and mediational models are the foundation of structural equation modeling. In fact, they can be considered as simple structural equation models themselves. Therefore, it is very important to understand how to analyze such models to understand more complex structural equation models that include latent variables. Generally, a mediation model like the one above can be implemented by doing a series of separate regressions. As described in latter sections of this document, the effects in a moderational model can be numerically described by using *path coefficients*, which are identical or similar to regression coefficients, depending on the specific choice of analysis you are performing.

g. **Covariance** and **correlation** are the building blocks of how your data will be represented when doing any programming or model specification within a software program that implements structural equation modeling. You should know how to obtain a correlation matrix or covariance matrix using PROC CORR in SAS, or use other menu tools from a statistical package of your choice, to specify that a correlation or covariance matrix be calculated. The covariance matrix in practice serves as your dataset to be analyzed. In the context of SEM, covariances and correlations between variables are essential because they allow you to include a relationship between two variables that is not necessarily causal. In practice, most structural equation models contain both causal and non-causal relationships. Obtaining covariance estimates between variables allows one to better estimate direct and indirect effects with other variables, particularly in complex models with many parameters to be estimated.

h. A **structural model** is a part of the entire structural equation model diagram that you will complete for every model you propose. It is used to relate all of the variables (both latent and manifest) you will need to account for in the model. There are a few important rules to follow when creating a structural model and they will be discussed in the second section of this document.

i. A **measurement model** is a part of the entire structural equation model diagram that you will complete for every model you propose. It is essential if you have latent variables in your model. This part of the diagram which is analogous to factor analysis: You need to include all individual items, variables, or observations that "load" onto the latent variable, their relationships, variances, and errors. There are a few important rules to follow when creating the measurement model and they will be discussed in the second section of this document.

j. Together, the structural model and the measurement model form the entire **structural equation model**. This model includes everything that has been measured, observed, or otherwise manipulated in the set of variables examined.

k. A **recursive** structural equation model is a model in which causation is directed in one single direction. A **nonrecursive** structural equation model has causation which flows in both directions at some parts of the model.

B. Types of Research Questions Answered by SEM.

SEM can conceptually be used to answer any research question involving the indirect or direct observation of one or more independent variables or one or more dependent variables. However, the primary goal of SEM is to determine and validify a proposed causal process and/or model. Therefore, SEM is a confirmatory technique. Like any other test or model, we have a sample and want to say something about the population that comprises the sample. We have a covariance matrix to serve as our dataset, which is based on the sample of collected measurements. The empirical question of SEM is therefore whether the proposed model produces a population covariance matrix that is consistent with the sample covariance matrix. Because one must specify *a priori* a model that will undergo validation testing, there are many questions SEM can answer.

SEM can tell you how if your model is adequate or not. Parameters are estimated and compared with the sample covariance matrix. Goodness of fit statistics can be calculated that will tell you whether your model is appropriate or needs further revision. SEM can also be used to compare multiple theories that are specified *a priori*.

SEM can tell you the amount of variance in the dependent variables (DVs) – both manifest and latent DVs – is accounted for by the IVs. It can also tell you the reliability of each measured variables. And, as previously mentioned, SEM allows you to examine mediation and moderation, which can include indirect effects.

SEM can also tell you about group differences. You can fit separate structural equation models for different groups and compare results. In addition, you can include both random and fixed effects in your models and thus include hierarchical modeling techniques in your analyses.

C. Limitations and Assumptions Regarding SEM.

Because SEM is a confirmatory technique, you must plan accordingly. You must specify a full model *a priori* and test that model based on the sample and variables included in your measurements. You must know the number of parameters you need to estimate – including covariances, path coefficients, and variances. You must know all relationships you want to specify in the model. Then, and only then, can you begin your analyses.

Because SEM has the ability to model complex relationships between multivariate data, sample size is an important (but unfortunately underemphasized) issue. Two popular assumptions are that you need more than 200 observations, or at least 50 more than 8 times the number of variables in the model. A larger sample size is always desired for SEM.

Like other multivariate statistical methodologies, most of the estimation techniques used in SEM require multivariate normality. Your data need to be examined for univariate and multivariate outliers. Transformations on the variables can be made. However, there are some estimation methods that do not require normality.

SEM techniques only look at first-order (linear) relationships between variables. Linear relationships can be explored by creating bivariate scatterplots for all of your variables. Power transformations can be made if a relationship between two variables seems quadratic.

Multicollinearity among the IVs for manifest variables can be an issue. Most programs will inspect the determinant of a section of your covariance matrix, or the whole covariance matrix. A very small determinant may be indicative of extreme multicollinearity.

The residuals of the covariances (not residual scores) need to be small and centered about zero. Some goodness of fit tests (like the Lagrange Multiplier test) remain robust against highly deviated residuals or non-normal residuals.

D. Preparing Your Data for Analysis.

Assuming you have checked for model assumptions, dealt with missing data, and imported your data into a software package, you should obtain a covariance matrix on your dataset. In SAS, you can run PROC CORR and create an output file that has the covariance matrix. Alternatively you can manually type in the covariance matrix or import it from a spreadsheet. You must specify in the data declaration that the set is a covariance matrix, so, in SAS for example, your code would appear as **data name type=COV**; effort must be made to ensure the decimals of the entries are aligned. Other programs like EQS or LISREL can handle full datasets and will automatically compute covariance matrices for you.

Developing the Model Hypothesis and Model Specification

Often, the most difficult part of SEM is correctly *identifying* your model. You must know exactly the number of latent and manifest variables you are including in your model, as well as the number of variances and covariances to be calculated, as well as the number of parameters you are estimating. This section details the rules and conventions that are used when specifying a model. At this time it does not cover a complete set of rules for each popular software programs that are available (EQS, LISREL, AMOS, and PROC CALIS in SAS). In general, though, the conventions that follow are generally compatible with the existing software programs.

A. Drawing your hypothesized model: procedures and notation.

The most important part of SEM analysis is the causal model you are required to draw before attempting an analysis. The following basic, general rules are used when drawing a model:

Rule 1. Latent variables/factors are represented with circles and measured/manifest variables are represented with squares.

Rule 2. Lines with an arrow in one direction show a hypothesized direct relationship between the two variables. It should originate at the causal variable and point to the variable that is caused. Absence of a line indicates there is no causal relationship between the variables.

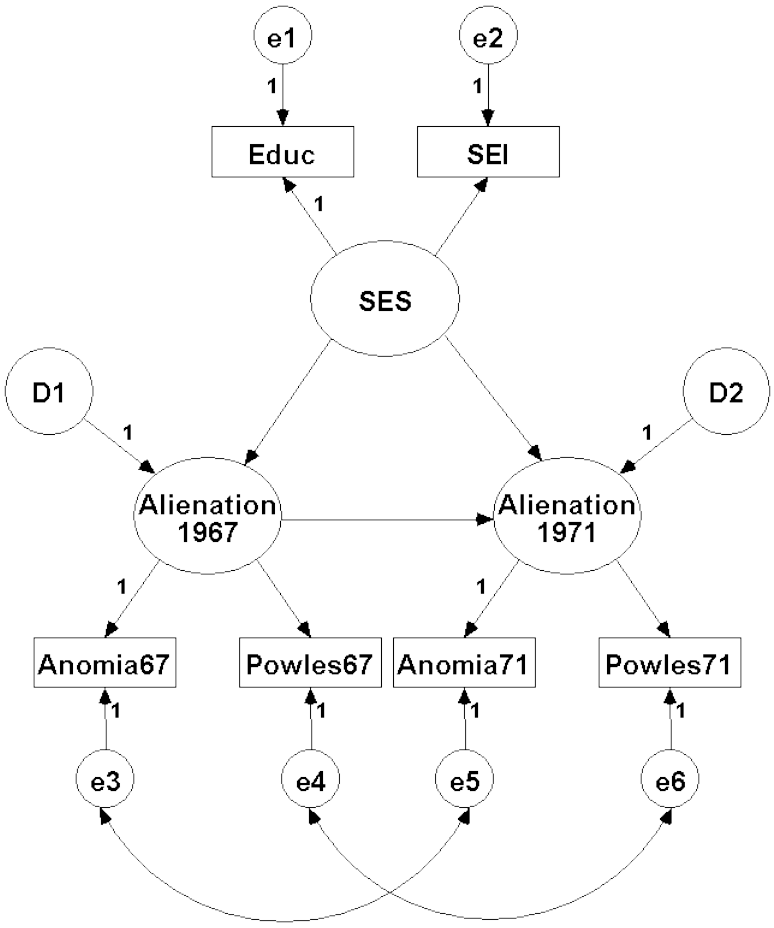
Rule 3. Lines with an arrow in both directions should be curved and this demonstrates a bi-directional relationship (i.e., a covariance).

Rule 3a. Covariance arrows should only be allowed for exogenous variables.

Rule 4. For every endogenous variable, a residual term should be added in the model. Generally, a residual term is a circle with the letter E written in it, which stands for error.

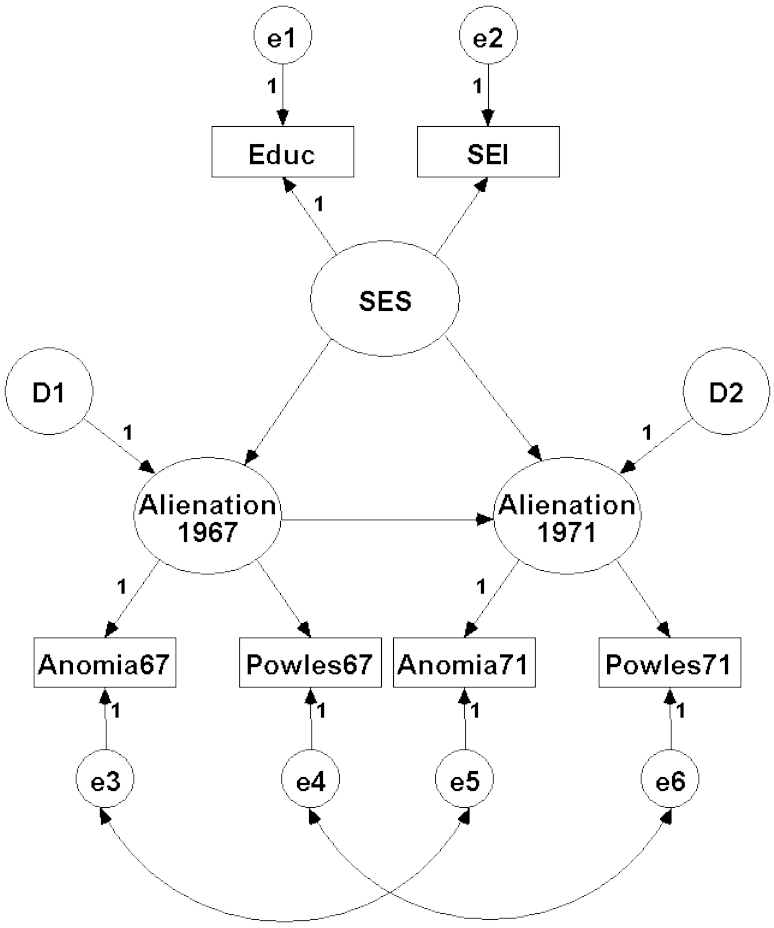
Rule 4a. For latent variables that are also endogenous, a residual term is not called error in the lingo of SEM. It is called a disturbance, and therefore the "error term" here would be a circle with a D written in it, standing for disturbance.

These rules need to be expanded and completed. An example diagram (that is analyzed later on) is included on the next page.



SEM example form AMOS:

The model is built in AMOS and the diagram is shown below, please see the SAS example for the explanation on the variables. The standardized parameter estimates are shown in the graph. The squares represent the observed variables and the circles are for the error terms. Three latent variables are assumed with 3 confirmatory factor analyses used to derive them. Ovals are used to indicate these latent variables. The correlation structure between error terms of the confirmatory factor analysis are suggested by AMOS after the initial model fitting without any correlated error terms. This helps improve the overall model fitting.



The goodness-of-fit test statistics are displayed below. Please note the Chi-square test statistic is not significant at 0.05, which suggest that the model fitting is only acceptable. Root mean square error of approximation (RMSEA) is 0.03202 and since it is less than 0.05, it indicates a good fit. Goodness of Fit Index (GFI) and Adjusted Goodness of Fit Index (AGFI) are larger than 0.9 which again reflect a good fit although GFI and AGFI may not be as informative as Chi-square test statistics and RMSEA.

Result (Default model)

Minimum was achieved
Chi-square = 7.81724
Degrees of freedom = 4
Probability level (p-value) = .09851

RMSEA

Model	RMSEA	LO 90	HI 90	PCLOSE

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	.03202	.00000	.06532	.78202
Independence model	.39072	.37687	.40475	.00001

RMR, GFI

Model	RMR	GFI	AGFI	PGFI
Default model	.08052	.99724	.98553	.18995
Saturated model	.00000	1.00000		
Independence model	3.98590	.48216	.27503	.34440

Regression Weights: (Group number 1 - Default model)

			Estimate	S.E.	C.R.	P	Label
Alienation1967	<---	SES	-.64495	.05350	-12.05418	***	
Alienation1971	<---	SES	-.22497	.05509	-4.08390	***	
Alienation1971	<---	Alienation1967	.58916	.05580	10.55811	***	
Educ	<---	SES	1.00000				
SEI	<---	SES	.58409	.04264	13.69760	***	
Powles67	<---	Alienation1967	1.00000				
Anomia67	<---	Alienation1967	1.12575	.06772	16.62422	***	
Powles71	<---	Alienation1971	1.00000				
Anomia71	<---	Alienation1971	1.13332	.07111	15.93816	***	

Covariances: (Group number 1 - Default model)

			Estimate	S.E.	C.R.	P	Label
e3	<-->	e5	1.61074	.32703	4.92541	***	
e4	<-->	e6	.53090	.24851	2.13634	.03265	

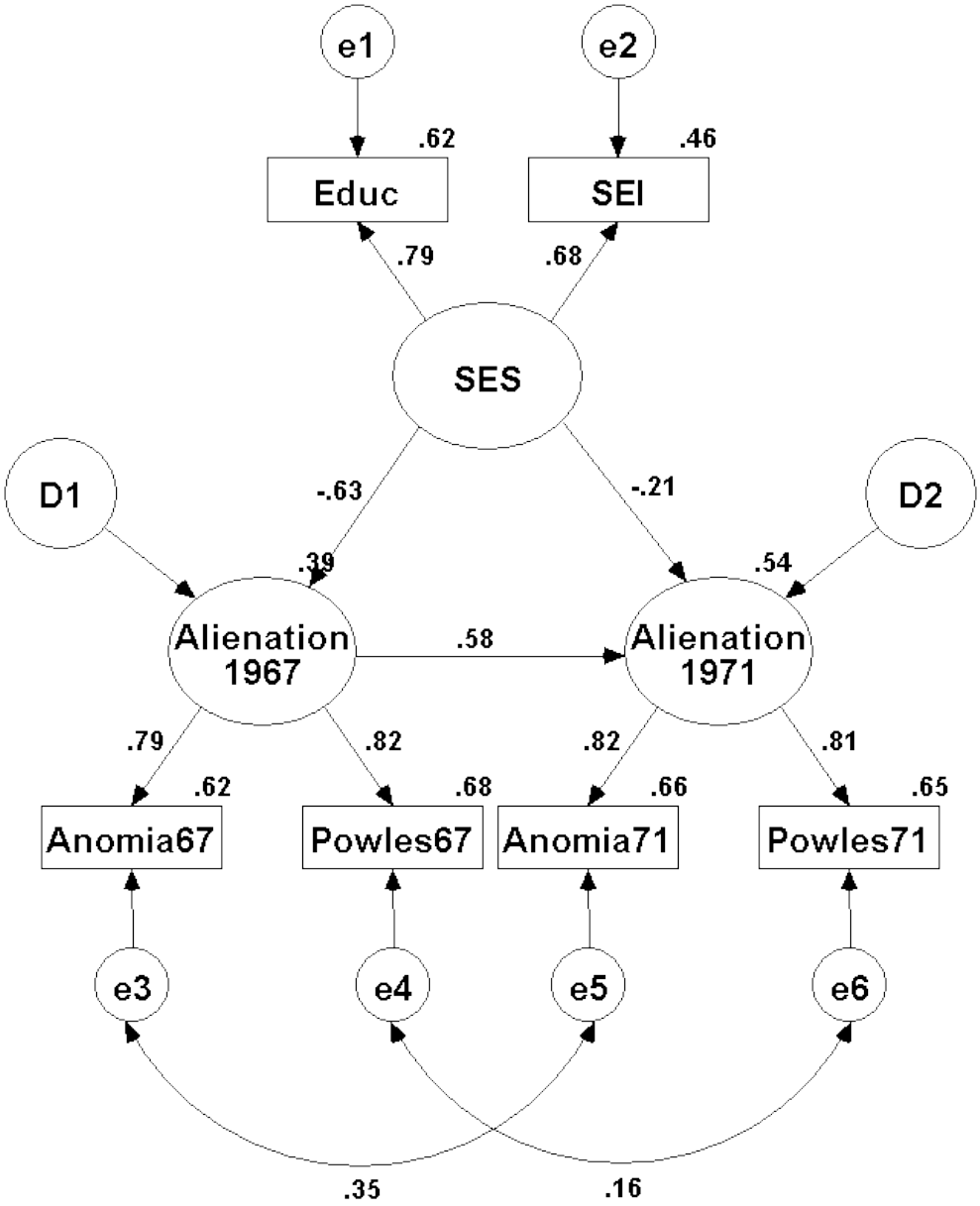
All the parameter estimates are high significant. In other words, all of them are significantly differently from 0. The interpretations on the parameter estimates are straight forward. For example, Alienation in 1967 decreases -.726 for each 1.00 increase in SES. The correlation structure between e3 and e5, e4 and e6 is also estimated by AMOS with significant results.

Standardized Regression Weights: (Group number 1 - Default model)

			Estimate
Alienation1967	<---	SES	-.62795
Alienation1971	<---	SES	-.21489
Alienation1971	<---	Alienation1967	.57797
Educ	<---	SES	.78908

			Estimate
SEI	<---	SES	.67865
Powles67	<---	Alienation1967	.82379
Anomia67	<---	Alienation1967	.78535
Powles71	<---	Alienation1971	.80590
Anomia71	<---	Alienation1971	.81502

The standardized the regression estimates are comparable, which may assist us to pick up more important factors and relationships.



3. SEM Examples in SAS

Wheaton, Muthen, Alwin, and Summers (1977) has served to illustrate the performed of several implementations for the analysis of structural equation models. Here two different models will be analyzed by SAS. The database has three continuous predictor variables: education level, a socioeconomic indicator, and feelings of powerlessness measured in 1967. There is one continuous dependent variable, feelings of powerless measured in 1971. The data used here is a simulated version distributed by http://www.utexas.edu/its/rc/tutorials/stat/amos/.

(1) A Path Analysis Example

The data were reanalyzed with PROC CALIS. Input data was data itself. A correlation or covariance matrix can be an input data. The path diagram, including unstandardized regression coefficients, appears the below.

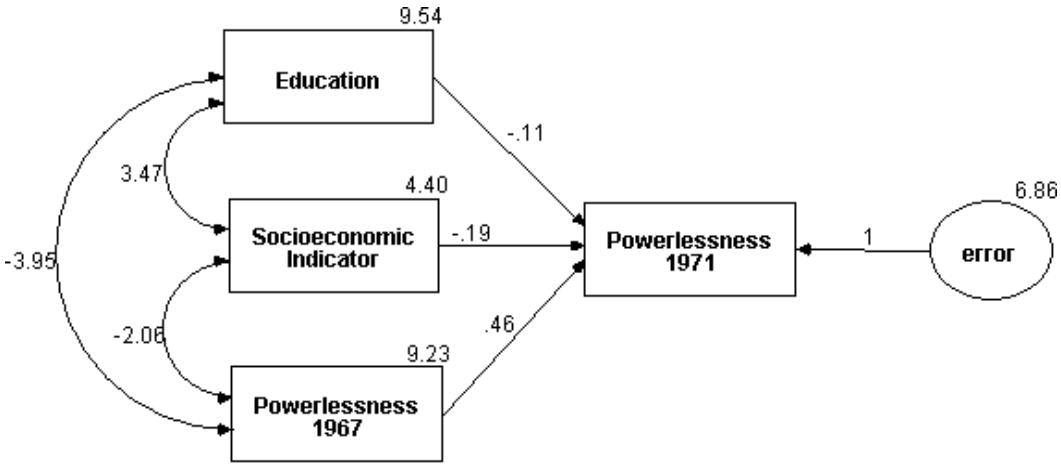


Figure 1. Path Analysis – Wheaton Data

i) Path Analysis Model

```
PROC CALIS data=Wheaton1 ucov aug;
   Lineqs
       Powles71 = b0 Intercept + b1 Educ + b2 SEI + b3 Powles67+ error;
   Std error = ve;
run;
```

PROC CALIS neglects the intercept parameter by default. To estimate the intercept, change the COV option to UNCOV, which analyzes uncorrelated covariance matrix, and use the AUG option, which adds a row and column for the intercept, called INTERCEP, to the matrix being analyzed. The COV option in the PROC CALIS statement requests the analysis of the covariance matrix. Without the COV option, the correlation matrix would be computed and analyzed. You can give an option by METHOD = option, maximum likelihood is the default.

ii) Results

The CALIS Procedure

Fit Function	0.0000
Goodness of Fit Index (GFI)	1.0000
Chi-Square	0.0000
Chi-Square DF	0
Pr > Chi-Square	<.0001
RMSEA Estimate	0.0000
Bentler's Comparative Fit Index	1.0000

Manifest Variable Equations with Estimates

```
Powles71   =    0.4596*Powles67   + -0.1125*Educ        + -0.1948*SEI          +
6.1129*Intercept
Std Err       0.0314 b3           0.0346 b1           0.0489 b2           1.0144
b0
t Value      14.6338             -3.2476             -3.9864              6.0263


+  1.0000 error
```

Variances of Exogenous Variables

Variable	Parameter	Estimate	Standard Error	t Value
Powles67		18.77442		
Educ		18.26784		
SEI		452.46517		
Intercept		1.00107		
error	ve	6.86743	0.31830	21.58

The path analysis showed some interesting results in terms of model fit: Chi-square = 0.000, df = 0, $p < 0.0001$ and CFI (Bentler's Comparative Fit Index) = 1.000 and RMSEA = 0.000.

Table 1. Unstandardized Parameter Estimates – Path Analysis

Parameter	Variable	Estimate	Standard Error	t-value
Coefficient	Education → Powles71	-0.1125	0.0346	-3.2476
	SEI → Powles71	-0.1948	0.0489	-3.9864
	Powles67 → Powles71	0.4596	0.0314	14.6338
Variance				
ve	Powles71 – unexplained variance	6.86743	0.31830	21.58

If input data is in the form of a correlation matrix, standardized and unstandardized parameter estimates will be equal. Note: The PROC REG and PROC CORR can be used to fit this simple model.

(2) SAS Code for Structural Equation Model with Latent variables

The data were reanalyzed with PROC CALIS. Input data was data itself. A correlation or covariance matrix can be an input data. Suppose you want to fit the following diagram in order to test a model of the stability of alienation over time, as measured by anomia and powerlessness feelings at two measurement occasions, 1967 and 1971, as well as education level and a socioeconomic index.

Figure 2. Structural Equation Model with Latent Variables – Wheaton Data

i) Path Analysis Model

Here F1 is 'Alienation 1967,' F2 is 'Alienation 1971,' and F3 is 'SES.' The names of the latent variables must start with the prefix letter F (for Factor); the names of the residuals must start with the prefix letters E (for Error) or D (for Disturbance).

```
PROC CALIS cov data=Wheaton1 method=ml;
   Lineqs
      Educ     =    F3 + e1,
      SEI      = a2 F3 + e2,
      Anomia67 =    F1 + e3,
      Powles67 = a4 F1 + e4,
      Anomia71 =    F2 + e5,
      Powles71 = a6 F2 + e6,
      F1       = c1 F3 + d1,
      F2       = c2 F1 + c3 F3 + d2;
   Std
      e1-e6 = The1-The6,
      d1-d2 = Psi1-Psi2,
      F3 = Ef3;
   Cov
      e3 e5 = The7,
      e4 e6 = The8;
   Var Educ SEI Anomia67 Powles67 Anomia71 Powles71;
RUN;
```

ii) Results

The model provided acceptable fit for CFI (0.9982 > 0.9) and RMSEA (0.0320 < 0.06). All parameter estimates were significant at the 0.01 level (z > 2.56). The p-value of Chi-square value (0.0985 > 0.05) is okay.

The CALIS Procedure

Fit Function	0.0084
Goodness of Fit Index (GFI)	0.9972
GFI Adjusted for Degrees of Freedom (AGFI)	0.9855

```
Chi-Square                                                7.8172
Chi-Square DF                                                4
Pr > Chi-Square                                         0.0985
RMSEA Estimate                                          0.0320
Bentler's Comparative Fit Index                         0.9982
Bentler & Bonett's (1980) Non-normed Index              0.9933
Bentler & Bonett's (1980) NFI                           0.9964
James, Mulaik, & Brett (1982) Parsimonious NFI          0.2657
Z-Test of Wilson & Hilferty (1931)                      1.2974
Bollen (1986) Normed Index Rho1                         0.9863
Bollen (1988) Non-normed Index Delta2                   0.9982

Manifest Variable Equations with Standardized Estimates
            Educ     =   0.7891 F3       +  0.6143 e1
            SEI      =   0.6786*F3       +  0.7345 e2
                             a2
            Anomia67 =   0.7853 F1       +  0.6191 e3
            Powles67 =   0.8238*F1       +  0.5669 e4
                             a4
            Anomia71 =   0.8150 F2       +  0.5794 e5
            Powles71 =   0.8059*F2       +  0.5921 e6
                             a6

 Latent Variable Equations with Standardized Estimates
  F1       =  -0.6280*F3       +  0.7783 d1
                    c1
  F2       =   0.5780*F1       + -0.2149*F3       +  0.6810 d2
                    c2                  c3

                  Variances of Exogenous Variables
                                                 Standard
Variable  Parameter      Estimate         Error    t Value
Powles67                 18.77442
Educ                     18.26784
SEI                     452.46517
Intercept                 1.00107
error     ve              6.86743       0.31830      21.58

        Correlations Among Exogenous Variables

              Var1     Var2     Parameter      Estimate
              e3       e5       The7            0.35284
              e4       e6       The8            0.16447
```

Table 2. Unstandardized Parameter Estimates – Structural Equation Model with Latent variables

Parameter	Variable	Estimate	Standard Error	t-value
Coefficient	SES → Education	1.0000		
	SES → SEI	0.5841	0.0426	13.6975
	Alienation67 → Anomia67	1.0000		
	Alienation67 → Powles67	0.8883	0.0534	16.6242
	Alienation71 →	1.0000		

	Anomia71			
	Alienation71 → Powles71	0.8824	0.0554	15.9381
Variance				
	E1	3.60241	0.41916	8.59
	E2	2.37520	0.17124	13.87
	E3	4.93748	0.47720	10.35
	E4	2.96934	0.36174	8.21
	E5	4.22987	0.50823	8.32
	E6	3.51642	0.40634	8.65
	D1	4.81299	0.49020	9.82
	D2	3.88116	0.39499	9.83
	F3	5.94428	0.56189	10.58
Covariance				
	E3 - E5	1.61249	0.32738	4.93
	E4 – E6	0.53145	0.24878	2.14

Reference

SAS/STAT User's Guide Version 8