Step your way through Path Analysis
Diana Suhr, Ph.D.
University of Northern Colorado

Abstract
This presentation provides a plan to step from regression to a path analysis. Regression analysis sometimes provides less than optimal results using a default model. Path analysis allows you to specify a model and relationships between variables. PROC CALIS provides a method to specify a path analysis model, examine relationships between variables, and provides a comprehensive analysis of your data.

Introduction
Do you want more information from your data? Would you like more than regression analysis? Path analysis allows specification of relationships between variables. This paper will review the basics of path analysis with structural equation modeling methodology and then present an example comparing results from regression analysis and path analysis

Purpose
The purpose of this presentation is to offer an alternative approach to regression analysis and a step-by-step approach to path analysis with structural equation modeling methodology (SEM). The goals are to present a powerful, flexible and comprehensive technique for investigating relationships between measured variables and to challenge you to design and plan research where SEM is an appropriate analysis tool.

Rationale
Analyzing research data and interpreting results could be complex and confusing. Traditional statistical approaches to data analysis specify default models, assume measurement occurs without error, and are somewhat inflexible. However, structural equation modeling requires specification of a model based on theory and research, is a multivariate technique incorporating measured variables and latent constructs, and explicitly specifies measurement error. A model (diagram) allows for specification of relationships between variables.

Structural equation modeling (SEM)
- is a comprehensive statistical approach to testing hypotheses about relations among observed and latent variables (Hoyle, 1995).
- is a methodology for representing, estimating, and testing a theoretical network of (mostly) linear relations between variables (Rigdon, 1998).
- tests hypothesized patterns of directional and nondirectional relationships among a set of observed (measured) and unobserved (latent) variables (MacCallum & Austin, 2000).

Two goals in SEM are
1) to understand the patterns of correlation/covariance among a set of variables and
2) to explain as much of their variance as possible with the model specified (Kline, 1998).

The purpose of the model, in the most common form of SEM, is to account for variation and covariation of the measured variables (MV). Path analysis (e.g., regression) tests models and relationships among MV. Confirmatory factor analysis tests models of relationships between latent variables (LVs or common factors) and MVs which are indicators of common factors.

Similarities between Regression and Path Analysis
Path Analysis with SEM is similar to traditional methods like correlation and regression in many ways. First, both regression and path analysis are based on linear statistical models. Second, statistical tests associated with both methods are valid if certain assumptions are met. Regression methods assume a normal distribution and Path Analysis assumes multivariate normality. Third, neither approach offers a test of causality.

Differences Between Regression and Path Analysis
Regression analysis differs from Path Analysis in several areas. First, Path Analysis is a highly flexible and comprehensive methodology. This methodology is appropriate for investigating achievement, economic trends, health issues, family and peer dynamics, self-concept, exercise, self-efficacy, depression, psychotherapy, and other phenomenon.
Second, regression methods specify a default model whereas Path Analysis with SEM requires formal specification of a model to be estimated and tested. SEM offers no default model and places few limitations on what types of relations can be specified. SEM model specification requires researchers to support hypothesis with theory or research and specify relations a priori.

Third, Path Analysis is a multivariate technique specifying relationships between observed (measured) variables. Multiple, related equations are solved simultaneously to determine parameter estimates. Variables in Path Analysis could be independent and dependent whereas variables in Regression Analysis are either independent or dependent.

Fourth, Path Analysis allows researchers to recognize the imperfect nature of their measures. SEM explicitly specifies error or unexplained variance while Regression Analysis assumes measurement occurs without error.

Fifth, traditional analysis provides straightforward significance tests to determine group differences, relationships between variables, or the amount of variance explained. Path Analysis provides no straightforward tests to determine model fit. Instead, the best strategy for evaluating model fit is to examine multiple tests (e.g., chi-square, Comparative Fit Index (CFI), Bentler-Bonett Nonnormed Fit Index (NNFI), Root Mean Squared Error of Approximation (RMSEA)).

Finally, a graphical language provides a convenient and powerful way to present complex relationships in Path Analysis. Model specification involves formulating statements about a set of variables. A diagram, a pictorial representation of a model, is transformed into a set of equations. The set of equations are solved simultaneously to test model fit and estimate parameters.

Statistics

Traditional statistical methods normally utilize one statistical test to determine the significance of the analysis, R^2 for Regression analysis. Structural Equation modeling, however, relies on several statistical tests to determine the adequacy of model fit to the data. The chi-square test indicates the amount of difference between expected and observed covariance matrices. A chi-square value close to zero indicates little difference between the expected and observed covariance matrices. In addition, the probability level must be greater than 0.05 when chi-square is close to zero.

The Comparative Fit Index (CFI) is equal to the discrepancy function adjusted for sample size. CFI ranges from 0 to 1 with a larger value indicating better model fit. Acceptable model fit is indicated by a CFI value of 0.90 or greater (Hu & Bentler, 1999).

Root Mean Square Error of Approximation (RMSEA) is related to residual in the model. RMSEA values range from 0 to 1 with a smaller RMSEA value indicating better model fit. Acceptable model fit is indicated by an RMSEA value of 0.06 or less (Hu & Bentler, 1999).

If model fit is acceptable, the parameter estimates are examined. The ratio of each parameter estimate to its standard error is distributed as a z statistic and is significant at the 0.05 level if its value exceeds 1.96 and at the 0.01 level if its value exceeds 2.56 (Hoyle, 1995). Unstandardized parameter estimates retain scaling information of variables and can only be interpreted with reference to the scales of the variables. Standardized parameter estimates are transformations of unstandardized estimates that remove scaling and can be used for informal comparisons of parameters throughout the model. Standardized estimates correspond to effect-size estimates.

If unacceptable model fit is found, the model could be revised when the modifications are meaningful. Model modification involves adjusting a specified and estimated model by either freeing parameters that were fixed or fixing parameters that were free. The Lagrange multiplier test provides information about the amount of chi-square change that results if fixed parameters are freed. The Wald test provides information about the change in chi-square that results if free parameters are fixed (Hoyle, 1995).

Considerations

The use of SEM could be impacted by

- the research hypothesis being testing
- the requirement of sufficient sample size
  A desirable goal is to have a 20 to 1 ratio for the number of subjects to the number of model parameters. However, a 10 to 1 ratio may be a realistic target. If the ratio is less than 5 to 1, the estimates may be unstable.
- measurement instruments
- multivariate normality
- parameter identification
• outliers
• missing data
• interpretation of model fit indices (Schumacker & Lomax, 1996).

**SEM Process**

A *suggested approach* to SEM analysis proceeds through the following process:

• review the relevant theory and research literature to support model specification
• specify a model (e.g., diagram, equations)
• determine model identification (e.g., if unique values can be found for parameter estimation; the number of
degrees of freedom, df, for model testing is positive)
• select measures for the variables represented in the model
• collect data
• conduct preliminary descriptive statistical analysis (e.g., scaling, missing data, collinearity issues, outlier
detection)
• estimate parameters in the model
• assess model fit
• respecify the model if meaningful
• interpret and present results.

**Definitions**

A *measured variable* (MV) is a variable that is directly measured whereas a latent variable (LV) is a construct that is
not directly or exactly measured.

**Relationships** between variables are of three types
• Association, e.g., correlation, covariance
• Direct effect is a directional relation between two variables, e.g., independent and dependent variables
• Indirect effect is the effect of an independent variable on a dependent variable through one or more
intervening or mediating variables

A *model* is a statistical statement about the relations among variables.

A *path diagram* is a pictorial representation of a model.

**Specification** is formulating a statement about a set of parameters and stating a model.

**Parameters** are specified as fixed or free.

**Fixed parameters** are not estimated from the data and their value is typically fixed to zero or one.

**Free parameters** are estimated from the data.

**Fit indices** indicate the degree to which a pattern of fixed and free parameters specified in the model are consistent
with the pattern of variances and covariances from a set of observed data. Examples of fit indices are chi-square,
CFI, NNFI, RMSEA.

The purpose of *estimation* is to obtain numerical values for the unknown (free) parameters.

The *criterion selected for parameter estimation* is known as the *discrepancy function*. It provides a guideline to minimize the differences between the population covariance matrix, $\Sigma$, as estimated by the sample covariance, $S$, and the covariance matrix derived from the hypothesized model, $\Sigma(0)$. For example, the discrepancy function for the ML method is $F_{ml} = \log |\Sigma(0)| + \text{Trace} [\Sigma(0)^{-1} S] - \log |S| - p$

**Evaluation of model fit**

• Chi-square is a “badness-of-fit” index, smaller values indicate better fit
• Other fit indices, e.g, CFI, NNFI, are “goodness-of-fit” indices where larger values mean better fit
• The Wald test provides information about the change in chi-square and determines the degree to which
model fit would deteriorate if free parameters were fixed.
• LaGrange Multiplier Test (LM) provides information about the amount of chi-square change and determines the degree to which model fit would improve if any selected subset of fixed parameters were converted into free parameters.

Model modification involves adjusting a specified and estimated model by either freeing parameters that were fixed or fixing parameters that were free. In SEM, model comparison is analogous to planned comparisons in ANOVA, and model modification is analogous to post-hoc comparisons in ANOVA. Model modification could sacrifice control over Type I error and lead to a situation where sample specific characteristics are generalized to a population

If a model is determined to have acceptable fit, then the focus moves to specific elements of fit.
• The ratio of each parameter estimate to its standard error is distributed as a z statistic and is significant at the 0.05 level if its value exceeds 1.96 and at the 0.01 level it its value exceeds 2.56 (Hoyle, 1995).
• Unstandardized parameter estimates retain scaling information of variables involved and can only be interpreted with reference to the scales of the variables
• Standardized parameter estimates are transformations of unstandardized estimates that remove scaling information and can be used for informal comparisons of parameters throughout the model. Standardized estimates correspond to effect-size estimates.

What indicates a “large” direct effect? A “small” one?
• Results of significance tests reflect not only the absolute magnitudes of path coefficients but also factors such as the sample size and intercorrelations among the variables
• Standardized path coefficients with absolute values less than 0.10 may indicate a “small” effect
• Values around 0.30, a “medium” effect
• Values greater than 0.50, a “large” effect

Note: SEM does nothing more than test the relations among variables as they were assessed. Researchers are often too quick to infer causality from statistically significant relations in SEM

Diagram Symbols

- Measured variable (V1), observed variable
- Direct relationship
- Covariance or correlation
- Error (e1) associated with measured variable (V1)
Example 1 – Regression/Path Analysis

Figure 1. Regression Model (math achievement at age 10, reading comprehension achievement at age 12, and mother’s educational level predicting math achievement at age 12).

Figure 2. Revised model (math achievement at age 10, reading comprehension at age 12 predict math achievement at age 12; indirect effect of mother’s educational level and math achievement at age 10).

Example 2 - Regression/Path Analysis with PROC REG and PROC CALIS

SEM techniques are easily applied to analyses in the health field. Application of SEM techniques have contributed to research on illness (Roth, Wiebe, Fillingim, & Shay, 1989), on exercise (Duncan, T. & McAuley, 1993; Duncan, T., Oman, & Duncan, S., 1994) and on substance use/abuse among adults (Curran, Harford, & Muthen, 1996) and adolescents (Curran, Stice, & Chassin, 1997; Duncan, T., Duncan, S., Alpert, Hops, Stoolmiller, & Muthen, 1997).

This path analysis example reanalyzes data from a study where researchers investigated the effects of hardiness, stress, fitness, and exercise on health problems (Roth, et al., 1989). College students (n=373) reported recent physical illness, recent stressful life events, current exercise participation levels, current perceived fitness levels, and hardiness components. Multiple regression and Path Analyses examined the effects related to illness. Subjects were 163 men and 210 women enrolled in an introductory psychology course at a southern United States university. The mean age of the subjects was 21.7 (sd = 5.5).

Assessments

Illness. Seriousness of Illness Rating Scale (Wyler, Masuda, & Holmes, 1968) is a self-report checklist of commonly recognized physical symptoms and diseases and provides a measure of current and recent physical health problems. Each item is associated with a severity level. A total illness score is obtained by adding the severity ratings of endorsed items (symptoms experienced within the last month).
Stress. Life Experience Survey (Sarason, Johnson, & Segal, 1978) is a measure used to access the occurrence and impact of stressful life experiences. Subjects indicate which events have occurred within the last month and rate the degree of impact on a 7-point scale (-3 = extremely negative impact, 0 = no impact, 3 = extremely positive impact). In the study, the total negative event score was used as an index of negative life stress (the absolute value of the sum of negative items).

Fitness. Fitness Questionnaire (Roth & Fillingim, 1988) is a measure of self-perceived physical fitness. Respondents rate themselves on 12 items related to fitness and exercise capacity. The items are on an 11-point scale of 0 = very poor fitness to 5 = average fitness to 10 = excellent fitness. A total fitness score is calculated using responses to 15 common exercise activities and providing blank spaces to write in additional activities.

Exercise. Exercise Participation Questionnaire (Roth & Fillingim, 1988) assessed current exercise activities, frequency, duration, and intensity. An aerobic exercise participation score was calculated using responses to 15 common exercise activities and providing blank spaces to write in additional activities.

Hardiness. In the study, hardiness included components of commitment, challenge, and control. A composite hardness score was obtained by summing Z scores from scales on each component. The challenge component included one scale whereas the other components included 2 scales. Therefore, the challenge Z score was doubled when calculating the hardness composite score. Commitment was assessed with the Alienation From Self and Alienation From Work scales of the Alienation Test (Maddi, Kobasa, & Hoover, 1979). Challenge was measured with the Security Scale of the California Life Goals Evaluation Schedule (Hahn, 1966). Control was assessed with the External Locus of Control Scale (Rotter, Seaman, & Liverant, 1962) and the Powerlessness Scale of the Alienation Test (Maddi, Kobasa, & Hoover, 1979).

Results
Tests were conducted to determine if variables as a whole predicted a significant proportion of the variance of the illness measure and whether each individual variance uniquely accounted for a significant proportion of that variance. A main effects regression model including stress, fitness, hardiness, exercise, and gender to predict illness accounted for approximately 20% of the variance. The SEM analysis excluding gender found hardiness mediated by stress and exercise mediated by fitness (Roth, et al., 1989).

SAS Code
```sas
   data illfl (type=corr);
      input _type_ $1-4     _name_ $6-13 exercise 15-20  hardy   22-27 fitness 29-34 stress 36-41 illness 43-48;
      cards;
      n                373    373    373    373    373
      mean           40.90   0.00  67.10   4.80  716.7
      std            66.50   3.80  18.40   6.70  624.8
      corr exercise   1.00  -0.03   0.39  -0.05  -0.08
      corr hardy     –0.03   1.00   0.07  -0.23  -0.16
      corr fitness    0.39   0.07   1.00  -0.13  -0.29
      corr stress     –0.05  –0.23  –0.13   1.00   0.34
      corr illness   –0.08  –0.16  –0.29   0.34   1.00
   ;;;
```

Figure 4. Regression/Path Analysis Model
**STEP 1 - Regression Analysis with PROC REG**

```plaintext
proc reg data=illfl;
model illness = exercise hardy fitness stress /selection = backward;
```

Note: PROC GLM procedure will not accept a correlation or covariance matrix as input. Therefore, the PROC REG procedure was run.

**STEP 2 – Regression/Path Analysis with PROC CALIS**

A regression analysis was run with PROC CALIS (path analysis). Input data was in the form of a correlation matrix, means, and standard deviations. A correlation matrix standardizes values and loses the metric of the scales. A covariance matrix preserved the metric of each scale (see Figure 4).

```plaintext
proc calis data=illfl corr stderr;
lineqs
  illness = p5p3 fitness + p5p4 stress + p5p1 exercise + p5p2 hardy + e5;
  std
    exercise = varex,
    hardy = varhr,
    fitness = varft,
    stress = varst,
    e5 = vare5;
var exercise fitness hardy stress illness;
```

**STEP 3 – Modified Path Analysis Model (see Figure 5)**

```plaintext
proc calis data=illfl corr stderr;
lineqs
  fitness = p3p1 exercise + e3,
  stress = p4p2 hardy + e4,
  illness = p5p3 fitness + p5p4 stress + e5;
  std
    exercise = vare1,
    hardy = vare2,
    e3-e5 = vare3-vare5;
var exercise fitness hardy stress illness;
```

Figure 5. Structural Equation Model - Illness
STEP 4 – Interpret Results

The regression analysis with backward elimination retained fitness and stress (p < 0.0001) in the model while removing exercise and then hardy. R-squared with all variables in the model is equal to 0.1835. R-squared for the revised model is equal to 0.177.

The path analysis model showed some interesting results in terms of model fit: chi-square = 0.000, df = 0, p < 0.0001 and CFI = 1.000 and RMSEA = 0.000. Similar to the regression analysis, significant parameter estimates were fitness (z = -5.067) and stress (z = 6.000) while parameter estimates for hardy (z = -1.530) and exercise (z = 0.663) were not significant.

<table>
<thead>
<tr>
<th>Table 1. Parameter Estimates – Path Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>p5p1</td>
</tr>
<tr>
<td>p5p2</td>
</tr>
<tr>
<td>p5p3</td>
</tr>
<tr>
<td>p5p4</td>
</tr>
<tr>
<td>e5</td>
</tr>
</tbody>
</table>

Standardized and unstandardized parameter estimates were equal because input was in the form of a correlation matrix (standardized values, mean = 0, std = 1). The standardized equation is

illness = 0.034*exercise – 0.260*fitness – 0.074*hardy + 0.291*stress + 0.817*e5

R-squared (1 - unexplained variance squared) from the path analysis model is equal to 0.1835 (1 – 0.9036^2 = 1 – 0.8165). The unexplained variance is the amount of variance that cannot be accounted for with the predictor variables.

The structural equation model with hardness mediated by stress and exercise mediated by fitness showed acceptable fit on three measures, chi-square (11.078, df = 5, p = 0.050), CFI (0.961), and RMSEA (0.057).

Unstandardized and standardized parameter estimates are equal due to input in the form of a correlation matrix (standardized, mean = 0, std = 1). The standardized equations are

Fitness = 0.390*exercise + 0.848*e3
Stress  = -0.230*hardy + 0.947*e4
Illness = -0.250*fitness + 0.308*stress + 0.823*e5

<table>
<thead>
<tr>
<th>Table 2. Parameter Estimates – Structural Equation Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter</td>
</tr>
<tr>
<td>p3p1</td>
</tr>
<tr>
<td>p4p2</td>
</tr>
<tr>
<td>p5p3</td>
</tr>
<tr>
<td>p5p4</td>
</tr>
<tr>
<td>varex</td>
</tr>
<tr>
<td>varhr</td>
</tr>
<tr>
<td>varft</td>
</tr>
<tr>
<td>varst</td>
</tr>
<tr>
<td>e5</td>
</tr>
</tbody>
</table>
Discussion
The regression and SEM model specify different relationships between variables in the model. The models include the same predictor and predicted variables in different configurations. Although statistical tests of significance differ, the amount of variance explained in each model is equal.

The development of theoretical models prior to SEM data analysis is critical. The results of the SEM analysis can serve to support or refute previous research. The direction of parameter estimates indicates effects on illness. More hardiness indicates less stress and less stress indicates less illness. More exercise indicates better fitness and less illness. The SEM analysis provides flexibility in determining the relationships between variables. Direct as well as indirect relationships between variables can be specified and estimated.

Conclusion
This presentation has compared regression analysis and path analysis. There are differences and similarities between SEM and “traditional” statistical techniques. With SEM methodology, models are specified a priori, measurement error is specified explicitly, and models are tested for acceptable fit with chi-square and several fit indices. SEM gives you the power not available with “traditional” statistical procedures. You are challenged to design and plan research where SEM is an appropriate analysis tool.

WAM
Path Analysis is a flexible and powerful statistical methodology used to examine the relationships between measured variables.

References


**Contact Information**

Diana Suhr is a Statistical Analyst in the Office of Budget and Institutional Analysis at the University of Northern Colorado. She holds a Ph.D. in Educational Psychology. She has been a SAS programmer since 1984.

email: diana.suhr@unco.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.
STEP analysis is a strategic method corporations use to understand four major external environmental factors of the business landscape. What are the factors included in STEP analysis? STEP analytical trends and relationships. How should you look at trends for the future? Deducing future trends on STEP analytical information. How often should you do STEP analysis? Are there other versions of STEP analysis? Make smarter business decisions with STEP analysis. Path analysis was developed as a method of decomposing correlations into different pieces for interpretation of effects (e.g., how does parental education influence children's income 40 years later?). Path analysis is closely related to multiple regression; you might say that regression is a special case of path analysis. Some people call this stuff (path analysis and related techniques) "causal modeling." The reason for this name is that the techniques allow us to test theoretical propositions about cause and effect without manipulating variables. However, the "causal" The four general steps to conducting a Path Analysis in R include: Read in your data (as a correlation matrix or raw data). Specify the model. Fit the model. View the results. Read in your data. For this tutorial, we will use the mtcars dataset to demonstrate how to conduct a path analysis. However, a covariance matrix can also be used if necessary. The semPaths() function provides a quick and easy way to generate a visual representation of your model and automatically calculates key statistics that describe the relationships between the dependent variable and each independent variable. The SEM produced below is that of the mtcars model we created earlier in this tutorial.