Paper 3765-2019 and 3240-2019

Analyzing Structural Causal Models Using the CALIS Procedure

Banoo Madhanagopal, John Amrhein, McDougall Scientific Ltd.

ABSTRACT

Structural Equation Modeling (SEM) is a statistical technique to model hypothesized relationships among observed (manifest) and unobserved (latent) variables. SEM is not only widely applied in the social sciences, but is also suitable in areas such business, ecology, engineering, finance, pharmaceutical, and research. Under certain assumptions, a SEM can support causal inference as a Structural Causal Model (SCM). Path diagrams, commonly used with SEM, are visual representations of the hypothesized associations and dependencies and are particularly useful when studying causality.

This paper describes how to formulate and interpret structural models as causal models. Using the PATH modeling language within the CALIS procedure, we fit SEMs for causal inference; we focus on model hypothesis and modification using fit statistics, but also briefly describe how to interpret model estimates to infer causality from direct and indirect effects. SEM is appropriate for both observational data and controlled experiments. Therefore, we support our discussion with two examples: the first application analyzes observational data from anonymized logs of a web site to infer the page causal dependencies i.e. which pages lead to visits of other pages; and the second example uses flow cytometry data from a cell signaling experiment to understand and discover the complex structure of the protein signaling pathways.

INTRODUCTION

The statistical terms correlation and causation are often misunderstood and used interchangeably. Correlation (or association) occurs when two or more variables' values change together in a measurable relationship. Causation is the effect that *changes* of one variable have on another variable's values. We have all heard many times that "correlation does not imply causation"; when two variables are correlated, it does not imply that *changing* one affects *change* in the other. For many research questions, correlation-based conclusions are provided based on the patterns observed, but we fail to investigate causal relations. Understanding the differences between the two goes a long way to support business decisions or developing a new intervention for a treatment, because the usefulness of causal results is far greater than correlational results.

Especially in an observational study, in which there was no experimental design giving rise to the data, correlational results are often "discovered". That is, we had no hypotheses of prior relationships among the variables. Although causal results may be discovered in a similar manner, it is more often the case that causal relationships are pre-specified, and the analyses are conducted to confirm or refute our hypotheses. The pre-specified causal relationships are learned from prior scientific studies or research and require the input of a subject matter expert.

Structural Equation Modeling (SEM) is a statistical technique to model hypothesized relationships among variables. We begin by hypothesizing or specifying assumed relationships between a set of variables. This can be done graphically (as we do in this paper) or by listing a set of functions is what is meant by "structural" in "structural equation modeling". We often rely on subject matter expertise to hypothesize the model structure. The purpose of the analysis is to confirm or refute the model structure. This is an important concept and might differ from the usual analyses of discovery with which we have become accustomed.

The variables in a SEM may be manifest (observed) or latent (unobserved). Variables are further classified as exogenous, which have no causes themselves but might affect the values of other variables, and endogenous, whose values are caused by other variables (which may be exogenous or endogenous).

Relationships between the variables belong to one of the following types:

- Correlational or Bidirectional
- Isolated or Conditionally Independent
- **Causal or Unidirectional (the focus of this paper)**

It is useful to visualize an SEM as a graphical model. Figure 1 shows a simple example of a SEM graph or pathway, illustrating the concepts of variable and relationship types discussed in this introduction. In Figure 1, we use a common standard by representing latent variables with ovals and manifest variables with rectangles. Variables A and W are exogenous variables because they have no single-headed arrows entering them. Variables X and Y are endogenous because they are the children of parents; A is the parent of X, and X and W are the parents of Y. The double-headed arrows represent covariances (between X and W) or variances (of Y). The three relationship types are also represented in Figure 1. X and W are assumed to have a correlational relationship, as indicated by the double-headed arrow. A and W are assumed to be conditionally independent; they are isolated from each other via the lack of a connecting arrow. Several causal relationships are assumed; A to X, X to Y, and W to Y. The omission of an arrow between X and W indicates a strong causal claim that there is no causal effect between the variables.

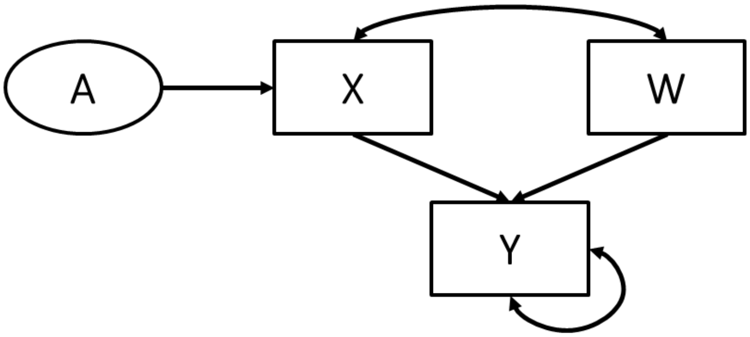


Figure 1 Graphical Representation of a Structural Equation Model

It is important that "causal" in SEM terminology does not mean "causal" as we defined it here; i.e. it does not mean that the change in a parent variable affects a change in a child variable. SEM "causal" is better understood as "predictive" or "explanatory", like regression modeling. But, in this paper, we use "causal" in the usual, non-SEM, definition. Beginning in the next section, we describe the conditions your SEM must meet to allow declaration of cause and effect.

WHAT MAKES A STRUCTURAL EQUATION MODEL A STRUCTURAL CAUSAL MODEL?

A SCM (Structural Causal Model), proposed by Pearl, integrates SEM and graphical models to help us understand causal relationships. SEMs are predominantly used to confirm a model rather than to explore a phenomenon. SEMs can be interpreted for cause and effect, that is, as SCMs, when the following conditions are met:

- The structure is a valid representation of reality
- The relationships are directed and acyclic
- Variables, conditioned on their parents, are independent of their ancestors
- There are no "back doors" from cause to effect

We discuss each of these in turn.

MODEL STRUCTURE IS A VALID REPRESENTATION OF REALITY

Causal modeling begins by drawing a graphical representation, like Figure 2, that represents all factors, that might affect the effect of interest. Subject matter experts should be consulted to ensure that no factors are omitted. You should not be concerned whether the factors have been measured; it is important to include all factors in a structure so that it reflects reality, or at least as you believe it to be. In this example, nutrition, motivation, and fitness are latent variables (we did not measure and record their values in the analysis data set), yet we believe them to be important variables to include so that our model *is a valid representation of reality*. Suppose we measure each person's activity level, perhaps in hours of rigorous physical activity per week. If we can assume that the measure of activity accounts for a person's general health and motivation, then activity captures unobserved attributes (of an individual) that affect heatstroke.

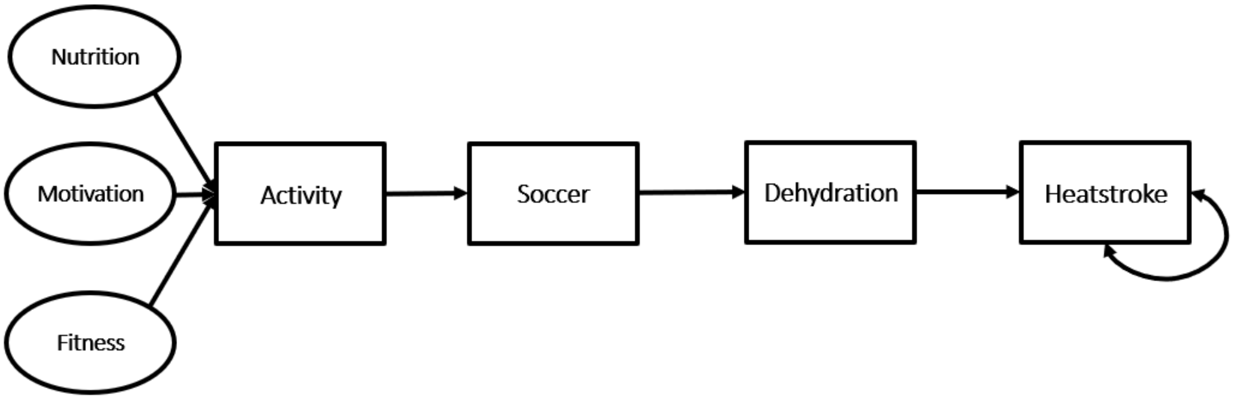


Figure 2 Hypothesized Dehydration-Heatstroke Model

RELATIONSHIPS BETWEEN VARIABLES ARE DIRECTED AND ACYCLIC

Directed Acyclic Graphs (DAGs) are a subset of all graphical models. Directed means that the relationships must be single-headed arrows, starting from one variable, (called the parent) and ending in another variable, (called the child). Acyclic means that no loops exist in the graph. To illustrate, consider the example in which heatstroke is assumed to be caused by dehydration from playing summer sports, like soccer. In Figure 2, we hypothesize a directed, acyclic causal path for heatstroke. Note that the double-headed arrow indicates the variance for heatstroke and not a relationship. Variables are independent, conditional on parents

Another condition for causal inference is that each variable is conditionally independent of its ancestors, given its parents. In Figure 2, soccer is an ancestor of heatstroke. If heatstroke is independent of soccer given that we observe dehydration, then the condition is met. To say another way, soccer affects the occurrence of heatstroke only through its cause of dehydration. If soccer affects heatstroke directly, or through another mediating variable, such as fatigue as shown in Figure 3 then heatstroke is not conditionally independent of soccer, our model is incomplete (does not reflect reality), and causal inferences are suspect.

If our SEM meets the above 3 conditions to interpret cause and effect, then we can conclude that it is a causal model. However, if conditional it is appropriate to mention one caveat to the criterion of conditional independence, known as the 'Back-Door' criterion.

THERE ARE NO "BACK-DOORS" FROM CAUSE TO EFFECT

The *backdoor criterion* in a DAG requires that we have accounted for all possible paths from a cause under study to its effect of interest. Alternatively, 'Which variables to control in the model to control for confounding?'. A back-door path can convey a spurious relationship between the cause and effect, but never *explains* causation.

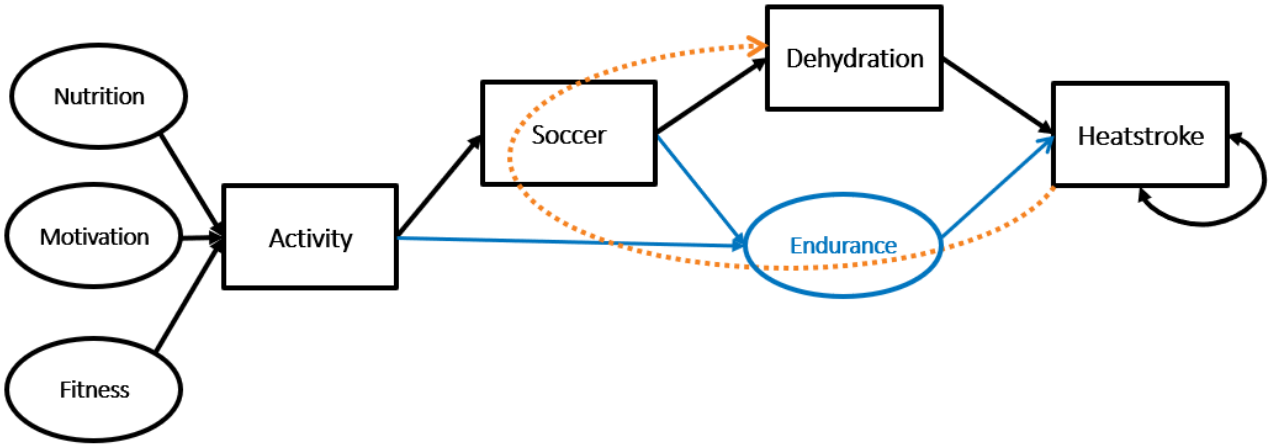


Figure 3 Back-Door Criterion

Consider the model in Figure 3. Suppose we are interested in the causal relationship between dehydration and heatstroke; i.e. dehydration is the cause under study and heatstroke is the effect of interest. On consultation with a subject matter expert, we decided that our model should include endurance as one of the causes of heatstroke. Do we need to control for endurance when estimating the causal effect of dehydration? In Figure 3, there *are* other variables that affect both dehydration and heatstroke; i.e. activity, endurance and soccer. Therefore, to correctly estimate the causal effect of dehydration on heatstroke, we need to "block" the path (i.e. close the backdoor) by controlling for a measured variable within the backdoor path to heatstroke; soccer in this example. Dehydration has backdoor access to heatstroke via soccer and endurance (the highlighted orange dashed arrow). However, if we control for soccer, then the backdoor is blocked and our causal conclusions about dehydration will be valid.

A thorough discussion of the backdoor criterion is beyond the scope of the paper. Readers are encouraged to consult one of the references by Pearl.

THE CALIS PROCEDURE

PATH MODELING LANGUAGE

PROC CALIS (Covariance Analysis and Linear Structural Equations) is the procedure in SAS/STAT for fitting SEMs. PROC CALIS incorporates eight different modeling languages, such as AMOS, COSAN, LINEQS, and LISREL, to appeal to a wide audience from different backgrounds. We use the PATH modeling language because it is an intuitive method to program graphical models. For example, the PATH statement used to code the model in Figure 2 is:

```
path
    heatstroke <--- dehydration,
    dehydration <--- soccer,
    soccer <--- activity,
    activity <--- individual,
    heatstroke <---> heatstroke;
```

or:

```
path
    dehydration ---> heatstroke,
    soccer ---> dehydration,
    activity ---> soccer,
    individual ---> activity,
    heatstroke <---> heatstroke;
```

You might choose the first syntax, as we do in this paper, because it mimics a MODEL statement in other SAS/STAT procedures. Or you might choose the second syntax because it is consistent with a graphical representation of a model that reads from left to right. Note that because nutrition, motivation, and fitness were not measured, they will not be variables in the input data set and will be represented by a single latent cause. These 3 variables, which are attributes of an individual, are collectively termed as 'individual' in our PATH statement.

Recall from our introduction that SEM is intended to confirm or refute a hypothesized model; SAS/STAT documentation refers to PROC CALIS as a "confirmatory analytic procedure". After fitting your hypothesized model, you might wish to refine the model based on the initial model fit. PROC CALIS provides capabilities for a process such as the following.

Step	PROC CALIS
1. Draw your hypothesized model diagram	1. Use a whiteboard, pencil and paper, or your favorite presentation software
2. Fit the model	2. PATH statement
3. Assess the fit	3. FITINDEX statement: goodness of fit statistics
4. Refine the model	4. MOD option on PROC statement: modification indices
5. Repeat steps 2, 3 and 4	5. PATH statement, FITINDEX statement, MOD option
6. Display final model diagram	6. PATHDIAGRAM statement
7. Assess causality	7. Evaluate the conditions for causal criteria

Table 1 Model Development Steps

COVARIANCE MATRIX

The fundamental unit of information in an SEM is the covariance matrix of the model variables. The number of unique elements within a covariance matrix with ‘k’ variables is equal to

$$i = \frac{1}{2}k(k + 1)$$

The number of unique observations including means is equal to

$$i = \frac{1}{2}k(k + 3)$$

For example, if we have 5 variables, then the variance-covariance matrix is 5x5 having 25 total elements. Out of these 25, 15 (5 variances and 10 covariances) are unique and capture the covariance structure of the data. The number of parameters that we estimate in our model cannot be greater than the number of unique elements, 15. If our analysis includes estimating means and intercepts, then we can estimate up to 20 parameters.

An ‘Under-Identified’ model is a model in which it is not possible to estimate all the model parameters because there are too few unique elements. A ‘Just-Identified’ model is a model in which the number of unique covariance elements equals the number of parameters being estimated. An ‘Over-Identified’ model is a model in which the number of unique covariance parameters is greater than the number of parameters being estimated. The difference is the degrees of freedom available for hypothesis tests. The total number of estimated parameters in the model should always be lower than fundamental unit of information in the data; i.e. the model should be over-identified.

There are some advantages to using a covariance matrix, rather than the raw data, as input, including;

- Covariance matrices preserve anonymity; e.g. protecting the identity of participants in clinical trials
- Ability to re-analyze a published covariance matrix
- Ability to analyze “big data” much more easily

GOODNESS OF FIT

PROC CALIS has more than two dozen different fit statistics that can be used to assess how well the model fits the data. Use the FITINDEX statement to specify which fit statistics to display in the Fit Summary table. Table 2 lists common indices, which fall into one of three categories;

1. Absolute or Standalone indices compare the fitted model to a saturated model and do not account for model complexity
2. Parsimony indices indicate how well the model fits the data, equivalently fits almost well any new data. These indices account for model complexity, penalizing complex models
3. Incremental indices compare the fitted model to the baseline model or null model containing only variance parameters, no covariances or coefficients

Symbol	Name	Description	Recommended Cut-offs
χ^2	Chi Square	An absolute index. Compares the hypothesized model to the full model with no constraints. Sensitive to sample size.	p-value >0.05
SRMR	Standardized Root Mean Square Residual	An absolute index. Root mean squared standardized residuals. Smaller is better.	< 0.08
RMSEA	Root Mean Square Error of Approximation	A parsimony index. If you use only one index, use this one. See Kelley and Lai (2011)	<.05=close fit <.08=mediocre >.1=poor fit
		RMSEA 90% Confidence Interval	Narrower is better
PROBCLFIT	Probability of Close Fit	A parsimony index. A chi-square test in which the null hypothesis is "close fit"	> 0.05
CAIC	Bozdogan Criterian AIC	Parsimony indices. Likelihood based. Penalizes for large samples and number of parameters.	Smaller is better
SBC	Schwarz Bayesian Criterion		
CFI	Bentler Comparative Fit Index	Incremental indices. Indexes amount of variance explained. Analogous to R^2. Preferable for smaller samples. NNFI is also known as Tucker Lewis Index (TLI).	>0.90
NNFI	Bentler-Bonett Non-normed Index		

Table 2 Common Goodness of Fit Statistics

Use a combination of fit indices to get a 'good-fitting' model before making causal inferences from the fitted SEM.

EXAMPLE 1: NAVIGATION WITHIN A WEBSITE

The weblogs example is a real-world data set of anonymized logs from a web site (Grozea, 2008). The data set consists of 20 variables whose values are the counts of daily visits to each of 20 web pages recorded over a period of 512 days; each row corresponds to one day's counts.

The web pages have links to other pages within the website that visitors use to navigate the site. The purpose of the SEM analysis is to infer page dependencies; i.e. which pages lead visitors to visit which other pages? For conciseness and to better visualize the path diagram, we limit the data to pages 1–7 in this example.

With the growing privacy concerns, weblog data such as cache, location, and IP addresses that are currently available to advertisers and ad platforms might become unavailable. More users are opting to prevent the tracking of their online navigation. Website owners and service providers may no longer be able to store or use such data. However, web designers and marketers will continue to be interested in the effectiveness of their website and ads in leading visitors to a target page such as a checkout page. SEMs using only the covariance matrices as input resolve this dilemma.

VISUALIZE THE HYPOTHESIZED MODEL

A DAG is drawn to visualize the relationships between the 7 variables in the weblogs data set. This graph helps us visualize our hypothesized dependencies. This first step is typically performed by subject matter experts; web designers and online marketing agents in this case. We did not have access to the SMEs, so we hypothesized that a visit to a page depended upon visits to the four preceding pages. For example, a visit to page 05 depended on visits to pages 01 – 04. The DAG for our hypothesized model, displayed in Figure 4, was generated using the PATHDIAGRAM statement in PROC CALIS.

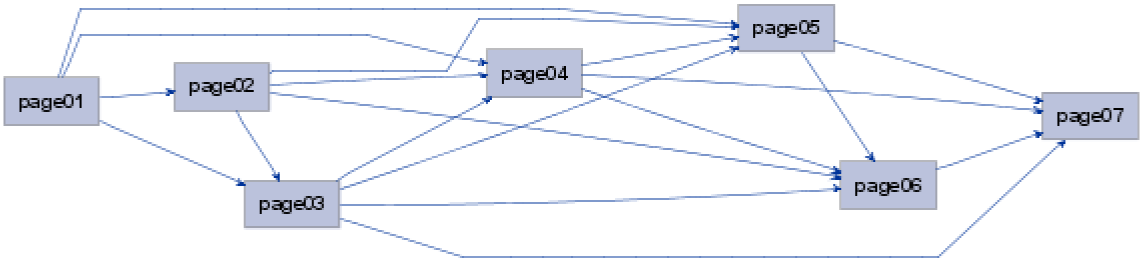


Figure 4 Graph of Weblogs Data

GENERATE THE COVARIANCE MATRIX

To demonstrate using a covariance matrix as input to PROC CALIS, we used the CORR Procedure to create the matrix. The following PROC CORR and DATA steps create two data sets; one containing the covariance matrix for pages 01 to 07, and the other containing the correlation matrix (which PROC CALIS also accepts as input):

```
 proc corr data=weblogs outp=corrout cov;
   var page01-page07;
run;

data webcorr(type=corr) webcov(type=cov);
  set corrout;
  if _type_ ne "COV" then output webcorr;
  if _type_ ne "CORR" then output webcov;
run;
```

PROC CALIS will read the metadata of the input data set to check the data set's type. Therefore, we set the type to CORR or COV using the TYPE= data set option. The OUTP= data set created by PROC CORR will contain a variable named _TYPE_. Rows corresponding to the covariance matrix will have _TYPE_ equal to "COV". Other rows will have _TYPE_ equal to "MEAN", "STD", or "N". PROC CALIS needs these statistics, so keep those rows in your data set.

Here we have 7 variables (page01-page07). Therefore, the variance-covariance matrix is 7x7. Out of 49 values in the matrix, 28 (7 variances, 21 covariances) are unique values representing the covariance structure. Therefore, our SEM will be over-identified if we are estimating fewer than 28 parameters.

TRANSLATE THE DAG TO A PATH STATEMENT

Each of the single headed arrows in Figure 4 represents a hypothesized dependency. For each of these paths, PROC CALIS will estimate a path coefficient and test whether the coefficient statistically differs from zero. A PROC CALIS step, using the PATH language, that fits our hypothesized model in Figure 4 is:

```
   proc calis data=webcov toteff mod;
   fitindex on(only) = [chisq df probchi bentlernnfi cfi rmsea rmsea_ll
     rmsea_ul probclfit srmsr caic sbc];
   path
    page02 <-- page01 = one_2,
    page03 <-- page01 page02 = one_3 two_3,
    page04 <-- page01 page02 page03 = one_4 two_4 three_4,
    page05 <-- page01 page02 page03 page04 = one_5 two_5 three_5 four_5,
    page06 <-- page02 page03 page04 page05 = two_6 three_6 four_6 five_6,
    page07 <-- page03 page04 page05 page06 = three_7 four_7 five_7 six_7;
   pathdiagram notitle fitindex=[chisq df probchi cfi rmsea srmsr caic sbc];
  run;
```

The TOTEFF=option on the PROC CALIS statement requests estimates and significance tests for total, direct and indirect effects (we discuss these later). The MOD= option requests modification indices, Lagrange Multipliers (LM) and Wald statistics, which we will use to guide model modifications. The FITINDEX statement ON(ONLY)= option limits the fit statistics to those specified in Table 2. The PATH statement specifies path coefficients to be estimated. Following the equal sign for each path we specify names that we want PROC CALIS to use to label parameters in the output. The names we use indicate the "from" and "to" pages in a path. For example, "one_3" will label the parameter associated with the path from page 01 to page 03. When there is more than one variable in a single path, the names must be in the same order as the variables. Otherwise you will misinterpret your output. The PATHDIAGRAM statement draws the path diagram of the model. The FITINDEX= option specifies which fit indices you want displayed within the diagram. This gives a complete picture of the model with fit indices in a single display.

ASSESS FIT STATISTICS

Only certain output from PROC CALIS is discussed here. The 'Modeling Information' table shows that the model was fit using 512 observations, which PROC CALIS read from the row of the input data set in which _TYPE_=N.

Modeling Information	
Maximum Likelihood Estimation	
Data Set	WEBCOV
N Obs	512
Model Type	PATH
Analysis	Covariances

Table 3 Modeling Information for Weblogs Data

The 'Variables in the Model' table can be used to verify if we correctly translated our hypothesized model on the PATH statement; i.e. if the endogenous, exogeneous, manifest, and latent variables are as we intended. In this model, all the variables are identified as manifest (observed) variables of which 6 are endogenous variables and 1 is an exogeneous variable (page 01 has no arrows going into it). Latent variables can be specified on the PATH statement but do not correspond to any variables represented in the raw data or input covariance matrix.

Variables in the Model		
Endogenous	Manifest	page02 page03 page04 page05 page06 page07
	Latent	
Exogenous	Manifest	page01
	Latent	
Number of Endogenous Variables = 6 Number of Exogenous Variables = 1		

Table 4 Variables Information for Weblogs Data

The model satisfied convergence criterion; confirmed by a note in the SAS log. The 'Path List' table, not shown here, displays the path coefficients for every path coded on the PATH statement. The 'Variance Parameter' table, also not shown here, displays the estimates of exogeneous variables and error variances of the endogenous variables in the model.

Fit Summary		
Absolute Index	Chi-Square	7.3652
	Chi-Square DF	3
	Pr > Chi-Square	0.0611
	Standardized RMR (SRMR)	0.0028
Parsimony Index	RMSEA Estimate	0.0534
	RMSEA Lower 90% Confidence Limit	0.0000
	RMSEA Upper 90% Confidence Limit	0.1035
	Probability of Close Fit	0.3779
	Bozdogan CAIC	188.3233
	Schwarz Bayesian Criterion	163.3233
Incremental Index	Bentler Comparative Fit Index	0.9993
	Bentler-Bonett Non-normed Index	0.9952

Table 5 Fit Summary for Weblogs Initial Model

The 'Fit Summary' table shows only the fit indices that we requested on the FITINDEX statement. The chi-square is non-significant, indicating that our hypothesized model is not statistically different from the saturated model. The SRMR is below 0.08, signifying a small deviation in residuals between the fitted model and hypothesized model. The RMSEA is in the mediocre range, signifying a medium amount of variance explained by the model. The probability of close fit is greater than 0.05 which suggests a good fit. The comparative fit index and non-normed index are both greater than 0.9. We will interpret SBC and BCAIC only after refitting a modified model; decreasing values will suggest a better fitting model.

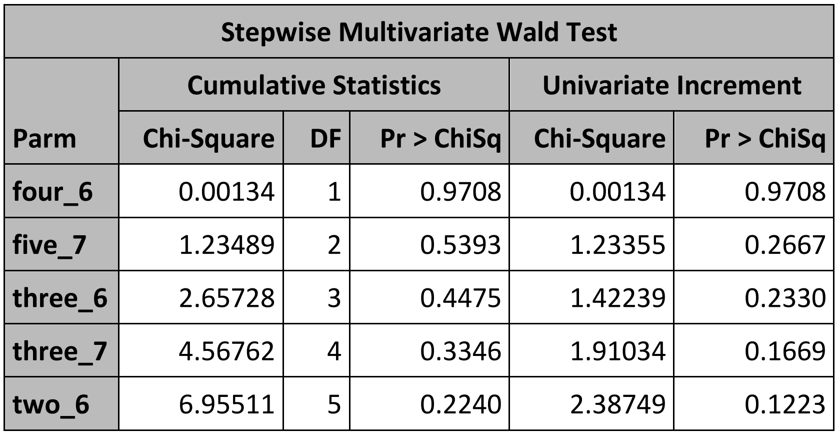
IMPROVE MODEL FIT USING MODIFICATION INDICES

The overall model fit was good, so we could stop here and conclude that our hypothesized model has been confirmed. However, the RMSEA = 0.0534, is a bit higher than we would like, so we decide to explore model modifications to improve the fit. The model can be modified by:

1. Increasing the number of paths (i.e. allowing the corresponding coefficients to be estimated)
2. Reducing the number of paths (i.e. constraining the corresponding coefficients to zero)

PROC CALIS computes Wald Test Indices that suggest paths to remove without affecting the chi-square statistic, and Lagrange Multiplier Indices that suggest paths to add to increase the chi-square statistic.

Wald Test



Stepwise Multivariate Wald Test					
	Cumulative Statistics			Univariate Increment	
Parm	Chi-Square	DF	Pr > ChiSq	Chi-Square	Pr > ChiSq
four_6	0.00134	1	0.9708	0.00134	0.9708
five_7	1.23489	2	0.5393	1.23355	0.2667
three_6	2.65728	3	0.4475	1.42239	0.2330
three_7	4.56762	4	0.3346	1.91034	0.1669
two_6	6.95511	5	0.2240	2.38749	0.1223

Table 6 Partial Results from Wald Test

Following the recommendations from Wald Test Indices, the model was refit without the direct paths page 02 to page 06, page 03 to page 06, and page 04 to page 06. We chose these 3 out of the top 5 because they are common to page 06. The revised PATH statement, which we label *'Deletion Model'* in Table 8 is:

```
path
page02 <-- page01 = one_2,
page03 <-- page01 page02 = one_3 two_3,
page04 <-- page01 page02 page03 = one_4 two_4 three_4,
page05 <-- page01 page02 page03 page04 = one_5 two_5 three_5 four_5,
page06 <-- page05 = five_6,
page07 <-- page03 page04 page05 page06 = three_7 four_7 five_7 six_7;
```

The overall chi-square is still non-significant, so we succeeded in simplifying our model without damaging its comparison to a saturated model. Note that the degrees of freedom increased by 3 because we are estimating 3 fewer parameters. We see improved error variance with the RMSEA=0.0410 now in the close-fit range. We can now use our incremental indices, BCAIC and SBC. For both, smaller values are better, and the values reduced significantly. This *'Deletion Model'* shows a better fit to the data than the '*Initial Model'.*

Lagrange Multiplier Test

Rank Order of the 10 Largest LM Stat for Path Relations				
To	From	LM Stat	Pr > ChiSq	Parm Change
page05	page07	6.21768	0.0126	0.02115
page02	page07	3.48662	0.0619	0.20863

Table 7 Partial Results from LM Statistics

The model fit might also be improved by using the Lagrange Multiplier (LM) Indices to guide us in adding paths. Subject matter experts should provide input to decisions about model modifications. In this example, we lack the subject matter expertise but, based on the LM indices, we added a direct path from page 07 to page 05. The revised PATH statement, which we label *'Addition Model'* in Table 8 is:

```
path
page02 <-- page01 = one_2,
page03 <-- page01 page02 = one_3 two_3,
page04 <-- page01 page02 page03 = one_4 two_4 three_4,
page05 <-- page01 page02 page03 page04 page07 = one_5 two_5 three_5
           four_5 seven_5,
page06 <-- page05 = five_6,
page07 <-- page03 page04 page05 page06 = three_7 four_7 five_7 six_7;
```

We can see that model fit is improved compared to the *'Deletion Model'*; the CFI and NNFI have values 1.0, the RMSEA value of zero indicates an acceptable model fit, and the chi-square p-value jumped to 0.43.

Fit Index	Initial Model	Deletion Model	Addition Model
Chi-Square	7.3652	11.1622	4.8937
Chi-Square DF	3	6	5
Pr > Chi-Square	0.0611	0.0835	0.4290
Standardized RMR (SRMR)	0.0028	0.0043	0.0045
RMSEA Estimate	0.0534	0.0410	0.0000
RMSEA 90% CI Lower	0.0000	0.0000	0.0000
RMSEA 90% CI Upper	0.1035	0.0780	0.0608
Probability of Close Fit	0.3779	0.6023	0.8818
Bozdogan CAIC	188.3233	170.4054	171.3751
Schwarz Bayesian Criterion	163.3233	148.4054	148.3751
Bentler Comparative Fit	0.9993	0.9992	1.0000
Bentler-Bonett Non-normed	0.9952	0.9972	1.0001

Table 8 Comparison of Fit Indices for Weblogs Data

CONCLUSION

Recall the conditions for a structural equation model to be a structural causal model. Let us now examine the final weblogs model to see if it met the criteria for a causal model.

Condition	Is Weblogs a Causal Model?
1. Reflect Reality?	1. No. Our model did not capture unobserved causes like page loading time (connectivity) or attributes of a visitor, such as socioeconomic status.
2. Directed and Acyclic	2. No. We added a path from page 7 back to page 5, which introduced a loop between the pages. That path is well supported by the data (indeed, it was suggested by the data). But a SME should assist in determining if this is reasonable or even possible given the website design.
3. Conditionally Independent	3. Maybe. Due to the lack of an available SME, we are unsure if our formulated hypothesis represents reality; i.e. whether we have omitted any paths.
4. "Back-Doors" Blocked?	4. Maybe. Yet again, we need a SME to confirm if there are any spurious relationships in the model across pages or via latent variables such as one representing 'computing environment'.

Table 9 Causal Criteria for the Weblogs Model

EXAMPLE 2: PROTEIN SIGNALING NETWORKS IN HUMAN T-CELLS

Our second example uses data from a designed experiment described in "Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data" published in *Science* (Sachs at al., April 2005). The data set contains approximately 700 to 900 single cell readings of 11 different phosphoproteins or phospholipids measured under 9 experimental conditions in human naïve CD4+ T cells. Simultaneous expression of multiple proteins is recorded via flow cytometry. The data was used to analyze signaling pathways.

A protein, in response to an extracellular signal, might trigger a response in a subsequent protein molecule, affecting its physiochemical properties. Proteins gain new functional capabilities via these signaling pathways. The extracellular signals are either stimulatory or inhibitory conditions. This experiment introduced extracellular signals and measured protein responses via flow cytometry. The purpose of the SEM analysis was to confirm hypothesized pathways, discover novel pathways, and to understand the protein signaling network causalities.

Cancer and autoimmune disease can occur at any age. The network connections between the molecules in a protein signaling pathway helps in understanding the underlying biological process, which, in turn, aids in developing new therapeutics for diseases that have abnormal signaling pathways.

Sachs et al. do not use SEMs to analyze their data. Instead, they use Bayesian Networks (BNs), which is an alternative method to understand causal pathways. BNs estimate conditional probabilities; i.e. the probabilities of outcome values conditioned on values of parent nodes. For an introduction to Bayesian Networks using SAS Enterprise Miner, see Wang and Amrhein, 2018.

VISUALIZE THE HYPOTHESIZED MODEL

The graph displayed in Figure 5 is the model hypothesized by Sachs et al. as shown in their Fig. 3.A. It represents relationships between 11 different signaling proteins. The pathways represented by blue arrows and the sole purple arrow are the pathways which we translated to a PATH statement in PROC CALIS.

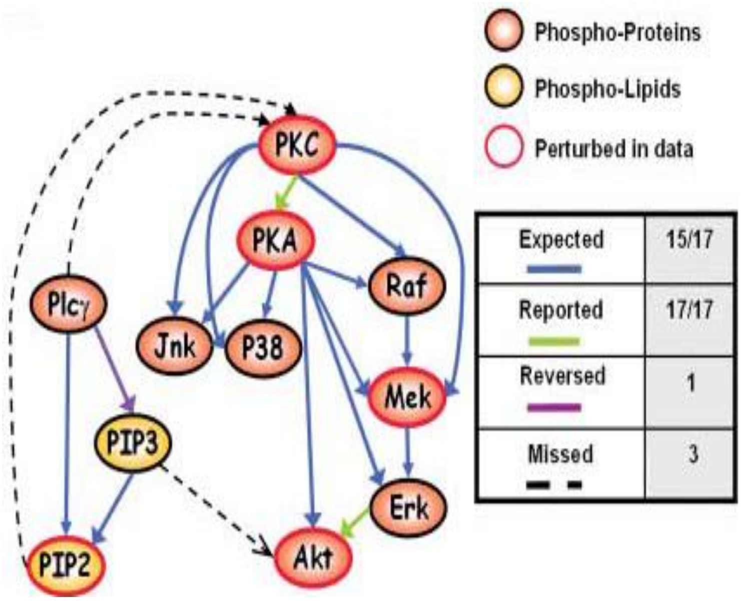


Figure 5 Protein Signaling Pathways (Fig. 3.A in Sachs et al)

TRANSLATE THE DAG TO A PATH STATEMENT

As per common practice with flow cytometry data, we transformed the values using the natural logarithm.

PROC CORR was again used to create the covariance matrix from the raw data. We have 11 variables (praf, pmek, plcg, PIP2, PIP3, p44/42, pakts473, PKA, PKC, P38, pjnk). Therefore, the variance-covariance matrix is 11x11. Out of 121 values in the matrix, 66 (11 variances, 55 covariances) are unique values representing the covariance structure. Therefore, our SEM will be over-identified if we are estimating fewer than 66 parameters.

Each of the single headed arrows in Figure 5 represents a hypothesized dependency. A PROC CALIS step, using the PATH language, that fits our hypothesized model in Figure 5 is:

```
 proc calis data=procov toteff mod;
   fitindex on(only) = [chisq df probchi bentlernnfi cfi rmsea rmsea_ll
                        rmsea_ul probclfit srmsr caic sbc];
   path
     pjnk p38 praf pmek <-- pkc,
     pjnk p38 pakts473 "p44/42"N pmek praf <-- pka,
     pmek <-- praf,
     "p44/42"N <-- pmek,
     plcg <-- pip3,
     pip2 <-- pip3 plcg;
   pathdiagram notitle fitindex= [chisq df probchi cfi rmsea srmsr caic sbc];
run;
```

The PROC CALIS options used in the code are described in the weblogs example. In this example we chose not to label the parameters associated with the paths. Note that, to retain the p44/42 column name from the article, we use the "*variable-name*"N naming construct. This allows non-standard characters, such as /, in variable names. This requires that you specify the VALIDVARNAME=ANY system option on the OPTIONS statement.

ASSESS FIT STATISTICS

Only certain output from PROC CALIS is discussed here. The 'Modeling Information' table shows that the model was fit assuming 7466 observations were used to create the covariance matrix.

Modeling Information	
Maximum Likelihood Estimation	
Data Set	PROCOV
N Obs	7466
Model Type	PATH
Analysis	Covariances

Table 10 Modeling Information for Protein Signaling Data

The 'Variables in the Model' table can be used to verify if we correctly translated our hypothesized model on the PATH statement. In this model, all the variables are identified as manifest (observed) variables of which 8 are endogenous and 3 are exogeneous.

Variables in the Model		
Endogenous	**Manifest**	P38 p44/42 pakts473 PIP2 pjnk plcg pmek praf
	Latent	
Exogenous	**Manifest**	PIP3 PKA PKC
	Latent	
Number of Endogenous Variables = 8 **Number of Exogenous Variables = 3**		

Table 11 Variables Information for Protein Signaling Data

The 'Fit Summary' table shows only the fit indices that we requested on the FITINDEX statement. The chi-square is significant, indicating that our hypothesized model differs significantly from the saturated model. The SRMR and RMSEA both indicate high residual error. The probability of close fit suggests a poor fitting model, as do the two incremental indices.

Fit Summary		
Absolute Index	**Chi-Square**	15800.9551
	Chi-Square DF	37
	Pr > Chi-Square	<.0001
	Standardized RMR (SRMR)	0.2114
Parsimony Index	**RMSEA Estimate**	0.2389

Fit Summary		
	RMSEA Lower 90% Confidence Limit	0.2358
	RMSEA Upper 90% Confidence Limit	0.2420
	Probability of Close Fit	<.0001
	Bozdogan CAIC	16088.5804
	Schwarz Bayesian Criterion	16059.5804
Incremental Index	Bentler Comparative Fit Index	0.6396
	Bentler-Bonett Non-normed Index	0.4643

Table 12 Fit Summary for Protein Signaling Initial Model

IMPROVE MODEL FIT USING MODIFICATION INDICES

Given that the initial model fit is not acceptable, we conclude that our hypothesized model is inadequate, and we therefore modify it to improve the fit.

Wald Test

Stepwise Multivariate Wald Test					
	Cumulative Statistics			Univariate Increment	
Parm	Chi-Square	DF	Pr > ChiSq	Chi-Square	Pr > ChiSq
_Parm08	2.42271	1	0.1196	2.42271	0.1196

Table 13 Results from Wald Test

All the estimates for the paths specified in the model were statistically significant except for the direct path from pka to p44/42 (Parm08) as recommended by the Wald Test Indices (see Table 13). Therefore, we refit the model without this path. The revised PATH statement is:

```
path
  pjnk p38 praf pmek <-- pkc,
  pjnk p38 pakts473 /*"p44/42"N*/ pmek praf <-- pka,
  pmek <-- praf,
  "p44/42"N <-- pmek,
  plcg <-- pip3,
  pip2 <-- pip3 plcg;
```

Lagrange Multiplier Test

Rank Order of the 10 Largest LM Stat for Path Relations				
To	**From**	**LM Stat**	**Pr > ChiSq**	**Parm Change**
p44/42	**pakts473**	3525	<.0001	0.76253
pakts473	**p44/42**	3510	<.0001	0.59219
plcg	**PKA**	2377	<.0001	-0.49547
PIP3	**plcg**	2169	<.0001	2.95210
PKA	**PKC**	2007	<.0001	0.36651
PKA	**plcg**	1710	<.0001	-0.51008
PKC	**PKA**	1652	<.0001	0.32101
plcg	**P38**	1568	<.0001	0.42232
plcg	**pakts473**	1432	<.0001	0.56034
plcg	**pjnk**	1256	<.0001	0.33952

Table 14 LM Statistics for *'Deletion Model'*

Removing path pka → p44/42 did not improve the fit appreciably (see *'Deletion Model'* in Table 16), so we used the Lagrange Multiplier (LM) Indices to guide us in adding parameters. Based on the LM statistics and subject matter expertise (Fig. 3.A in Sachs et al.), we made the following changes:

- Added path p44/42 → pakts473
- Reversed path pip3 → plcg to plcg → pip3
- Added pkc →pka

The new path statement is:

```
path
 pjnk p38 praf pmek pka <-- pkc,
 pjnk p38 pakts473 /*"p44/42"N*/ pmek praf <-- pka,
 pmek <-- praf,
 "p44/42"N <-- pmek,
 pip3 <-- plcg,
 pip2 <-- pip3 plcg,
 pakts473 <-- "p44/42"N;
```

The fit indices for this model are labeled *'Addition Model'* in Table 16. Overall, the fit statistics improved but still indicate a poor fit.

Ranked Error Variances and Covariances

You can continue to modify your model using the modification indices and input from subject matter experts until you are satisfied with the final model. Keep in mind that you should not treat your modified model as the original hypothesized model, but rather as a "discovered" model that will need to be confirmed using new data.

So far, we have discussed only the addition or deletion of paths. However, you can also consider variances and covariances. PROC CALIS automatically estimates error variances for all manifest and latent variables, and all covariances between *exogenous* variables. The LM test indices might suggest adding covariances between *endogenous* variables. Covariance between endogenous variables is between their residuals, which might be correlated. Accounting for this correlation might improve the model fit.

Rank Order of the 10 Largest LM Stat for Error Variances and Covariances				
Error of	**Error of**	**LM Stat**	**Pr > ChiSq**	**Parm Change**
pakts473	**p44/42**	2500	<.0001	6.76646
praf	**pakts473**	1394	<.0001	0.28986
pmek	**pakts473**	1129	<.0001	0.24614

Table 15 Partial Results for highest ranked covariances from '*Addition Model*'

Based on the LM test indices for the *'Addition Model'* (Table 15), we added the 3 highest ranked error covariances to the model using the PCOV statement. On the PCOV statement, you specify the two endogenous variables followed by an equal sign and then a name to identify the covariance parameter:

```
pcov
  pakts473 "p44/42"N = cov_pakt_p44,
  praf pakts473 = cov_praf_pakt,
  pmek pakts473 = cov_pmek_pakt;
```

It is evident from the fit statistics, which we label *'Covariance Model'* in Table 16, that this model does not fit the data much better than 'Addition Model'. Again, the subject matter expert should determine if this a reasonable approach.

Fit Indices	Initial Model	Deletion Model	Addition Model	Covariance Model
Chi-Square	15800.9551	15803.3774	10393.8438	7235.1369
Chi-Square DF	37	38	38	35
Pr > Chi-Square	<.0001	<.0001	<.0001	<.0001
SRMR	0.2114	0.2120	0.1632	0.1541
RMSEA Estimate	0.2389	0.2357	0.1911	0.1660
RMSEA Lower CL	0.2358	0.2327	0.1880	0.1628
RMSEA Upper CL	0.2420	0.2388	0.1942	0.1692
Prob(Close Fit)	<.0001	<.0001	<.0001	<.0001
Bozdogan CAIC	16088.5804	16081.0846	10671.5510	7542.5985
SBC	16059.5804	16053.0846	10643.5510	7511.5985
Bentler CFI	0.6396	0.6396	0.7633	0.8354
NNFI	0.4643	0.4784	0.6573	0.7413

Table 16 Comparison of Fit Indices for Protein Signaling Data

Latent Variables

Our two examples have not included any latent variables. In the weblogs example, we did conclude that some latent constructs, such as the visitor's socioeconomic status, might affect their navigation behavior. But, in the absence of subject matter expertise, it is difficult to insert this into the model.

In the protein experiment, Sachs et al. discuss the possible effects of unmeasured proteins. For example, they illustrate in their Figure 3.B the hypothesized influence of unmeasured proteins mediating certain pathways (i.e. getting in between the parent and child), such as pkc to pjnk. We added latent variables to our model at the hypothesized positions, but these models failed to converge.

CONCLUSION

As we did in the weblog example, we can evaluate the four conditions required to interpret our protein signaling model as a causal model.

Condition	Is Protein Signaling a Causal Model?
1. Reflect Reality?	1. Yes. The experiment was designed by subject matter experts. If no possible causes to protein inhibition or stimulation are omitted, then the network reflects reality.
2. Directed, and Acyclic	2. Maybe. As discussed by Sachs et al., protein signaling pathways contain feedback loops and cyclic paths, which confounds cause-effect relationships, making it difficult to estimate the magnitude of a cause. However, if the feedback is lagged rather than simultaneous, then we may be able to conclude that the model is a DAG.
3. Conditionally Independent	3. Yes. These are classical signaling pathways that connect proteins in human T-cell, which were developed by subject matter experts (such as, cell biologists and geneticists).
4. "Back-Doors" Blocked?	4. If we meet the conditional independence criterion then there are no spurious backdoor paths. Again, this requires insight from a subject matter expert.

Table 17 Causal Criteria for the Protein Signaling Model

DISCUSSION

MEDIATION

The protein network in example 2 contains both direct and indirect connections between the signaling molecules. A direct connection does not have any variables between the parent and child variables; e.g. the direct connection between pip3 and pip2 (see Figure 6). An indirect connection is one that passes through another variable or variables; i.e. from ancestor to parent to child such as pip3 to plcg to pip2. PLCG is known as a mediator because it alters the effect that pip3 has on pip2. This begs the question of whether pip3 effects pip2 primarily on its own via the direct connection or through its mediator, plcg.

Mediation can be complete or partial. Complete mediation occurs when an indirect effect is significant, and the direct effect is not significant so that the effect is only through the mediator. Partial mediation is observed when both indirect effect and direct effects are significant, which means part of the effect is mediated and the remaining is direct.

A total effect does not have to be significant to validate mediation due to a property known as inconsistent mediation. Inconsistent mediation occurs when direct and indirect effects have different signs. For example: a positive indirect effect and a negative direct effect will cancel out each other resulting in a total effect that is not significant.

PROC CALIS provides estimates for direct, indirect and total effects for any two exogenous variables. The option TOTEFF on the PROC CALIS statement requests partitioning of total effects into direct indirect effects.

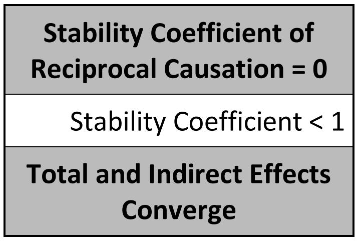
The Stability coefficient

Partitioning total effects into direct and indirect effects relies on a condition regarding the convergence of total effects. This condition can be assessed using a measure known as the stability coefficient and might be in question with models containing reciprocal or cyclic paths. Therefore, before we analyze total and indirect effects in our model, we should check this measure either in the SAS log or in the output table "Stability Coefficient". The stability coefficient must be less than 1.

In the SAS log:

> *"NOTE: The stability coefficient is 0, which is less than one. The condition for converged total and indirect effects is satisfied"*

In the PROC CALIS output:



Stability Coefficient of Reciprocal Causation = 0
Stability Coefficient < 1
Total and Indirect Effects Converge

Table 18 Stability Coefficient for Protein Signaling Data

Direct, Indirect, and Total Effects

Consider the following mediation example, protein pip3 has a direct influence on protein pip2 and the influence of pip3 on pip2 is mediated by protein plcg. This suggests that pip3 might directly or indirectly, through plcg, affect pip2.

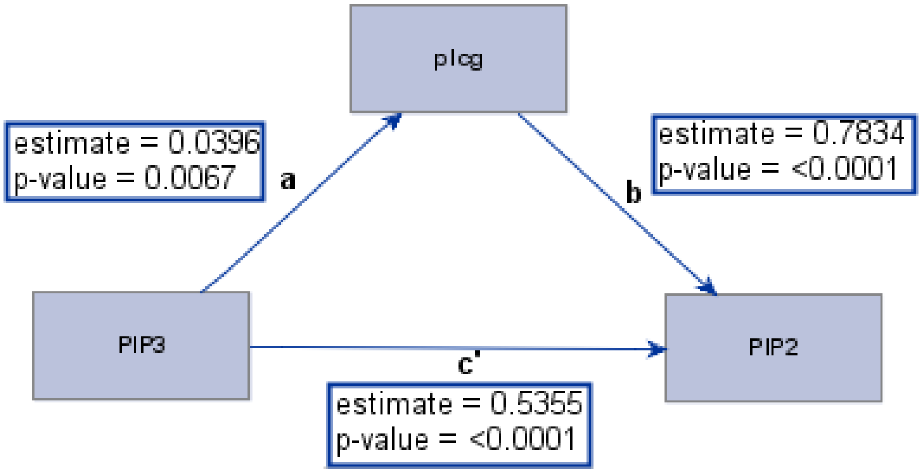


Figure 6 Mediation for '*Deletion Model*' in Protein Signaling Pathway

where, a = First component of the indirect effect of pip3 on pip2
b = Second component of the indirect effect of pip3 on pip2
c' = Direct effect of pip3 on pip2

The estimates a, b, and c' in Figure 8 are from the *Deletion Model* of the protein signaling example. The indirect effect of pip3 on pip2 is mediated by plcg and is estimated by the product of a and b (0.0396*0.7834) = 0.031 (p-value = .0068). The total effect of pip3 on pip2 is the sum of the indirect and direct effects; 0.031 + 0.5355 = 0.5665 (p-value < .0001). Because both the direct and indirect effects are significant, the mediation is partial. The interpretation is that a unit *change* in pip3 causes a 0.5665-unit change in pip2.

CATEGORICAL VARIABLES

Categorical variables are not supported by PROC CALIS unless an input covariance matrix is derived using polychoric or polyserial correlations. A polyserial correlation measures the correlation between a continuous variable and a categorical variable with a bivariate normal distribution. A polychoric correlation measures the correlation between any two categorical variables having bivariate normal distributions.

A simple SAS code to do polyserial correlation using PROC CORR:

```
proc corr data=sashelp.cars polyserial;
with type; /* Categorical Variable */
var weight horsepower; /* Continuous Variables */
run;
```

A simple SAS code to do polychoric correlation using PROC FREQ:

```
proc freq data=sashelp.cars;
tables make*origin/plcorr;
run;
```

PROC CALIS treats this input matrix as usual covariance matrix for continuous variables and estimates the coefficients for the parameters. This approach can be used with any estimation method. The standard errors may not be correct, but the parameter estimates are reasonably close.

NORMALITY

PROC CALIS requires that the data used for analysis follow a multivariate normal distribution. When the data are non-normal, parameter estimates are not affected, but standard errors are under estimated, and the probability of type 1 error is high, goodness of fit chi square is over estimated and other fit statistics may not be meaningful. Significant skewness and kurtosis might indicate that the data is not normal; therefore, multivariate measures of skewness and kurtosis are available in PROC CALIS.

MOORE-PENROSE INVERSE MATRIX

When fitting the protein signaling models using the data on the original scale, we encountered the following note and warning in the SAS log:

> *NOTE: The Moore-Penrose inverse is used in computing the covariance matrix for parameter estimates.*
>
> *WARNING: Standard errors and t values might not be accurate with the use of the Moore-Penrose inverse."*

The Moore-Penrose inverse is a pseudo inverse, which can be used to find an approximate solution that minimizes the error when a unique inverse cannot be found. Computed standard errors, t values and modification indices are likely to be approximate values. Although the Moore-Penrose provides a usable solution, interpret the results with caution.

We applied the natural log transformation to the raw data prior to creating the covariance matrix. Log transformations are a common method for handling skewed data. By doing so, we no longer had the computing issue with inverse matrix.

CONCLUSION

The purpose of this paper is to introduce the reader to interpret Structural Equation Models (SEMs) as Structural Causal Models (SCM); i.e. for causal relationships. To interpret an SEM as an SCM, you focus on the model structure, which is guided by subject matter experts. We described four conditions that a graphical SEM must meet to allow interpretation as a SCM.

Using the PATH modeling language within PROC CALIS, a flexible approach whose syntax is closely related to the path diagrams representations, we suggested a modeling process beginning with drawing a path diagram of a hypothesized model, fitting the model, and assessing the model's fit with the data. We also described the strategies used to improve overall model fit by using modification indices and understand the mediation effects in the model.

We provided two examples of the model fitting process, one from observational data and one from a controlled experiment. Only after model-fitting did we evaluate the conditions that must be met to declare cause-and-effect pathways or relationships. We reinforced the difficulty of making causal inferences and the importance of subject matter expertise.

While writing this paper, we learned about the CAUSALGRAPH Procedure introduced in SAS/STAT 15.1. This procedure provides capabilities to assess whether a cause-effect relationship within a graphical model meets the criteria to declare a causal relationship; i.e. is ‘identifiable’.

REFERENCES

Grozea, C. “Causal Discovery in Weblogs.” Accessed December 7, 2008. http://www.causality.inf.ethz.ch/repository.php?id=13

Kelley, Ken and Keke Lai. 2011. “Accuracy in Parameter Estimation for the Root Mean Square Error of Approximation: Sample Size Planning for Narrow Confidence Intervals.” Multivariate Behavioral Research, 46:1–32.

Sachs, K., Prez, O., Pe’er, D., Lauffenburger, D.A., Nolan, G.P. April 2005. “Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data”. Science, 308.

Thompson, Clay, 2019. “Causal Graph Analysis with the CAUSALGRAPH Procedure”. Proceedings of SAS Global Forum 2019, Paper SAS2998-2019. Dallas, Texas. SAS Institute, Cary NC.

Wang, F., Amrhein, J., 2018. “Bayesian Networks for Causal Analysis”. Proceedings of SAS Global Forum 2018, Paper 2776-2018. Denver, Colorado. SAS Institute, Cary NC.

ACKNOWLEDGMENTS

We thank Blue Neustifter at McDougall Scientific Ltd. for her support and advice.

RECOMMENDED READING

- *SAS/STAT® 15.1 User’s Guide, The CALIS Procedure, Overview: CALIS Procedure*
- Pearl, Judea. 2015. An Introduction to Causal Inference. Self-published.
- Pearl, J., Glymour, M., Jewell, N. 2016. Causal Inference in Statistics. A Primer. Wiley

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Banoo Madhanagopal
McDougall Scientific Ltd.
bmadhanagopal@mcdougallscientific.com
http://www.mcdougallscientific.com/

John Amrhein
McDougall Scientific Ltd.
jamrhein@mcdougallscientific.com
http://www.mcdougallscientific.com/