SC705: Advanced Statistics
Instructor: Natasha Sarkisian
Class notes: Introduction to Structural Equation Modeling (SEM)

SEM is a family of statistical techniques which builds upon multiple regression, and incorporates and integrates path analysis and factor analysis.

Advantages of SEM compared to multiple regression:

- more flexible assumptions (particularly helpful to deal with multicollinearity)
- use of confirmatory factor analysis to explicitly account for measurement error by having multiple indicators per latent variable
- graphical modeling interface (diagrams)
- ability to test models overall rather than coefficients individually
- ability to test models with multiple dependent variables
- ability to model mediating variables
- ability to test coefficients across multiple groups
- ability to handle difficult data (longitudinal with autocorrelated errors, non-normal data, incomplete data)

SEM simultaneously:
(a) models causal processes represented by a series of regression equations, and
(b) provides the ability to include unobserved (latent) variables and takes into account measurement error. In line with that, the structural equation modeling process centers around two steps:

1. Validating the measurement model -- accomplished through confirmatory factor analysis.
2. Fitting the structural model -- accomplished through path analysis with latent variables.

Sometimes, SEM can be used for only one of these two:

(a) SEM software can be used to estimate a model in which each variable has only one indicator – i.e., to conduct path analysis
(b) SEM software can be used to estimate a model in which each variable has multiple indicators (i.e., all variables are latent) but there are no direct effects (arrows) connecting the variables – i.e., to conduct confirmatory factor analysis

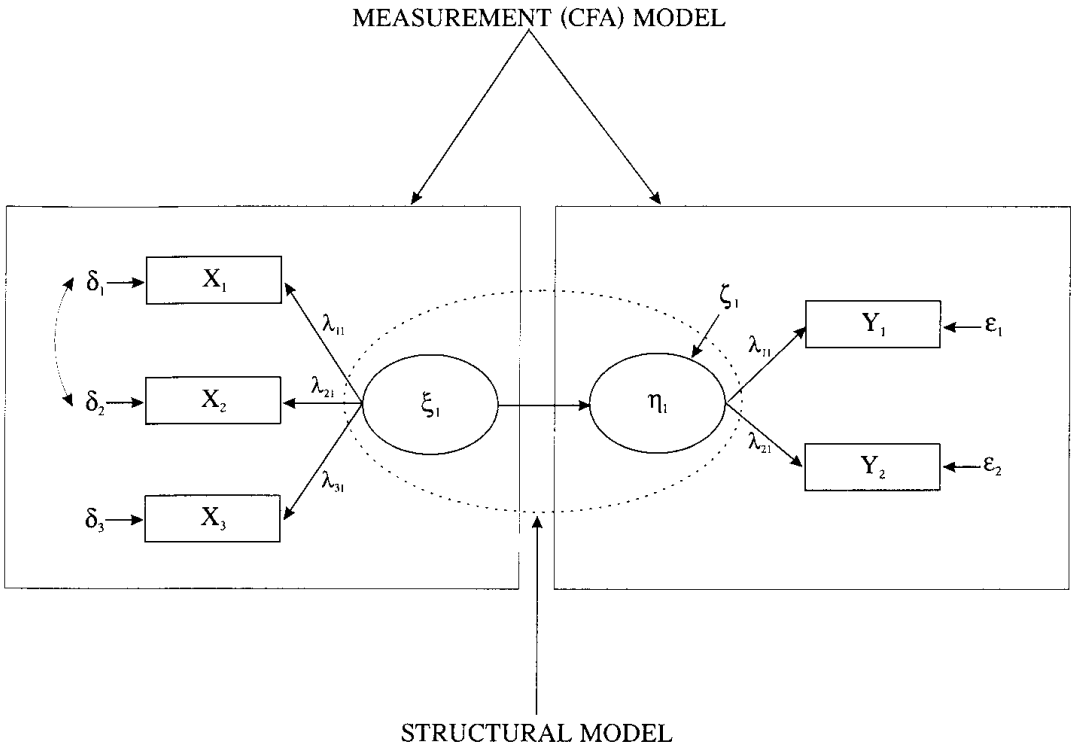
Usually, however, the term SEM refers to hybrid models with both multiple indicators for each latent variable (sometimes called factor), and directional paths specified connecting these latent variables.

1. Measurement model.

The measurement model is the part of an SEM model that deals with the latent variables and their indicators.

- Latent variables are the *unobserved variables* also called *constructs* or *factors* which are measured by their respective indicators.

- Indicators are *observed variables*, sometimes called *manifest variables* or *reference variables*, such as items in a survey instrument. Four or more indicators per latent variable is recommended, three is acceptable and common practice, two is problematic, and with one indicator, measurement error cannot be modeled (unless it is known prior to the analysis). Models using only two indicators per latent variable are more likely to fail to converge, and error estimates may be unreliable. Note: indicator variables cannot be combined arbitrarily to form latent variables. For instance, combining gender, race, or other demographic variables to form a latent variable called "background factors" would be improper because it would not represent any single underlying continuum of meaning.



A pure measurement model is a confirmatory factor analysis (CFA) model in there are straight arrows from the latent variables to their respective indicators, straight arrows from the error terms to their respective variables, but there are no direct effects (straight arrows) connecting the latent variables. In such a measurement model, we assume freely estimated covariance between each possible pair of latent variables, so we connect them with two-headed covariance arrows.

We start by specifying a model on the basis of theory. Each variable in the model is conceptualized as a latent one, even if we have to use a single indicator to measure it. Confirmatory factor analysis is used to verify that indicators seem to measure the corresponding latent variables. Note that we use common factor analysis (or principal axis factoring) rather than principal components analysis when conducting confirmatory factor analysis within SEM framework. This is important because common factor

analysis assumes a separation of variance into common variance and measurement error, while principal components analysis includes all of the variance into factors it creates.

The measurement model is evaluated like any other SEM model, using goodness of fit measures (we'll discuss them in detail later). We only proceed to the structural model when we confirmed that the measurement model is valid.

2. Structural model. This step involves fitting a structural model, or multiple models if we want to compare. We can evaluate these models in terms of "model fit," which measures the extent to which the covariances predicted by the model correspond to the observed covariances in the data. If the fit is not good, we can use modification indexes to alter the model and therefore to improve fit.

Two types of variables can be identified in a structural model.

- Exogenous variables are independent variables – i.e. variables with no prior causal variables determining them (though they are usually correlated with other exogenous variables -- depicted by a double-headed arrow -- unless there is strong theoretical reason not to do so).
- Endogenous variables are dependent variables in a broad sense – these can be pure dependent variables or mediating variables (variables which are both effects of other exogenous or mediating variables, and are causes of other mediating and dependent variables).

Therefore, the structural model includes a set of exogenous and endogenous variables in the model, together with the direct effects (straight arrows) connecting them, and the disturbance terms (residual variance) for endogenous variables.

Two broad types of structural models exist -- *recursive* and *nonrecursive* models. A structural model that specifies direction of cause from one direction only is termed a recursive model; one that allows for reciprocal or feedback effects is termed a nonrecursive model. We will mostly deal with recursive models in this course, but we will address the nonrecursive ones as one of more advanced topics.

SEM is usually viewed as a confirmatory rather than exploratory procedure, using one of three approaches:

1. *Strictly confirmatory approach*: A model is generated by theory (and prior research); it is then estimated and tested using SEM goodness-of-fit tests to determine if the pattern of variances and covariances of variances in the data is consistent with the theoretical model. Because other (unexamined) models may fit the data just as well or even better, declaring that the model fits does not mean we confirm that it's the correct model – it just means we can't disconfirm it.
2. *Alternative models approach*: One may test two or more causal models to determine which has the best fit. There are many goodness-of-fit measures, reflecting different considerations, and usually three or four are reported by the researcher. The problem here is that it is rare to be able to find in the literature

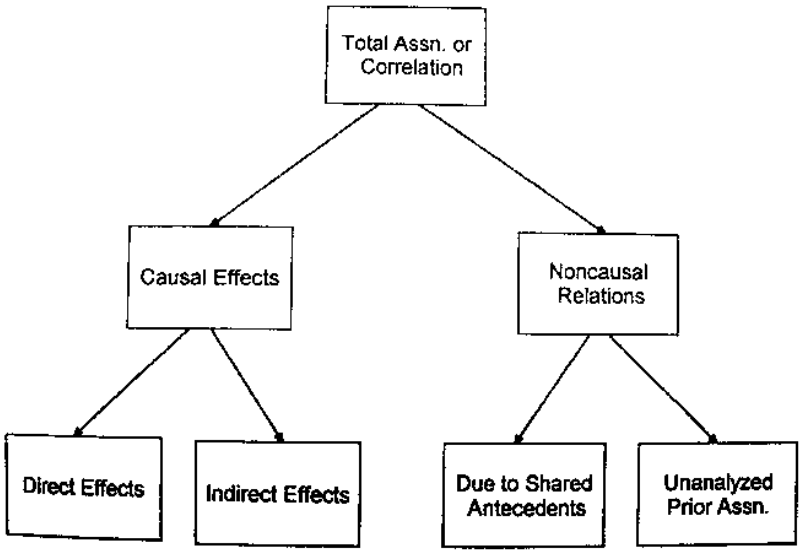
two well-developed alternative models to test – indeed, it's often not easy to find even one well-developed model.

3. *Model development approach*: In practice, much SEM research combines confirmatory and exploratory purposes: a model is tested using SEM procedures, found to be deficient, and an alternative model is then tested based on changes suggested by SEM modification indexes. This is the most common approach found in the literature. The problem with this approach is that models confirmed in this manner are post-hoc so they may be unstable (may not fit new data, as they were based on the uniqueness of the initial dataset). Researchers may attempt to overcome this problem by using a *cross-validation* strategy -- the model is developed using one sample and then confirmed using another sample (e.g., you can split the original sample in half).

For any of these approaches, theoretical insight is key to SEM! It is especially important to realize that causal directions in SEM are inferred from the theory rather than established from the data.

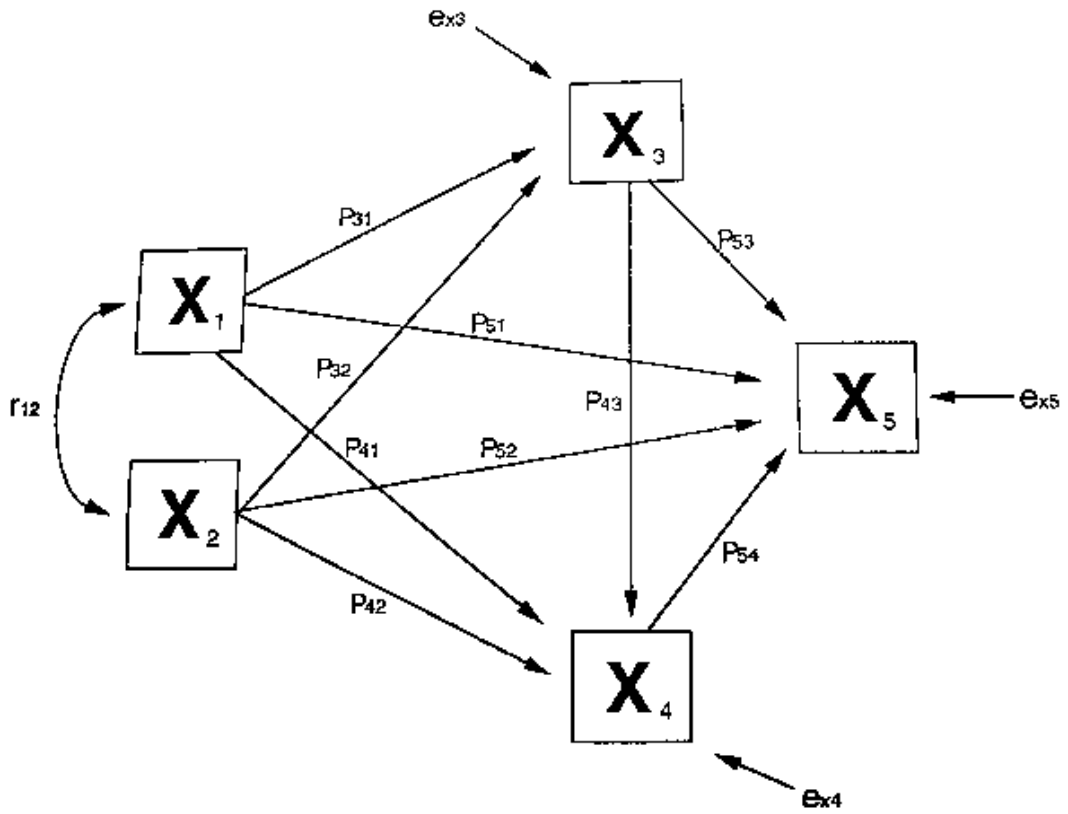
Path Analysis

SEM, and especially the structural model, is based on path analysis. Path analysis methods transform the variance-covariance (or correlation) matrix into a set of regression coefficients. Path analysis partitions the variance and performs the decomposition of effects:



(Diagram from Maruyama 1998, Basics of Structural Equation Modeling, p.37)

We are most interested in direct and indirect effects. Direct causal effects are represented as regression coefficients for the specified arrow. Indirect causal effects are represented as a sum of products of indirect paths. Unanalyzed prior associations are depicted with double-headed arrows.



(Diagram from Maruyama 1998, Basics of Structural Equation Modeling, p.38)

Path analysis does not require SEM software – it can be done using OLS. We would estimate three OLS models, and then take the resulting coefficients and fill them into the diagram.

Regression equations:
$X_3 = p_{31}*X_1 + p_{32}*X_2 + e_{X3}$
$X_4 = p_{41}*X_1 + p_{42}*X_2 + P_{43}*X_3 + e_{X4}$
$X_5 = p_{51}*X_1 + p_{52}*X_2 + p_{53}*X_3 + p_{54}*X_4 + e_{X5}$

Direct effect of X_1 on X_5: p_{51}
Indirect effect of X_1 on X_5: $p_{31}*p_{53} + p_{41}*p_{54}$
Total effect of X_1 on X_5: $p_{51} + p_{31}*p_{53} + p_{41}*p_{54}$
Unanalyzed prior association between X_1 and X_2: r_{12}

Unanalyzed prior association is modeled by including variables in regression models simultaneously; that is, all the coefficients used in the path diagram are partial regression coefficients – regression coefficients for X → Y while controlling for all other predictors of Y specified in the diagram.

The analysis of indirect effects is often called mediation analysis. Mediation is a causal sequence in which one variable (X) affects a second variable (Z) that, in turn, affects a

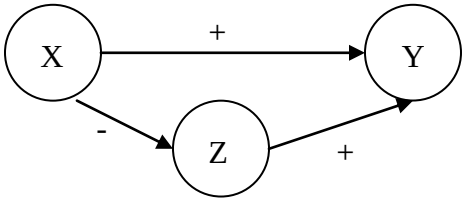
third variable (Y). The intervening variable, Z, is the mediator; it mediates the relationship between a predictor, X, and an outcome, Y.

We can distinguish full and partial mediation. Full mediation occurs when, if mediator Z is in the model, there is no longer a relationship between X and Y. Partial mediation occurs when the relationship between X and Y is reduced in size when Z is included, but does not disappear entirely. Full mediation rarely happens, although if we have multiple mediators, it can be possible to fully "explain away" the link between X and Y.

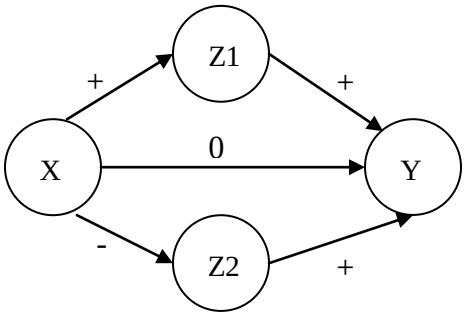
Suppressors

It is also possible to have significant indirect effects even if there was no significant direct effect without any mediators. This type of mediation involves mediators that are also called suppressors. In general, suppressors are variables that increase the effect of another variable (or a set of variables) if we include them in a regression equation. Suppressors can be either included in the mediation process, or they can be confounding factors (we'll discuss those later).

For example, once a suppressor variable is in the model, we can observe a negative indirect effect and a positive direct effect, or vice versa. If Z is omitted, however, X → Y relationship might be null; it only appears when we control for Z. For example:



Different mediators also can potentially "cancel out" each other's effects, so that the total effect (direct effect + all indirect effects) is zero; for example:

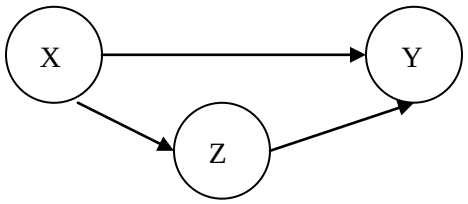


It is also possible for a mediator to change not only the direct effect, but also the indirect effects of other mediators. Thus, when examining mediation, you should make sure to avoid the omitted variable bias (make sure that all theoretically relevant variables are included).

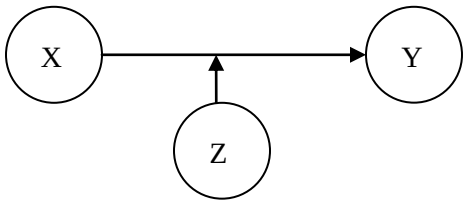
Mediation versus moderation

We have to be careful to distinguish mediation from moderation. Moderation always involves interaction terms.

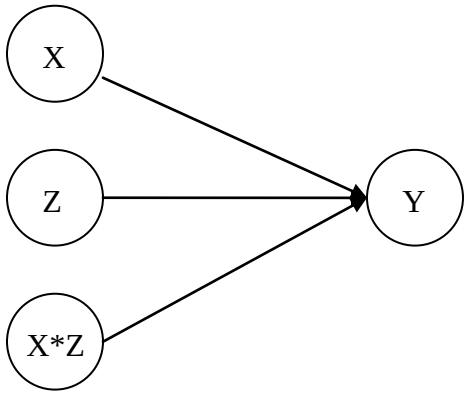
Mediation (Z is a mediator):



Moderation (Z is a moderator):



Or if depicted with an interaction term:

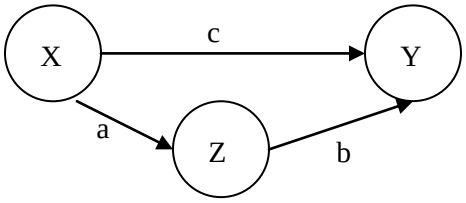


Establishing mediation

To fully establish mediation, we have to follow these steps identified by Baron and Kenny (1986):
(1) ensure that X → Y (run a model with Y as an outcome and X as a predictor: Y = const + c1*X + e)
(2) establish that X → Z (run a model with Z as an outcome and X as a predictor: Z = const + a*X + e
(3) establish that Z → Y while controlling for X (run a model with Y as an outcome and Z and X as predictors: Y = const + c*X + b*Z + e)
(4) observe that when we introduce Z (in the model specified in step 3: Y = const + c*X + b*Z + e), the relationship between X and Y is either entirely non-existent (full mediation, where c is zero) or diminished (partial mediation, where both b and c are non-zero but c is substantially reduced). That is, in full mediation, X →Z → Y and there is no direct path between X and Y – only the indirect one, calculated as a*b.

Some suggest that step 1 of this process can be skipped; and indeed, there can be indirect effects of X even in the absence of X → Y direct link. Although this is not mediation in a classic sense, it is still essentially a mediation analysis as it examines indirect effects.

Ultimately, we should examine X→ Y bivariate link to know what it is, but it doesn't have to be significant for us to proceed.



While it is easy to calculate the indirect effect itself, its statistical significance is more difficult to establish. Originally, to calculate the statistical significance of the indirect path, a*b, Sobel test was used. It relies on calculating the standard error of the product a*b using Sobel's formula:

$s_{a*b} = \sqrt{b^2 s_a{}^2 + a^2 s_b{}^2}$, where a and b are the unstandardized regression coefficients and s_a and s_b are their standard errors. Then the t-test statistic is calculated as

$$t = \frac{a*b}{s_{a*b}}$$

There are also more complex ways to calculate that standard error that also include the produce of two variances (either add it or subtract it); those versions were proposed by Goodman and Aroyan. The results are typically similar, and simulations seem to suggest that Sobel and Aroyan versions perform better than the Goodman version. You can also get the significance test based on all three formulas using an online Sobel calculator: http://quantpsy.org/sobel/sobel.htm . You need to know a, b, s_a, and s_b to use it.

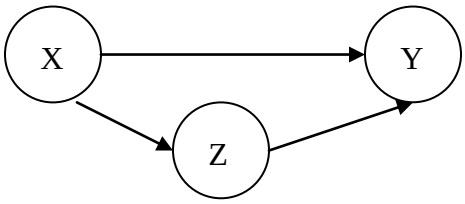
Whichever formula you use, however, this approach assumes the distribution of conditional indirect effects is normal, but it is usually non-normal. Therefore, significance testing that relies on the assumption of normality is no longer recommended for final models with indirect effects, unless you have a very large sample. The second approach that is more robust is to use bootstrapping to obtain standard errors and confidence intervals. Bootstrapping involves drawing many random samples (with replacement) from one's current sample and calculating coefficients of interest for each of them, which allows us to construct a distribution of those estimates and determine standard errors and confidence intervals based on that distribution. There are macros written for both SPSS and SAS to bootstrap standard errors and confidence intervals for mediation models; see the links to macros here: http://quantpsy.org/sobel/sobel.htm or here http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html

In Stata, there are user-written commands that allow you to calculate standard errors and create confidence intervals using bootstrapping: see sgmediation command (available from http://www.ats.ucla.edu/stat/stata/ado/analysis) and medeff command (st0243_1 from http://www.stata-journal.com/software/sj12-2).

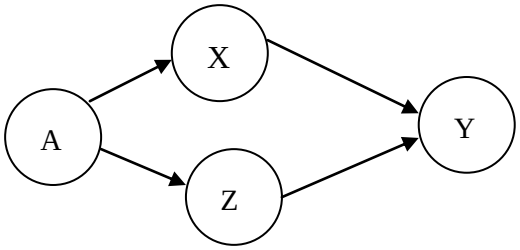
Mediation and causality

In general, when conducting mediation analysis, we make claims about causal relationships among variables. To make such claims more appropriate and more convincing, we need to make sure that we:

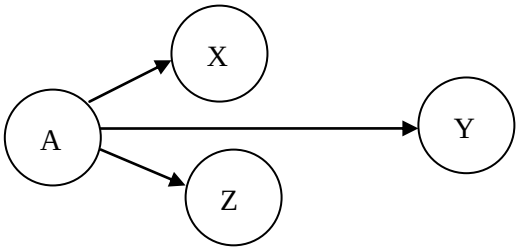
1. Rule out potential spurious relationships and confounding variables. For example, we should attempt to identify potential "third variables" – confounders -- that influence some of the variables involved in our proposed mediation model and change the effects we are interested in. For instance, we need to make sure that if we think our model is this:



That there is no such A variable that the true causal sequence is in fact this:



or even this:



Thus, if we can think of such potential "third" variables, we might need to include them to empirically rule out such possibilities by demonstrating that even if we control for potential confounders, the relationships of mediation are still there. Importantly, confounding variables may either "explain away" some paths in our mediation model and make them non-significant, or they can make other coefficients larger and more prominent (in this case, confounding variables serve as suppressors).

2. When constructing a mediation model, we should be really sure of causal directions based on our theory; for example, it should be really clear to us that it is X → Z rather

than Z → X, etc. If we cannot make distinct causal claims based on theory, then mediation analysis is not appropriate.

3. Our causal claims will be stronger if we can establish temporal precedence. That is, if X precedes Y in time, this would provide additional evidence that the link between them is directional, and potentially causal.

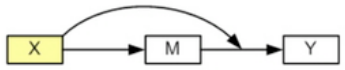
4. It is important to be sure that independent variable X and mediator Z do not interact; otherwise, we would need to examine moderation (or moderated mediation).

Moderated mediation

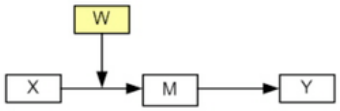
Moderated mediation takes place if a moderator variable interacts with a mediator variable which means that the size of the indirect effect depends on the value of the moderator. There is a Stata-based website that provides an excellent guide to how to estimate and test different types of moderated mediation models using regression analysis in Stata: http://www.ats.ucla.edu/stat/stata/faq/modmed.htm

The types of models addressed are presented in this diagram.

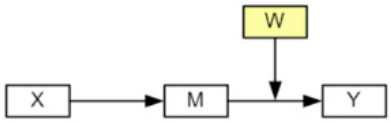
Model 1



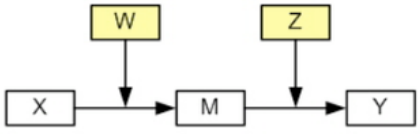
Model 2



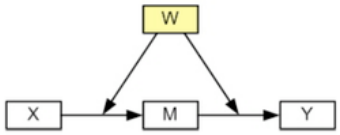
Model 3



Model 4



Model 5



(Diagram from http://www.ats.ucla.edu/stat/stata/faq/modmed.htm)