REN R 690 - Path Analysis

My variables are correlated so a linear regression won't work...Now what?
A crash course on Path Analysis

Vocabulary:

In this handout, we are assuming that the reader has some familiarity with Structural Equation Modeling vocabulary. However, if this is not the case, we highly recommend flipping to the back of this handout and first reading the "Background information" before reading anything else.

Some Theory:

Path Analyses are:

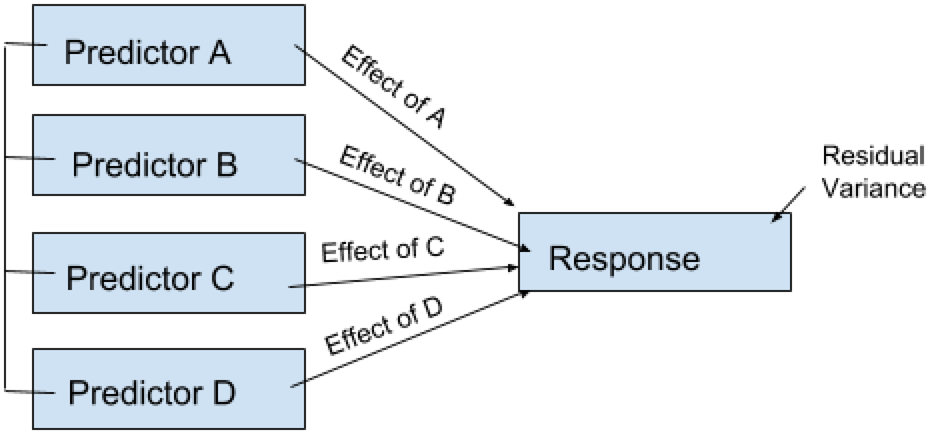
- One of the tools which falls under the Structural Equation Modelling umbrella.
- Also known as "simultaneous equations" or "structural equations with observed variables".
- Analyses which includes manifest variables with mediated mechanisms. In other words, the path analysis application of structural equation modeling has only observed values, and tries to establish causal relationships between those observed variables.
- Depicted by SEM path diagrams: input hypothesized relationships and output "actual" relationships (following statistical analyses) can both be presented by SEM path diagrams.

Limitations of Path Analysis: Assumptions

1. Normality of residuals is necessary (as the null hypothesis assumes normal distribution)
2. No loops between variables
3. All direct influences between specified variables are specified in the formal hypothesis
4. The directional nature of the relationship between variables is constant for all components within each variable.
5. Obviously, you want more observations than parameters."More rows than columns" as Andreas says.
6. All variables are measured without error (this becomes tricky with social studies for example, if you're trying to measure stress levels of an individual). If this is the case, you will have to look into using a latent factor in an more complex SEM model.
7. While path analysis is useful for evaluating causal hypotheses, this method cannot determine the direction of causality. It clarifies correlation and indicates the strength of a causal hypothesis, but does not prove direction of causation.

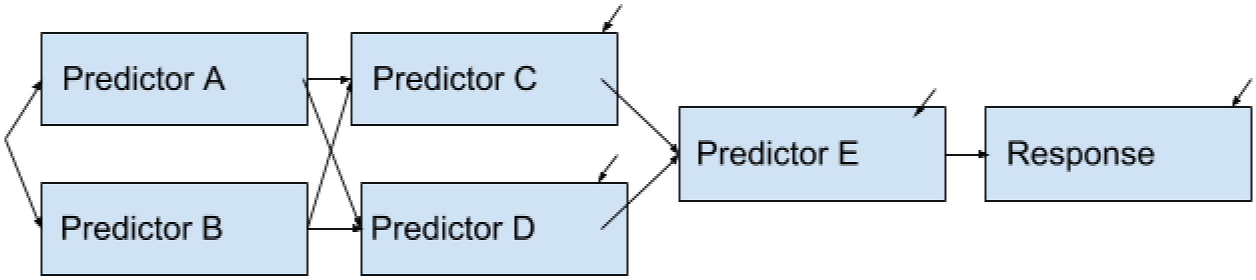
Explaining path analysis as a buildup from multiple regression.

(1) What are the effects of my predictor variables on my dependant variables?: multiple regression question



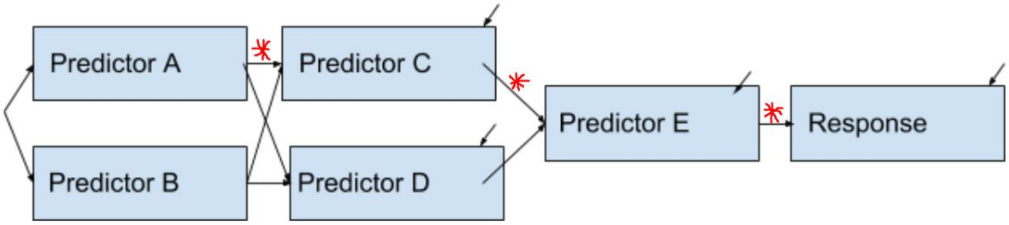
- Note:
  - In this form, the multiple regression is “saturated”. All variables are associated to all other variables in the model, and there are no restrictions to relationships: All the predictors correlate with one another and the dependant variable is regressed on all the predictors.
- Information we can get from this model:
  - Joint contribution of predictors: How much of the variance in the dependant variable is explained by the set of all the predictor variables
  - Unique predictability: How much of the variance in the dependant variable is explained by one of the predictor variables, while controlling for all the others.

(2) Is the effect of one of the predictor variables on the dependant variable, mediated by other predictor variables?: simple path analysis question



- Note:
  - Compared to question (1), we now have 4 dependant variables (Predictors C, D, and E + Response)
  - There is an equation for every dependant variable
  - There are *structured* relationships, where:
    - A leads to C which leads to E which leads to Response (we would say that A has an “indirect” effect on Response)

    - A does not lead directly to E, nor does A lead directly to Response (if we did hypothesize that A also leads directly to Response, there would be a single headed arrow from A to response and this would called a "direct" effect)
    - This structure is advantageous because it gives us a formal, testable hypothesis: Here we are saying that our hypothesis lays out this particular structure, where we allow the influences to go through certain pathways, but not in others. To use SEM modelling language, we have now *specified* the model.
- Once we have specified the model, we can *evaluate* the model: Does my hypothesized model fit the data well? Can I use it to draw inferences in other systems?
- If we find the model does not fit well, and we do not believe we have identified the underlying mediating mechanism leading from A to Response, then we can do some *respecification* to try and make it fit better.
- Once we have an well fitting model,
    - We may stop and conclude that we have a more subtly nuanced understanding of how the factors relate to each other over the multiple regression model
    - Or, we may do a formal test of mediation: If we have significant links in a mediating chan, that is not enough to infer that the entire indirect effect is significant. We have to test if that entire "chain" as a unit is significantly than 0. This is sometimes called a "specific indirect effect" (an indirect effect of A on Response that goes through C and E). We take the product of the individual coefficients (stars) so that we can do a z-test.



(3) More complex Path Analyses models can include:

- More predictor variables
- More mediator variables
- More response variables/ outcomes
- Multiple response groups (ex: how would the outcome differ if my species of interest is a lodgepole pine vs. a white spruce?)
- Moderation (ex1: how different if the effect of Predictor C if my species of interest of lodgepole pine vs. white spruce? Ex2: how does the relationship between depression and physical health scores change if Poverty status (yes/no) is included ?)

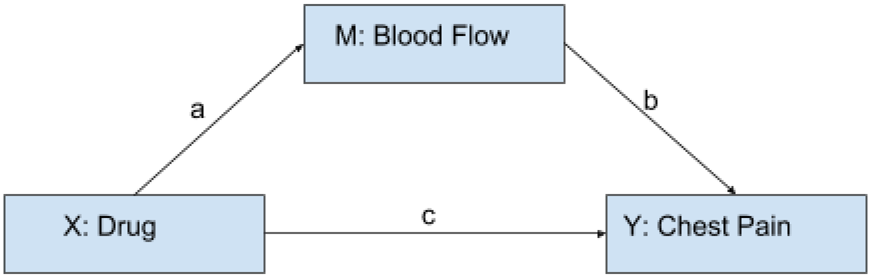
Example dataset: `Optional Lab: Path Analysis by J.G.and M. de K.`

```
install.packages("lavaan", dependencies=TRUE)
library(lavaan)
```

Lavaan step 1= SEM Step 1: Specification

NB: for the purpose of this lab, we are only showing a very simple "mediation" path analysis. This dataset comes from Rossel (2012) section 9.2.5, with re-named variables from Jordan C YouTube tutorial, to facilitate interpretation. For R coding of path analysis with more complex datasets, refer to Rossel(2012)

In path analysis, the way that you estimate the parameters is based on the results of a series of regressions; the number of regression equations you need to estimate equals the number of arrows that are pointing to the response variable. So for a situation with 1X, 1M, and 1Y with arrows going from X to M, M to Y, and X to Y, we need three regression equations to get the parameters for the model



```
DRUG <- rnorm(100) #predictor X--> GTN Drug (relieves chest pain)#
X=DRUG

FLOW <- 0.5*X + rnorm(100) #moderator M --> Blood flow#
M=FLOW

PAIN <- 0.7*M + rnorm(100) #response Y --> Angina (chest pain)#
Y=PAIN

DATA <- data.frame(X = DRUG, Y = PAIN, M = FLOW)

model <- 'Y ~ c*X #direct effect: DRUG has direct effect on Pain#
          M ~ a*X #mediator DRUG has on effect on FLOW #
          Y ~ b*M #mediator: FLOW has an effect on PAIN#
          indirect := a*b #indirect effect#
          direct :=c
          total:= c +(a*b)'#total effect = sum of direct and
indirect#
```

In Lavaan,

- <- means "is defined by",
- you specify the model between two single quotes,
- ~ indicates a regression,
- := ‘defines’ new parameters which take on values that are an arbitrary function of the original model parameters. In our example, these are represented by the a and b, which are calculated but are not in the original dataset

<u>Lavaan step 2 = SEM Step 3: Estimation</u>

“fit” is the name of the object that will hold the results of the sem model, where the syntax is: sem(*name of the model* , *which data set to use*)

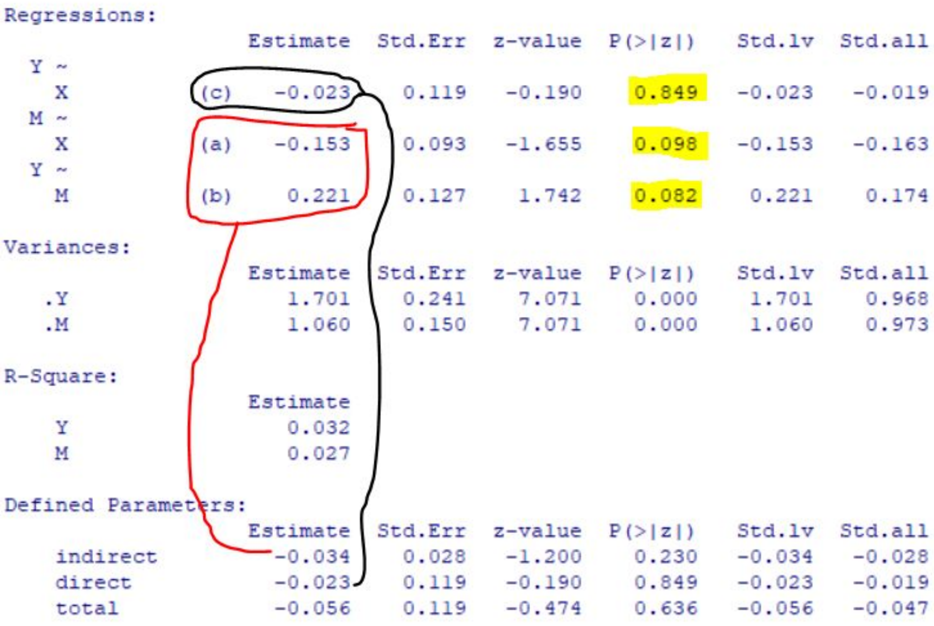
```
fit <- sem(model, data=DATA)
summary(fit)
```

Here we get estimates of the path as well as results from significance tests. By default, Lavaan does not give all the results it has stored. Ex: if we wanted standardized coefficients, fit indices, and R2 values, the `summary(fit)` code would look like this instead:

```
summary(fit, standardized=T, fit.measures=T, rsq=T)
```

<u>Lavaan step 3 = SEM Step 4: Evaluation</u>

This is where you view results and decide if you need to restructure. First, we would note that the **product** of a and b regression estimates do, in fact, equal the estimate of the indirect defined parameter (red circle leading to -0.034) and that the regression estimate c is, in fact, the same as the direct defined parameter (black circle leading to -0.023).



```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  Y ~
    X         (c)    -0.023    0.119   -0.190    0.849   -0.023   -0.019
  M ~
    X         (a)    -0.153    0.093   -1.655    0.098   -0.153   -0.163
  Y ~
    M         (b)     0.221    0.127    1.742    0.082    0.221    0.174

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
   .Y                 1.701    0.241    7.071    0.000    1.701    0.968
   .M                 1.060    0.150    7.071    0.000    1.060    0.973

R-Square:
                   Estimate
    Y                 0.032
    M                 0.027

Defined Parameters:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
    indirect         -0.034    0.028   -1.200    0.230   -0.034   -0.028
    direct           -0.023    0.119   -0.190    0.849   -0.023   -0.019
    total            -0.056    0.119   -0.474    0.636   -0.056   -0.047
```

Next, based on this output, we would say (Yellow highlights) that X (Drug) is neither a significant indicator of Y (Pain), nor of M (Blood flow). Also, Y is not significantly indicated by M, so Blood flow does not seem to be related to chest pain in these patients.

Visualization with semPlot (leading to to SEM Step 6 of interpretation)
Once we have our results from Lavaan, we can use another package, semPlot, to visualize these relationships. In keeping with the extreme complexity and potential flexibility of SEM models themselves, it is beyond the scope of this short tutorial to cover all potential nuances of this package. For more information, see “semplot resources” below.

```
install.packages("semPlot") # First install and load the “semPlot” package. #
library(semPlot)
```

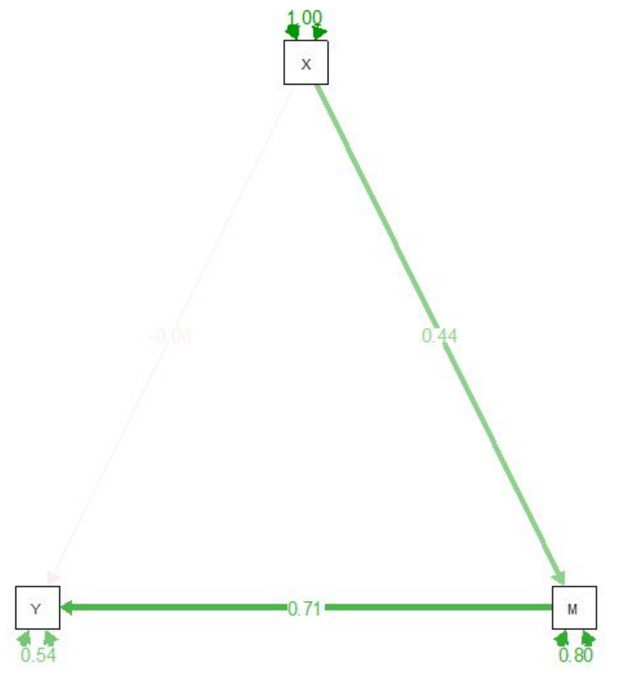
# Now we will generate a simple visual summary of our “fit” model created in step 3. #

```
semPaths(fit)
```

Without getting too fancy, we can add some basic information about our model using the “standardized parameters” tool. This places the standardized path coefficient for each component of the defined relationships on the graph, and colour- and size-codes the lines. Note also in the following command that the “edge.label.cex” allows for adjustment of the label size.

```
semPaths(fit, "std",  edge.label.cex = 1)
```

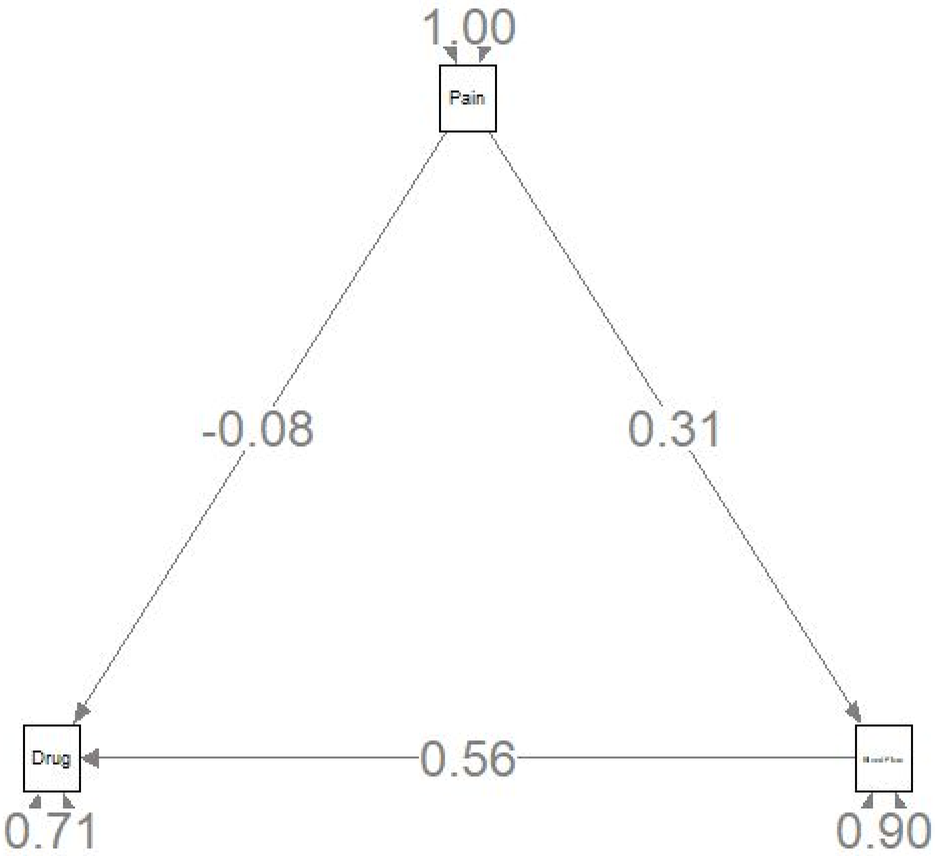
- Thicker lines= larger path coefficients
- Greener lines =positive path coefficients
- Red lines = negative path coefficients.
- Values showing on arrows between two variables represent the standardized path coefficient for the regressions between those variables
- Values on the circular arrows unique to each variable show the standardized chi-square value for variances of each variable alone



To further customize the figure above:

- Keep paths black and of uniform size using “model”;
- increase the size of the path labels using “edge.label.cex”,
- increase the category labels using “label.cex”,
- and override the variable labels using “nodeLabels”.

```
semPaths(fit, "model", "std",  edge.label.cex = 2, label.cex = 1,
nodeLabels = c("Drug", "Blood Flow", "Pain"))
```



Appendix: Background Information

Structural Equation Model (SEM) It would be neglectful to explain path analysis without providing some (VERY SIMPLE) explanation of SEM: SEM is not a technique per se, but a collection of techniques which can be used together. Structural Equation Modeling (SEM) is an umbrella term that captures other models you've already seen, and extends them to be combined in unique ways. This "umbrella" includes for example: General Linear Model (GLM), T-test, correlation, regression,and analysis of variance. Structural Equation Models helps to explore the mediating mechanisms between the predictor variables, and the outcome. This starts to lead to an explanatory "chain of events" which can explain the outcome, as a result of the predictor(s).

SEM models can includes both latent and manifest variables, and can be used to explain very complex systems.

- Predictor Variable Independent Variable. Also known as exogeneous variables in SEM
- Outcome Variable Result/ Dependant Variable. Also known as endogenous variables in SEM
- Latent variables Non-observed variables (in the SEM models, these are drawn as circles)
- Manifest variables Observed or measured variables. May also be known as indicator (in the SEM models, these are drawn as squares or rectangles)
- Mediator variables Variables which explain the relation between a predictor and an outcome...this indicates a specific causal pathway. It occurs when at least part of the reason X affects Y is *through* Z : X affects Z and Z affects Y.
- Moderator variables indicates that the effect of X on Y is different for different values of Z. In other words, Z moderates (affects) the effect of X on Y.

There are four main Structural Equation Models:

1. Path Analysis is an application of SEM without latent variables. Parameters: regression coefficients and residual variances
2. Confirmatory Factor Analysis a.k.a. Measurement models are used when you estimate the paths that link each manifest variable to their corresponding latent variable. Parameters: item residuals, factor loadings, and residual variances
3. Latent Variable Structural Models are made when latent variables are introduced within the path analysis framework. Parameters: regression coefficients, item residuals, factor loadings, covariances, and residual variances
4. Growth Curve Models a.k.a. Longitudinal models are an application of SEM than can be done if you have multiple observations of the same variable *over time.*

The steps (Bollen 1989) to all SEM models are as follows:

(1) Specification: What variables do you have in your dataset, and how do you relate those to one another? (What are the equations?; what do we believe the independent

variables are; what are the dependant variables; the mediators; the moderators; the latent; the manifest.)

(2) Identification: tricky and not often done in practice because it gets complex, but basically, this is: Did you have enough observed information in your data, to estimate the parameters in the model?

(3) Estimation: estimate sample parameters based on the characteristics of the data. (Maximum likelihood, least squares, asymptotic variation, ect…)

(4) Evaluation: how do we determine if the model fits the data? We want to determining if the model is appropriate enough to draw inferences from. N.B.: There are <u>multiple</u> ways to do this, and no one set value as a cutoff of “acceptable”. For a step-by-step tutorial on how to do this, we highly recommend “SEM Episode 5: Evaluating Model Fit” (YouTube).

(5) Respecification: Let’s say that we evaluated our model to be inappropriate to use to draw inferences. This means that the specifications we made in (step a) do not fit well. Thus, our hypothesized model does not fit well. To decide whether to add or remove variables, we can:

   (a) Turn back to the theory and add parameters in an a-priori way and conduct a formal likelihood ratio test. This makes the model less parsimonious; we must evaluate if the addition of the extra parameters significantly improves the fit of our model to the data.

   (b) Next, if option a did not significantly improve the fit of the model to the data, we would remove those parameters we had added in, and turn to the characteristics of the data and see what IT suggests. Note: This is very powerful and very cool, but this is NOT hypothesis testing: we may be adding chance parameters which decrease the replicability of our model in future studies. However, if you’re still interested, this respecification method is based on modification indices.

(6) Interpretation: Is tricky. Researchers may be tempted to make inferences beyond what the data allows.

Path diagrams for SEM are used to convey broader aspects of the model.

- <u>Rectangles or Squares</u>: represent variables we observe directly, also called manifest variables.
- <u>Circles</u>: represent variables we do not observe, also called latent variables
- <u>Single-headed arrows between two variables</u>: denote regression coefficients (regression of the dependant variable in the independent variable). Also sometimes called a directed effect
- <u>Small single-headed arrow pointing into the dependant variable</u>: represents residual variance. This is sometimes called the disturbance in the equation. This is the part of the dependant variable not explained by the predictors.
- <u>Double headed curved arrow:</u> the two predictor variables are correlated to each other
- <u>Lines drawn between multiple predictor variables</u>: represent that all the predictor variables along that line covaries with everything else

SEM and Lavaan Resources:

- The Analysis Factor website:
    - The four models you meet in Structural Equation modeling by M.R. Escobar https://www.theanalysisfactor.com/four-types-sem/
    - Five common relationships among three variables in a statistical model by K. Grace-Martin http://www.theanalysisfactor.com/five-common-relationships-among-three-variables-in-a-statistical-model/
- University of Exeter course handout:
    - "Topic 3: Path Analysis" http://people.exeter.ac.uk/SEGLea/multvar2/pathanal.html
- YouTube Series by Curran-Bauer Analytics, "Office Hours" :
    - SEM Episode 1: Introduction to Structural Equation Models https://www.youtube.com/watch?v=wHFrgp3SQMI&t=7s
    - SEM Episode 2: Path Analysis https://www.youtube.com/watch?v=QkXTdTCF3PI&t=346s
    - SEM Episode 5: Evaluating Model Fit https://www.youtube.com/watch?v=gFXIRoj49OI&t=330s
- YouTube video by Jordan C :
    - R Tutorial: Path Analysis and Mediation using Lavaan https://www.youtube.com/watch?v=-B37sK9NTfI

semPlot Resources:

- The author of the package provides some simple examples of commonly used semPlot analyses.
    - http://sachaepskamp.com/semPlot/examples
- A quick overview of semPlot, useful if you're setting up more complex relationships including latent variables:
    - https://biologyforfun.wordpress.com/2014/08/10/ploting-sems-in-r-using-semplot/
- Official semPlot manual- pretty dense to begin with, but useful once you're familiar with the basics and would like to start customizing:
    - https://cran.r-project.org/web/packages/semPlot/semPlot.pdf

Resources for more formal reading:

- Kenneth Bollen 1989 original book on SEM "Structural equations with latent variables"
- Yves Rossel 2012 document for running SEM in R "Lavaan: an R package for structural equation modeling and more Version 0.5-12 (BETA)" http://users.ugent.be/~yrosseel/lavaan/lavaanIntroduction.pdf
- James B. Grace, "Structural Equation Modeling and Natural Systems" - super helpful complete overview of SEM with a bend towards ecological analysis.