Multivariate Statistics and Methodology with R

Path analysis

Aja Murray; Aja.Murray@ed.ac.uk

This week

- Techniques
  - *Path analysis (esp. path mediation models)*
- Functions
  - *sem( ) from lavaan*
- Reading
  - *lavaan tutorial: http://lavaan.ugent.be/tutorial/tutorial.pdf (section 13)*

Learning outcomes

- Know how to specify, estimate, and interpret path analysis models in R
- Have a sense of the range of different models that can be fit using path analysis
- Know how to test, interpret and report path mediation models in particular

What is path analysis?

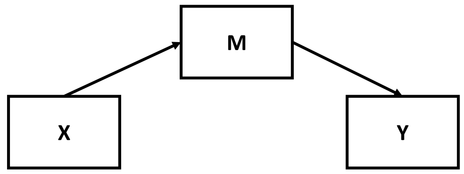
- Links several regression models together
- Tests the set of regression models as a whole
- Useful for situations where there are multiple outcome variables in sequence or parallel
- Models the relations between observed variables (i.e., does not involve latent variables)
- Common example: **path mediation model**

Mediation

- Is when a predictor X, has an effect on an outcome Y, via a mediating variable M
- The mediator **transmits** the effect of X to Y
- Examples of mediation hypotheses:
  - *Conscientiousness (X) affects health (Y) via health behaviours (M)*
  - *Conduct problems (X) increase the risk of depression (Y) via peer problems (M)*
  - *Attitudes to smoking (X) predict intentions to smoke (M) which in turn predicts smoking behaviour (Y)*
  - *An intervention (X) to reduce youth crime (Y) works by increasing youth self-contol (M)*

Visualising a mediation model

- In a SEM diagram we can represent mediation as:



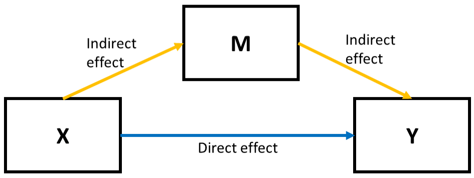
Mediation… not to be confused with moderation

- Mediation is commonly confused with **moderation**
- Moderation is when a moderator z modifies the effect of X on Y
  - *e.g., the effect of X on Y is stronger at higher levels of Z*
- Also known as an **interaction** between X and Z
- Examples of moderation could be:
  - *An intervention (X) works better to reduce bullying (Y) at older ages (Z) of school pupil*
  - *The relation between stress (X) and depression (Y) is lower for those scoring higher on spirituality (Z)*

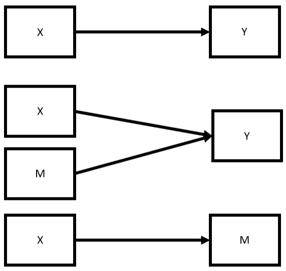
Direct and indirect effects in mediation

- We seldom hypothesise that a mediator completely explains the relation between X and Y
- More commonly, we expect both **indirect effects** and **direct effects** of X on Y
  - *The indirect effects of X on Y are those transmitted via the mediator*
  - *The direct effect of X on Y is the remaining effect of X on Y*

Visualing direct and indirect effects in mediation



Testing mediation



- Traditionally, mediation was tested using a series of separate regression models:

    1. *Y~X*
    2. *Y~X+M*
    3. *M~X*

Traditional methods of testing mediation

- The three regression models:

  1. *Y~X*
  2. *Y~X+M*
  3. *M~X*

- Model 1 estimates the overall effect of X on Y
- Model 2 estimates the partial effects of X and M on Y
- Model 3 estimates the effect of X on M
- If the following conditions were met, mediation was assumed to hold:
  - *The effect of X on Y (eq.1) is significant*
  - *The effect of M on x (eq.3) is significant*
  - *The effect of X on Y becomes reduced when M is added into the model (eq.2)*

Limitations of traditional methods of testing mediation

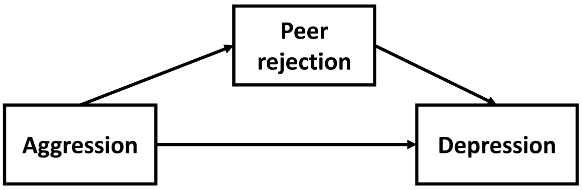
- Low power
- Very cumbersome for multiple mediators, predictors, or outcomes
- You don’t get an estimate of the magnitude of the indirect effect
- Much better way: **path mediation model**

Testing a path mediation model in lavaan

- Specification
  - *Create a lavaan syntax object*
- Estimation
  - *Estimate the model using e.g., maximum likelihood estimation*
- Evaluation/interpretation
  - *Inspect the model to judge how good it is*
  - *Interpret the parameter estimates*

Example

- Does peer rejection mediate the association between aggression and depression?



The data

```r
library(psych)
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:lavaan':
##
##     cor2cov
```

```r
describe(agg.data2)
```

```
##     vars   n  mean   sd median trimmed  mad   min  max range skew kurtosis   se
## Dep    1 500 -0.07 1.08  -0.09   -0.08 1.15 -2.94 2.67  5.61 0.08    -0.33 0.05
## PR     2 500  0.02 1.04   0.05    0.01 1.01 -2.69 3.12  5.81 0.07    -0.08 0.05
## Agg    3 500 -0.02 0.98   0.01   -0.03 0.98 -3.16 2.70  5.86 0.02    -0.07 0.04
```

```r
#PR = peer rejection, Agg= aggression, Dep= depression
```

Mediation example

- Does peer rejection mediate the association between aggression and depression?

```
#Create the model syntax

model1<-'Dep~PR      # Depression predicted by peer rejection
        Dep~Agg     # Depression predicted by aggression (the direct effect)
        PR~Agg      # Peer rejection predicted by aggression'


#estimate the model

model1.est<-sem(model1, data=agg.data2)
```

The model output

```r
summary(model1.est, fit.measures=T)
```

```
## lavaan 0.6-5 ended normally after 13 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of free parameters                          5
##
##   Number of observations                           500
##
## Model Test User Model:
##
##   Test statistic                                 0.000
##   Degrees of freedom                                 0
##
## Model Test Baseline Model:
##
##   Test statistic                               253.745
##   Degrees of freedom                                 3
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    1.000
##   Tucker-Lewis Index (TLI)                       1.000
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)              -1347.036
##   Loglikelihood unrestricted model (H1)      -1347.036
##
##   Akaike (AIC)                                2704.073
##   Bayesian (BIC)                              2725.146
##   Sample-size adjusted Bayesian (BIC)         2709.276
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.000
##   90 Percent confidence interval - lower         0.000
##   90 Percent confidence interval - upper         0.000
##   P-value RMSEA <= 0.05                             NA
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.000
##
## Parameter Estimates:
##
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##   Standard errors                             Standard
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   Dep ~
##     PR                0.289    0.048    6.009    0.000
##     Agg               0.256    0.051    5.033    0.000
##   PR ~
##     Agg               0.530    0.041   12.932    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
```

```
##    .Dep                 0.932    0.059   15.811    0.000
##    .PR                  0.805    0.051   15.811    0.000
```

Things to note from the model output

- All three regressions paths are statistically significant
- The model is **just-identified**
  - *The degrees of freedom are equal to 0*
  - *The model fit cannot be tested*
  - *The model fit statistics (TLI, CFI, RMSEA, SRMR) all suggest perfect fit but this is meaningless*

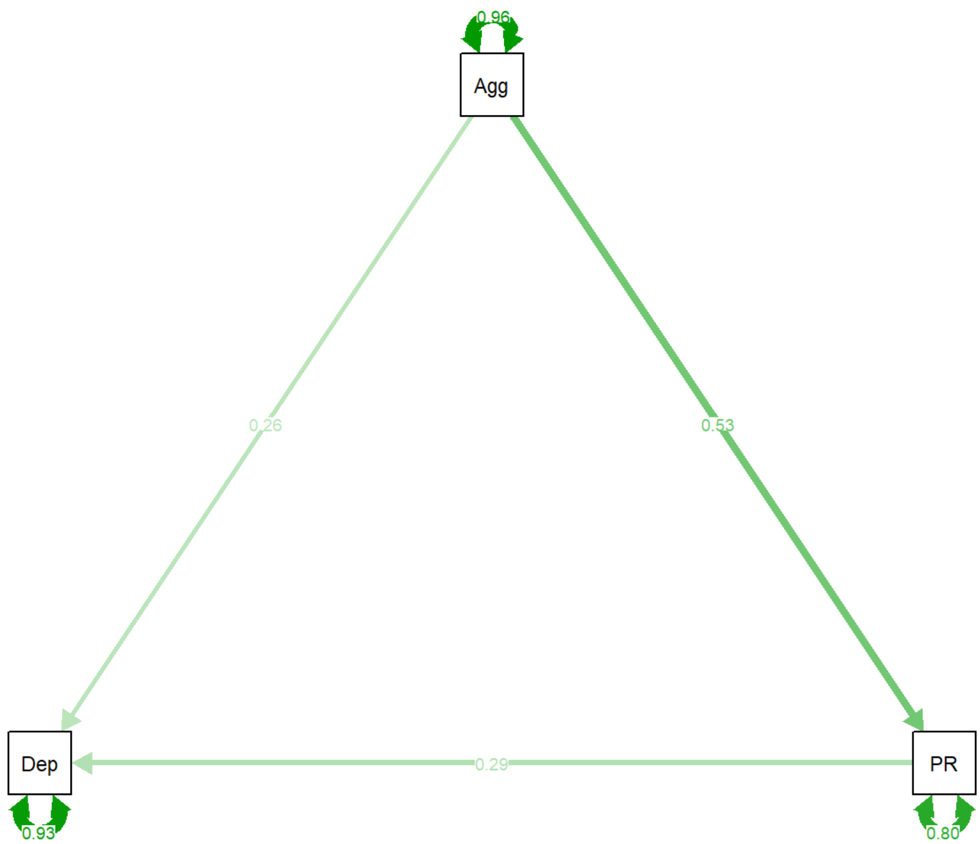
Visualising the model

- We can use semPaths() from the semPlot package to help us visualise the model
  - *Shows the parameter estimates within an SEM diagram*

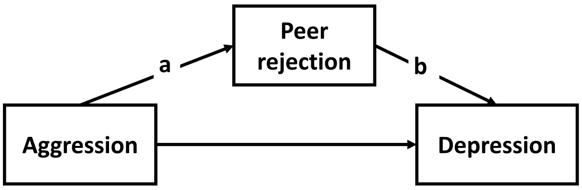
```
library(semPlot)
```

```
## Registered S3 methods overwritten by 'huge':
##   method    from
##   plot.sim  BDgraph
##   print.sim BDgraph
```

```
semPaths(model1.est, what='est')
```



Calculating the indirect effects



- To calculate the indirect effect of X on Y in path mediation, we need to create some new parameters
- The indirect effect of X on Y via M is:
  - $a * b$
  - a = *the regression coefficient for M~X*
  - b = *the regression coefficient for Y~M*

Calculating indirect effects in lavaan

- To calculate the indirect effect of X on Y in lavaan wD:
  - *Use parameter labels ‘a’ and ‘b’ to label the relevant paths*
    - a is for the effect of X on M
    - b is for the effect of M on Y
  - *Use the ‘:=’ operator to create a new parameter ‘ind’*
    - ‘ind’ represents our indirec effect

```
  model1<-'Dep~b*PR      # Add b label here
           Dep~Agg
           PR~a*Agg      # Add a label here

ind:=a*b                 # create a new parameter ind which is the product of a and b'
```

Indirect effects in the output

```r
model1.est<-sem(model1, data=agg.data2)
summary(model1.est)
```

```
## lavaan 0.6-5 ended normally after 13 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of free parameters                          5
##
##   Number of observations                           500
##
## Model Test User Model:
##
##   Test statistic                                 0.000
##   Degrees of freedom                                 0
##
## Parameter Estimates:
##
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##   Standard errors                             Standard
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   Dep ~
##     PR        (b)     0.289    0.048    6.009    0.000
##     Agg               0.256    0.051    5.033    0.000
##   PR ~
##     Agg       (a)     0.530    0.041   12.932    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .Dep               0.932    0.059   15.811    0.000
##    .PR                0.805    0.051   15.811    0.000
##
## Defined Parameters:
##                    Estimate  Std.Err  z-value  P(>|z|)
##     ind               0.153    0.028    5.449    0.000
```

Statistical significance of the indirect effects

- Default method of assessing the statistical significance of indirect effects assume normal sampling distribution
- May not hold for indirect effects which are the product of regression coefficients
- Instead we can use **bootstrapping**
  - *Provides an estimate of the sampling variance of a coefficient based on the actual data*
    - as opposed to a theoretical sampling distribution
  - *Resamples with replacement repeatedly from the observed data*
  - *Calculates the sampling variance based on variation of the coefficient across resamples*
    - Number of resamples usually between 1000 and 10000
  - *Allows 95% confidence intervals (CIs) to be computed*
  - *If 95% CI includes 0, the indirect effect is not significant at alpha=.05*

Bootstapped CIs for indirect effect in lavaan

```
  model1<-'Dep~b*PR
           Dep~Agg
           PR~a*Agg
ind:=a*b'

model1.est<-sem(model1, data=agg.data2, se='bootstrap') #we add the argument se='bootstrap'
```

Output for bootstrapped CIs for an indirect effect in lavaan

```r
summary(model1.est, ci=T) # we add the argument ci=T to see the confidence intervals in the output
```

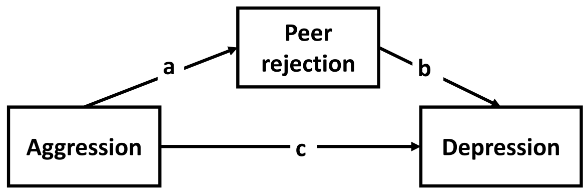
```
## lavaan 0.6-5 ended normally after 13 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of free parameters                          5
##
##   Number of observations                           500
##
## Model Test User Model:
##
##   Test statistic                                 0.000
##   Degrees of freedom                                 0
##
## Parameter Estimates:
##
##   Standard errors                            Bootstrap
##   Number of requested bootstrap draws             1000
##   Number of successful bootstrap draws            1000
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper
##   Dep ~
##     PR        (b)     0.289    0.050    5.791    0.000    0.192    0.390
##     Agg               0.256    0.050    5.149    0.000    0.153    0.347
##   PR ~
##     Agg       (a)     0.530    0.039   13.422    0.000    0.449    0.607
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper
##    .Dep               0.932    0.055   17.020    0.000    0.821    1.042
##    .PR                0.805    0.053   15.193    0.000    0.705    0.908
##
## Defined Parameters:
##                    Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper
##     ind               0.153    0.028    5.423    0.000    0.101    0.212
```

Total effects in path mediation

- As well as the direct and indirect effect, it is often of interest to know the **total** effect of X on Y

$$Total = Indirect + Direct$$

Total effects in path mediation



$$Total = a * b + c$$

Total effect in lavaan

```
  model1<-'Dep~b*PR
           Dep~c*Agg          # we add the label c for our direct effect
           PR~a*Agg
ind:=a*b
total:=a*b+c                  # we add a new parameter for the total effect'

model1.est<-sem(model1, data=agg.data2, se='bootstrap') #we add the argument se='bootstrap'
```

Total effect in lavaan output

```
summary(model1.est, ci=T)
```

```
## lavaan 0.6-5 ended normally after 13 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of free parameters                          5
##
##   Number of observations                           500
##
## Model Test User Model:
##
##   Test statistic                                 0.000
##   Degrees of freedom                                 0
##
## Parameter Estimates:
##
##   Standard errors                            Bootstrap
##   Number of requested bootstrap draws             1000
##   Number of successful bootstrap draws            1000
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper
##   Dep ~
##     PR         (b)    0.289    0.049    5.943    0.000    0.191    0.387
##     Agg        (c)    0.256    0.052    4.946    0.000    0.154    0.364
##   PR ~
##     Agg        (a)    0.530    0.040   13.098    0.000    0.446    0.610
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper
##    .Dep               0.932    0.054   17.230    0.000    0.820    1.032
##    .PR                0.805    0.054   14.898    0.000    0.699    0.916
##
## Defined Parameters:
##                    Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper
##     ind               0.153    0.028    5.483    0.000    0.099    0.209
##     total             0.410    0.046    8.855    0.000    0.313    0.503
```

Why code the total effect in lavaan?

- We could have just added up the coefficients for the direct and indirect effects
- By coding it in lavaan, however, we can assess the statistical significance of the total effect
- Useful because sometimes the direct and indirect effects are not individually significant but the total effect is
  - *May be especially relevant in cases where there are many mediators of small effect*

Interpreting the total, direct, and indirect effect coefficients

- The total effect can be interpreted as the *unit increase in Y expected to occur when X increases by one unit*
- The indirect effect can be interpreted as the *unit increase in Y expected to occur via M when X increases by one unit*
- The direct effect can be interpreted as the *unit increase in Y expected to occur with a unit increase in X over and above the increase transmitted by M*
  - *Note: 'direct' effect may not actually be direct - it may be acting via other mediators not included in our model*

Standardised parameters

- As with CFA models, standardised parameters can be obtained using:

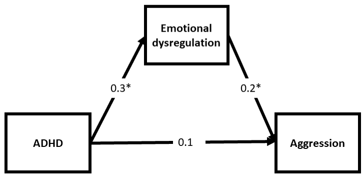
```
summary(model1.est, ci=T, standardized=T)
```

```
## lavaan 0.6-5 ended normally after 13 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of free parameters                          5
##
##   Number of observations                           500
##
## Model Test User Model:
##
##   Test statistic                                 0.000
##   Degrees of freedom                                 0
##
## Parameter Estimates:
##
##   Standard errors                            Bootstrap
##   Number of requested bootstrap draws             1000
##   Number of successful bootstrap draws            1000
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper
##   Dep ~
##     PR         (b)    0.289    0.049    5.943    0.000    0.191    0.387
##     Agg        (c)    0.256    0.052    4.946    0.000    0.154    0.364
##   PR ~
##     Agg        (a)    0.530    0.040   13.098    0.000    0.446    0.610
##    Std.lv  Std.all
##
##     0.289    0.278
##     0.256    0.233
##
##     0.530    0.501
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper
##    .Dep               0.932    0.054   17.230    0.000    0.820    1.032
##    .PR                0.805    0.054   14.898    0.000    0.699    0.916
##    Std.lv  Std.all
##     0.932    0.803
##     0.805    0.749
##
## Defined Parameters:
##                    Estimate  Std.Err  z-value  P(>|z|) ci.lower ci.upper
##     ind               0.153    0.028    5.483    0.000    0.099    0.209
##     total             0.410    0.046    8.855    0.000    0.313    0.503
##    Std.lv  Std.all
##     0.153    0.139
##     0.410    0.372
```

Reporting path mediation models

- Methods
  - *The model being tested*
  - *e.g. ‘Y was regressed on both X and M and M was regressed on X’*
  - *The estimator used (e.g., maximum likelihood estimation)*
  - *The method used to test the significance of indirect effects ('bootstrapped 95% confidence)*
- Results
  - *Model fit (for over-identified models)*
  - *The parameter estimates for the path mediation and their statistical significance*
    - Can be useful to present these in a SEM diagram
    - Helps reader better visualise the model
    - The diagrams from R not considered ‘publication quality’ - draw in powerpoint or similar

Reporting path mediation models - example of SEM diagram with results



Note. *=significant at p<.05

- Include the key parameter estimates
- Indicate statistically significant paths (e.g. with an '*')
- Include a figure note that explains how statistically significant paths (and at what level) are signified

Reporting path mediation models - the indirect effects

- Results
  - *The coefficient for the indirect effect and the bootstrapped 95% confidence intervals*
  - *Common to also report **proportion mediation**:*

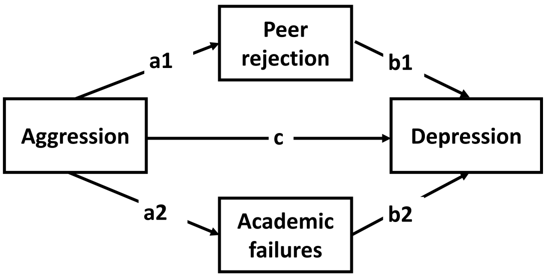
$$\frac{indirect}{total}$$

- However, important to be aware of limitations:
  - *Big propotion mediation possible when total effect is small - makes effect seem more impressive*
  - *Small proportion mediation even when total effect is big - can underplay importance of effect*
  - *Should be interpreted in context of total effect*
- Tricky interpretation if there are a mix of negative and positive effects involved

Extensions of path mediation models

- We can extend our path mediation model in various ways:
  - *Several mediators in sequence or parallel*
  - *Multiple outcomes*
  - *Multiple predictors*
  - *Multiple groups (e.g., comparing direct and indirect effects across males and females)*
  - *Add covariates to adjust for potential confounders*

Example: multiple mediation model




```
model2<-'Dep~b2*Aca
         Aca~a2*Agg

         Dep~b1*PR
         PR~a1*Agg

         Dep~c*Agg  # direct effect

ind1:=a1*b1
ind2:=a2*b2
total=a1*b1+a2*b2+c'
```

Other path analysis models

- Path mediation models are a common application of path models
- But they are just one example
- Anything that can be expressed in terms of regressions between observed variables can be tested as a path model
- Can include ordinal or binary variables
- Can include moderation
- Other common path analysis models include:
  - *Autoregressive models for longitudinal data*
  - *Cross-lagged panel models for longitudinal data*

Other path analysis models - AR

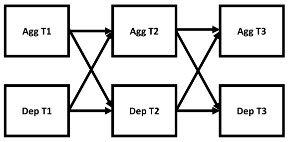
- autoregressive models to examine the stability of a construct over time


```
##creating a lavaan syntax object for an autoregressive model

Autoregressive<-'AggT3~AggT2
                 AggT2~AggT1'
```

Other path analysis models - CLPM

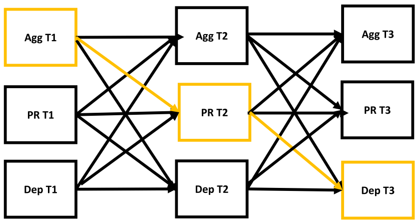
- cross-lagged panel models to examine the relations between constructs over time
  - *autoregressive paths control for previous levels of each construct*
  - *cross-lagged paths capture the relations between the two constructs*




```
# creating a lavaan syntax objecr for a CLPM
CLPM<-'AggT3~AggT2+DepT2
       AggT2~AggT1+DepT1
       DepT3~DepT2+AggT2
       DepT2~DepT1+AggT1'
```

Other path analysis models - CLPM with mediation

- longitudinal mediation models using a cross-lagged panel model




```
#creating a lavaan syntax object for a longitudinal mediation model

CLPM.med<-'AggT3~AggT2+DepT2+PRT2
           AggT2~AggT1+DepT1+PRT1
           DepT3~DepT2+AggT2+b*PRT2   # label the effect of M on Y as b
           DepT2~DepT1+AggT1+PRT1
           PRT3~PRT2+AggT2+DepT2
           PRT2~PRT1+a*AggT1+DepT1    # label the effect of X on M as a

ind:=a*b'
```

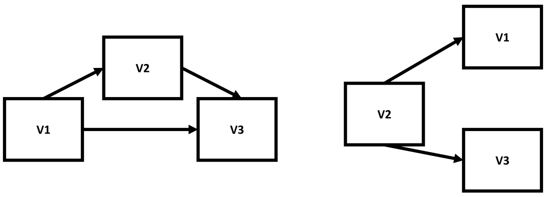
Making model modifications

- As in CFA models, you *may* want to make some modifications to your initially hypothesised model
  - *non-significant paths that you want to trim*
  - *include some additional paths not initially included*
- Remember that this now moves us into exploratory territory where:
  - *Model modifications should be substantively as well as statistically justifiable*
  - *You must be aware of the possibility that you are capitalising on chance*
  - *You should aim to replicate the modifications in independent data*

Cautions regarding path analysis models

- **Assumption** that the paths represent causal effects is only an assumption
  - *Especially if using cross-sectional data*
- The parameters are only accurate if the model is correctly specified

Cautions regarding path analysis models - indistinguishable models



Mediation | Common cause

Your assumed model | The actual model

Measurement error in path analysis

- Path analysis models use observed variables
- Assumes no measurement error in these variables
- Path coefficients likely to be attenuated due to unmodelled measurement error
- Structural equation models solve this issue
- They are path analysis models where the paths are between latent rather than observed variables
- …more on this next week

Path analysis summary

- Path analysis can be used to fit sets of regression models
  - *Common path analysis model is the path mediation model*
  - *But very flexible - huge range of models that can be tested*
- In R, path analysis can be done using the sem( ) function in lavaan
- Need to be aware that we aren't *testing* causality but assuming it