Statistical Analysis Using Structural Equation Models

EPsy 8266

Christopher David Desjardins

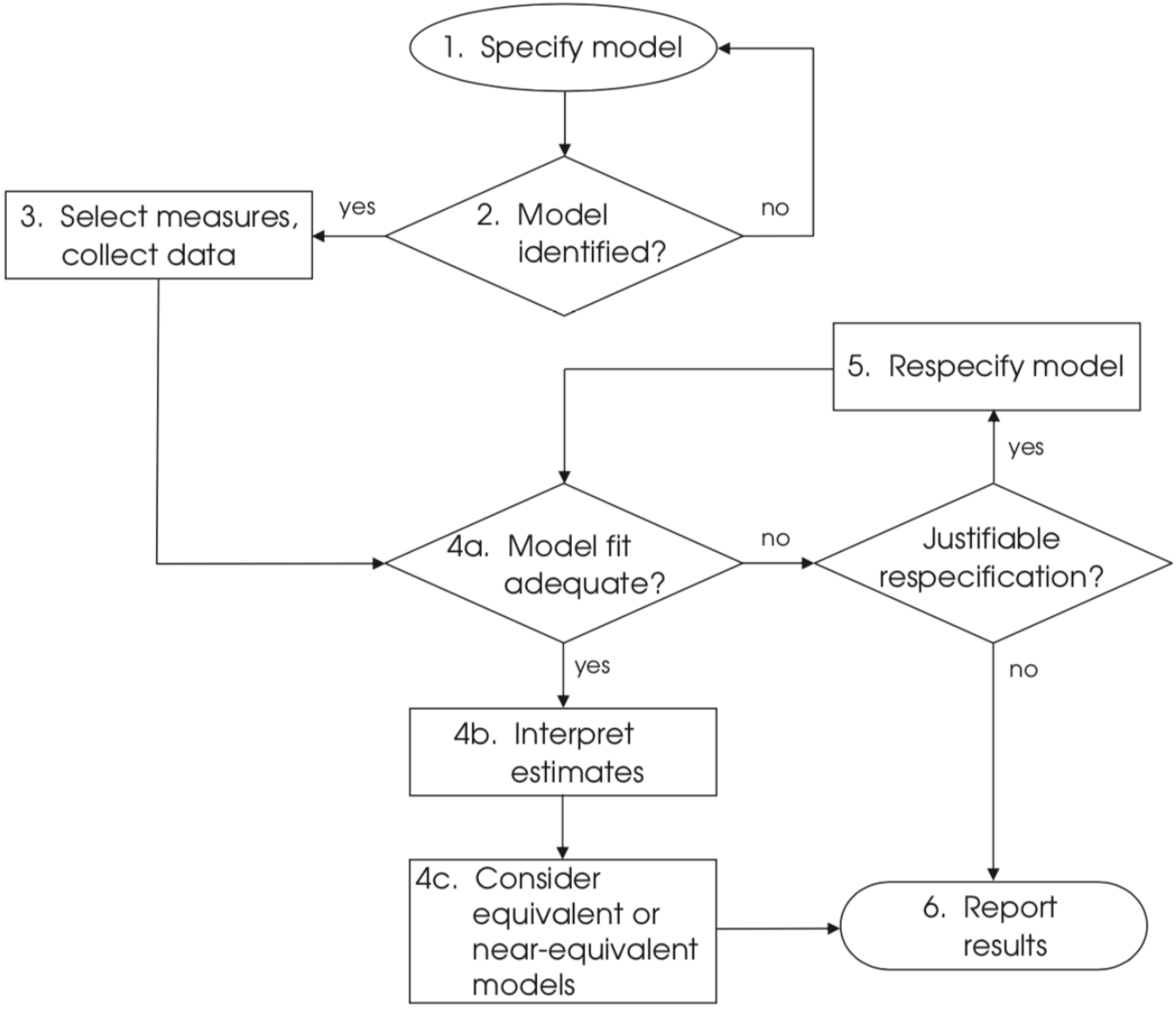
Research Methodology Consulting Center

2/19/19

Finding a paper

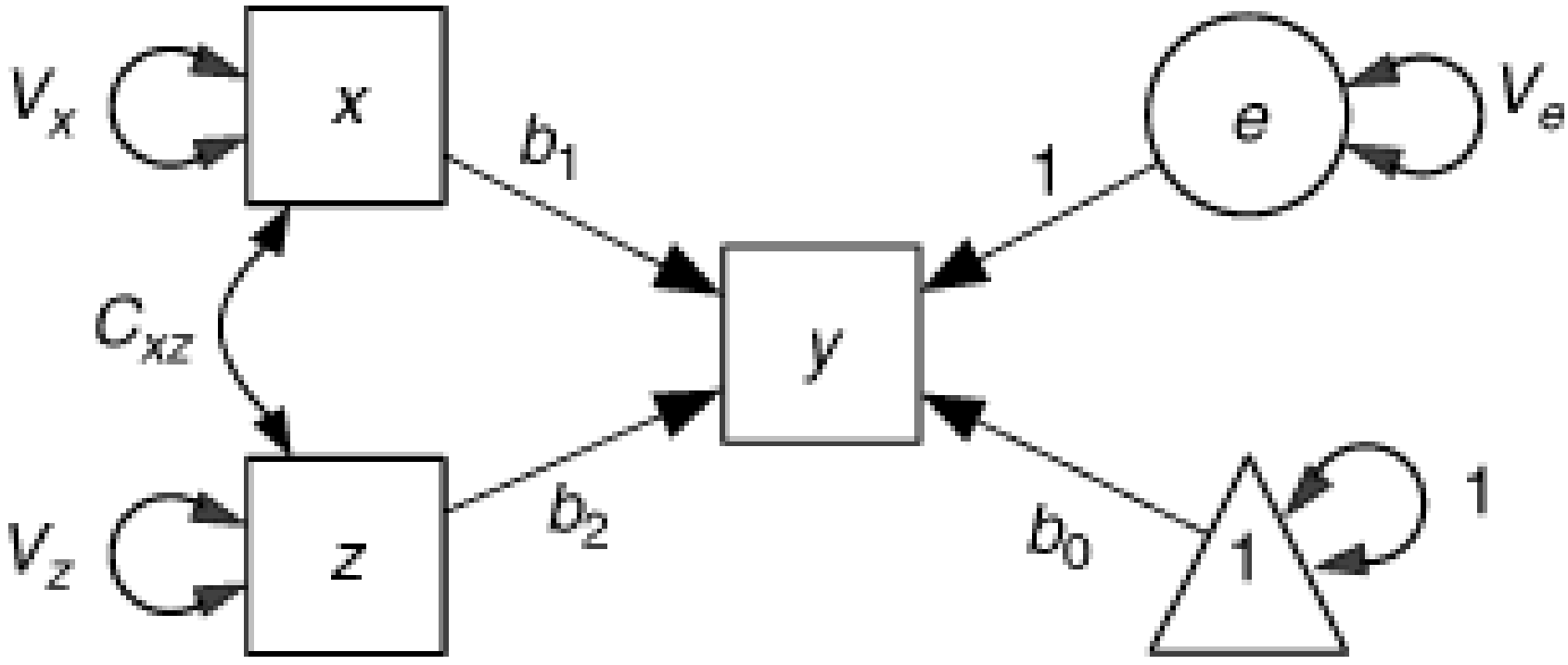
1. **Find a paper with a covariance matrix or a correlation matrix and standard deviations (possibly with means).**
2. **Read the data into R.**
3. Briefly (a few sentences) explain their RQs and what their model/analyses is
4. Goal: To recreate their results and I'll host it on Github here: https://github.com/cddesja/lavaan-reproducible

Kline's SEM workflow



RAM Symbols

- Observed variables are squares
- Latent variables are circles (this includes errors and disturbances)
- X $\rightarrow$ Y. "X causes Y". If X is changed (intervened upon) then Y will change. but nothing is assumed about direction, sometimes called "spurious relationship". X and Y might have a common cause(s) not included in the model explicitly.
- X $\leftrightarrow$ Y. "X and Y are simply correlated", but nothing is assumed about direction, sometimes called "spurious relationship". X and Y might have a common cause(s) not included in the model explicitly.


V_x
x
b_1
e
V_e
1
C_{xz}
y
V_z
z
b_2
b_0
1
1

Review

How do we determine the number of unique elements in a covariance matrix?

How do we calculate degrees of freedom?

Types of parameters

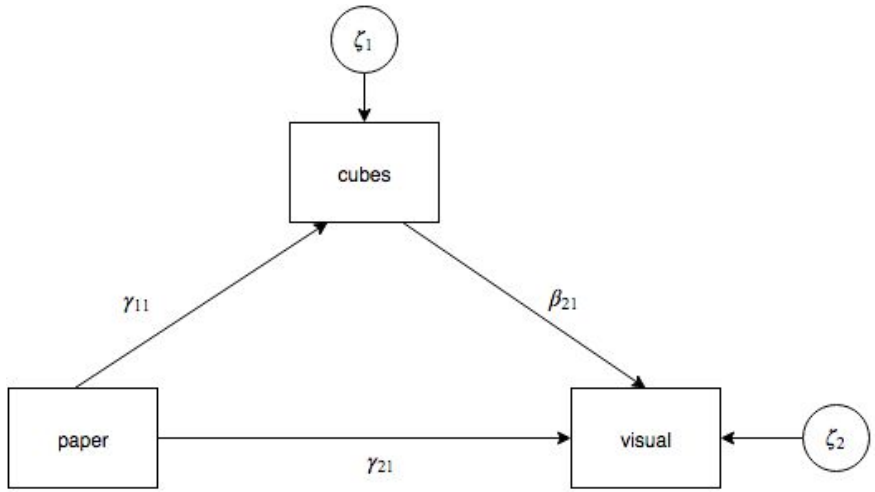
Parameters can be ...

- Fixed
- Free
- Constrained

Basic idea of path analysis

Bivariate correlation between any two variables can be broken down into a series of effects: direct causal effects, indirect causal effects, and noncausal or spurious components.

Consider the following:



The effects in this model are estimated using the following two regression equations:

$$cubes = \gamma_{11} paper + \zeta_1$$
$$visual = \gamma_{21} paper + \beta_{21} cubes + \zeta_2$$

The bivariate correlation between *cubes* and *visual* can be reproduced from these standardized regression coefficients. That is, $\rho_{13} = \gamma_{21} + \gamma_{11}\beta_{21}$.

```r
hs.data <- read.csv("https://tinyurl.com/y5crk8ur")
cor(hs.data[,c("visual", "cubes", "paper")])

##           visual     cubes     paper
## visual 1.0000000 0.2973455 0.3652928
## cubes  0.2973455 1.0000000 0.2379818
## paper  0.3652928 0.2379818 1.0000000

library("lavaan")
mod <- '
cubes ~ paper
visual ~ paper + cubes
'
fit <- sem(mod, data = hs.data)
```

```r
standardizedSolution(fit, ci = FALSE)
```

```
##      lhs op    rhs est.std    se      z pvalue
## 1  cubes  ~  paper   0.238 0.054  4.440      0
## 2 visual  ~  paper   0.312 0.050  6.226      0
## 3 visual  ~  cubes   0.223 0.053  4.241      0
## 4  cubes ~~  cubes   0.943 0.026 36.978      0
## 5 visual ~~ visual   0.820 0.039 20.949      0
## 6  paper ~~  paper   1.000 0.000     NA     NA
```

The correlation between *paper* and *visual* was 0.36.

Questions

How many unique elements in our covariance matrix?

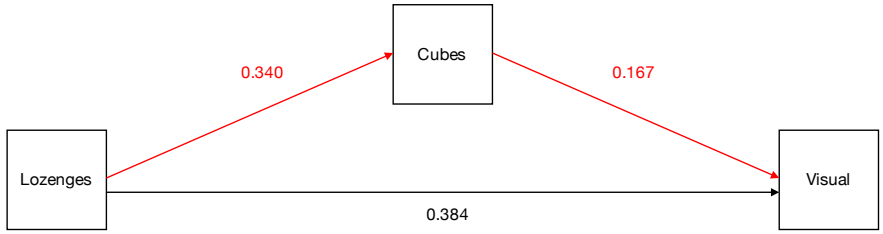
Is our model above under identified, just identified, or over identified?

How many parameters are we estimating?

```r
fitMeasures(fit, c("npar", "df"))

## npar   df
##    5    0
```

Calculation of indirect effects

Path Multiplication Rule - The value of the effect associated with a compound path is the product of its path coefficients (*this works for standardized regression coefficients or unstandardized*).



Standardized regression coefficient of Cubes on Lozenges is $\gamma_{CL} = 0.340$, and the regression of Visual on Cubes yields a regression slope of $\beta_{VC} = 0.167$.

What is the indirect effect of Visual on Lozenges?

For a 1 standard deviation increase in Lozenges, Cubes goes up 0.340 SD.

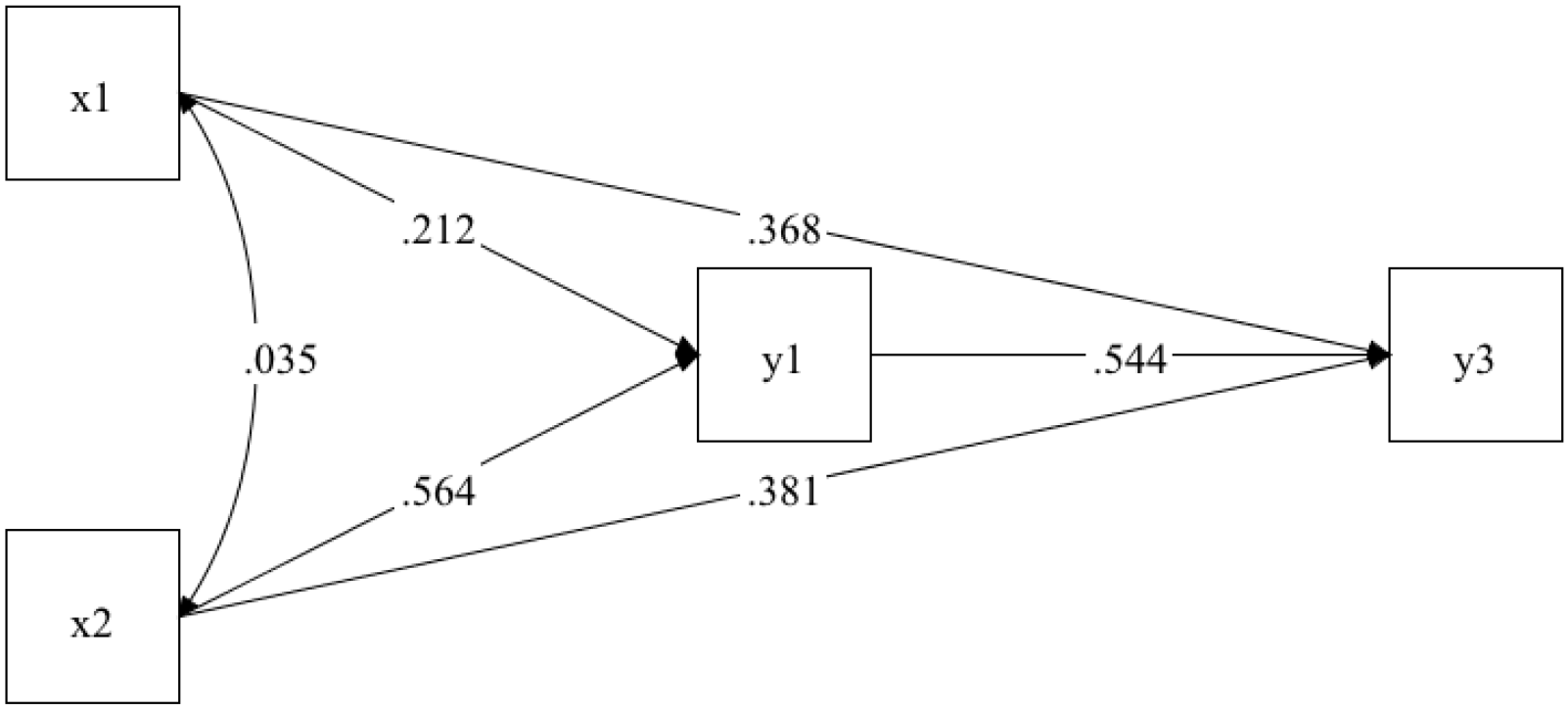
If Cubes goes up 0.340 SD then Visual goes up 0.340 * 0.167 = 0.057 SD.

So, the indirect effect of a 1 standard deviation increase in Lozenges through Cubes on Visual is a .057 SD increase in Visual.

Decomposing and calculating effects

- **Direct effect** - direct influence of variable of interest (X) on another variable (DV) that is *unmediated* by any other variable, i.e. each single headed arrow represents a direct effect.
- **Indirect effects** - start from the DV later in the path diagram (on the right). Trace backwards (right to left) against arrows passing intervening variables until you get to X. Each combination of intervening variables is a separate indirect effect.
- **Spurious effects (due to common causes)** Start from DV. Trace backwards to a variable, Z, that has a direct or indirect effect on both X and DV. Move from Z to X. There are as many spurious effects of X on DV due to Z as many ways you can get from DV to X through Z following the rule above.
- **Correlated (unanalyzed) effects**
  - If X is exogenous, find variable Z that is both exogenous and has a direct or indirect effect on DV. Start from variable DV. Trace back to Z. Make the last step through the double headed arrow to X.
  - If X is endogenous, find an exogenous variable Z that has a direct or indirect effect on DV and is correlated to another exogenous variable W that has a direct or indirect effect on X. Start from variable DV. Trace back to Z. Travel through the double headed arrow to W. Move from W to X.

Decomposing effects activity



	y1	x1	x2
y3	0.847	0.508	0.705

Correlations

Mediation or indirect effects?

- Mediation implies causation.
  - X causes a change in M, which changes Y.
- For this to happen, really need temporal precedence.
- Best to just refer to indirect effects.

Is relationship of between X and Y mediated by M?

How do we assess this?

1. Is there a statistically significant relationship between X and Y? If yes, record the effect as c. Then
2. Is there a statistically significant relationship between X and M? If yes, record the effect as a, Then.
3. Is there a statistically significant relationship between M and Y? If yes, then.
4. Estimate the effect that X and M have simultaneously on Y. Record the M on Y effect as b, record the X on Y effect as c'

If $'$ is not significantly, then the causal effect of X on Y is fully mediated by M. If $|c'|$ is less than $|c|$, then the causal effect of X on Y is partially mediated by M.

Sometimes, presented as $\frac{|c-c'|}{c}$, the percent of the total causal effect of X on Y explained by the mediator M.

Approaches to testing indirect effects

- Baron & Kenny's causal steps approach (Bad)
  - Do a series of regressions and determine if c' is $< c$.
  - Not a test of mediation (ab pathway)
- Product of coefficients approach (Sobel test) (Not so good)
  - Determine if ab is significant using a Wald test
  - Significance test of ab is problematic. Assumes normal distributions of SEs.
- SEM estimation of indirect effect (Not great)
  - Determine if ab is significant.

```
lavaan.mod <- "
y1 ~ a*x1 + c*x2 + e*x3;
y3 ~ b*y1 + d*x1 + f*x2;
# indirect effect of x1 on y3
ab := a*b
"
```

  - Use ML estimation to determine significance of indirect effect (still assumes normality assumption)

The bootstrap

- In traditional statistics, standard errors and confidence intervals are based on theoretical sampling distributions of the parameter estimates.
- These have certain distributional assumptions and/or rely on asymptotic theory (i.e. when we have a large sample)
- Often we violate these assumptions (i.e., we have non-normal data) or our sample size is really too small for these statistics to be accurate.
- One thing we can do is to use the bootstrap developed by Efron (1982).
- It creates an **empirical sampling distribution of parameter estimates** that you can use in an analogous fashion to the theoretical sampling distribution.

How it works

1. Let's say your data set is size n. Fit the model and obtain some parameter estimate β.
2. Take a sample size of n from your data set with replacement. This is the first bootstrap sample.
3. Use this sample, fit your model again and obtain your parameter estimate for β.
4. Repeat steps 2 - 3, a large number of times. Let's call this k times.
5. The distribution of your k estimates of β is your empirical sampling distribution. From this you can calculate:
   - Standard error of β - This is just the standard deviation of this distribution.
   - 95% confidence interval - Locate the 2.5% percentile and the 97.5% percentile.
   - A bias correction is recommended.

This whole resampling procedure is known as **non-parametric bootstrapping**.

Method 1: The non-parametric bootstrap

This toy example is using the `mtcars` data set (`?mtcars`). Let's say we want to predict mile per gallon given the weight of the car. We want to use the bootstrap on the slope of weight.

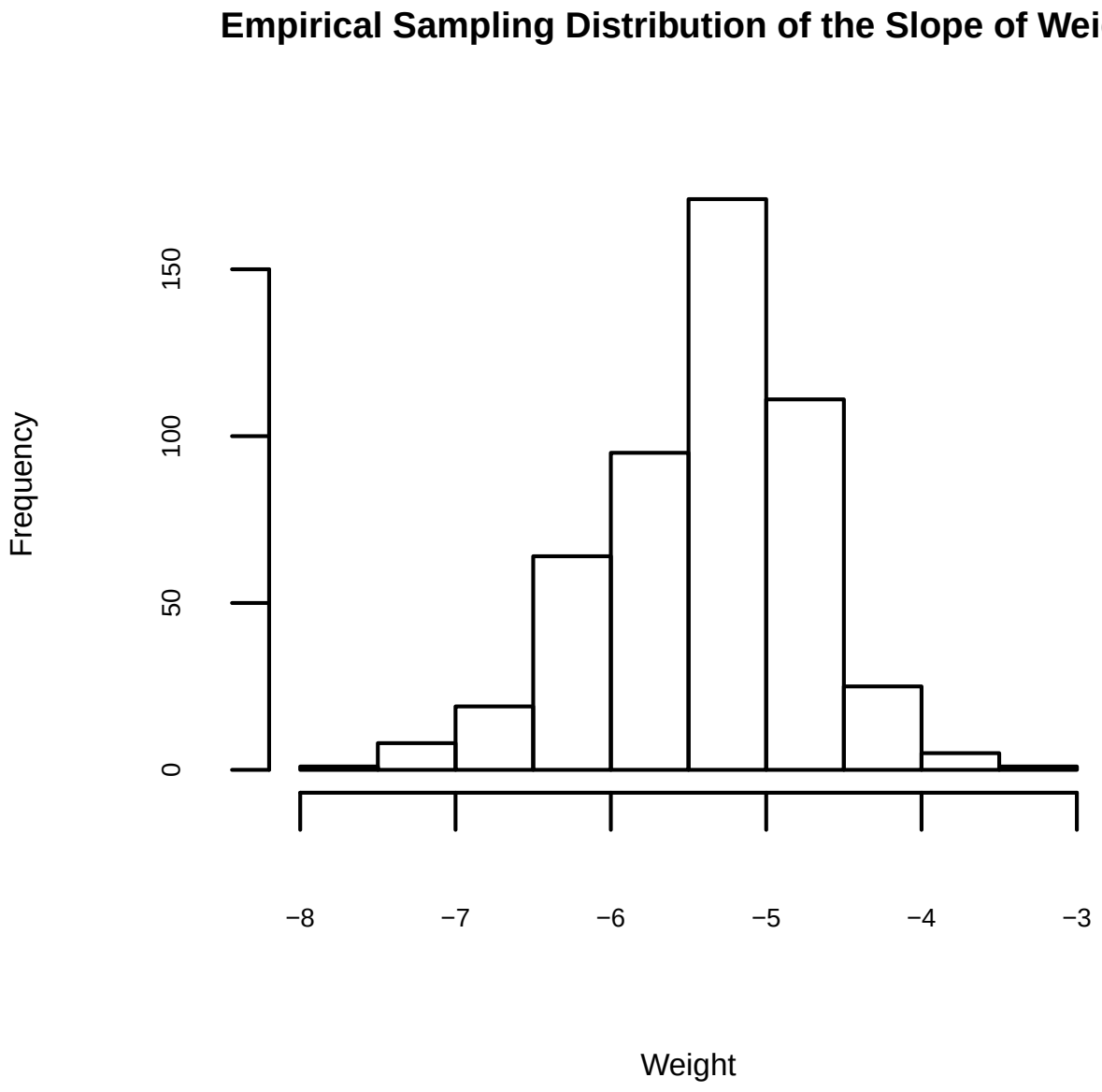
```
step1 <- coef(lm(mpg ~ wt, data = mtcars))[[2]]
step1

## [1] -5.344472
```

Our estimate is that for 1000 lbs, we expect that miles per gallon to decrease by 5.34.

Method 1: The R code

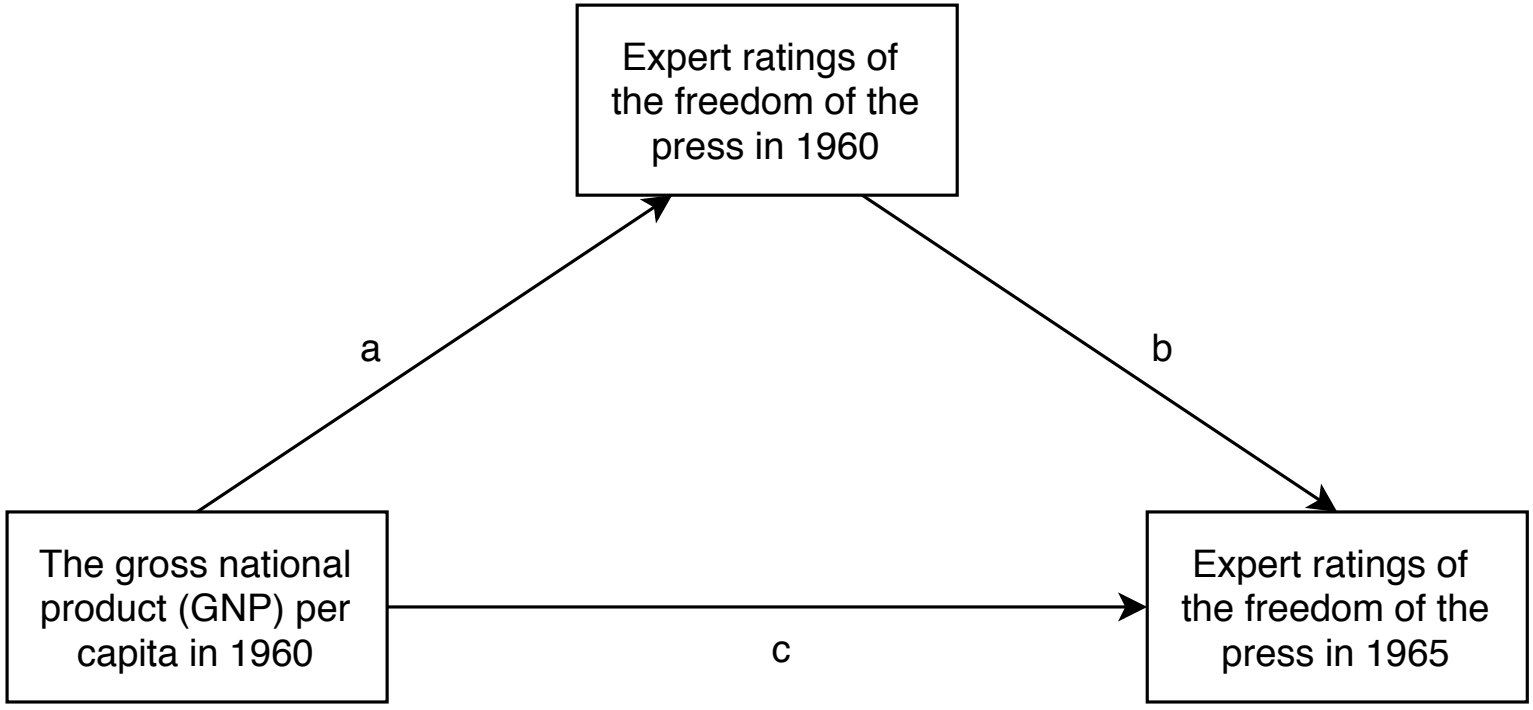
```
k <- 500
n <- nrow(mtcars)
beta.weight <- rep(NA, k)
for(i in 1:k){
  obs <- sample(x = n, size = n, replace = T)
  tmp <- mtcars[obs,]
  beta.weight[i] <- coef(lm(mpg ~ wt, data = tmp))[[2]]
}
```

Method 1: Plotting the empirical distribution



Method 1: Using the empirical distribution

```
# Calculate a 95% confidence interval
quantile(beta.weight, probs = c(.025, .975))

##      2.5%     97.5%
## -6.796561 -4.189005
```


Expert ratings of the freedom of the press in 1960
a
b
The gross national product (GNP) per capita in 1960
c
Expert ratings of the freedom of the press in 1965

Bootstrapping the indirect effect

```
model <- '
  y1 ~ a*x1
  y5 ~ c*x1 + b*y1

  # indirect effects
  ab := a*b
  totl := c + a*b
'
fit <- sem(model, data = PoliticalDemocracy)
summary(fit)

# define the function to extract the indirect effect
boot.fun <- function(x) {
  parameterEstimates(x)[parameterEstimates(x)$label == "ab", "est"]
}

# run the bootstrap
# - by default this is nonparameteric
ab.dist <- bootstrapLavaan(fit, R = 5000, FUN = boot.fun)
hist(ab.dist)
quantile(ab.dist, probs = c(.025, .975))
```

Bootstrapping a covariance matrix

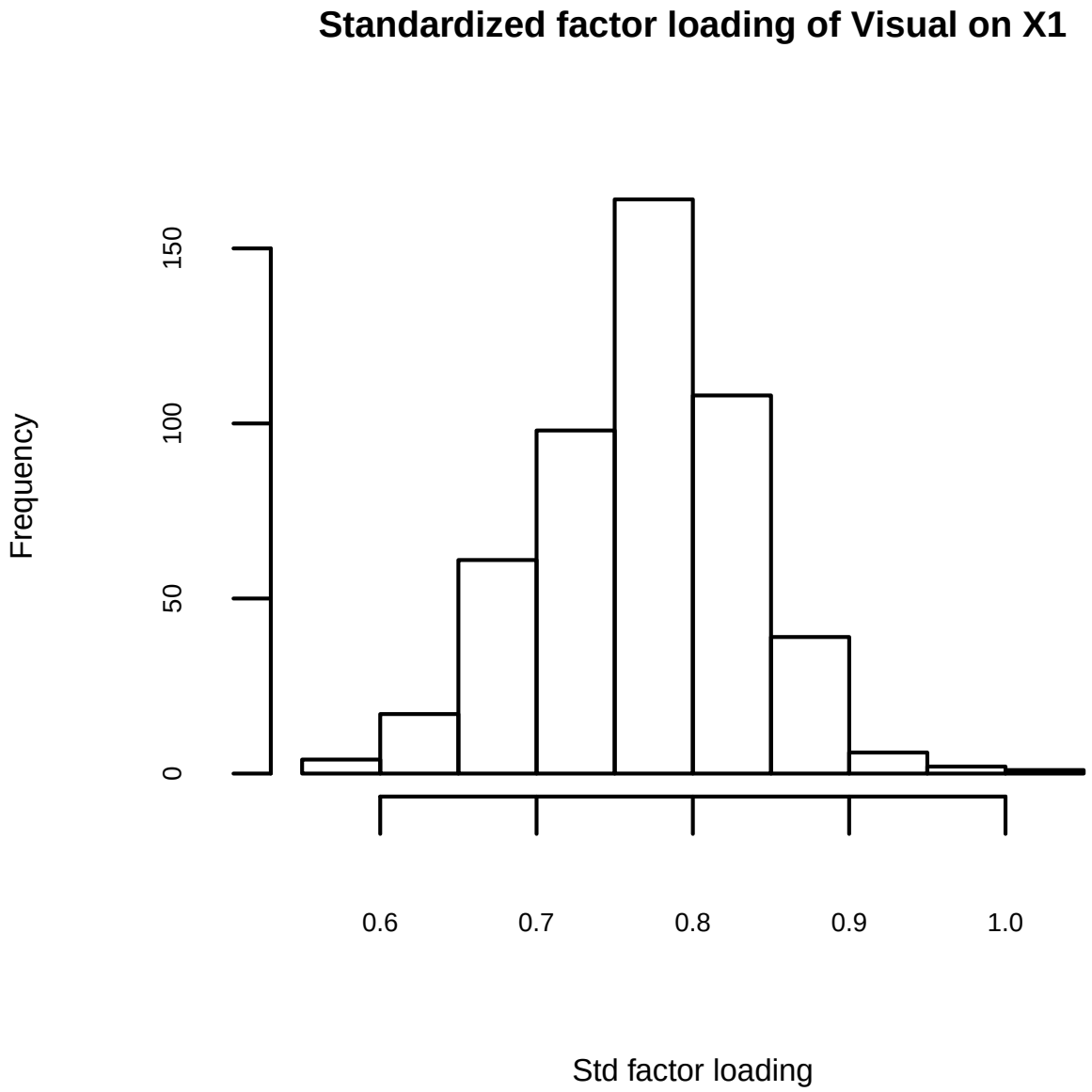
- If you have just a covariance matrix because the data are not available, then you can use a **parametric bootstrap, aka a Monte Carlo bootstrap**.
  1. To do this, you generate a sample of size n from a multivariate normal distribution with a covariance matrix equal to the covariance matrix you've obtained.
  2. This can be done in R using `MASS::mvrnorm`
  3. We we aren't resampling k, we are generating k sets of new data with our observed covariance matrix.
  4. Note, this relies on multivariate normality! But we don't have the data, so maybe it's multivariate normal?
  5. This is very useful for obtaining standard errors when we don't have a theoretical distribution. So, with latent variables, this can be useful getting SEs for standardized loadings and indirect paths.
  6. We want a large k, the larger the better. Computing is fast and cheap nowadays!
  7. The larger the k, the more smoother the distribution will be.

Method 2: The parametric bootstrap from the HZ covariance

In this example, we are trying to get a 95% CI around the standardized coefficient for `visual` on `x1`.

```
library("lavaan")
library("MASS")
hs.cov <- cov(HolzingerSwineford1939[,7:ncol(HolzingerSwineford1939)])
k <- 500
n <- 301
HS.model <- 'visual  =~ x1 + x2 + x3
             textual =~ x4 + x5 + x6
             speed   =~ x7 + x8 + x9'
est.par <- rep(NA, k)
for(i in 1:k){
  tmp <- mvrnorm(n, mu = rep(0, ncol(hs.cov)), Sigma = hs.cov)
  tmp <- as.data.frame(tmp)
  fit.tmp <- cfa(HS.model, tmp)
  est.par[i] <- standardizedSolution(fit.tmp)[1,4]
}
```

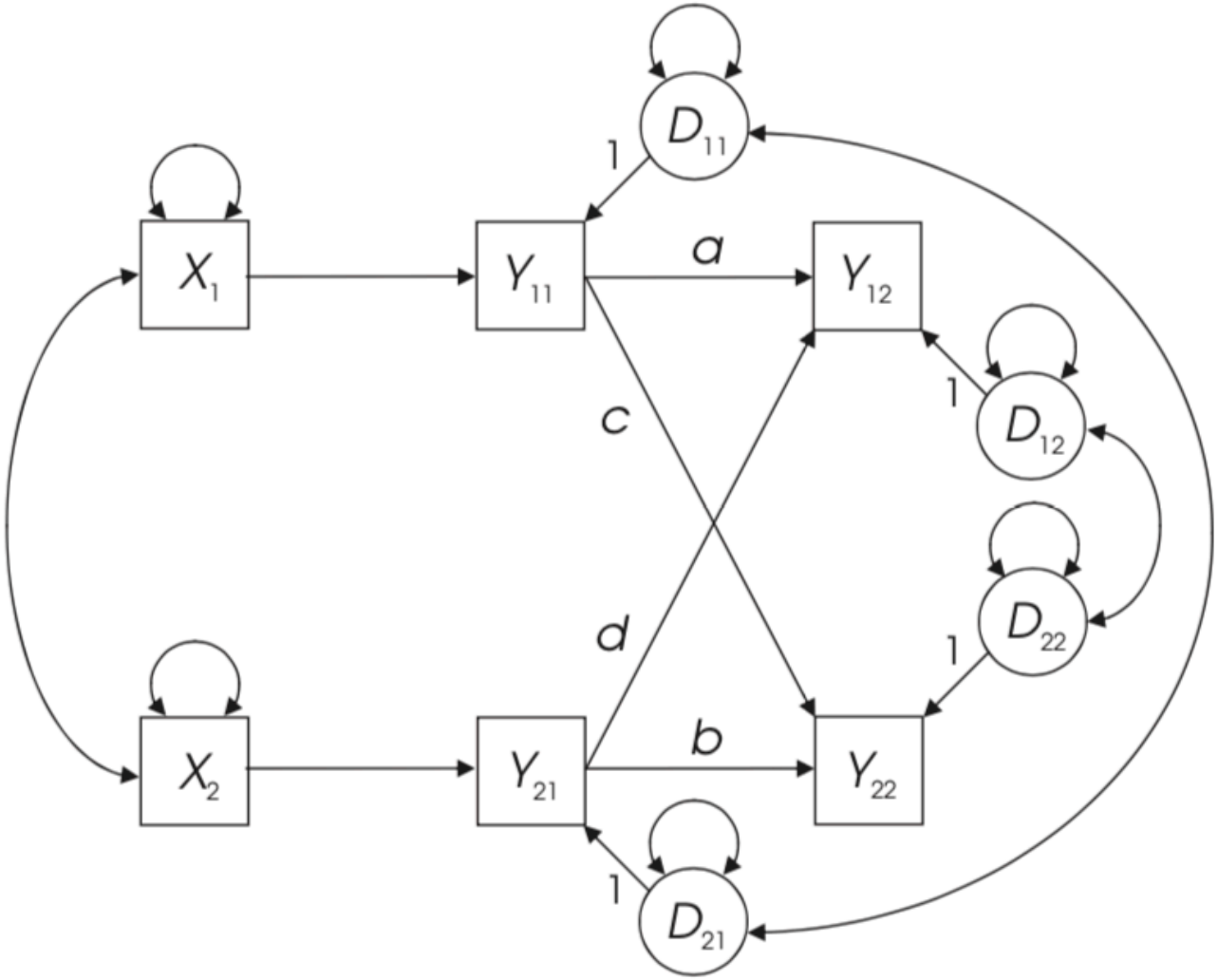
Method 2: Plotting the empirical distribution



Method 2: Using the empirical distribution

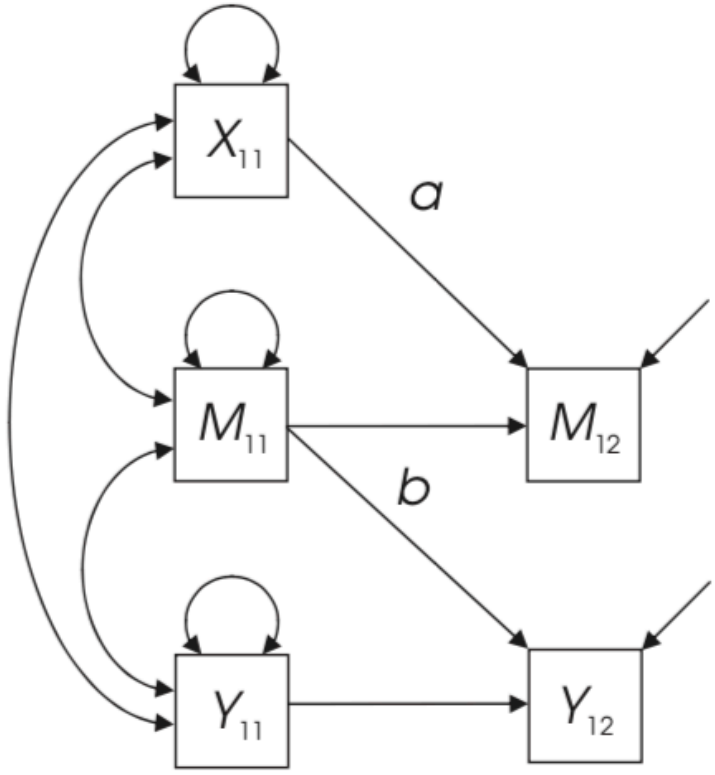
```
# Point estimate
mean(est.par)

## [1] 0.7693568

# Calculate a 95% confidence interval
quantile(est.par, probs = c(.025, .975))

##      2.5%     97.5%
## 0.6356613 0.8898625

# Truth?
fit.true <- cfa(HS.model, HolzingerSwineford1939)
standardizedSolution(fit.true)[1,]

##      lhs op rhs est.std    se      z pvalue
## 1 visual =~  x1   0.772 0.055 14.041      0
```

Panel model

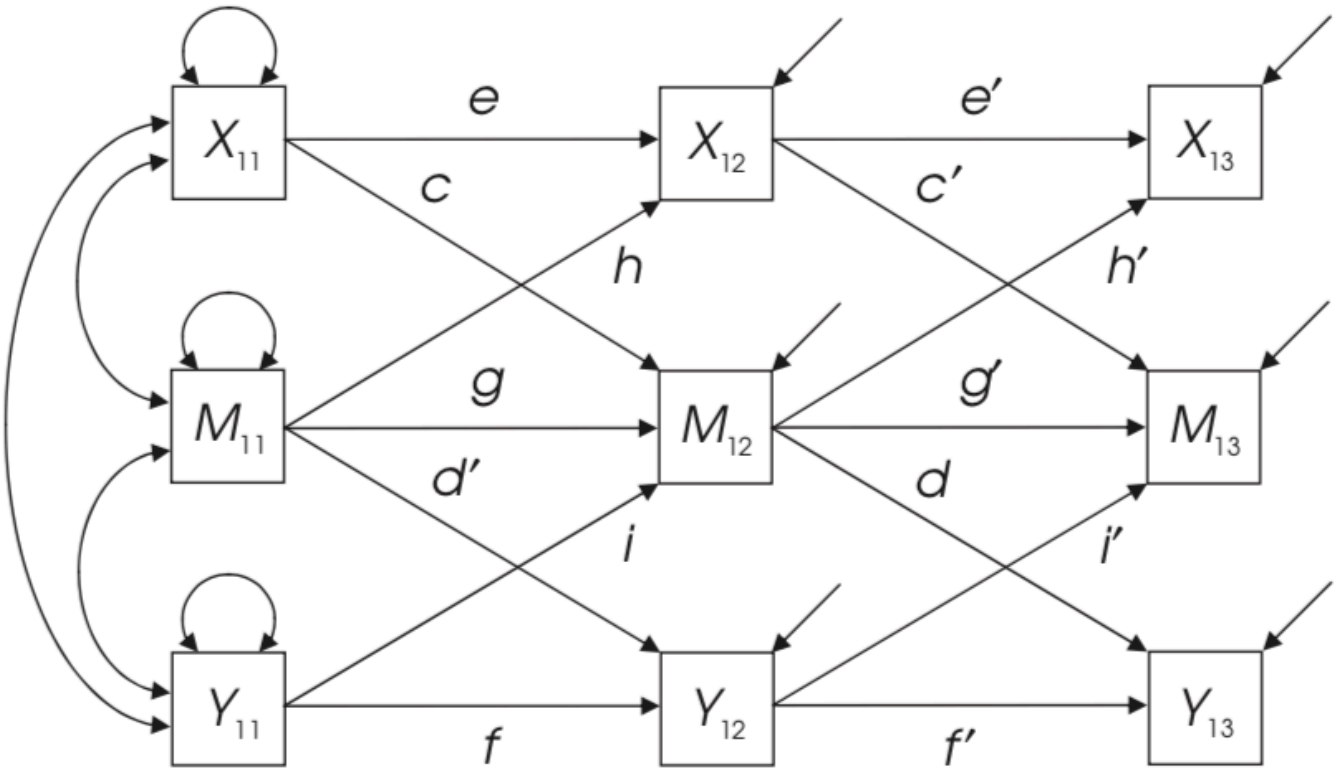


Panel mediation model 1

(a) Half longitudinal mediation



Panel mediation model 2



Non-recursive models

- Models with reciprocal paths (feedback loops)
- Models with correlated errors

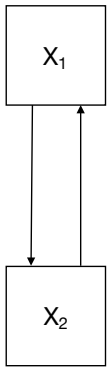
When models include one of these, then the assumption of predictors being uncorrelated with errors and/or errors being uncorrelated is broken

OLS regression will not yield unbiased estimators

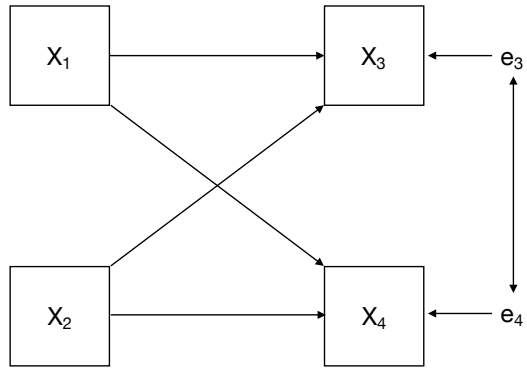
Estimation for Path Analysis

A path analysis model is recursive if it doesn't contain:

Feedback loops



Correlated errors



- If the path analysis model has feedback loops and/or has correlated errors then a method that estimates the paths simultaneously must be used. For example maximum likelihood for the multivariate vector of all observed variables
- Note that the simultaneous ML method can also be used when the model is recursive. This is simpler than performing several different regressions because it is done all in one step.

Reciprocal paths

1. One variable causes another and it in turn causes the first and so on.
2. Equilibrium must be assumed, i.e. direct effects must remain stable throughout an infinite number of reverberations.
3. Total effect is calculated as the infinite sum

Identifiability

- Necessary and sufficient conditions for the identifiability of models with reciprocal paths are complex (Bollen, 1989).
- Each endogenous variable in a feedback loop must have its own unique predictor (**instrumental variable**).
- Some people have argued that reciprocal paths can be used to test competing models of causation by identifying the primary cause.