Introduction to structural equation modeling and mixed models in R

Day 6

Oksana Buzhdygan

oksana.buzh@fu-berlin.de

27.02.2023

Day 6 – Part 1

Day 6 – Part 1

Outline

- Assumptions of Covariance-Based Estimation

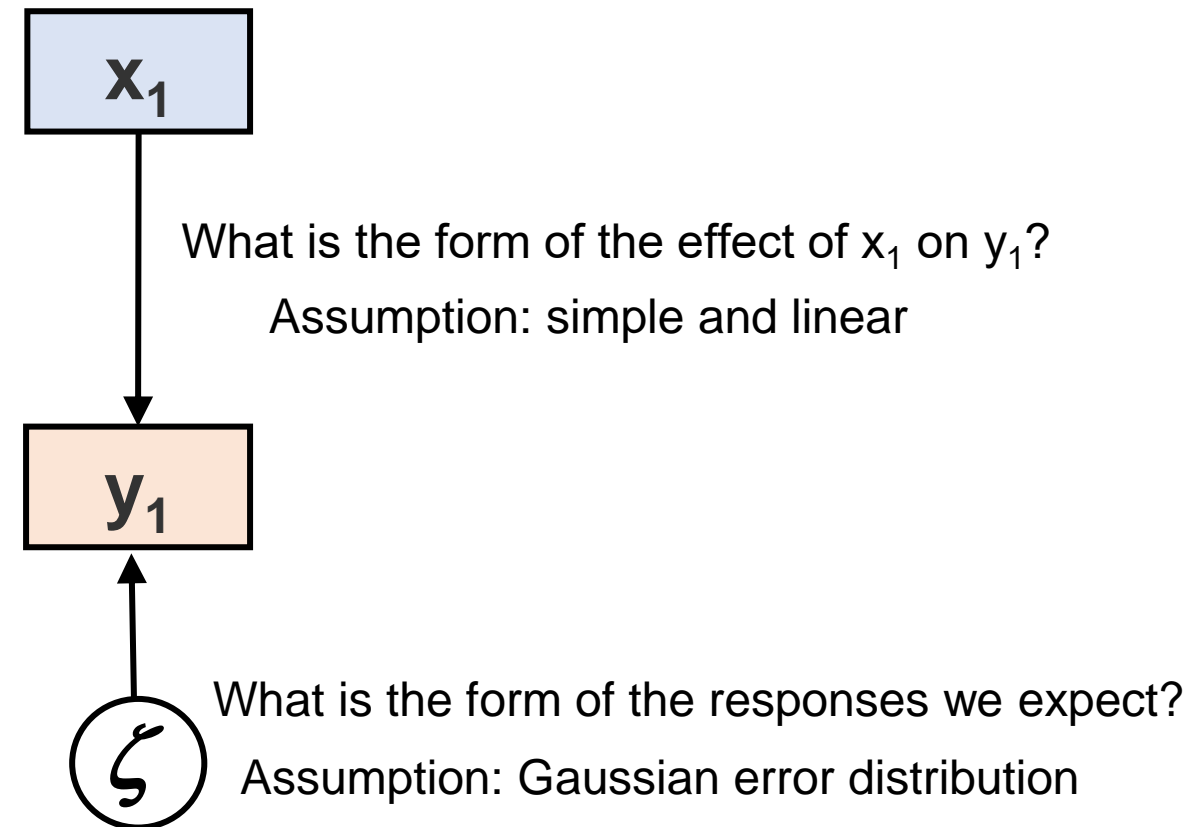 - Adjusting for Violated Assumptions

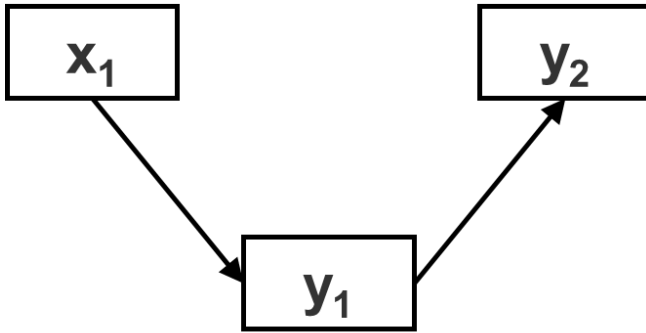# Assumptions

Two Major Assumptions of Covariance-Based Estimation:

1. **Residuals** are <u>normal</u>

2. **Data** are <u>multivariate normal</u>

1. Residuals are normal

- This is a linear modeling technique

- Assumption of Gaussian error distribution

- Violations require corrections

x_1

What is the form of the effect of x_1 on y_1?

Assumption: simple and linear

y_1

ζ

What is the form of the responses we expect?

Assumption: Gaussian error distribution

Assumptions

Test the distribution of residuals

```
# Read data
library(tidyverse)

data1 <-read_tsv("Data/SEMdata1.txt")


# SEM model in lavaan

library(lavaan)

sem_mod1 <- ' y1 ~ x1
              y2 ~  y1
'
```
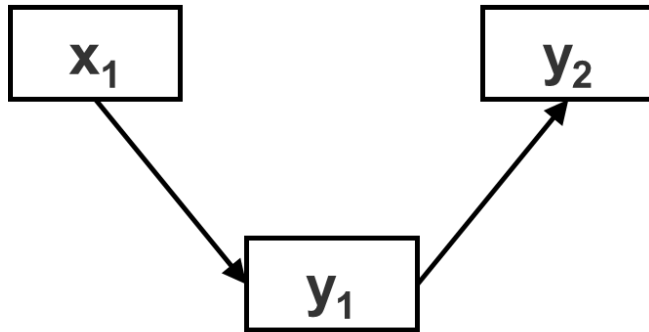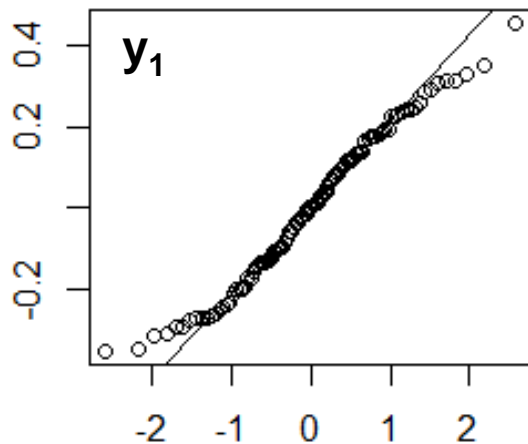
Assumptions

Test the distribution of residuals

Normal Q-Q Plot

y₁
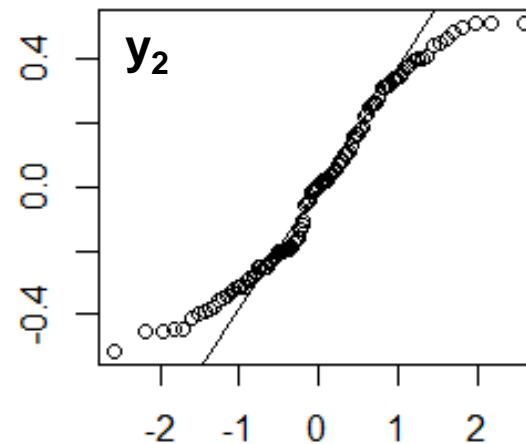

Theoretical Quantiles

Normal Q-Q Plot

y₂

Theoretical Quantiles

```
# SEM model in lavaan
sem_mod1 <- ' y1 ~ x1
              y2 ~  y1
'

# get casewise residuals
mod1 <-lm(y1 ~ x1, data1)
mod2 <- lm (y2 ~  y1, data1)

res_y1 <- resid(mod1)
res_y2 <- resid(mod2)

# Q-Q Plots
qqnorm(res_y1)
qqline(res_y1)
qqnorm(res_y2)
qqline(res_y2)
```
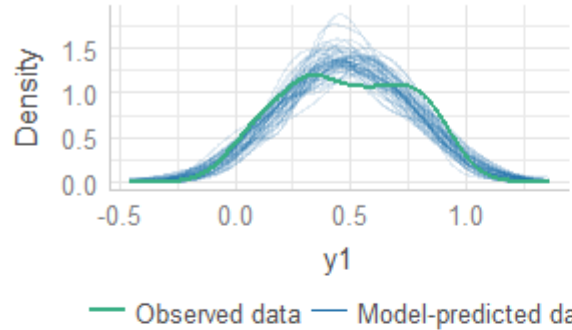
7

Assumptions

Posterior Predictive Check
Model-predicted lines should resemble observed data

Linearity
Reference line should be flat and horizontal
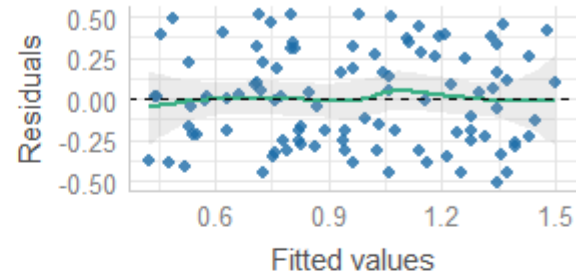
Homogeneity of Variance
Reference line should be flat and horizontal
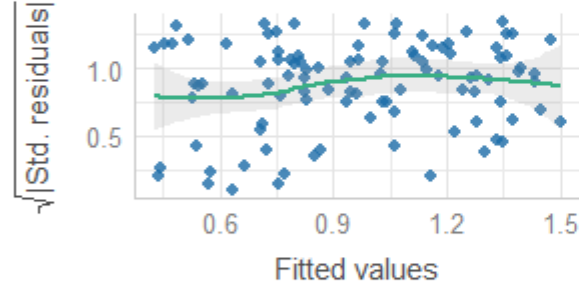
Influential Observations
Points should be inside the contour lines
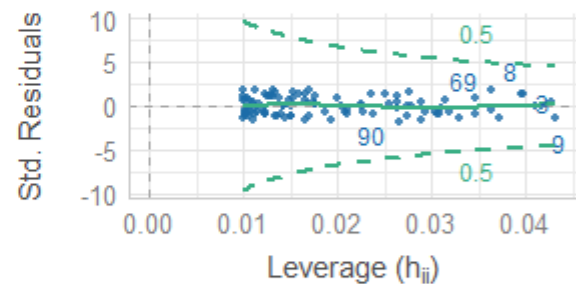
Normality of Residuals
Dots should fall along the line
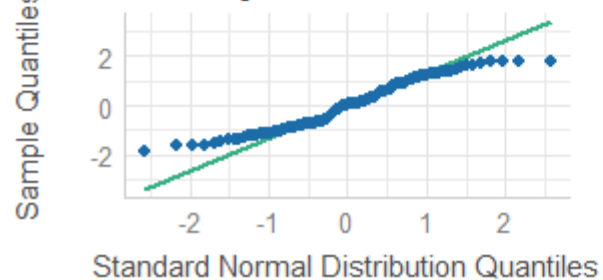
mod1 <-lm(y1 ~ x1, data1)

```
# or

library(performance)

x11(height=20,width=20)

check_model(mod1)
# and
check_model(mod2)
```
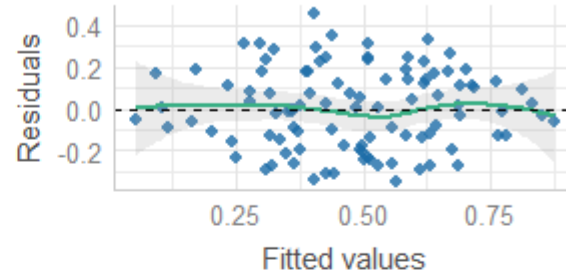
Assumptions

Posterior Predictive Check
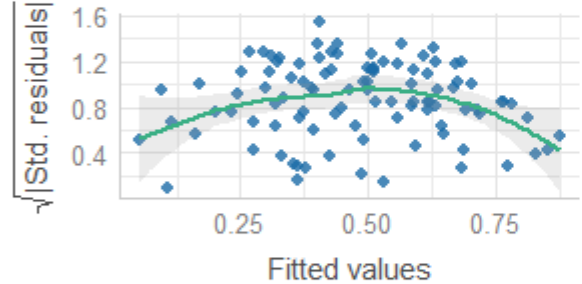Model-predicted lines should resemble observed da

Linearity
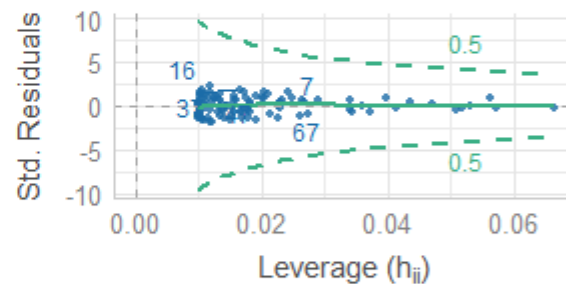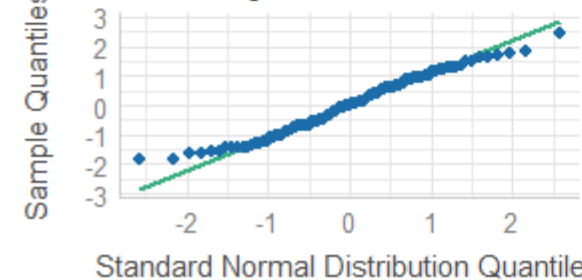Reference line should be flat and horizontal

Homogeneity of Variance
Reference line should be flat and horizontal

Influential Observations
Points should be inside the contour lines

Normality of Residuals
Dots should fall along the line

```
# or

library(performance)

x11(height=20,width=20)

check_model(mod1)
# and
check_model(mod2)
```

mod2 <- lm (y2 ~ y1, data1)
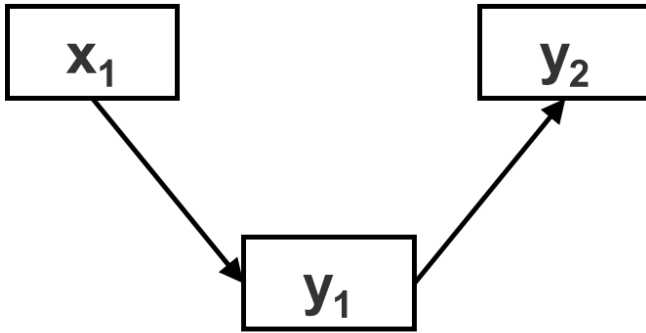
 Other options:

Multivariate Shapiro-Wilks Test

```
# Multivariate Test with Shapiro-Wilks
library(mvnormtest)

res <- cbind(res_y1, res_y2)


mshapiro.test(t(res))
```

Assumptions
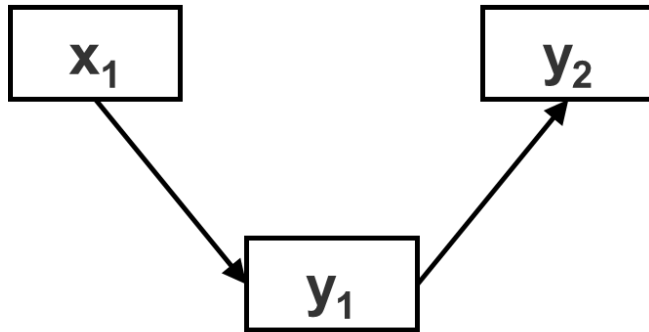
 Other options:

Multivariate Shapiro-Wilks Test

```
# Test with Shapiro-Wilks
library(mvnormtest)

res <- cbind(res_y1, res_y2)

mshapiro.test(t(res))


    >

            Shapiro-Wilk normality test

    data:  Z
    W = 0.98828, p-value = 0.5288
```

Often too sensitive of a test

Residuals seems fine

Assumptions

1. Residuals are normal

- This is a linear modeling technique

- Assumption of Gaussian error distribution

- Violations require corrections

  - Data transformation: e.g. log, square root

  - GLM: package **piecewiseSEM**

2. Data are multivariate normal

Multivariate normality - multiple normally distributed variables that have joint normal distribution (any linear combination of the variables is normally distributed).

Figures: wikipedia.org

13

Assumptions
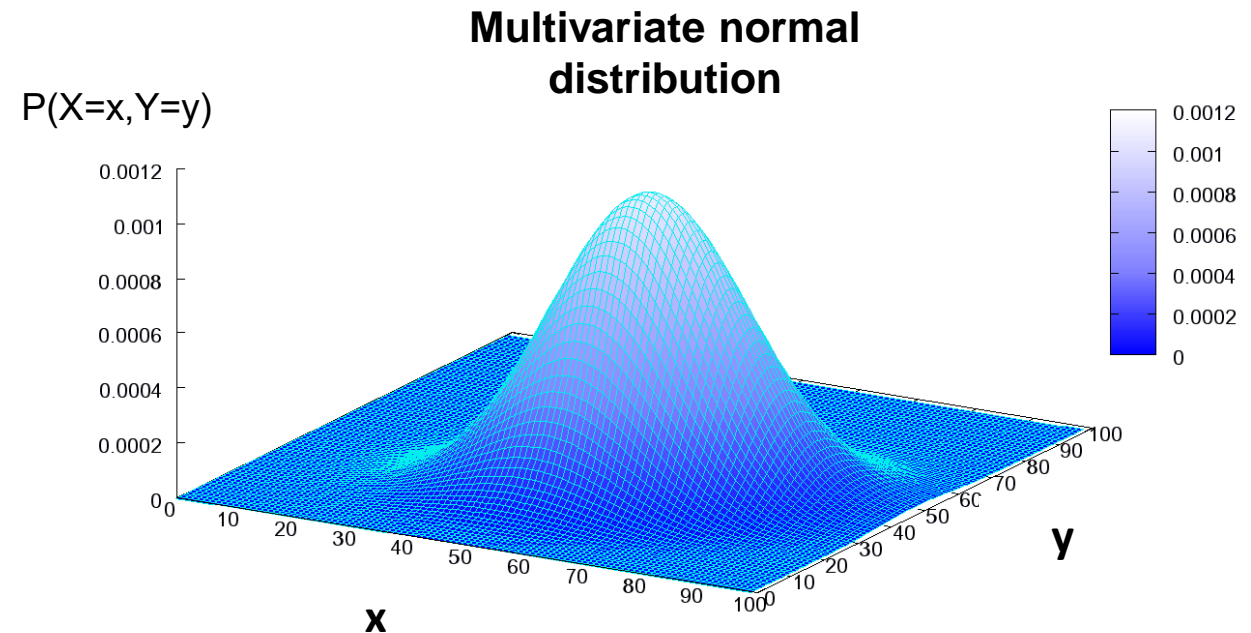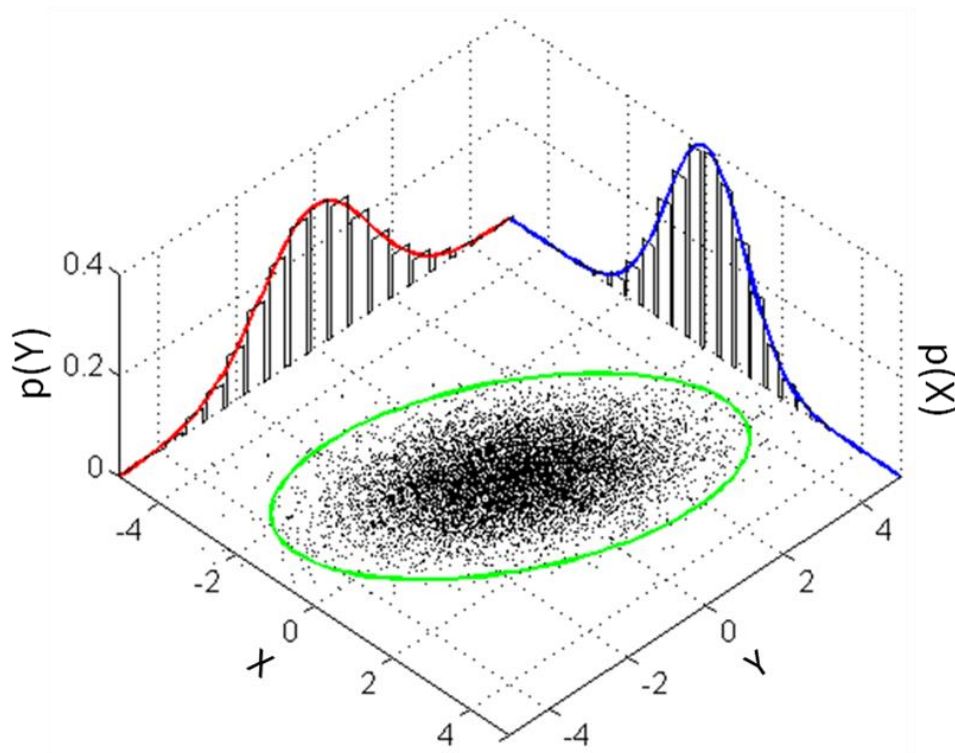
2. Data are multivariate normal

- We are fitting based on a covariance matrix:

  - the variables have a multivariate normal distribution.

- Fairly robust to violations

  (especially with increasing sample size)

- Severe violations result in

  - inflated test of model fit

  - underestimated parameter errors

x_1

y_2

y_1
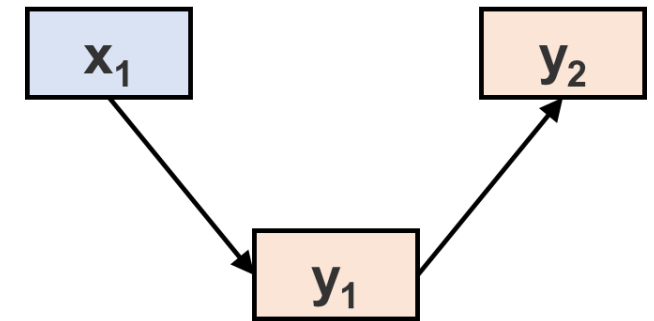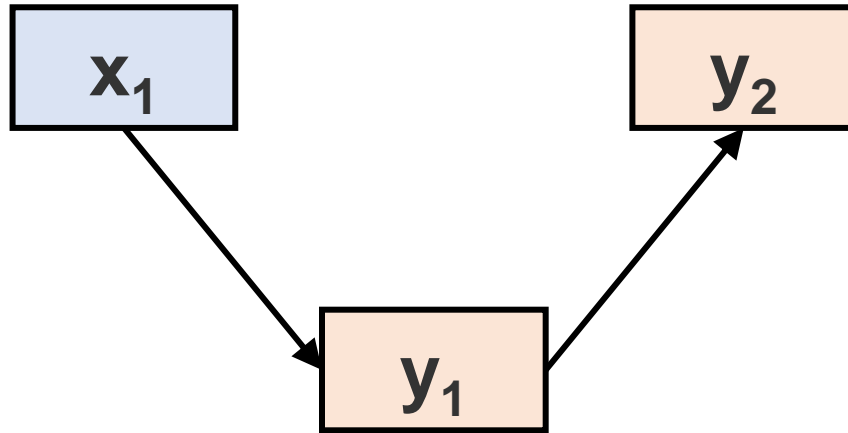
Are these Data Multivariate Normal?

x_1

y_2

y_1
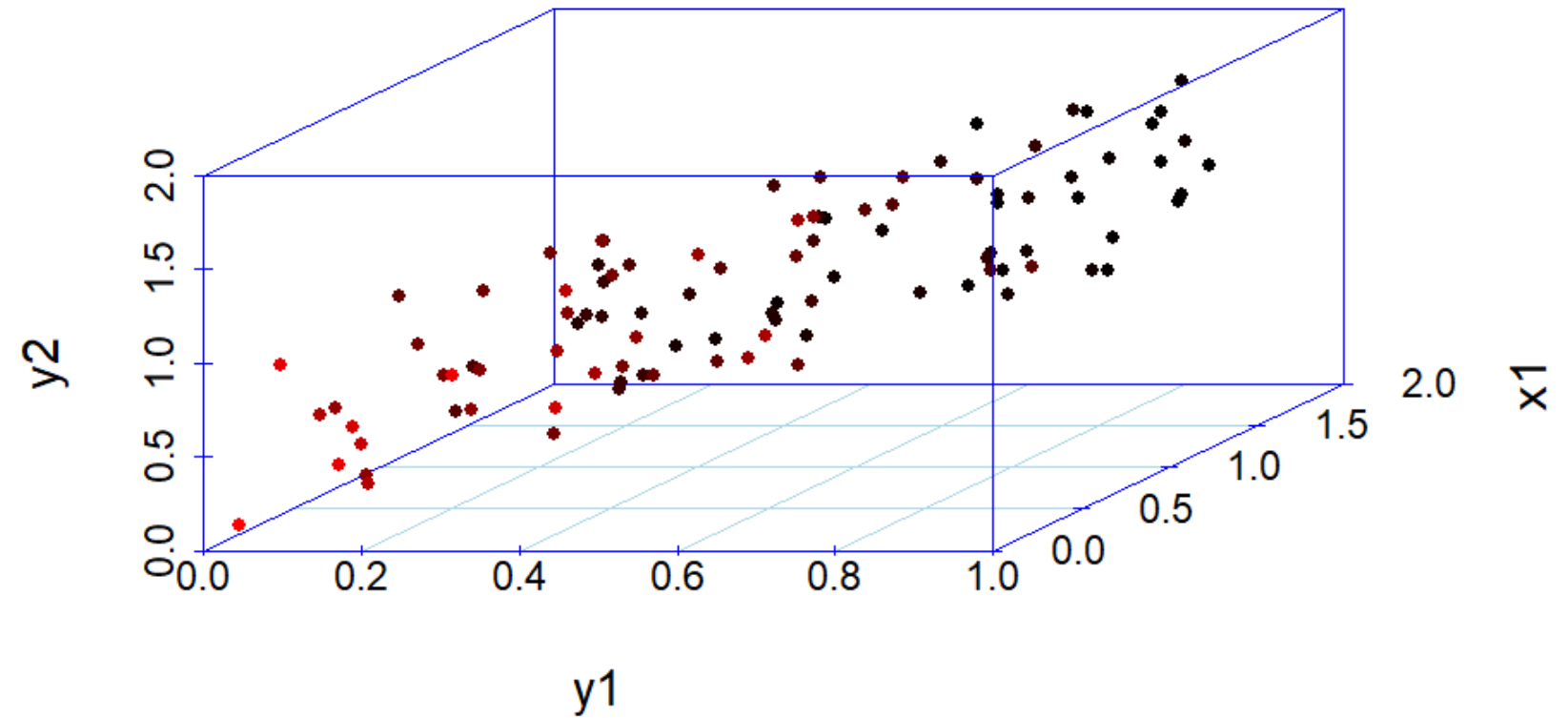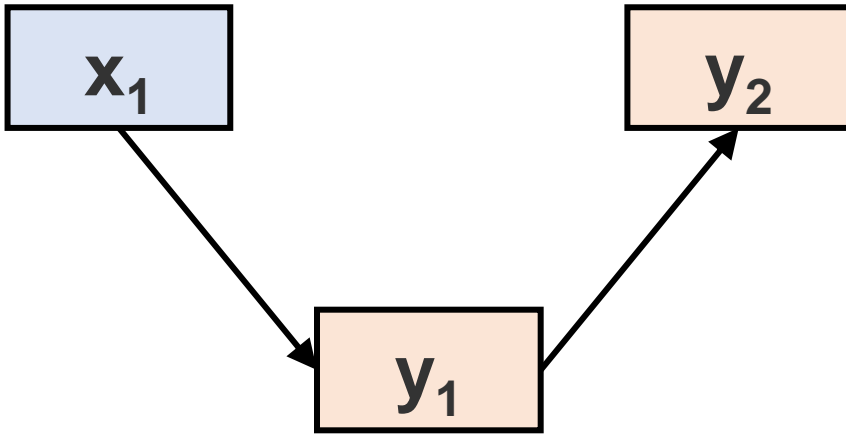
Multivariate Mardia's Test

```
library(MVN)
 mvn(data1,mvnTest="mardia")
```

Assumptions

```
$multivariateNormality
            Test          Statistic                    p value Result
1 Mardia Skewness   3.25985146525359      0.974630684427374      YES
2 Mardia Kurtosis  -3.33768747709889 0.000844787094267163       NO
3             MVN                <NA>                    <NA>    NO


$univariateNormality
              Test  Variable Statistic    p value Normality
1 Anderson-Darling       y1    1.0871     0.0072     NO
2 Anderson-Darling       x1    0.2286     0.8059     YES
3 Anderson-Darling       y2    0.2959     0.5878     YES
```

Multivariate normality of data

```
$multivariateNormality
            Test         Statistic                  p value Result
1 Mardia Skewness  3.25985146525359     0.974630684427374     YES
2 Mardia Kurtosis -3.33768747709889 0.000844787094267163      NO
3             MVN              <NA>                  <NA>      NO


$univariateNormality
             Test  Variable Statistic    p value Normality
1 Anderson-Darling     y1      1.0871     0.0072     NO
2 Anderson-Darling     x1      0.2286     0.8059     YES
3 Anderson-Darling     y2      0.2959     0.5878     YES
```

```
# Shapiro-Wilk Univariate normality test
mvn(newdata,mvnTest="mardia", univariateTest="SW")
?mvn()
```

Multivariate normality of **data**

```
# plots for Multivariate Normality

mvn(data1,multivariatePlot="qq")

mvn(data1, multivariateOutlierMethod="quan")
```

Chi-Square Q-Q Plot

Chi-Square Q-Q Plot

Assumptions

Normal Q-Q Plot (y1)

Normal Q-Q Plot (x1)


```
# univariate plots

mvn(data1,univariatePlot="qqplot")

par(mfrow=c(1,1))
```

Normal Q-Q Plot (y2)


```
> mvn(data1,mvnTest="mardia", univariateTest="SW")

$univariateNormality
          Test  Variable  Statistic   p value  Normality
1 Shapiro-Wilk    y1        0.9593     0.0036      NO
2 Shapiro-Wilk    x1        0.9909     0.7353      YES
3 Shapiro-Wilk    y2        0.9885     0.5472      YES
```

Assumptions

My data is not normal!

- This can just be a feature of the data, and residuals may still be normal.

- Severe violations result in:

    - incorrect standard errors

    - inaccurate χ^2

- There are techniques to get unbiased fit and error statistics.

    - The Satorra-Bentler Chi Square Test

    - Bollen-Stine Bootstrap

Assumptions

The Satorra-Bentler Chi Square:

- Correction coefficient for χ^2 and Standard Errors
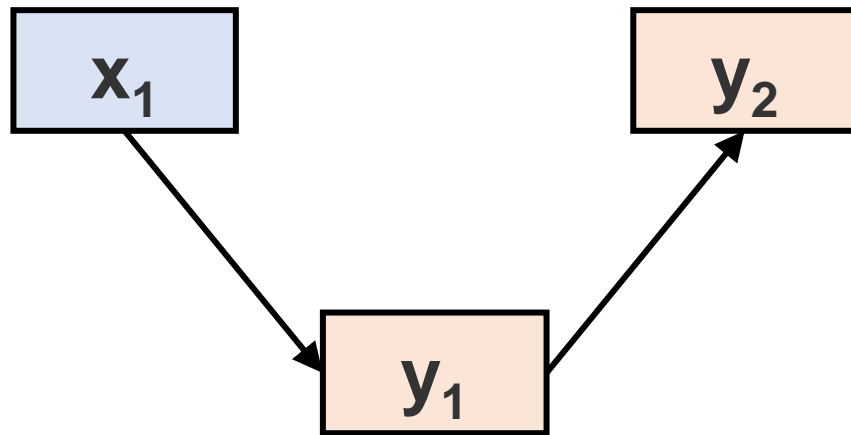

```
# Model specification in lavaan
sem_mod1 <- '
    y2 ~ y1
    y1 ~ x1
`
# MLM estimation with robust SE and/or test statistic
sem.fit1 <- sem(sem_mod1, data=data1,
                estimator="MLM", se="robust")
# or
sem.fit1 <- sem(sem_mod1, data=data1,
                test="Satorra-Bentler")
```

The Satorra-Bentler Chi Square:

- Correction coefficient for χ^2 and Standard Errors

```
> summary(sem.fit1, standardize = T)


Model Test User Model:

                                   Standard        Robust

  Test Statistic                      1.064         1.269
  Degrees of freedom                      1             1
  P-value (Chi-square)                0.302         0.260
  Scaling correction factor                         0.838
        Satorra-Bentler correction
```

23

The Satorra-Bentler Chi Square:

- Correction coefficient for χ^2 and Standard Errors

```r
# Model specification in lavaan
sem_mod1 <- '
    y2 ~ y1
    y1 ~ x1
`
# MLM estimation with robust SE and/or test statistic
sem.fit1 <- sem(sem_mod1, data=data1,
                estimator="MLM", se="robust")


summary(sem.fit1, fit.measures=TRUE)
```

The Satorra-Bentler Chi Square:

- Correction coefficient for χ^2 and Standard Errors

```
summary(sem.fit1,fit.measures=TRUE)

...

   Robust Comparative Fit Index (CFI)                    0.998
   Robust Tucker-Lewis Index (TLI)                       0.995
...
Root Mean Square Error of Approximation:

...
Robust RMSEA                                             0.048
   90 Percent confidence interval - lower                0.000
   90 Percent confidence interval - upper                0.254


Standardized Root Mean Square Residual:
   SRMR                                      0.021        0.021
```

Bollen-Stine Bootstrap

```
# Model specification in lavaan
sem_mod1 <- '
    y2 ~ y1
    y1 ~ x1
`

# Bootstapping
sem.fit1 <- sem(sem_mod1, data=data1,
                test="bollen.stine", se="bootstrap",
                    bootstrap=1000)
```

Typically want ~ 1000 bootstrap replicates

Assumptions

Bollen-Stine Bootstrap

```
# Bollen-Stine Bootstrap results
Model Test User Model:


  Test statistic                                        1.064
  Degrees of freedom                                        1
  P-value (Chi-square)                                  0.302


  Test statistic                                        1.064
  Degrees of freedom                                        1
  P-value (Bollen-Stine bootstrap)                      0.509


Parameter Estimates:


  Standard errors                                   Bootstrap
  Number of successful bootstrap draws                   1000
```

Assumptions

Assumptions:

1. Residuals are normal

2. Data are multivariate normal

3. No missing data

- NA in data bias parameter estimates

```
# Checking for missing data
> which(is.na(data1))
integer(0)
```

```
# Full-information maximum likelihood (FIML) estimation
# adjusting for incomplete data (if any)
sem(sem_mod1, data=data1, missing="fiml")


# Adjusting for incomplete data and non-normality in data
sem(sem_mod1, data=data1, estimator="MLR", missing="fiml")
```

Assumptions

Assumptions:

1. Residuals are normal

2. Data are multivariate normal

3. No missing data

4. **No redundant variables**

- **Covariance matrix must be positive definite**

No singular determinants from high correlation (r=0.99) or when one variable is a linear function of another

ERROR **There are non-positive definite elements in the matrices!**

vif < 2 (no collinearity)

```
m1 <- lm(y1 ~ x1, data1)
m2 <- lm(y2 ~ x1 + y1, data1)
library(car)
vif(m2)
>
 x1        y1
1.907226 1.907226
```

Assumptions

Assumptions:

1. Residuals are normal

2. Data are multivariate normal

3. No missing data

4. No redundant variables

5. Sample size is sufficiently "large"

n sample size

Minimum requirement $$n = p \times 5$$

p number of path coefficients

Not sufficient sample size?

Try local estimation:
package ***piecewiseSEM***

Assumptions

Assumptions:

1. Residuals are normal

2. Data are multivariate normal

3. No missing data

4. No redundant variables

5. Sample size is sufficiently "large"

6. **Samples are independent**

For dependant (hierarchical) data use LMM or GLMM: package *piecewiseSEM*

Protocol for violated assumptions of covariance-based SEM

Violated assumptions	Steps for Corrections
Non-normality of Residuals	Data transformation: e.g. *log*, *square root*
	Local estimation with GLM: package **piecewiseSEM**
Data are not multivariate normal	MLM estimation with robust SE & test statistic: **library(lavaan) # Always report results for 'robust' test statistics** **sem(…, estimator="MLM", se="robust"** **#or test="Satorra-Bentler")**
	Bootstapping: **# Always report results for 'robust' test statistics** **library(lavaan)** **sem(…, test="bollen.stine", se="bootstrap)**
Missing data	Full information maximum likelihood: **library(lavaan)** **sem(…, missing="fiml") #for normal data** **sem(…, missing="fiml", estimator="MLR")#for non-normal data**
Positive definite S matrix	Check for multicolinearity in each single regression model: **library(car)** **vif(m2) # vif ≤ 2 (no collinearity)**
Dependant samples (hierarchical)	Local estimation with LMM or GLMM: package **piecewiseSEM**
Not sufficient sample size	Local estimation: package **piecewiseSEM**

Day 6 Task 1

California, USA.

Photos credit: USFS, and Jon Keeley, USGS

doi.org/10.1186/s42408-019-0041-0

doi.org/10.1071/WF07049

Postfire recovery of plant communities in California shrublands

A number of measures were taken, including:

- Vegetation cover **"cover"**

- Age of stands that burned **"age"**

- Fire severity **"firesev"**

```
# Keeley data
library(piecewiseSEM)
data(keeley)
```

Data: Grace, J.B. and Keeley, J.E. 2006. A structural equation model analysis of postfire plant diversity in California shrublands. Ecological Applications 16:503-514

Day 6 Task 1

Postfire recovery of plant communities in California shrublands

Other measurements:

- Vegetation species richness **"richness"**

- Local abiotic conditions (aspect, soils) **"abiotic"**

- Spatial heterogenity **"hetero"**

- Distance from coast **"distance"**

Measurements:

- Vegetation cover **"cover"**

- Age of stands that burned **"age"**

- Fire severity **"firesev"**

```
# Keeley data
library(piecewiseSEM)
data(keeley)
```

Data: Grace, J.B. and Keeley, J.E. 2006. A structural equation model analysis of postfire plant diversity
        in California shrublands. Ecological Applications 16:503-514

Day 6 Task 1

```
# Keeley data
library(piecewiseSEM)
data(keeley)
```

Other measurements:

- Vegetation species richness **"richness"**

- Local abiotic conditions (aspect, soils) **"abiotic"**

- Spatial heterogeneity **"hetero"**

- Distance from coast **"distance"**

Data: Grace, J.B. and Keeley, J.E. 2006. A structural equation model analysis of postfire plant diversity
    in California shrublands. Ecological Applications 16:503-514

Day 6 Task 1

abiotic
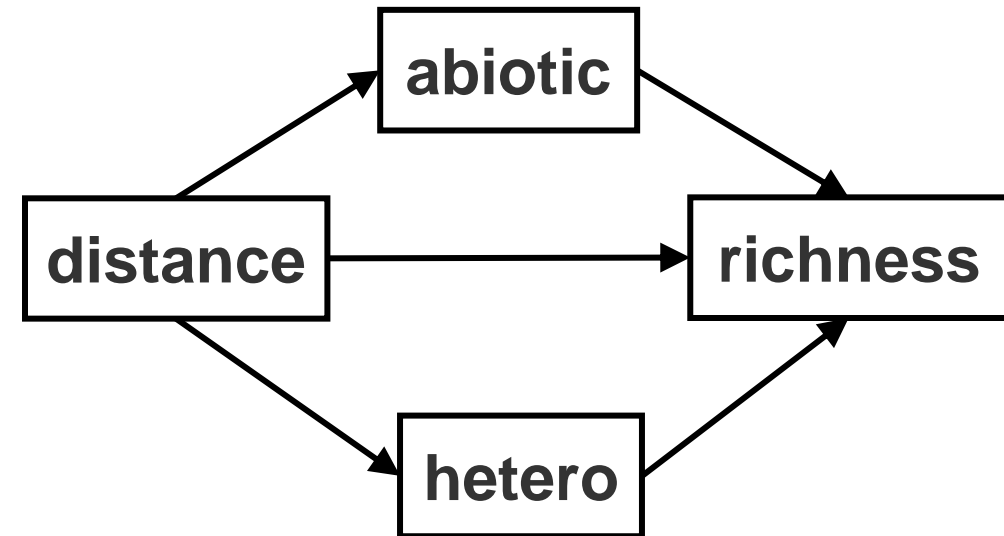
distance

richness

hetero

Other measurements:

- Vegetation species richness **"richness"**

- Local abiotic conditions (aspect, soils) **"abiotic"**

- Spatial heterogeneity **"hetero"**
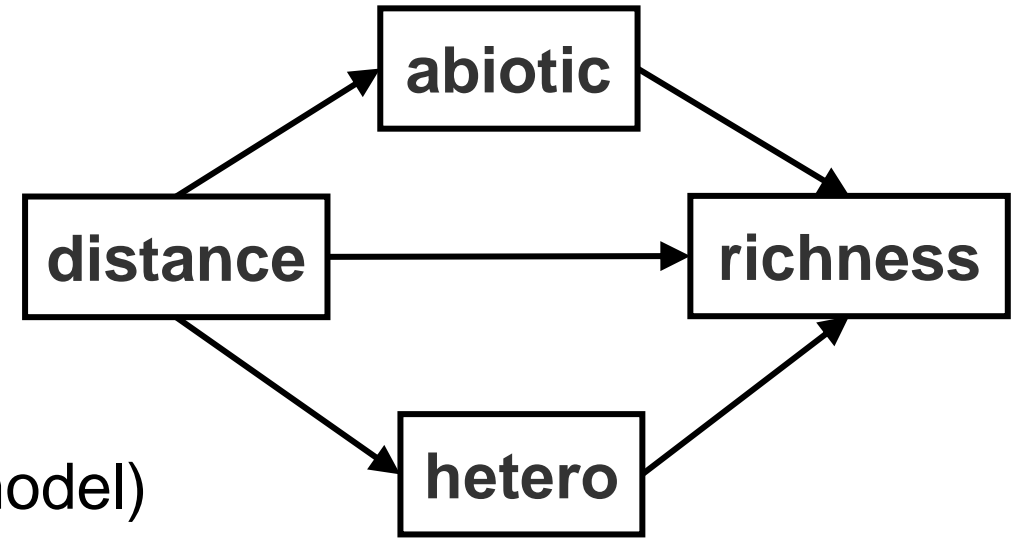
- Distance from coast **"distance"**
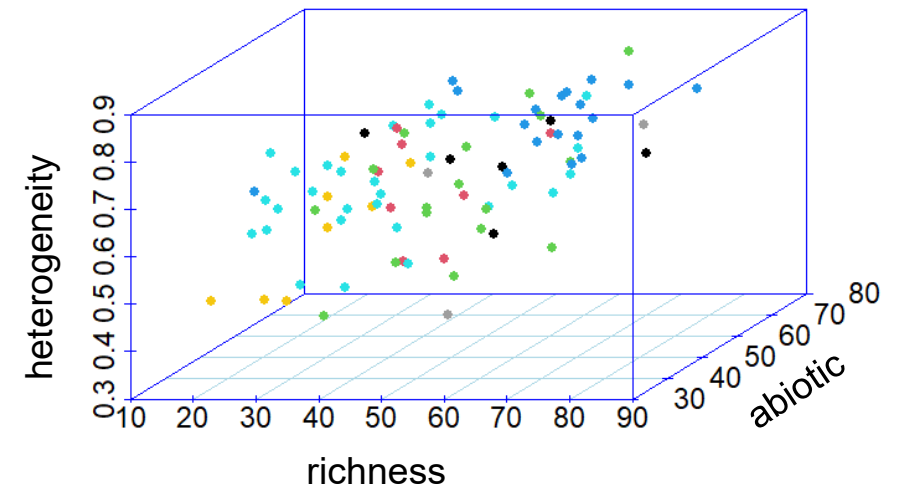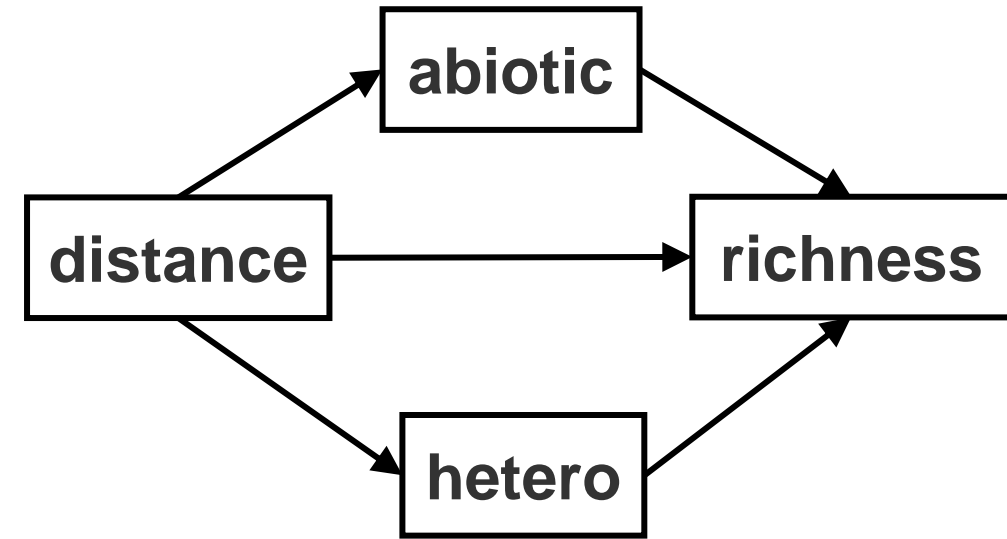
Day 6 Task 1

1. Specify the following model in lavaan

2. Check assumptions for covariance-based SEM

   • normality of residuals

   • multivariate normality of data

   • multicolinearity
     (function `vif(lm_model)` for each regression model)

3. Fit the model using data `data(keeley)`

4. Get the fit indices

5. Fill in Standardized Coeficients and R^2 for the model

6. Calculate indirect and total effects of distance on plant richness.
   What would you say about direct and indirect effects in this system?

Day 6 Task 1

When you fit the model

```
# Error about data scales
Warning message:
In lav_data_full(data = data, group = group, cluster = cluster,  :
  lavaan WARNING: some observed variances are (at least) a factor 1000 times larger than
others; use varTable(fit) to investigate
```

Day 6 Task 1

```r
# Call the model-implied covariance matrix


lavInspect(SemFit, "obs")$cov



  # Check the data scales

  varTable(SemFit)
```

Day 6 – Part 2

Outline

- Model Comparison in SEM

The concept of Goodness of Fit

1) **Are we ignoring important links?**

   Links = Processes

   When we are missing important paths:

   - our parameter estimates may be incorrect
   - our model is misspecified

Does the model fit the data?
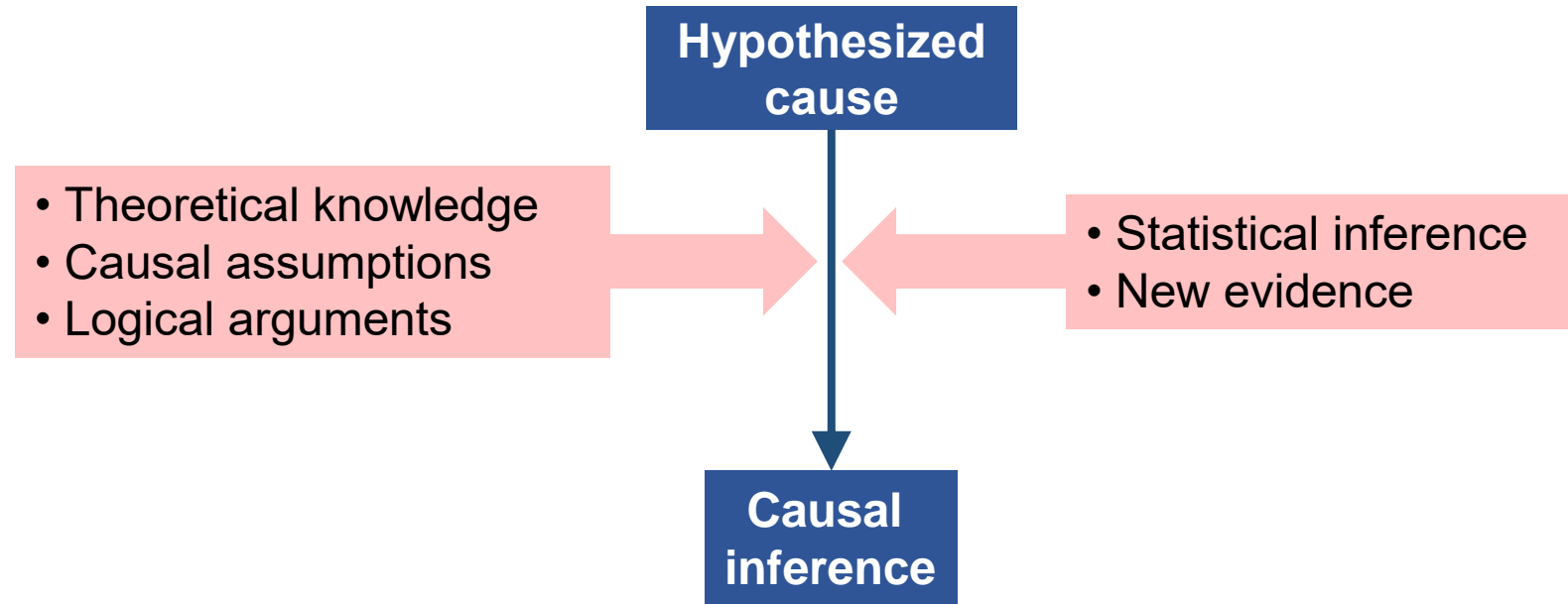
=

Does the model represent the data well?

=

Are we missing important information?

2) Are all the included links supported by the data?

Model Comparison

Model selection is a "Decision Problem"

Hypothesized cause

- Theoretical knowledge
- Causal assumptions
- Logical arguments

- Statistical inference
- New evidence

Causal inference

- SEM is not null hypothesis testing,

  but evaluating theory

- Decision should be based on

  1) a priory causal knowledge **<u>and</u>**

  2) the weight of data

Model Comparison

Question 1. Are we ignoring important links?

Model Comparison

California, USA.

Photos credit: USFS, and Jon Keeley, USGS

doi.org/10.1186/s42408-019-0041-0

doi.org/10.1071/WF07049

```
# Keeley data
library(piecewiseSEM)
data(keeley)
```
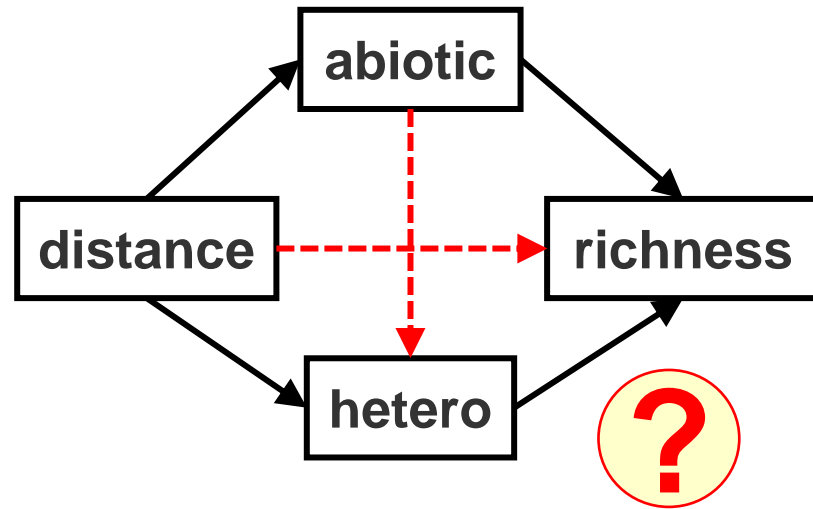
Data: Grace, J.B. and Keeley, J.E. 2006. Ecological Applications 16:503-514

Postfire recovery of plant communities in California shrublands

- Vegetation species richness **"richness"**

- Local abiotic conditions (aspect, soils) **"abiotic"**

- Spatial heterogenity **"hetero"**

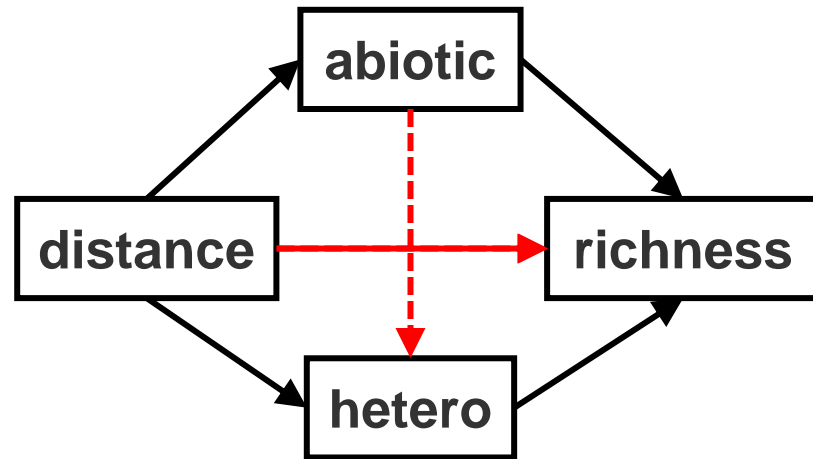- Distance from coast **"distance"**
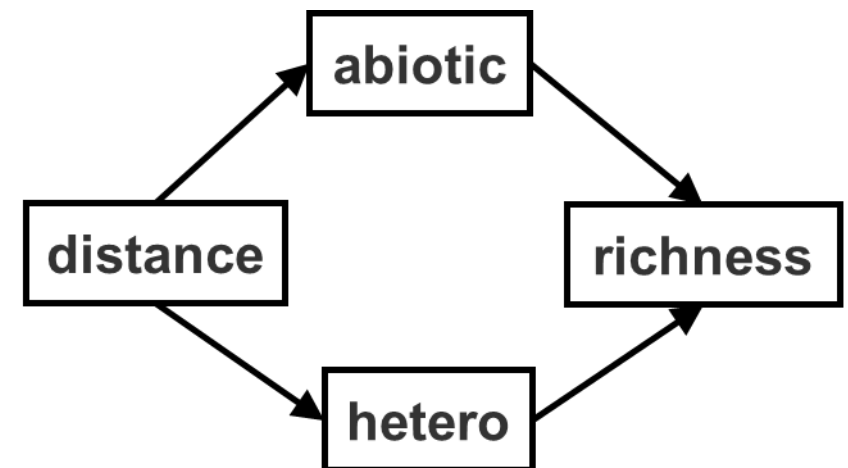
Model Comparison


```
library(piecewiseSEM)
data(keeley)
# Specify the model in lavaan
sem_m1 <- '
  rich ~ abiotic + hetero
  hetero ~ distance
  abiotic ~ distance'

# Fit the model
sem_fit1 <- sem(sem_m1, data=keeley)
summary(sem_fit1, standardize = T)
```

```
>
...
Model Test User Model:

  Test statistic                    17.831
  Degrees of freedom                     2
  P-value (Chi-square)               0.000
```
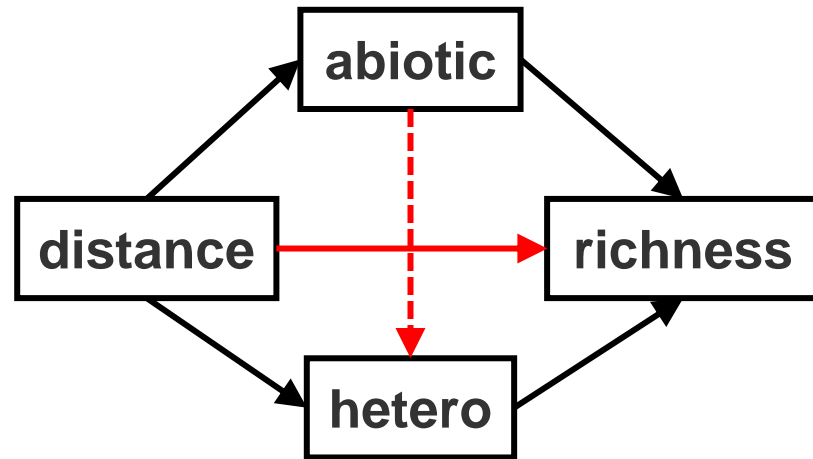
Model Comparison


```
# residual correlation

resid(sem_fit1, "cor")$cov

>

        rich  hetero abiotc distnc
rich    0.000
hetero  0.042 0.000
abiotic 0.032 0.118  0.000
distance 0.271 0.000  0.000  0.000
```

Large Residual Correlation for
"distance" – "rich"

Model Comparison

**Modification Indices
(Lagrange Multipliers)**


```
# Request modification indices

summary(sem_fit1, standardize = T,
                        modindices=T)
# or

modificationIndices(sem_fit1, standardized=F)
```

- the (approximate) amount that χ^2 would decrease due to including a path.

Model Comparison

Modification index

Expected parameter change

```
>
        lhs    op       rhs       mi      epc
8   distance  ~~  distance   0.000   0.000
9       rich   ~  distance  15.181   0.662
10    hetero   ~      rich   1.141  -0.002
11    hetero   ~   abiotic   1.792   0.002
12   abiotic   ~      rich   3.811  -0.196
13   abiotic   ~    hetero   1.792   8.933
14  distance   ~      rich  14.728   0.347
16  distance   ~   abiotic   0.000   0.000
```

Modification Indices (Lagrange Multipliers)

```
# Request modification indices
summary(sem_fit1, standardize = T,
                          modindices=T)
# or
modificationIndices(sem_fit1, standardized=F)
```

- **mi** - the (approximate) amount that χ^2 would decrease due to including a path.

- Looking for 'mi' values > 3.84

The "single-degree-of-freedom chi-square criterion" = 3.84

Model Comparison

Modification Indices
(Lagrange Multipliers)

```
>
          lhs   op     rhs     mi      epc
9        rich  ~ distance 15.181   0.662

12   abiotic  ~      rich   3.811  -0.196

14  distance  ~      rich  14.728   0.347
```
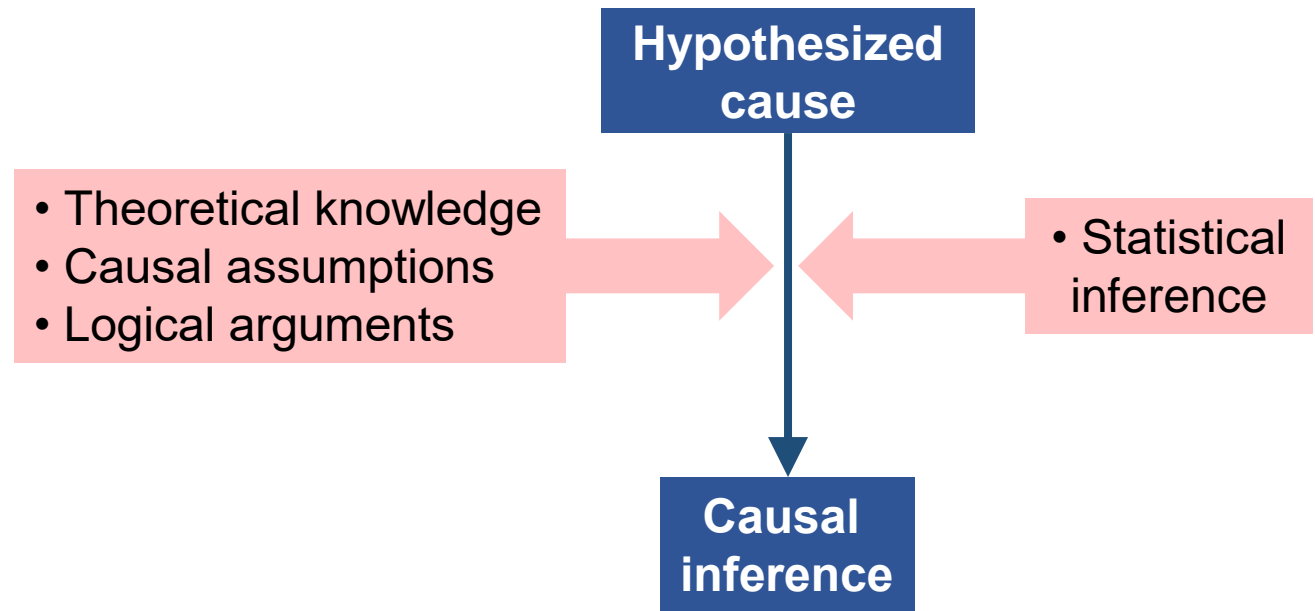
```
# Look for indices >3.84


modificationIndices(sem_fit, standardized=F,
                         minimum.value=3.8)
```

- **mi** - the (approximate) amount that χ^2 would decrease due to including a path.

- Looking for **mi** values > 3.84

- **Be very careful here for data dredging.**

Model Comparison

**Modification Indices
(Lagrange Multipliers)**

```
>

        lhs   op    rhs       mi       epc
9      rich  ~ distance    15.181    0.662
12   abiotic  ~      rich    3.811   -0.196
14 distance   ~      rich   14.728    0.347
```

- Theoretical knowledge
- Causal assumptions
- Logical arguments

- Statistical inference

Hypothesized cause

Causal inference

- **Not a good practice for confirmatory modelling**

- **Ok for exploratory modelling**

- Looking for **mi** values > 3.84

- **Be very careful here for data dredging.**

Model Comparison

Modification Indices (Lagrange Multipliers)

```
>

        lhs    op     rhs       mi      epc
9       rich  ~ distance 15.181   0.662
12   abiotic  ~         rich  3.811  -0.196
14  distance  ~         rich 14.728   0.347
```
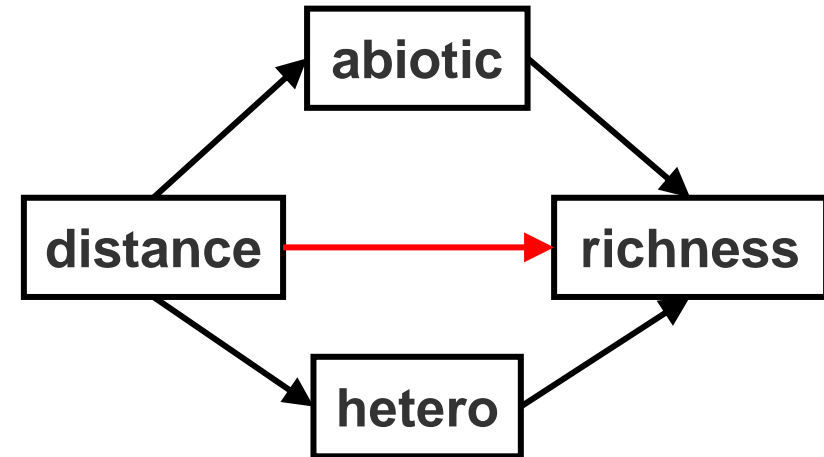

- **mi** - the (approximate) amount that χ^2 would decrease due to including a path.

- Looking for **mi** values > 3.84

- Be very careful here for data dredging.

Model Comparison
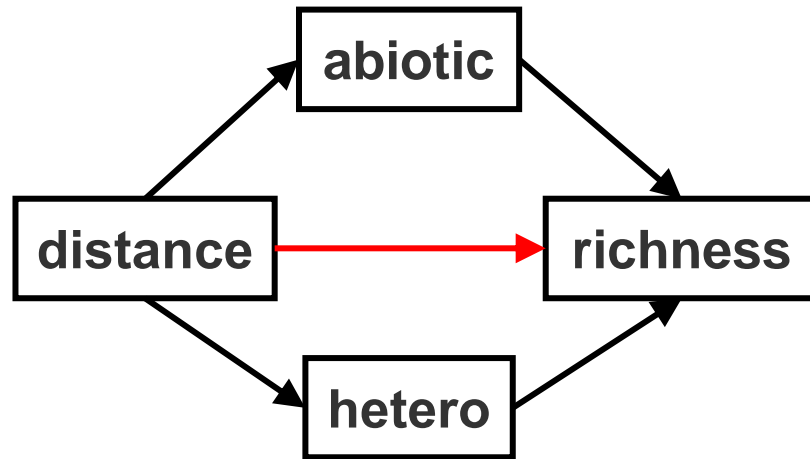

```
# Specify the model 2 in lavaan
sem_m2 <- '
  rich ~ abiotic + hetero + distance
  hetero ~ distance
  abiotic ~ distance`

# Fit the model
sem_fit2 <- sem(sem_m2, data=keeley)
summary(sem_fit2, standardize = T)
```

```
# Model 2
Model Test User Model:

  Test statistic                    1.810
  Degrees of freedom                    1
  P-value (Chi-square)              0.178
```

```
# Model 1


17.831
2
0.000
```

Model Comparison

Model 1

Model 2

The likelihood-ratio test
$$(\chi^2 - \text{difference test})$$

```
# The likelihood-ratio test
anova(sem_fit1, sem_fit2)
>

Chi-Squared Difference Test


        Df    AIC    BIC   Chisq   Chisq diff   Df diff    Pr(>Chisq)
sem_fit2  1 1155.3 1175.3  1.8104
sem_fit1  2 1169.3 1186.8 17.8307      16.02          1    6.267e-05 ***
```

Model Comparison

Question 1. Are we ignoring important links?
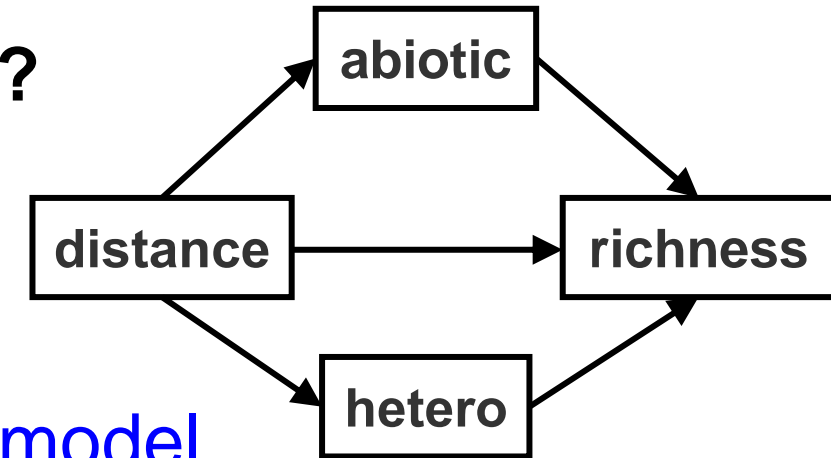
We are not missing major links in our model

Question 2. Are all the included links supported by the data?

Model Comparison

1. Does a simpler model still reproduce the more complex model's covariance matrix? Models are required to be nested.

   – Likelihood Ratio Testing (LR)

2. Compare the weight of evidence across multiple models (models are not required to be nested)

   – Information Theoretic Approaches (e.g. AIC)

The two can be used complementarily, however!

Model Comparison

How many parameters it takes to fit your model?

Remember the principle of **parsimony**
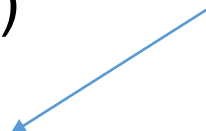
57

Model Comparison

Akaike information criterion (AIC)

- Based on model fit χ^2

- Parsimony adjusted for model complexity q

- Models are not required to be nested (like in LR test)

q number of estimated parameters in a model

$$AIC = \chi^2 + 2q$$

Model Comparison

Akaike information criterion (AIC)

- Based on model fit χ^2

- Parsimony adjusted for model complexity q

- Models are not required to be nested (like in LR test)

q number of estimated parameters in a model

$$AIC = \chi^2 + 2q$$

$$AIC_c = AIC + \left(\frac{2q(q+1)}{n-q-1} \right)$$

AIC_c

- Small Sample-Size Adjusted

- More complex parsimony correction term
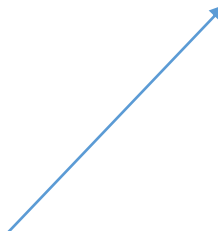
n number of samples
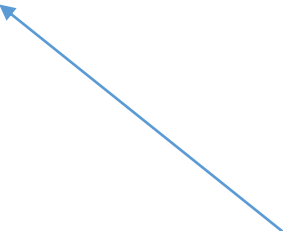
Model Comparison

Bayesian Information Criterion (BIC)

$$BIC = \chi^2 - DF \times log(n)$$

DF degrees of freedom

n number of samples

Model Comparison

AIC difference criteria

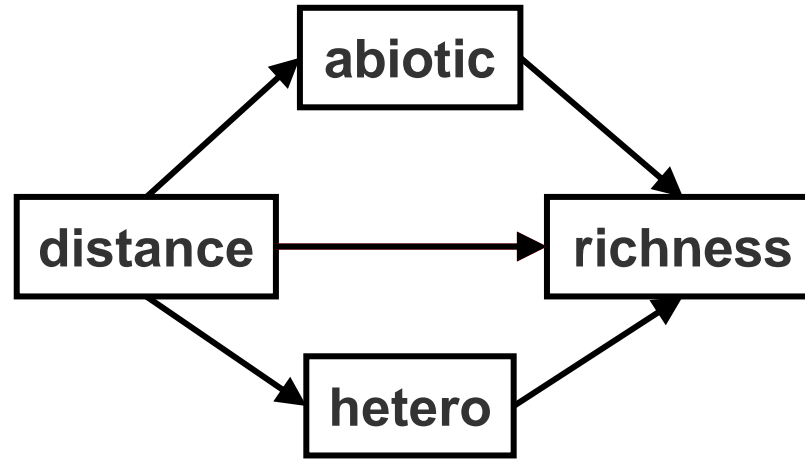| AIC difference	Support for equivalency of models
0-2	substantial
4-7	weak
> 10	none

Burnham, K.P. and Anderson, D.R. 2002. Model Selection and Multimodel Inference. Springer Verlag. (second edition), p 70.
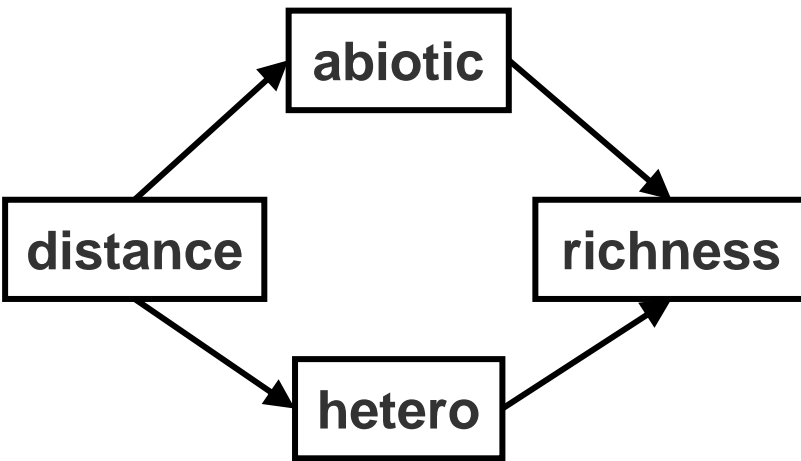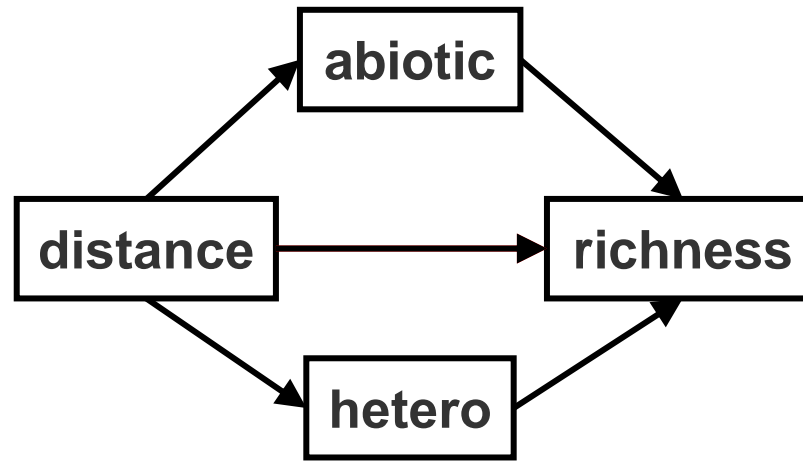
Model Comparison

Model 1. Fully mediated

Model 2.   Partially mediated

Model Comparison with AIC$_c$

Model Comparison

Model 1. Fully mediated

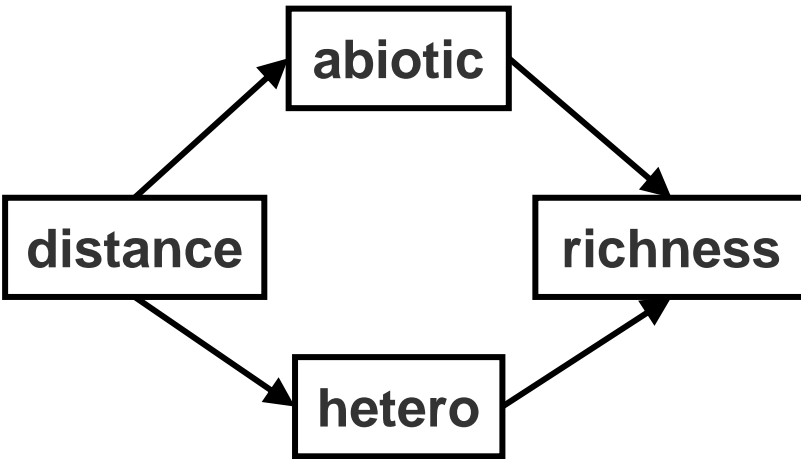Model 2. Partially mediated

Model Comparison with AIC_c

```
# Model selection based on AICc:

library(AICcmodavg)

aictab(cand.set = list(sem_fit1, sem_fit2),
        modnames = c("Full", "Partial"))

> Model selection based on AICc:

        K    AICc    Delta_AICc   AICcWt  Cum.Wt     LL
Partial 8   1157.05      0.00        1       1     -569.63
Full    7   1170.66     13.61        0       1     -577.64
```
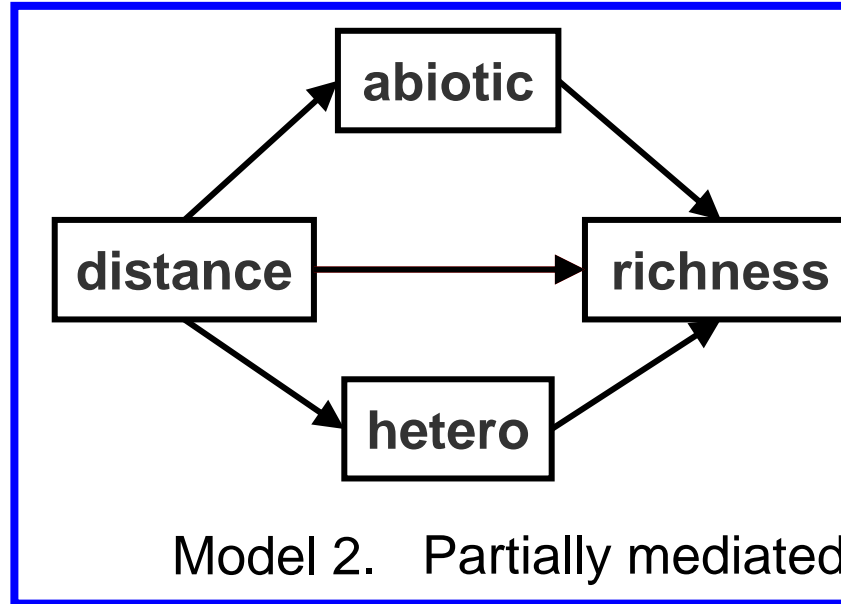
Model Comparison

Model 1. Fully mediated

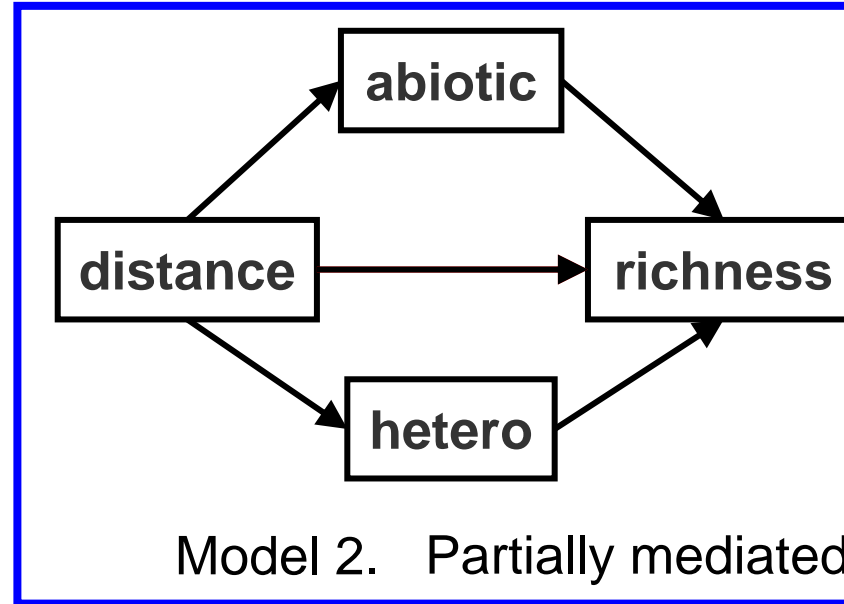Model 2.   Partially mediated

Model Comparison with AIC$_c$
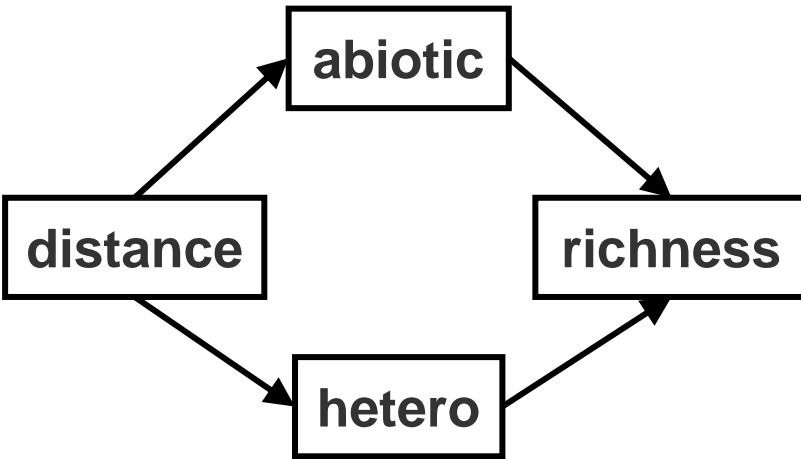
```
# Model selection based on AICc:

library(AICcmodavg)

aictab(cand.set = list(sem_fit1, sem_fit2),

        modnames = c("Full", "Partial"))

> Model selection based on AICc:
```

	K	AICc	Delta_AICc	AICcWt	Cum.Wt	LL
Partial	8	1157.05	0.00	1	1	-569.63
Full	7	1170.66	13.61	0	1	-577.64

64

Model Comparison

Model 1. Fully mediated

Model 2.   Partially mediated

Model Comparison with AIC_c

```
# Model selection based on AICc:

library(AICcmodavg)

aictab(cand.set = list(sem_fit1, sem_fit2),

        modnames = c("Full", "Partial"))

> Model selection based on AICc:
           K      AICc    Delta_AICc    AICcWt    Cum.Wt        LL
Partial    8    1157.05         0.00         1         1    -569.63
Full       7    1170.66        13.61         0         1    -577.64
```

Model Comparison

```r
# Model Comparison with AIC
aic <- AIC(sem_fit1, sem_fit2)
> aic
        df       AIC
sem_fit1  7 1169.29
sem_fit2  8 1155.27

# Calculate delta AIC
d_aic <- aic[2] - min(aic[2])
>           AIC
sem_fit1  14.02032
sem_fit2  0.00000

# Rank delta AIC in ascending order
library(dplyr)
arrange(d_aic, +AIC)
             AIC
sem_fit2  0.00000
sem_fit1 14.02032
```

**Model Comparison
with AIC**

66

Protocol for model comparison

Model type	Targeted question / aim	Test	Rule of thumb
Two nested models	Does a simpler model still reproduce the more complex model's covariance matrix?	Likelihood Ratio Testing (χ^2 – difference test) `anova(fit1, fit2)`	`Pr(>Chisq)` larger than .05 means that more complex model is significantly more accurate than the simpler model

Not nested models
 	
Multiple models
	Compare the weight of evidence across multiple models and / or non-nested models	Information Theoretic (IC Approaches:
		
		AIC (sample size > 250)
		`aic <- AIC(fit1, fit2)`
		`aic[2] - min(aic[2]) # Δ AIC`
		
		AICc (sample size $\leq$ 250)
		`library(AICcmodavg)`
		`aictab(cand.set = list(fit1, fit2),`
		`          modnames = c("mod1", "mod2"))`
		`# Δ AICc Is calculated in the output`
		
		BIC
		`bic <- BIC(sem_fit1, sem_fit2)`
		`# Δ BIC`
		`bic[2] - min(bic[2])`
			Δ IC (difference) criteria
			
			| IC difference
				Support for equivalency of models

---	---
0-2	substantial
4-7	weak
> 10	none	

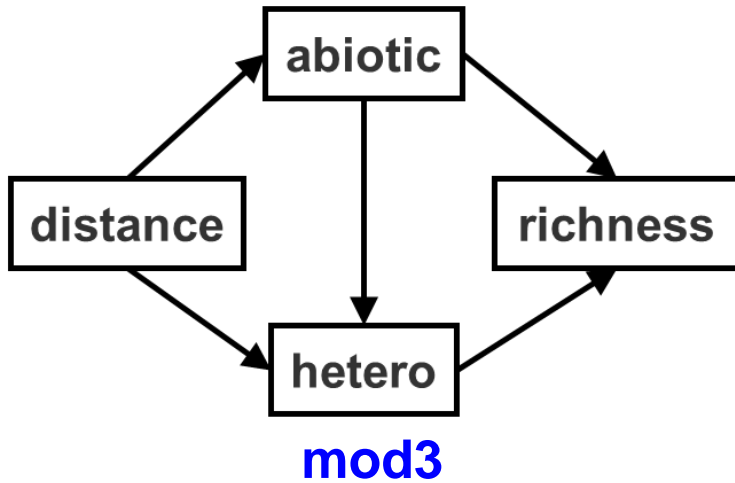# Model Comparison

Day 6 Task 2

Model Comparison

```
# Keeley data
library(piecewiseSEM)
data(keeley)
```


mod1

mod2

mod3

mod4

mod5
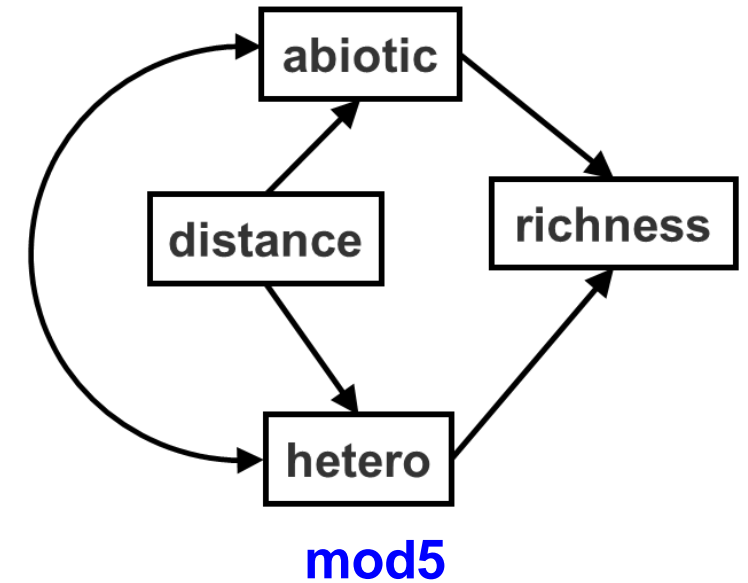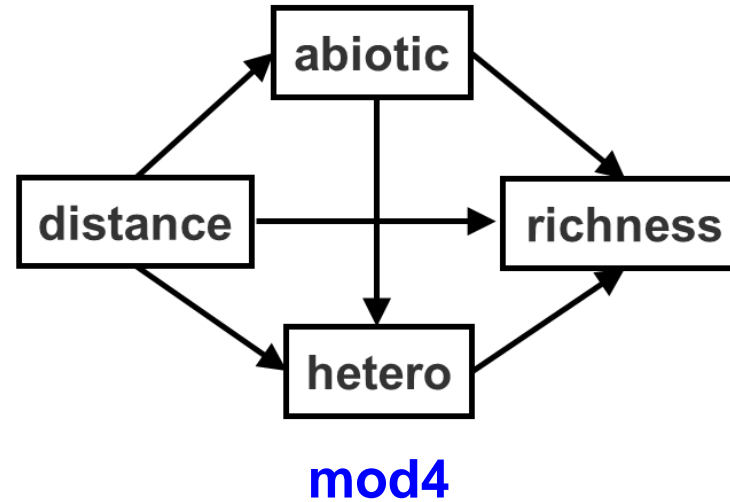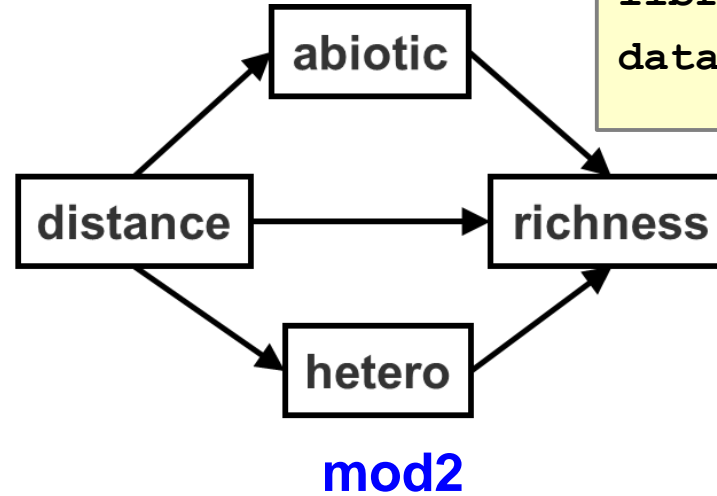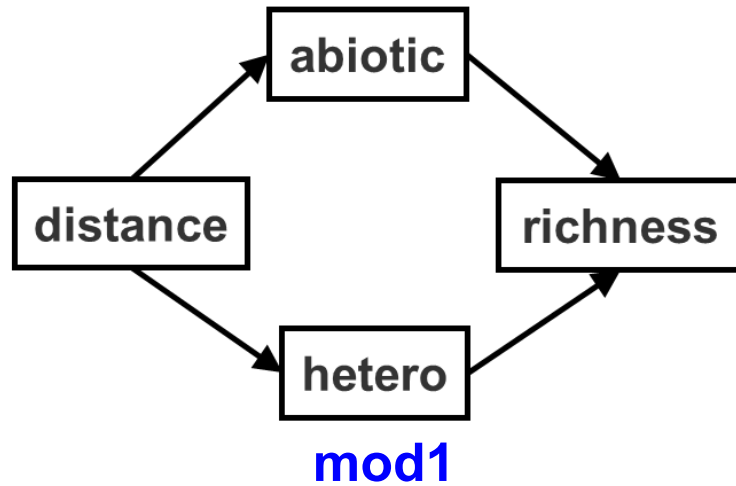
Day 6 Task 2

1. Think about the scientific meaning when we include the correlations among "hetero" and "abiotic" (in model 5)

2. Specify each model in lavaan

3. Fit each model using data `data(keeley)`

4. Compare the models using AICc and select the best model
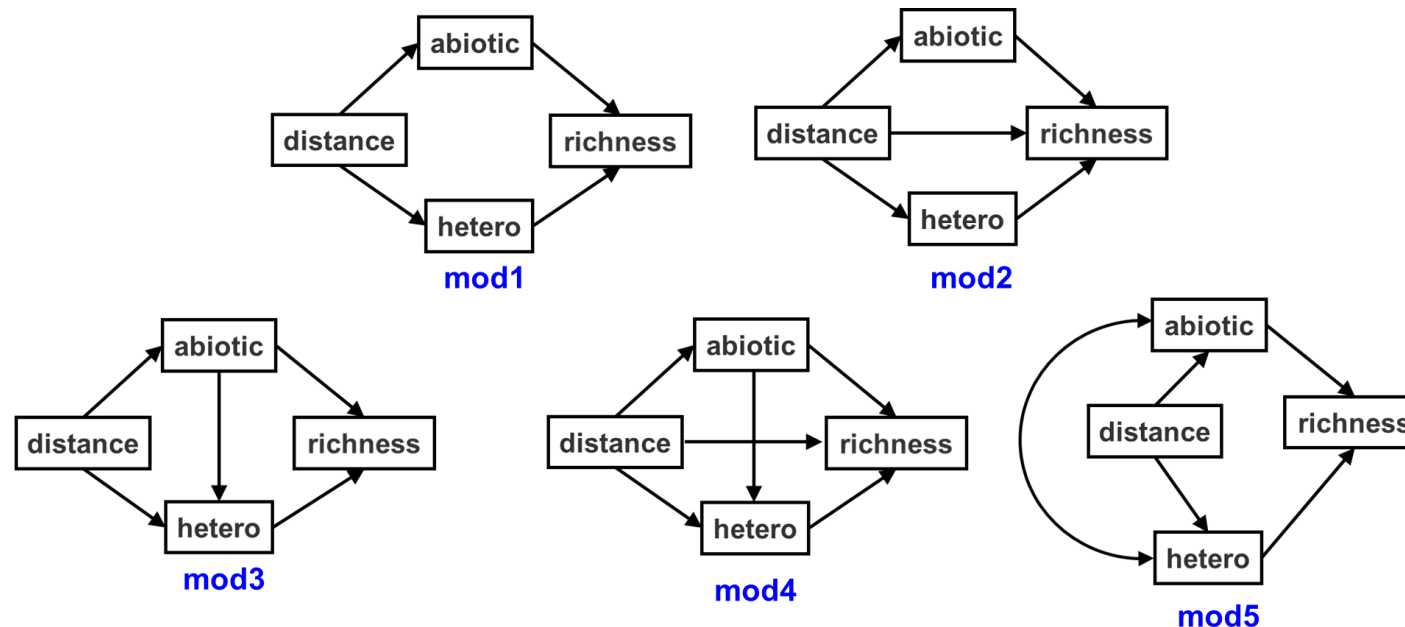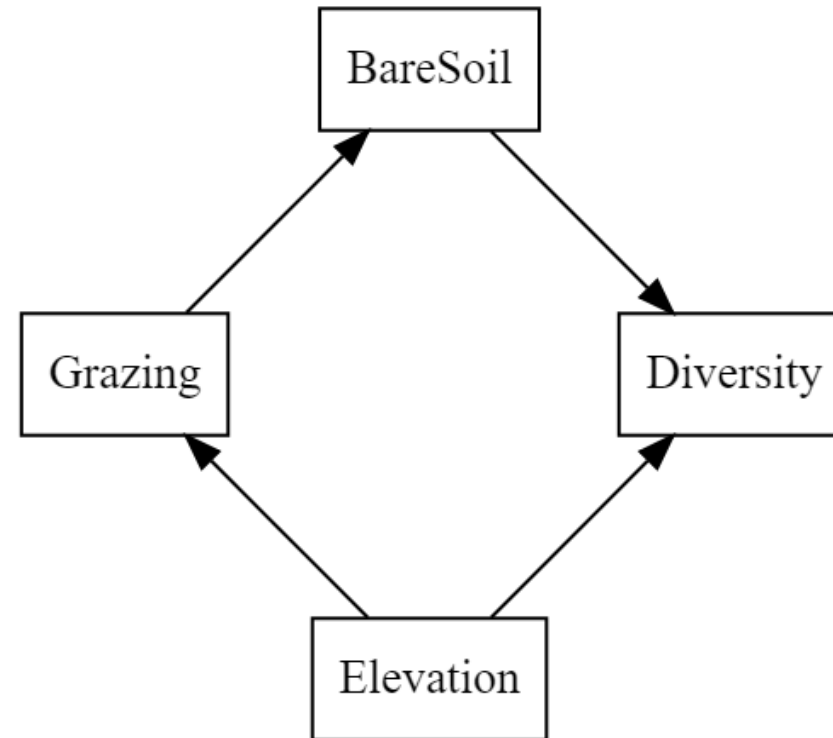
Test of Mediation

Day 6 Task 3

Day 6 Task 3

Test of mediation

Effects of grazing on plant diversity along elevation gradient


```
library(tidyverse)
# data
data <- read_csv("Grassl_data_2.csv")
```

Day 6 Task 3

For the model on Fig. 1:

1. For the variables on Fig 1 and data "Grassl_data_**2**.csv" check the assumptions of the covariance-based SEM

2. If there are any violations of the assumptions use the needed procedures to adjust for these violations while you fit the model (part 1 of this lecture) .

3. Get the fit indices.

4. Follow the procedures in this lecture (part 2) and test the following questions:

 - Are we ignoring important links?
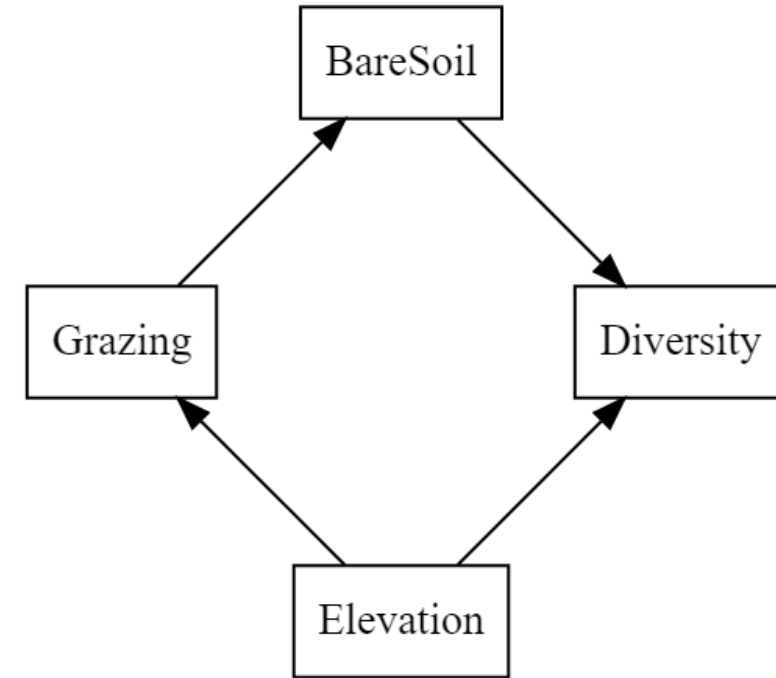
 - Are all the included links supported by the data?

Fig. 1

Protocol for violated assumptions of covariance-based SEM

Violated assumptions	Steps for Corrections
Non-normality of Residuals	Data transformation: e.g. *log*, *square root*
	Local estimation with GLM: package `piecewiseSEM`

Data are not multivariate normal	MLM estimation with robust SE & test statistic:
	`library(lavaan) # Always report results for 'robust' test statistics`
	`sem(…, estimator="MLM", se="robust`
	`        #or test="Satorra-Bentler")`

	Bootstapping: `# Always report results for 'robust' test statistics`
	`library(lavaan)`
	`sem(…, test="bollen.stine", se="bootstrap)`

Missing data	Full information maximum likelihood:
	`library(lavaan)`
	`sem(…, missing="fiml") #for normal data`
	`sem(…, missing="fiml", estimator="MLR")#for non-normal data`

Positive definite S matrix	Check for multicolinearity in each single regression model:
	`library(car)`
	`vif(m2) # vif ≤ 2 (no collinearity)`

Dependant samples (hierarchical)	Local estimation with LMM or GLMM: package `piecewiseSEM`
Not sufficient sample size	Local estimation: package `piecewiseSEM`

Recap: Protocol for model comparison

Model type	Targeted question / aim	Test	Rule of thumb
Two nested models	Does a simpler model still reproduce the more complex model's covariance matrix?	Likelihood Ratio Testing ($\chi^2 -$ difference test) `anova(fit1, fit2)`	`Pr(>Chisq)` larger than .05 means that more complex model is significantly more accurate than the simpler model

Not nested models	Compare the weight of evidence across multiple models and / or non-nested models	Information Theoretic (IC) Approaches:
		
		AIC (sample size > 250)
		
		`aic <- AIC(fit1, fit2)`
		
		`aic[2] - min(aic[2]) # Δ AIC`
		
		AICc (sample size $\leq$ 250)
		
		`library(AICcmodavg)`
		
		`aictab(cand.set = list(fit1, fit2),`
		`         modnames = c("mod1", "mod2"))`
		`# Δ AICc Is calculated in the output`
		
		BIC
		`bic <- BIC(sem_fit1, sem_fit2)`
		`# Δ BIC`
		`bic[2] - min(bic[2])`
			Δ IC (difference) criteria
			
			<table><tr><th>IC difference</th><th>Support for equivalency of models</th></tr><tr><td>0-2</td><td>substantial</td></tr><tr><td>4-7</td><td>weak</td></tr><tr><td>> 10</td><td>none</td></tr></table>

Multiple models			