# Introduction to structural equation modeling and mixed models in 

## Day 4 – Part 1: SEM

Oksana Buzhdygan

oksana.buzh@fu-berlin.de

# Day 4 – Part 1 Outline

- Assumptions of Covariance-Based Estimation
  - Adjusting for Violated Assumptions

# Assumptions

**Two Major Assumptions of Covariance-Based Estimation:**

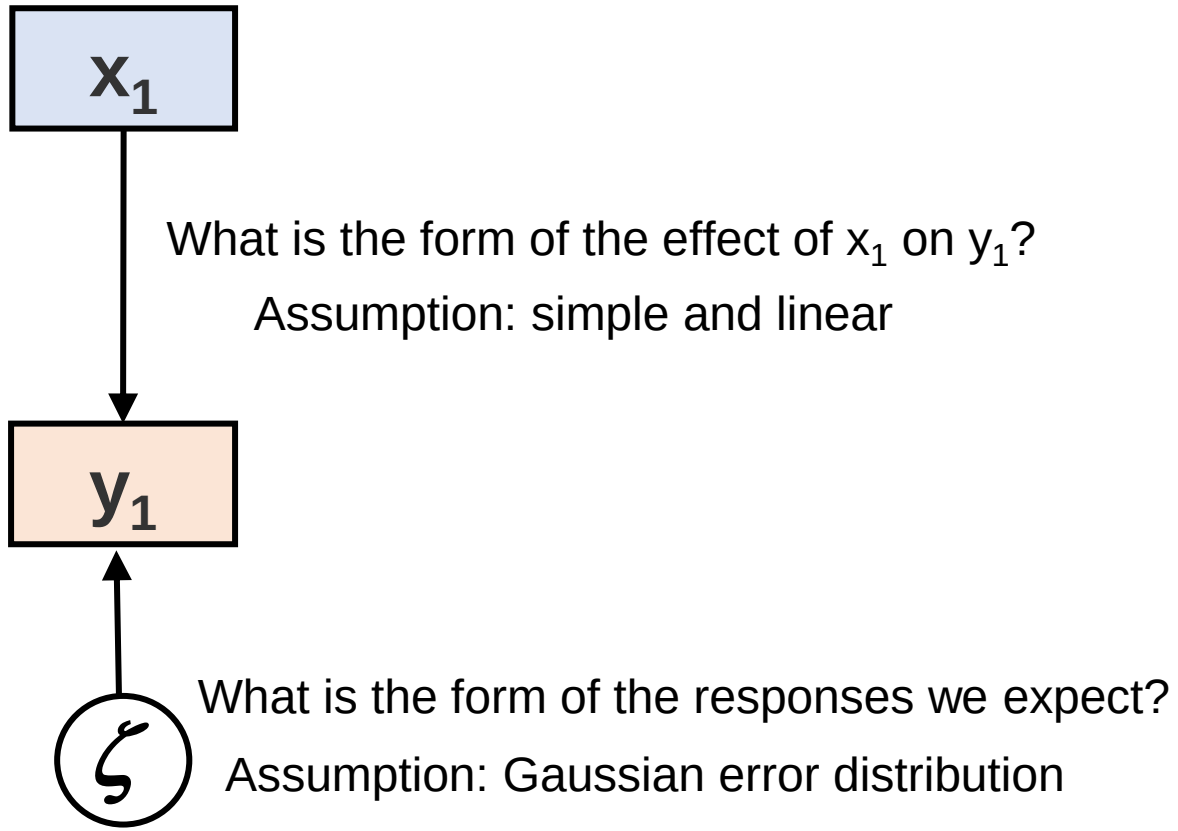
1. Residuals are normal
2. Data are multivariate normal

# Assumptions



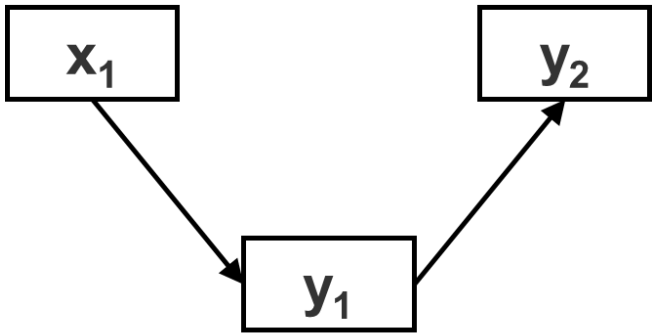
## 1. Residuals are normal

- This is a linear modeling technique
- Assumption of Gaussian error distribution
- Violations require corrections



# Assumptions

## Normality of Residuals

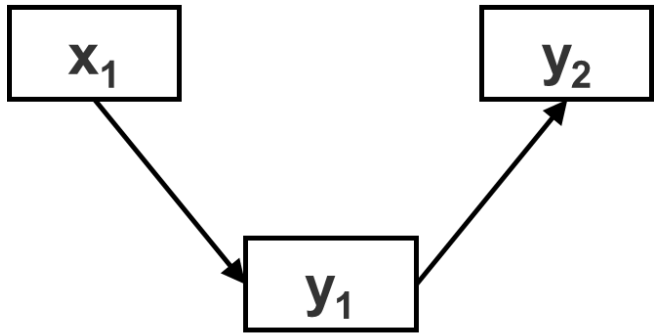



**Test the distribution of residuals**

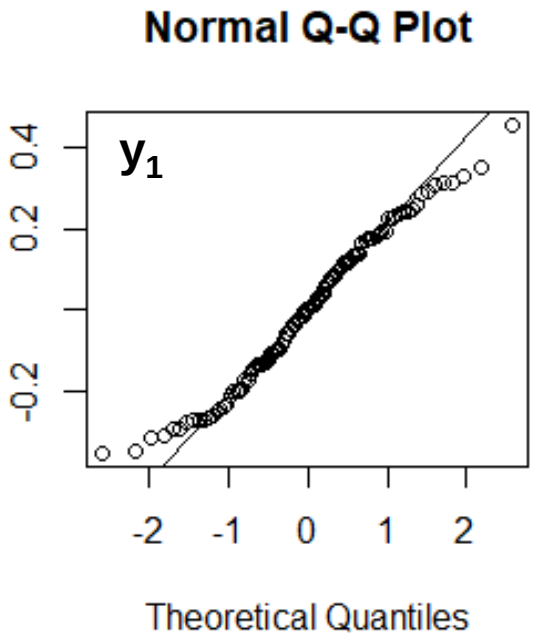
```
# SEM model in lavaan
sem_mod1 <- ' y1 ~ x1
              y2 ~  y1
'
```

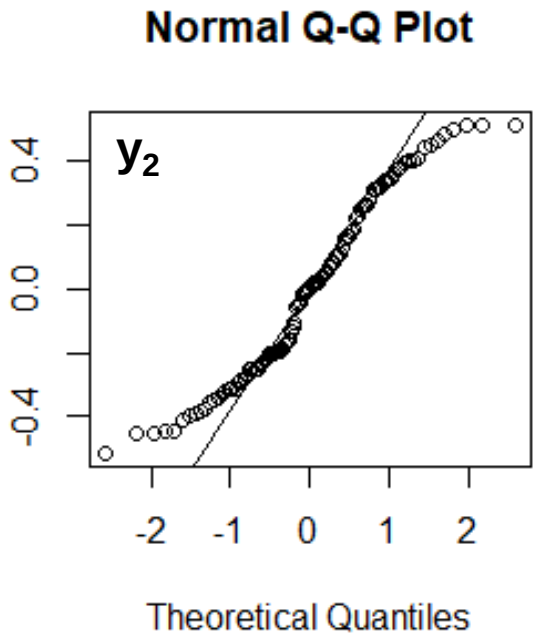
# Assumptions

## Normality of Residuals




## Test the distribution of residuals

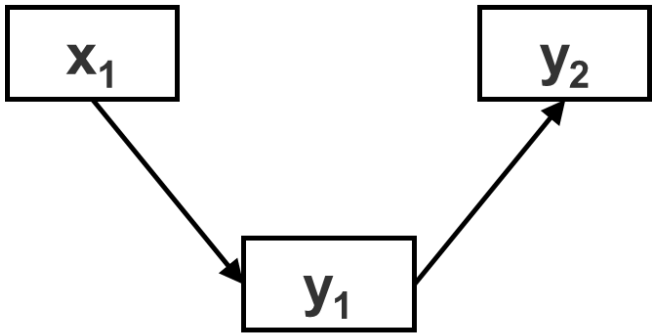



```
# SEM model in lavaan
sem_mod1 <- ' y1 ~ x1
              y2 ~  y1
'

 # 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)
```

# Assumptions

## Normality of Residuals




Additional options:

**Multivariate Shapiro-Wilks Test**

**Often too sensitive of a 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
```

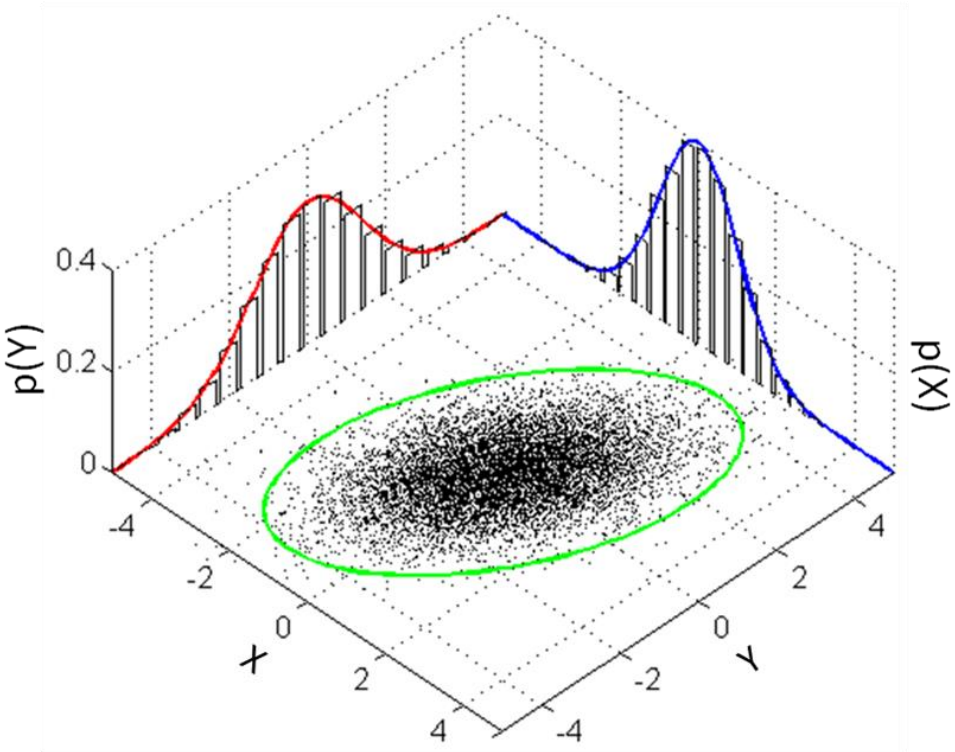
**Residuals seems fine**

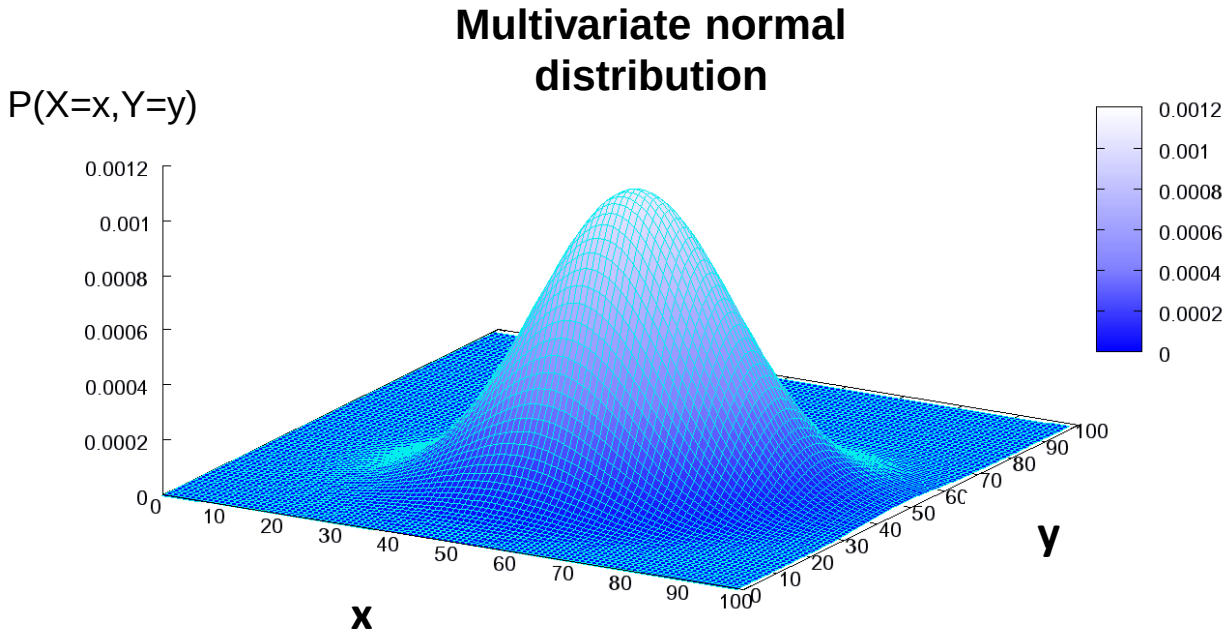
## 1. Residuals are normal

- 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

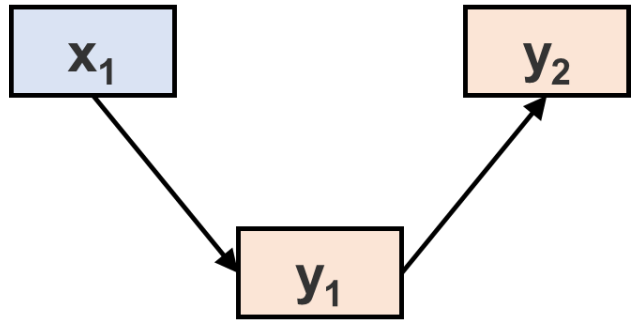
**Multivariate normality** - multiple normally distributed variables that have joint normal distribution (any linear combination of the variables is normally distributed).




Figures: wikipedia.org

## 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)

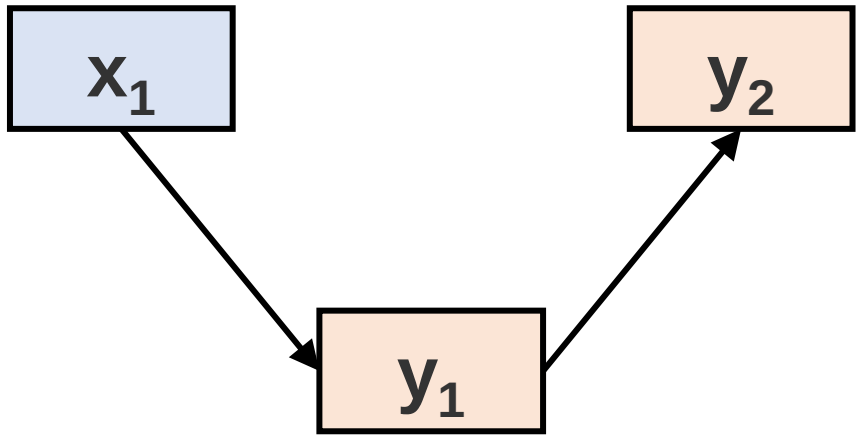
- Severe violations result in

  - inflated test of model fit

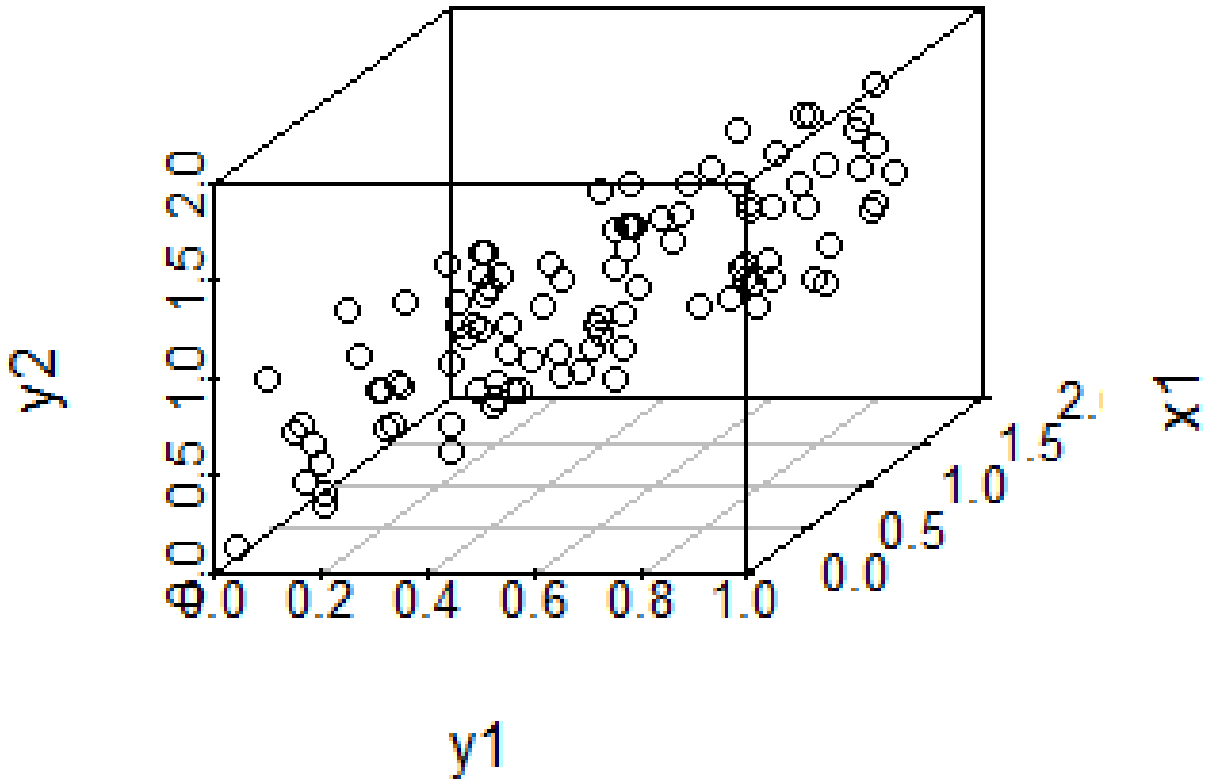
  - underestimated parameter errors

# Assumptions

## Multivariate normality of data

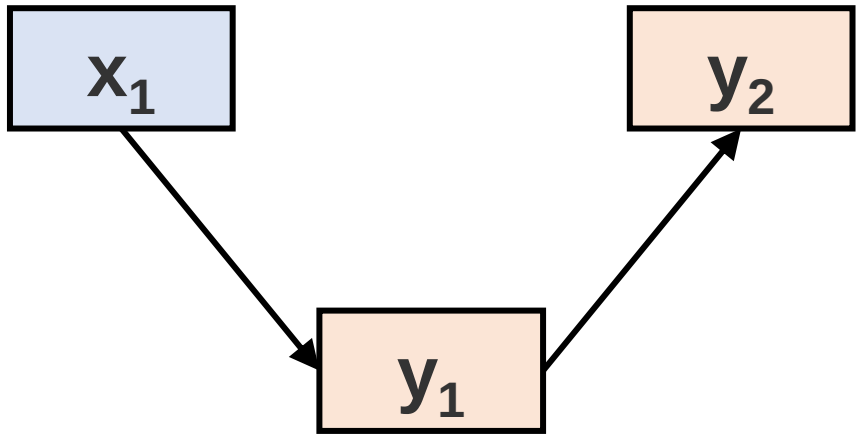



**Are these Data Multivariate Normal?**



# Assumptions

## Multivariate normality of data




### Multivariate Mardia's Test

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

# Assumptions

## Multivariate normality of data

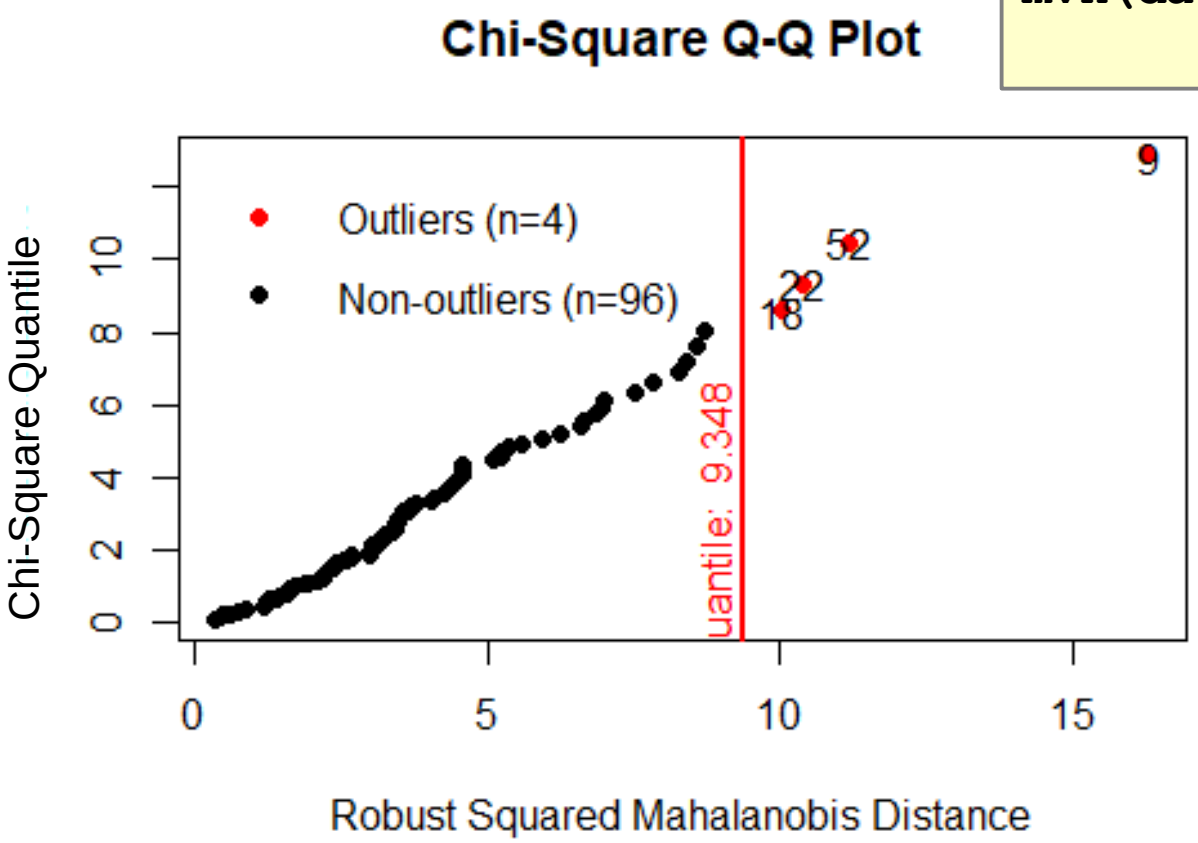
```
$multivariateNormality
              Test         Statistic               p value Result
1 Mardia Skewness  3.25985146525359   0.97463068427374     YES
2 Mardia Kurtosis -3.33768747709889 0.000844787094267163      NO
3             MVN              <NA>                  <NA>      NO

$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
```

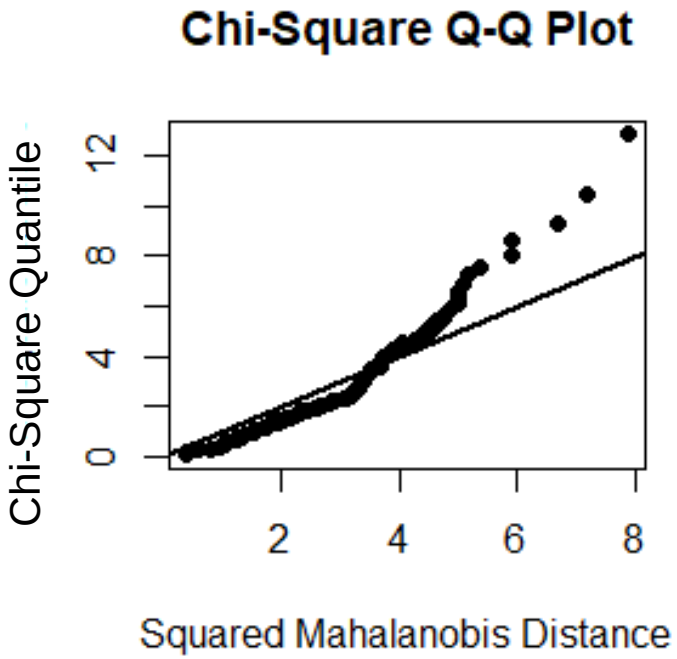
```
# Shapiro-Wilk Univariate normality test
mvn(newdata,mvnTest="mardia", univariateTest="SW")
?mvn()
```

# Assumptions

## Multivariate normality of data

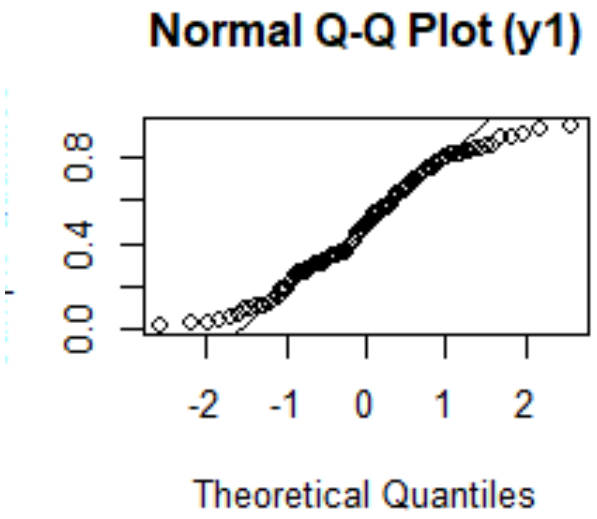
```
# plots for Multivariate Normality
mvn(data1,multivariatePlot="qq")
mvn(data1, multivariateOutlierMethod="quan")
```

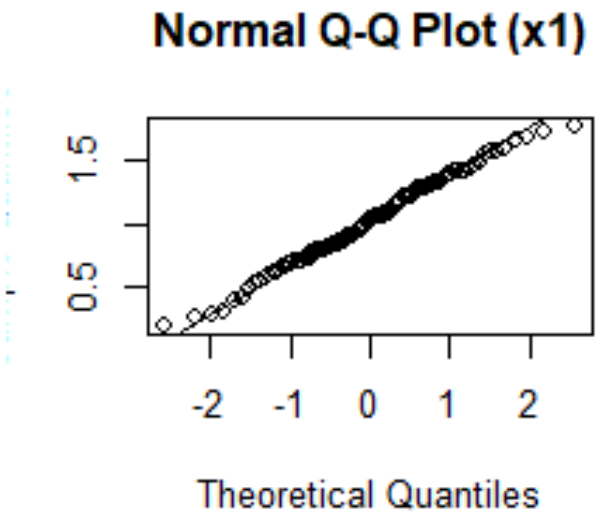


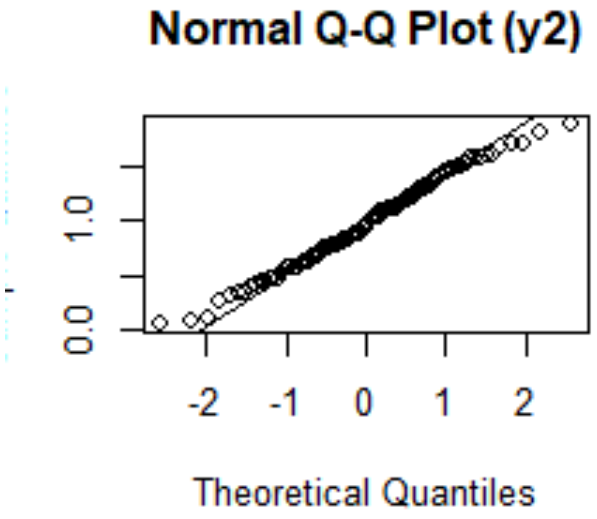
# Assumptions

## Multivariate normality of data




```
# univariate plots
mvn(data1,univariatePlot="qqplot")
```




```
> 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

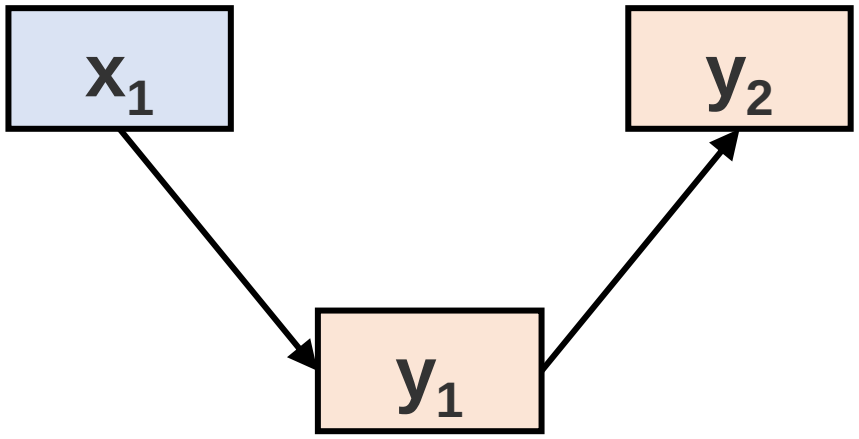
## Adjusting for non-normality of data

**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 $\chi^2$
- There are techniques to get unbiased fit and error statistics.
  - The Satorra-Bentler Chi Square Test
  - Bollen-Stine Bootstrap

**The Satorra-Bentler Chi Square**:

- Correction coefficient for $\chi^2$ 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")
```

# Adjusting for non-normality of data

**The Satorra-Bentler Chi Square**:

- Correction coefficient for $\chi^2$ 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
```

# Adjusting for non-normality of data

**The Satorra-Bentler Chi Square**:

- Correction coefficient for $\chi^2$ 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")

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

# Adjusting for non-normality of data

**The Satorra-Bentler Chi Square**:

- Correction coefficient for $\chi^2$ 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

# Adjusting for non-normality of data

## 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:**

1. Residuals are normal
2. Data are multivariate normal
3. **No missing data**

- NA in data bias parameter estimates

```
# Full-information maximum likelihood (FIML) estimation
# adjusting for incomplete data
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

No singular determinants

**Assumptions:**

1. Residuals are normal
2. Data are multivariate normal
3. No missing data
4. **No redundant variables**

vif < 2 (no collinearity)

- **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


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

```
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"**

Minimum requirement $n = p \times 5$

$n$ sample size

$p$ 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***

# 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:<br>***lavaan****:* estimator="MLM", se="robust“ or test="Satorra-Bentler" |
| | Bootstapping:<br>***lavaan****:* test="bollen.stine", se="bootstrap |
| Missing data | Full information maximum likelihood:<br>***lavaan****:* missing="fiml“ (normal data)<br>missing="fiml“, estimator="MLR" (non-normal data) |
| Positive definite S matrix | Check for multicolinearity in each single regression model: vif() |
| Dependant samples (hierarchical) | Local estimation with LMM or GLMM: package ***piecewiseSEM*** |
| Not sufficient sample size | Local estimation: package ***piecewiseSEM*** |

# Day 4 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 4 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)
```

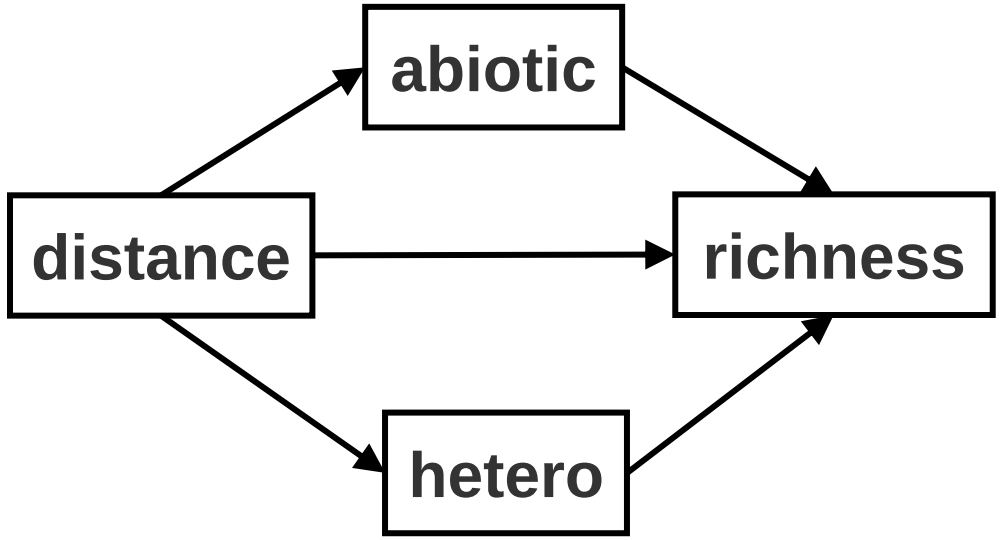
Data: Grace, J.B. and Keeley, J.E. 2006. A structural equation model analysis of postfire plant diversity in California shrublands. Ecological Applications 16:503-514

# Day 4 Task 1

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

**Other measurements:**

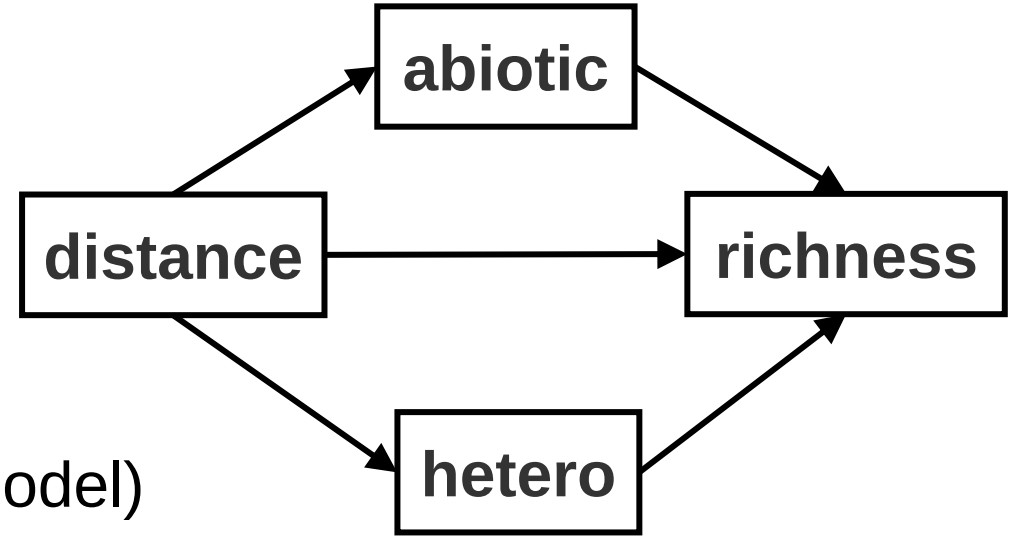
- Vegetation species richness **"richness"**
- Local abiotic conditions (aspect, soils) **"abiotic"**
- Spatial heterogenity **"hetero"**
- Distance from coast **"distance"**




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 4 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 4 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 4 Task 1

```
# Call the model-implied covariance matrix

lavInspect(SemFit, "obs")$cov



  # Check the data scales

  varTable(SemFit)
```