Introduction to statistics in R

Linear models

Lecture by Caterina Penone

Introduction: why statistics?

- Describe data

their distribution, their mean values, their relationships among each other, etc.

- Test hypotheses

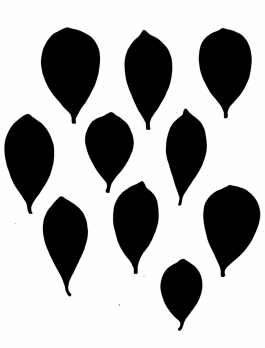
whether differences among groups or relationships are produced by chance or whether there is a real effect which requires a biological interpretation

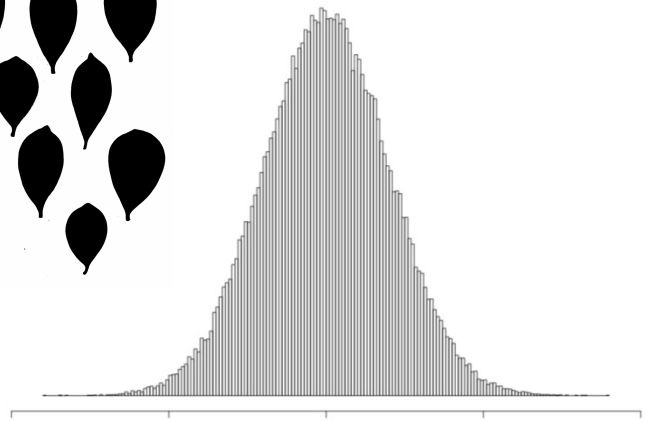
- Create a model

a summary of the data which allows to predict future outcomes

Major types of data

Continuous
Normally distributed





Discrete
Non-normally distributed

Counts

Proportions
Successes-failures

B<u>efore</u> fitting a linear model

1. What is my **hypothesis**?

2. What is my **response** (a.k.a. **dependent**) variable? Is it a continuous measurement, a count, a proportion, a category?

3. What are my **explanatory** (a.k.a. **independent**) variables? Are they continuous or categorical? Do they interact?

4. Are my **data points independent** or **grouped** in some way?
   If non-independent, include random effects (mixed models – tomorrow)

Before fitting a model, in practice

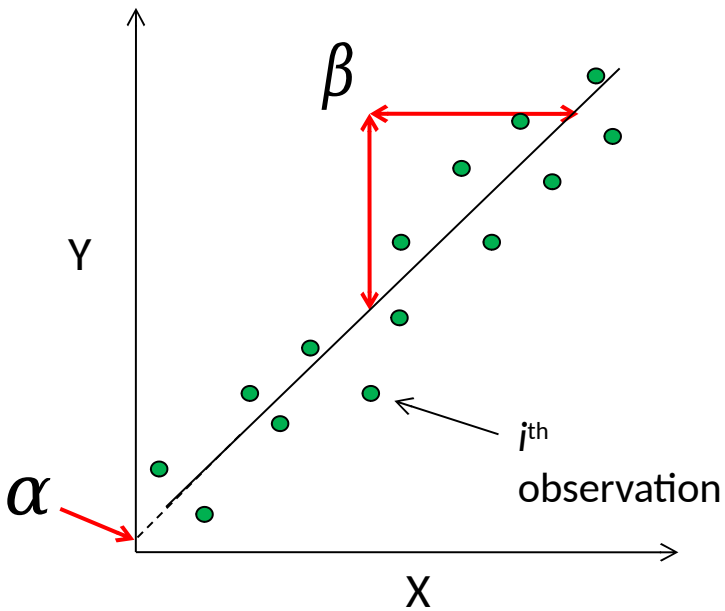
- Explore the dataset
- Check / think about data type
- Plot data
- Check correlations between variables

- Points to check:
  - How many observations
  - If categories (e.g. regions): how many observations per region
  - How my explanatory variables (x) are related between them?

The Linear Model

- A linear model **describes** the relationship between one variable, and one or more other variables.

$$Y_i = \alpha + \beta * X_i + \varepsilon_i$$



Y_i = Response (dependent variable)

X_i = Explanatory (independent variable)

α = The intercept (value of y when x=0)

β = Value by which y changes with x
= the slope/ the strength of the x-y relationship

ε_i = Error – unexplained, **normally distributed information** (residuals)

A linear model in R

- $y = a + bx$ in R: y ~ b
- $y = a + bx + cz$ in R: y ~ b + c

response variable
dependent variable
explained variable
measured variable

explanatory variable
independent variable
predictor variable
manipulated variable

Linear model: how does it work

Central assumption of linear models

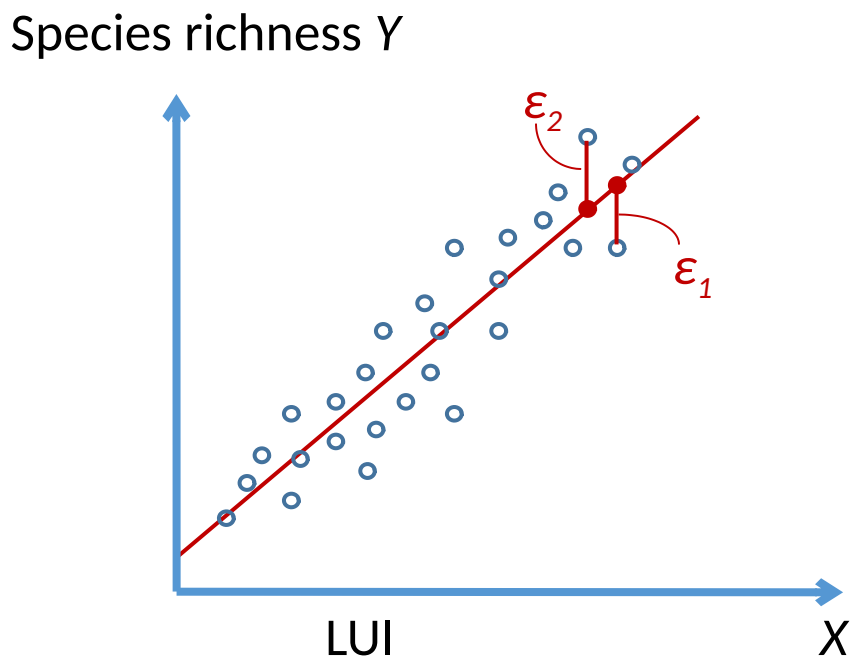
Describing data with few parameters

$$Y_i = a + b \cdot X_i + \varepsilon_i \qquad \varepsilon_i \sim Normal(0, \sigma^2)$$

ε_i are called the residuals

$\hat{Y}$ are the fitted or predicted values

The residuals ε_i are the difference between the observed (°) and predicted (•) values

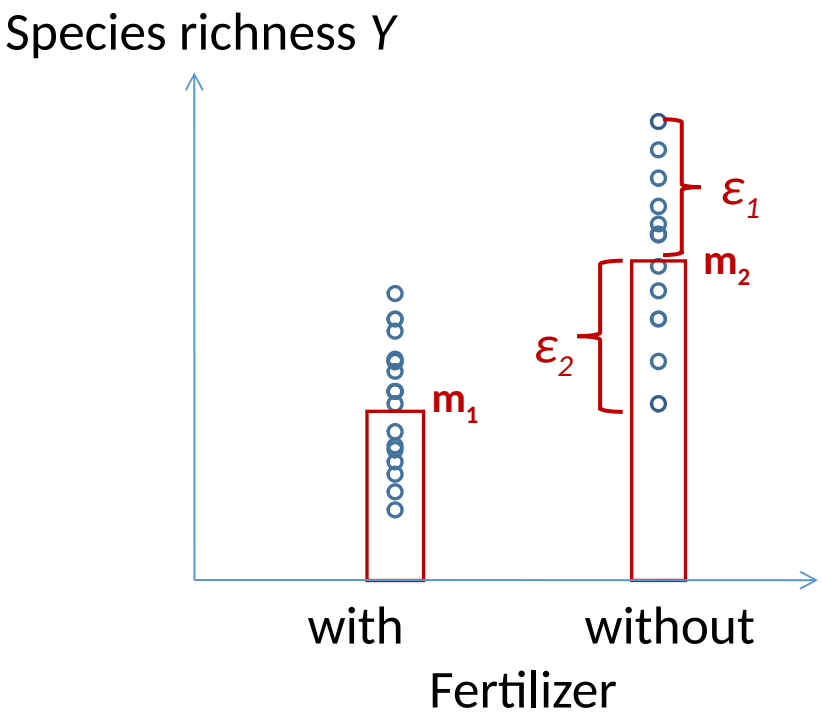


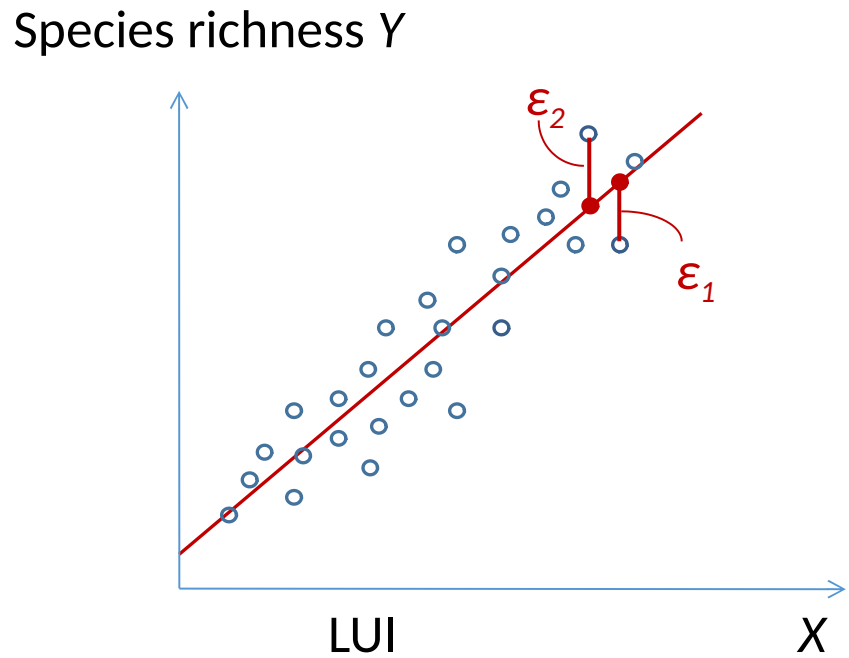
Linear model: how does it work

Criteria for parameter estimates: Least square approach

Alternative: Maximum likelihood (later)

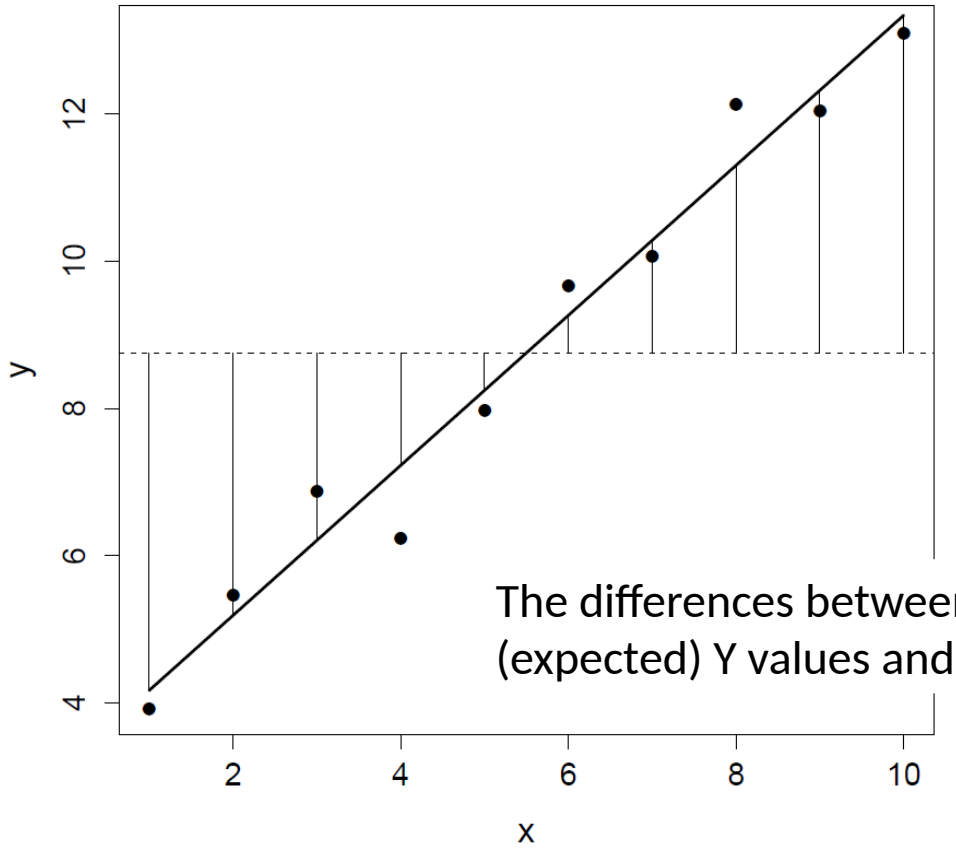
The sum of squared residuals is minimal: $$Min\left(\sum_{i=1}^{n} \varepsilon_i^{\,2}\right)$$





Linear regression: how it works

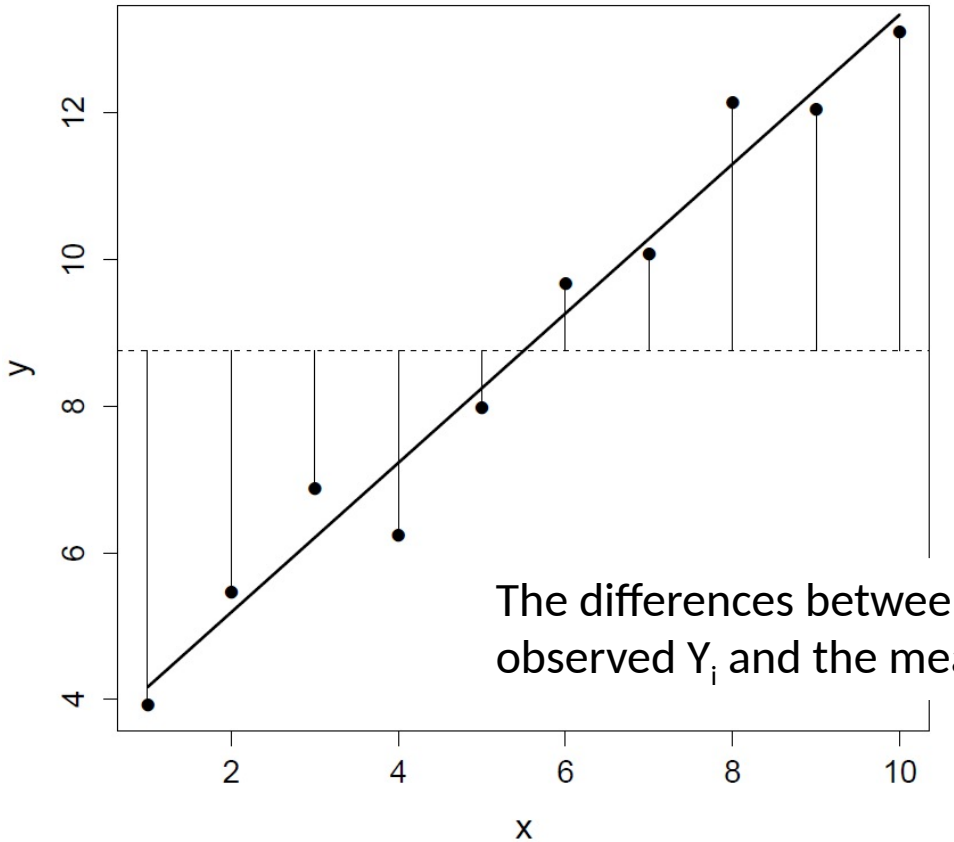
$$SS_{regression} = \Sigma\ (\hat{y}_i - \bar{y})^2$$



The differences between the fitted (expected) Y values and the mean of Y

Linear regression: how it works

$$SS_{total} = \Sigma\ (y_i - \bar{y})^2$$



The differences between the observed Y_i and the mean of Y

Linear regression: how it works

	df	Sum of squares	Mean SS	F-value	P(>F)
Model	1				
Residuals	8				

- Model df = # parameters estimated – 1
- Residual df = # observations – 1 – Regression df

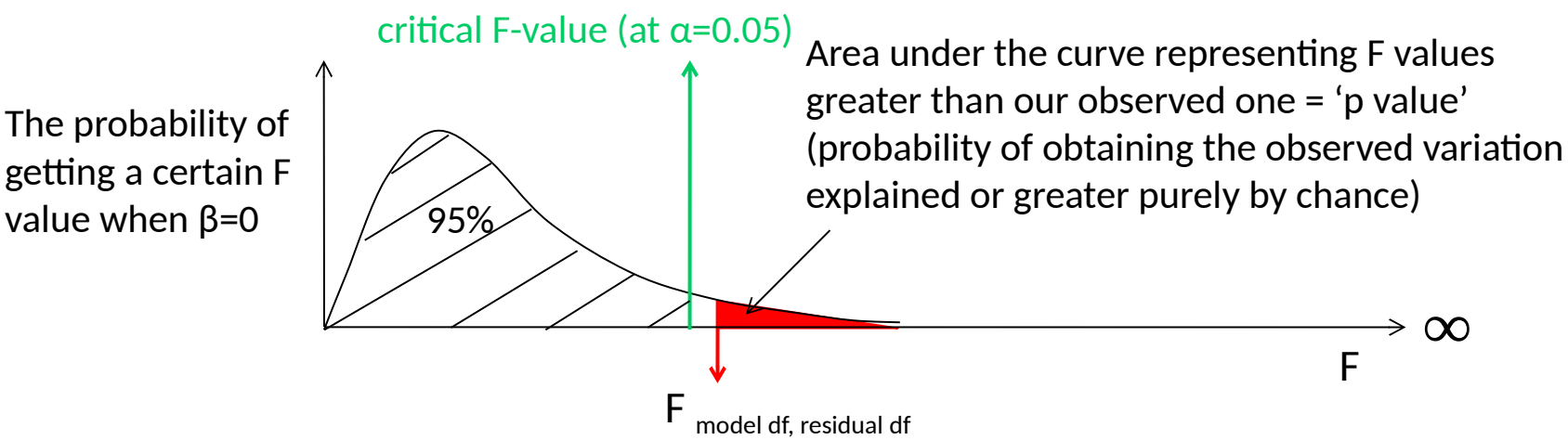
- The **F-ratio** = $\frac{SS_{resid} / df_{resid}}{SS_{regression} / df_{Model}}$

- F-ratio ~ $F_{\text{dfmodel, dfresiduals}} = F_{1,8}$

Linear regression: how it works

	df	Sum of squares	Mean SS	F-value	P(>F)
Model	1				
Residuals	8				

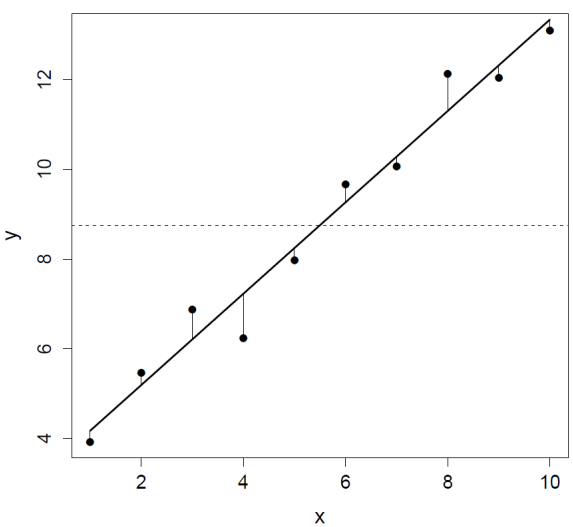
- Using the **F-distribution**, determined by our df, we find the critical F-value (at α=0.05) above which our F-ratio needs to be, to reject the null hypothesis (β = 0).



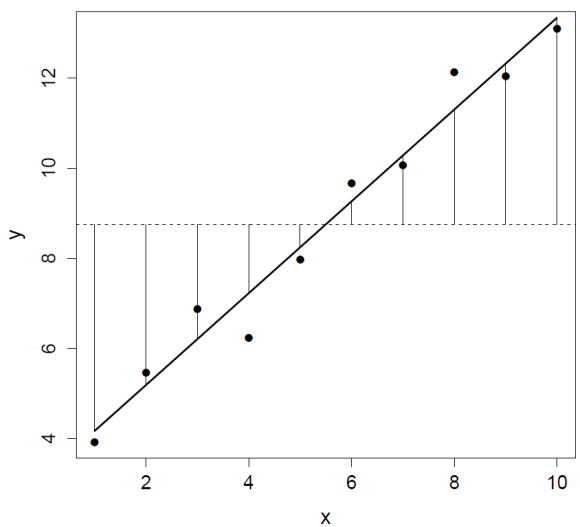
Linear model and coefficient of determination R^2

- The coefficient of determination R^2 is the "**proportion of variance explained by the model**", i.e. the proportion of variance in the response variable that is predictable from the explanatory variable(s).
- It is based on the Sum of Squares

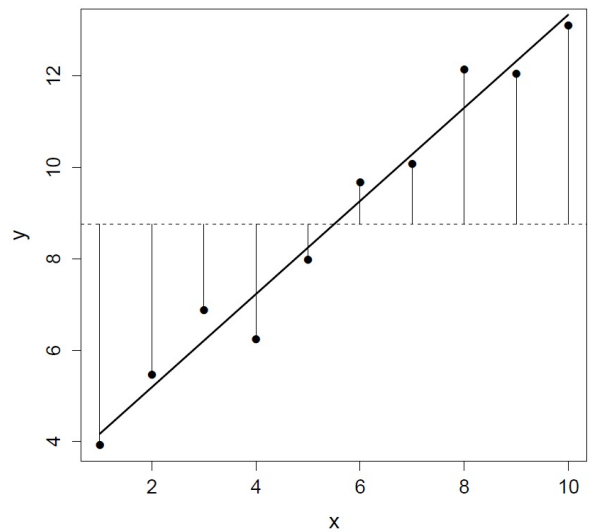
$SS_{residual} = \Sigma_i e_i^2 = \Sigma\,(y_i - \hat{y}_i)^2$



$SS_{regression} = \Sigma\,(\hat{y}_i - \bar{y})^2$



$SS_{total} = \Sigma\,(y_i - \bar{y})^2$



Linear model and coefficient of determination R^2

- The coefficient of determination R^2 is the "**proportion of variance explained by the model**", i.e. the proportion of variance in the response variable that is predictable from the explanatory variable(s).

$SS_{residual} = \Sigma_i e_i^2 = \Sigma\,(y_i - \hat{y}_i)^2$

$SS_{total} = \Sigma\,(y_i - \bar{y})^2$

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

Linear model and coefficient of determination R^2

- As we can always get a better fit (or at least not a worse fit) by adding more explanatory variables to the model, we often use an '**adjusted R^2**'; which is **penalized for including more variables into the model**:

$$adj.R^2 = 1 - \left(\frac{n-1}{n-p-1}\right)\left(\frac{SS_{residual}}{SS_{total}}\right)$$

Degrees of freedom

The degrees of freedom is the number of values in a statistic that are free to vary

The total degrees of freedom df_{total} is equal the number of observations minus one $(N\text{-}1)$

The degree of freedom of the model df_{model} is equal the number of parameters minus one $(P\text{-}1)$

The residual degree of freedom $df_{Residual}$ is the difference between them

-> $df_{Residual} = df_{total} - df_{model}$

The residual degree of freedom determine the power of a statistical test, i.e. the ability to detect significant differences

Statistical significance

The statistical significance is the probability to observe an effect just *by chance,*

- assuming that there is no such effect in reality
- knowing the probability of an event (as in throwing dice)
- or assuming a specified data distribution (e.g. normal distribution)

⇒ we conclude that there is an effect and quantify the probability that this conclusion is wrong

In ecology, a threshold error probability $\alpha = 0.05 = 5\ \%$ is generally accepted

-> 1 out of 20 (= 5%) tests becomes significant just by chance

-> a problem when you do many statistical tests

-> reduce the threshold probability α when many test are involved (see: Verhoeven et al. Oikos 2005)

Statistical significance

Levels of significance: some people prefer „exact“ error probabilities, some only different levels

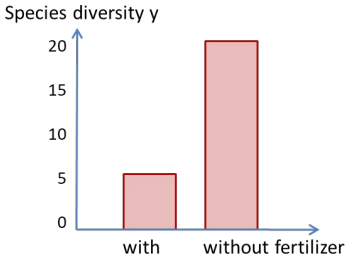
Stars	Levels	Exact	Terms
ns	$p > 0.1$	$p > 0.13452$	non significant
.	$p < 0.1$	$p < 0.08745$	marginally significant
*	$p < 0.05$	$p < 0.04256$	significant
**	$p < 0.01$	$p < 0.00127$	highly significant
***	$p < 0.001$	$p < 0.00028$	

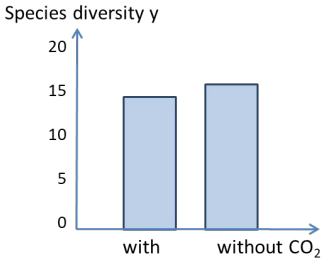
What affects significance?

The statistical significance is influenced by three factors:

- The effect size: how large is the effect

Big differences are more likely statistically significant





- The remaining variability of the data

The higher the variability, the less likely it will become statistically significant

- The number of replications

The larger the sample size, the more likely it will become statistically significant

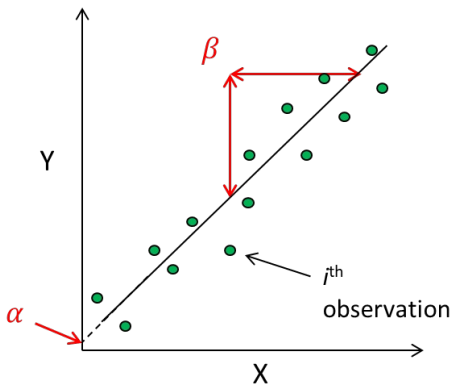
All three factors are part of the experimental design and can be, at least partly, determined by scientists

Nine commandments on p-values

1. Do not use the term ‘‘significance’’ outside the statistical meaning to prevent confusion.
2. Always say “statistically significant’’ to make clear that you do not make any judgment on magnitude of an effect; use the expression ‘‘biological relevance’’ when you refer to biological significance
3. Do not write “we found a significant effect of x on y”; write instead “higher levels of x increased the amount of y ($p < 0.05$)”. Thereby you put the statistics where it belongs: in parenthesis!
4. “Statistically significant” does *not* entail the that magnitude of the effect is biologically relevant
5. “Statistically significant’’ only means that your p-value is lower than a cut-off point (usually 5 %)
6. Biological relevance is decided a priori and it is not necessarily a large value. You have to judge the magnitude you consider biologically relevant on a case by case basis.
7. Provide confidence intervals or standard errors for your parameter estimates.
8. A p-value is *not* a direct measure of evidence against the null hypothesis. The smaller the p-value does not mean the ‘‘better’’, simply “less probable” a chance result.
9. A p-value provides you with is the probability of having obtained your data, or more extreme data, only if the null hypothesis is true. *No other interpretation of what a p-values is correct*.

Modified from: Martinez-Abrain (2008) Acta Oecologica 34: 9-11

Formulas



- `y ~ 1` just the intercept
- `y ~ a` one main effect
- `y ~ a + b` two main effects
- `y ~ a + b + a:b` two main effects and interaction between a and b
- `y ~ a*b` same as previous
- `y ~ factor(a)` transform a into factor
- `log(y) ~ a` log transform y

- + inclusion of an explanatory variable in the model (not addition)
- - deletion of an explanatory variable from the model (not subtraction);
- * inclusion of explanatory variables and interactions (not multiplication);

A linear model in R

- Create a model corresponding to our first hypothesis: plant richness varies with LUI

```
> mod0 <- lm(Plant_SpeciesRichness ~ LUI, data=dat)
> summary(mod0)

Call:
lm(formula = Plant_SpeciesRichness ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.165  -7.605  -1.167   7.038  26.069 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   35.182      2.261  15.558  < 2e-16 ***
LUI           -6.317      1.306  -4.838 3.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.11 on 121 degrees of freedom
Multiple R-squared:  0.1621,	Adjusted R-squared:  0.1552 
F-statistic: 23.41 on 1 and 121 DF,  p-value: 3.905e-06
```

Dissecting the model summary

The formula. Stored in: mod0$terms

```
Call:
lm(formula = Plant_SpeciesRichness ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.165  -7.605  -1.167   7.038  26.069 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   35.182      2.261  15.558  < 2e-16 ***
LUI           -6.317      1.306  -4.838 3.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.11 on 121 degrees of freedom
Multiple R-squared:  0.1621,	Adjusted R-squared:  0.1552 
F-statistic: 23.41 on 1 and 121 DF,  p-value: 3.905e-06
```

Dissecting the model summary

```
Call:
lm(formula = Plant_SpeciesRichness ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.165  -7.605  -1.167   7.038  26.069 
```

A summary of the residuals.

Same as: summary(mod0$residuals)

Residuals are accessible: mod0$residuals

Residuals can be calculated: Plant_SpeciesRichness-mod0$fitted.values

Dissecting the model summary

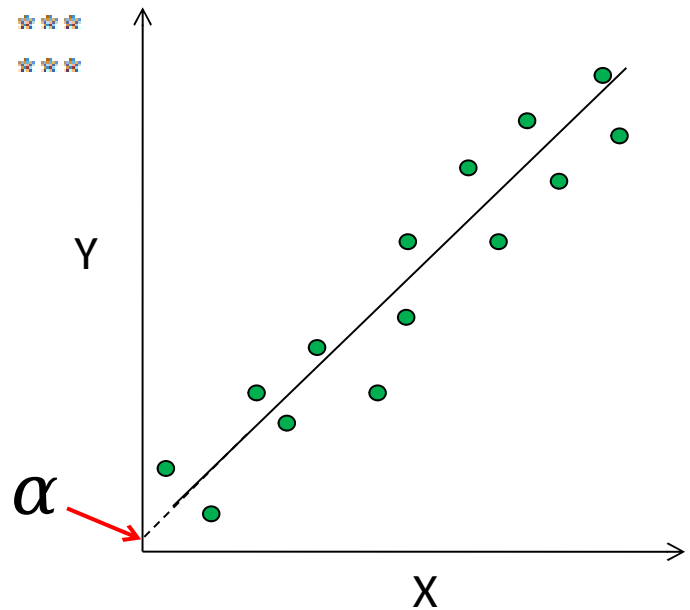
```
Call:
lm(formula = Plant_SpeciesRichness ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.165  -7.605  -1.167   7.038  26.069 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   35.182      2.261  15.558  < 2e-16 ***
LUI           -6.317      1.306  -4.838 3.91e-06 ***
```

The model coefficients

Dissecting the model summary

```
Call:
lm(formula = Plant_SpeciesRichness ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.165  -7.605  -1.167   7.038  26.069 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   35.182      2.261  15.558  < 2e-16 ***
LUI           -6.317      1.306  -4.838 3.91e-06 ***
```

The intercept

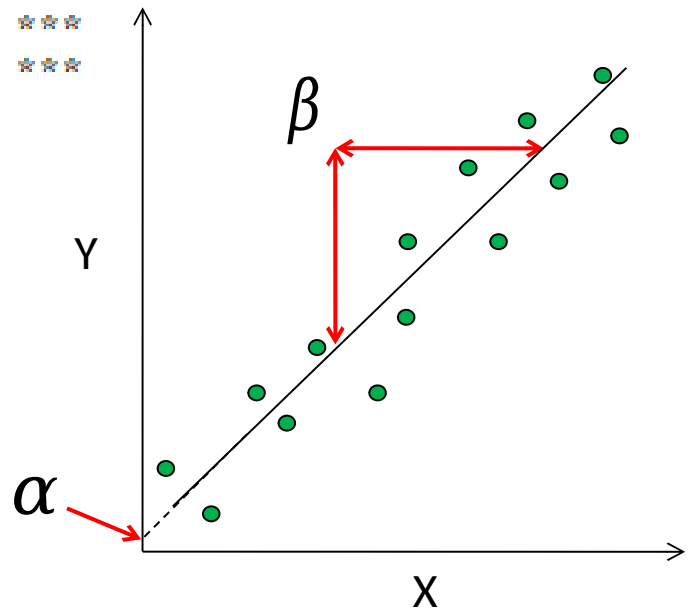


Dissecting the model summary

```
Call:
lm(formula = Plant_SpeciesRichness ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.165  -7.605  -1.167   7.038  26.069 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   35.182      2.261  15.558  < 2e-16 ***
LUI           -6.317      1.306  -4.838 3.91e-06 ***
```

The slope

Slope and intercept can be extracted using:
mod0$coefficients



Dissecting the model summary

```
Call:
lm(formula = Plant_SpeciesRichness ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.165  -7.605  -1.167   7.038  26.069 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   35.182      2.261  15.558  < 2e-16 ***
LUI           -6.317      1.306  -4.838 3.91e-06 ***
```

Standard errors

Dissecting the model summary

```
Call:
lm(formula = Plant_SpeciesRichness ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.165  -7.605  -1.167   7.038  26.069 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   35.182      2.261  15.558  < 2e-16 ***
LUI           -6.317      1.306  -4.838 3.91e-06 ***
```

t values: Estimate divided by Std. Error

Dissecting the model summary

```
Call:
lm(formula = Plant_SpeciesRichness ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.165  -7.605  -1.167   7.038  26.069 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   35.182      2.261  15.558  < 2e-16 ***
LUI           -6.317      1.306  -4.838 3.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.11 on 121 degrees of freedom
Multiple R-squared:  0.1621,	Adjusted R-squared:  0.1552 
F-statistic: 23.41 on 1 and 121 DF,  p-value: 3.905e-06
```

p-value: depends on t-value and degrees of freedom

Dissecting the model summary

Residual standard error: can be calculated

```
> k <- length(mod0$coefficients)-1 #Subtract one to ignore intercept
> SSE <- sum(mod0$residuals**2)
> n <- length(mod0$residuals)
> sqrt(SSE/(n-(1+k))) #Residual Standard Error
[1] 10.1132
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   35.182      2.261  15.558  < 2e-16 ***
LUI           -6.317      1.306  -4.838 3.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.11 on 121 degrees of freedom
Multiple R-squared:  0.1621,    Adjusted R-squared:  0.1552
F-statistic: 23.41 on 1 and 121 DF,  p-value: 3.905e-06
```

Dissecting the model summary

Multiple and adjusted R-squared: model fit

```
> SSyy <- sum((dat$Plant_SpeciesRichness-mean(dat$Plant_SpeciesRichness))**2)
> SSE <- sum(mod0$residuals**2)
> (SSyy-SSE)/SSyy
[1] 0.162086
> n <- length(dat$Plant_SpeciesRichness)
> k <- length(mod0$coefficients)-1 #Subtract one to ignore intercept
> 1-(SSE/SSyy)*(n-1)/(n-(k+1))
[1] 0.1551611
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   35.182      2.261  15.558  < 2e-16 ***
LUI           -6.317      1.306  -4.838 3.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.11 on 121 degrees of freedom
Multiple R-squared:  0.1621,    Adjusted R-squared:  0.1552
F-statistic: 23.41 on 1 and 121 DF,  p-value: 3.905e-06
```

Dissecting the model summary

F-statistic : a test that checks if at least one of the coefficients is nonzero (significantly different from 0)
#Ho: All coefficients are zero
#Ha: At least one coefficient is nonzero

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   35.182      2.261  15.558  < 2e-16 ***
LUI           -6.317      1.306  -4.838 3.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.11 on 121 degrees of freedom
Multiple R-squared:  0.1621,    Adjusted R-squared:  0.1552
F-statistic: 23.41 on 1 and 121 DF,  p-value: 3.905e-06
```

<u>After</u> fitting a linear model

1. Are the linear model **assumptions** met?
   - Test the model assumptions

2. How **good** is the model for my data?
   - Investigate how well the model describes the data
     - % of variance explained

3. Is my **hypothesis** confirmed?

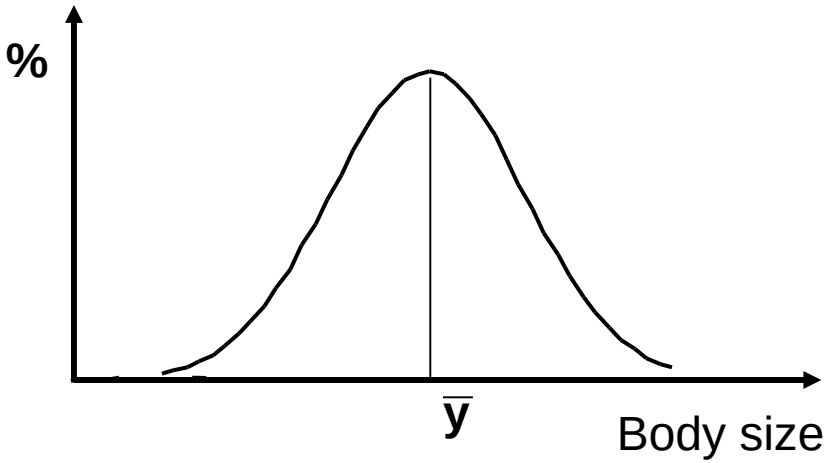
Assumptions of linear models

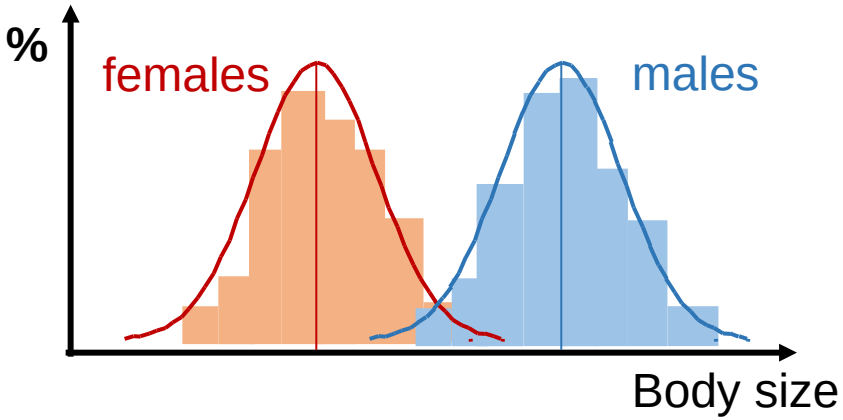
- Normality of errors / residuals

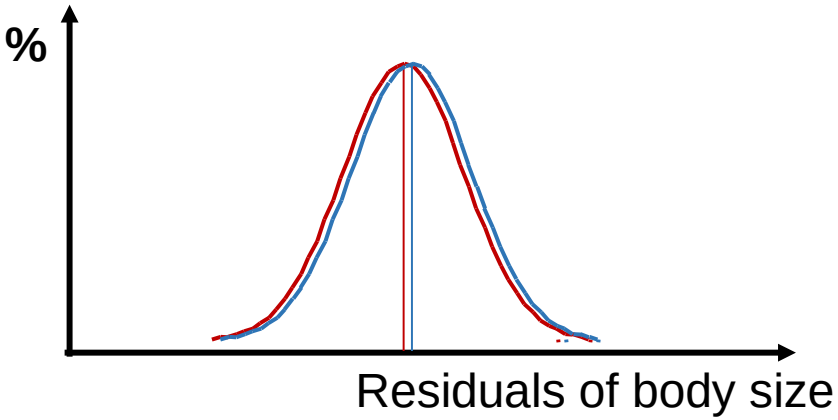
- Homoscedasticity / constant variance

The variance in y is constant (i.e. the variance does not change as y gets bigger)

Assumptions: normality of residuals



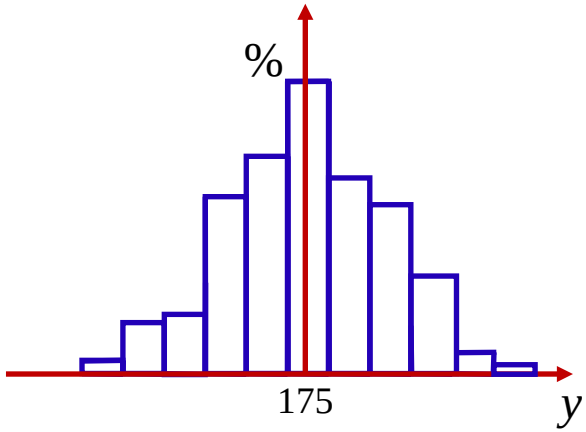




- > Not the raw data needs to be normally distributed, but the residuals

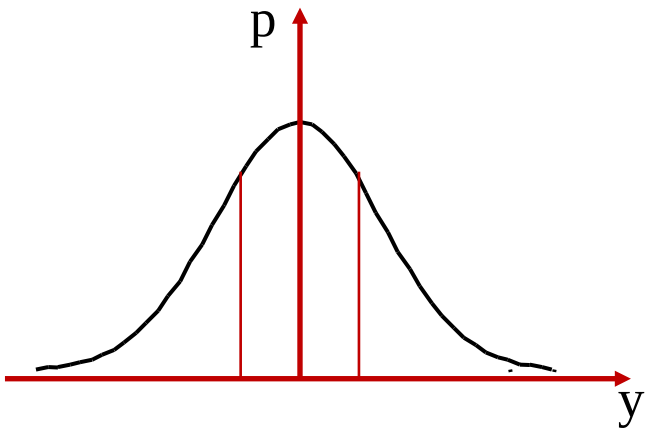
Test the model assumptions - QQplot

Frequency distribution



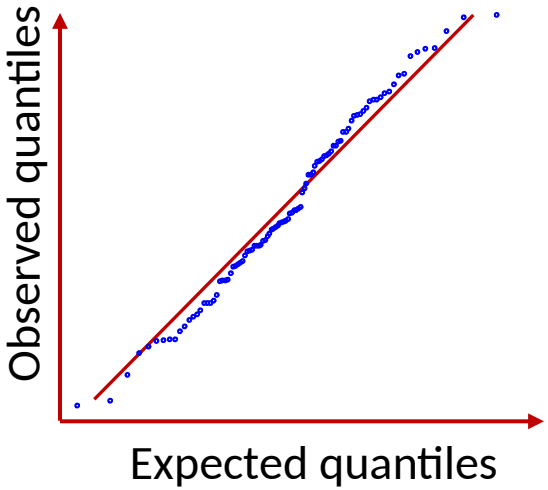
Observed quantiles

Probability distribution



Expected quantiles

Quantile-quantile QQ-plot

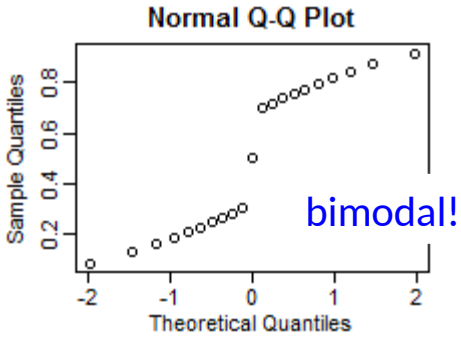


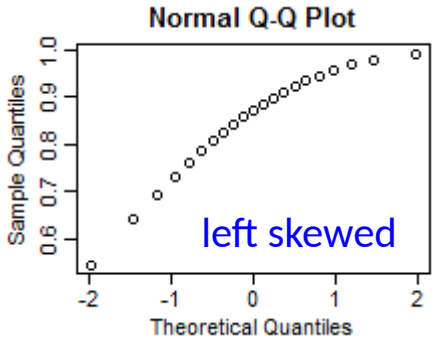
Expected quantiles in a perfect normal distribution with mean and sd corresponding to the ones of the residuals

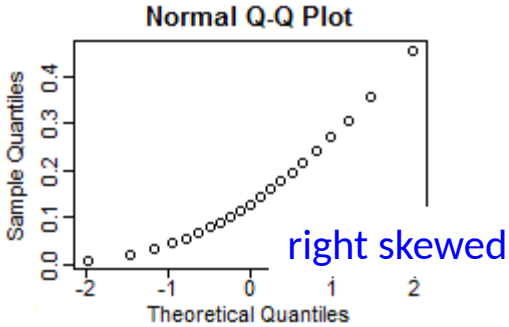
```
> qqnorm(residuals(mod0))
> qqline(residuals(mod0))
```

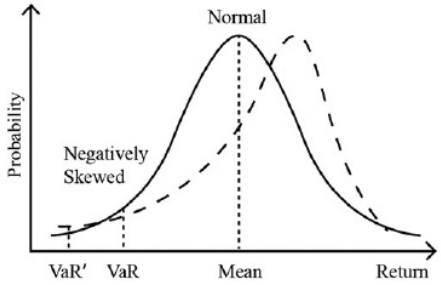
Model validation: normality of residuals

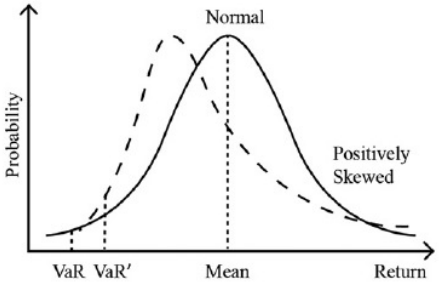
- What a qq-plot should not look like:







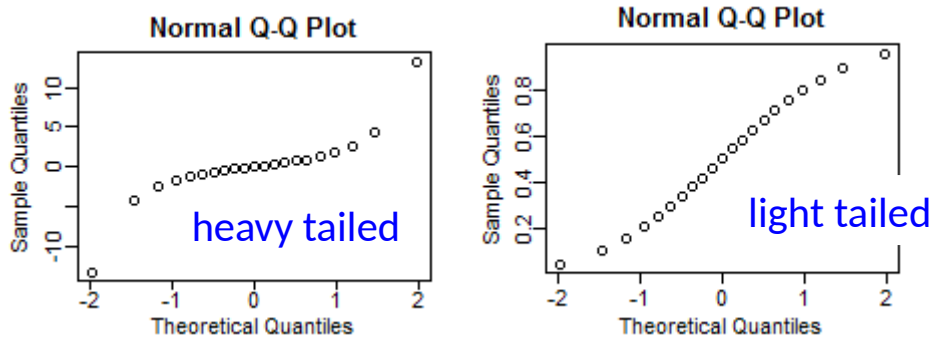


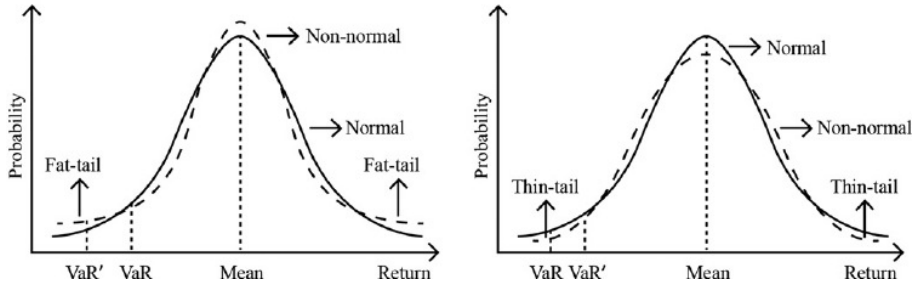


skewness

Model validation: normality of residuals

- What a qq-plot should not look like:





Positive
Leptokurtic

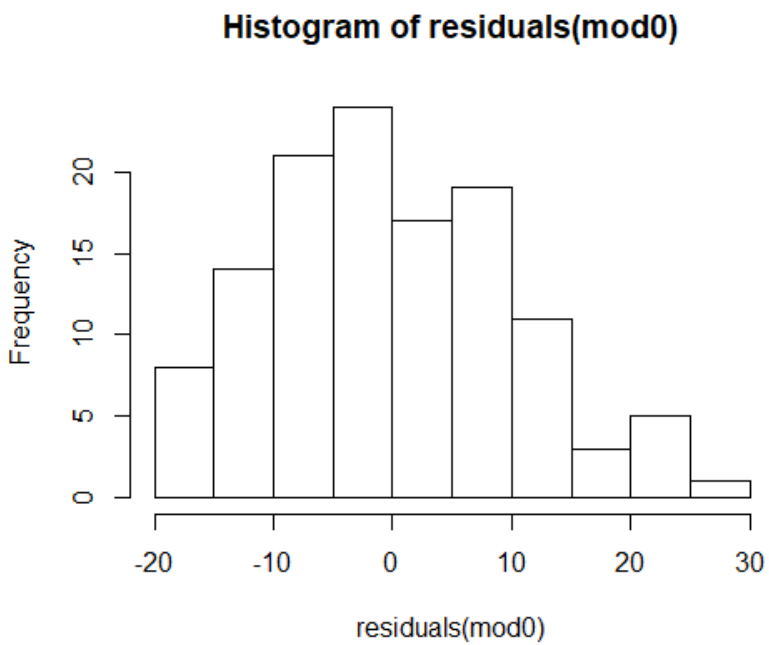
Negative
Platykurtic

kurtosis

Test the model assumptions in R

- Graphically

```
> hist(residuals(mod0))
```



- Statistically

```
> shapiro.test(residuals(mod0))

        Shapiro-Wilk normality test

data:  residuals(mod0)
W = 0.98096, p-value = 0.0799
```

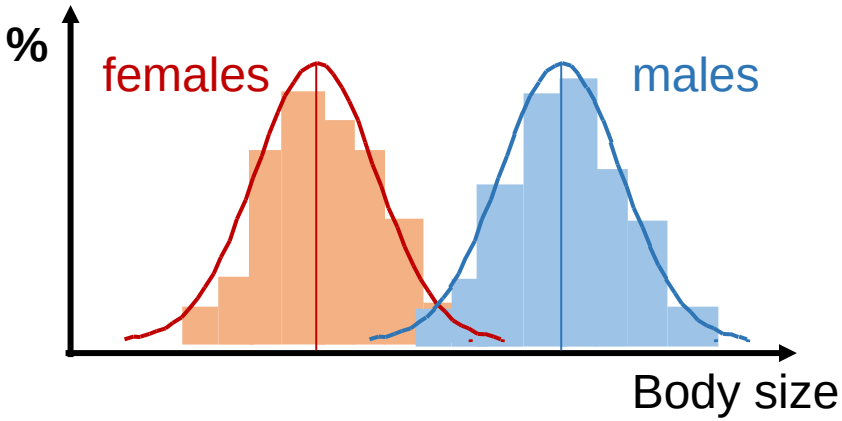
TAKE-HOME : Assumptions of linear models

- 2 Assumptions of linear models
  - normality of residuals
    - → check residual distribution with QQ plot
    - QQ-plot : points should +- follow the diagonal

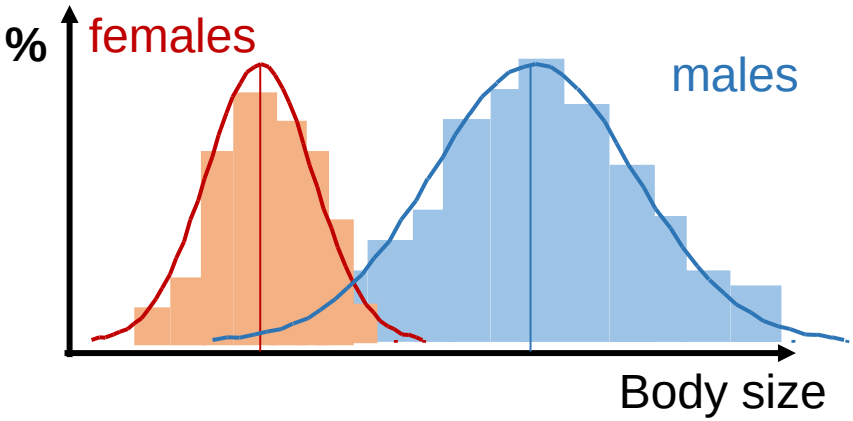
- Homoscedasticity / constant variance

Assumptions: homoscedasticity

- The variance is homogeneous



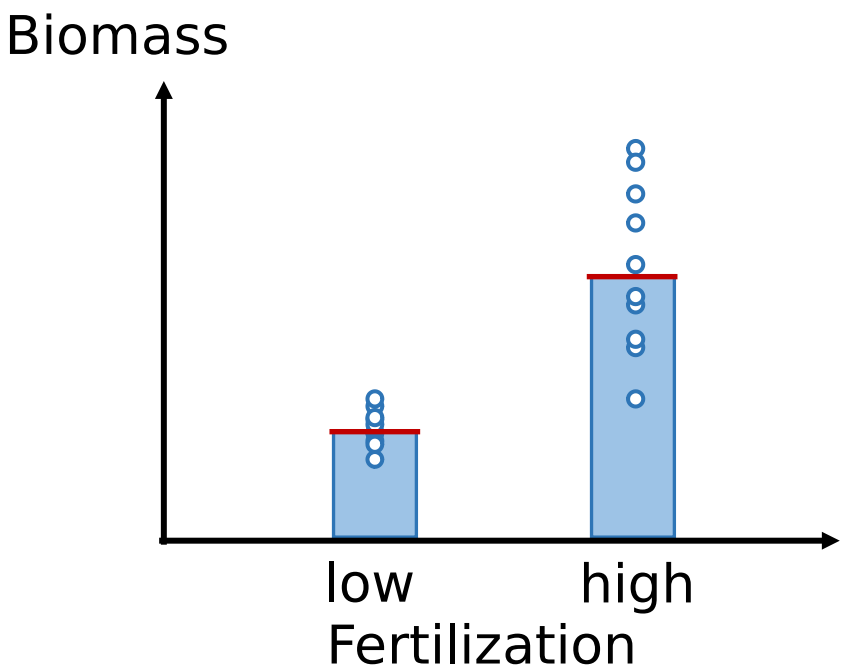
Homoscedastic - the variances of the groups is ≈ equal

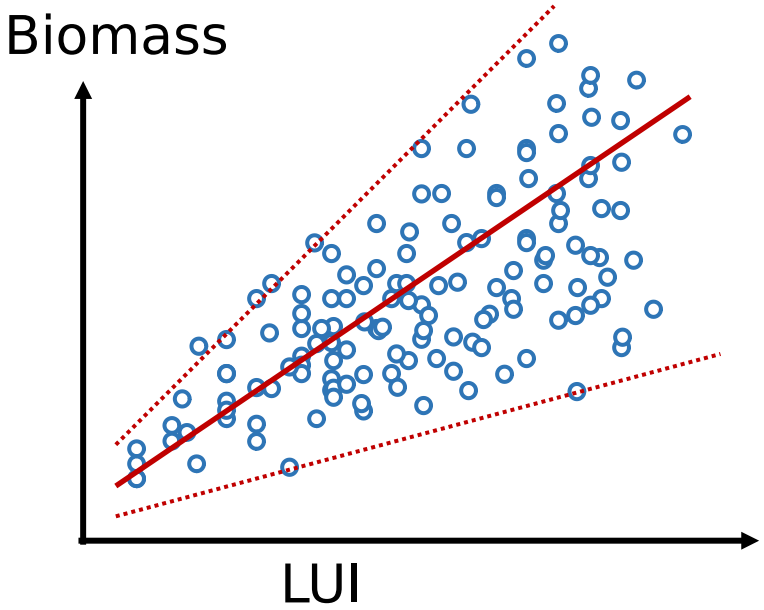


Heteroscedastic - the variances of the groups differs

Assumptions: homoscedasticity

- The variance increases with the predicted value



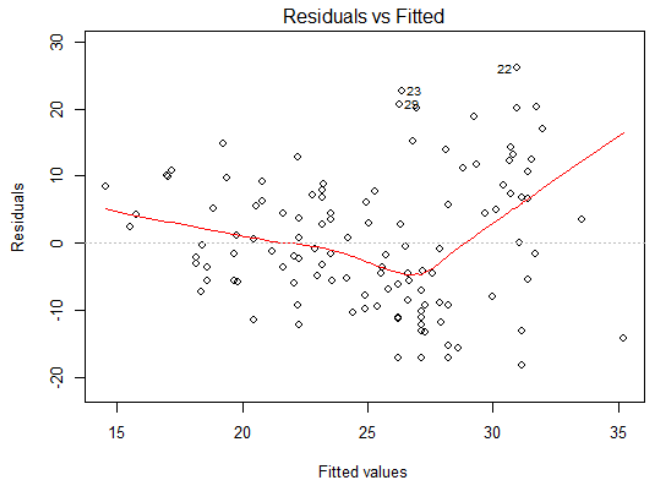


Test the model assumptions in R - homoscedasticity

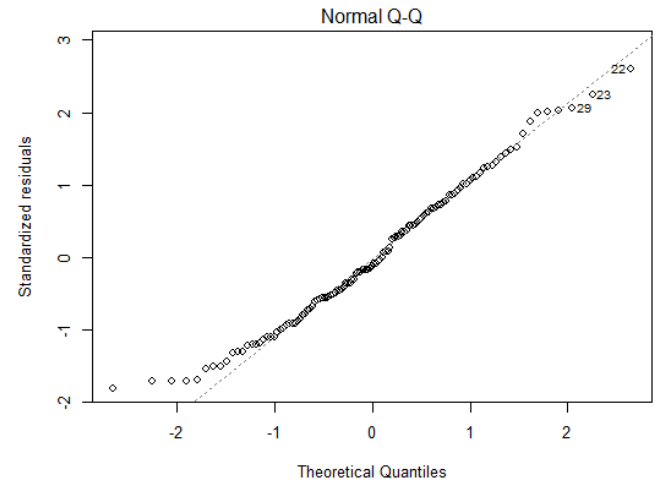
- Graphically

plot(mod0)

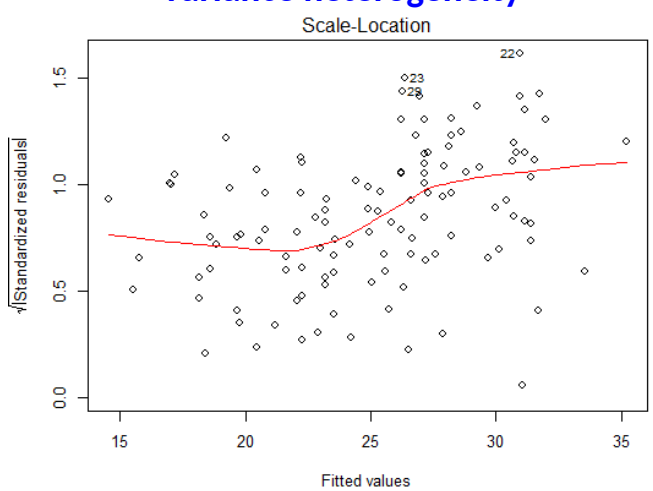
Variance heterogeneity/outliers



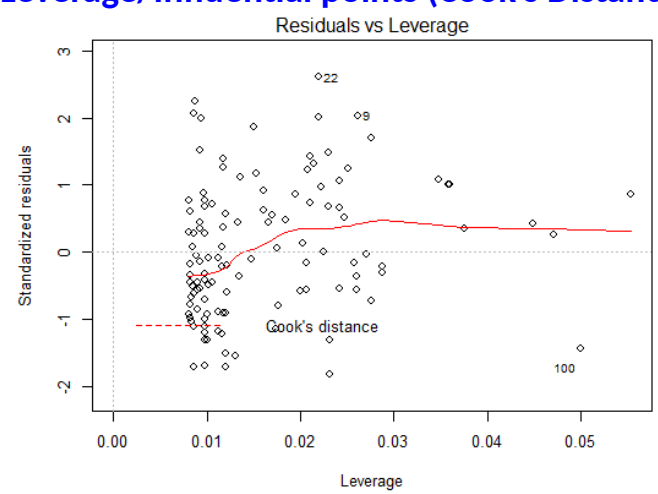
Normality/outliers



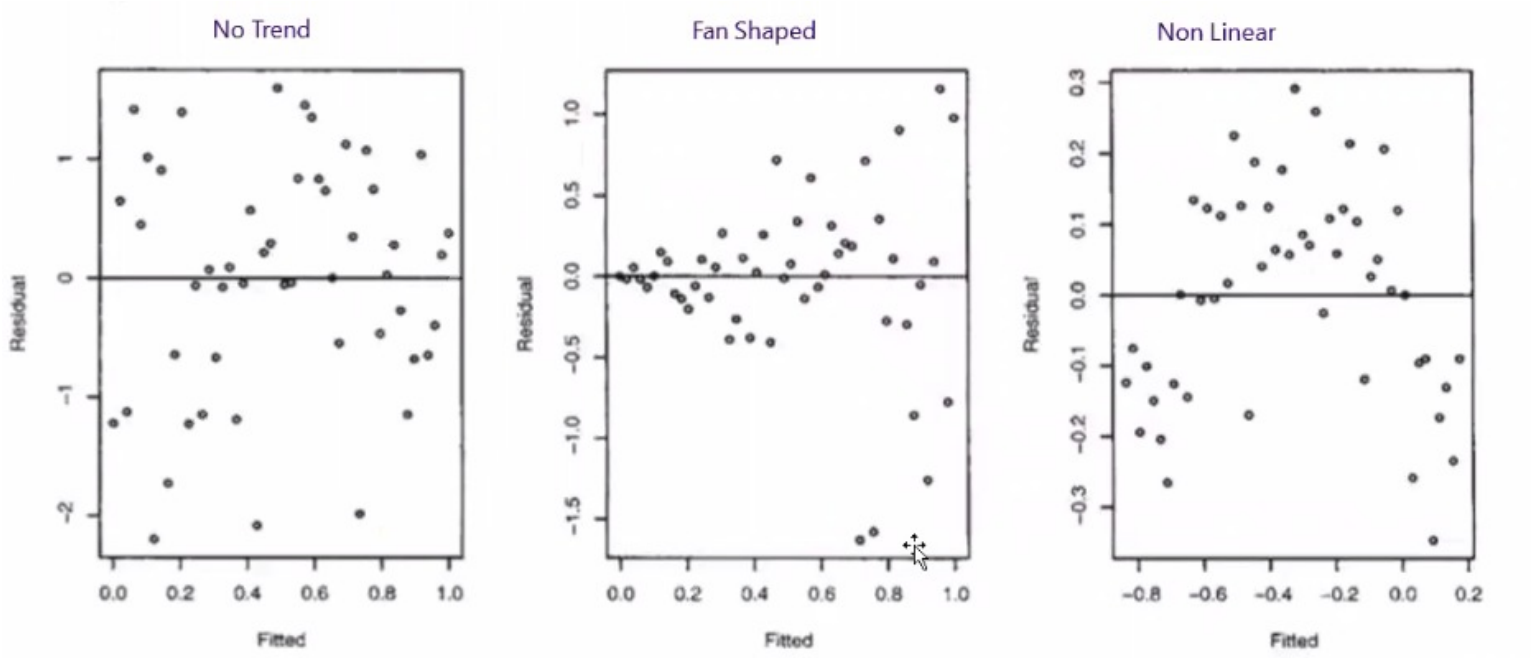
Variance heterogeneity



Leverage/Influential points (Cook's Distance)



Test the model assumptions in R - homoscedasticity



Additional information : Test the model assumptions in R

- Statistically

Shapiro-Wilk test for normality

- `> shapiro.test(residuals (mod))`

Kolmogorov-Smirnov test for any distribution

- `> ks.test(residuals (mod), „pnorm“)`

Levene test for constant variances (categorical variables)

- `> leveneTest(mod)`

All these tests are highly sensitive, in particular with large sample sizes

-> they tell that you often have not-normal residuals and unequal variances

TAKE-HOME : Assumptions of linear models

- 2 Assumptions of linear models
  - normality of residuals
    - → check residual distribution with QQ plot
    - QQ-plot : points should +- follow the diagonal

- Homoscedasticity / constant variance
  - check with `plot(mod)`
    distribution should look +- uniform (straight horizontal line)

- prefer visual inspection over formal tests

What if the assumptions are not met?

- Transformation of the response variable
- Generalized linear models (later)
- Missing explanatory variable (later)

Transformation of the response variable

- Another example

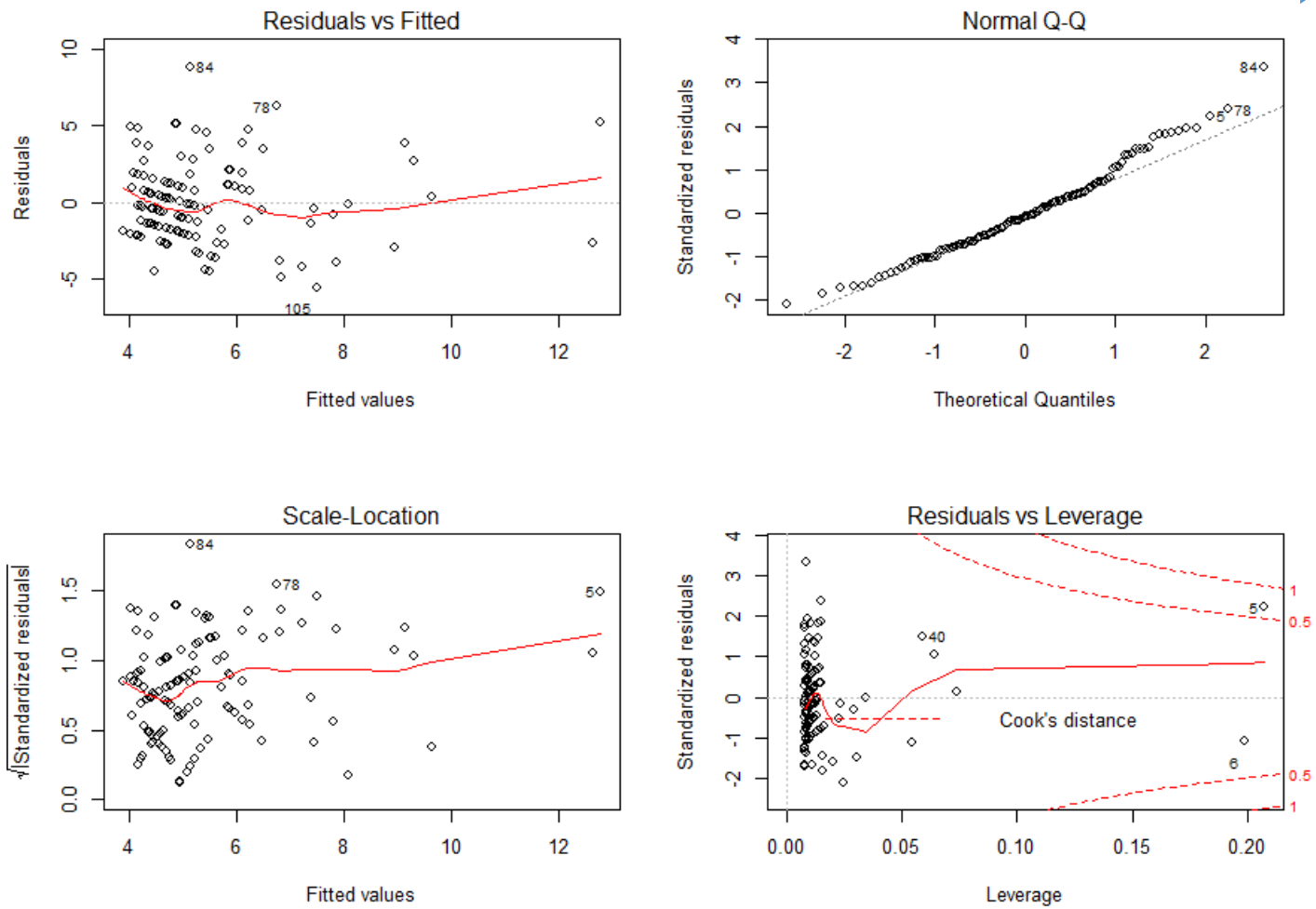
Predators richness ~ herbivores biomass

```
> mod0.1 <- lm(Predator_SpeciesRichness ~ Herbivore_biomass, data=dat)
> summary(mod0.1)

Call:
lm(formula = Predator_SpeciesRichness ~ Herbivore_biomass, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.5113 -1.8948 -0.2293  1.2920  8.8656 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3.781e+00  3.521e-01  10.740  < 2e-16 ***
Herbivore_biomass 1.425e-04  2.282e-05   6.246 6.51e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.645 on 121 degrees of freedom
Multiple R-squared:  0.2438,	Adjusted R-squared:  0.2375 
F-statistic: 39.01 on 1 and 121 DF,  p-value: 6.511e-09
```

Transformation of the response variable

- Model assumptions

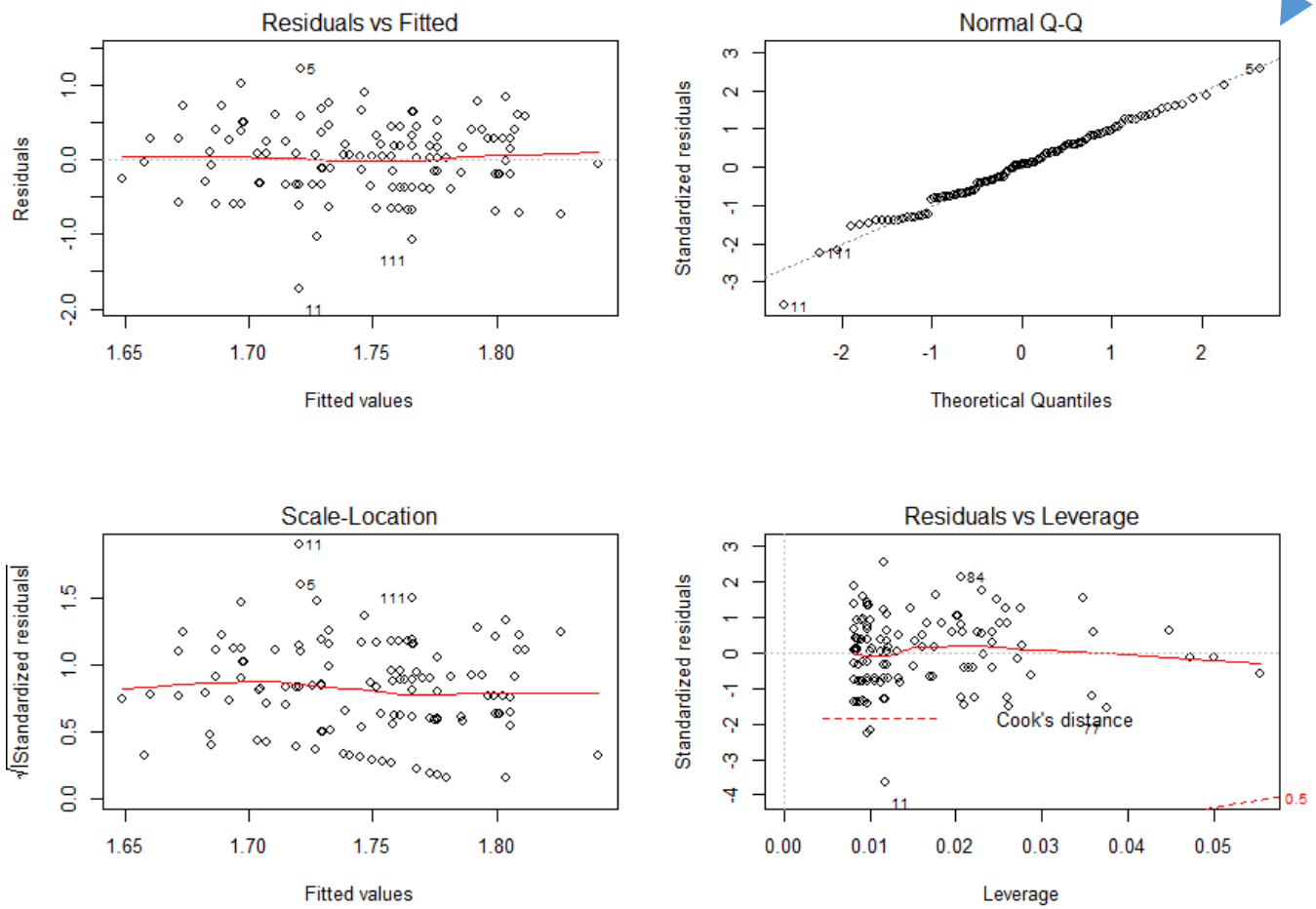
```
> plot(mod0.1)
```



Transformation of the response variable

- After Y transformation - model assumptions

```
> plot(lm(log(Predator_SpeciesRichness+1)~LUI,data=dat))
```



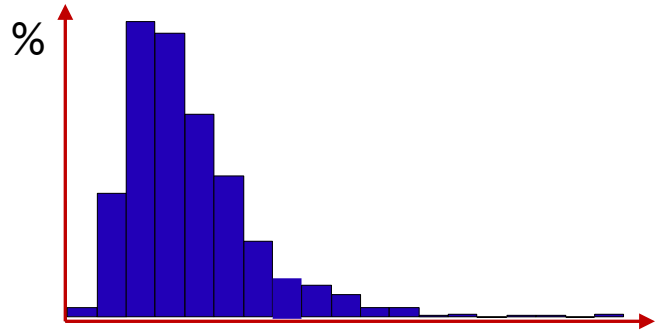
Common transformations

Arcsine-transformation `asin(x)` or arcsine-square root - transformation `asin(sqrt(x))` for proportion data

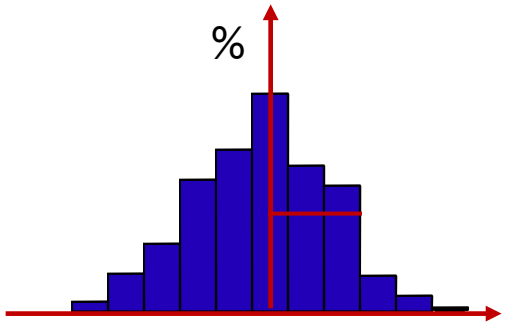
Square-root transformation `sqrt(x + 0.5)` for count data

Log-transformation `log(x)` or `log (x+1)` for continuous, positive data

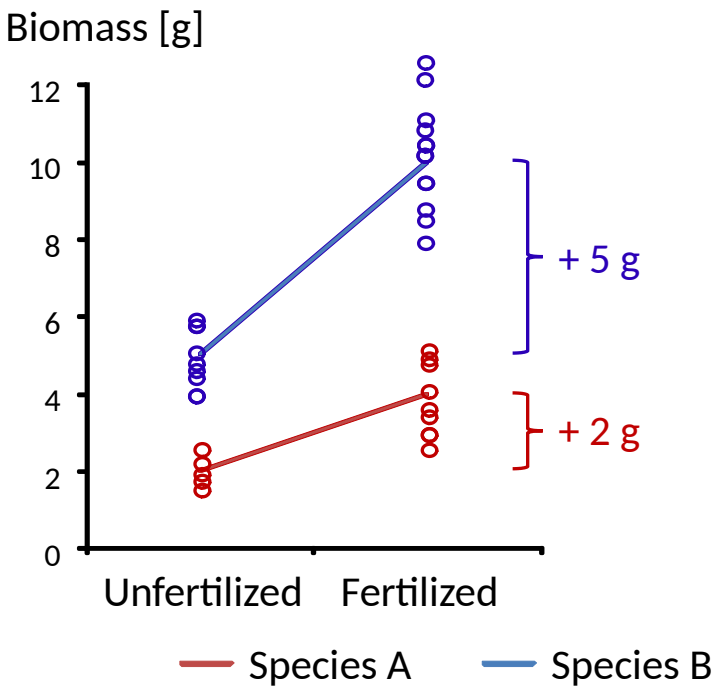
Log-Normal distribution

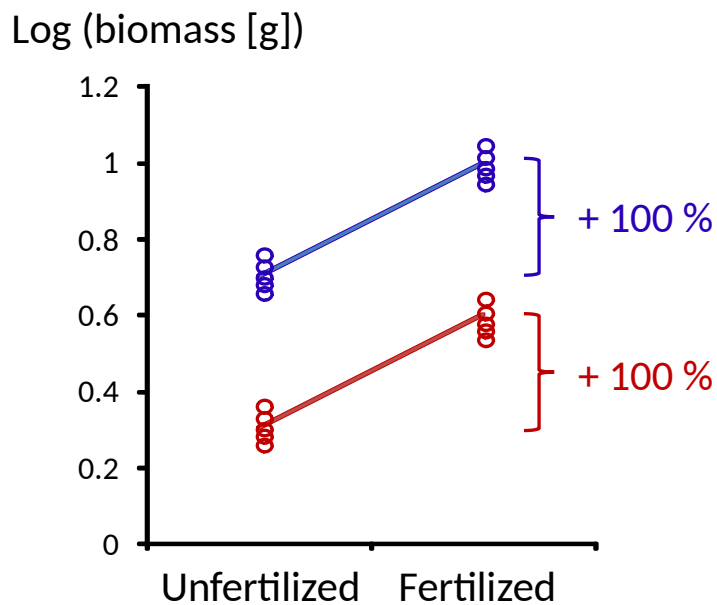


Normal distribution



Warning





Species x fertilizer interaction:
Species respond differently to fertilizer

No species x fertilizer interaction:
Species respond the same to fertilizer

The logarithm changes multiplicative into additive interactions

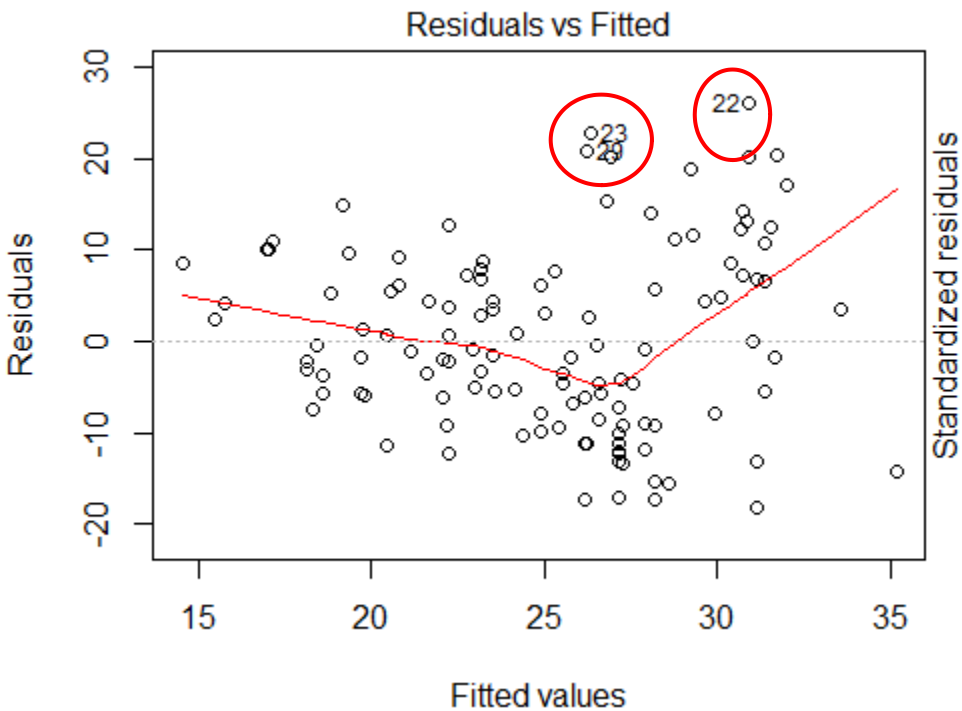
$$\log(a \times b) = \log(a) + \log(b)$$

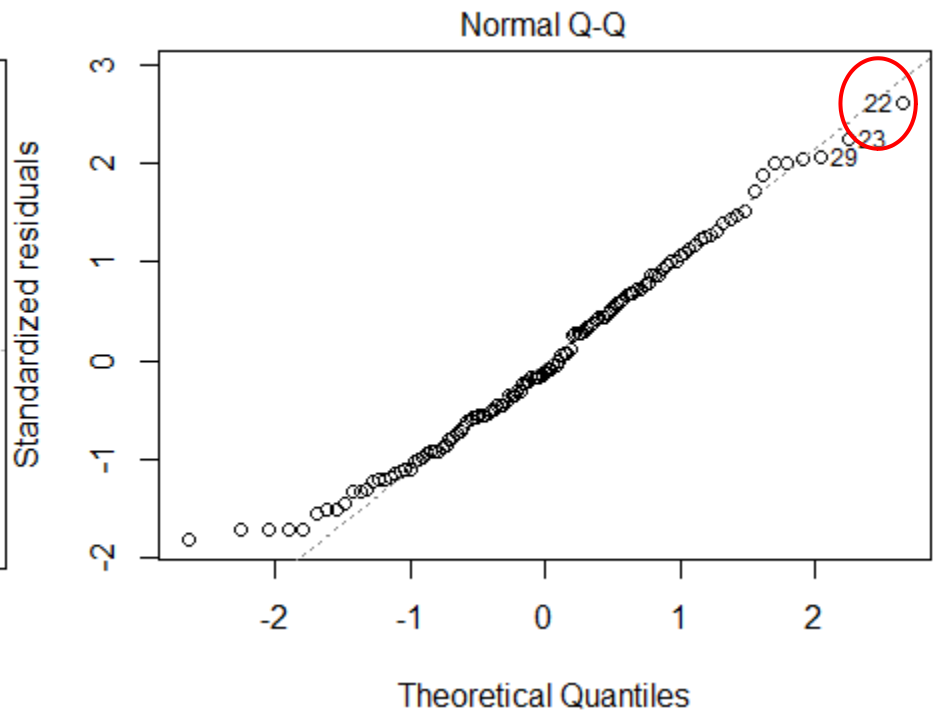
Absolute differences are transformed into relative differences

Another important point: outliers and leverage points

Outliers

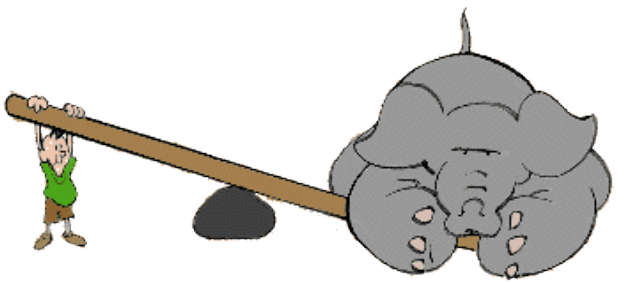
- **Outliers** are **extreme values** in the **response variable (y)**, either as actual values or as residuals.
  Models cannot predict outliers as well as other data points.

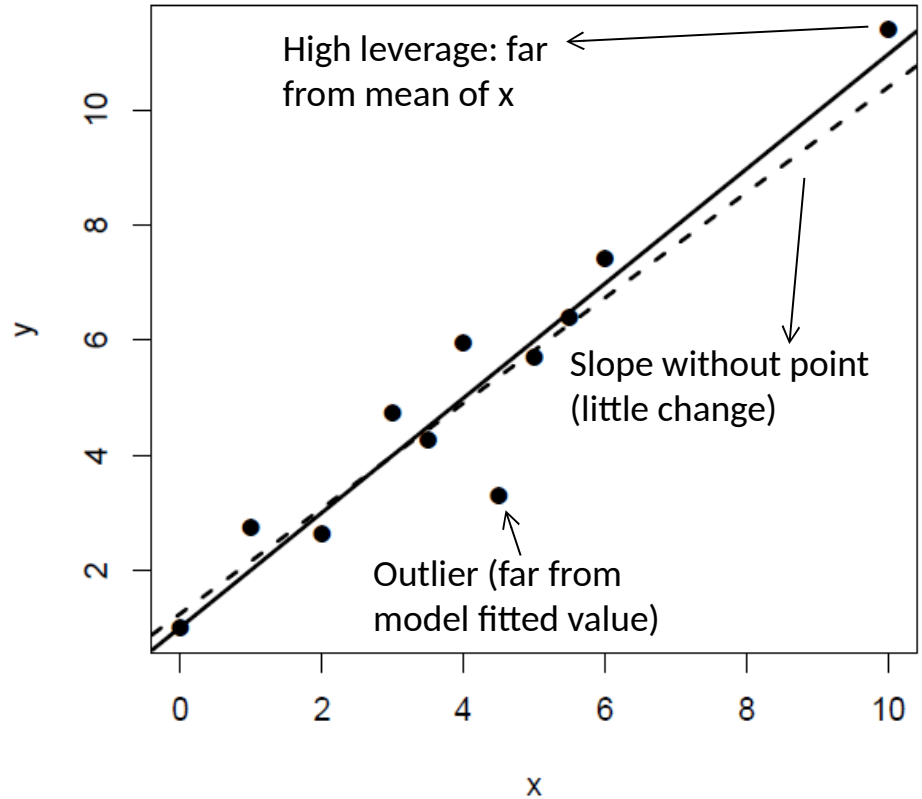




Leverage, outliers and influence

- Data points with high **leverage** are those with **extreme values** in the **explanatory variable (x)**.



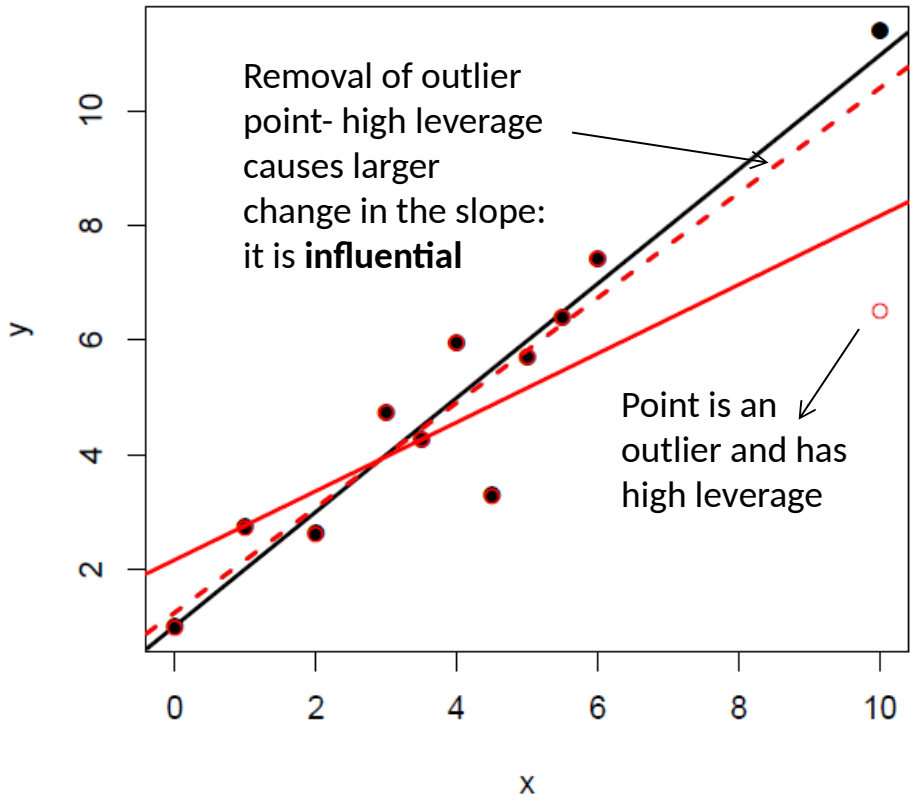


- Leverage measures how extreme values of x are. It is measured as a 'hat' value (h_i), and is between 0 and 1:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1){s_x}^2}$$

Leverage, outliers and influence

- When points have a high leverage AND are outliers for y, they are likely to be influential (having a larger effect on parameter estimate).

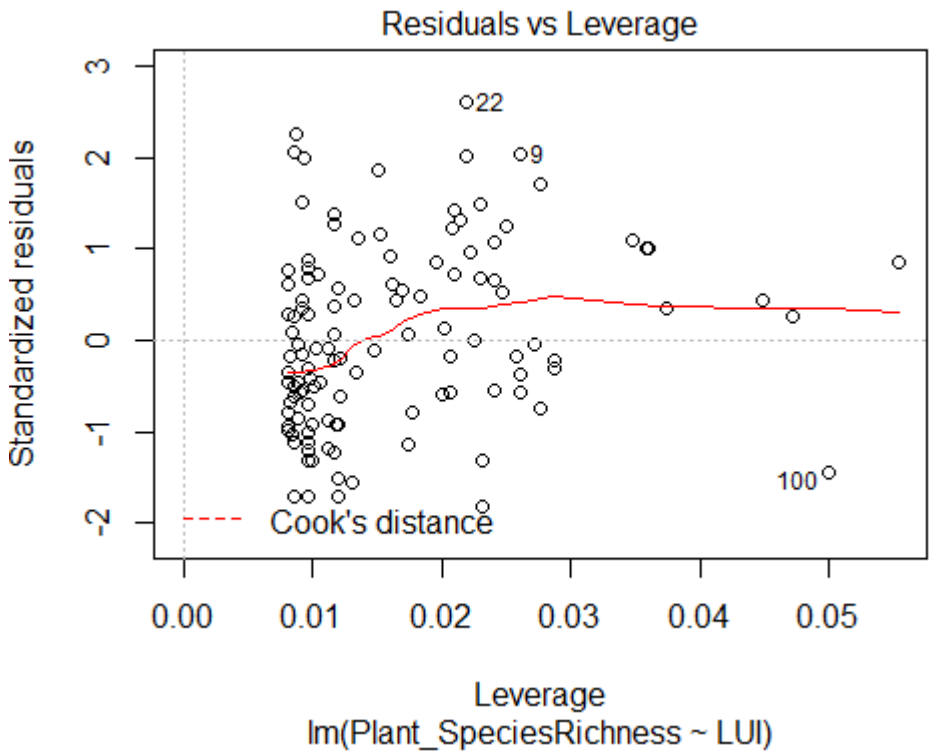


Cook's Distance, a measure of influenc

$$D_i = \frac{\sum_{j=1}^{n}(\widehat{Y}_j - \widehat{Y_{j(i)}})^2}{pMS_E}$$

- Default 4th plot shows Cook's Distances of observations as contours

- How influential is too influential?
  - CDi > 4/(n-k-1)
    k = number of regression slopes
    n = number of observations
  - Points noticeably different from the majority

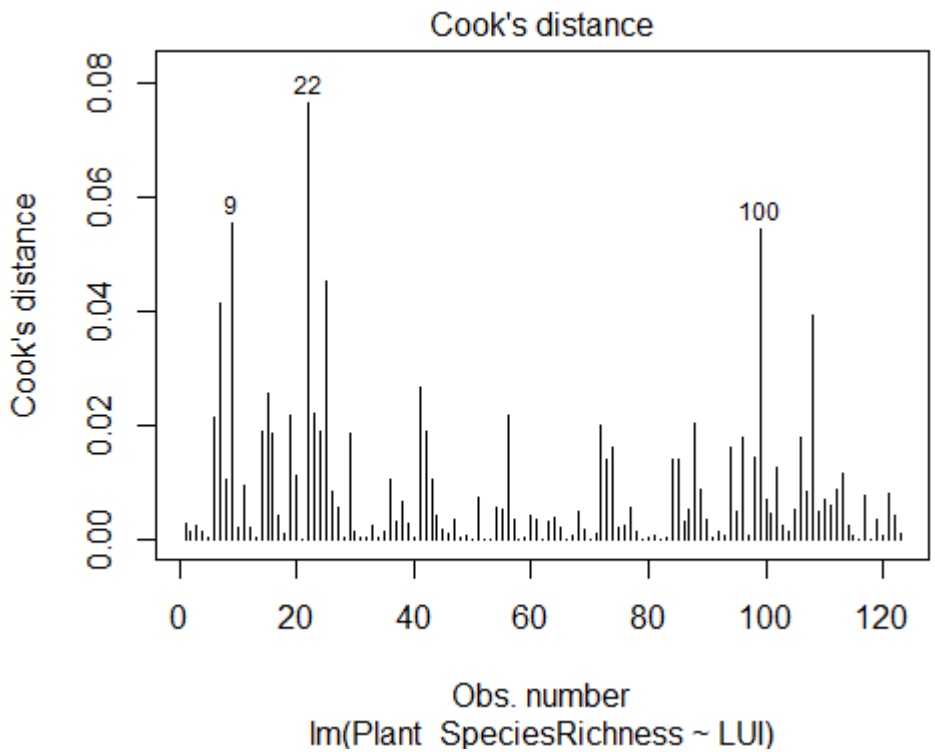


Cook's Distance, a measure of influence

- Alternative plot shows Cook's Distances per observation.

```
> cutoff <- 4/((nrow(dat)-length(mod0$coefficients)-2))
> plot(mod0, which=4, cook.levels=cutoff)
```

Not -1 because we also don't count the intercept



TAKE-HOME : Outliers

- There are ways to detect (influential) outliers

- Inspect your outliers closely
  - get to know your data
  - try to find out why they are outliers

- BUT : don’t remove them, they belong to your data
  - except you have a very very good reason

Multiple explanatory variables

Explanatory variables are corrected for each other

```
Call:
lm(formula = Plant_SpeciesRichness ~ LUI, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.165  -7.605  -1.167   7.038  26.069 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   35.182      2.261  15.558  < 2e-16 ***
LUI           -6.317      1.306  -4.838 3.91e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.11 on 121 degrees of freedom
Multiple R-squared:  0.1621,	Adjusted R-squared:  0.1552 
F-statistic: 23.41 on 1 and 121 DF,  p-value: 3.905e-06
```

```
Call:
lm(formula = Plant_SpeciesRichness ~ LUI + Herbivore_SpeciesRichness, 
    data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-18.260  -6.479  -1.371   4.485  23.158 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)               20.43141    3.65666   5.587 1.46e-07 ***
LUI                       -4.75092    1.23885  -3.835 0.000202 ***
Herbivore_SpeciesRichness  0.40180    0.08205   4.897 3.07e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.271 on 120 degrees of freedom
Multiple R-squared:  0.3016,	Adjusted R-squared:   0.29 
F-statistic: 25.92 on 2 and 120 DF,  p-value: 4.411e-10
```

anova() versus summary()

The anova() output is sequence-dependent!
The summary() output gives sequence-independent parameter estimates

```
mod1 <- lm(PlantR ~ LUI+ herbivoreR)
anova(mod1)
```

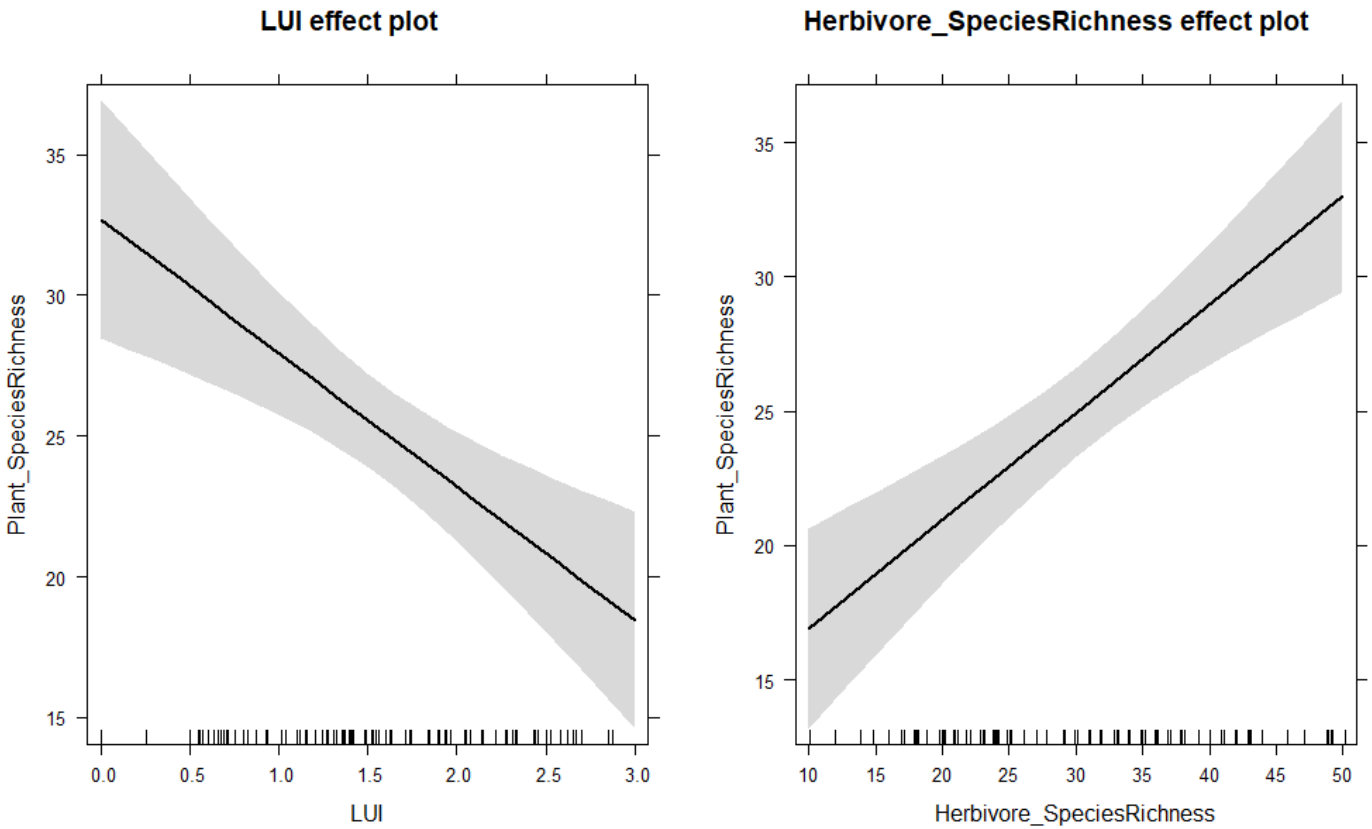
```
Analysis of Variance Table

Response: Plant_SpeciesRichness
                           Df  Sum Sq Mean Sq F value    Pr(>F)
LUI                         1  2393.9 2393.92  27.852 5.903e-07 ***
Herbivore_SpeciesRichness   1  2061.2 2061.21  23.981 3.069e-06 ***
Residuals                 120 10314.3   85.95
```

```
mod1 <- lm(PlantR ~ herbivore + LUI)
anova(mod1)
```

```
Analysis of Variance Table

Response: Plant_SpeciesRichness
                           Df  Sum Sq Mean Sq F value    Pr(>F)
Herbivore_SpeciesRichness   1  3191.0  3191.0  37.126  1.38e-08 ***
LUI                         1  1264.1  1264.1  14.707 0.0002016 ***
Residuals                 120 10314.3    86.0
```

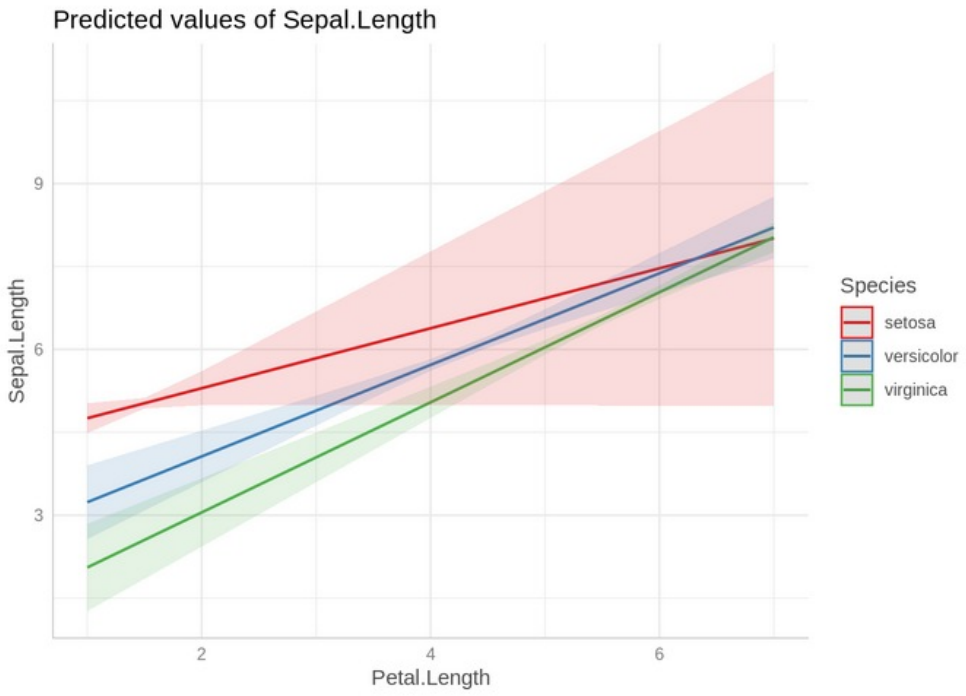
Effect package for multiple regressions

```
> plot(effect("LUI", mod1))
> plot(allEffects(mod1))
>
```



Effect package for multiple regressions

- `ggeffects` package creates nicer plots



https://rstudio-pubs-static.s3.amazonaws.com/490532_5b04406f4d154f2d9b5a99ff488d644a.html

Models with continuous and categorical explanatory variables - same approach!

```
> mod2 <- lm(Plant_SpeciesRichness ~ Region + LUI + Herbivore_SpeciesRichness, data=dat)
> summary(mod2)
```

```
Call:
lm(formula = Plant_SpeciesRichness ~ Region + LUI + Herbivore_SpeciesRichness,
    data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-14.3183  -4.3133  -0.3375   5.0681  14.7320

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               29.57873    2.69456  10.977  < 2e-16 ***
RegionSCH                 -8.61440    1.42981  -6.025 1.98e-08 ***
RegionALB                  8.50926    1.42975   5.952 2.79e-08 ***
LUI                       -6.50518    0.85397  -7.618 7.09e-12 ***
Herbivore_SpeciesRichness  0.19450    0.05798   3.355  0.00107 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.252 on 118 degrees of freedom
Multiple R-squared:  0.6877,    Adjusted R-squared:  0.6771
F-statistic: 64.96 on 4 and 118 DF,  p-value: < 2.2e-16
```

1. Factors order!

```
> levels(dat$Region)
[1] "HAI" "SCH" "ALB"
```

Default is **alphabetical order**

Models with continuous and categorical explanatory variables - same approach!

```
> mod2 <- lm(Plant_SpeciesRichness ~ Region + LUI + Herbivore_SpeciesRichness, data=dat)
> summary(mod2)
```

```
Call:
lm(formula = Plant_SpeciesRichness ~ Region + LUI + Herbivore_SpeciesRichness,
    data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-14.3183  -4.3133  -0.3375   5.0681  14.7320

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               29.57873    2.69456  10.977  < 2e-16 ***
RegionSCH                 -8.61440    1.42981  -6.025 1.98e-08 ***
RegionALB                  8.50926    1.42975   5.952 2.79e-08 ***
LUI                       -6.50518    0.85397  -7.618 7.09e-12 ***
Herbivore_SpeciesRichness  0.19450    0.05798   3.355  0.00107 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.252 on 118 degrees of freedom
Multiple R-squared:  0.6877,    Adjusted R-squared:  0.6771
F-statistic: 64.96 on 4 and 118 DF,  p-value: < 2.2e-16
```

1. Factors order!
2. Intercept refers to the first group (HAI)
3. For the other groups, estimates are deviations from the values of the **first** group

TAKE-HOME : Models with multiple explanatory variabes

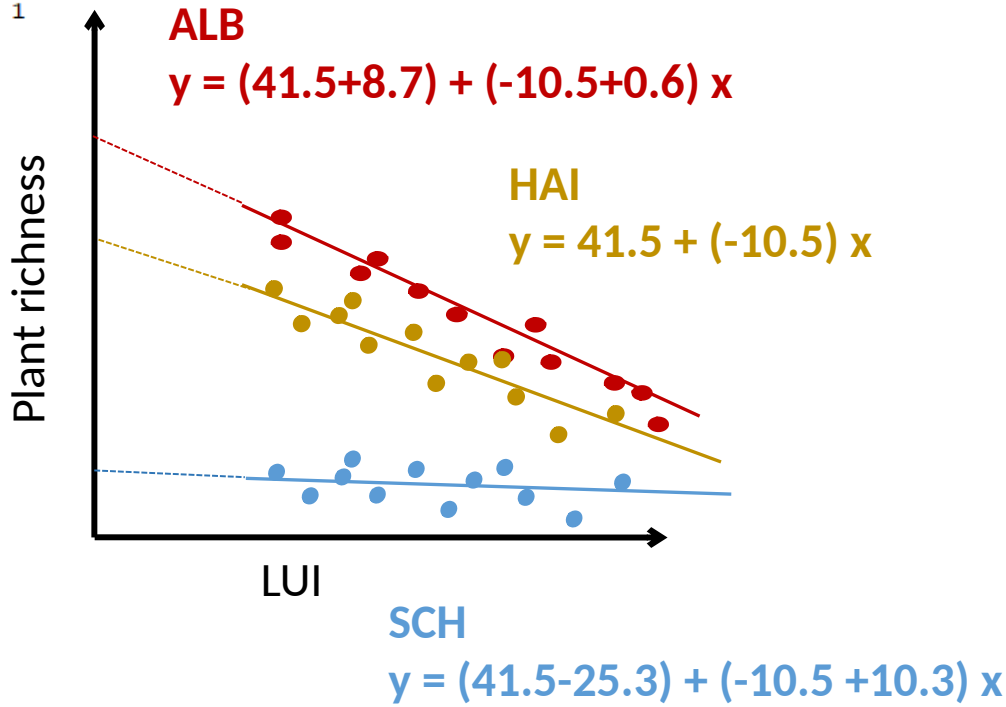
- variables are corrected for each other
  - prefer one model with several explanatory variables over individual models

Models with interactions

```
> mod3 <- lm(Plant_SpeciesRichness ~ Region*LUI, data=dat)
> summary(mod3)

Call:
lm(formula = Plant_SpeciesRichness ~ Region * LUI, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.8683  -3.4265  -0.1381   4.0117  13.2839 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    41.5734     2.2652  18.353  < 2e-16 ***
RegionSCH     -25.3738     3.2981  -7.693 4.97e-12 ***
RegionALB       8.7798     3.1188   2.815  0.00572 ** 
LUI           -10.5412     1.4278  -7.383 2.46e-11 ***
RegionSCH:LUI  10.3845     2.0148   5.154 1.05e-06 ***
RegionALB:LUI   0.6793     1.8235   0.373  0.71018    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.756 on 117 degrees of freedom
Multiple R-squared:  0.7376,	Adjusted R-squared:  0.7263 
F-statistic: 65.76 on 5 and 117 DF,  p-value: < 2.2e-16
```

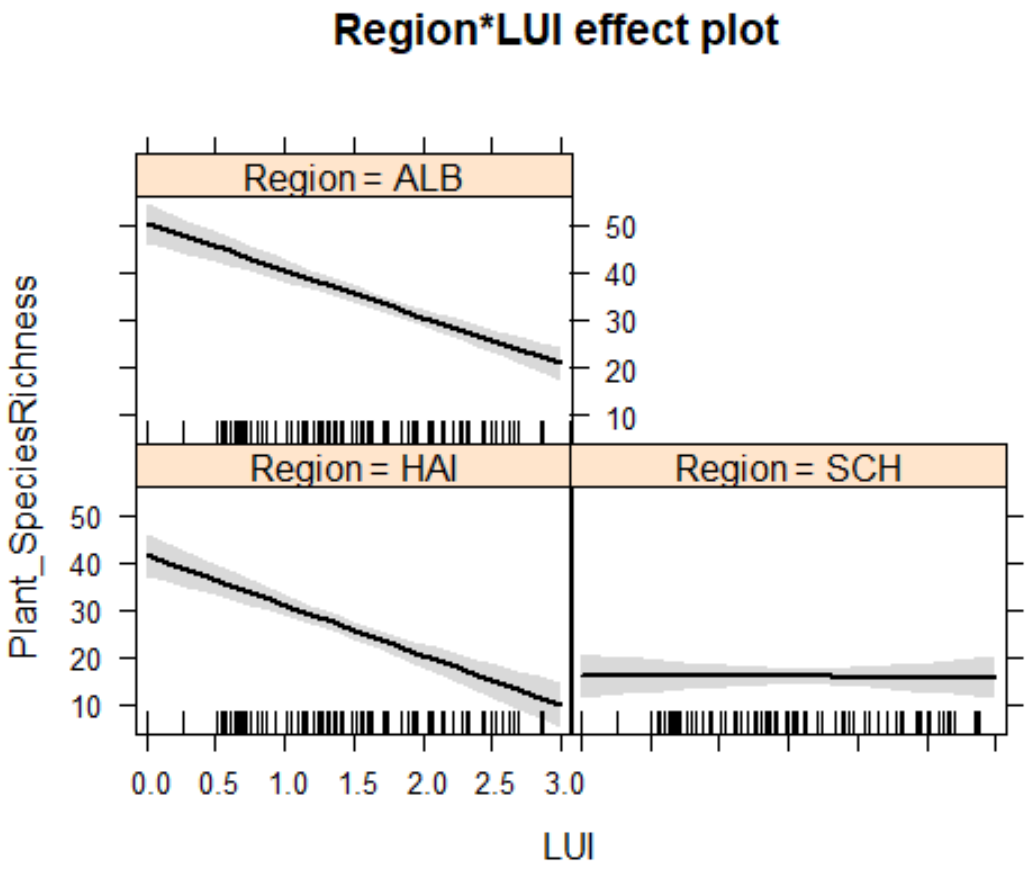
Models with interactions

```
> mod3 <- lm(Plant_SpeciesRichness ~ Region*LUI, data=dat)
> summary(mod3)

Call:
lm(formula = Plant_SpeciesRichness ~ Region * LUI, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.8683  -3.4265  -0.1381   4.0117  13.2839 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    41.5734     2.2652  18.353  < 2e-16 ***
RegionSCH     -25.3738     3.2981  -7.693 4.97e-12 ***
RegionALB       8.7798     3.1188   2.815  0.00572 ** 
LUI           -10.5412     1.4278  -7.383 2.46e-11 ***
RegionSCH:LUI  10.3845     2.0148   5.154 1.05e-06 ***
RegionALB:LUI   0.6793     1.8235   0.373  0.71018    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.756 on 117 degrees of freedom
Multiple R-squared:  0.7376,	Adjusted R-squared:  0.7263 
F-statistic: 65.76 on 5 and 117 DF,  p-value: < 2.2e-16
```



Visualize interactions

```
> plot(allEffects(mod3))
```



TAKE-HOME : Interactions

- Visualisation helps to understand them
- Can only tell you that “there is a difference”, not why

Centering and scaling = Standardizing (z-scores)

Standardized variable = (observation-mean)/SD

- **Mean=0 and SD=1**

- helps **getting rid of collinearity** (centering)
- **makes regression coefficients comparable** when original variables were on a different scale (compare slope strength) (scaling)

- Use original scale for predictions!

- **Don't be confused by the function name in R:**

```
scale(x, center=TRUE, scale=TRUE)
```

Effects of centering and scaling on estimates

Centering affects intercepts, scaling affects slopes.

```
> mod4 <- lm(Plant_SpeciesRichness ~ LUI + Herbivore_SpeciesRichness, data=dat)
> summary(mod4)

Call:
lm(formula = Plant_SpeciesRichness ~ LUI + Herbivore_SpeciesRichness,
    data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-18.260  -6.479  -1.371   4.485  23.158

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               20.43141    3.65666   5.587 1.46e-07 ***
LUI                       -4.75092    1.23885  -3.835 0.000202 ***
Herbivore_SpeciesRichness  0.40180    0.08205   4.897 3.07e-06 ***
```

ORIGINAL VARIABLES

```
Call:
lm(formula = Plant_SpeciesRichness ~ LUI_s + Herbivore_SpeciesRichness_s,
    data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-18.260  -6.479  -1.371   4.485  23.158

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                  25.1707     0.8359  30.111  < 2e-16 ***
LUI_s                        -3.3318     0.8688  -3.835 0.000202 ***
Herbivore_SpeciesRichness_s   4.2545     0.8688   4.897 3.07e-06 ***
```

STANDARDIZED

TAKE-HOME : Standardisation

- If you want to compare slopes of explanatory variables
- If you want to get rid of collinearity
  → standardise

- If you want to do predictions
  → don’t standardise

Models with quadratic terms

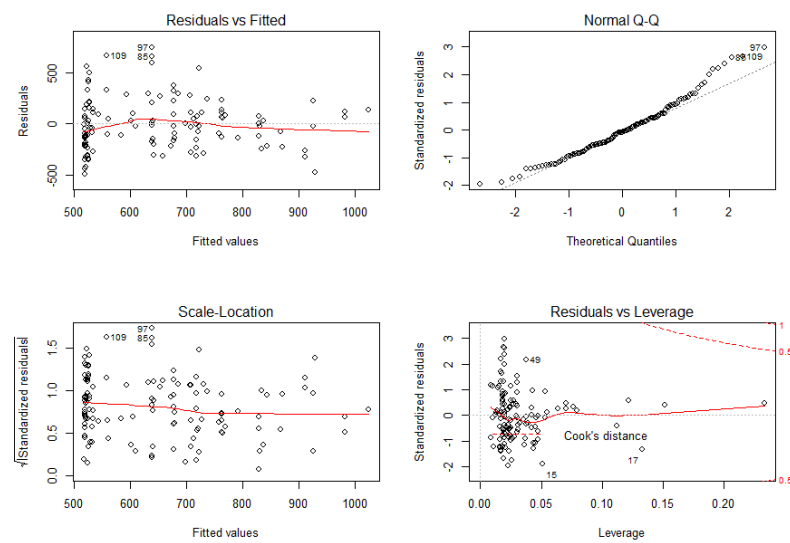
```
> mod5 <- lm(Plant_SpeciesRichness ~ Region + LUI + I(LUI^2), data=dat)
> summary(mod5)

Call:
lm(formula = Plant_SpeciesRichness ~ Region + LUI + I(LUI^2),
    data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2900  -4.7291  -0.4768   4.9542  15.1762

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   41.724      2.889  14.444  < 2e-16 ***
RegionSCH     -9.284      1.451  -6.400 3.28e-09 ***
RegionALB      8.793      1.468   5.990 2.33e-08 ***
LUI          -14.342      3.616  -3.967 0.000125 ***
I(LUI^2)       2.117      1.061   1.995 0.048303 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.436 on 118 degrees of freedom
Multiple R-squared:  0.6691,    Adjusted R-squared:  0.6579
F-statistic: 59.65 on 4 and 118 DF,  p-value: < 2.2e-16
```

It is still a LINEAR model : the parameters are linearly related to the response variable

Potential problems in linear models

- What is the effect of the single LUI components?

```
> mod6 <- lm(Plant_biomass ~ Fstd + Gstd + Mstd, data=dat)
> summary(mod6)

Call:
lm(formula = Plant_biomass ~ Fstd + Gstd + Mstd, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max 
-490.60 -177.48  -20.05  127.62  756.57 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  518.872     43.702  11.873  < 2e-16 ***
Fstd          28.623     22.375   1.279  0.20331    
Gstd           3.638     17.266   0.211  0.83348    
Mstd          98.501     36.744   2.681  0.00839 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 

Residual standard error: 255.3 on 119 degrees of freedom
Multiple R-squared:  0.2093,	Adjusted R-squared:  0.1894 
F-statistic:  10.5 on 3 and 119 DF,  p-value: 3.529e-06
```



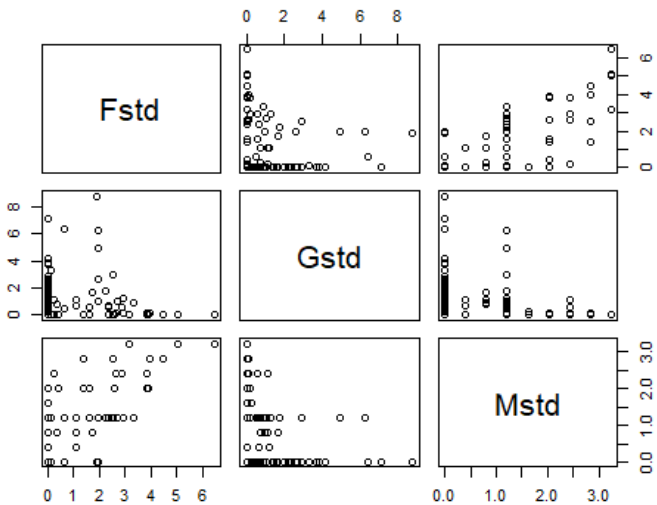
Model assumptions are met

But.. No effect of fertilisation?

Potential problems in linear models

- What is relationship between the single LUI components?

```
> pairs(dat[,c("Fstd","Gstd", "Mstd")])
```




```
> cor(dat[,c("Fstd","Gstd", "Mstd")])
           Fstd       Gstd       Mstd
Fstd  1.0000000 -0.1253154  0.6937203
Gstd -0.1253154  1.0000000 -0.4343348
Mstd  0.6937203 -0.4343348  1.0000000
```

Multicollinearity: Variance inflation factors

- Variance inflation factor: how much the variance of a regression coefficient is inflated due to multicollinearity in the model
- function vif() in car package):

```
> vif(mod6)
    Fstd     Gstd     Mstd
2.080829 1.330413 2.524365
```

- VIFs > 2 indicate multicollinearity (roughly, some people use 5)
- Choose variables to keep based on biology
- Never throw all possible variables in a model, think first

Generic functions

`anova (my.model)` : returns the ANOVA table
`summary()` : returns the model parameters, including t-test

`residuals()` : returns the residuals of the model
`predict()` : returns the predicted values of the model

`plot()` : creates four graphs for model checking

`lm()` : linear models; assumes normal residuals
`aov()` : analysis of variance; assumes normal **and balanced** data

Use `anova(lm())`

Limitations of linear models

- non-normal distribution of data
  - if data transformation not applicable
  - e.g. counts, proportions, success-failures, binary response
  - use glm
- data are not independent
  - nested design, e.g. multiple measures per plot/ individual
  - use mixed models : fixed and random effects
    - fixed term : like in lm
    - random term : account for non-independence of data