Kurs Bio144:
Datenanalyse in der Biologie

Stefanie Muff & Owen L. Petchey

Lecture 4: Multiple linear regression

11. March 2019

9th March 2020

Overview

- Checking the assumptions of linear regression: the QQ-plot
- Multiple predictors x_1, x_2, ..., x_m
- R^2 in multiple linear regression
- t-tests, F-tests and p-values
- Binary and factor covariates

↳ explanatory variable
predictor
independent

Course material covered today

The lecture material of today is based on the following literature:

- Chapters 3.1, 3.2a-q of *Lineare Regression*
- Chapters 4.1 4.2f, 4.3a-e of *Lineare Regression*

Recap of last week I

- The linear regression model for the data $\mathbf{y} = (y_1, \ldots, y_n)$ given $\mathbf{x} = (x_1, \ldots, x_n)$ is

$$y_i = \alpha + \beta x_i + \epsilon_i \,, \qquad \epsilon_i \sim \mathrm{N}(0, \sigma^2) \text{ independent.}$$

- Estimate the parameters α, β and σ^2 by least squares.
- The estimated parameters $\hat{\alpha}$, $\hat{\beta}$ contain uncertainty and are normally distributed around the true values.
- Use the knowledge about the distribution to formulate statistical tests, such as: Is $\beta = 0$?
  $\rightarrow$ T-test with $n - 2$ degrees of freedom.
- All this is done automatically by R:

```
summary(r.bodyfat)$coef

##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -26.984368  2.7689004 -9.745518 3.921511e-19
## bmi           1.818778  0.1083411 16.787522 2.063854e-42
```

Recap of last week II

Remember: The assumption in linear regression is that the residuals follow a $N(0, \sigma^2)$ distribution, implying that :

a) The expected value of ϵ_i is 0: $E(\epsilon_i) = 0$.

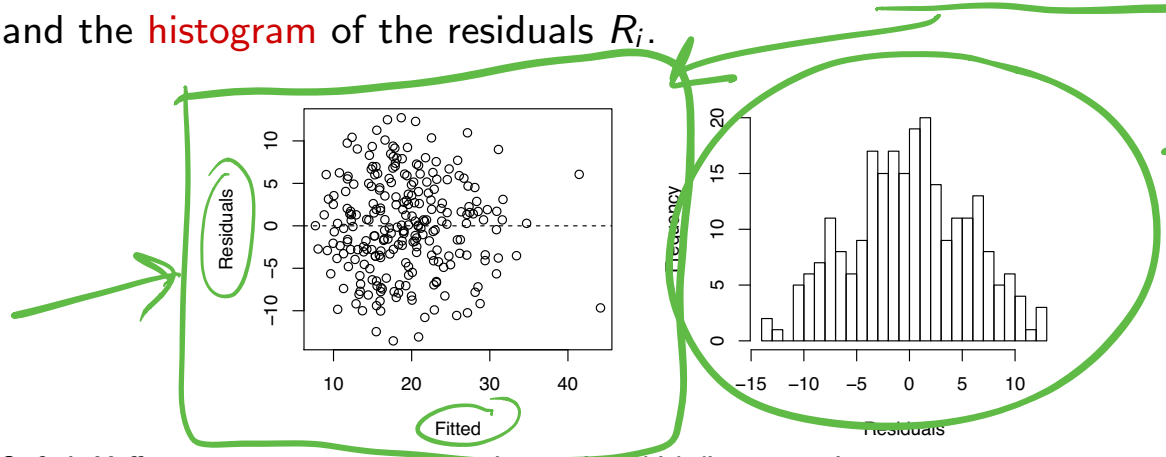
b) All ϵ_i have the same variance: $Var(\epsilon_i) = \sigma^2$.

c) The ϵ_i are normally distributed.

d) The ϵ_i are independent of each other.

non-linear relationship.

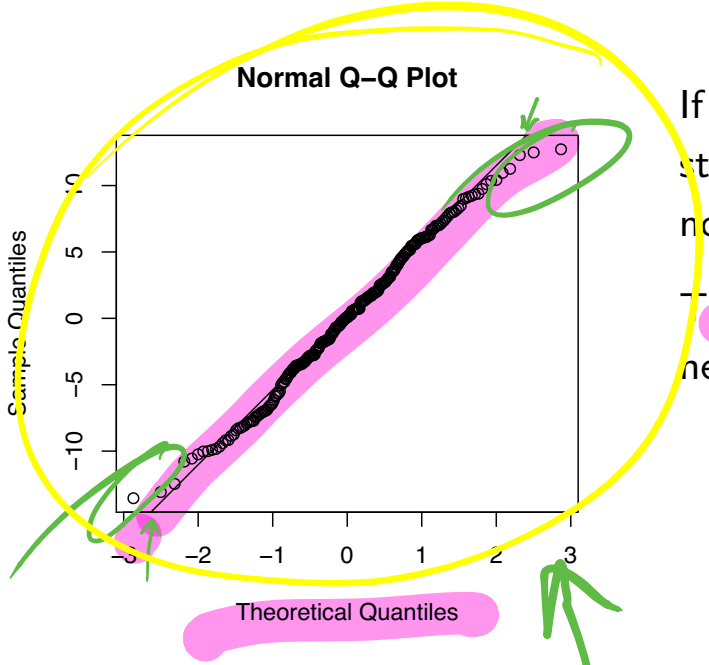
We started to do some residual analysis using the Tukey-Anscombe plot and the histogram of the residuals R_i.



residuals ε_i

Another useful diagnostic plot: The QQ-plot

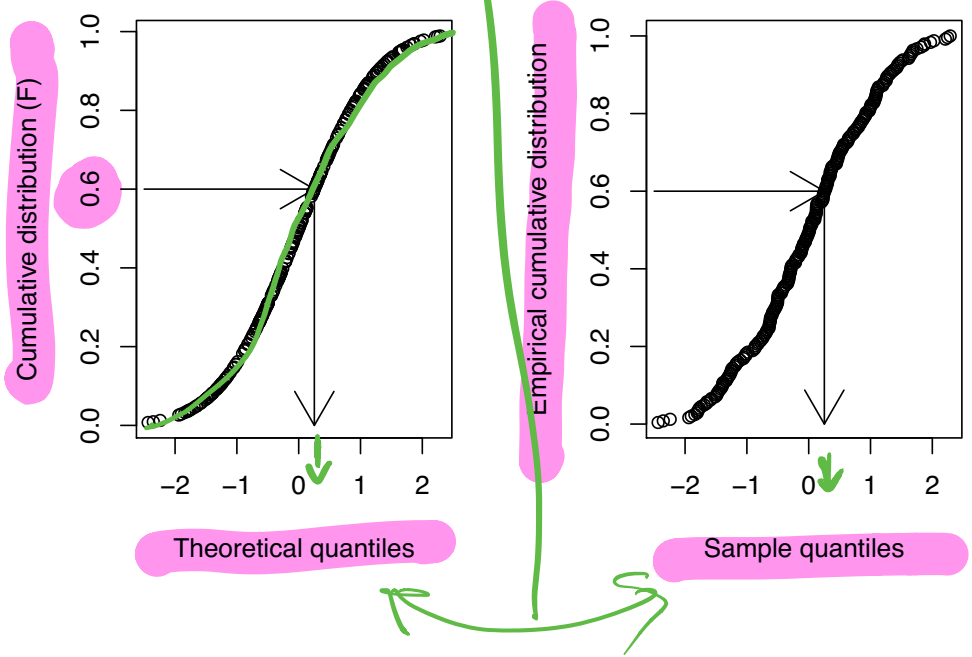
Usually, not the histogram of the residuals is plotted, but the so-called quantile-quantile (QQ) plot. The quantiles of the observed distribution are plotted against the quantiles of the respective theoretical (normal) distribution:



If the points lie approximately on a straight line, the data is fairly normally distributed.

This is often "tested" by eye, and needs some experience.

The idea is that, for each observed point, theoretical quantiles are plotted against the sample quantiles.



Optional:You may want to watch the youtube video for more explanation **given here**.

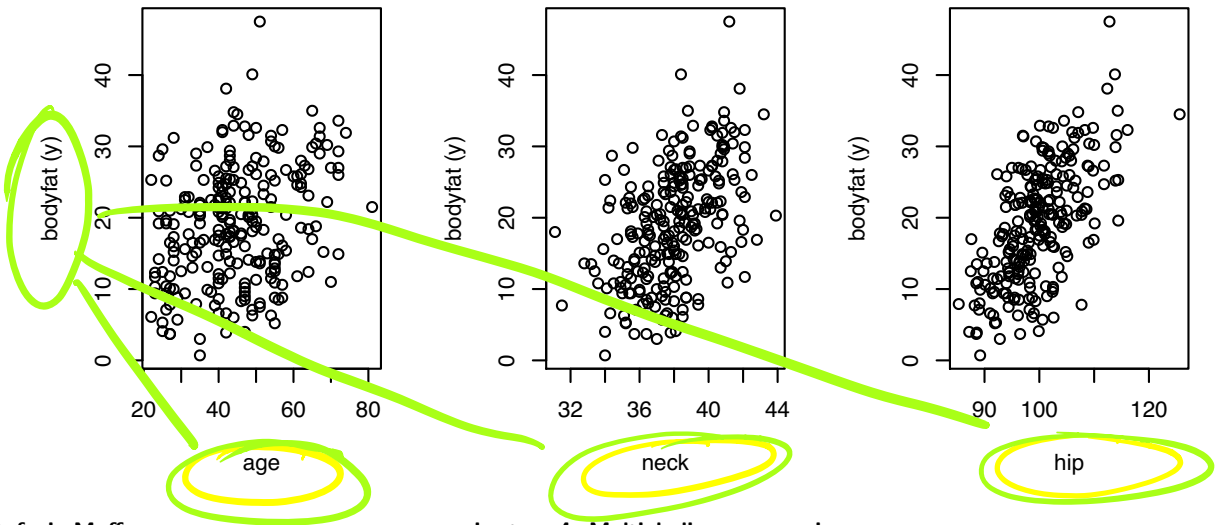
Multiple linear regression

Bodyfat example

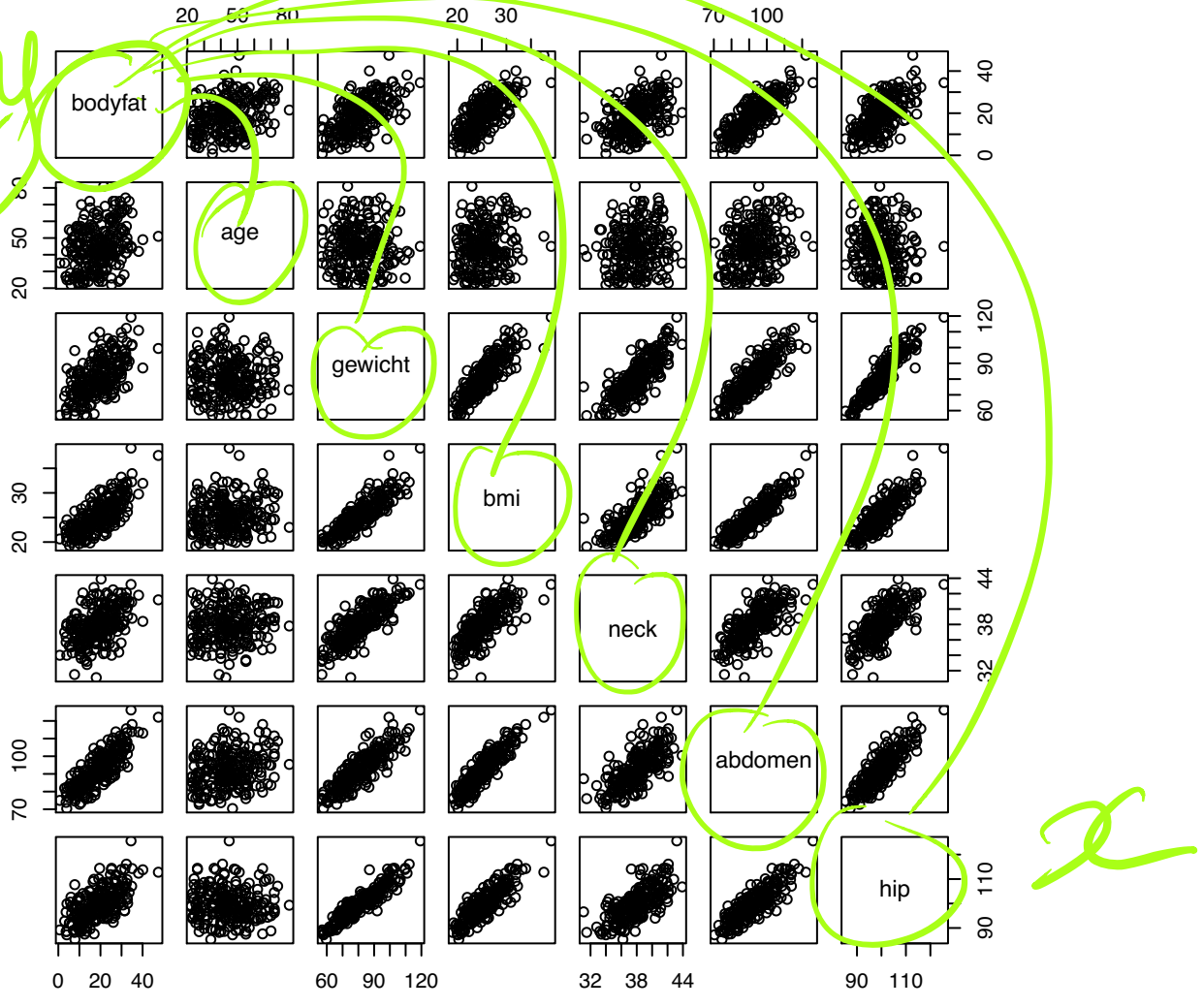
We have so far modeled bodyfat in dependence of bmi, that is:
$(bodyfat)_i = \alpha + \beta \cdot bmi_i + \epsilon_i$.

However, other predictors might also be relevant for an accurate prediction of bodyfat.

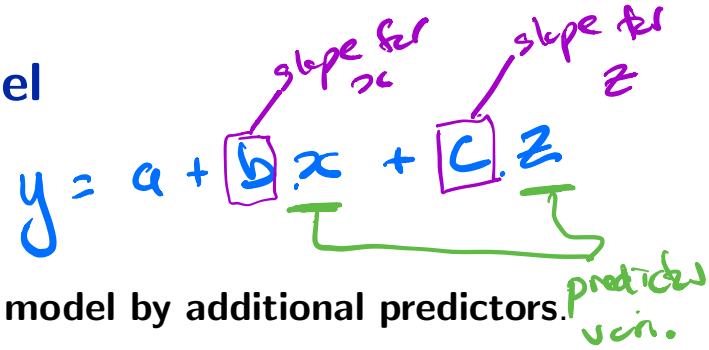
Examples: Age, neck fat (Nackenfalte), hip circumference, abdomen circumference etc.



Or again the pairs plot:



Multiple linear regression model



The idea is simple: Just **extend the linear model by additional predictors**.

- Given several influence factors $x_i^{(1)}, \ldots, x_i^{(m)}$, the straightforward extension of the simple linear model is

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)} + \epsilon_i$$

$$\text{with } \epsilon_i \sim N(0, \sigma^2).$$

- The parameters of this model are $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m)$ and σ^2.

The components of β are again estimated using the **least squares** method. Basically, the idea is (again) to minimize

$$\sum_{i=1}^{n} e_i^2$$

with

$$e_i = y_i - (\beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_m x_i^{(m)})$$

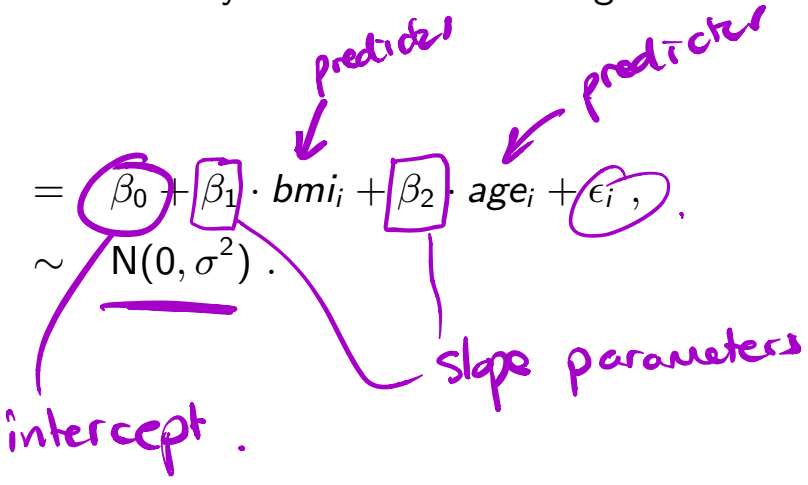
It is a bit more complicated than for simple linear regression, see Section 3.4 of the Stahel script.

→ Some **linear algebra** is needed to understand these sections, but we do not look into this for the moment.

Multiple linear regression for bodyfat

Let us regress the proportion (%) of bodyfat (from last week) on the predictors **bmi** and **age** simultaneously. The model is thus given as

$$
\begin{aligned}
(bodyfat)_i &= \beta_0 + \beta_1 \cdot bmi_i + \beta_2 \cdot age_i + \epsilon_i \,, \\
\text{with} \quad \epsilon_i &\sim N(0, \sigma^2) \,.
\end{aligned}
$$



Before we estimate the parameters, let us ask the questions that we intend to answer:

whole model

F-test

1. Is the **ensemble** of all covariates associated with the response?
2. If yes, which covariates are associated with the response? — t-test
3. Which proportion of response variability (σ_y^2) is explained by the model? — r^2

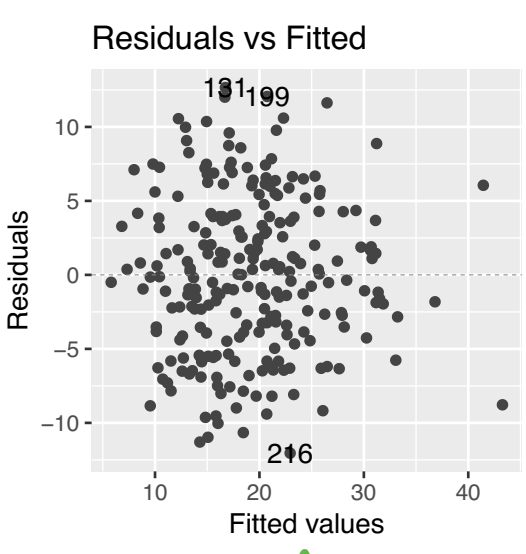
Multiple linear regression with R

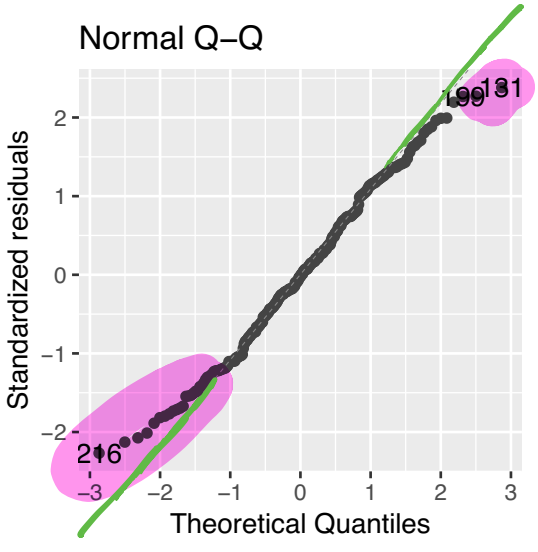
Let's now fit the model with R, and quickly glance at the output:

```
r.bodyfatM <- lm(bodyfat ~ bmi + age, d.bodyfat)

summary(r.bodyfatM)

##
## Call:
## lm(formula = bodyfat ~ bmi + age, data = d.bodyfat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0415  -3.8725  -0.1237   3.9193  12.6599
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -31.25451    2.78973 -11.203  < 2e-16 ***
## bmi           1.75257    0.10449  16.773  < 2e-16 ***
## age           0.13268    0.02732   4.857 2.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.329 on 240 degrees of freedom
## Multiple R-squared:  0.5803,Adjusted R-squared:  0.5768
## F-statistic: 165.9 on 2 and 240 DF,  p-value: < 2.2e-16
```

Model checking

Before we look at the results, we have to check if the modelling assumptions are fulfilled:





ε: $\hat{y}$

nothing to worry about ☺

This seems ok, so continue with answering questions 1-3.

Question 1: Are the covariates associated with the response?

ensemble of covariates

To answer question 1, we need to perform a so-called *F*-test. The results of the test are displayed in the final line of the regression summary. Here, it says:

```
F-statistic:  165.9 on 2 and 240 DF, p-value:  < 2.2e-16
```

So apparently (and we already suspected that) the model has some explanatory power.

*The *F*-statistic and -test is briefly recaptured in 3.1.f) of the Stahel script, but see also Mat183 chapter 6.2.5. It uses the fact that

$$\frac{SSQ^{(R)}/m}{SSQ^{(E)}/(n-p)} \sim F_{m,n-p}$$

follows an *F*-distribution with m and $(n-p)$ degrees of freedom, where m are the number of variables, n the number of data points, p the number of β-parameters (typically $m+1$). $SSQ^{(E)} = \sum_{i=1}^{n} R_i^2$ is the squared sum of the residuals, and $SSQ^{(R)} = SSQ^{(Y)} - SSQ^{(E)}$ with $SSQ^{(y)} = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

Question 2: Which variables are associated with the response?

```
summary(r.bodyfatM)$coef

##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -31.2545057 2.78973238 -11.203406 1.039096e-23
## bmi           1.7525705 0.10448723  16.773060 2.600646e-42
## age           0.1326767 0.02731582   4.857137 2.149482e-06
```

slope

To answer this question, again look at the *t*-tests, for which the *p*-values are given in the final column. Each *p*-value refers to the test for the null hypothesis $\beta_0^{(j)} = 0$ for covariate $x^{(j)}$.

As in simple linear regression, the T-statistic for the j-th covariate is calculated as

$$T_j = \frac{\hat{\beta}_j}{se^{(\beta_j)}}, \qquad (1)$$

with $se^{(\beta_j)}$ given in the second column of the regression output.

The distribution of this statistic is $T_j \sim t_{n-p}$.

Therefore: A "small" p-value indicates that the variable is relevant in the model.

Here, we have

- $p < 0.001$ for bmi
- $p < 0.001$ for age

Thus both, bmi and age seem to be associated with bodyfat.

Again, a 95% CI for $\hat{\beta}_j$ can be calculated with R:

```
confint(r.bodyfatM)

##                 2.5 %      97.5 %
## (Intercept) -36.7499929 -25.7590185
## bmi           1.5467413   1.9583996
## age           0.0788673   0.1864861
```

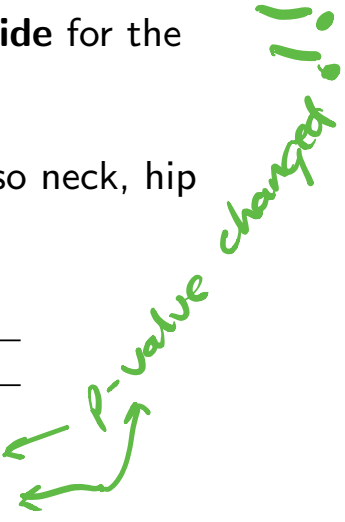
(The CI is again $[\hat{\beta} - c \cdot \sigma^{(\beta)}; \hat{\beta} + c \cdot \sigma^{(\beta)}]$, where c is the 97.5% quantile of the t-distribution with $n - p$ degrees of freedom; compare to slides 38-40 of last week).

!However!:

The p-value and T-statistics should only be used as a **rough guide** for the "significance" of the coefficients.

For illustration, let us extend the model a bit more, including also neck, hip and abdomen:

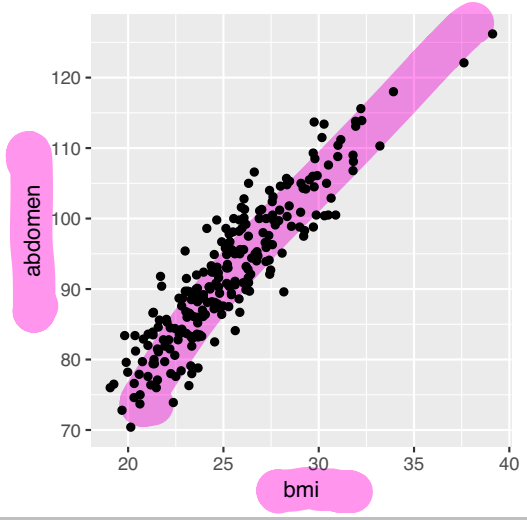
	Coefficient	95%-confidence interval	p-value
Intercept	-7.75	from -22.13 to 6.63	0.29
bmi	0.43	from -0.03 to 0.88	0.066
age	0.015	from -0.04 to 0.07	0.60
neck	-0.80	from -1.18 to -0.43	< 0.0001
hip	-0.32	from -0.53 to -0.11	0.003
abdomen	0.84	from 0.67 to 1.00	< 0.0001



It is now much less clear how strongly age ($p = 0.60$) and bmi ($p = 0.07$) are associated with bodyfat.

Basically, the problem is that the variables in the model are correlated and therefore explain similar aspects of bodyfat.

Example: Abdomen (Bauchumfang) seems to be a relevant predictor and it is obvious that abdomen and BMI are correlated:



This problem of collinearity is at the heart of many confusions of regression analysis, and we will talk about such issues later in the course (lectures 8 and 9).

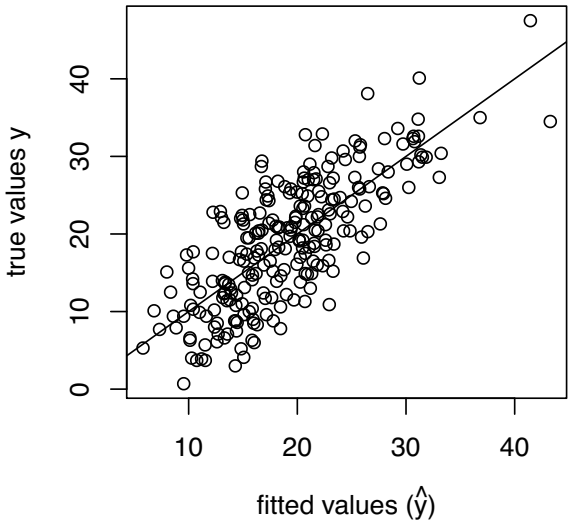
Please see also IC: practical 4 (milk example) for an analysis and more thoughts.

Question 3: Which proportion of variability is explained?

To answer this question, we can look at the multiple R^2 (see Stahel 3.1.h). It is a generalized version of R^2 for simple linear regression:

R^2 for multiple linear regression is defined as the squared correlation between $(y_1, \ldots, y_n)$ and $(\hat{y}_1, \ldots, \hat{y}_n)$, where the $\hat{y}$ are the fitted values

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x^{(1)} + \ldots + \hat{\beta}_m x^{(m)}$$



R^2 is also called the *coefficient of determination* or "Bestimmtheitsmass", because it measures the proportion of the reponse's variability that is explained by the ensemble of all covariates:

- multiple r-squared

$$R^2 = SSQ^{(R)}/SSQ^{(Y)} = 1 - SSQ^{(E)}/SSQ^{(Y)}$$

With

amount of variability explained

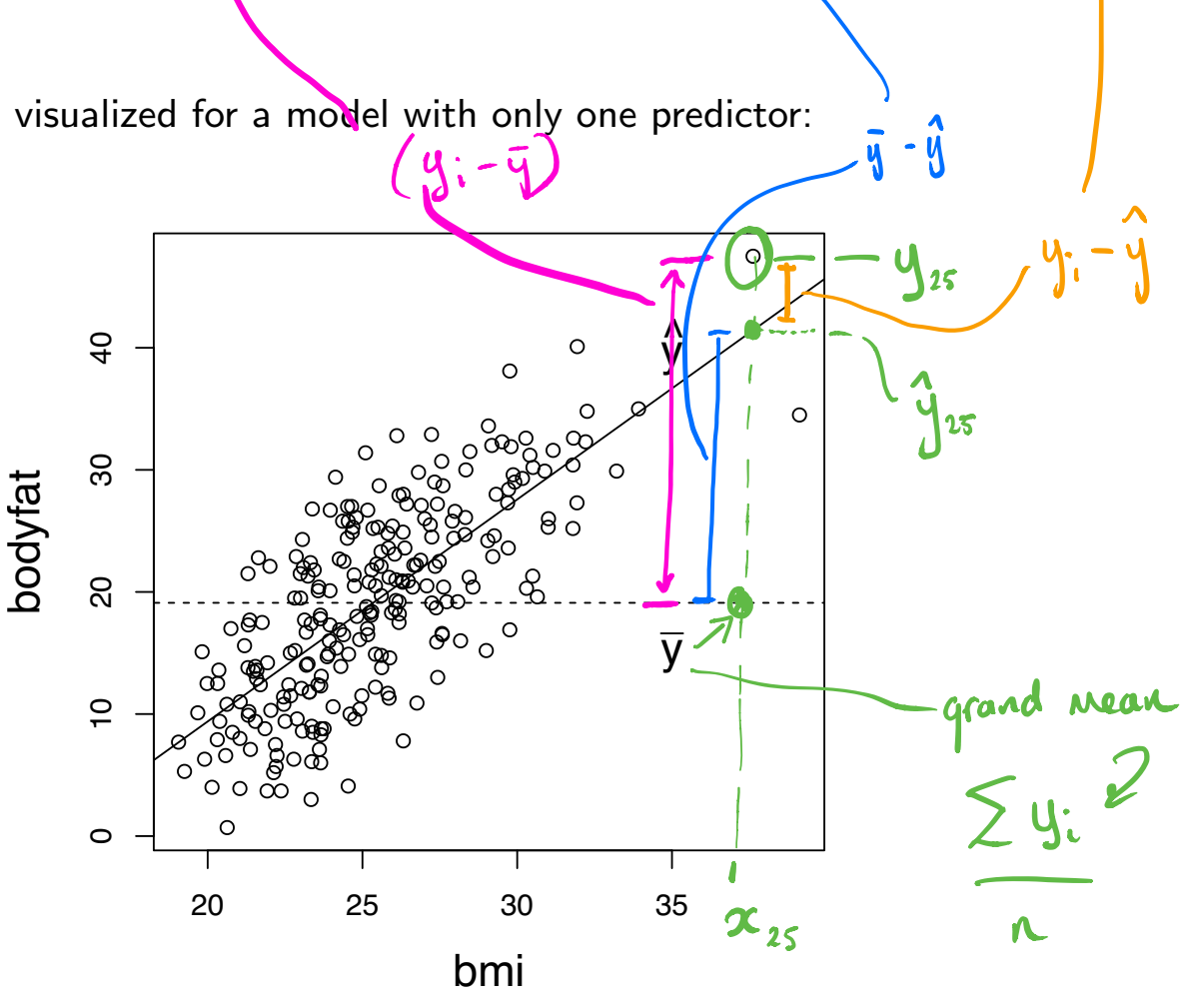
total var. in y

$$\text{total variability} = \text{explained variability} + \text{residual variability}$$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SSQ^{(Y)} = SSQ^{(R)} + SSQ^{(E)}$$

This can be visualized for a model with only one predictor:



Let us look at the R^2s from the three bodyfat models
(model 1: $y \sim bmi$
model 2: $y \sim bmi + age$
model 3: $y \sim bmi + age + neck + hip + abdomen$):

```
summary(r.bodyfat)$r.squared

## [1] 0.5390391

summary(r.bodyfatM)$r.squared

## [1] 0.5802956

summary(r.bodyfatM2)$r.squared

## [1] 0.718497
```

only bmi

bmi + age

bmi, age, neck, hip, abdomen.

The models explain 54%, 58% and 72% of the total variability of y.

It thus *seems* that larger models are "better". However, R^2 does always increase when new variables are included, but this does not mean that the model is more reasonable.

Model selection is a topic that will be treated in more detail later in this course (week 8).

Adjusted R^2

When the sample size n is small with respect to the number of variables m included in the model, an adjusted R^2 gives a better ("fairer") estimation of the actual variability that is explained by the covariates:

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1}$$

Why R_a^2?

It **penalizes for adding more variables** if they do not really improve the model!

Note: R_a^2 may decrease when a new variable is added.

Interpretation of the coefficients

Apart from model checking and thinking about questions 1-3, it is probably even **more important to understand what you see.** Look at the output and ask yourself:

What does the regression output actually mean?

slope = $\frac{\Delta y}{\Delta x}$

	Coefficient	95%-confidence interval	p-value
Intercept	-31.25	from -36.75 to -25.76	< 0.0001
bmi	1.75	from 1.55 to 1.96	< 0.0001
age	0.13	from 0.08 to 0.19	< 0.0001

Table: Parameter estimates of model 2.

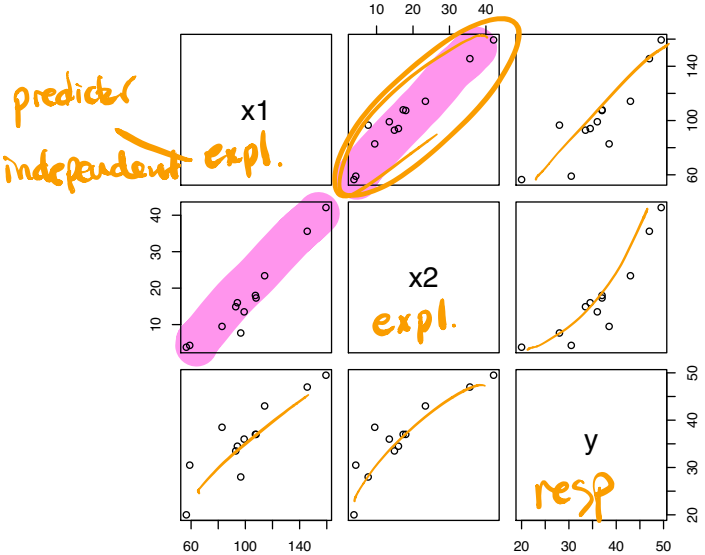
Task in teams: Interpret the coefficients, 95% CIs and p-values.

Example: Catheter Data

Catheter length (y) for heart surgeries depending on two characteristic variables $x^{(1)}$ and $x^{(2)}$ of the patients.

Aim: estimate y from $x^{(1)}$ and $x^{(2)}$ ($n = 12$).

Again look at the data first ($x^{(1)}$ and $x^{(2)}$ are highly correlated!):



Regression results with both variables: $R^2 = 0.81$, $R_a^2 = 0.76$, F-test $p = 0.0006$.

	Coefficient	95%-confidence interval	p-value
Intercept	21.09	from 1.25 to 40.93	0.04
x1	0.077	from -0.25 to 0.40	0.61
x2	0.43	from -0.41 to 1.26	0.28

With x_1 only: $R^2 = 0.78$, $R_a^2 = 0.75$, F-test $p = 0.0002$

	Coefficient	95% confidence interval	p-value
Intercept	12.13	from 2.66 to 21.59	0.017
x1	0.24	from 0.15 to 0.33	0.0002

With x_2 only: $R^2 = 0.80$, $R_a^2 = 0.78$, F-test $p = 0.0001$

	Coefficient	95%-confidence interval	p-value
Intercept	25.63	from 21.16 to 30.09	< 0.0001
x2	0.62	from 0.40 to 0.83	< 0.0001

Questions:

1. Is x_1 an influential covariate? 80% yes
2. Is x_2 an influential covariate? 61% yes
3. Are both covariates needed in the model? 65% yes
4. Interpretation of the results? ~ 40 yes.

NO

→ Go to the klicker link http://www.klicker.uzh.ch/~~k~~ to answer the questions.

nfn

Binary covariates

predictor variables

eg. male / female.

lm()

So far, the covariates x were always continuous.

In reality, there are **no restrictions assumed with respect to the x variables**.

One very frequent data type are **binary** variables, that is, variables that can only attain values 0 or 1.

See section 3.2c of the Stahel script:

If the binary variable x is the only variable in the model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, the model has only two predicted outcomes (plus error):

$$y_i = \begin{cases} \beta_0 + \epsilon_i & \text{if } x_i = 0 \\ \beta_0 + \beta_1 + \epsilon_i & \text{if } x_i = 1 \end{cases}$$

male : $x_i = 0$

female : $x_i = 1$

y_i for males

difference b/t being male + female

Example: Smoking variable in Hg Study

For the 59 mothers in the Hg study, check if their smoking status (0=no,1=yes) influences the Hg-concentration in their urine.

We fit the following linear regression model:

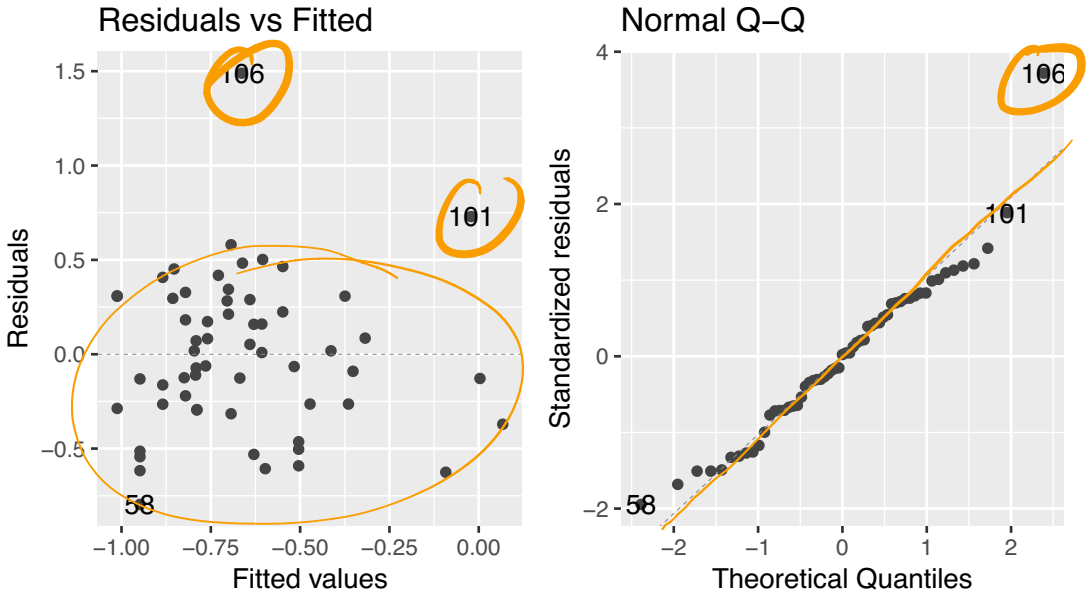
$$\log(Hg_{urin}) = \beta_0 + \beta_1 \cdot x_i^{(1)} + \beta_2 \cdot x_i^{(2)} + \beta_3 \cdot x_i^{(3)} + \epsilon_i \ ,$$

Where

- $\log(Hg_{urin})$ is the urine mercury concentration.
- $x^{(1)}$ is the binary smoking indicator (0/1), denoted as **dummy variable**.
- $x^{(2)}$ the number of amalgam fillings.
- $x^{(3)}$ the monthly number of marine fish meals.

(Remember from week 1 that the log of Hg concentrations is needed to obtain "useful" distributions.)

First check the modelling assumptions:



It seems ok, apart from one point (106) that could be categorized as an outlier. We ignore it for the moment.

The results table is given as follows:

	Coefficient	95%-confidence interval	p-value
Intercept	-1.01	from -1.22 to -0.80	< 0.0001
smoking	0.22	from -0.06 to 0.50	0.12
amalgam	0.092	from 0.05 to 0.14	0.0001
fish	0.032	from 0.01 to 0.06	0.015

There is some weak ($p = 0.12$) indication that smokers have an increased Hg concentration in their body. Their $\log(Hg_{urin})$ is in average by 0.22 higher than for non smokers.

In principle, we have now – at the same time – fitted **two models:** one for smokers and one for non-smokers, assuming that the slopes of the remaining covariates are the same for both groups.

Smokers: $y_i = -1.01 + 0.22 + 0.092 \cdot amalgam_i + 0.032 \cdot fish_i + \epsilon_i$

Non-smokers: $y_i = -1.01 + 0.092 \cdot amalgam_i + 0.032 \cdot fish_i + \epsilon_i$

Factor covariates – categorical predictor variables

Some covariates indicate a **category**, for instance the species of an animal or a plant. This type of covariate is called a **factor**. The trick: convert a factor with k levels (for instance 3 species) into k dummy variables $x_i^{(j)}$ with

$$x_i^{(j)} = \begin{cases} 1, & \text{if the } i\text{th observation belongs to group } j. \\ 0, & \text{otherwise.} \end{cases}$$

Each of the covariates $x^{(1)}, \ldots, x^{(k)}$ can then be included as a binary variable in the model

$$y_i = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_k x^{(k)} + \epsilon_i \ .$$

$\uparrow$ 1 ... 5

species A B C D E

$k = 5$

However: this model is not identifiable*.

* What does that mean? I could add a constant to $\beta_1, \beta_2, \ldots \beta_k$ and subtract it from β_0, and the model would fit equally well to the data, so it cannot be decided which set of the parameters is best.

Solution: One of the k categories must be selected as a *reference category* and is *not included in the model*. Typically: the first category is the reference, thus $\beta_1 = 0$.

The model thus discriminates between the factor levels, such that (assuming $\beta_1 = 0$)

$$\hat{y}_i = \begin{cases} \beta_0 & \text{if } x_i^{(1)} = 1 \\ \beta_0 + \beta_2, & \text{if } x_i^{(2)} = 1 \\ \dots & \\ \beta_0 + \beta_k, & \text{if } x_i^{(k)} = 1 \end{cases}.$$

Please also consult Stahel 3.2e and g.

!!Important to remember!!

(Common aspect that leads to confusion!)

Please note that a factor covariate with k factor levels requires $k - 1$ parameters!

$\rightarrow$ The degrees of freedom are also reduced by $k - 1$.

Example: Earthworm study

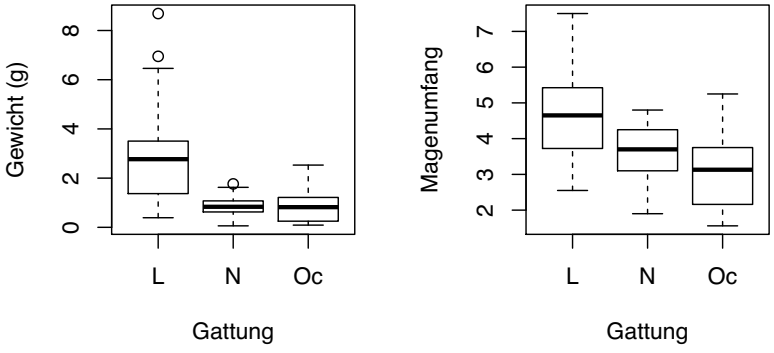
(Angelika von Förster und Burgi Liebst)

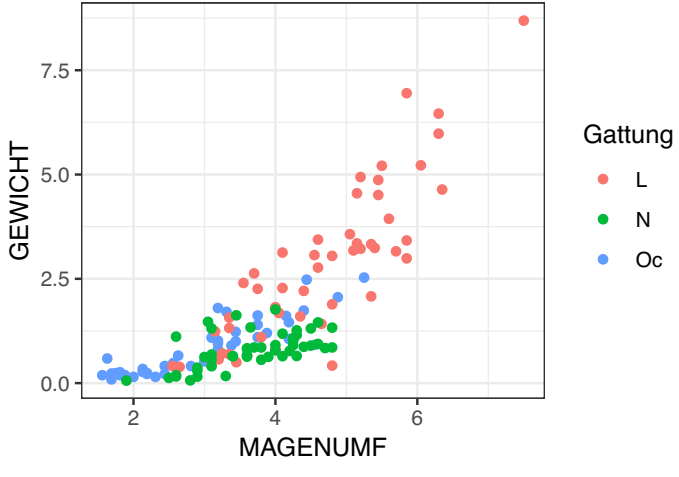
Die Dachse im Sihlwald ernähren sich zu einem grossen Prozentsatz von Regenwürmern. Ein Teil des Muskelmagens der Regenwürmer wird während der Passage durch den Dachsdarm nicht verdaut und mit dem Kot ausgeschieden. Wenn man aus der Grösse des Muskelmagenteilchens auf das Gewicht des Regenwurms schliessen kann, ist die Energiemenge berechnenbar, die der Dachs aufgenommen hat.

Frage: Besteht eine Beziehung zwischen dem Umfang des Muskelmagenteilchens und dem Gewicht des Regenwurms?

Data set of $n = 143$ worms with three species (Lumbricus, Nicodrilus, Octolasion), weight, stomach circumference (Magenumfang).

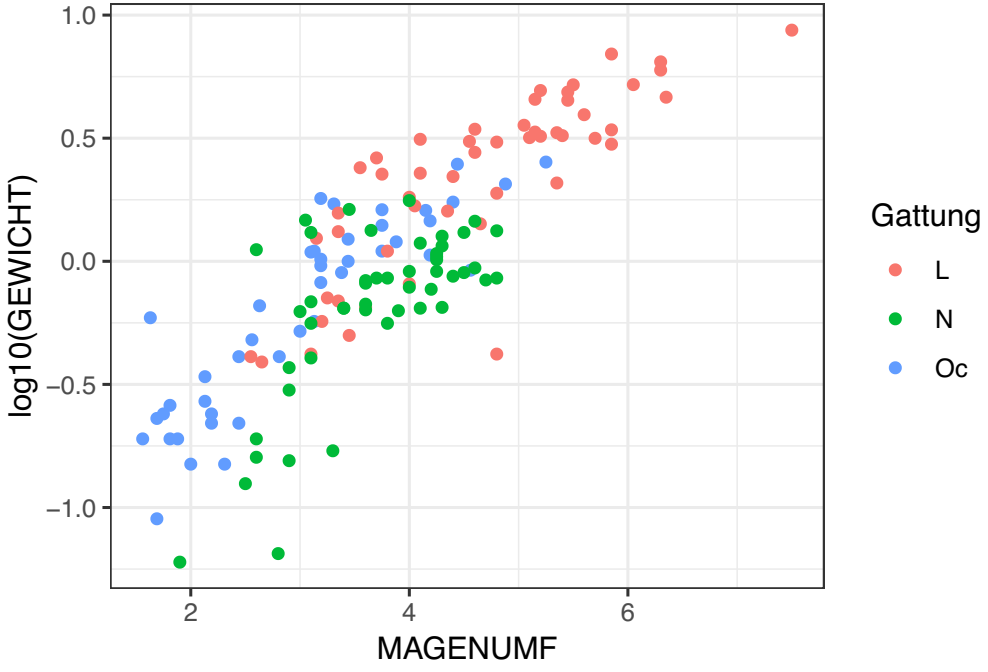
Data inspection suggests that the three species have different weight and stomach sizes:





However, data inspection also suggests that there is not really a linear relationship between weight and stomach size – rather it looks exponential!

Therefore, log-transform the response (weight), and it looks much better:

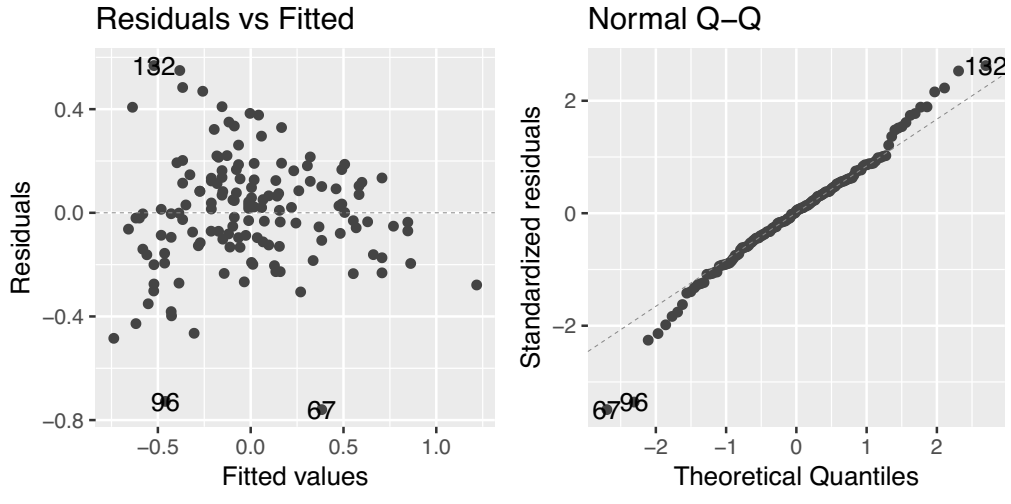


Formulate a model with $\log_{10}$(Gewicht) as response and `Magenumfang` and `Gattung` as covariates. This is simple in R:

(**Hint:** Make sure that `Gattung` is stored as a factor in R (check by `glimpse(d.wurm)`))

```
r.lm <- lm(log10(GEWICHT) ~  MAGENUMF + Gattung,d.wurm)
```

Before doing anything else, check the modeling assumptions:



→ This seems ok (although the TA plot is a bit funnel-like).

Results:

	Coefficient	95%-confidence interval	p-value
Intercept	-1.10	from -1.29 to -0.91	< 0.0001
MAGENUMF	0.31	from 0.27 to 0.35	< 0.0001
GattungN	-0.22	from -0.32 to -0.13	< 0.0001
GattungOc	-0.039	from -0.15 to 0.07	0.48

$R^2 = 0.76$, $R^2_a = 0.75$.

- Question: Why is Gattung Lumbricus (L) not in the results table?
- Answer: L was chosen as the "reference category", thus $\beta_L = 0$.

Degrees of freedom: We had 143 data points. How many degrees of freedom are left for the residual error?

Answer: We estimated 4 parameters, thus $143 - 4 = 139$.

Interpreting the results I

- $\beta_0 = -1.10$ is the intercept.
- $\beta_1 = 0.31$ is the slope for `MAGENUMF`.
- $\beta_2 = -0.22$ is the coefficient for Gattung=Nicodrilus.
- $\beta_3 = -0.039$ is the coefficient for Gattung =Octolasion.
- No coefficient is needed for Gattung Lumbricus, because $\beta_L = 0$.

We have now actually fitted **three** models, one model for each species:

Lumbricus: $\hat{y}_i = -1.10 + 0.31 \cdot \text{MAGENUMF}$

Nicodrilus: $\hat{y}_i = -1.10 + (-0.22) + 0.31 \cdot \text{MAGENUMF}$

Octolasion: $\hat{y}_i = -1.10 + (-0.039) + 0.31 \cdot \text{MAGENUMF}$

Interpreting the results II

Main question: Is there a relation between stomach size and body mass?

Results: `MAGENUMF` has a positive slope estimate with $p < 0.0001$, thus very strong evidence that the relation exists. Increasing `MAGENUMF` by 1 unit increases $\log_{10}$(`GEWICHT`) by $+0.31$.

Moreover, the $R^2 = 0.76$ is relatively high and almost identical to R^2_a.

Interpreting the results III

Question: Is the "Gattung" covariate relevant in the model, that is, do the model intercepts differ for the three species?

Problem: The p-values of the worm species are not very meaningful. They belong to tests that compare the intercept of a factor level with the intercept of the reference level (i.e., the *difference* in intercepts!). However, the question is whether the variable `Gattung` has an effect in total.

Solution: When a factor covariate with k levels is in the model, it occupies $k - 1$ parameters. Therefore, the t-test needs to be replaced by the F-test.

F-test to compare models

F-Test zum Vergleich von Modellen. Die Frage sei, ob die q Koeffzienten β_{j_1}, β_{j_2}, ..., β_{j_q} in einem linearen Regressionsmodell gleich null sein könnten.

- Nullhypothese: $\beta_{j_1} = 0$ und $\beta_{j_2} = 0$ und ... und $\beta_{j_q} = 0$
- Teststatistik:

$$T = \frac{(\mathrm{SSQ}^{(E)*} - \mathrm{SSQ}^{(E)})/q}{\mathrm{SSQ}^{(E)}/(n-p)};$$

  $\mathrm{SSQ}^{(E)*}$ ist die Quadratsumme des Fehlers im „kleinen" Modell, die man aus einer Regression mit den verbleibenden $m-q$ X-Variablen erhält, und p die Anzahl Koeffizienten im „grossen" Modell ($= m+1$, falls das Modell einen Achsenabschnitt enthält, $= m$ sonst).
- Verteilung von T unter der Nullhypolthese: $T \sim \mathcal{F}_{q,n-p}$, F-Verteilung mit q und $n-p$ Freiheitsgraden.

Der Test heisst F-Test zum Vergleich von Modellen. Allerdings kann nur ein kleineres Modell mit einem grösseren verglichen werden, in dem alle X-Variablen des kleinen wieder vorkommen, also mit einem „umfassenderen" Modell. Der früher besprochene F-Test für das gesamte Modell (3.1.e) ist ein Spezialfall: das „kleine" Modell besteht dort nur aus dem Achsenabschnitt β_0.

Remember: $F_{1,n-p} = t^2_{n-p}$

F-test for the earthworms

The function `anova()` in R does the F-test for categorical variables.

```
anova(r.lm)

## Analysis of Variance Table
##
## Response: log10(GEWICHT)
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## MAGENUMF    1 19.7790 19.7790  409.69 < 2.2e-16 ***
## Gattung     2  1.3537  0.6768   14.02 2.842e-06 ***
## Residuals 139  6.7106  0.0483
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: Here, the F-value for `Gattung` is distributed as $F_{2,139}$ under the Null-Hypothesis.

This gives $p = 2.842e - 06$, thus a clear difference in the regression models for the three species ("Gattung is relevant").

Plotting the earthworms results

All species have the same slope (this is a modeling assumption), but different intercepts: