Kurs Bio144:
Datenanalyse in der Biologie

Stefanie Muff & Owen L. Petchey

Week 3: Multiple linear regression
9./10. March 2017

Overview (todo: check)

- Checking the assumptions of linear regression
- Tukey-Anscombe diagram, QQ-plot, leverage plot
- Multiple predictors x_1, x_2, ..., x_p
- R^2 in multiple linear regression
- t-tests, F-tests and p-values
- Transformation of variables

Course material covered today

- Chapters 3.1, 3.2a-q of *Lineare Regression*
- Chapters 4.1 4.2f, 4.3a-e of *Lineare Regression*
- Chapter 11.2 in *Statistische Datenanalyse*

Recap of last week I

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

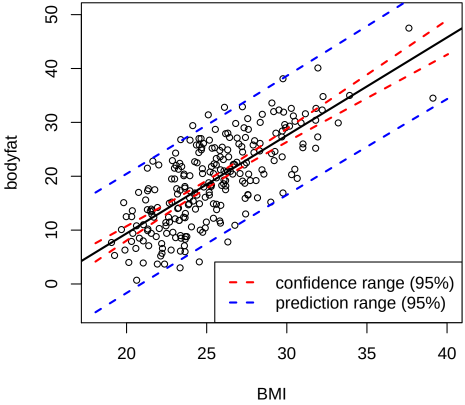
$$y_i = \alpha + \beta x_i + e_i \ , \qquad e_i \sim \mathrm{N}(0, \sigma_e^2) \text{ independent.}$$

- Estimate the parameters α, β and σ_e^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 $\beta = 0$?
- 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

- Confidence and prediction ranges:



Recap of last week III

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

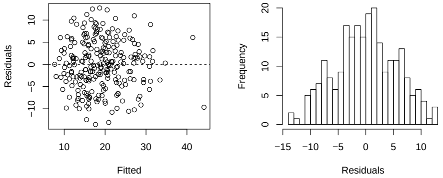
a) The expected value of e_i is 0: $E(e_i) = 0$.

b) All e_i have the same variance: $Var(e_i) = \sigma_e^2$.

c) The e_i are normally distributed.

d) The e_i are independent of each other.

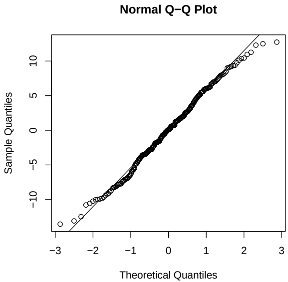
We started to do some residual analysis using the Tukey-Anscombe plot and the Histogram of the residuals R_i.



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:

```
> qqnorm(r.bodyfat$residuals)
> qqline(r.bodyfat$residuals)
```

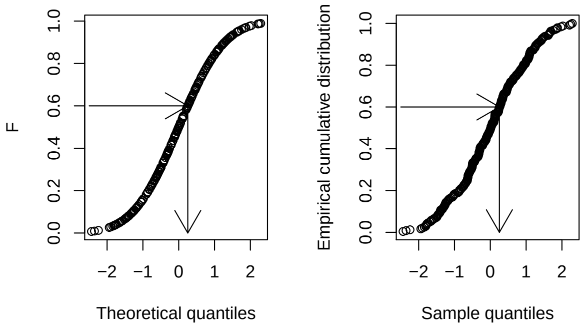


If the points lie approximately on a straight line, the data is fairly normally distributed.

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

Please read "Quantil-Quantil-Diagramme", Chapter 11.2., p.258-261, in "Statistische Datenenalyse" by W. Stahel (Mat183 literature).

It gives a very nice and intuitive description of QQ diagrams!



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

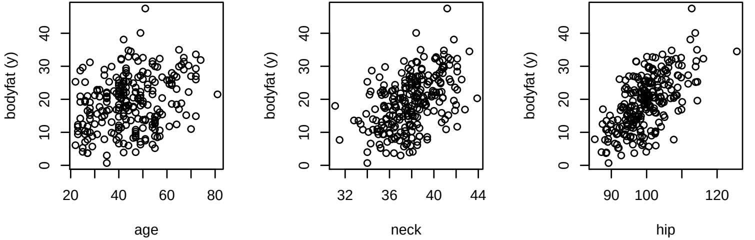
Multiple linear regression

Bodyfat example

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

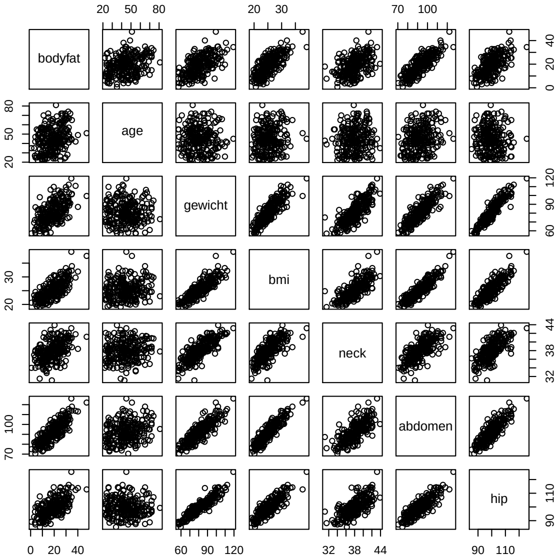
However, other predictors might also be relevant for an accurate prediction of bodyfat.

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



Or again the pairs plot:

```
> pairs(d.bodyfat)
```



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

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \ldots + \beta_2 x_i^{(m)} + e_i \\ \text{with} \quad e_i &\sim \mathrm{N}(0, \sigma_e^2). \end{aligned}$$

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

The components of $\boldsymbol{\beta}$ are again estimated using the **least squares** method. Basically, the idea is (again) to minimize

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

with

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

It is a bit more complicated than for simple linear regression, see Sections 3.3 and 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. (It will come later in week 6.)

Multiple linear regression for bodyfat

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

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

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

1. Does the **ensemble** of all covariates explain a relevant part of the variability of the response?
2. If yes, which influence variables are good predictors of bodyfat?
3. How good is the overall model fit?

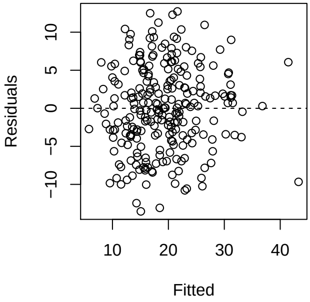
Multiple linear regression with R

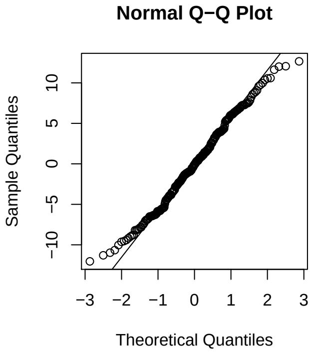
Let's now fit the model with R, and quickly glance at the output:

```
> 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:





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

Question 1: Does the model have some explanatory power?

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 (`df()` in R) 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 $SSQ^{(y)} = \sum_{i=1}^n (y_i - \overline{y})^2$.

Question 2: Which variables influence 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
```

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

$$T_j = \frac{\hat{\beta}_j - \beta_{j_0}}{se^{(\beta_j)}} \underbrace{=}_{if\, \beta_{j_0}=0} \frac{\hat{\beta}_j}{se^{(\beta_j)}} \,, \tag{1}$$

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

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 have some predictive power for bodyfat.

!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:

```
> r.bodyfatM2 <- lm(bodyfat ~ bmi + age + neck + hip + abdomen,d.bodyfat)
> summary(r.bodyfatM2)$coef
```

```
               Estimate Std. Error    t value     Pr(>|t|)
(Intercept) -7.74964673 7.29830233 -1.0618424 2.893881e-01
bmi          0.42647368 0.23132902  1.8435805 6.649276e-02
age          0.01457356 0.02782994  0.5236649 6.010010e-01
neck        -0.80206081 0.19096606 -4.2000177 3.779800e-05
hip         -0.31764315 0.10751209 -2.9544876 3.447492e-03
abdomen      0.83909391 0.08417902  9.9679702 9.035870e-20
```

It is now much less clear what the influences of age ($p = 0.60$) and bmi ($p = 0.06$) are.

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

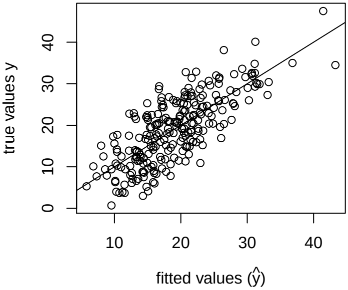
This problem is at the heart of many confusions of regression analysis, and we will talk about such issues later in the course.

Question 3: How good is the overall model fit?

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

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

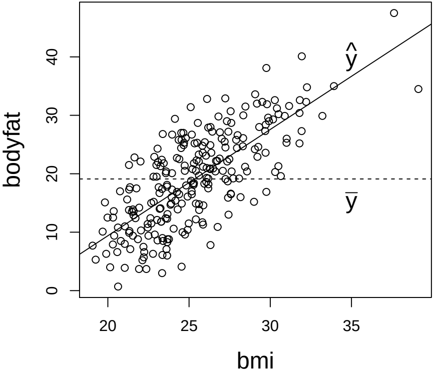


R^2 is also called the *coefficient of determination* or "Bestimmtheitsmass", because it measures the proportion of the reponse's variability that is explained by the ensemble of all covariates:

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

Remembering that

$$\begin{aligned}
\text{total variability} &= \text{explained variability} + \text{residual variability} \\
\sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \qquad + \qquad \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \\
SSQ^{(Y)} &= SSQ^{(R)} \qquad + \qquad SSQ^{(E)}
\end{aligned}$$


ŷ
ȳ
bodyfat
bmi
0
10
20
30
40
20
25
30
35

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: $y \sim bmi + age + neck + hip + abdomen$):

```
[1] 0.5390391
```

```
[1] 0.5802956
```

```
[1] 0.718497
```

The models thus 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.

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 (or "fairer", i.e. unbiased) estimation of the actual variability that is explained by the covariates:

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

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?

	Coefficent	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 3.

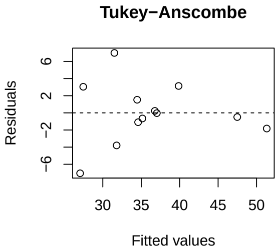
Task in teams: Interpret the coefficients, 95% CIs and p-values.

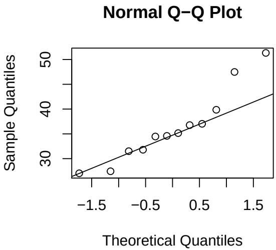
Example: Catheter Data

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 $x^{(2)}$ ($n = 12$). Regression results:

	Coefficent	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 $R^2 = 0.81$, $R_a^2 = 0.76$, p-value of the F-test $p = 0.0006$, and diagnostic residual plots





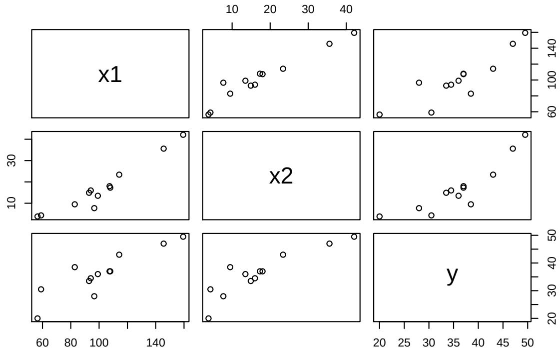
1. Are the modelling assumptions fulfilled?
2. Does the model have some predictive power?
3. Which variable(s) influence(s) the response?
4. How good is the overall fit of the model?
5. Interpretation of the results?

Also look at the regression table when using $x^{(1)}$ and $x^{(2)}$ alone:

	Coefficent	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

	Coefficent	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

```
> pairs(d.cath)
```



Note that $x^{(1)}$ and $x^{(2)}$ are highly correlated!

Binary covariates

So far, the covariates x were always continuous.

However, in our regression models there are **no restrictions assumed with respect to the x variables**.

One very frequent data type of covariates 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 $y_i = \beta_0 + \beta_1 x_i + e_i$, the model has only two predicted outcomes

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

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 the urin.

We fit the following linear regression model:

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

Where

- $\log(Hg_{urin})_i$ is the urine mercury concentration.
- $x^{(1)}$ is the binary smoking indicator (0/1).
- $x^{(2)}$ 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.)

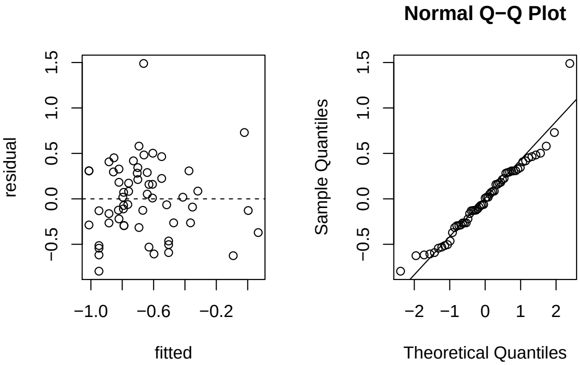
The results table is given as follows:

	Coefficent	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 a slightly increased Hg concentration in their body.
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 other covariates are the same for both groups.

Smokers: $y_i = -1.01 + 0.22 + 0.092 \cdot amalgam_i + 0.032 \cdot fish_i + e_i$
Non-smokers: $y_i = -1.01 + 0.092 \cdot amalgam_i + 0.032 \cdot fish_i + e_i$

...and for completeness, again a short check of the modelling assumptions:



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

Categorical covariates

An even more general
(Stahel 3.2e)