```
> library(reporttools)
> library(biostatUZH)
```

Kurs Bio144:
Datenanalyse in der Biologie

Stefanie Muff & Owen L. Petchey

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

Overview (todo: check)

- Multiple predictors x_1, x_2, . . . , x_p
- R^2 in multiple linear regression
- t-tests, F-tests and p-values
- Binary and factor covariates
- Interactions between covariates
- Multiple vs. many single regressions

Course material covered today

- Chapters 3.1 - 3.3 of *Lineare Regression*, p.25-39 (Stahel script)
- Other?

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

$$y_i = \alpha + \beta x_i + E_i \ , \qquad E_i \sim \mathsf{N}(0, \sigma_E^2) \text{ independent.}$$

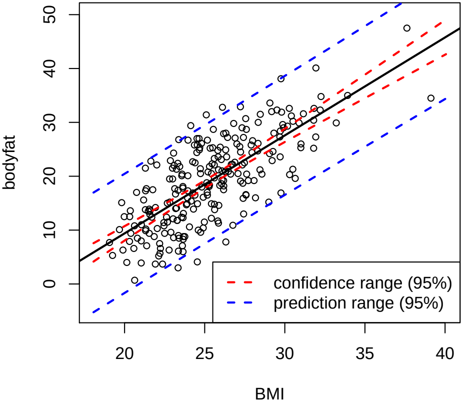
- Estimate the parameters α, β and σ_E^2 by least squares.
- The estimated parameters $\hat{\alpha}$, $\hat{\beta}$ contain uncertainty and are normally distributed around the true values.
- Knowing this helps to deduce statistical tests, such as: Is $\beta = 0$?
- Testing is simple using R:

```
> 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:

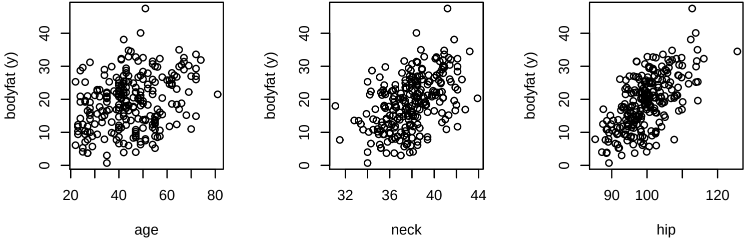


Bodyfat example

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.



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

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

- 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 with R

Let us fit bodyfat (from last week) to the predictors **bmi** and **age** simultaneously. The R code to fit the model is

```
> r.bodyfatM <- lm(bodyfat ~ bmi + age ,d.bodyfat)#+ neck + hip + abdomen
> 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
```

Which questions should be asked? (In principle, we should have asked the questions *before*!)

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

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

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

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:

```
> 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 already correlated and therefore explain similar aspects that influence the proportion of bodyfat.

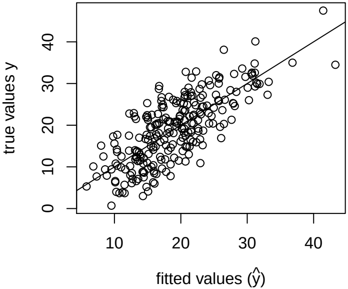
This problem is at the heart of many confusions of regression analysis, and we will talk about such issues later in the course.

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

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

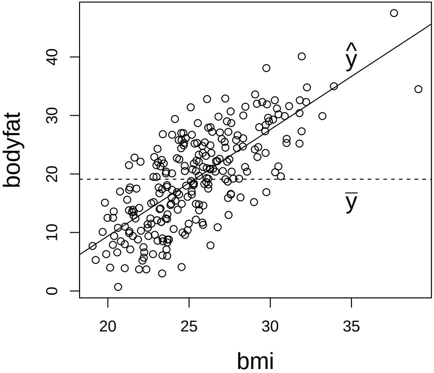


R^2 is also called the *coefficient of determination* or “Bestimmtheitsmass”, because it measures the proportion of the reponse’s variability that is explained by the ensemble of all covariates:

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

Remembering that

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


bodyfat
bmi
0
10
20
30
40
20
25
30
35
$\hat{y}$
$\bar{y}$

Let us look at the R^2s from the two bodyfat models (model 1 with bmi + age, model 2 with bmi + age + neck + hip + abdomen):

```
[1] 0.5802956
```

```
[1] 0.718497
```

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

It thus *seems* that the larger model is "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 will be a topic that comes later in this course...

Interpretation of the coefficients

What does the regression output actually *mean*?