

# Lecture 4: Regression (continued) and multiple regression

## BIO144 Data Analysis in Biology

Stephanie Muff, Owen Petchey & Uriah Daugaard

University of Zurich

01 March, 2024

## Recap of last week

- ▶ Why use linear regression?
- ▶ Fitting the line (least squares).
- ▶ Is the linear model good enough – the five assumptions.
- ▶ What if something goes wrong (transformations and handling outliers)?

# Overview of this week

Regression continued...

- ▶ How well does the model describe the data: Correlation and  $R^2$
- ▶ Are the parameter estimates compatible with some specific value (t-test)?
- ▶ What range of parameters values are compatible with the data (confidence intervals)?
- ▶ What regression lines are compatible with the data (confidence band)?
- ▶ What are plausible values of other data (prediction band)?

Multiple regression:

- ▶ Multiple linear regression  $x_1, x_2, \dots, x_m$
- ▶ Checking assumptions
- ▶  $R^2$  in multiple linear regression
- ▶  $t$ -tests,  $F$ -tests and  $p$ -values

# 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*

## How good is the regression model?

This is, per se, a difficult question. . . .

One often considered index is the **coefficient of determination** (**Bestimmtheitsmass**)  $R^2$ . Let us again look at the regression output from the bodyfat example:

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

## [1] 0.5390391 53.9%

Compare this to the squared correlation between the two variables:

`cor(d.bodyfat$bodyfat,d.bodyfat$bmi)^2`

## [1] 0.5390391

→ In simple linear regression,  $R^2$  is the squared correlation between the independent and the dependent variable.

- ▶  $R^2$  indicates the proportion of variability of the response variable  $y$  that is **explained by the ensemble of all covariates**.
- ▶ Its value lies between 0 and 1.

The **larger**  $R^2$

- ⇒ the **more** variability of  $y$  is captured (“explained”) by the covariate
- ⇒ the **"better"** is the model.

(However, it's a bit more complicated, as we will see in the multiple regression later in the lecture today)

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

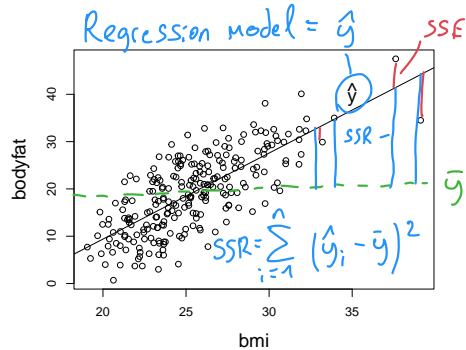
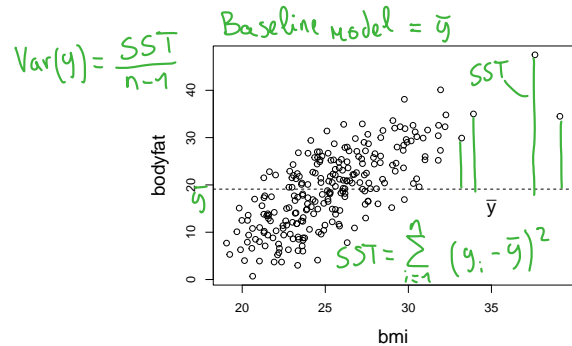
$$R^2 = SSQ^{(R)} / SSQ^{(Y)} = 1 - SSQ^{(E)} / SSQ^{(Y)}$$

With

$$\begin{aligned}
 R^2 &= \frac{\text{explained variability}}{\text{total variability}} = \text{proportion of explained variability} \\
 \underbrace{\text{total variability}}_{\substack{\text{in } y: \\ \parallel \\ \sum_{i=1}^n (y_i - \bar{y})^2 \\ \parallel \\ SSQ^{(Y)}}} &= \underbrace{\text{explained variability}}_{\substack{\text{by the regression} \\ \parallel \\ \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \\ \parallel \\ SSQ^{(R)}}} + \underbrace{\text{residual variability}}_{\substack{\text{in the errors} \\ \parallel \\ \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ \parallel \\ SSQ^{(E)}}} \\
 SST_{\text{total}} &= SSR_{\text{regression}} + SSE_{\text{errors}}
 \end{aligned}$$

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

$$SST = SSR + SSE$$





# Are the parameter estimates compatible with some specific value (t-test)?

*Important:*  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are themselves **random variables** and as such contain **uncertainty**!

Let us look again at the regression output, this time only for the coefficients. The second column shows the standard error of the estimate:

```
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

→ The logical next question is: what is the distribution of the estimates?

## Distribution of the estimators for $\hat{\beta}_0$ and $\hat{\beta}_1$

To obtain an idea, we generate data points according to model

$$y_i = 4 - 2x_i + \epsilon_i, \quad \epsilon_i \sim N(0, 0.5^2).$$

In each round, we estimate the parameters and store them:

```
niter <- 1000
pars <- matrix(NA, nrow=niter, ncol=2)
for (ii in 1:niter){
  x <- rnorm(100)
  y <- 4 - 2*x + rnorm(100, 0, sd=0.5)
  pars[ii,] <- lm(y~x)$coef
}
```

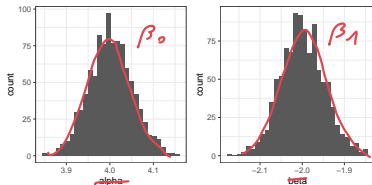
generate x values

compute y values

estimate and save parameters

Do this 1000 times  $\Rightarrow$  Distribution of beta parameters

Doing it 1000 times, we obtain the following distributions for  $\hat{\beta}_0$  and  $\hat{\beta}_1$ :



This looks suspiciously normal!

In fact, from theory it is known that

$$\hat{\beta}_1 \sim N(\beta_1, \overset{\text{estimate}}{\sigma^{(\beta_1)^2}}) \quad \text{and} \quad \hat{\beta}_0 \sim N(\beta_0, \overset{\text{estimate}}{\sigma^{(\beta_0)^2}})$$

For formulas of the variances  $\sigma^{(\beta_1)^2}$  and  $\sigma^{(\beta_0)^2}$ , please consult Stahel 2.2.h.

### To remember:

- ▶  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are **unbiased estimators** of  $\beta_0$  and  $\beta_1$ .
- ▶ the parameters estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are **normally distributed**.
- ▶ the formulas for the variances depend on the residual variance  $\sigma^2$ , the sample size  $n$  and the variability of  $X$  ( $SSQ^{(X)(*)}$ ).

(\*)

$$SSQ^{(X)} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$se(\beta_1) = \text{estimate of } \sigma(\beta_1) = \hat{\sigma}(\beta_1)$$

With all this, we can calculate a standardised measure of the uncertainty in the parameter estimates, known as the *standard error*, or *SE*:

Standard error of parameter estimate:  $se(\beta_1) = \sqrt{\frac{\hat{\sigma}^2}{SSQ(X)}}$

Estimated residual variance:  $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n R_i^2$

Residuals (also sometimes  $e_i$ ):  $R_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$

Sum of squares of  $X$ :  $SSQ(X) = \sum_{i=1}^n (x_i - \bar{x})^2$

# Are the parameter estimates compatible with some specific value (t-test)?

Let's first go back to the output from the bodyfat example:

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

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

Besides the estimate and the standard error (which we discussed before), there is a **t value** and a probability **Pr(>|t|)** that we need to understand.

How do these things help us?

# Testing the “effect” of a covariate

Remember: in a statistical test you first need to specify the *null hypothesis*. Here, typically, the null hypothesis is

$$H_0 : \beta_1 = 0 .$$

In words:  $H_0 =$  "no association"

Here, the *alternative hypothesis* is given by

$$H_A : \beta_1 \neq 0$$

Remember: To carry out a statistical test, we need a *test statistic*.

What is a test statistic?

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2(\beta_1)^2)$$

→ It is some type of **summary statistic** that follows a known distribution under  $H_0$ .  
For our purpose, we use the so-called ***T*-statistic**

$$T = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{se(\beta_1)} . \quad (1)$$

Again: typically,  $\beta_{1,H_0} = 0$ , so the formula simplifies to

$$T = \frac{\hat{\beta}_1}{se(\beta_1)} .$$

$\sim t\text{-distr}$

Under  $H_0$ ,  $T$  has a *t*-distribution with  $n - 2$  degrees of freedom ( $n$  = number of data points).

(You should try to recall the *t*-distribution. Check Mat183, keyword: *t*-test.)

So let's again go back to the bodyfat regression output:

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

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

$$T = \frac{1.82}{0.11} = 16.8$$

$$e^{-42} = 10^{-42}$$

Task:

→ Please use equation (1) to find out how the first three columns (Estimate, Std. Error and t value) are related! Check by a calculation...

Note: The last column contains the **p-value** of the test of the null hypothesis of  $\beta_1 = 0$ .



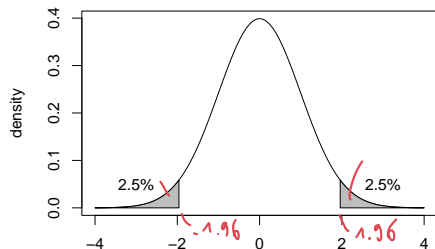
## Recap: Formal definition of the $p$ -value

The **formal definition of  $p$ -value** is the probability to observe a data summary (e.g., an average) that is at least as extreme as the one observed, given that the Null Hypothesis is correct.

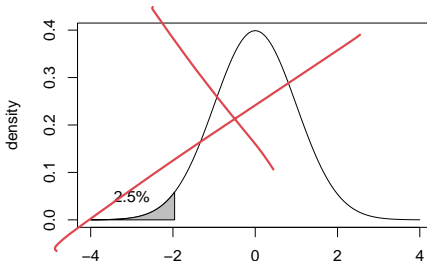
Example (normal distribution): Assume that we calculated that  $t$ -value = -1.96

$$\Rightarrow \Pr(|t| \geq 1.96) = 0.05 \text{ and } \cancel{\Pr(t \leq -1.96) = 0.025}.$$

Two-sided  $p$ -value (0.05)



One-sided  $p$ -value (0.025)



The regression output from R indicates that the  $p$ -value for BMI is very small ( $p < 0.0001$ ).

Conclusion: there is **very strong evidence** that the BMI is associated with bodyfat, because  $p$  is extremely small (thus it is very unlikely that such a slope  $\hat{\beta}_1$  would be seen if there was no association of BMI and body fat).

This basically answers question 1: “Are the parameters compatible with some specific value?”

## A cautionary note on the use of $p$ -values

Maybe you have seen that in statistical testing, often the criterion  $p \leq 0.05$  is used to test whether  $H_0$  should be rejected. This is often done in a black-or-white manner.

However, we will put a lot of attention to a more reasonable and cautionary interpretation of  $p$ -values in this course!

# What range of parameters values are compatible with the data (confidence intervals)?

To answer this question, we can determine the confidence intervals of the regression parameters.

## Facts we know about $\hat{\beta}_1$

- ▶  $\hat{\beta}_1$  is estimated with a standard error of  $\sigma^{(\beta_1)}$
- ▶ The distribution of  $\hat{\beta}_1$  is normal, namely  $\hat{\beta}_1 \sim N(\beta_1, \sigma^{(\beta_1)^2})$ .
- ▶ However, since we need to estimate  $\sigma^{(\beta_1)}$  from the data, we have a  $t$ -distribution.

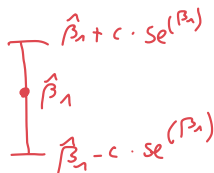
Doing some calculations (similar to those in chapter 8.2.2 of Mat183 script) leads us to the 95% confidence interval

confidence level

lower bound

upper bound

$$[\hat{\beta}_1 - c \cdot \hat{\sigma}^{(\beta_1)}; \hat{\beta}_1 + c \cdot \hat{\sigma}^{(\beta_1)}],$$



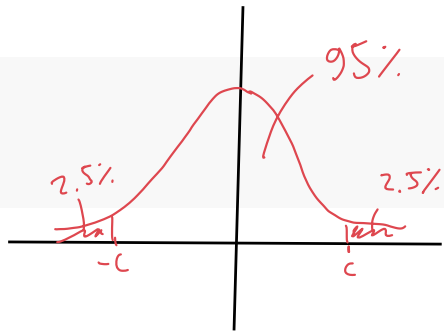
where  $c$  is the 97.5% quantile of the  $t$ -distribution with  $n - 2$  degrees of freedom.

Doing this for the bodfat example “by hand” is not hard. We have 241 degrees of freedom:

```
coefs <- summary(r.bodyfat)$coef
beta <- coefs[2,1]
sdbeta <- coefs[2,2]
beta + c(-1,1) * qt(0.975,241) * sdbeta
```

$\hat{\beta}_1 \quad \pm \quad c \approx 2 \quad \cdot \quad se(\beta_1)$

```
## [1] 1.605362 2.032195
```



Even easier: directly ask R to give you the CIs.

```
confint(r.bodyfat,level=c(0.95))
```

```
##                2.5 %      97.5 %
## (Intercept) -32.438703 -21.530032
## bmi         1.605362   2.032195
```

In summary,

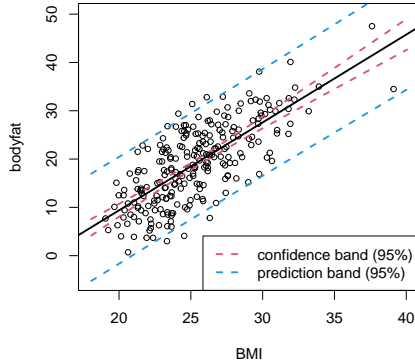
	Coefficient	95%-confidence interval	<i>p</i> -value
Intercept	-26.98	from -32.44 to -21.53	< 0.0001
bmi	1.82	from 1.61 to 2.03	< 0.0001

Interpretation: for an increase in the bmi by one index point, roughly 1.82% percentage points more bodyfat are expected, and all true values for  $\beta_1$  between 1.61 and 2.03 are compatible with the observed data.

# Confidence and Prediction Bands

- ▶ Remember: When another sample from the same population was taken, the regression line would look slightly different.
- ▶ There are two questions to be asked:
  1. Which other regression lines are compatible with the observed data?  
⇒ This leads to the **confidence band**.
  2. Where do future observations with a given  $x$  coordinate lie?  
⇒ This leads to the **prediction band**.

# Bodyfat example



Note: The prediction band is much broader than the confidence band.



# Calculation of the confidence band

Given a fixed value of  $x$ , say  $x_0$ . The question is:

Where does  $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$  lie with a certain confidence (i.e., 95%)?

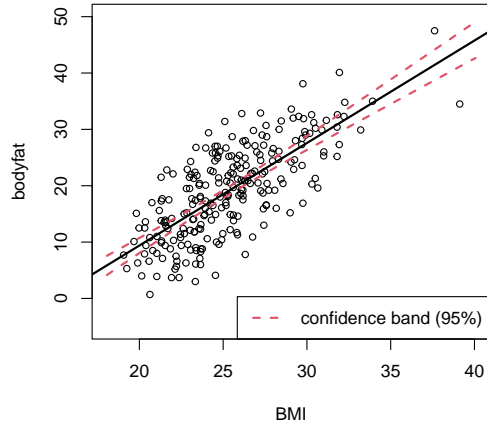
This question is not trivial, because both  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are estimates from the data and contain uncertainty.

The details of the calculation are given in Stahel 2.4b.

Plotting the confidence interval around all  $\hat{y}_0$  values one obtains the **confidence band** or **confidence band for the expected values** of  $y$ .

Note: For the confidence band, only the uncertainty in the estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$  matters.

### Confidence band



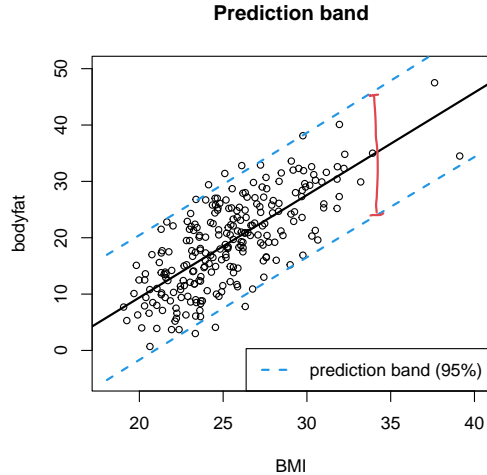
# Calculations of the prediction band

Given a fixed value of  $x$ , say  $x_0$ . The question is:

Where does a **future observation** lie with a certain confidence (i.e., 95%)?

To answer this question, we have to **consider not only the uncertainty in the predicted value**  $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ , but also the **error in the equation**  $\epsilon_i \sim N(0, \sigma^2)$ .

This is the reason why the **prediction band is always wider than the confidence band**.



## That is regression done (at least for our current purposes)

- ▶ Why use (linear) regression?
- ▶ Fitting the line (= parameter estimation)
- ▶ Is linear regression good enough model to use?
- ▶ What to do when things go wrong?
- ▶ Transformation of variables/the response.
- ▶ Handling of outliers.
- ▶ Goodness of the model: Correlation and  $R^2$
- ▶ Tests and confidence intervals
- ▶ Confidence and prediction bands

(Homework and Practical class: Presentation of findings)

# Multiple linear regression

Multiple continuous explanatory variables.

- ▶ Question 1: Are the explanatory variables (i.e. more than one) associated with the response?
- ▶ Question 2: Which variables are associated with the response?
- ▶ Question 3: What proportion of variability is explained?

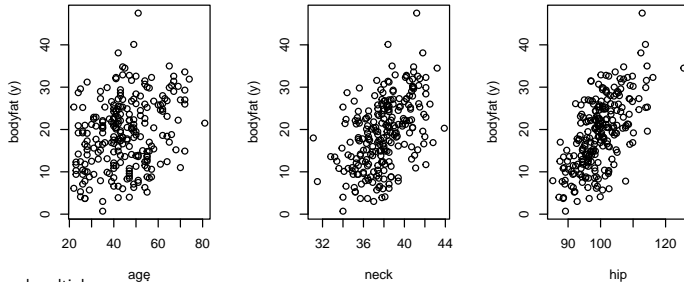
## Bodyfat example

We have so far modeled bodyfat in dependence of bmi, that is:

$$(bodyfat)_i = \beta_0 + \beta_1 \cdot bmi_i + \epsilon_i.$$

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

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



**Multiple linear regression** is when we have more than one explanatory variable. We can then ask three questions:

1. Is the **ensemble** of all explanatory variables associated with the response?
2. If yes, which explanatory variables are associated with the response?
3. What proportion of response variability ( $SSQ^{(Y)}$ ) is explained by the model?



# Multiple linear regression model

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

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

$$y_i = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1 x_i^{(1)}}_{\text{slope of } x_i^{(1)}} + \underbrace{\beta_2 x_i^{(2)}}_{\text{slope of } x_i^{(2)}} + \dots + \underbrace{\beta_m x_i^{(m)}}_{\text{slope of } x_i^{(m)}} + \underbrace{\epsilon_i}_{\text{random error}}$$

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

*m explanatory variables*

- The parameters of this model are  $\beta = (\beta_0, \beta_1, \dots, \beta_m)$  and  $\sigma^2$ .
- m + 1*

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

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

with

$$e_i = \overset{\text{observed}}{y_i} - \underbrace{(\beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + \dots + \beta_m x_i^{(m)})}_{\hat{y}_i: \text{expected}}$$

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; we look at this in Lecture 7.

## 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

$$(bodyfat)_i = \beta_0 + \beta_1 \cdot bmi_i + \beta_2 \cdot age_i + \epsilon_i ,$$

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

# 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:
```

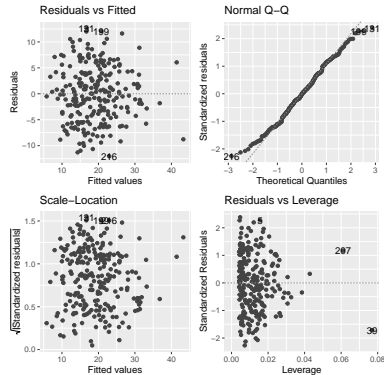
	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	-31.25451	2.78973	-11.203	< 2e-16 ***
## bmi	1.75257	0.10449	16.773	< 2e-16 ***
## age	0.13268	0.02732	4.857	2.15e-06 ***

```
## ---
## 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

$$\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

Before we look at the results, we must check if the modelling assumptions are fulfilled (check our 'chute before we jump):



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

## Question 1: Are the explanatory variables associated with the response?

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

$$F = \frac{\overset{\text{explained variability}}{SSQ^{(R)} / m}}{\underset{\text{unexplained variability}}{SSQ^{(E)} / (n - p)}} \sim F_{m, n-p}$$

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  $\beta$ -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  $SSQ^{(Y)} = \sum_{i=1}^n (y_i - \bar{y})^2$ .

$n$  is the number of data points 243

$m$  is the number of explanatory variables in the regression model  $m = 2$

$p$  is the number of beta parameters estimated (e.g. intercept, plus a slope for each explanatory variable, hence  $p = m + 1$ )  $p = 3$

And the degrees of freedom for error are  $n - p$  240

## 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
```

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 explanatory variable  $x^{(j)}$ .

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

$$T_j = \frac{\hat{\beta}_j}{se(\beta_j)} , \quad (2)$$

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.

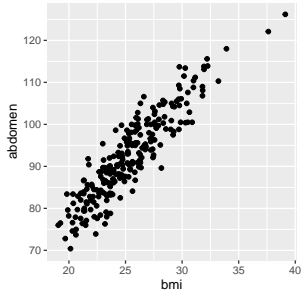
For illustration, let us extend the model a bit more, including also neck, hip and abdomen:

	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

larger  
then before

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).

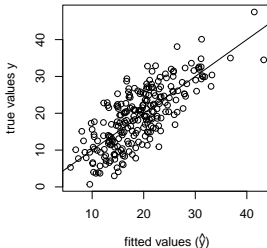
Please see also IC: practical 4 (milk example) for an analysis and more thoughts.

### 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  $(y_1, \dots, y_n)$  and  $(\hat{y}_1, \dots, \hat{y}_n)$ , where the  $\hat{y}$  are the fitted values

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



Let us look at the  $R^2$ s from the three bodyfat models

model r.bodyfat:  $y \sim bmi$

model r.bodyfatM:  $y \sim bmi + age$

model r.bodyfatM2:  $y \sim bmi + age + neck + hip + abdomen$ :

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

```
## [1] 0.5390391
```

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

```
## [1] 0.5802956
```

```
summary(r.bodyfatM2)$r.squared
```

```
## [1] 0.718497
```

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.

## 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 explanatory variables:

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

| if  $n \gg m$   
 $R_a^2 = R^2$

Why  $R_a^2$ ?

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

**Note:**  $R_a$  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*?**

	Coefficient	95%-confidence interval	<i>p</i> -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 1: Parameter estimates of model 2.

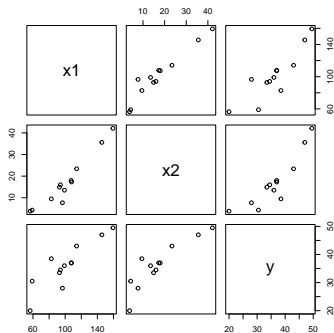
Task in teams: Interpret the coefficients, 95% CIs and *p*-values.

## 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 to consider:

1. Is  $x_1$  an important explanatory variable?
2. Is  $x_2$  an important explanatory variable?
3. Are both explanatory variables needed in the model?
4. Interpretation of the results?

# Recap

- ▶ How well does the model describe the data: Correlation and  $R^2$
- ▶ Are the parameter estimates compatible with some specific value (t-test)?
- ▶ What range of parameters values are compatible with the data (confidence intervals)?
- ▶ What regression lines are compatible with the data (confidence band)?
- ▶ What are plausible values of other data (prediction band)?

## Multiple regression:

- ▶ Multiple linear regression  $x_1, x_2, \dots, x_m$
- ▶ Checking assumptions
- ▶  $R^2$  in multiple linear regression
- ▶  $t$ -tests,  $F$ -tests and  $p$ -values

## Next steps

- ▶ Homework.
- ▶ Practical.
- ▶ Then week 5: Binary/categorical explanatory variables, and interactions