Kurs Bio144: Datenanalyse in der Biologie

Lecture 3: Simple linear regression

Stefanie Muff (Lecture) & Owen L.Petchey (Practical)

University of Zurich

31 December, 2020

Overview

- Introduction of the linear regression model
- Parameter estimation
- Simple model checking
- Goodness of the model: Correlation and R^2
- Tests and confidence intervals
- Confidence and prediction ranges

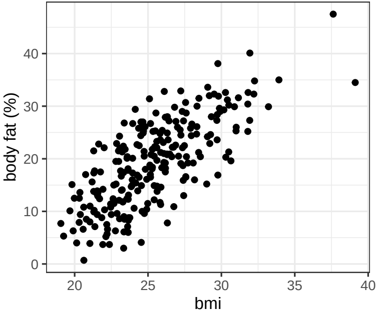
Course material covered today

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

- Chapter 2 of *Lineare Regression*, p.7-20 (Stahel script)

The body fat example

Remember: Aim is to find prognostic factors for body fat, without actually measuring it.
Even simpler question: How good is BMI as a predictor for body fat?



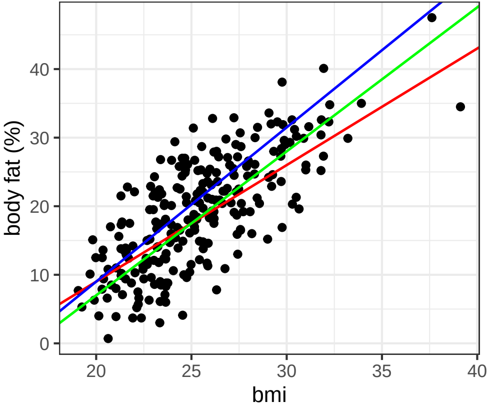
Linear relationship

- The most simple relationship between an *explanatory variable* (X) and a *target/outcome variable* (Y) is a linear relationship. All points (x_i, y_i), $i = 1, \ldots, n$, on a straight line follow the equation

$$y_i = \alpha + \beta x_i \ .$$

- Here, α is the axis intercept and β the slope of the line. β is also denoted as the regression coefficient of X.
- If $\alpha = 0$ the line goes through the origin $(x, y) = (0, 0)$.
- Interpretation of linear dependency: proportional increase in y with increase (decrease) in x.

But which is the “true” or “best” line?



Task: Estimate the regression parameters α and β (by “eye”) and write them down.

It is obvious that

- the linear relationship does not describe the data perfectly
- another realization of the data (other 243 males) would lead to a slightly different picture.

$\Rightarrow$ We need a **model** that describes the relationship between BMI and bodyfat.

The simple linear regression model

In the linear regression model the dependent variable Y is related to the independent variable x as

$$Y = \alpha + \beta x + \epsilon \ , \qquad \epsilon \sim N(0, \sigma^2)$$

In this formulation Y is a random variable $Y \sim N(\alpha + \beta x, \sigma^2)$ where

$$Y \quad = \quad \underbrace{\text{expected value}}_{E(Y)=\alpha+\beta x} \quad + \quad \underbrace{\text{random error}}_{\epsilon} \ .$$

Note:

- The model for Y given x has three parameters: α, β and σ^2 .
- x is the independent / explanatory / regressor variable.
- Y is the dependent / outcome / response variable.

Note

- The linear model propagates the most simple relationship between two variables. When using it, please always think if such a relationship is meaningful/reasonable/plausible.
- Always look at the data before you start with model fitting.

Visualization of the regression assumptions

The assumptions about the linear regression model lie in the error term

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

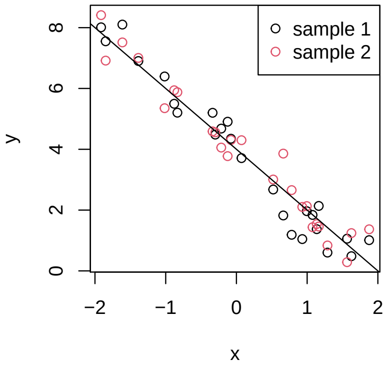
Note: The true regression line goes through $E(Y)$.

Insights from data simulation

(Simulation are *always* a great way to understand statistics!!)

Generate an independent (explanatory) variable **x** and **two** samples of a dependent variable **y** assuming that

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



→ Random variation is always present. This leads us to the next question.

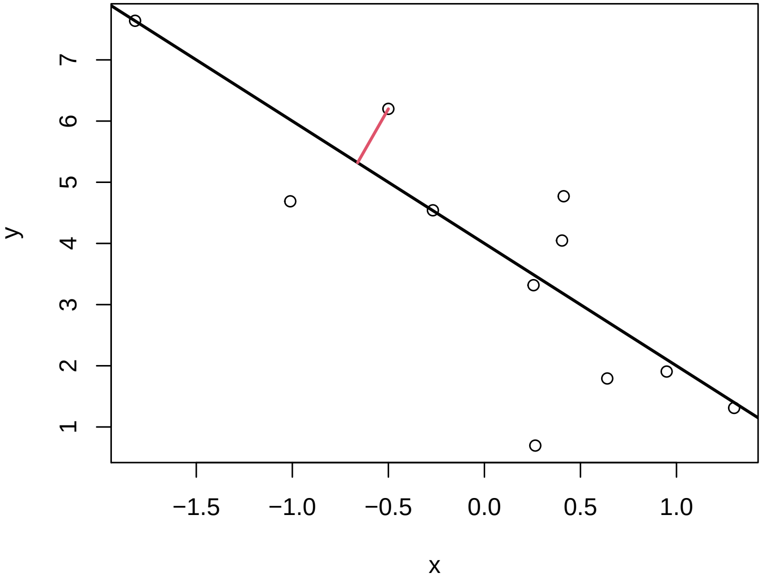
Parameter estimation

In a regression analysis, the task is to estimate the regression coefficients α, β and the residual variance σ^2 for a given set of (x, y) data.

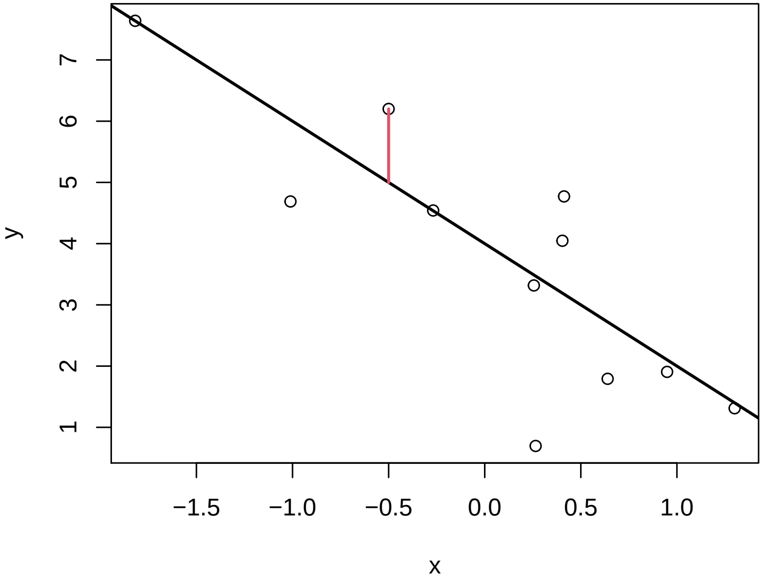
- **Problem:** For more than two points (x_i, y_i), $i = 1, \ldots, n$, there is generally no perfectly fitting line.
- **Aim:** We want to find the parameters (a, b) of the best fitting line $Y = a + bx$.
- **Idea:** Minimize the deviations between the data points (x_i, y_i) and the regression line.

But how?

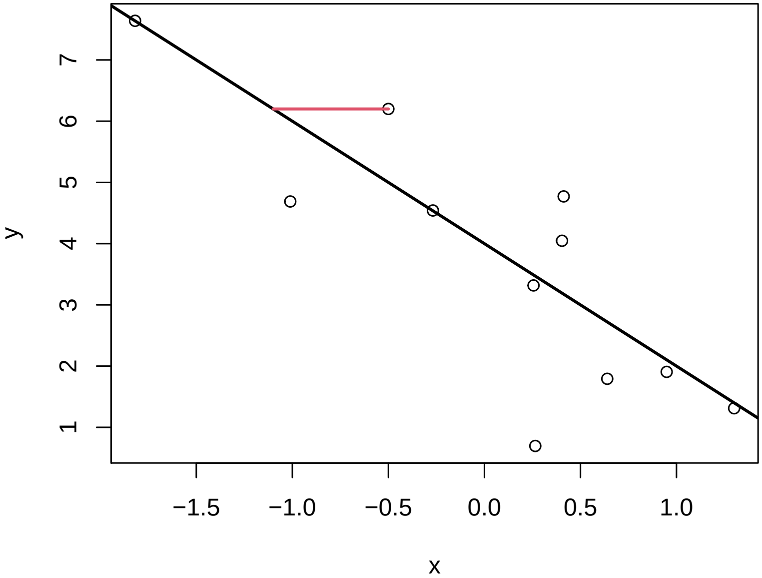
Should we minimize these distances. . .



Or these?



Or maybe even these?



Least squares

For multiple reasons (theoretical aspects and mathematical convenience), the parameters are estimated using the least squares approach. In this, yet something else is minimized:

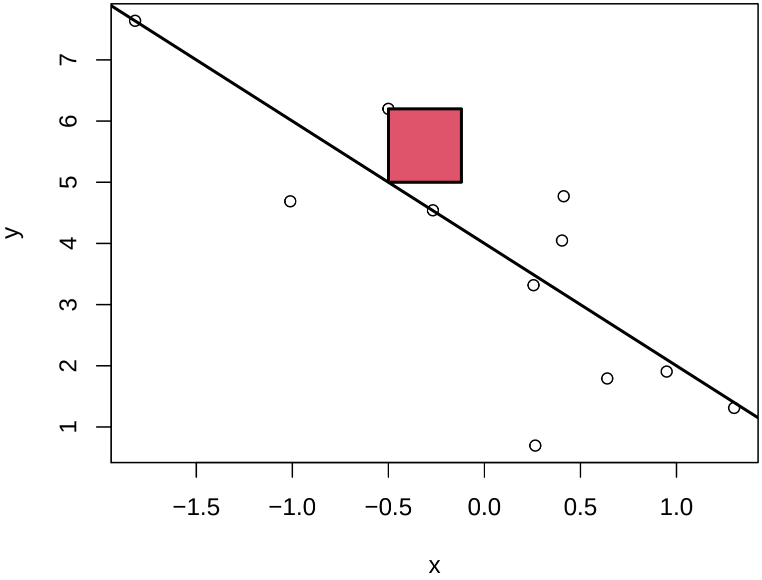
The parameters α and β are estimated such that the sum of squared vertical distances (sum of squared residuals)

$$SSE = \sum_{i=1}^{n} e_i^2 \ , \qquad \text{where} \quad e_i = y_i - \underbrace{(a + bx_i)}_{=\hat{y}_i}$$

is being minimized.

Note: $\hat{y}_i = a + bx_i$ are the predicted values.

So we minimize the sum of these areas!



Least squares estimates

For a given sample $(x_i, y_i), i = 1, .., n$, with mean values $\overline{x}$ and $\overline{y}$, the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$ are computed as

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{cov(x,y)}{var(x)} ,$$

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x} .$$

Moreover,

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 \quad \text{with residuals } e_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$$

is an unbiased estimate of the residual variance σ^2.

(The derivation of the parameters can be looked up in the Stahel script 2.A b. Idea: Minimization through derivating equations and setting them =0.)

Do-it-yourself “by hand”

Go to the Shiny gallery and try to “estimate’ ’ the correct parameters.

You can do this here:

https://gallery.shinyapps.io/simple_regression/

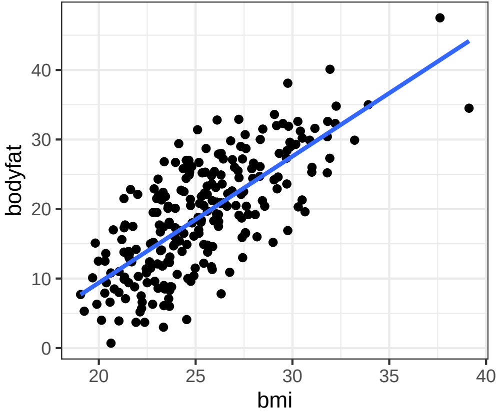
Estimation using R

Let's estimate the regression parameters from the bodyfat example

```r
r.bodyfat <- lm(bodyfat ~ bmi,d.bodyfat)
summary(r.bodyfat)
```

```
##
## Call:
## lm(formula = bodyfat ~ bmi, data = d.bodyfat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5485  -3.5583   0.0785   4.0384  12.7330
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -26.9844     2.7689  -9.746   <2e-16 ***
## bmi           1.8188     0.1083  16.788   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.573 on 241 degrees of freedom
## Multiple R-squared:  0.539,  Adjusted R-squared:  0.5371
## F-statistic: 281.8 on 1 and 241 DF,  p-value: < 2.2e-16
```

The resulting line can be added to the scatterplot:



Interpretation: for an increase in the BMI by one index point, we roughly expect a 1.82% percentage increase in bodyfat.

Uncertainty in the estimates $\hat{\alpha}$ and $\hat{\beta}$

Important: $\hat{\alpha}$ and $\hat{\beta}$ 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
```

$\rightarrow$ The logical next question is: what is the distribution of the estimates?

Distribution of the estimators for $\hat{\alpha}$ and $\hat{\beta}$

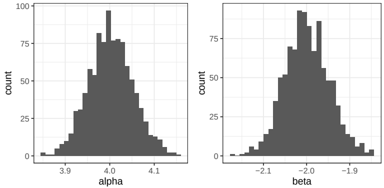
To obtain an idea, we generate data points according to model

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

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
}
```

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



This looks suspiciously normal!

In fact, from theory it is known that

$$\hat{\beta} \sim N(\beta, \sigma^{(\beta)2}) \qquad \text{and} \qquad \hat{\alpha} \sim N(\alpha, \sigma^{(\alpha)2})$$

For formulas of the standard deviations $\sigma^{(\beta)2}$ and $\sigma^{(\alpha)2}$, please consult Stahel 2.2.h.

To remember:

- $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of α and β.
- the parameters estimates $\hat{\alpha}$ and $\hat{\beta}$ are normally distributed.
- the formulas for the variances depend on the residual variance σ^2, the sample size n and the variability of X ($\text{SSQ}^{(X)(\star)}$).

$(\star)$

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

Are the modelling assumptions met?

In practice, it is advisable to check if all our modelling assumptions are met.

$\rightarrow$ Otherwise we might draw invalid conclusions from the results.

Remember: Our assumption is that $\epsilon_i \sim N(0, \sigma^2)$. This implies

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) All ϵ_i are normally distributed.

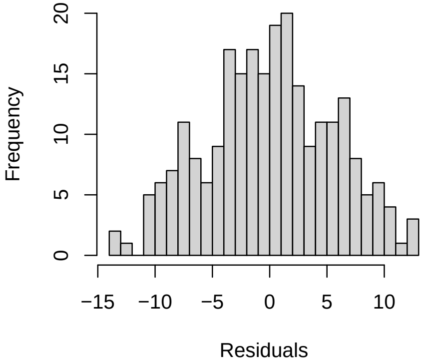
In addition, it is assumed that

d) $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent.

Note: We do not actually observe ϵ_i, but only the residuals e_i. Let us introduce two simple graphical model checking tools for our residuals e_i.

Model checking tool II: Histogram of residuals

Look at the histogram of the residuals:



The normal distribution assumption (c) seems ok as well.

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 form the bodyfat example:

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

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

Compare this to the squared correlation between the two variables:

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

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

$\rightarrow$ 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

$\Rightarrow$ the **more** variability of y is captured ("explained") by the covariate
$\Rightarrow$ the **"better"** is the model.

(However, it's a bit more complicated, see later in the course. . .)

Testing and Confidence Intervals

After the regression parameters and their uncertainties have been estimated, there are typically two fundamental questions:

1. **"Are the parameters compatible with some specific value?"**

   Typically, the question is whether the slope β might be 0 or not, that is: "Is there an effect of the covariate x or not?"

   $\Rightarrow$ This leads to a **statistical test**.

2. **"Which values of the parameters are compatible with the data?"**

   $\Rightarrow$ This leads us to determine **confidence intervals**.

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

```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
```

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 to answer the two questions above?

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

$$H_0: \quad \beta = \beta_0 = 0 \ .$$

In words: $H_0 =$ "no effect"

(Included in H_0 is the assumption that the data follow the simple linear regression model!)

Here, the *alternative hypothesis* is given by

$$H_A: \quad \beta \neq 0$$

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

What is a test statistic?

→ 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**

$$T = \frac{\hat{\beta} - \beta_0}{se^{(\beta)}} \ . \qquad (1)$$

Again: typically, $\beta_0 = 0$, so the formula simplifies to $T = \frac{\hat{\beta}}{se^{(\beta)}}$.

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     Pr(>|t|)
## (Intercept) -26.984368  2.7689004 -9.745518 3.921511e-19
## bmi           1.818778  0.1083411 16.787522 2.063854e-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…

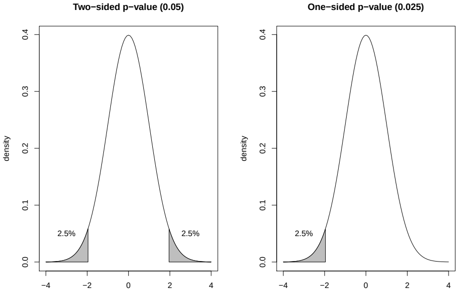
Note: The last column contains the ***p*-value** of the test $\beta = 0$.

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 the observed test-statistic leads to a z-value = -1.96

$\Rightarrow Pr(|z| \geq 1.96) = 0.05$ and $Pr(z \leq -1.96) = 0.025$.



The regression output on slide 33 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 $\hat{\beta}$ would be seen if there was no effect of BMI on body fat).

This basically answers question 1 from slide 29.

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!

Confidence intervals of regression parameters

Question 2 from slide 29:

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

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

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

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

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

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
```

```
## [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	p-value
Intercept	-26.98	from -32.44 to -21.53	< 0.0001
bmi	1.82	from 1.61 to 2.03	< 0.0001

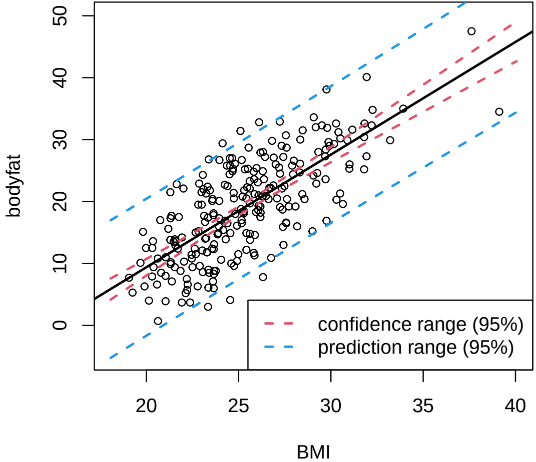
Interpretation: for an increase in the bmi by one index point, roughly 1.82% percentage points more bodyfat are expected, and all true values for β between 1.61 and 2.03 are compatible with the observed data.

Confidence and Prediction Ranges

- 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?
   $\Rightarrow$ This leads to the **confidence range**.
2. Where do future observations with a given x coordinate lie?
   $\Rightarrow$ This leads to the **prediction range**.

Bodyfat example



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

Calculation of the confidence range

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

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

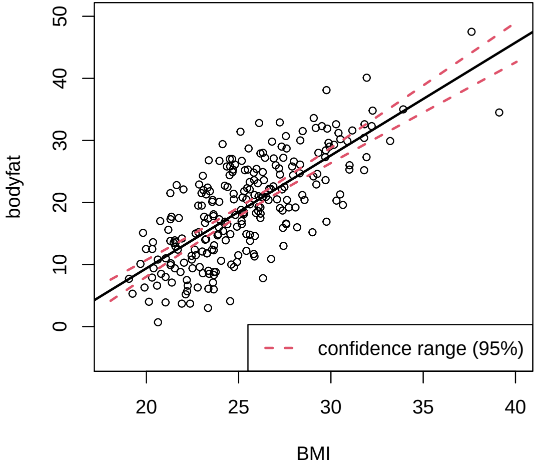
This question is not trivial, because both $\hat{\alpha}$ and $\hat{\beta}$ are estimates from the data and contain uncertainty.

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 range** or **confidence band for the expected values** of y.

Note: For the confidence range, only the uncertainty in the estimates $\hat{\alpha}$ and $\hat{\beta}$ matters.

Confidence range



Calculations of the prediction range

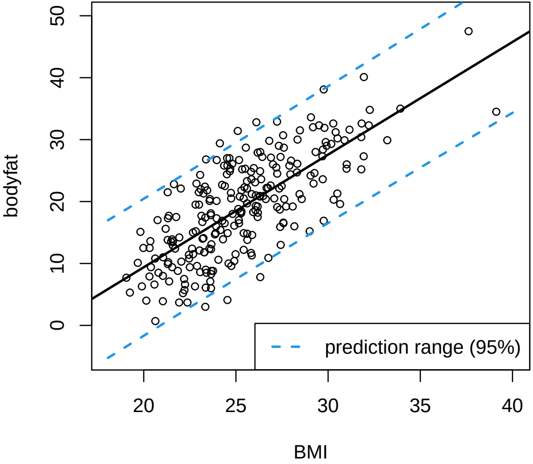
Given a fixed value of x, say x_0. The question is:

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 $\hat{y}_0 = \hat{\alpha} + \hat{\beta}x_0$, but also the error in the equation $\epsilon_i \sim N(0, \sigma^2)$.

This is the reason why the $\{$**prediction range is always wider than the confidence range**$\}$.

Prediction range



Tasks until the next practical (Thu/Fri)

The idea of the course is that as a preparation for the practical part you will do the following:

- Understand what today's lecture was about. You will certainly need to click through it again.
- Go to openedX and do all the "Before class (BC)" tasks.

$\rightarrow$ **The same procedure applies to all course weeks.**