Lecture 3: Regression

BIO144 Data Analysis in Biology

Stephanie Muff & Owen Petchey

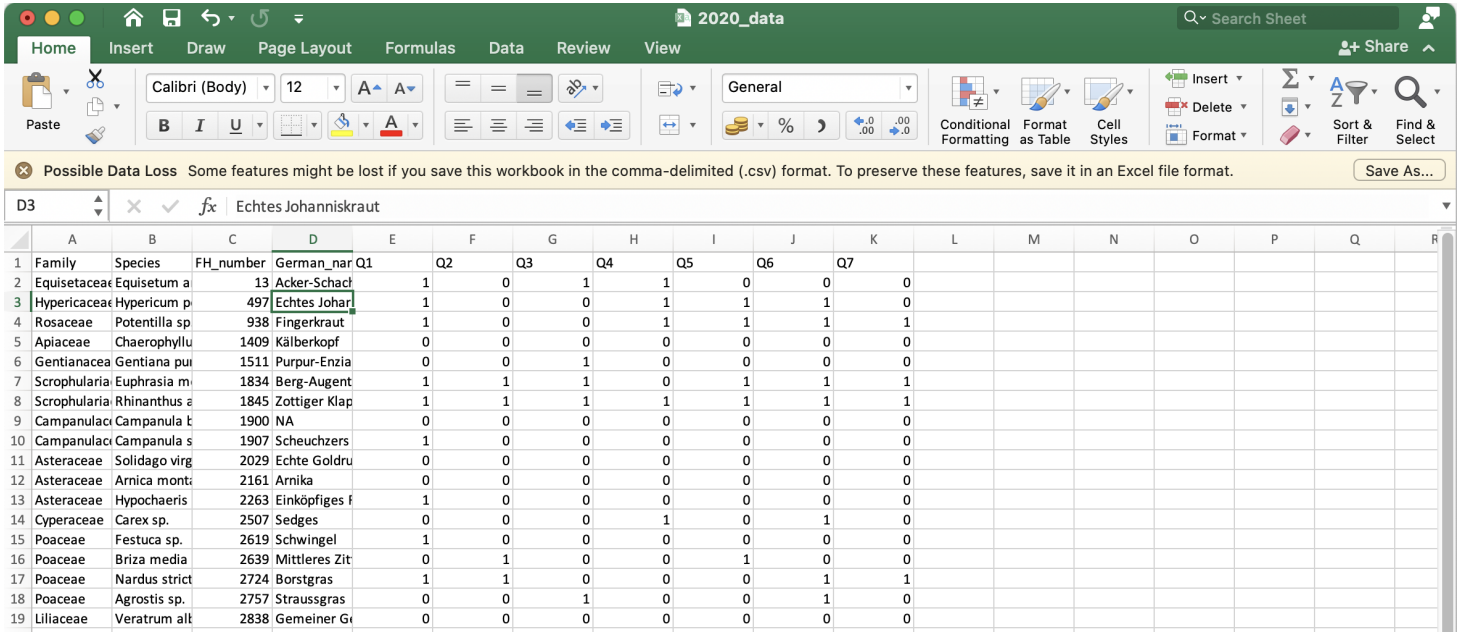
University of Zurich

05 March, 2023

First an alert!

Downloading and opening data (CSV) files

Excel will try to be helpful, but in fact it can break things. Do not accept its help, unless you are sure.



	A	B	C	D	E	F	G	H	I	J	K
1	Family	Species	FH_number	German_nar	Q1	Q2	Q3	Q4	Q5	Q6	Q7
2	Equisetaceae	Equisetum a	13	Acker-Schach	1	0	1	1	0	0	0
3	Hypericaceae	Hypericum p	497	Echtes Johar	1	0	0	1	1	1	0
4	Rosaceae	Potentilla sp	938	Fingerkraut	1	0	0	1	1	1	1
5	Apiaceae	Chaerophyllu	1409	Kälberkopf	0	0	0	0	0	0	0
6	Gentianacea	Gentiana pur	1511	Purpur-Enzia	0	0	1	0	0	0	0
7	Scrophularia	Euphrasia m	1834	Berg-Augent	1	1	1	0	1	1	1
8	Scrophularia	Rhinanthus a	1845	Zottiger Klap	1	1	1	1	1	1	1
9	Campanulac	Campanula b	1900	NA	0	0	0	0	0	0	0
10	Campanulac	Campanula s	1907	Scheuchzers	1	0	0	0	0	0	0
11	Asteraceae	Solidago virg	2029	Echte Goldru	0	0	0	0	0	0	0
12	Asteraceae	Arnica monta	2161	Arnika	0	0	0	0	0	0	0
13	Asteraceae	Hypochaeris	2263	Einköpfiges F	1	0	0	0	0	0	0
14	Cyperaceae	Carex sp.	2507	Sedges	0	0	0	1	0	1	0
15	Poaceae	Festuca sp.	2619	Schwingel	1	0	0	0	0	0	0
16	Poaceae	Briza media	2639	Mittleres Zit	0	1	0	0	1	0	0
17	Poaceae	Nardus strict	2724	Borstgras	1	1	0	0	0	1	1
18	Poaceae	Agrostis sp.	2757	Straussgras	0	0	1	0	0	1	0
19	Liliaceae	Veratrum alt	2838	Gemeiner Ge	0	0	0	0	0	0	0

Overview

- 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

(Next lecture will be interpretation and use of the model. We don't use it until we know it is "working well").

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

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

A head's up...

- Regression is a type of "linear model".
- A t-test is a linear model.

Over the next five lectures you will learn about other types of linear model.

- Multiple regression.
- ANOVA.
- ANCOVA.

All are mathematically very similar, with many similar properties.

The same function in R 'lm' is used to make them all.

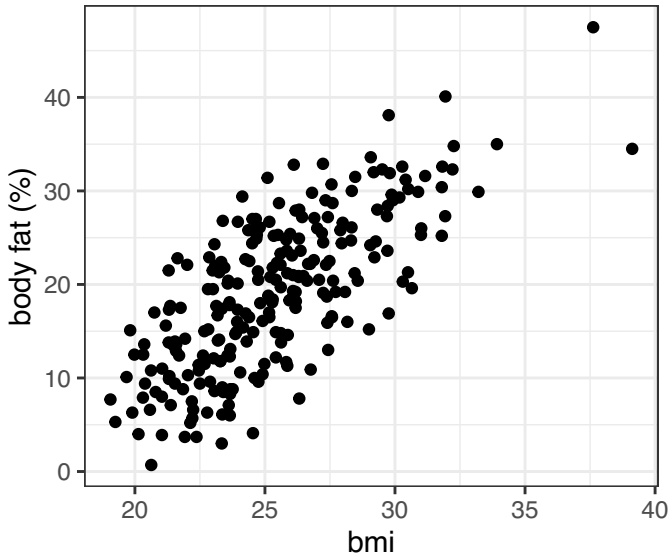
Why use regression?

Imagine we want to know if BMI (easily measured; height and weight) is a good indicator of percentage body fat (less easily measured; specialist instrument required). If it is, then we save some time and effort, and still get the knowledge we need (e.g. percentage body fat).

Put another way: is BMI a good predictor for percentage body fat?

We can address this by looking at the relationship between the two continuous variables, BMI and percentage body fat (next slide).

Does BMI related to percent body fat?



Happy? Not quite? There are many questions we may want to answer:

- What is a good mathematical representation of the relationship?
- Is the relationship different from what we would expect by change?
- How good is the mathematical representation?
- How much uncertainty is there in any predictions?

chance

Linear regression

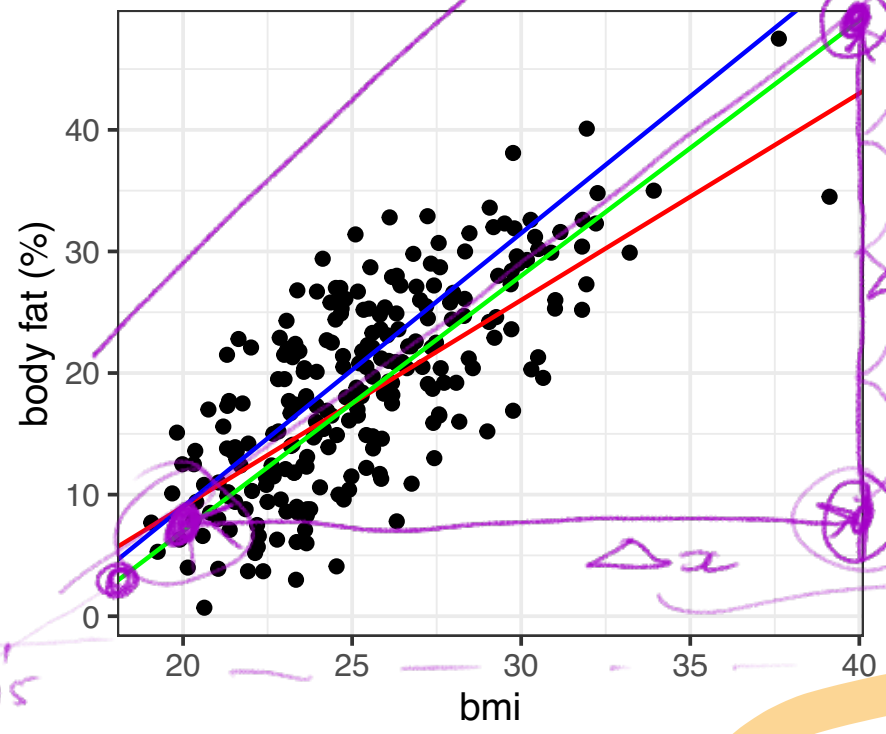
We often use linear regression because:

- A linear relationship is mathematically simple.
- Prior knowledge and observed data suggest it is at least a good starting point.
- Given an *explanatory variable* (X) and a *response variable* (Y) all points (x_i, y_i), $i = 1, \ldots, n$, on a straight line follow the equation

$$y_i = \beta_0 + \beta_1 x_i \ .$$

- β_0 is the intercept - the value of Y when $x_i = 0$
- β_1 the slope of the line, also known as the regression coefficient of X.
- If $\beta_0 = 0$ the line goes through the origin $(x, y) = (0, 0)$.
- Interpretation of linear dependency: proportional increase in y with increase (decrease) in x.

Which is the "best" line?



Task: Estimate the regression parameters β_0 and β_1 (by "eye") and write them down.

Not a perfect fit to the data

It is obvious that

- the linear relationship does not describe the data perfectly
- another realization of the data (a different group of people) 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

$$Y = \beta_0 + \beta_1 x + \epsilon \,, \qquad \epsilon \sim N(0, \sigma^2)$$

In this formulation Y is a random variable $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$ where

$$Y = \underbrace{\text{expected value}}_{E(Y)=\beta_0+\beta_1 x} + \underbrace{\text{random error}}_{\epsilon} \,.$$

Note:

- The model for Y given x has three parameters: β_0, β_1 and σ^2 .
- x is the independent / explanatory / regressor variable.
- Y is the dependent / outcome / response variable.

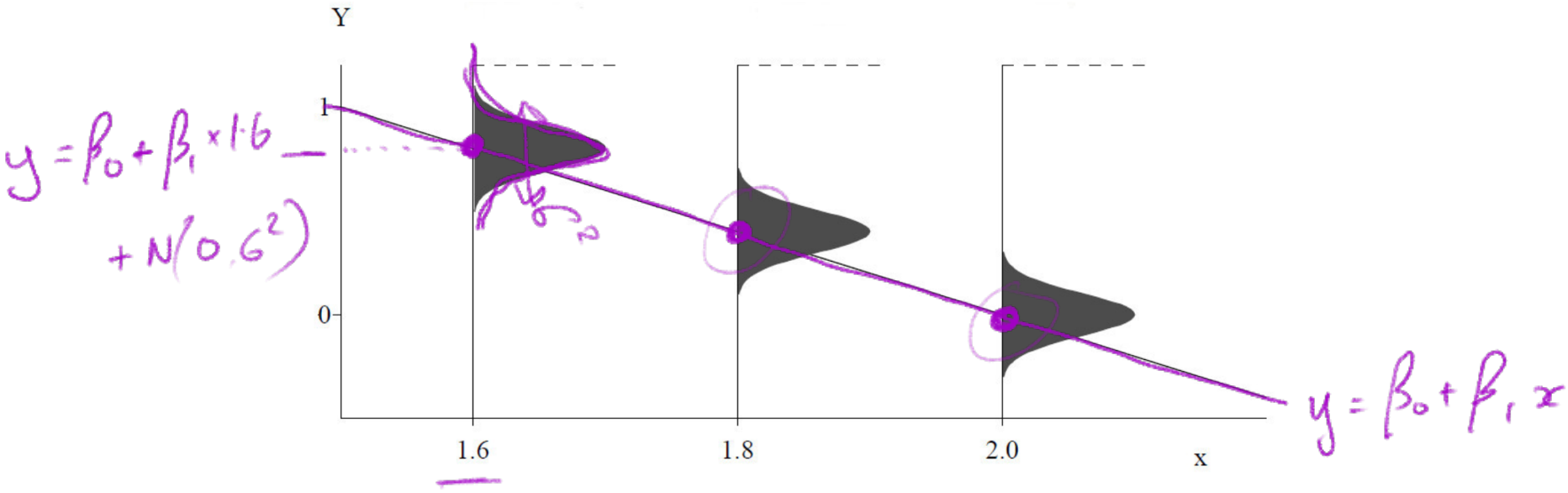
Note

- The linear model propagates the most simple relationship between two variables.
- It is often a good starting point.
- But before 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

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



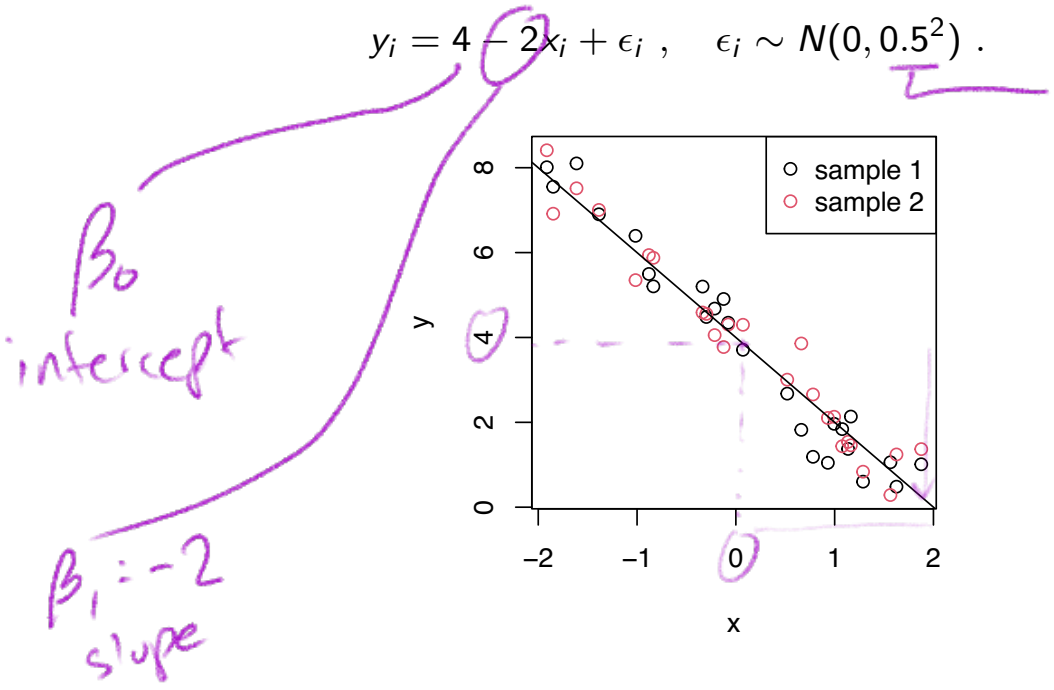
Note: The regression line goes through $E(Y)$.

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.

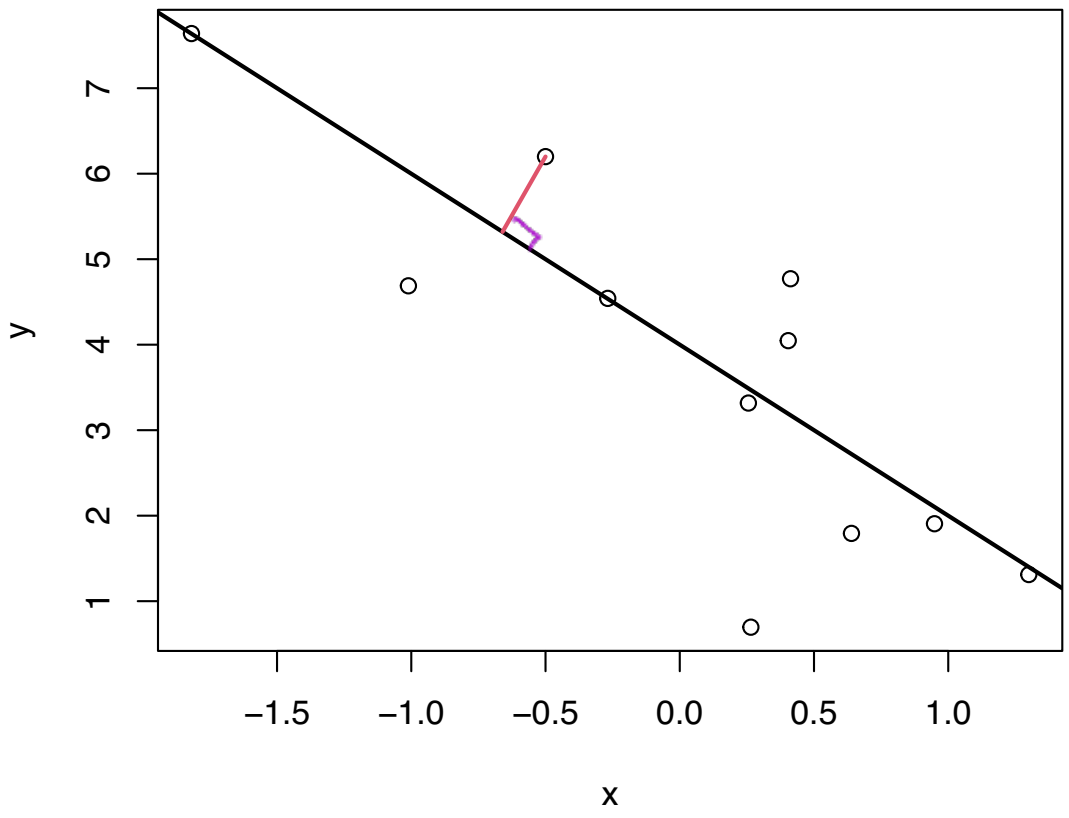
Parameter estimation

In a regression analysis, the task is to estimate the regression coefficients β_0, β_1 and the residual variance σ^2 for a given set of (x, y) data.

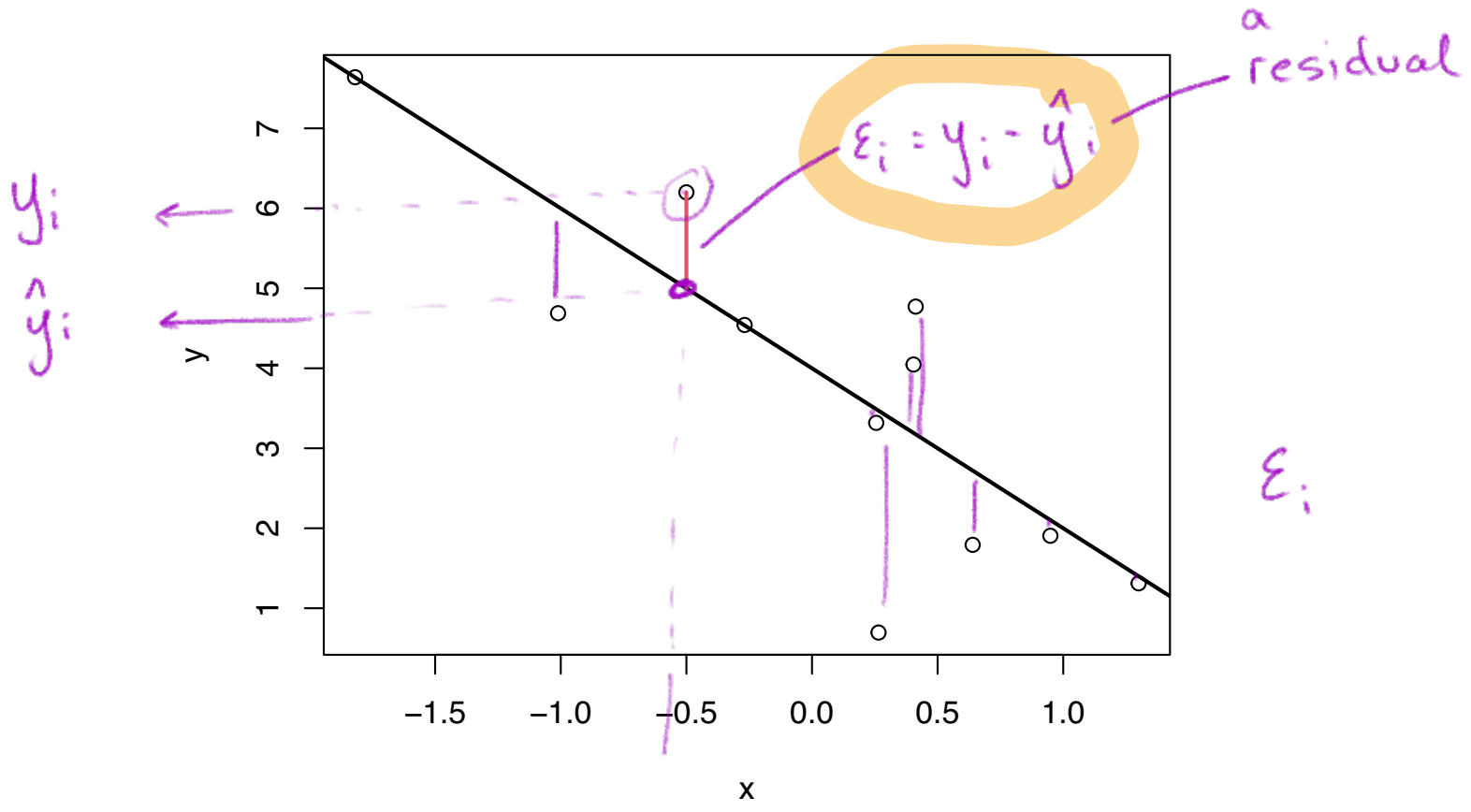
- **Problem:** For more than two points (x_i, y_i), $i = 1, \ldots, n$, there is generally no perfectly fitting line.
- **Aim:** We want to estimate the parameters (β_0, β_1) of the best fitting line $Y = \beta_0 + \beta_1 x$.
- **Idea:** Minimize the deviations between the data points (x_i, y_i) and the regression line.

But how?

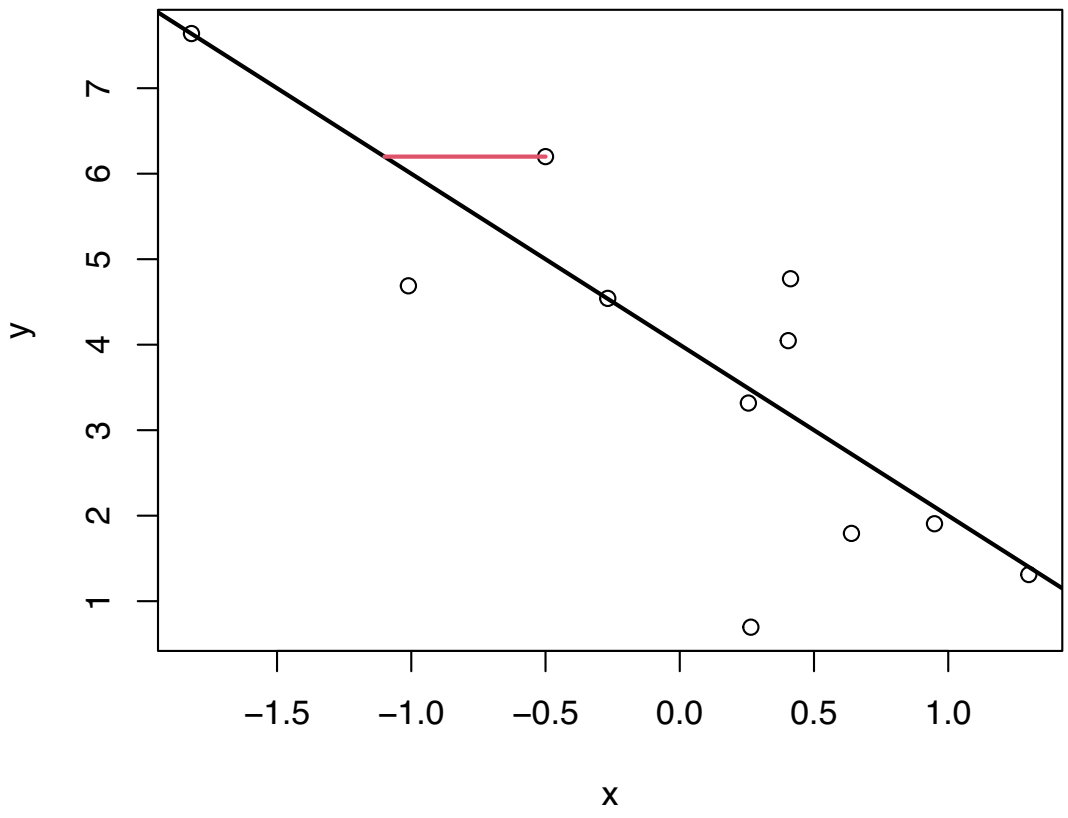
Should we minimize these distances...



Or these?


a residual
$\varepsilon_i = y_i - \hat{y}_i$
y_i
$\hat{y}_i$
ε_i
y
x
1
2
3
4
5
6
7
−1.5
−1.0
−0.5
0.0
0.5
1.0

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:

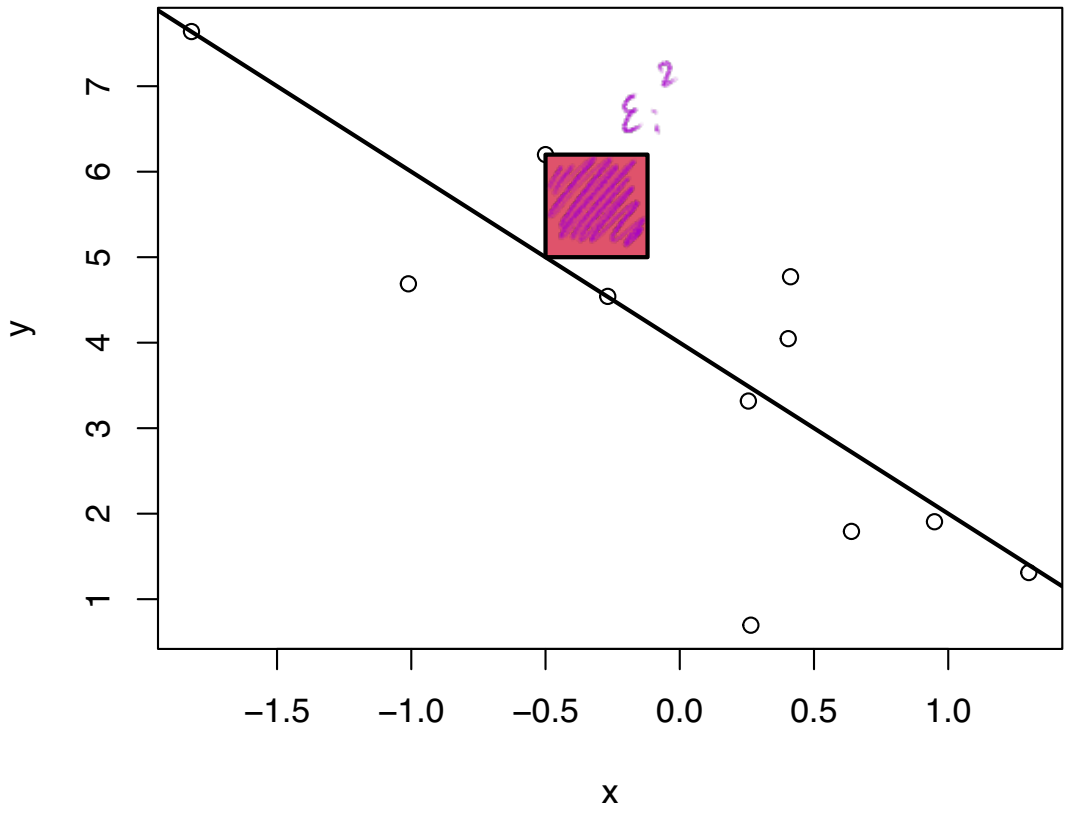
The parameters β_0 and β_1 are estimated such that the sum of squared vertical distances (sum of squared residuals)

$$SSE = \sum_{i=1}^{n} e_i^2, \quad \text{where} \quad e_i = y_i - \underbrace{(\beta_0 + \beta_1 x_i)}_{=\hat{y}_i}$$

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 $\bar{x}$ and $\bar{y}$, the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are computed as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{cov(x, y)}{var(x)},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

Moreover,

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n e_i^2 \quad \text{with residuals } e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

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

(Derivations are in the Stahel script 2.A b. Hint: differentiate, set to zero, solve.)

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

the commands we use in R to do a regression

Let's estimate the regression parameters from the bodyfat example

response variable

explanatory variable

dataset

```
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
```

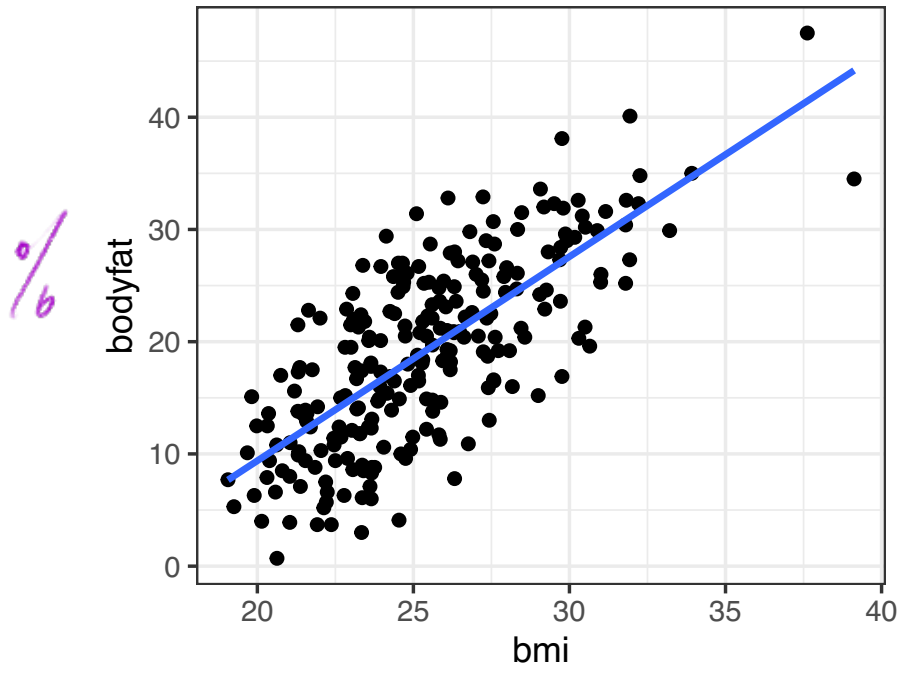
intercept = -26.98

slope = 1.82

rounded to two decimal places

1/2

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.

Is the model good enough to use?

- All models are wrong, but is ours good enough to be useful.
- Are the assumption of the model justified?
- It would be very unwise to use the model before we know if it is good enough to use.
- Don't jump out of an aeroplane until you know your parachute is good enough!

What assumptions to we make?

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

We make this assumption because it is often well enough met, and it gives great mathematical tractibility.

This assumption implies that:

(a) The ϵ_i are normally distributed.

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

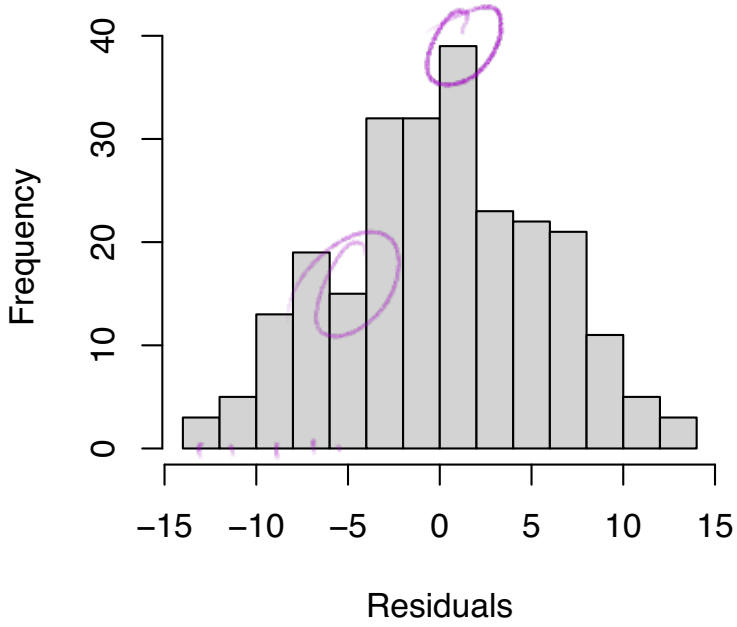
(c) The ϵ_i are independent of each other.

Furthermore:

- (d) we assumed a linear relationship.
- (e) implies there are no outliers (implied by (a) above)

(a) Normally distributed residuals

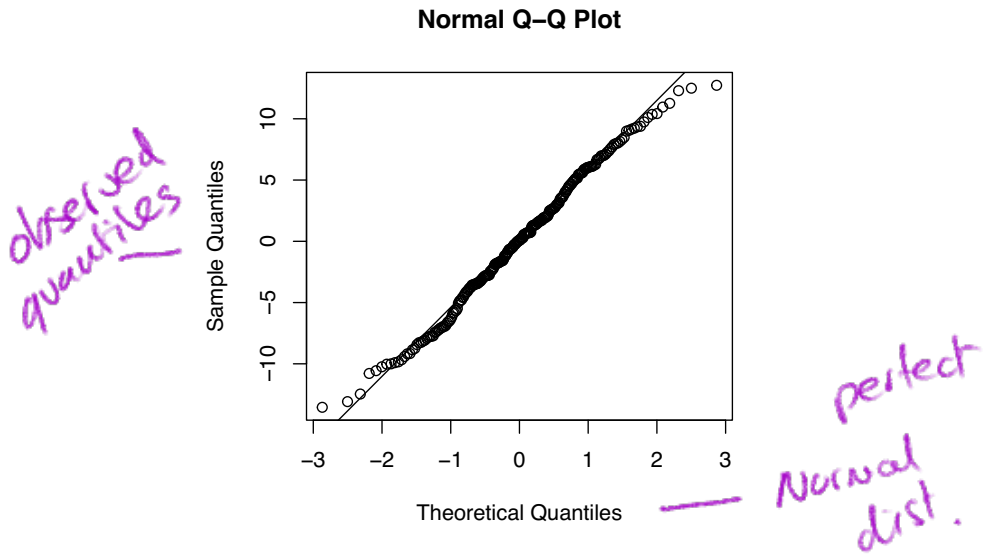
Look at the histogram of the residuals:



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

(a) Normally distributed residuals: 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:

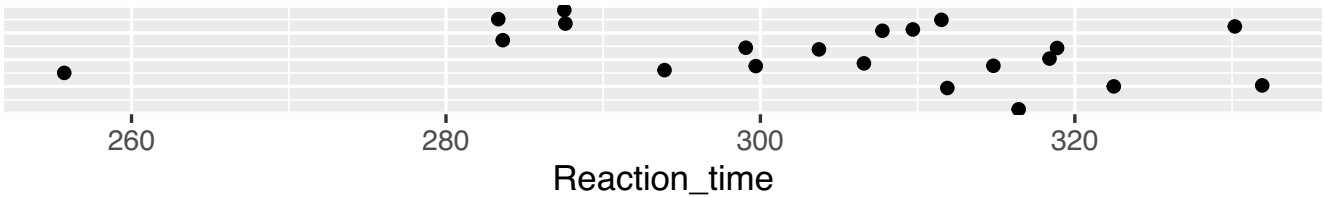


If the points lie approximately on a straight line, the data is fairly normally distributed.

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

What on earth is a quantile???

Imagine we make 21 measures of something, say 21 reaction times:



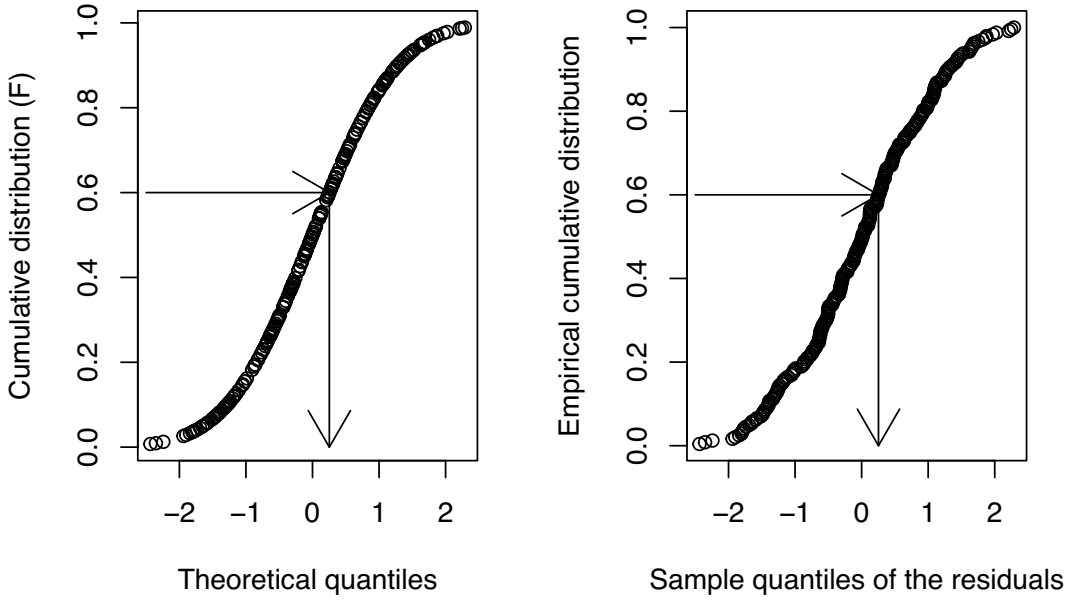
The median of these is 307.8. The median is the 50% or 0.5 quantile, because half the data points are above it, and half below.

```
quantile(dd$Reaction_time)
```

```
##     0%    25%    50%    75%   100%
## 255.7 293.9 307.8 316.4 331.9
```

The QQ-plot continued...

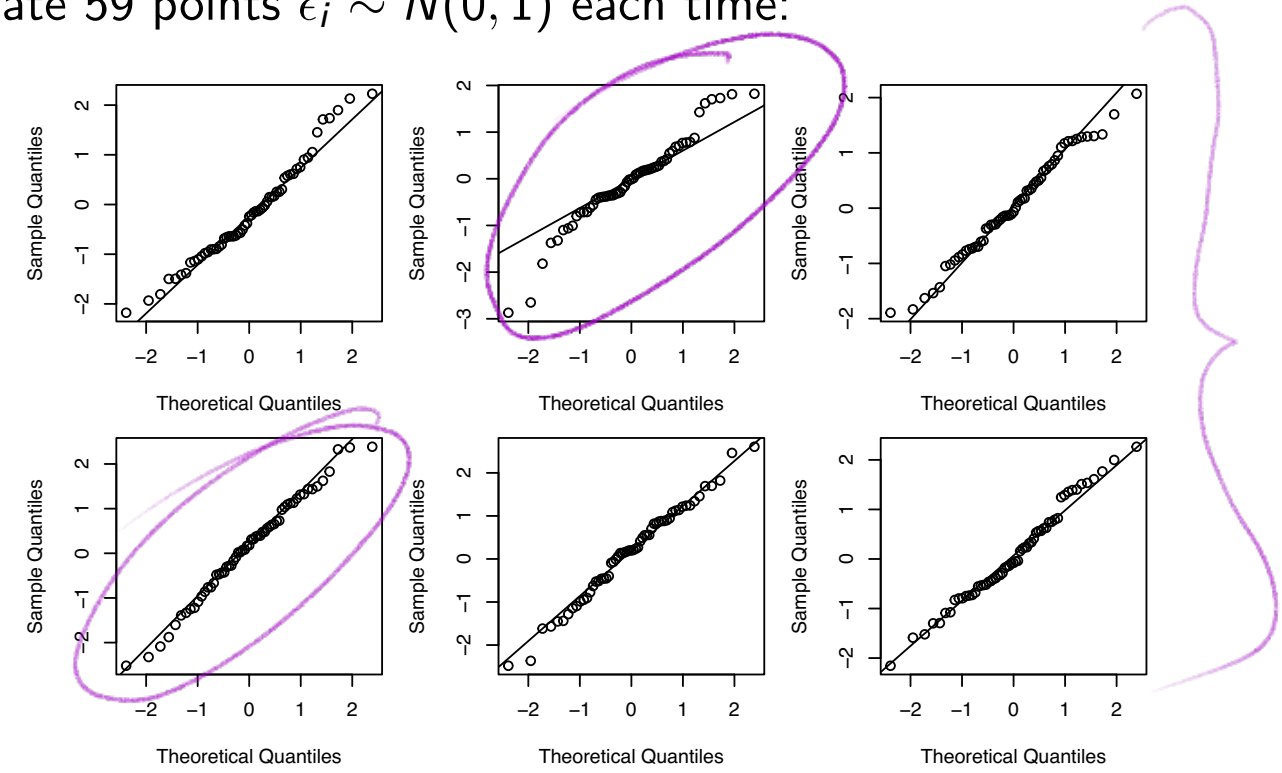
The *theoretical quantiles* come from the normal distribution. The *sample quantiles* come from the distribution of our residuals.



How do I know if a QQ-plot looks "good"?

There is **no quantitative rule** to answer this question, experience is needed. However, you can gain this experience from simulations. To this end, generate the same number of data points of a normally distributed variable and compare to your plot.

Example: Generate 59 points $\epsilon_i \sim N(0, 1)$ each time:

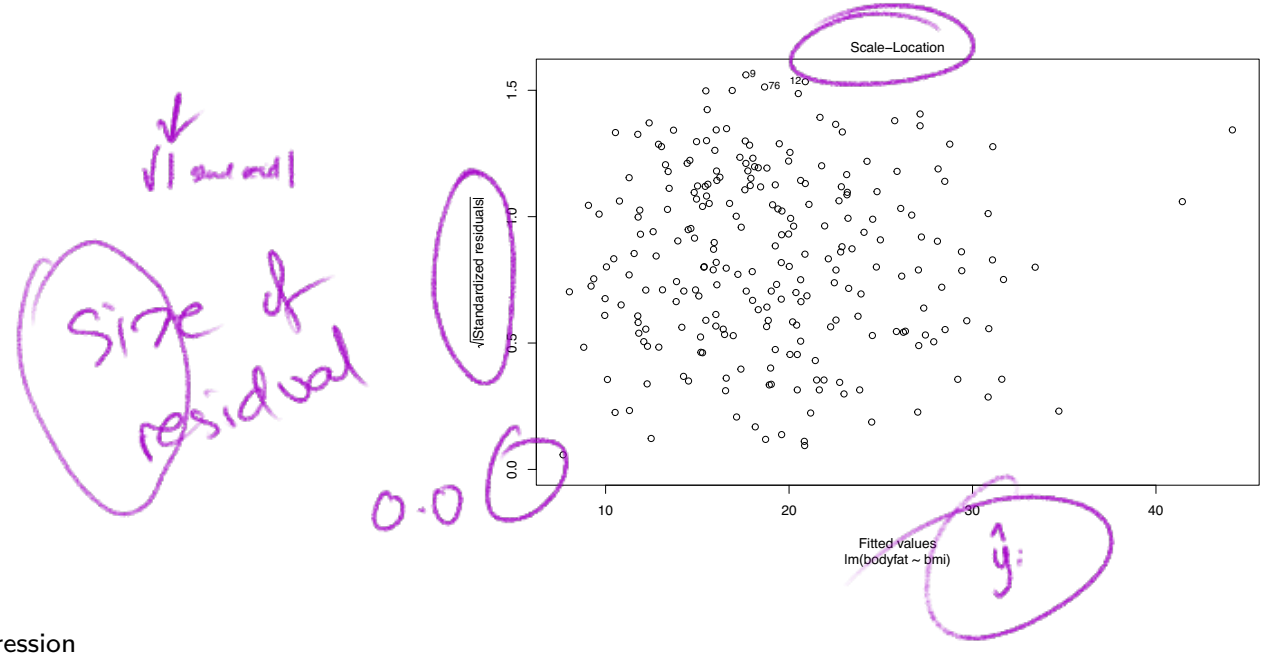


(b) Equal variance (all ϵ_i have the same variance)

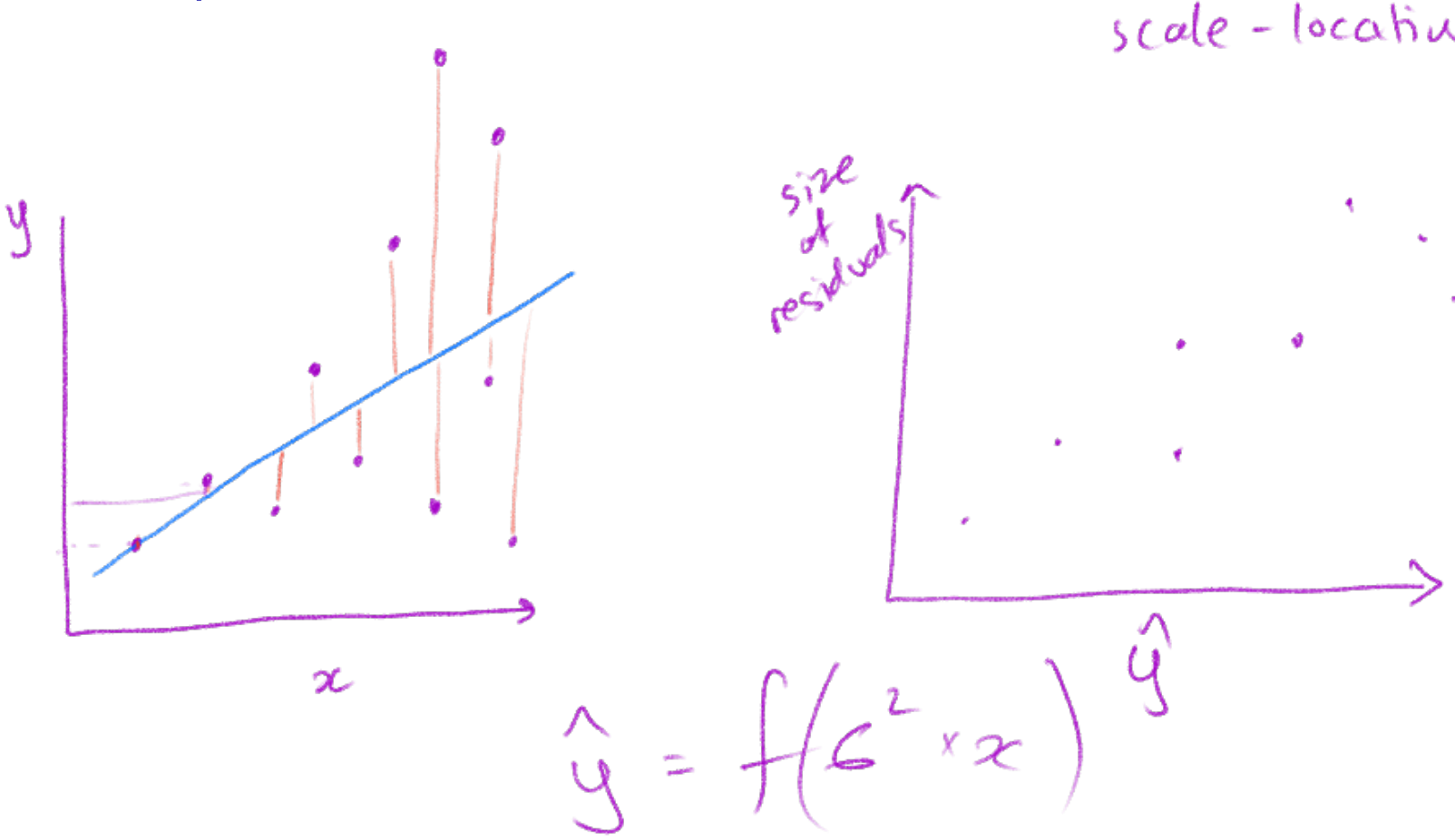
Scale-location plot (Streuungs-Diagramm).

The scale-location plot is particularly suited to check the assumption of equal variances (**homoscedasticity / Homoskedastizität**).

The idea is to plot the square root of the (standardized) residuals $\sqrt{|R_i|}$ against the fitted values $\hat{y}_i$. There should be **no trend**:



Sketch explanation of scale-location



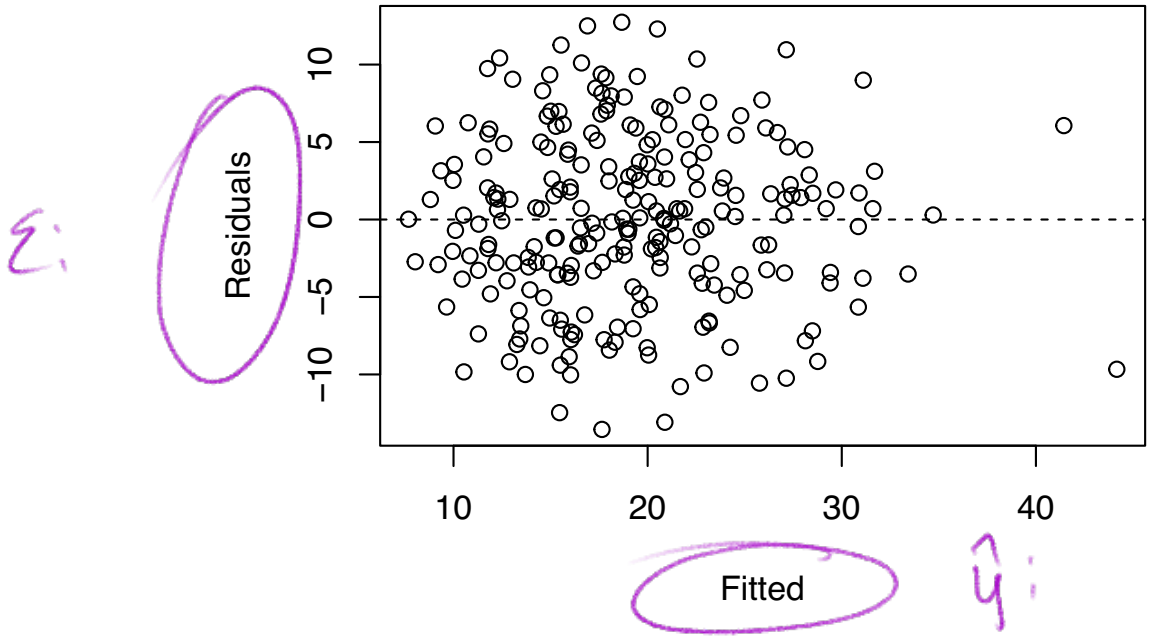
(c) Independence (the ϵ_i are independent of each other)

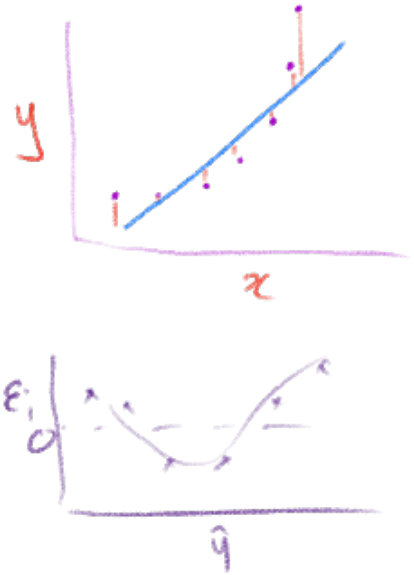
- Think carefully about how the data were collected.
- Think carefully about any structure and / or groupings in the data that are not described in it.

(d) Linearity

With simple regression, we can assess this directly on the graph with fitted line (scroll to an example).

Even then, it can be useful to look at a graph called the **Tukey-Anscombe plot*. It is a graph of the residuals versus the fitted values.





It is sometimes useful to enrich the TA-plot by adding a “running mean” or a “smoothed

(e) No outliers

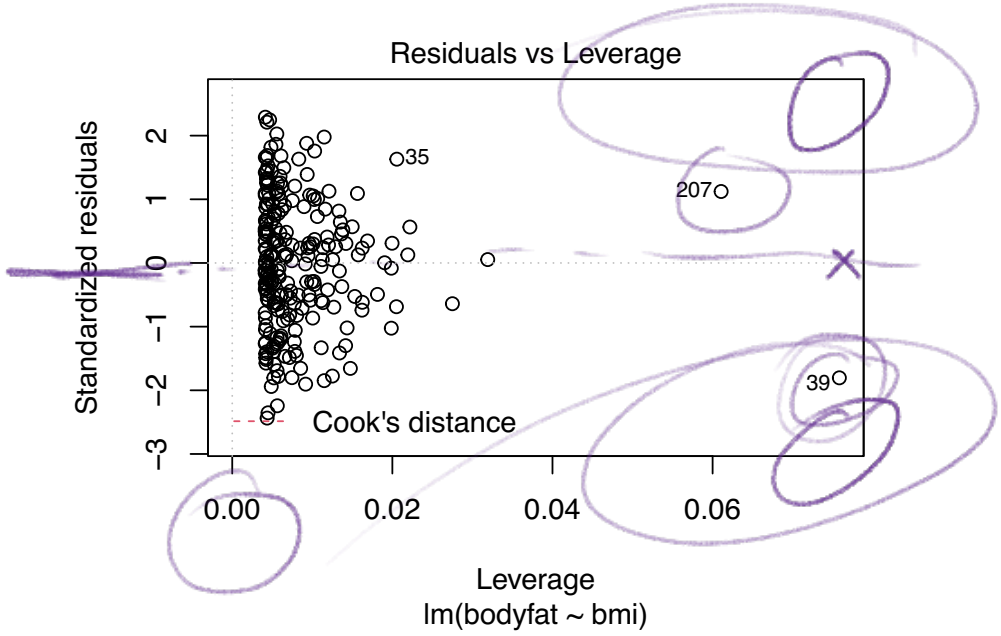
An outlier is a data point (value) that is an abnormal distance from the others.

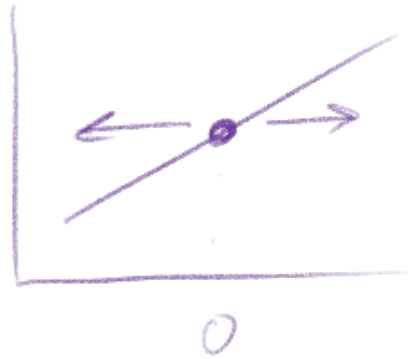
How to identify outliers:

- Look at the histograms of the data, and scatter plots (OK, but not foolproof).
- Look at the distribution of the residuals.

How to identify important outliers: * Look at "leverage".

Leverages ("Hebel")



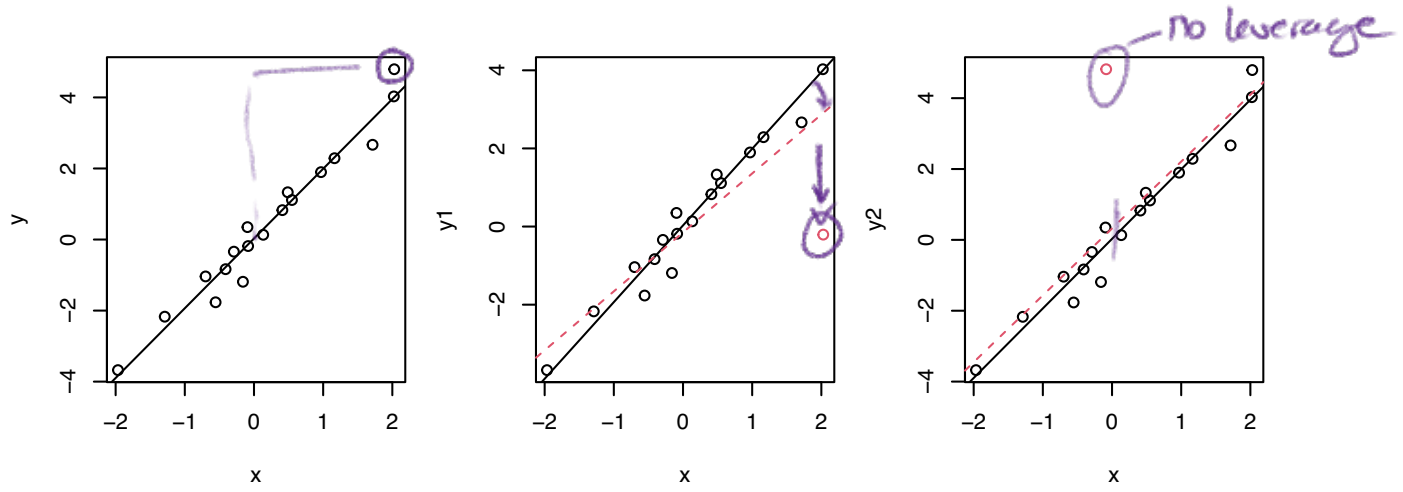


To understand the leverage plot, we need to introduce the idea of the *leverage* ("Hebel").

In simple regression, the leverage of individual i is defined as
$H_{ii} = (1/n) + (x_i - \overline{x})^2 / SSQ^{(X)}$.

Graphical illustration of the leverage effect

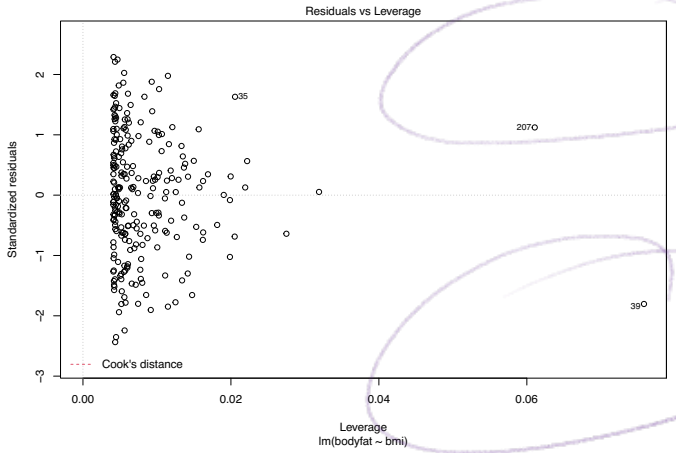
Data points with x_i values far from the mean have a stronger leverage effect than when $x_i \approx \overline{x}$:



The outlier in the middle plot "pulls" the regression line in its direction and biases the slope.

Leverage plot (Hebelarm-Diagramm)

In the leverage plot, (standardized) residuals $\tilde{R}_i$ are plotted against the leverage H_{ii} :

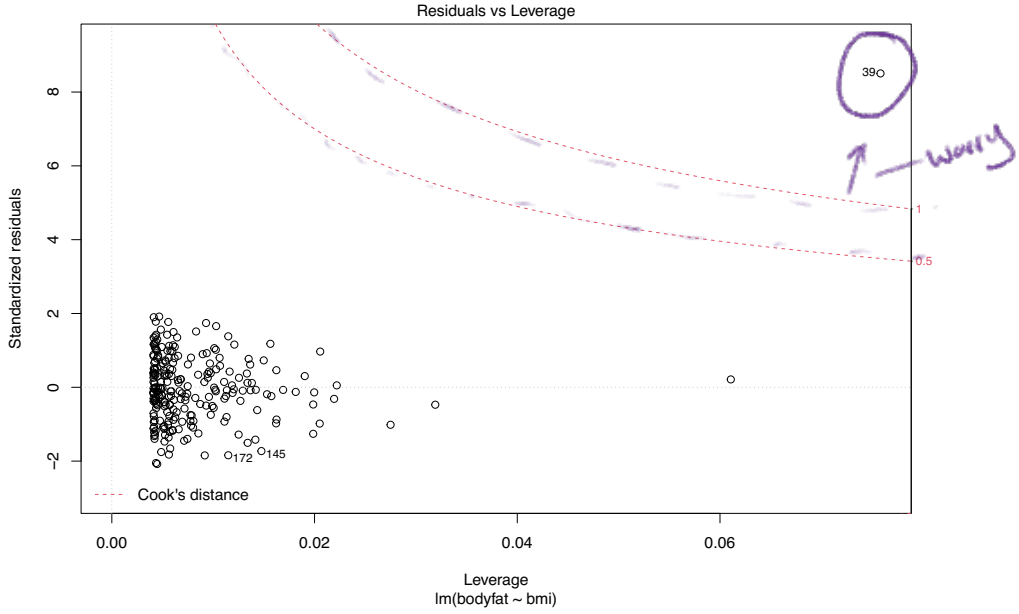


Critical ranges are the top and bottom right corners!!

Here, observations 36, 39, and 207 are labelled as potential outliers.

An extreme outlier...

Now I multiplied by 3 the observed value of bodyfat in observation 39.

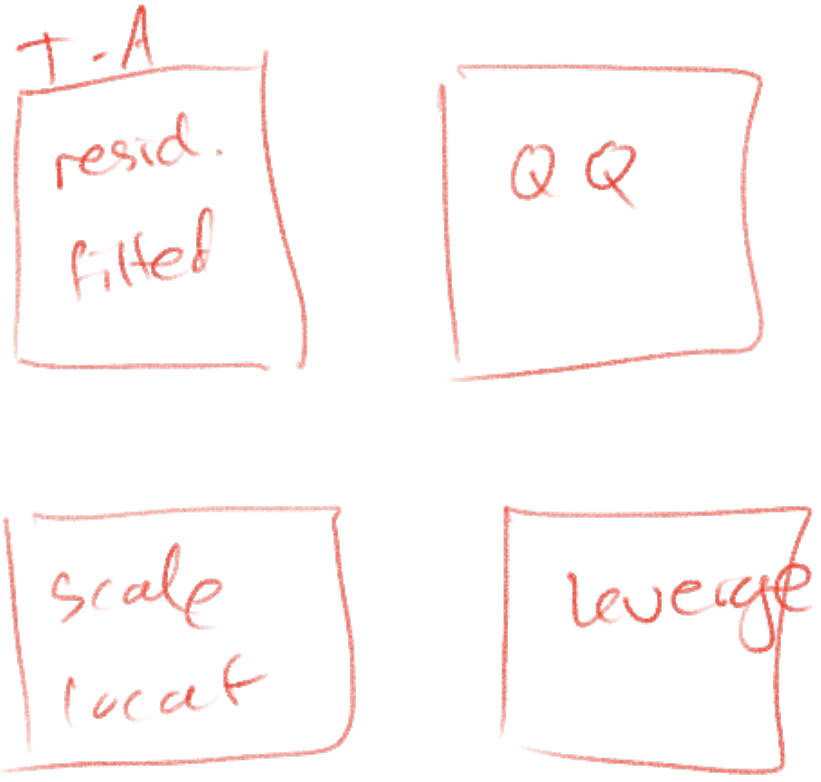


Some texts will give a rule of thumb that points with Cook's distances greater than 1 should be considered influential, while other books claim a reasonable rule of thumb is $4/(n-p-1)$ where n is the sample size, and p is the number of *beta* parameters.

What can go "wrong" during the modeling process?

Answer: a lot of things!

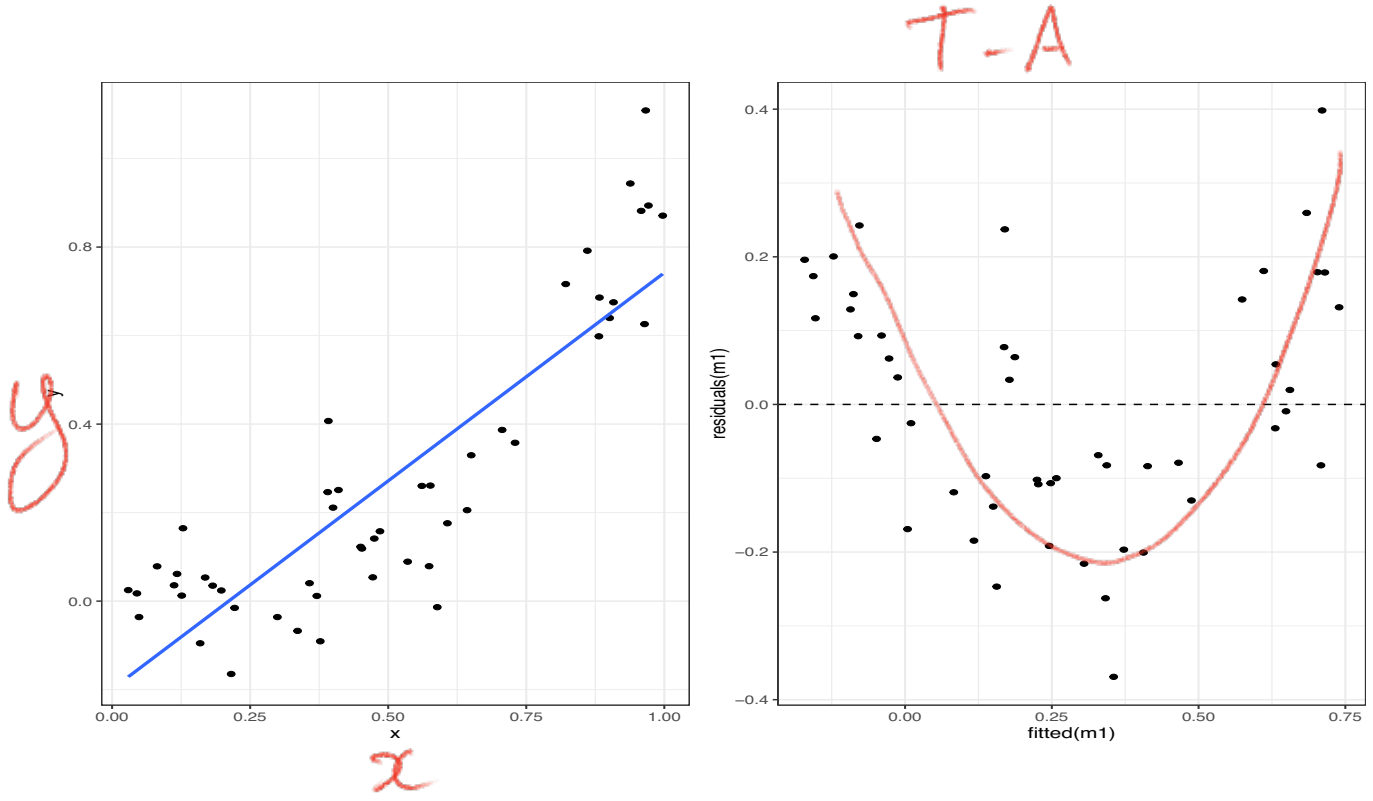
- Non-linearity.
- Non-normal distribution of residuals.
- Heteroscadisticity (unequal variance).
- Important outliers.



What to do when things "go wrong"?

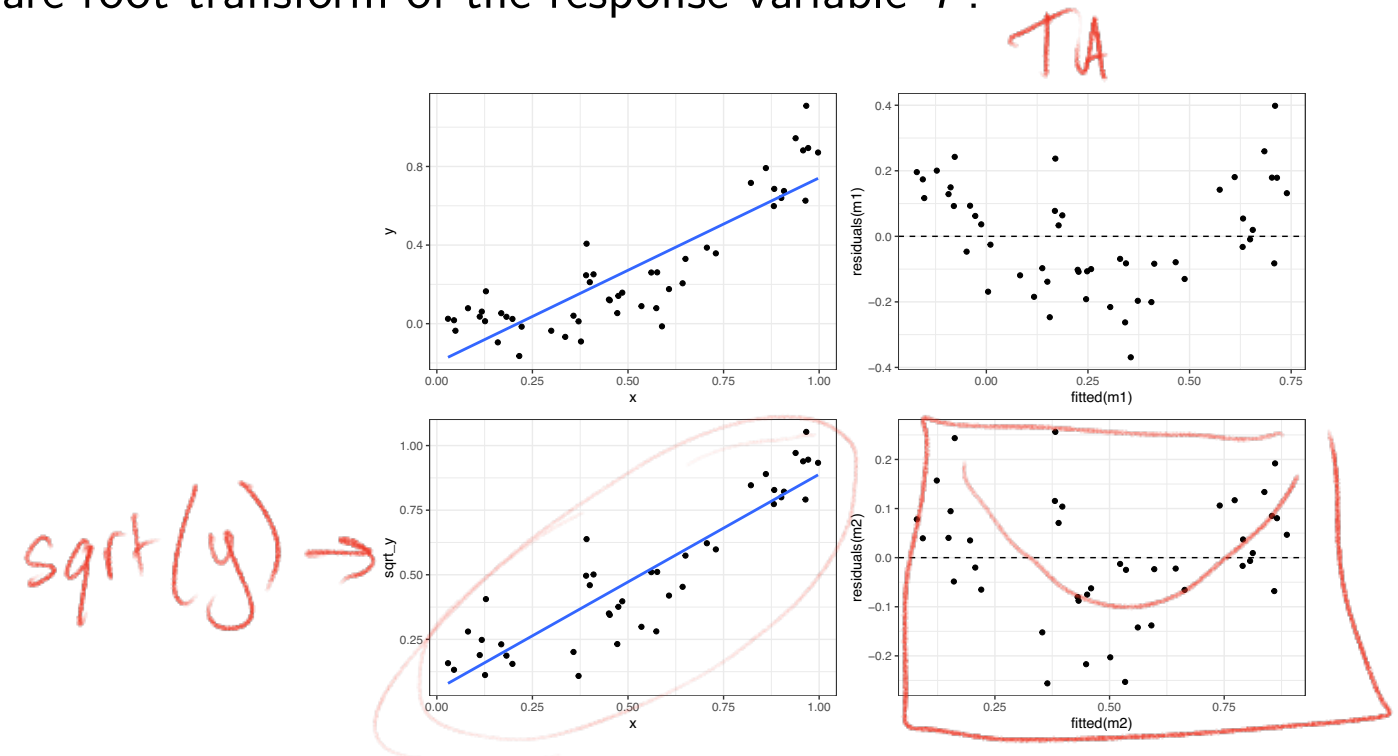
1. Now: Transform the response and/or explanatory variables.
2. Now: Take care of outliers.
3. Later in the course: Improve the model, e.g., by adding additional terms or interactions.
4. Later in the course: Use another model family (generalized or nonlinear regression model).
5. Not in this course: Use weighted regression.

Non-linearity



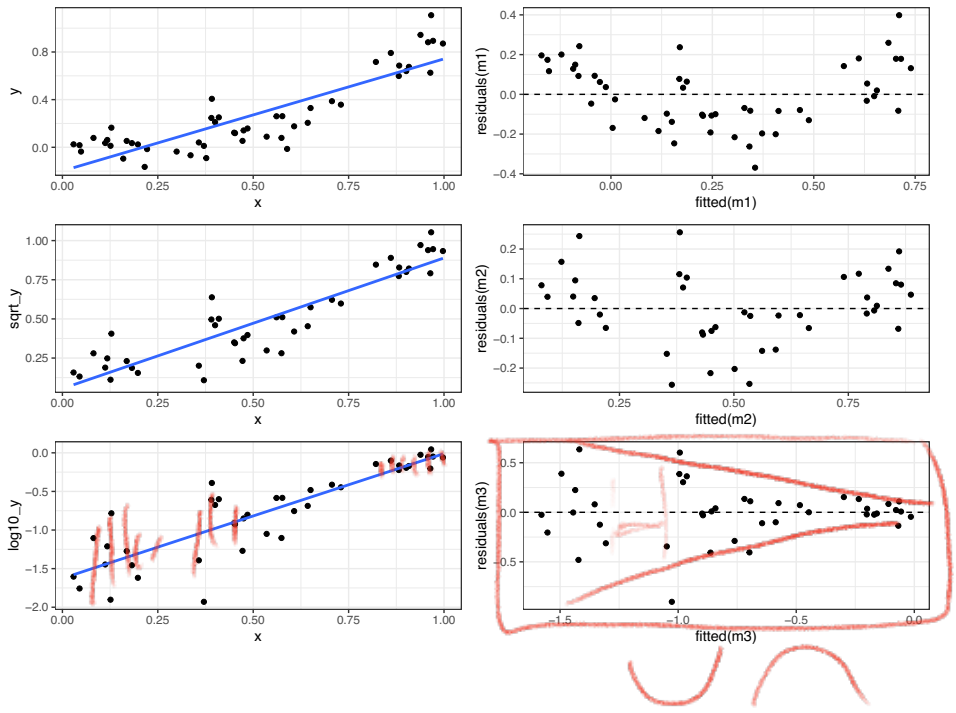
Transformation of the response?

Square root transform of the response variable Y:



Another transformation

Log transformation of the response variable Y:



Common transformations

Which transformations should be considered to cure model deviation symptoms? There is no simple answer. But some guidelines. E.g. if we see nonlinearity and increasing variance with increasing fitted values, then a log transform may improve matter.

- This is why we look at if the model is good / appropriate before we start using it. I.e. we make sure our parachute is in good working order *before* we jump from the aeroplane.

Some common and useful first aid transformations are:

- The log transformation for **concentrations** and **absolute values**.
- The square-root ($\sqrt{\cdot}$) transformation for **count data**.
- The arcsin($\sqrt{\cdot}$) transformation for **proportions/percentages**.

These transformations can also be applied on explanatory variables!

Outliers

What do we do when we identify the presence of one or more outliers?

1. Start by checking the "correctness" of the data. Is there a typo or a digital point that was shifted by mistake? Check both the response and explanatory variables.
2. If not, ask whether the model has been mis-specified. Do reasonable transformations of the response and/or explanatory variables eliminate the outlier? Do the residuals have a distribution with a long tail (which makes it more likely that extreme observations occur)?
3. Sometimes, an outlier may be the most interesting observation in a dataset!
4. Consider that outliers can also occur by chance!
5. Was the outlier created by some interesting but different process from the other data points.
6. Only if you decide to report the results of both scenario can you check if inclusion/exclusion changes the qualitative conclusion, and by how much it changes the quantitative conclusion.

Deleting outliers

It might seem tempting to delete observations that apparently don't fit into the picture. However:

- Do this **only with absolute care** e.g., if an observation has extremely implausible values!
- Before deleting outliers, check points 1-6 from the previous slide.
- When deleting outliers or the x% of most extreme observations, you **must mention this in your report**.

Overview

- 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

During the course we'll see many more examples of things going at least a bit wrong. And we'll do our best to improve the model, so we can be confident in it, and start to use it. Which we will start to do next week in Lecture 4.

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:

- Consolidate your understanding of today's lecture.
- Go to OLAT and do all the "Homework" tasks.

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