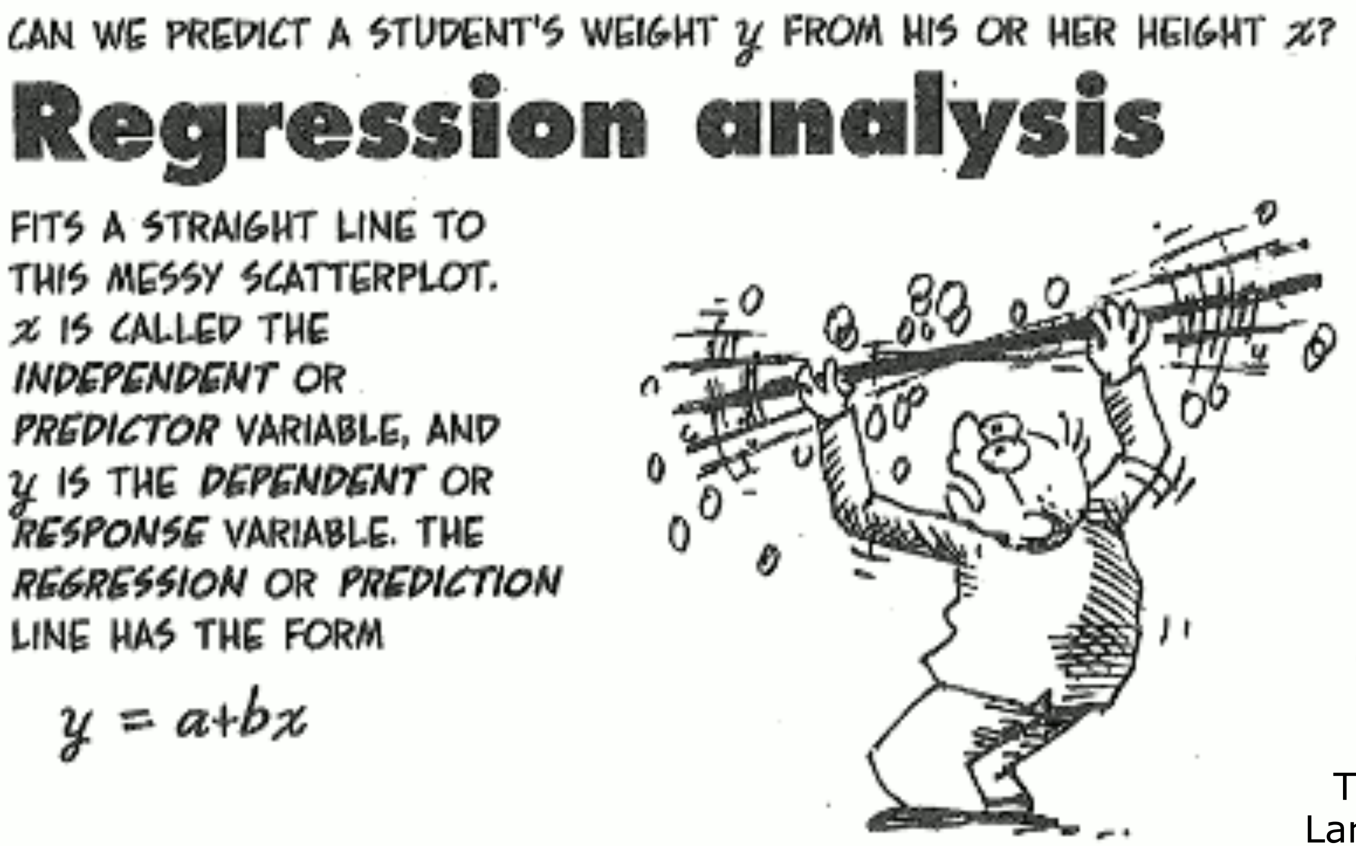


The Cartoon Guide to Statistics,
Larry Gonick and Woollcott Smith

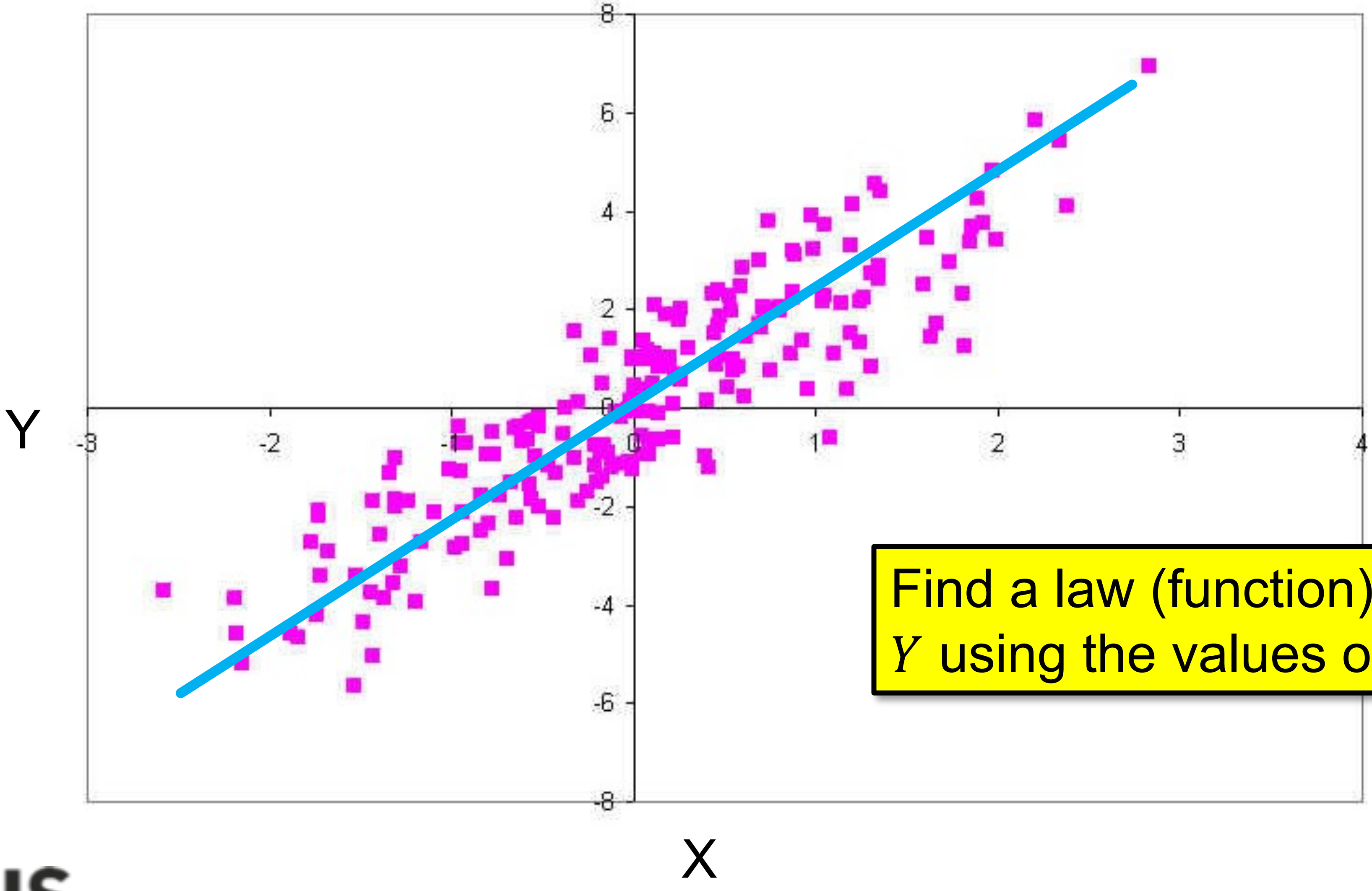
Linear regression

Biostatistics, ETH HG E 21

Goals today

- Get an intuition for (simple) linear regression
- Parameter estimation
- Checking the assumptions of a linear regression

Relation between two variables

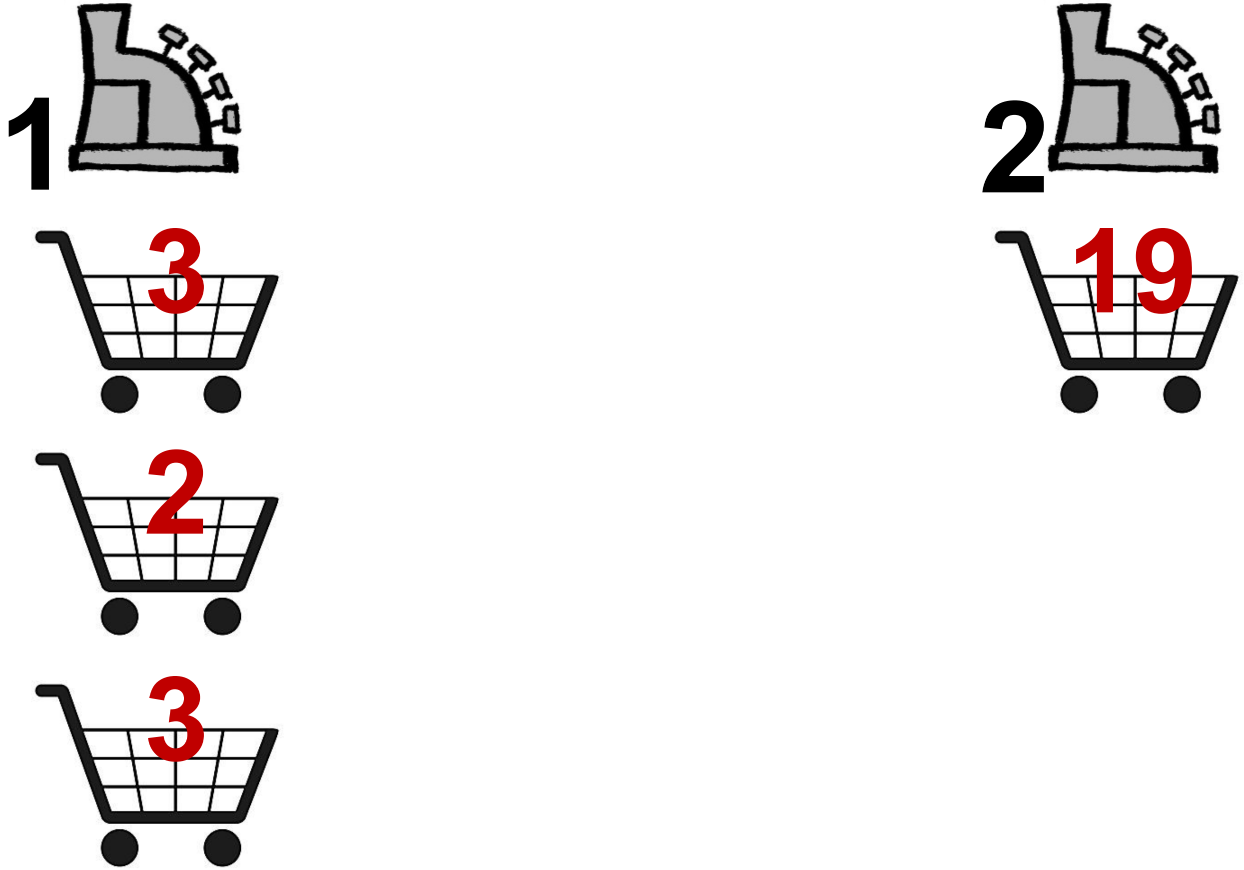


Find a law (function) explaining Y using the values of X

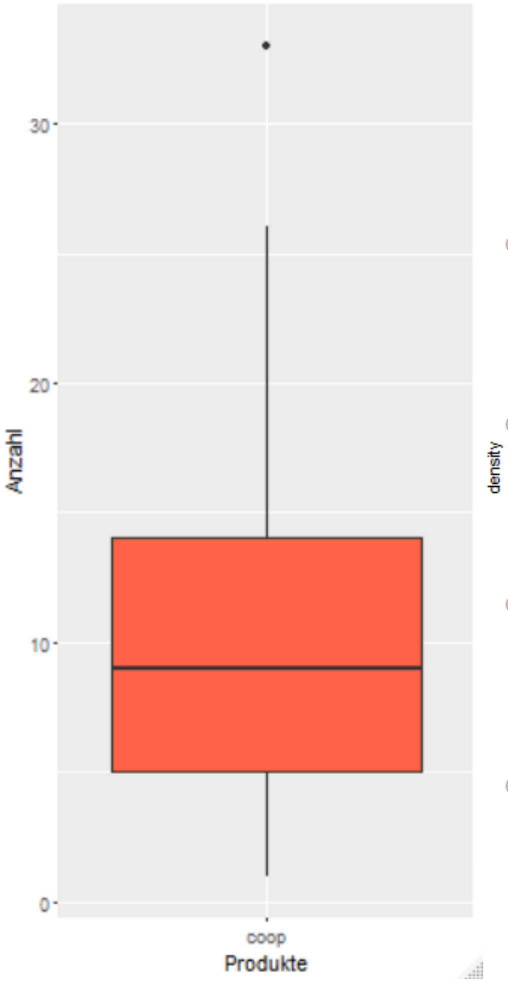
nexus
Personalized Health Technologies

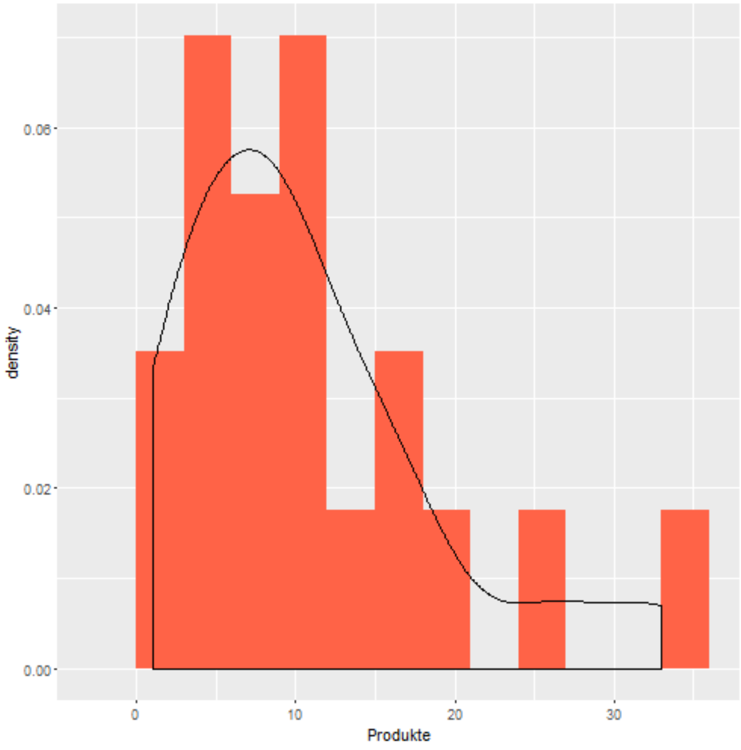
Example adapted from Statistik I, Dr. Markus Kalisch, Seminar for Statistics, ETH Zurich

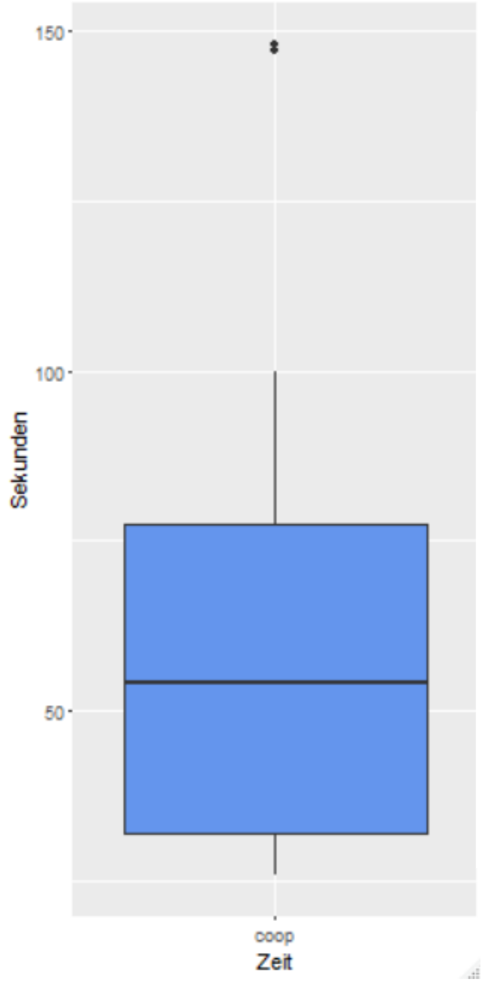
Where to que?

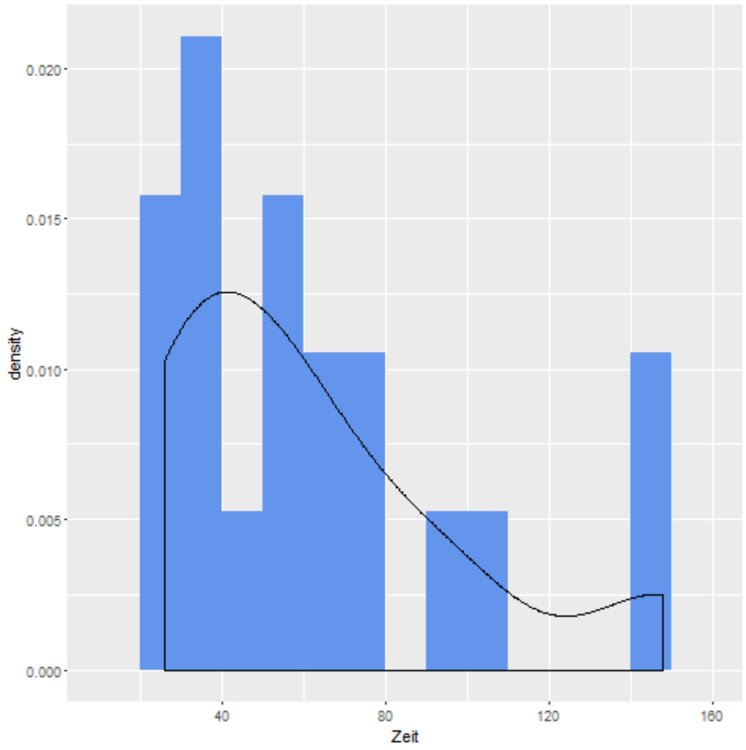


Coop Zurich Main Station – one cashier from 17:40-18:00

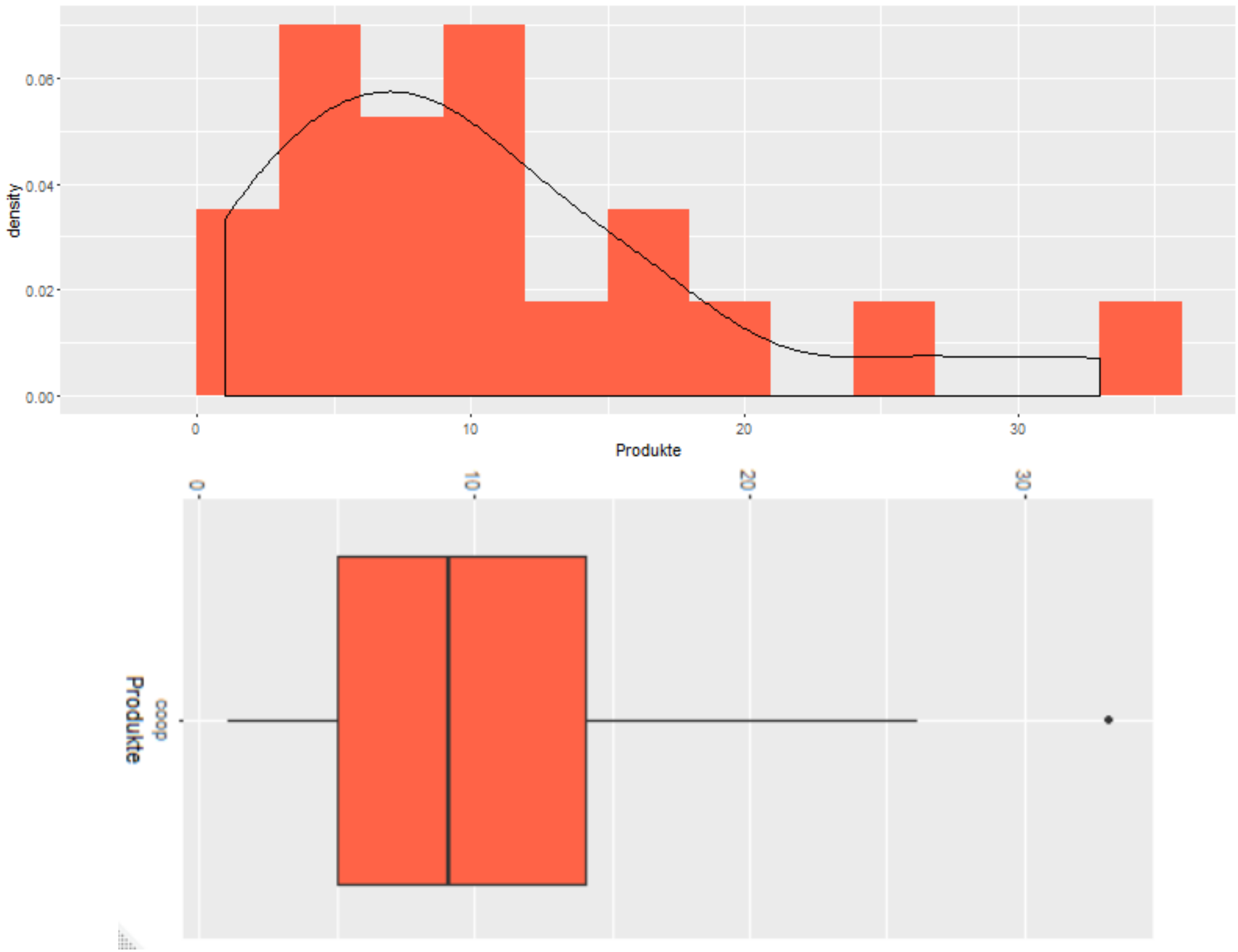




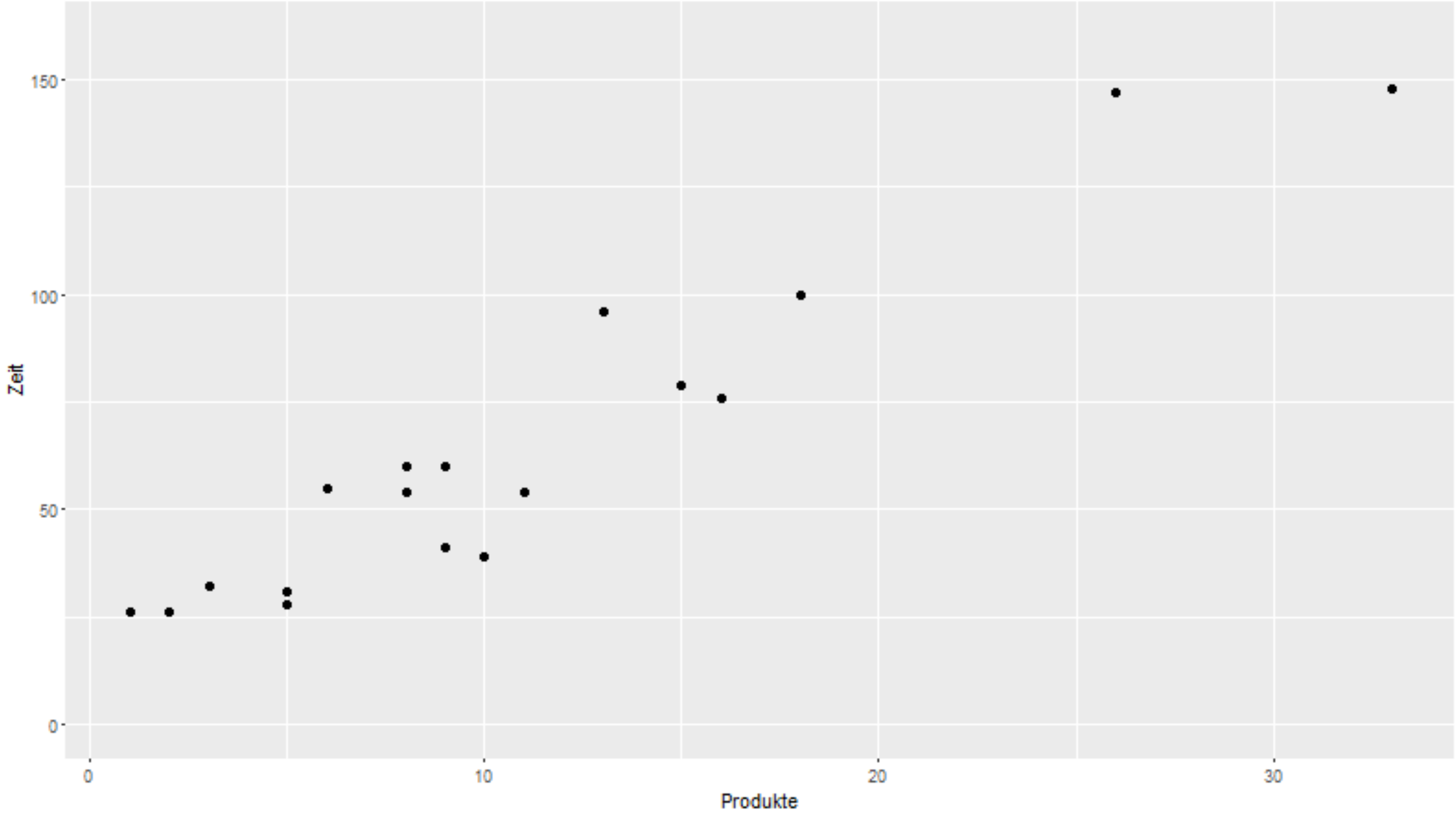




nexus Personalized Health Technologies


density
0.00
0.02
0.04
0.06
0
10
20
30
Produkte
0
10
20
30
Produkte
coop

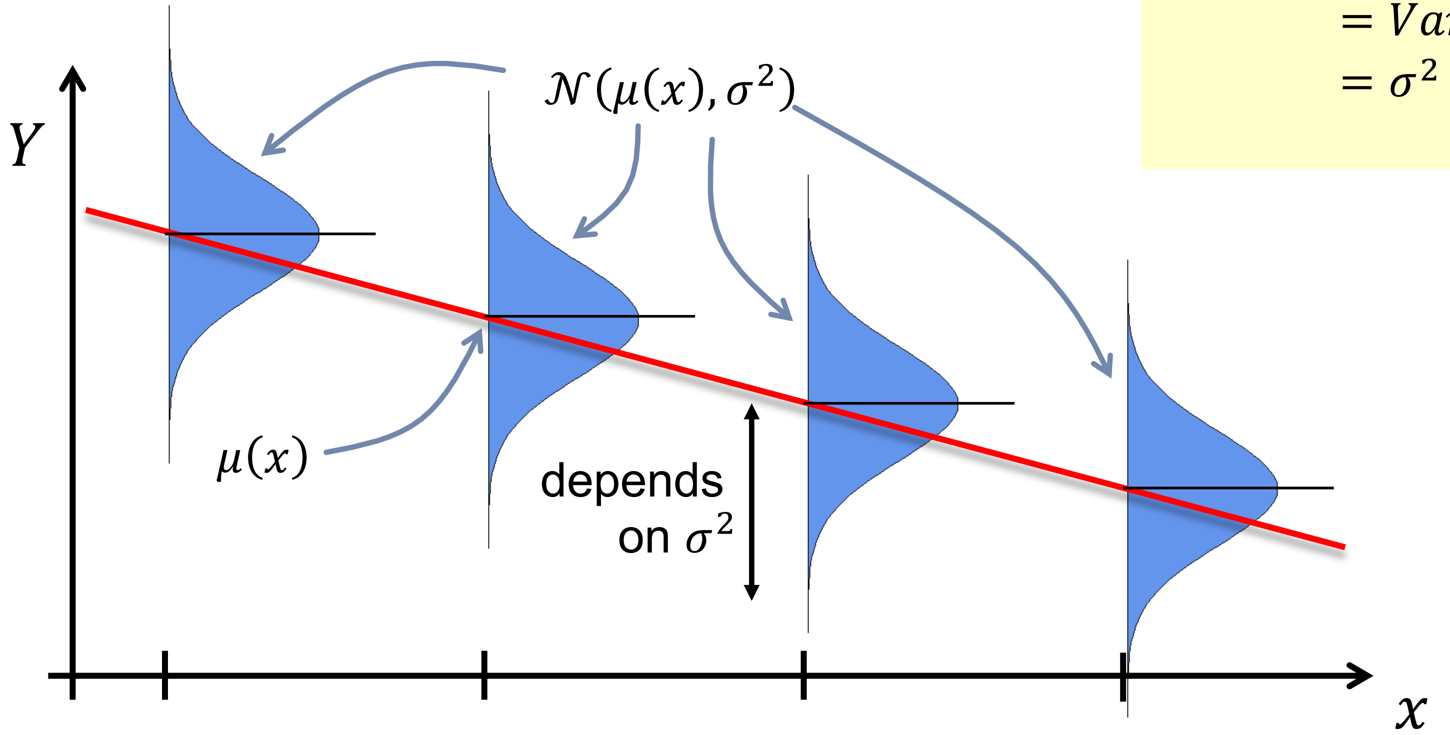
Scatter plot of the data



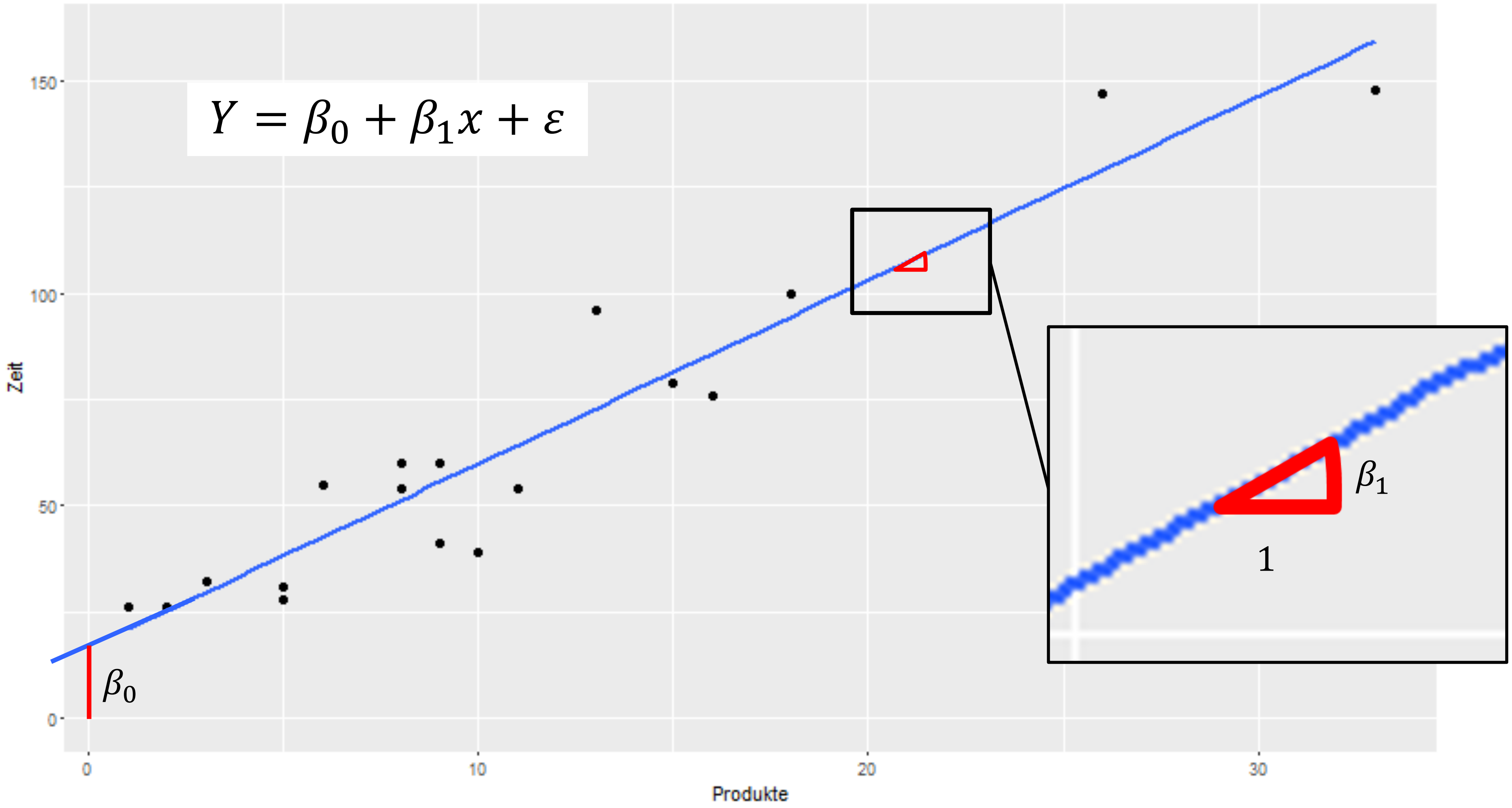
Linear regression: two definitions

1. $Y \sim \mathcal{N}(\mu(x), \sigma^2)$
   - $\mu(x) = \beta_0 + \beta_1 x$
2. $Y = \beta_0 + \beta_1 x + \varepsilon$
   - $\varepsilon \sim N(0, \sigma^2)$

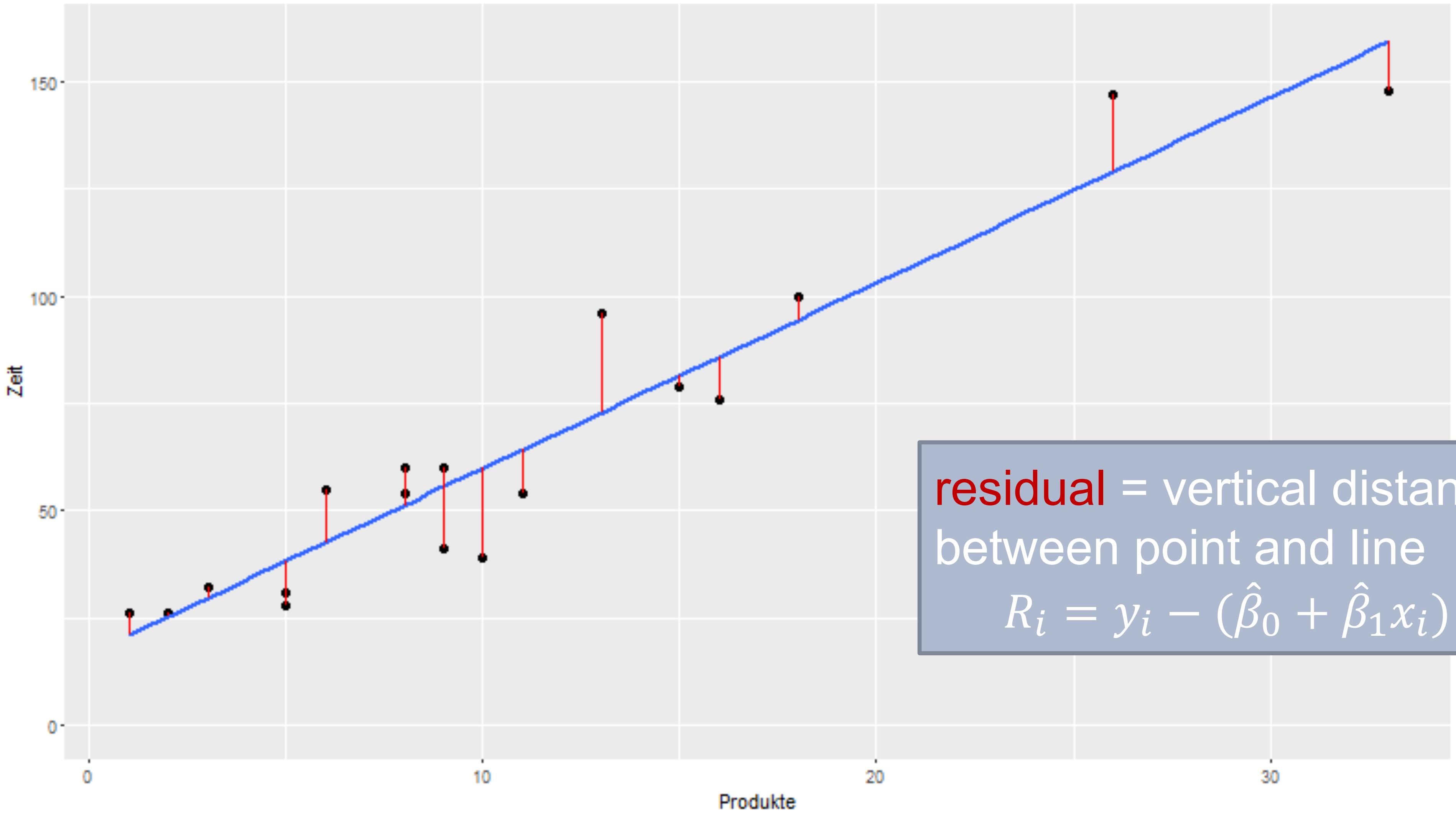
$$
\begin{aligned}
E(Y) &= E(\beta_0 + \beta_1 x + \varepsilon) \\
&= \beta_0 + \beta_1 x + E(\varepsilon) \\
&= \beta_0 + \beta_1 x \\
Var(Y) &= Var(\beta_0 + \beta_1 x + \varepsilon) \\
&= Var(\varepsilon) \\
&= \sigma^2
\end{aligned}
$$



Regression line



Residuen



residual = vertical distance between point and line

$$R_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Parameter estimation – option 1
Ordinary Least Squares (OLS)

- Which line fits into the points the best?
- Choose $\hat\beta_0, \hat\beta_1$ to minimise the sum of squared residuals:

$$\hat\beta_0, \hat\beta_1 \text{ minimise } \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

Parameter estimation – option 2
Maximum Likelihood Estimation (MLE)

- $Y_i \sim \mathcal{N}(\mu(x_i), \sigma^2)\ i.i.d.$
- Likelihood: $\mathcal{L}(\beta_0, \beta_1) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\left(y_i - \mu(x_i)\right)^2}{\sigma^2}\right)\right)$
- log-Likelihood:

$$\ell(\beta_0, \beta_1) = \log\left(\mathcal{L}(\beta_0, \beta_1)\right) = -n\pi\sigma^2 - \frac{1}{2}\frac{\left(\sum_{i=1}^{n}\left(y_i - \mu(x_i)\right)^2\right)}{\sigma^2}$$
$$= -n\pi\sigma^2 - \frac{1}{2}\frac{\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right)}{\sigma^2}$$

- log-Likelihood is maximised, if $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ is minimised
- In the situation of simple linear regression MLE is equivalent to OLS

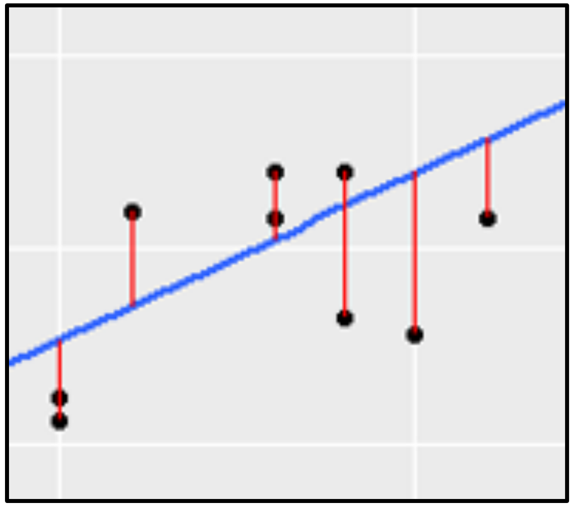
Estimating σ^2

- Once again, the residuals are

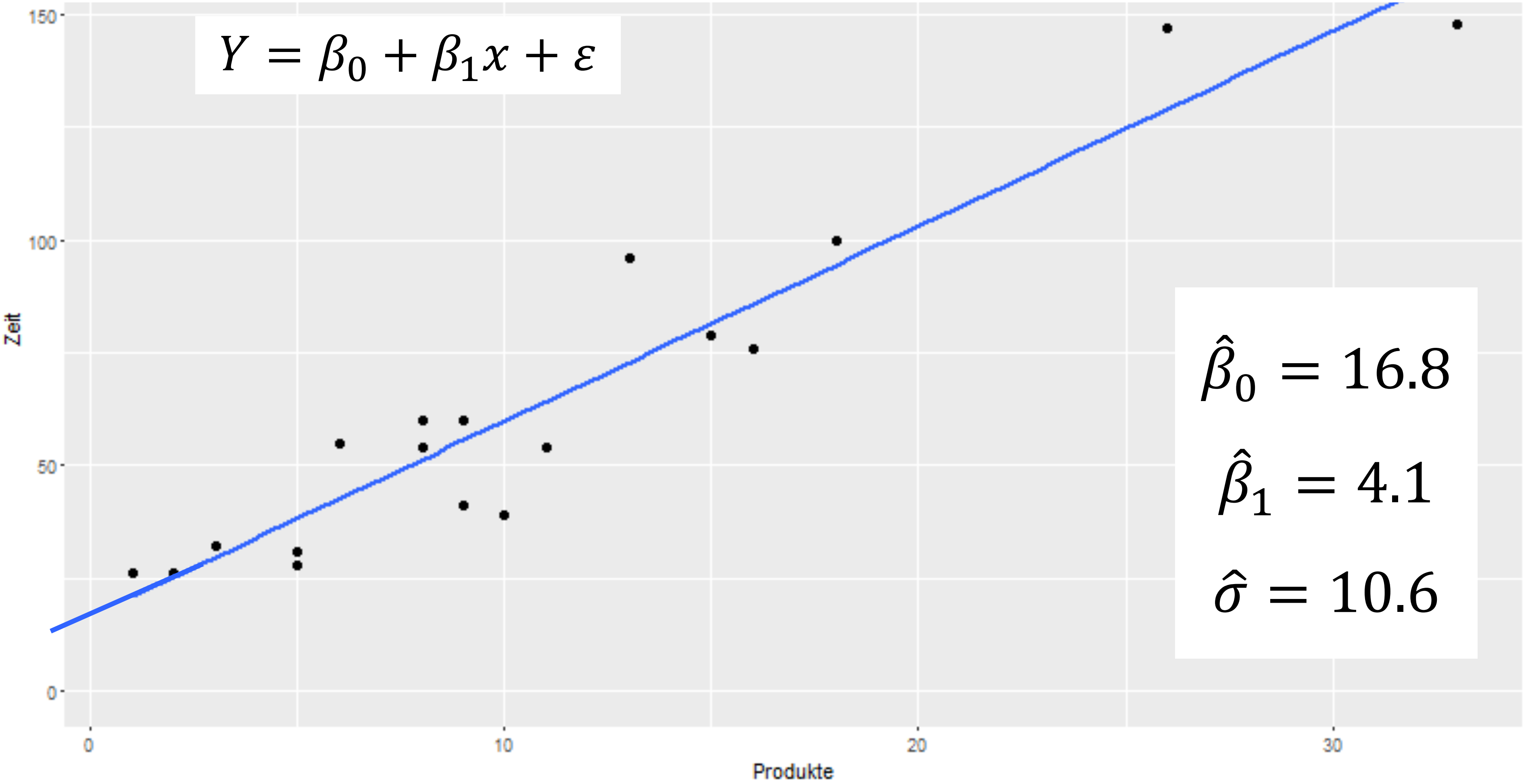
$$R_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right), \qquad i = 1, 2, \dots, n$$

- We simply estimate the variance of the residuals:

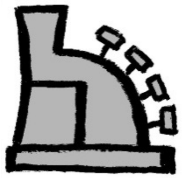
$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} R_i^2$$



Regressionslinie



Wo anstehen?

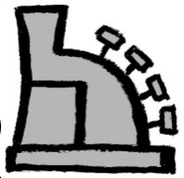
1 

83.2

16.8 + **3**·4.1= 29.1 s

16.8 + **2**·4.1= 25 s

16.8 + **3**·4.1= 29.1 s

2 

94.7

16.8 + **19**·4.1= 94.7 s

$\hat{\beta}_0 = 16.8$

$\hat{\beta}_1 = 4.1$

$\hat{\sigma} = 10.6$

Example adapted from Statistik I, Dr. Markus Kalisch, Seminar for Statistics, ETH Zurich

Test für $\boldsymbol{\beta_0}$ und $\boldsymbol{\beta_1}$

- X, Y are random variables & $\hat{\beta}_0, \hat{\beta}_1$ are functions of X and Y
  $\Rightarrow$ $\hat{\beta}_0, \hat{\beta}_1$ are also random variables

One can proof that $\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2_{\beta_i})$, using…

$$\hat{\sigma}_{\beta_1} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}$$

…and also

$$\frac{\hat{\beta}_1 - \beta_{1,\mathcal{H}_0}}{\hat{\sigma}_{\beta_1}} \sim t_{n-2}$$

Requirements for linear regression

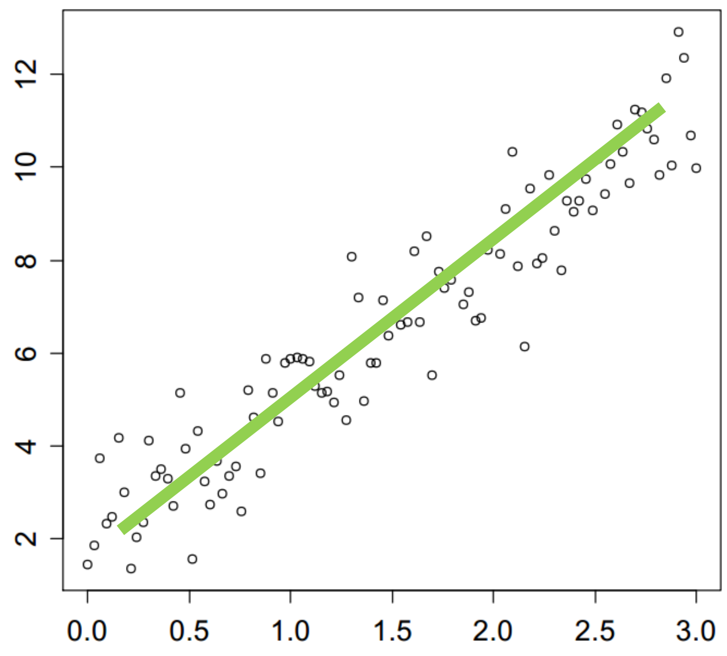
Two definitions:

1. $Y \sim \mathcal{N}(\mu(x), \sigma^2)$
   - $\mu(x) = \beta_0 + \beta_1 x$
2. $Y = \beta_0 + \beta_1 x + \varepsilon$
   - $\varepsilon \sim N(0, \sigma^2)$

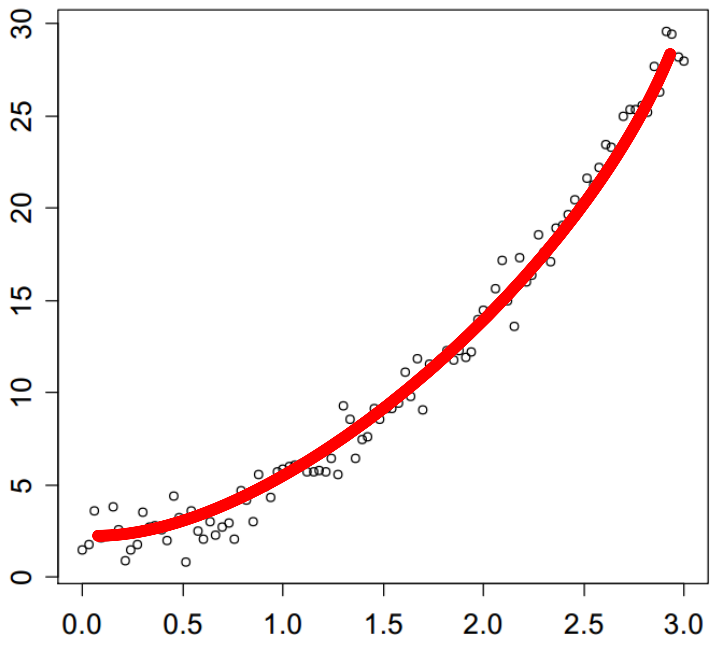
- **Linearity** – Y can be explained using a linear combination, e.g., $Y = \beta_0 + \beta_1 \cdot X + \varepsilon$
- **Constant variance** – error has a constant variance (independent of X)
- **Normality** – error ε needs to be normal distributed

If these assumptions are strongly violated the model is not valid

Linearity: Scatter plot for simple linear regression



OK

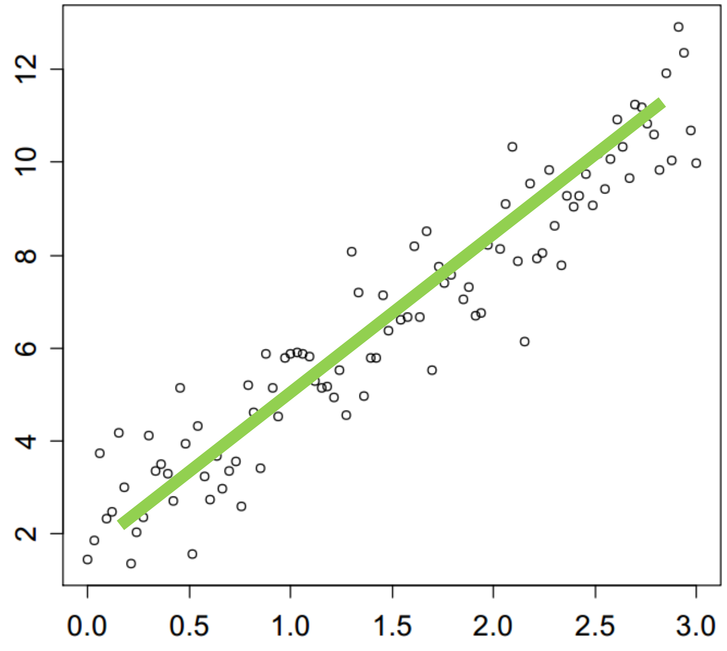


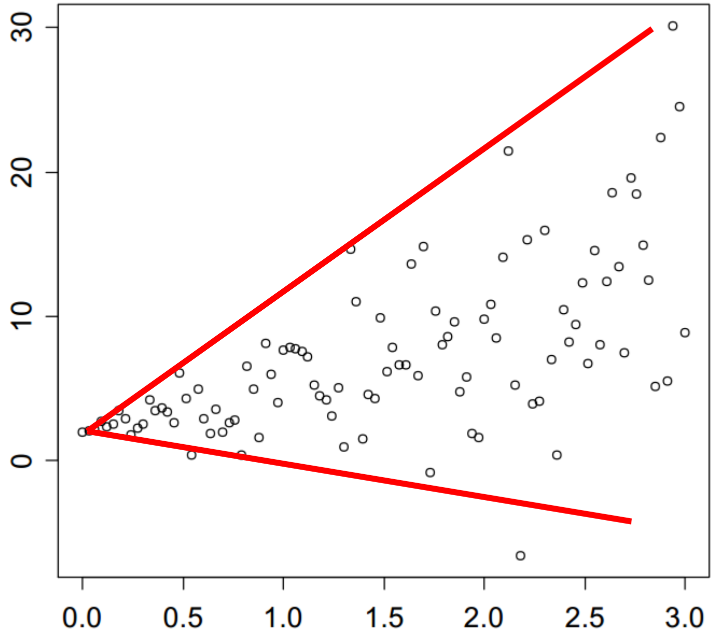
Systematic error

Curvature:

$$y = b_0 + b_1 x + b_2 x^2$$

Linearity: Scatter plot for simple linear regression

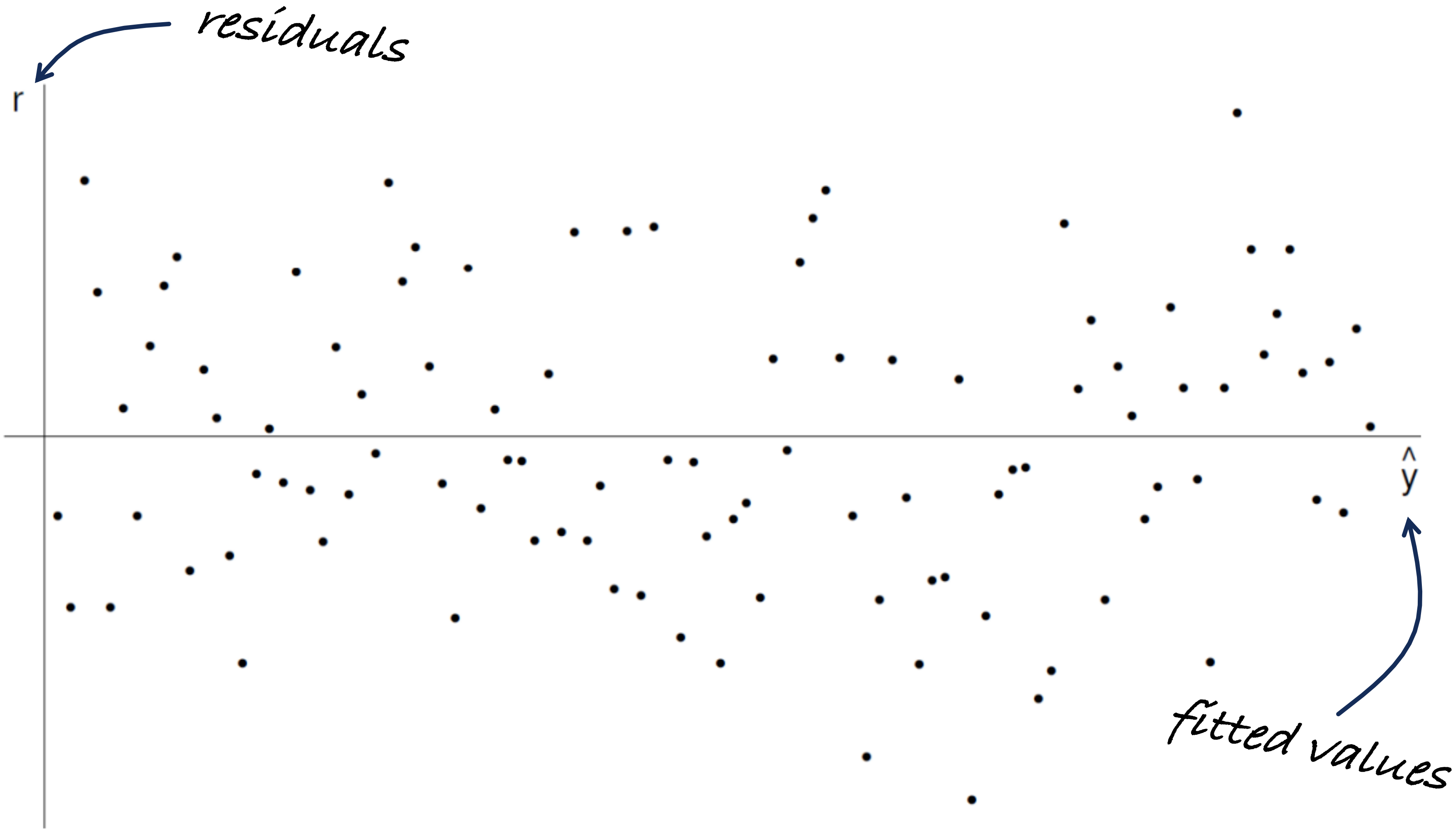




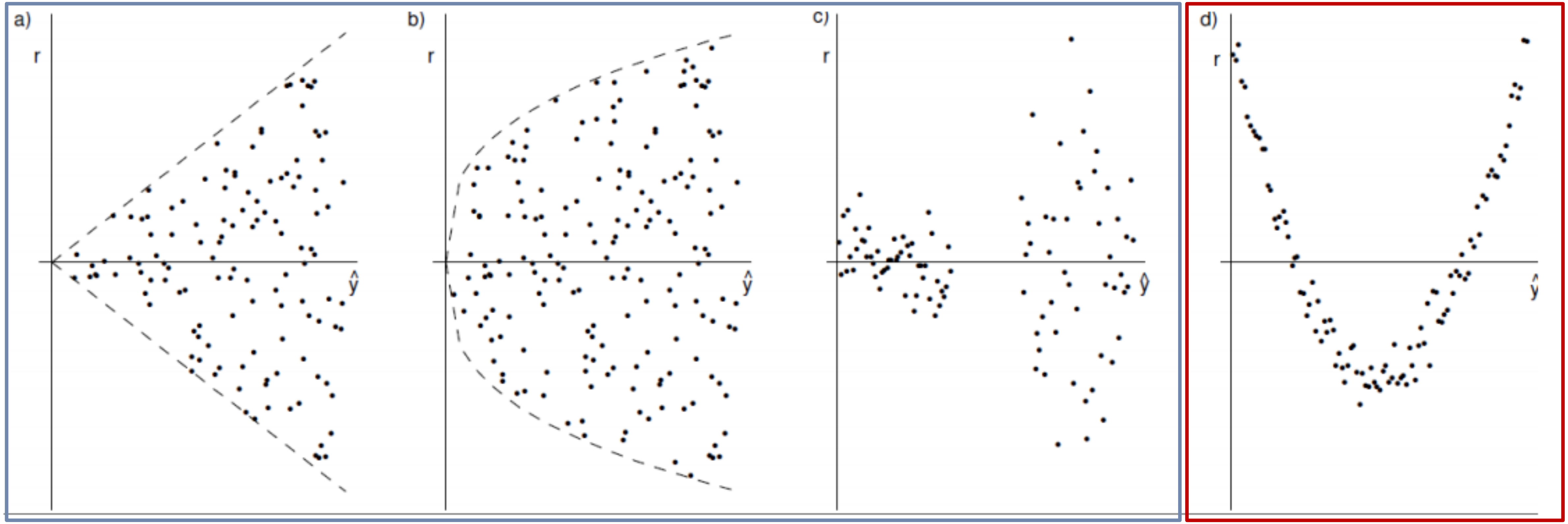
OK

Error variance is not constant

Constant variance: Example for a good TA-plot



Constant variance: Examples for bad TA-plots



Error variance not constant

Systematic error

QQ-Plot

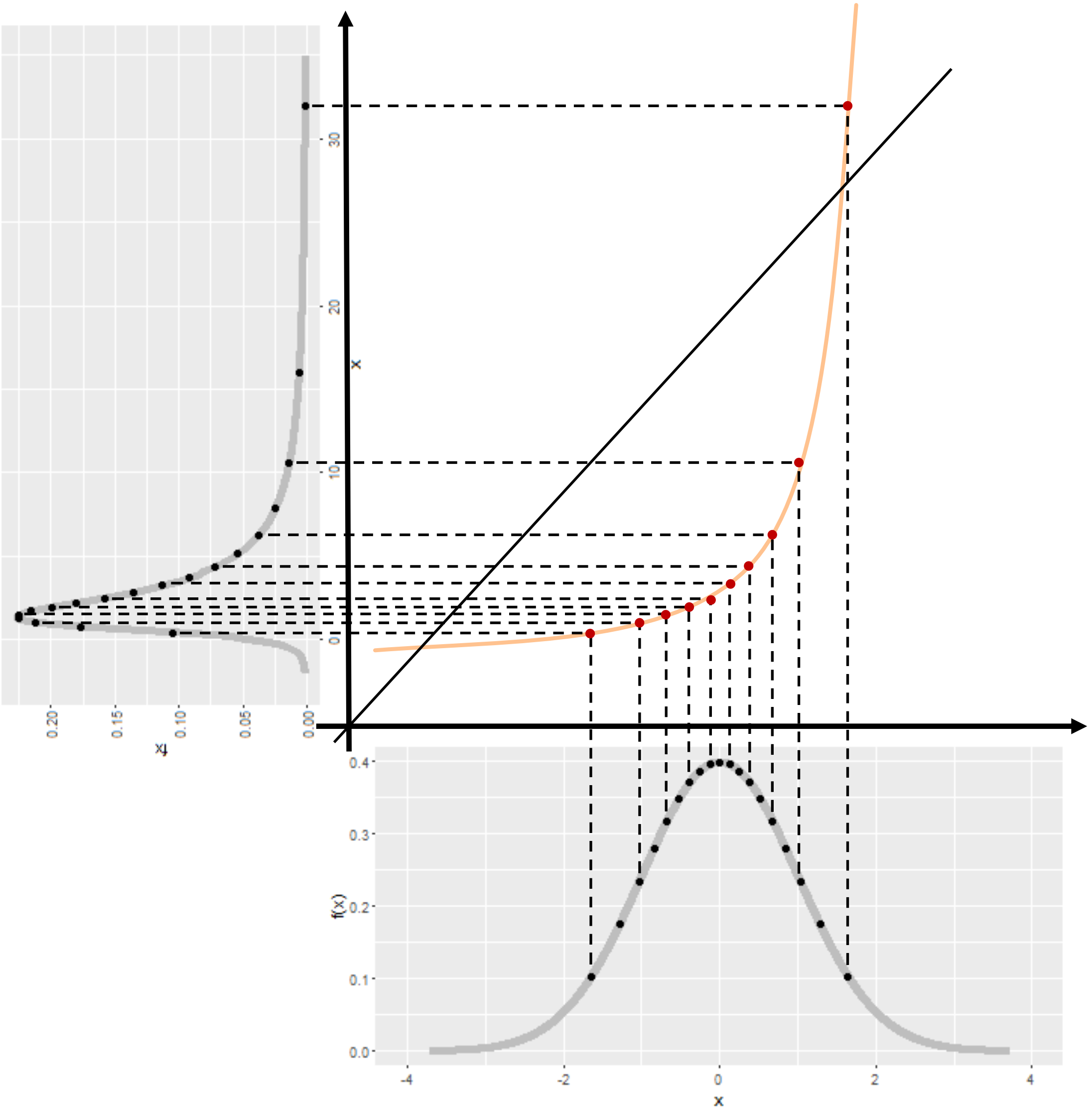
- Plot on the x-axis the theoretical quantiles:

$$q_1 = \frac{0.5}{n}, q_2 = \frac{1.5}{n}, \ldots, q_n = \frac{n-0.5}{n}$$

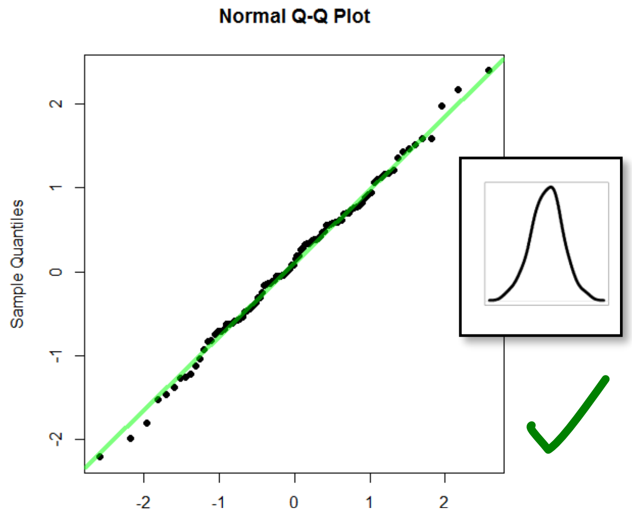
- Plot on the y-axis the empiric quantiles:

$$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$$

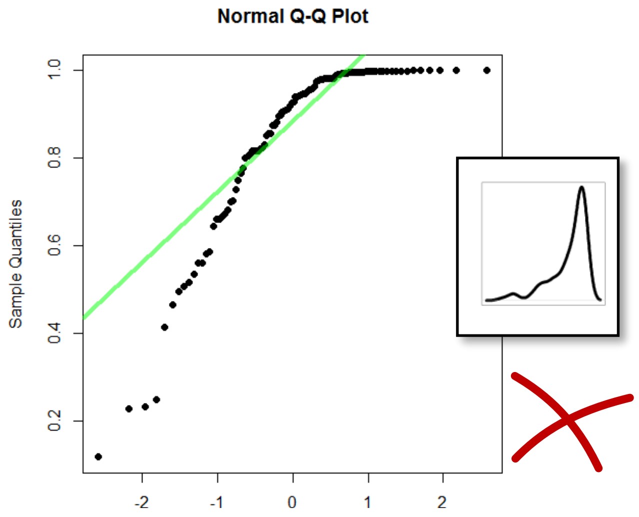
  - $x_{(i)}$: ordered observations



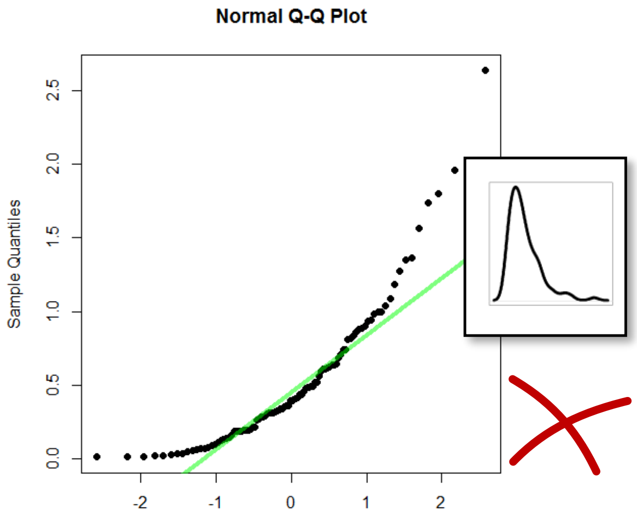
Normality: QQ-plot



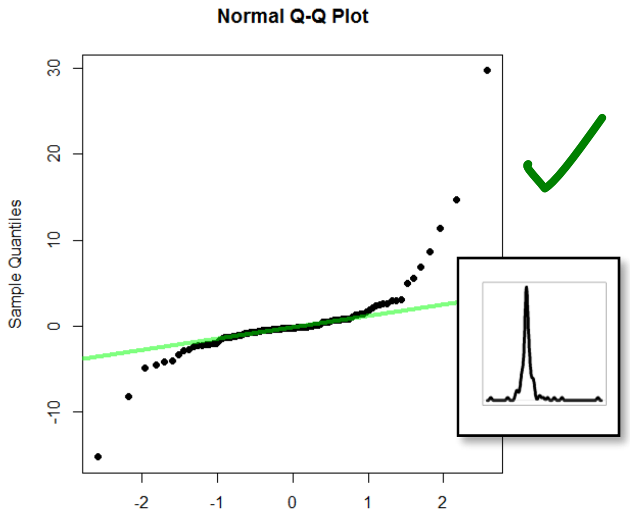
normal



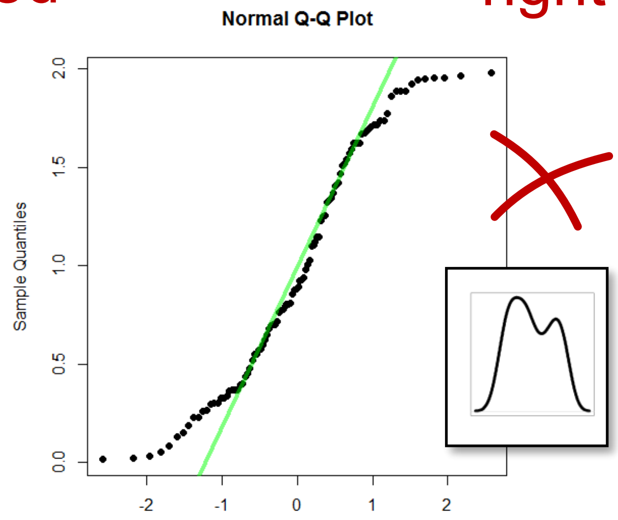
left skewed



right skewed

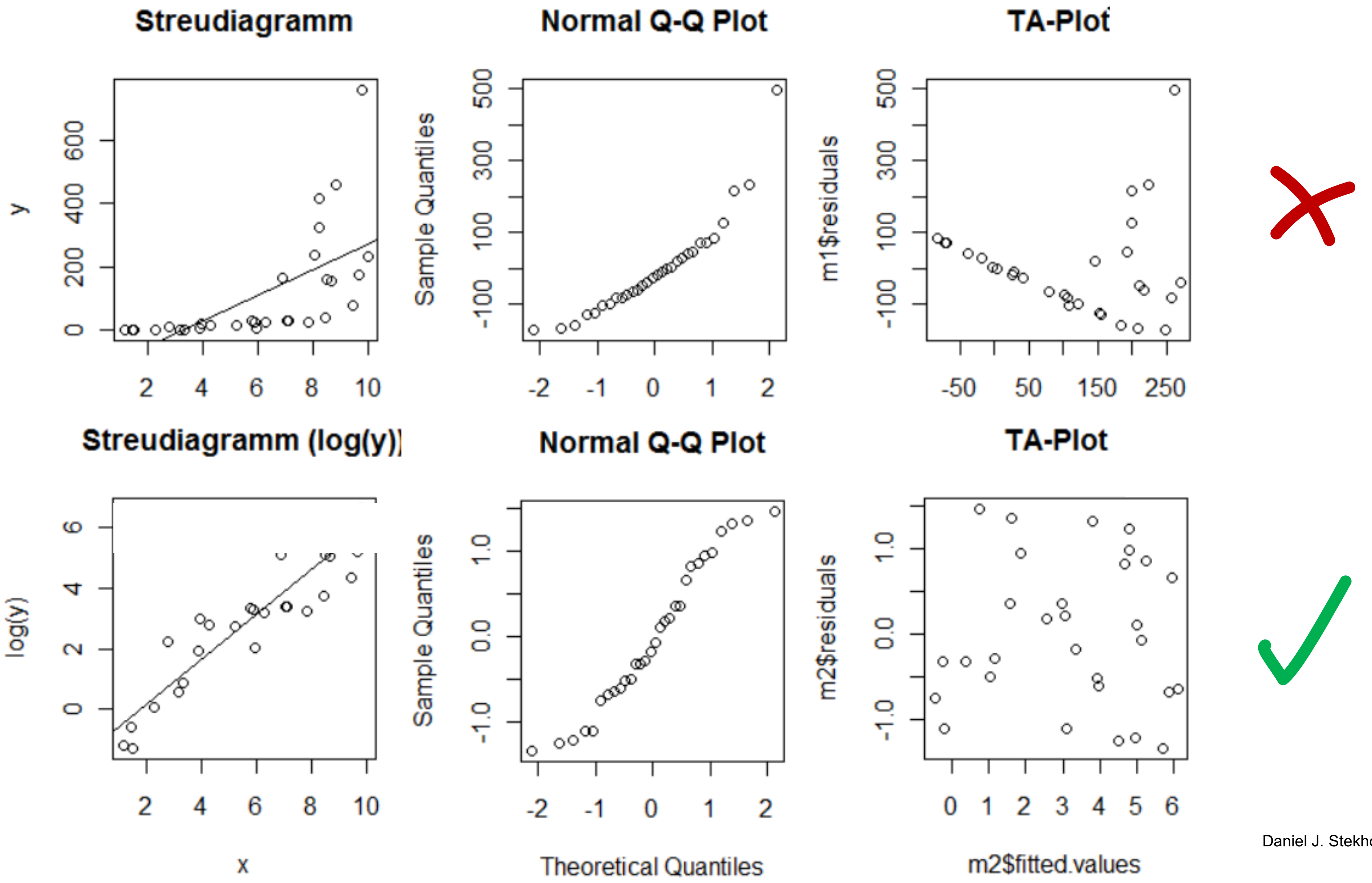


heavy tailed



short tailed

…what if residual analysis is haywire



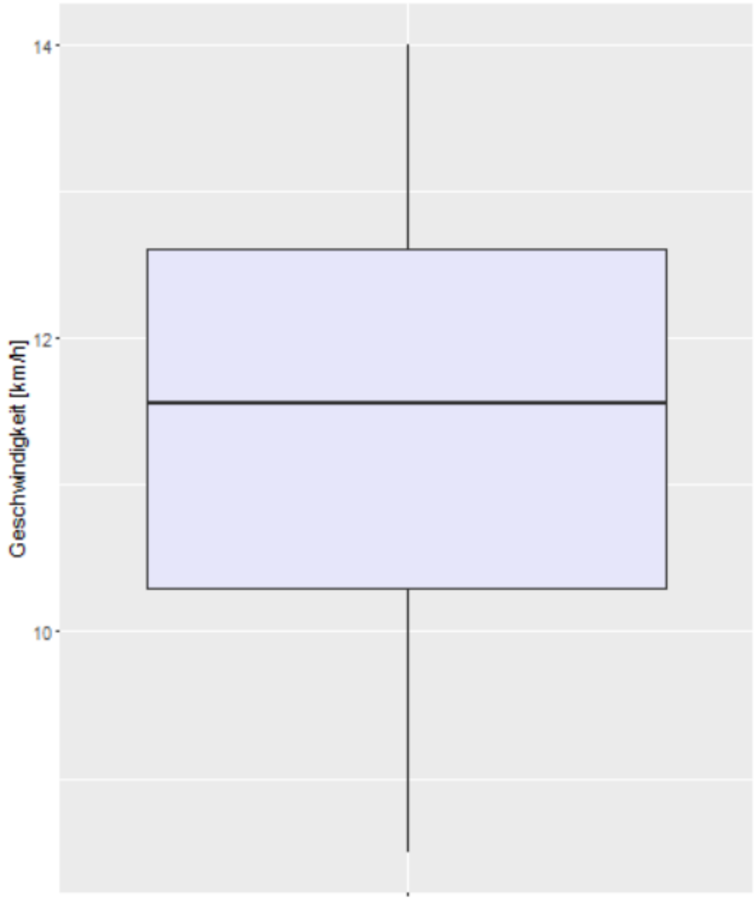
Aerobic performance

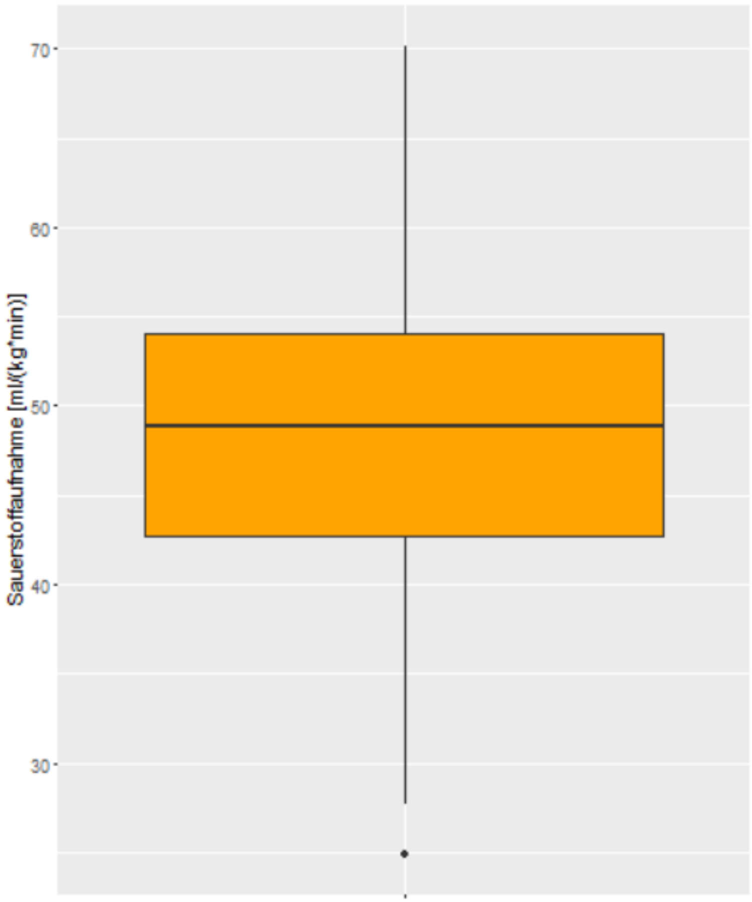
- VO_2max: amount of oxygen, the body can absorb per kg mass and minute
- Test is expensive and effortful
- Not meant for the broad community
- Alternative?

Example adapted from Statistik I, Dr. Markus Kalisch, Seminar for Statitics, ETH Zurich

Léger et al., 1983

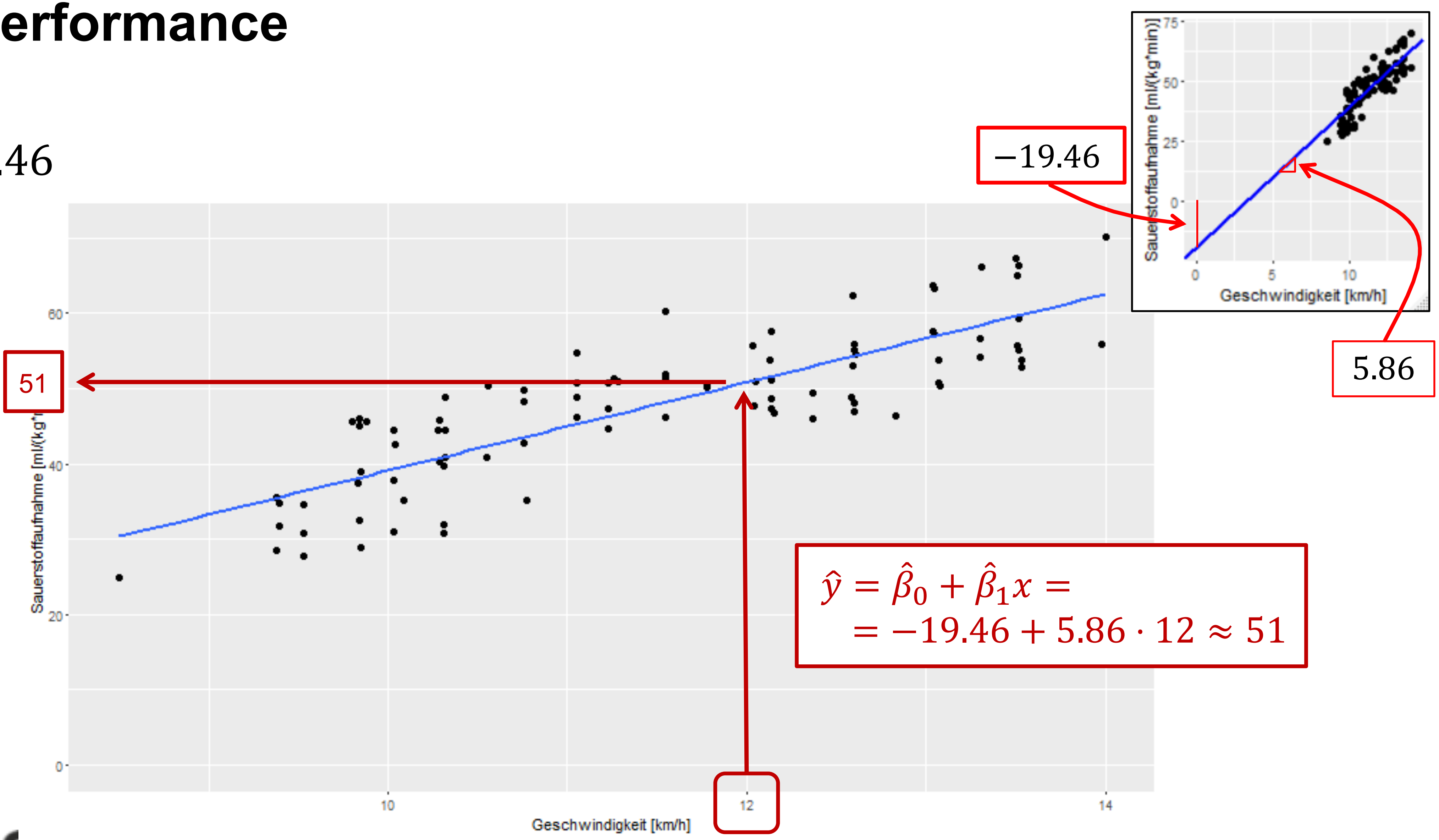
- 91 subjects, 20m-shuttle-test and VO_2max measurement





Aerobic Performance

- $\hat{\beta}_0 = -19.46$
- $\hat{\beta}_1 = 5.86$
- $\hat{\sigma} = 5.4$



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \\ = -19.46 + 5.86 \cdot 12 \approx 51$$

Linear regression in R

- Model: $Y_i = \beta_0 + \beta_1 x_i + E_i, E_i \sim \mathcal{N}(0, \sigma^2)\ \ i.i.d.$
- Model: $Y_i = \mathbf{-19.46} + \mathbf{5.86} \cdot x_i + E_i, E_i \sim \mathcal{N}(0, \mathbf{5.43}^2)\ \ i.i.d$

Standard error of $\hat{\beta}_1$
approx. 95%-CI:
$5.86 \pm 2 \cdot 0.41$
exact 95%-CI:
$5.86 \pm 1.99 \cdot 0.41$

$t_{89}; 0.975$

```
> fit <- lm(vo2max ~ vmax, data = dat)
> summary(fit)

Call:
lm(formula = vo2max ~ vmax, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2230  -4.3976  -0.2016   4.7026  12.0348

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.4582     4.7239  -4.119  8.5e-05 ***
vmax          5.8566     0.4082  14.347  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.433 on 89 degrees of freedom
Multiple R-squared:  0.6981, Adjusted R-squared:  0.6948
F-statistic: 205.8 on 1 and 89 DF,  p-value: < 2.2e-16
```

Observed Test statistic t
in the test:
$\mathcal{H}_0: \beta_1 = 0$ vs $\mathcal{H}_A: \beta_1 \neq 0$

P-value:
Assume $\beta_1 = 0$; what is the probability of t or an even more extreme value?

Degrees of freedom:
$n - (\text{Number of } \beta' \text{s})$
$= 91 - 2 = 89$

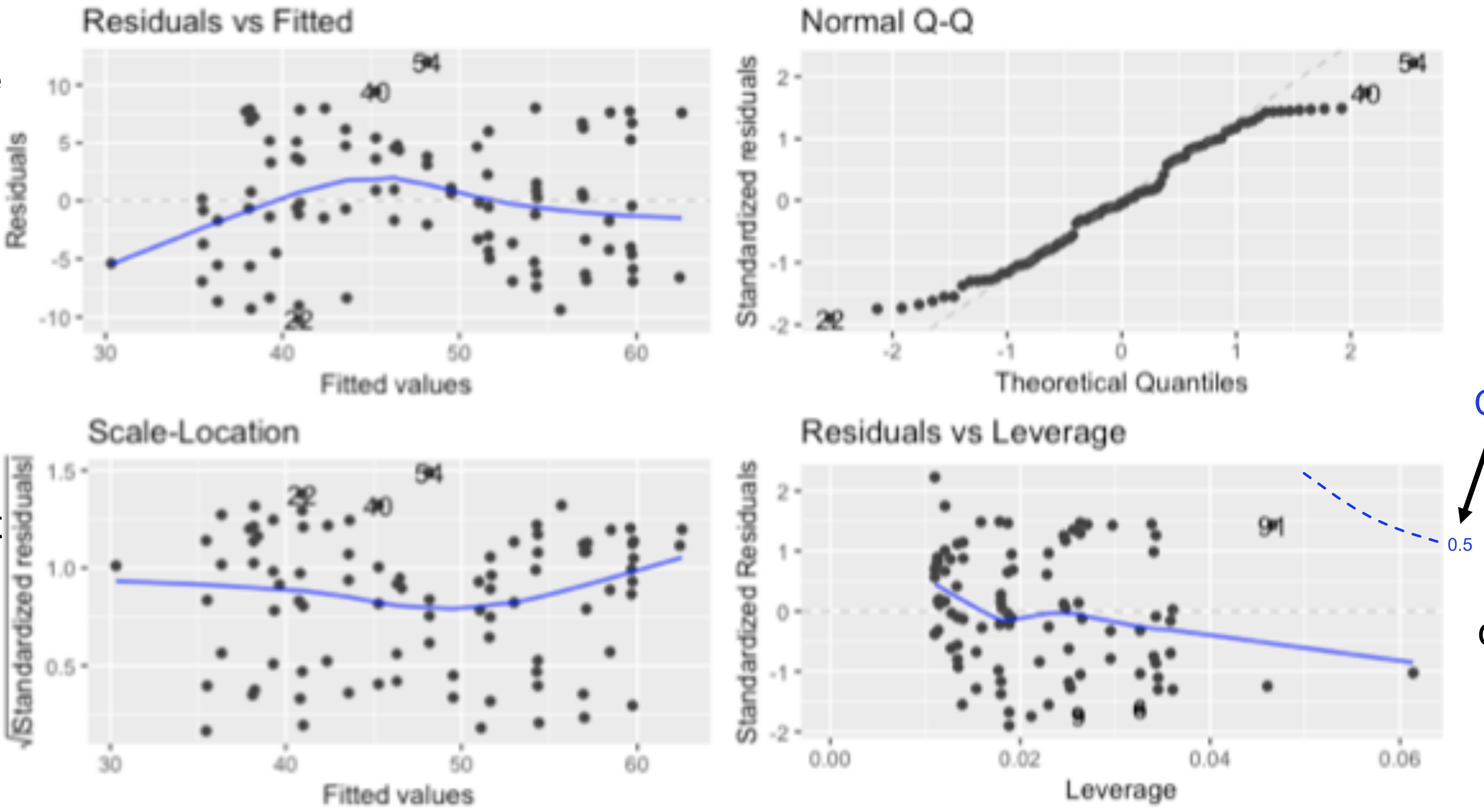
nexus
Personalized Health Technologies

Residual analysis in R: plot(fit)

OK, I admit, you will get something else, but the idea is the same…

In a perfect TA plot the blue smoother would be horizontal at 0 and the points spread in an even band along

The scale-location plot is alright, if it shows a smooth and horizon-tal blue line



The QQ plot is OK, if all the points are more or less on the diagonal

Cook's distance

A leverage plot with no points behind the dashed lines (Cook's distance) is fine

Summary

- Get an intuition for (simple) linear regression
  - Estimate an intercept and a slope to get a line
  - Just like waiting in line at the cashiers
- Parameter estimation
  - MLE and OLS, minimising the squared distance between the line and the points
- Checking the assumptions of a linear regression
  - Diagnostic plots of the structure, the error, and ... well, the error