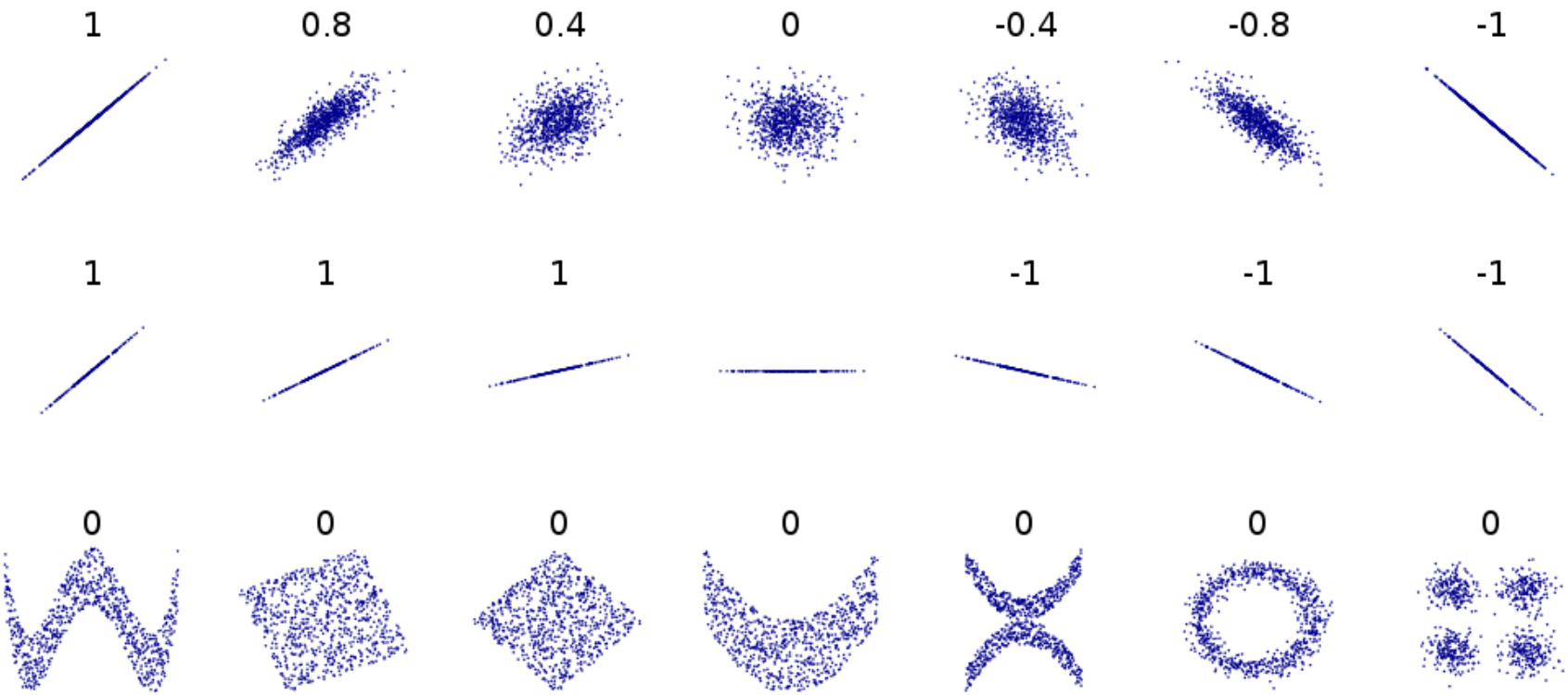
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Einfache lineare Regression

Statistik (Biol./Pharm.) – Herbst 2012

Wdh: Korrelation



Big picture: Generalized Linear Models (GLMs)

- **Bisher:** Population wird mit einer Verteilung beschrieben
  Bsp: Medikament wirkt mit 30% Wa. Wie wa. ist es, dass bei 10 Patienten mindestens 5 gesund werden?
- **Neu:** Population wird mit einer Verteilung beschrieben, die von einem (oder mehreren) Parametern abhängt
  Bsp: Wirkwa. hängt von Dosis ab. Bei welcher Dosis werden im Mittel 90% der Patienten gesund?

- **Generalized Linear Models:** Zshg zw. erklärenden Variablen (z.B. Dosis) und Parametern einer Verteilung (z.B. Erfolgswa. in Binomialverteilung)

Bsp 1: Wirkung von Medikament

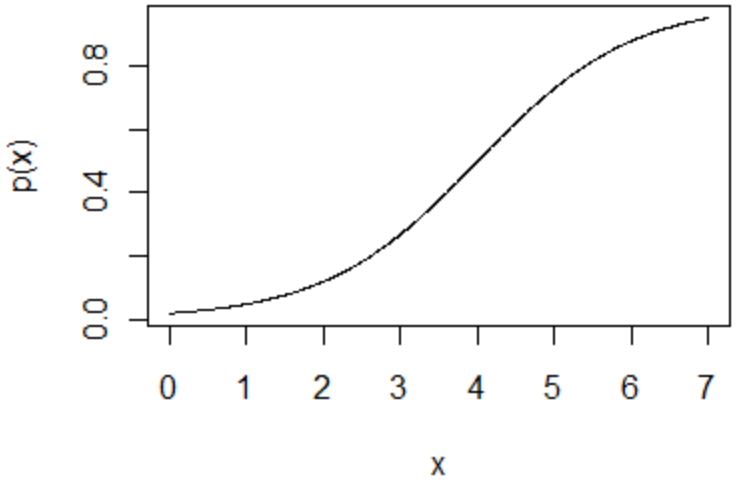
- X: Dosis des Wirkstoffs; n: Patienten, p: Genesungswa.
  Y: Anz. gesunder Patienten nach Behandlung
- $Y \sim Bin(n, p(x))$
- Zshg. zwischen p und x z.B.:

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

- Kann man umformen zu:

$$\underbrace{\log\left(\frac{p(x)}{1 - p(x)}\right)}_{\text{Logistische Funktion}} = \beta_0 + \beta_1 x$$

Linear in β's



"Bei welcher Dosis ist die Genesungswa. 80%?"

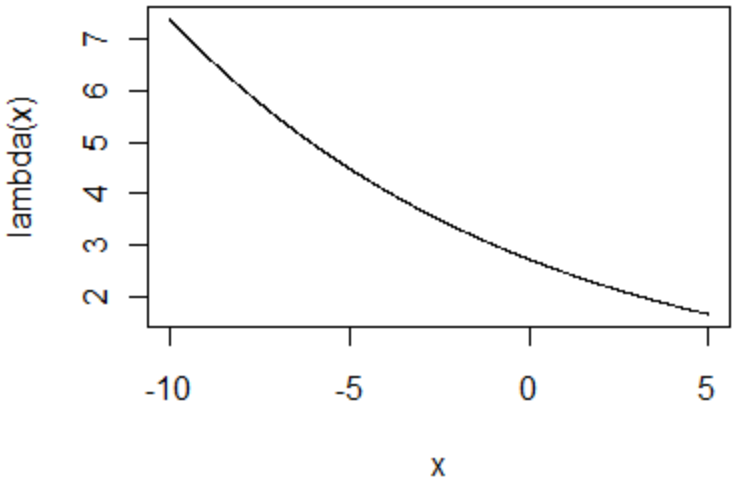
- "Logistische Regression", "Binomialregression"

Bsp 2: Anzahl Autounfälle im Winter

- Y: Anz. Autounfälle pro Tag in ZH
  X: Temperatur in Celsius
- $Y \sim Pois(\lambda(x))$
- Zshg. zw. λ und x z.B.:
  $\lambda(x) = \exp(\beta_0 + \beta_1 x)$
- Kann man umformen zu:
  $\log(\lambda(x)) = \beta_0 + \beta_1 x$

Linear in β's

- "Poissonregression"



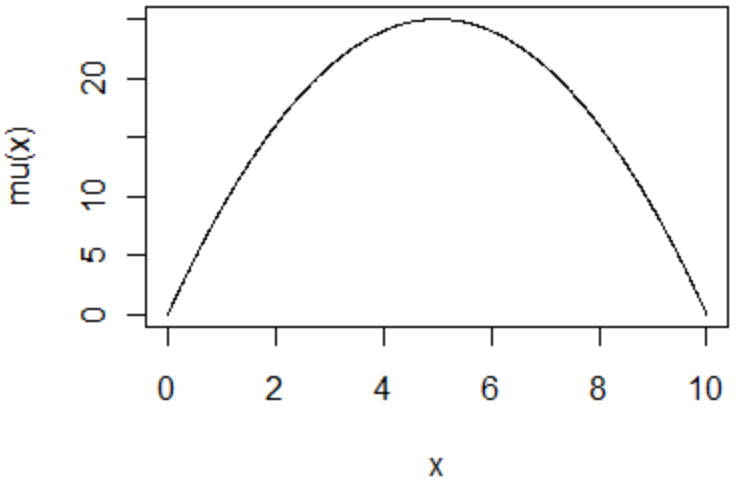
Morgen wird es -5 C.

Was ist das 95%-Quantil der Unfälle morgen?

Bsp 3: Kraftzuwachs bei Training

- Y: Kraftzuwachs nach 6 Wochen Training bei Anfängern
  X: Trainingszeit pro Woche
- $Y \sim N(\mu(x), \sigma^2)$
- Zshg. zw. μ und x z.B.:
  $\mu(x) = \beta_0 + \beta_1 x + \beta_2 x^2$

Linear in β's



Welche Trainingsdauer pro Woche bringt optimalen Kraftzuwachs?

- "Lineare Regression"
- Einfache Lineare Regression
  $\mu(x) = \beta_0 + \beta_1 x$ (eine Erklärende)
- Multiple Lineare Regression
  $\mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + etc$ (mehrere Erklärende)

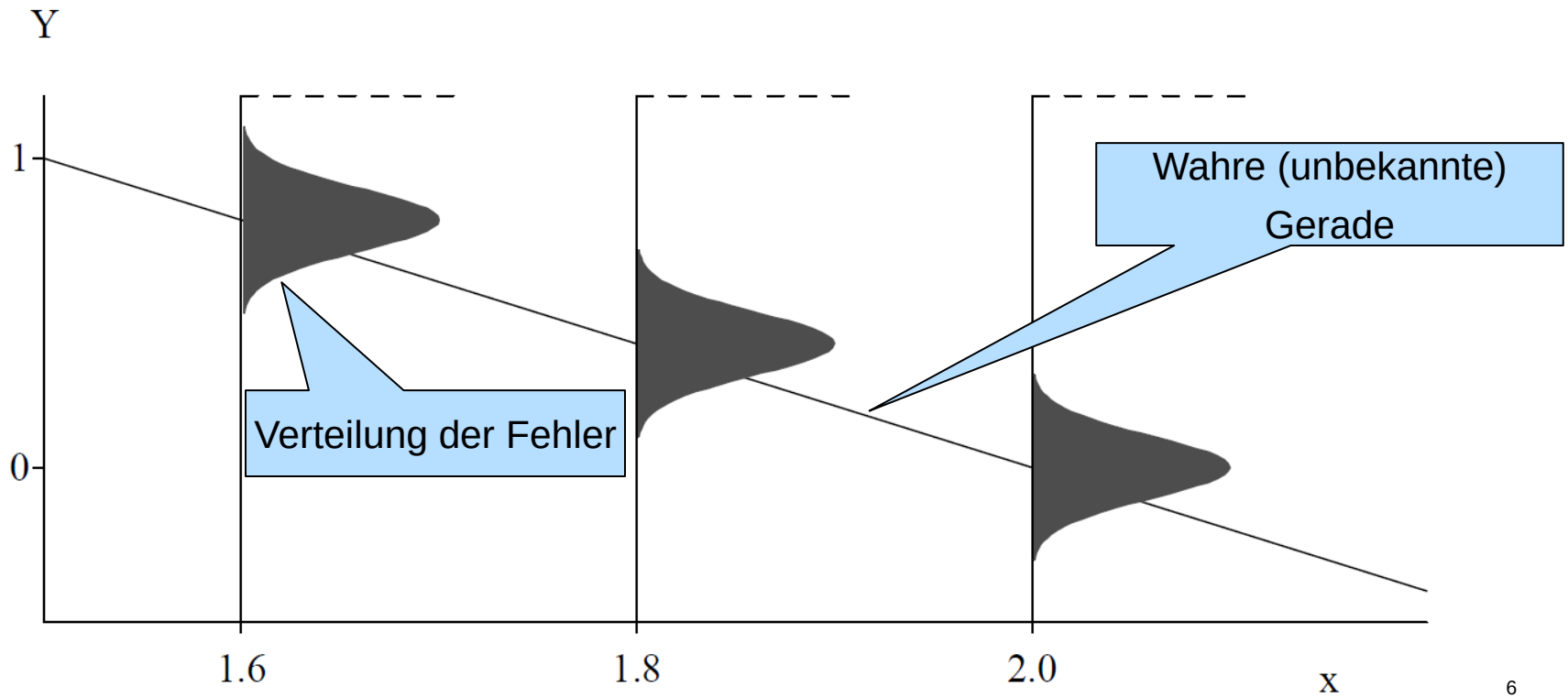
Lineare Regression: Zwei Definitionen

1. $Y \sim N(\mu(x), \sigma^2)$
   $\mu(x) = \beta_0 + \beta_1 x$

Def 1 und Def 2 sind äquivalent

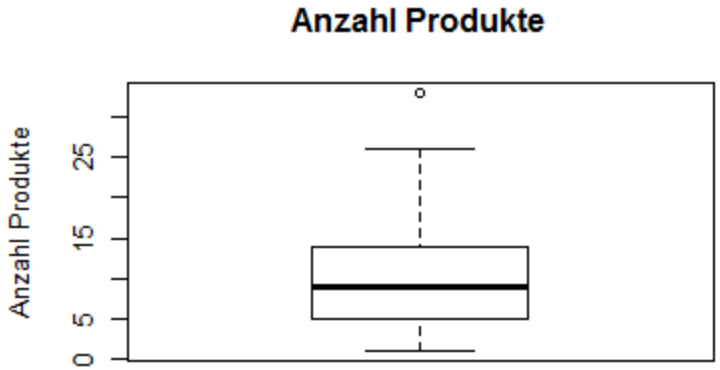
2. $Y = \beta_0 + \beta_1 x + \varepsilon$
   $\varepsilon \sim N(0, \sigma^2)$

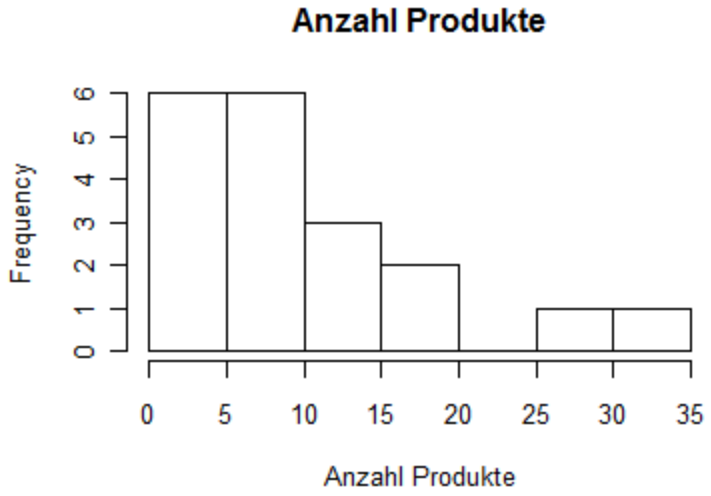
$$E(Y) = E(\beta_0 + \beta_1 x + \varepsilon) = \beta_0 + \beta_1 x + E(\varepsilon) = \beta_0 + \beta_1 x$$
$$Var(Y) = Var(\beta_0 + \beta_1 x + \varepsilon) = Var(\varepsilon) = \sigma^2$$

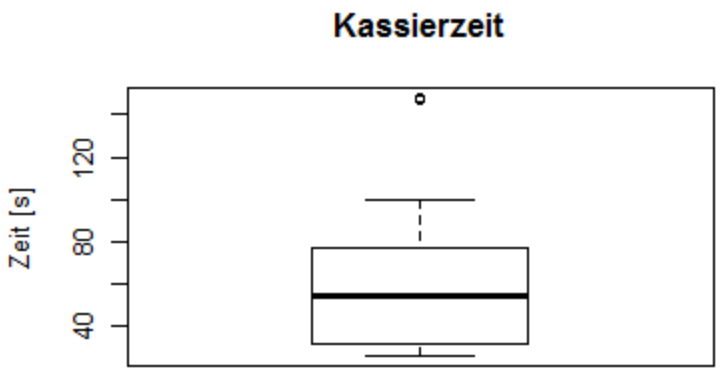


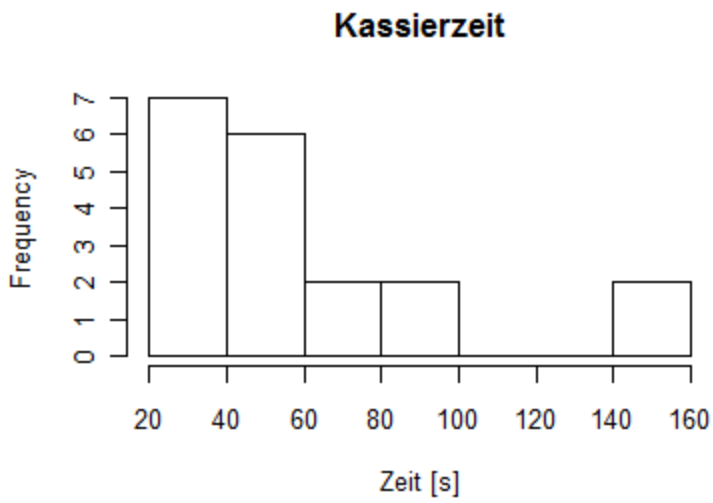
Welche Schlange?

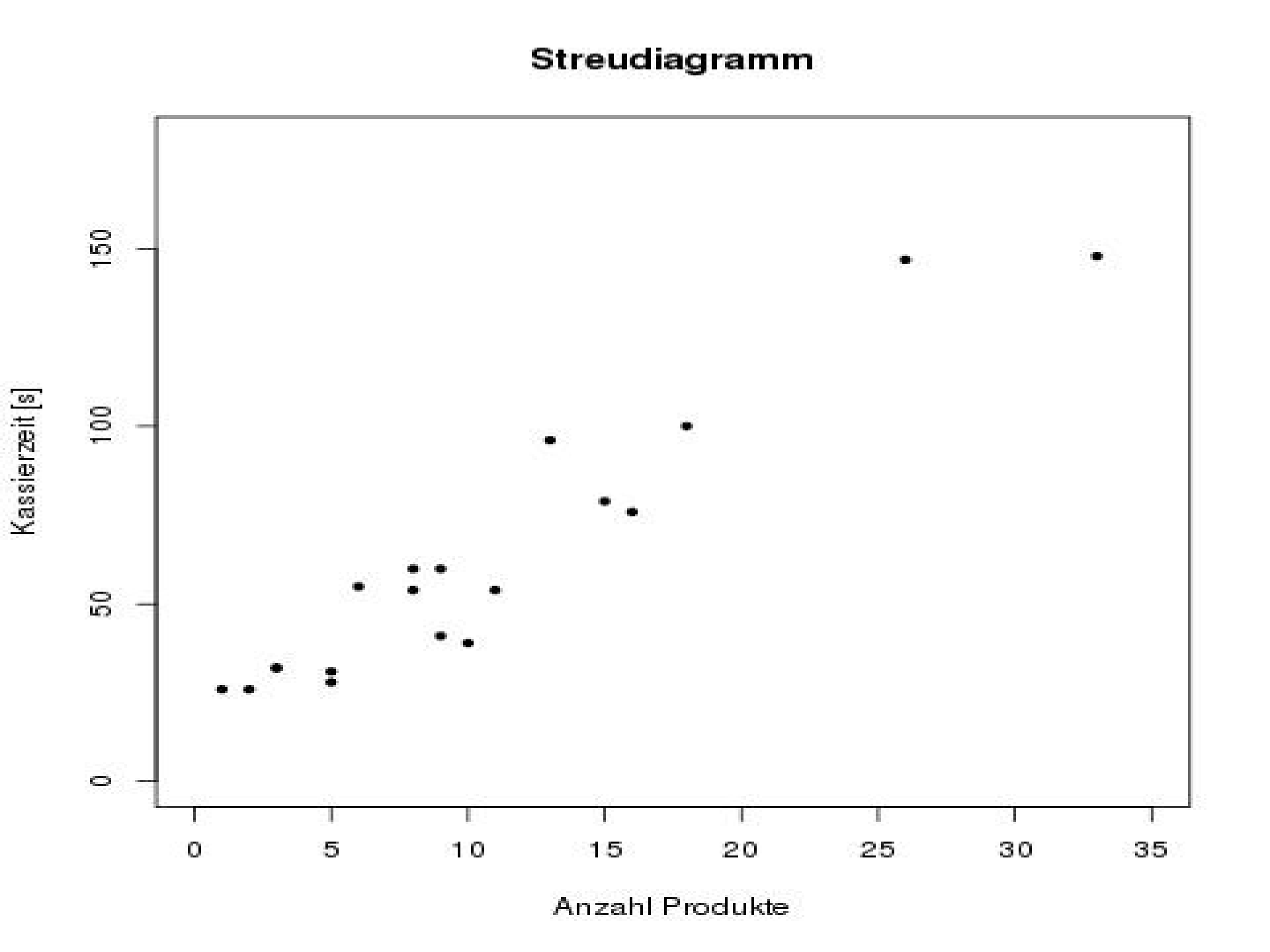
Coop Hauptbahnhof
Di, 22.11.2011, 17:40 – 18:00
(eine Kassiererin)

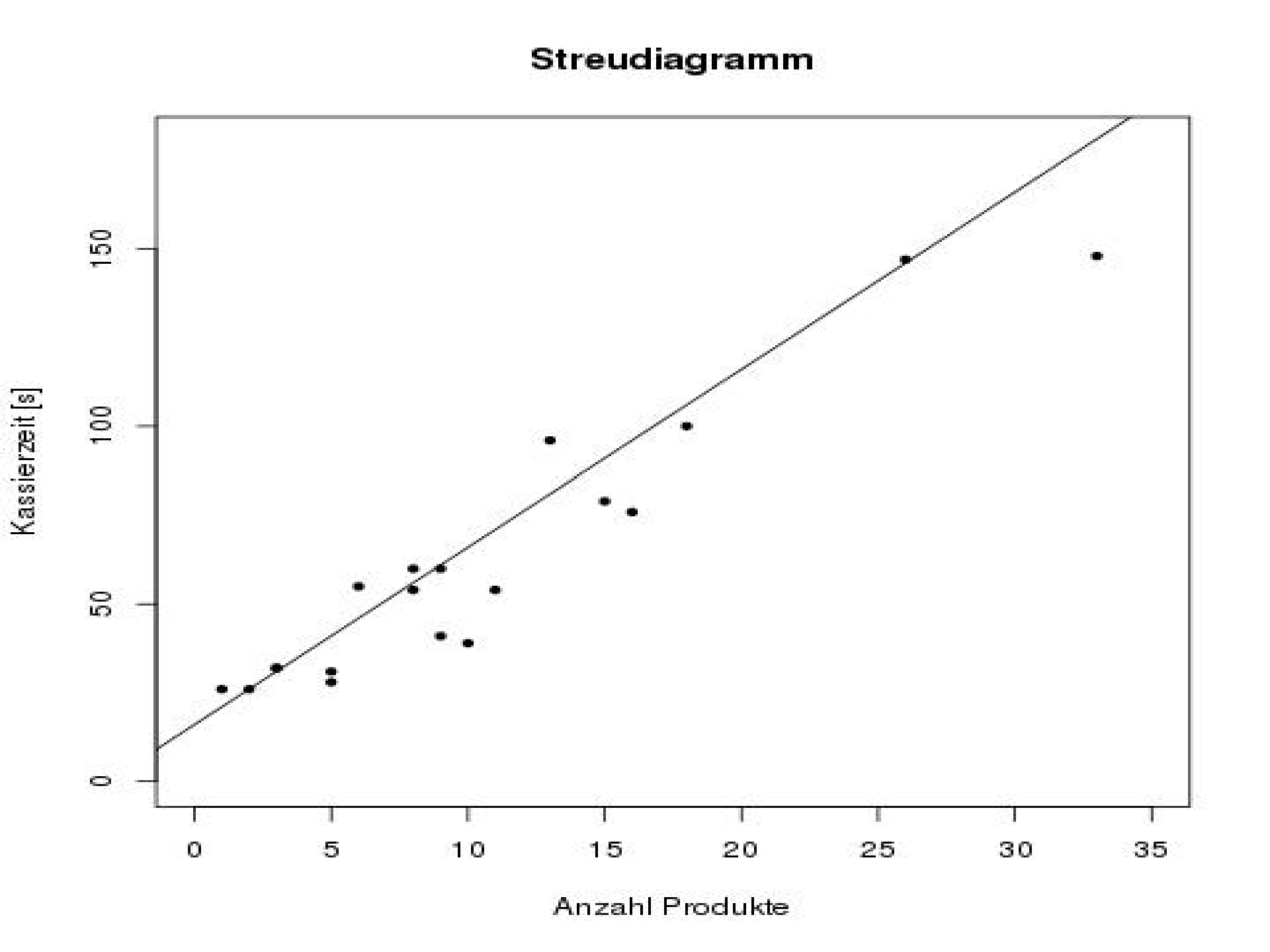


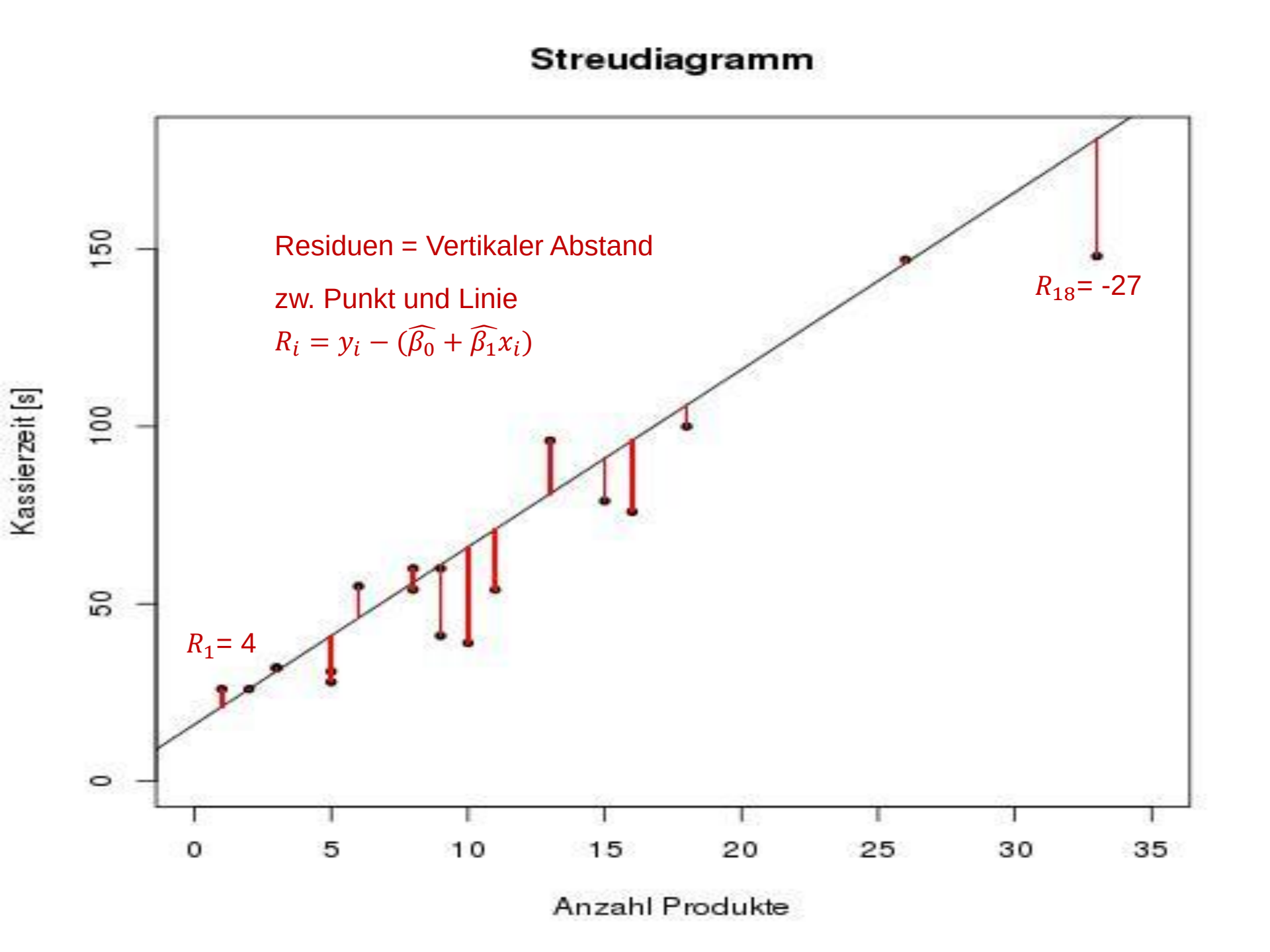







Streudiagramm
Kassierzeit [s]
0
50
100
150
0
5
10
15
20
25
30
35
Anzahl Produkte


Streudiagramm
Kassierzeit [s]
0
50
100
150
0
5
10
15
20
25
30
35
Anzahl Produkte


Streudiagramm
Residuen = Vertikaler Abstand
zw. Punkt und Linie
$R_i = y_i - (\widehat{\beta_0} + \widehat{\beta_1} x_i)$
R_1= 4
R_{18}= -27
Kassierzeit [s]
0
50
100
150
0
5
10
15
20
25
30
35
Anzahl Produkte

Parameterschätzung – Variante 1: Methode der kleinsten Quadrate (“Least Squares”, LS)

- Welche Gerade passt am besten zu den Punkten?
- Wähle $\widehat{\beta_0}, \widehat{\beta_1}$ so, dass Summe der quadrierten Residuen minimal ist:

$$\widehat{\beta_0}, \widehat{\beta_1} \; minimieren \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

Parameterschätzung: Variante 2
Maximum Likelihood Methode (ML)

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

$$= -n\pi\sigma^2 - \frac{1}{2}\frac{\left(\sum_{i=1}^{n}(y_i - \mu(x_i))^2\right)}{\sigma^2} =$$

$$= -n\pi\sigma^2 - \frac{1}{2}\frac{\left(\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2\right)}{\sigma^2}$$

- Log-Likelihood ist maximal, wenn $\sum_{i=1}^{n}(x_i - \beta_0 - \beta_1 x_i)$ minimal ist.

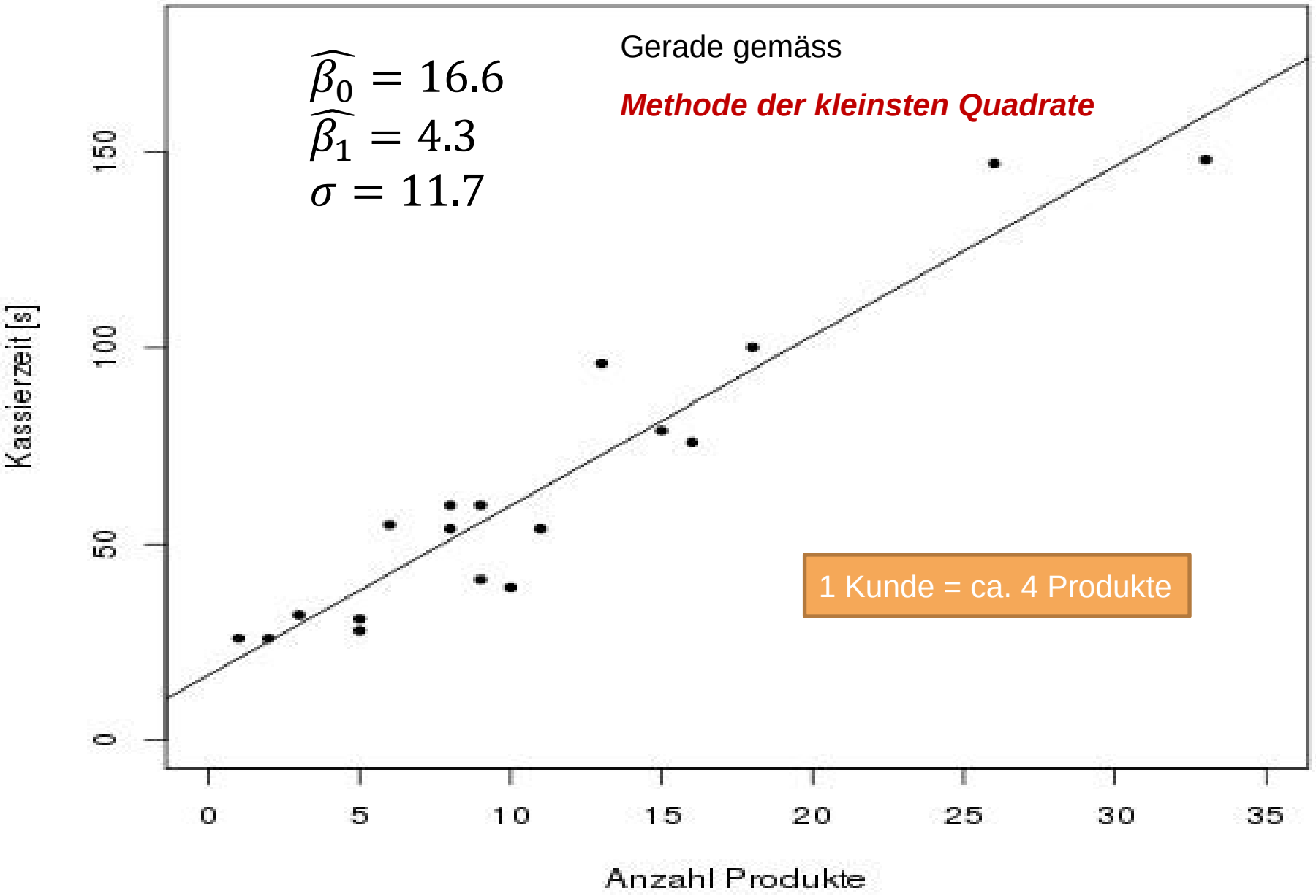
Methode der kleinsten Quadrate

und

Maximum Likelihood Methode

sind äquivalent !

Streudiagramm



Welche Schlange?

Kasse 1

3 3 + 4 = 7

2 2 + 4 = 6

3 3 + 4 = 7

20

Kasse 2

19 19 + 4 = 23

23

Aerobe Leistungsfähigkeit

VO2max: Menge Sauerstoff, die der Körper pro kg maximal pro Minute verwerten kann

- Teuer, aufwändig
- Nicht für breite Masse geeignet

Ersatz: Cooper & Shuttle

- 12-Minuten Test nach Cooper (1968)
- 20m-Shuttle-Test nach Leger (1983)

Eur J Appl Physiol (1982) 49: 1–12

European Journal of
Applied
Physiology
and Occupational Physiology

A Maximal Multistage 20-m Shuttle Run Test to Predict $\dot{V}O_2$ max*

Luc A. Léger[1] and J. Lambert[2]

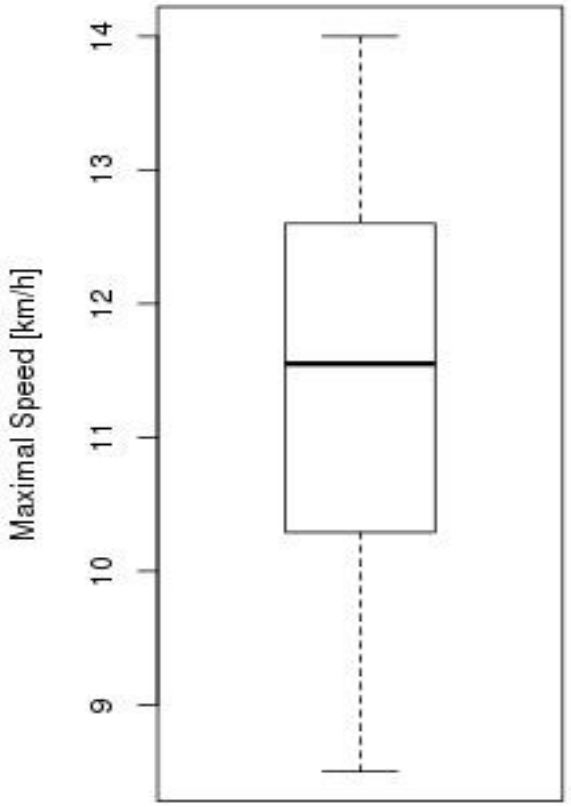
[1] Département d'éducation physique, Université de Montréal, CEPSUM, C.P. 6128, Succ. "A", Montréal (Québec), Canada, H3C 3J7
[2] Département de Médecine sociale et préventive, Université de Montréal, Canada

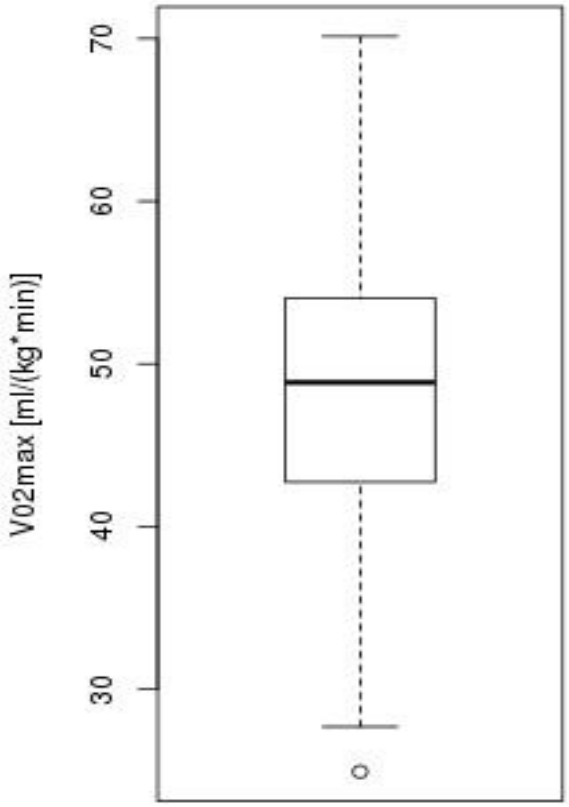
Ersatz: Cooper & Shuttle

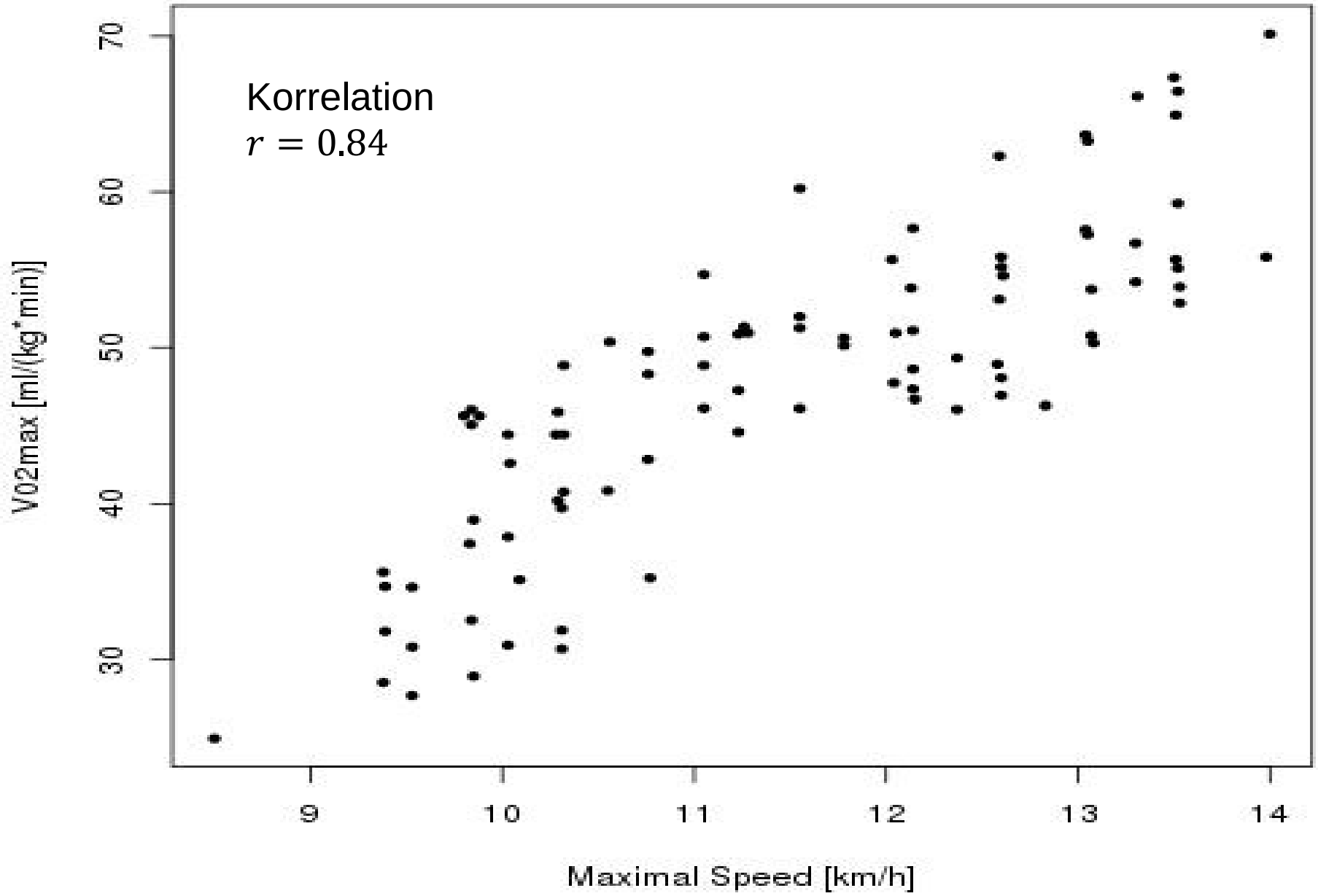
- 12-Minuten Test nach Cooper (1968)
- 20m-Shuttle-Test nach Leger (1983)

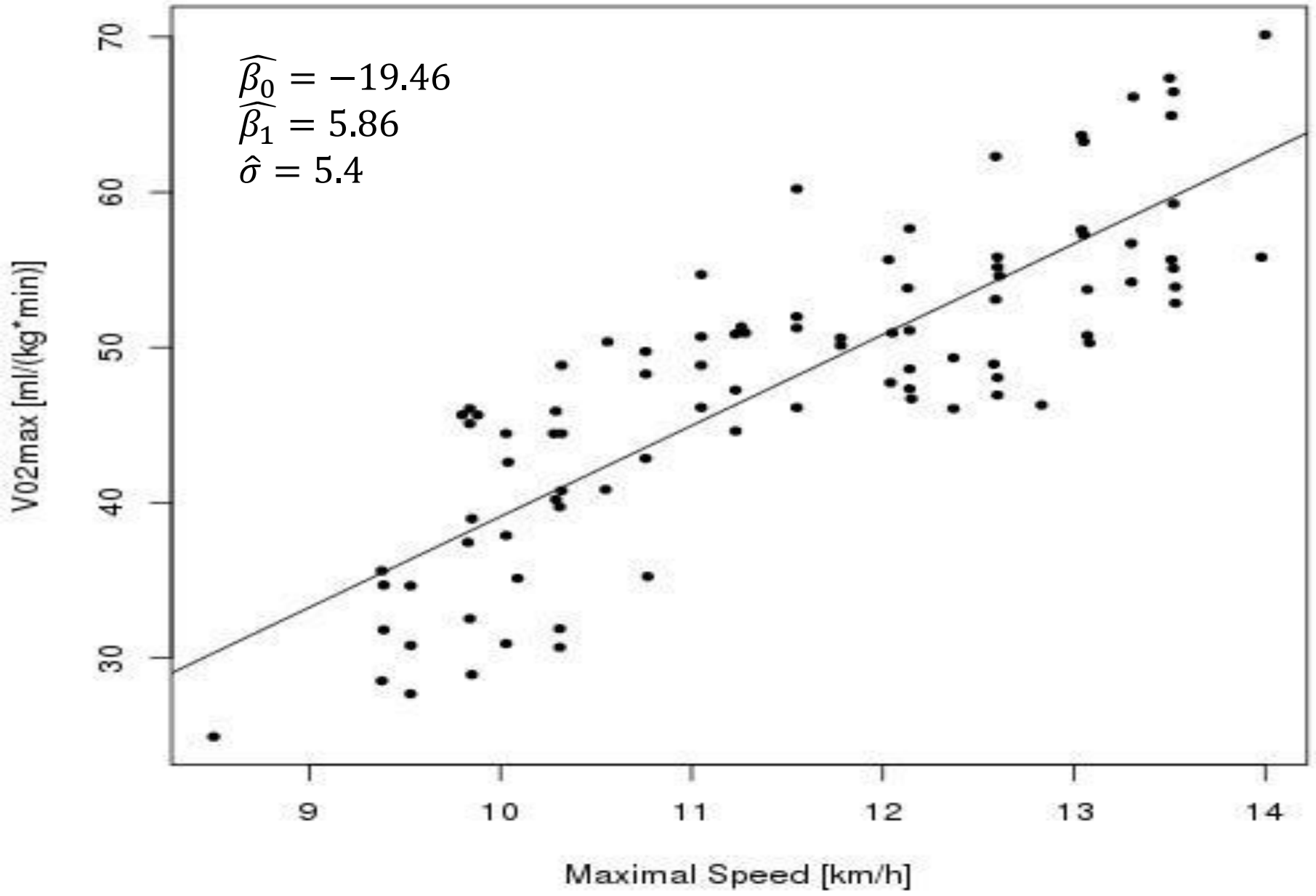
- Kann Shuttle-Test den VO2max-Wert vorhersagen?
- Falls ja: Einfache Testmöglichkeit für breite Bevölkerung

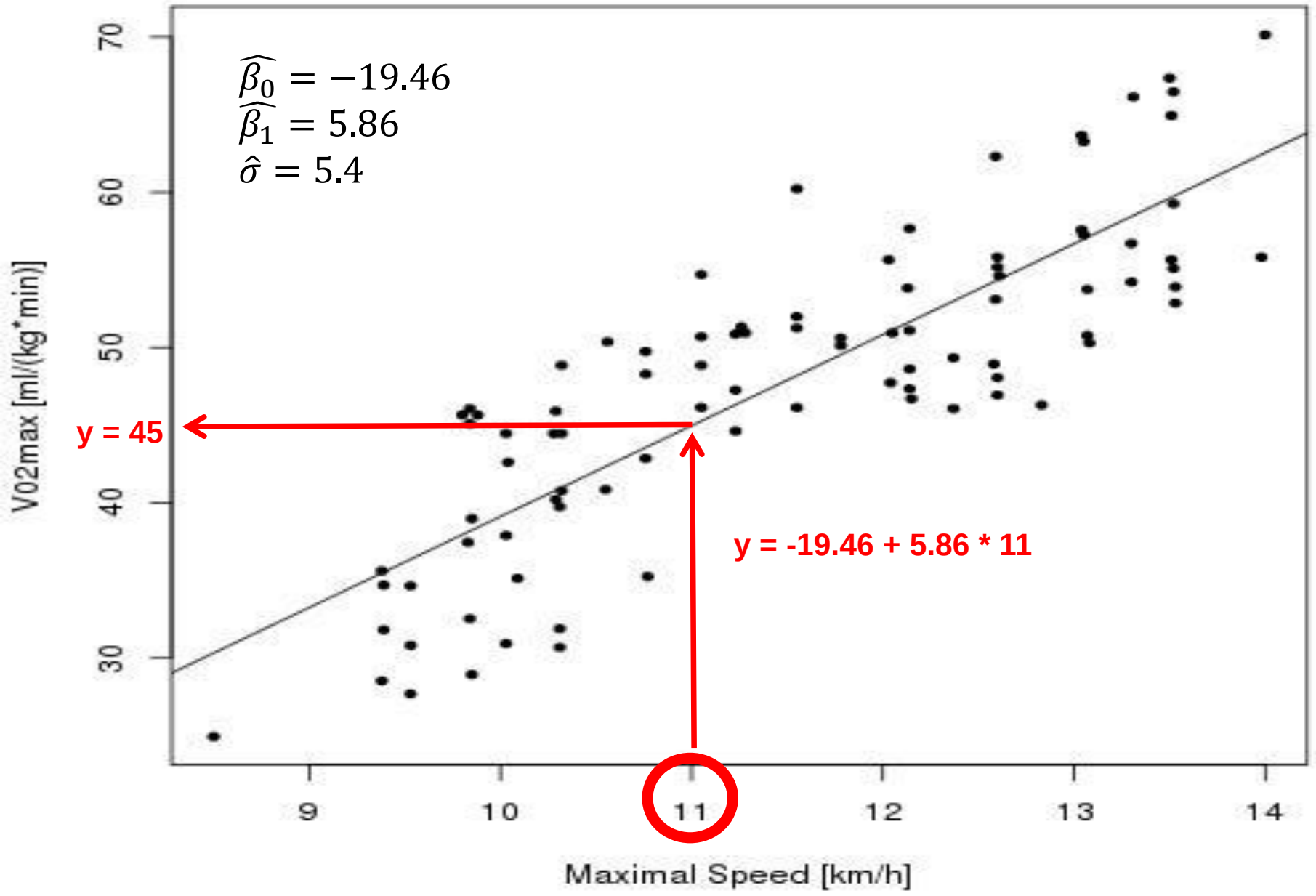
Leger et. al., 1983:
91 Personen
Shuttle test & VO2max



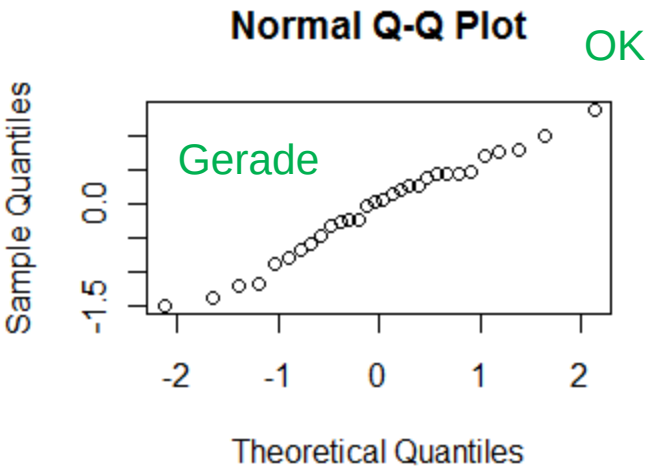


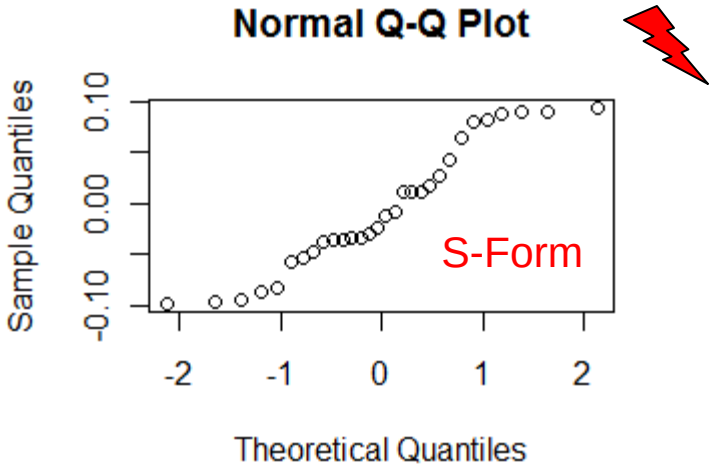

Korrelation
$r = 0.84$
V02max [ml/(kg*min)]
30
40
50
60
70
9
10
11
12
13
14
Maximal Speed [km/h]

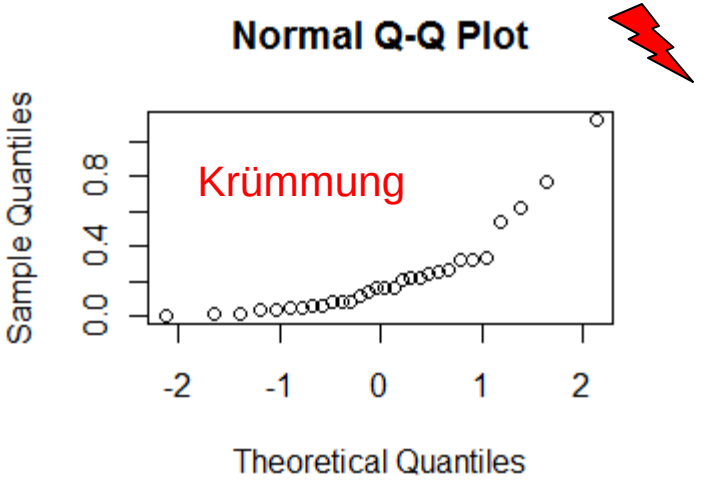

$\widehat{\beta_0} = -19.46$
$\widehat{\beta_1} = 5.86$
$\hat{\sigma} = 5.4$
V02max [ml/(kg*min)]
30
40
50
60
70
9
10
11
12
13
14
Maximal Speed [km/h]


$\widehat{\beta_0} = -19.46$
$\widehat{\beta_1} = 5.86$
$\hat{\sigma} = 5.4$
y = 45
y = -19.46 + 5.86 * 11
V02max [ml/(kg*min)]
30
40
50
60
70
9
10
11
12
13
14
Maximal Speed [km/h]

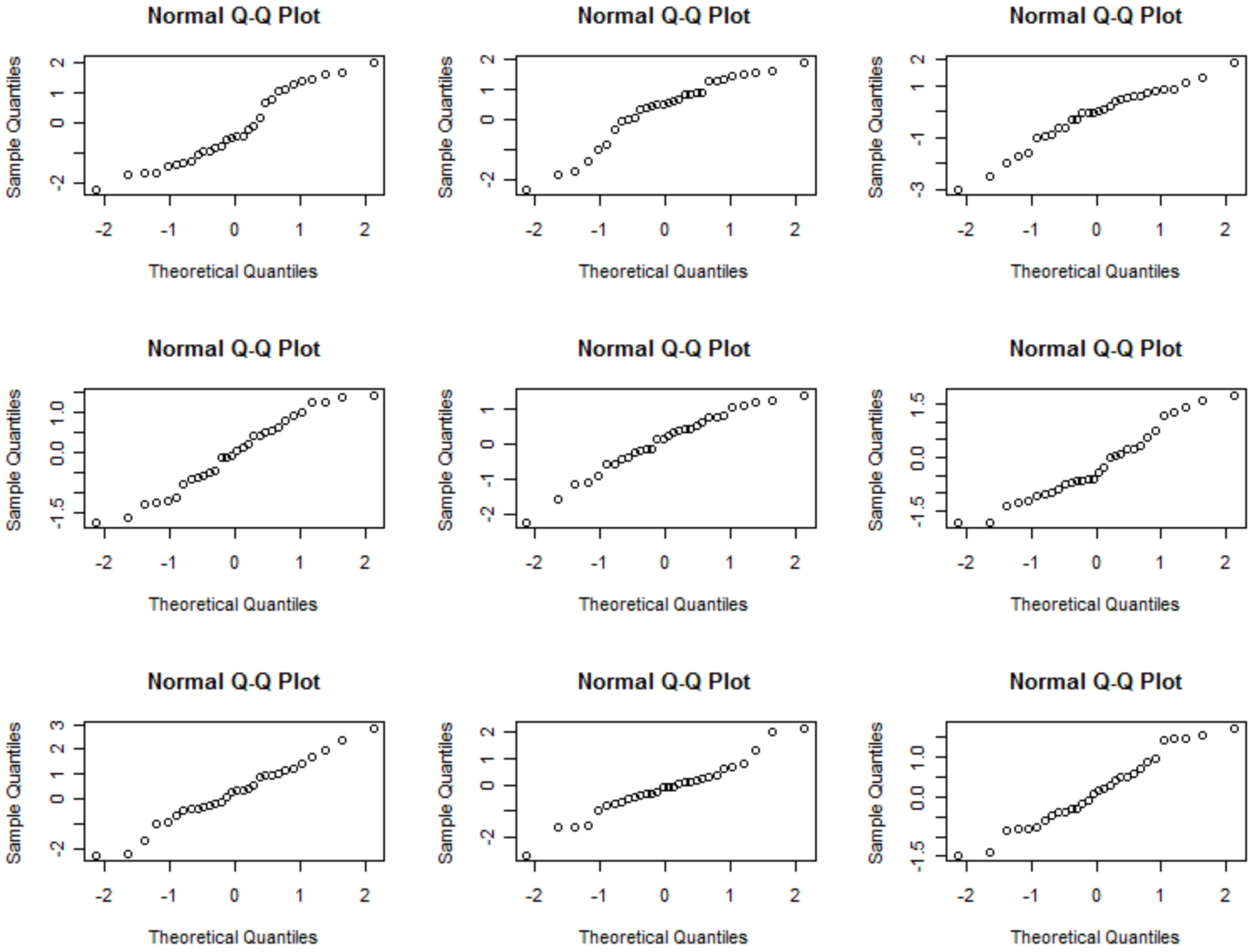
Residuenanalyse: QQ-Plot



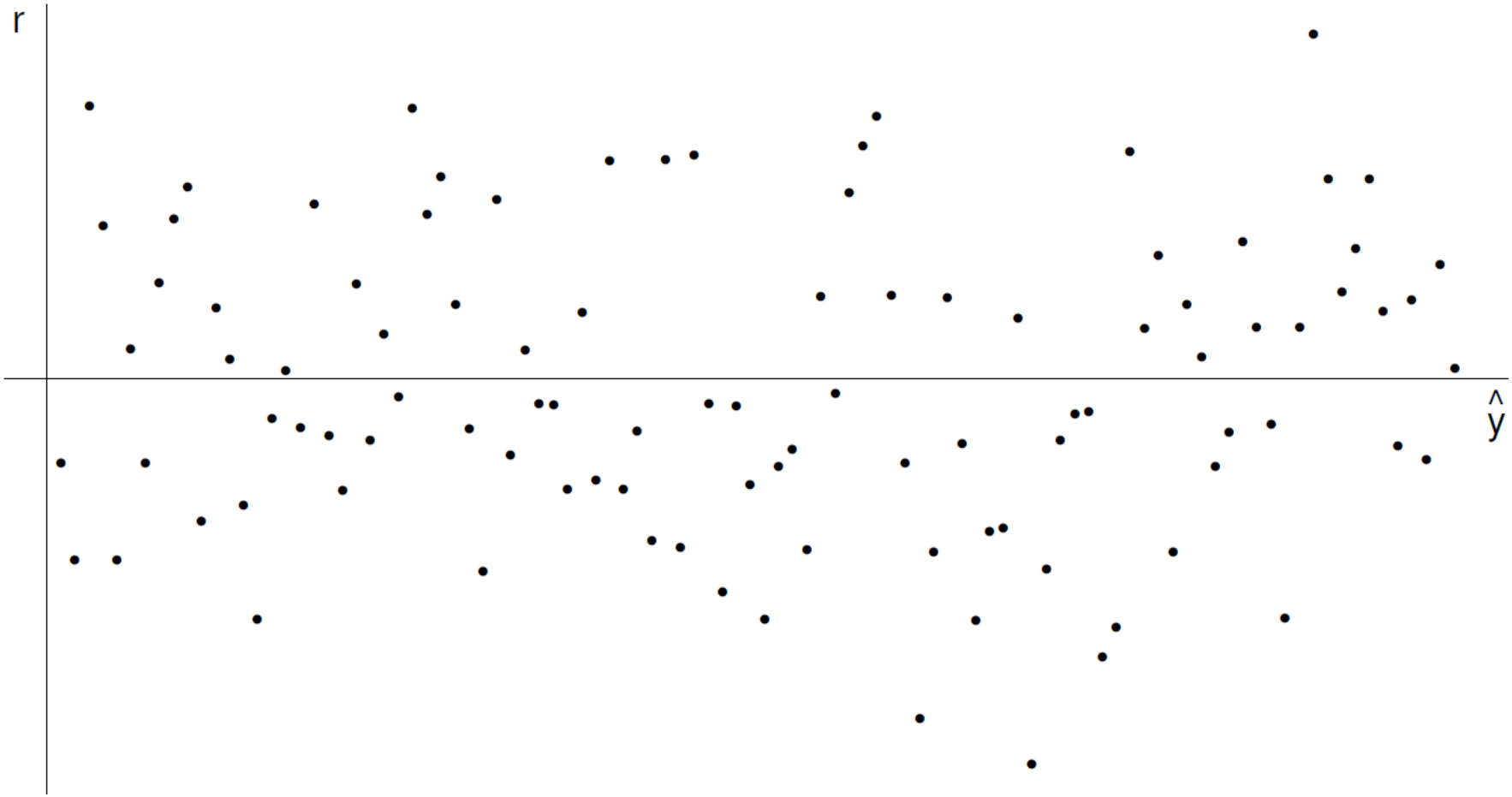




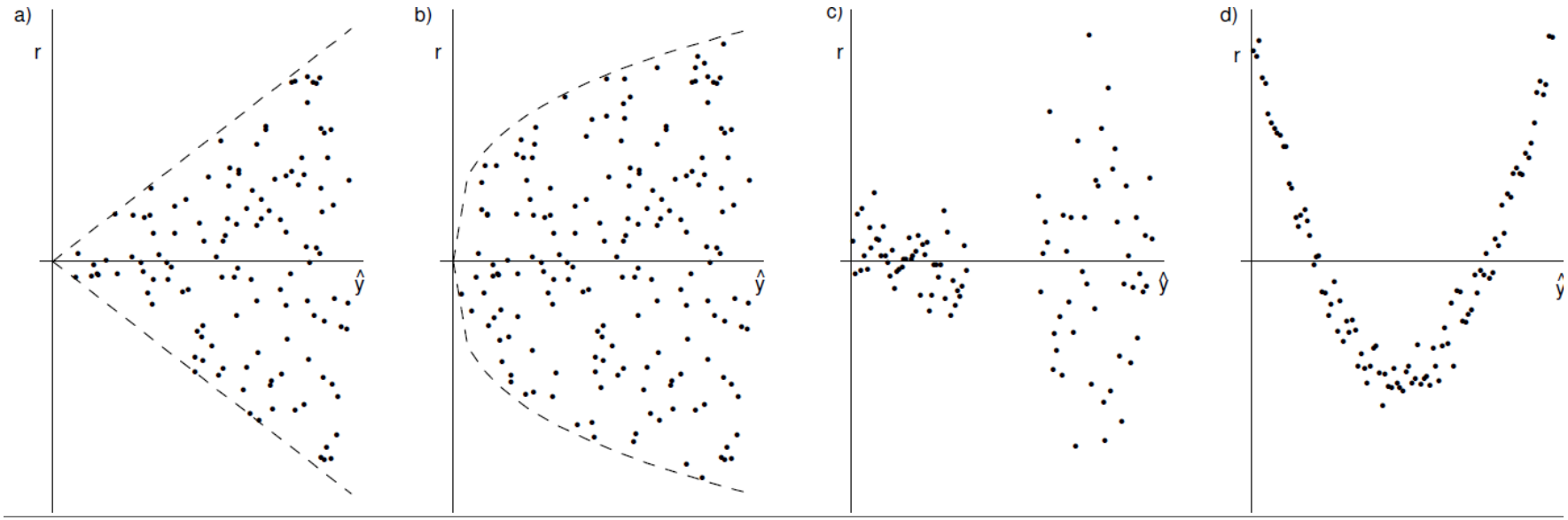
QQ-Plots: Streuung von “guten” QQ-Plots ($n = 30, R_i \sim N(0,1)$)



Beispiel für guten Tukey-Anscombe Plot



Beispiele für schlechte Tukey-Anscombe Plots



Falls Residuenplots schlecht

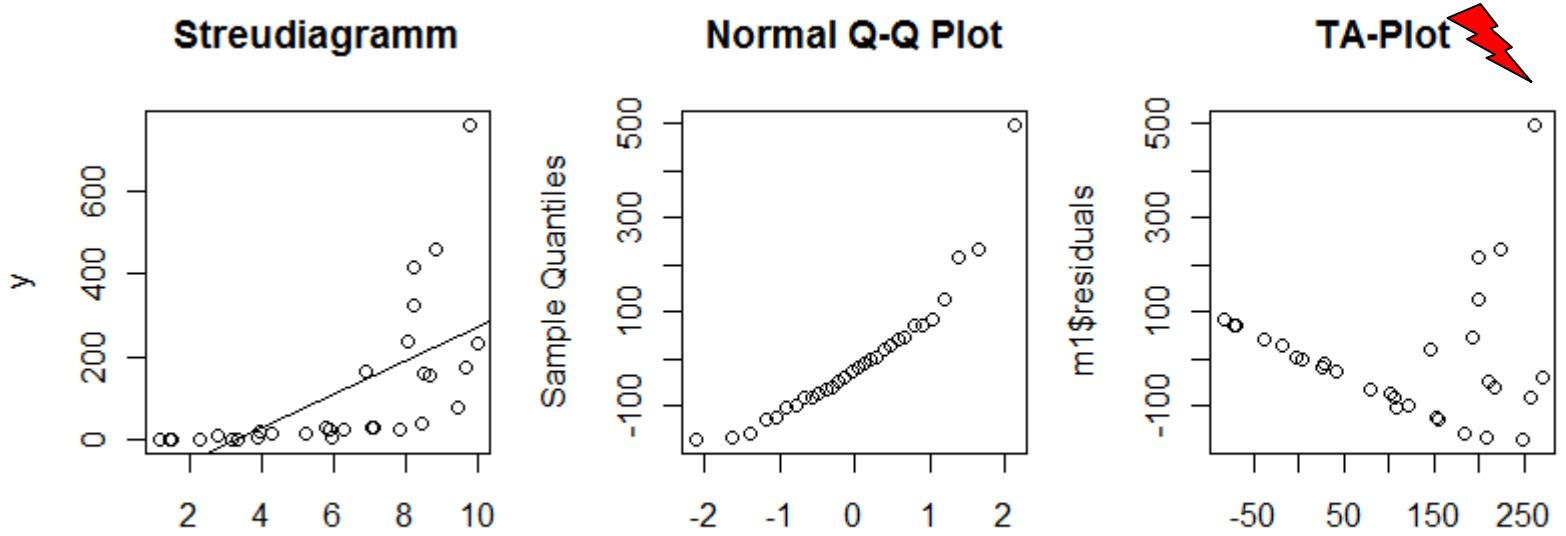
- Oft helfen Transformationen von x oder y
- Achtung: Vorsicht beim Interpretieren der neuen Parameter

- Bsp: $\log(y)$ statt y

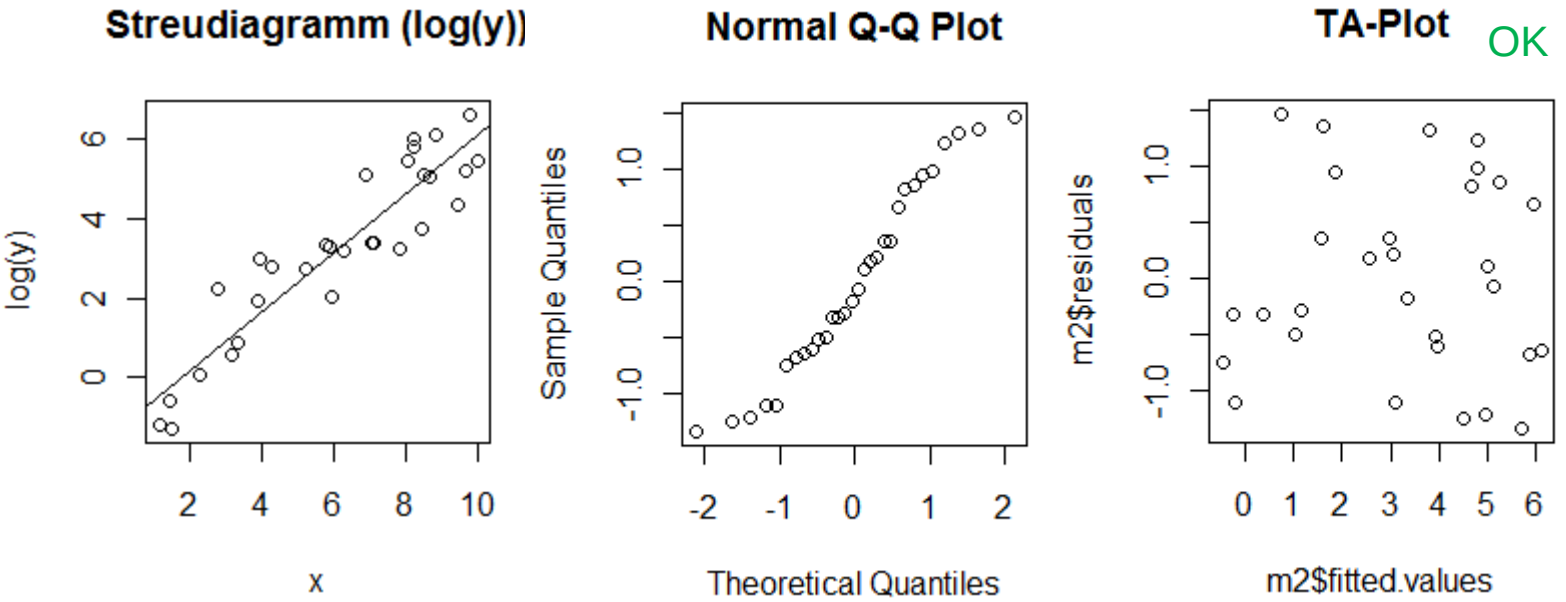
  Vorher: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
  Wenn x durch x+1 ersetzt wird, ändert sich Y im Mittel zu $Y + \beta_1$

  Nachher:
  $\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i \leftrightarrow Y_i = \exp(\beta_0 + \beta_1 x_i + \varepsilon_i)$
  Wenn x durch x+1 erstetzt wird, ändert sich Y “im Mittel” zu $Y * \exp(\beta_1)$

Bsp: Ohne Log-Transformation



y

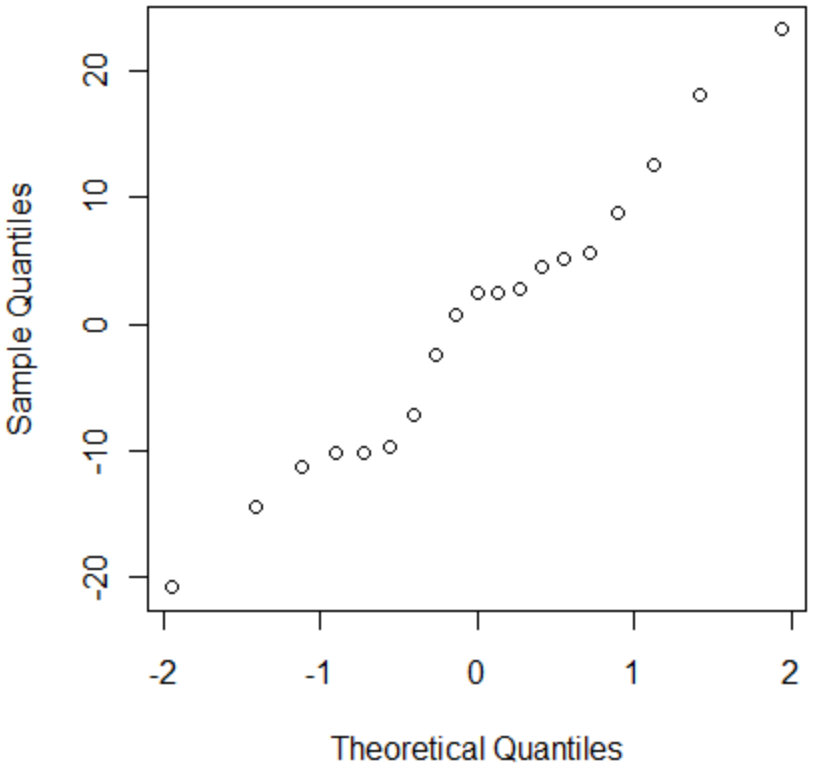


OK

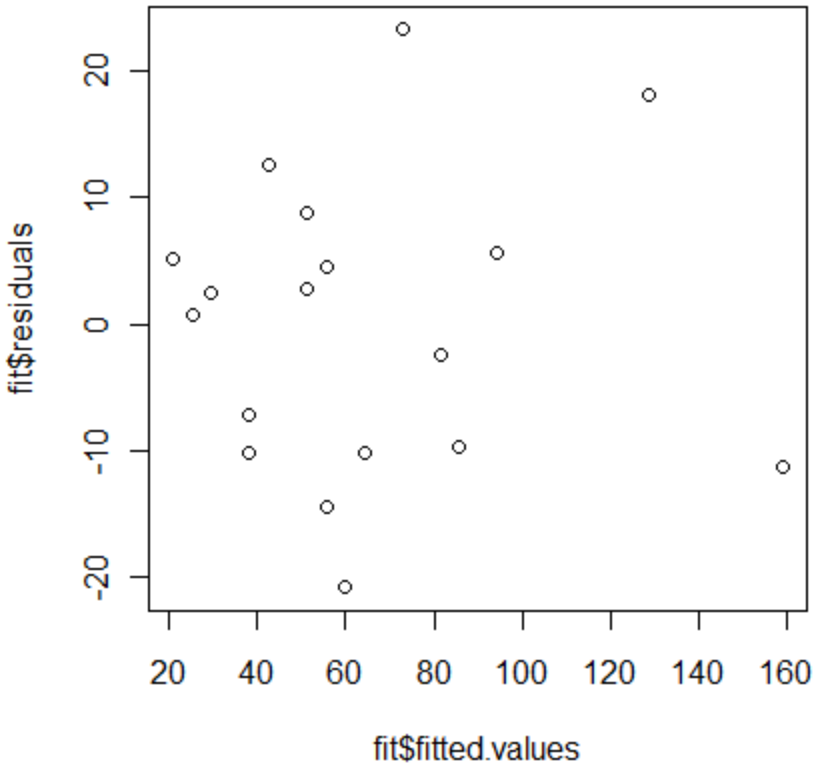
$\log(y)$

Residuenanalyse: Supermarkt

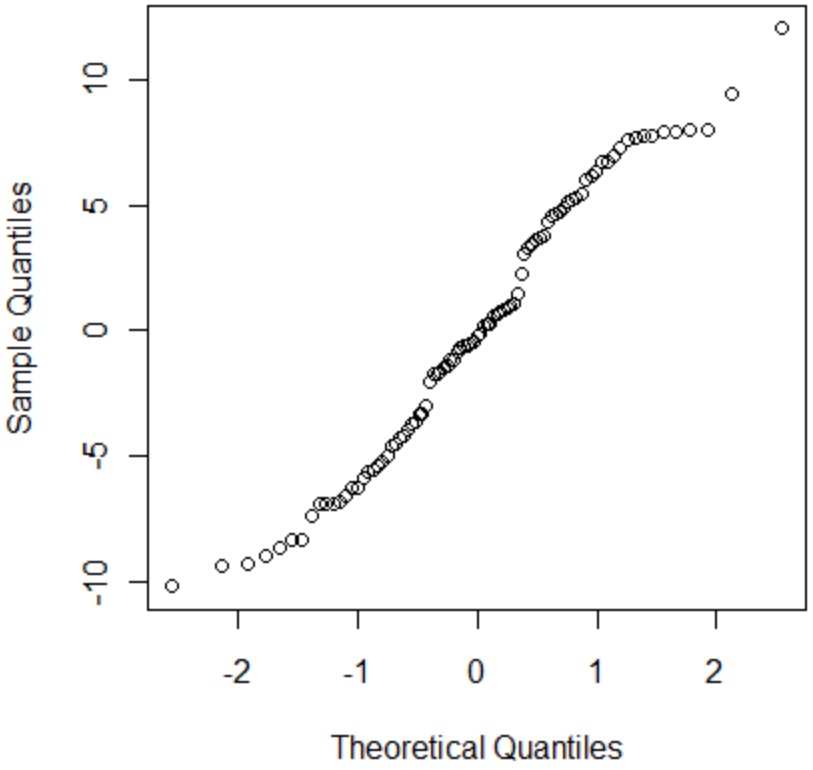
OK



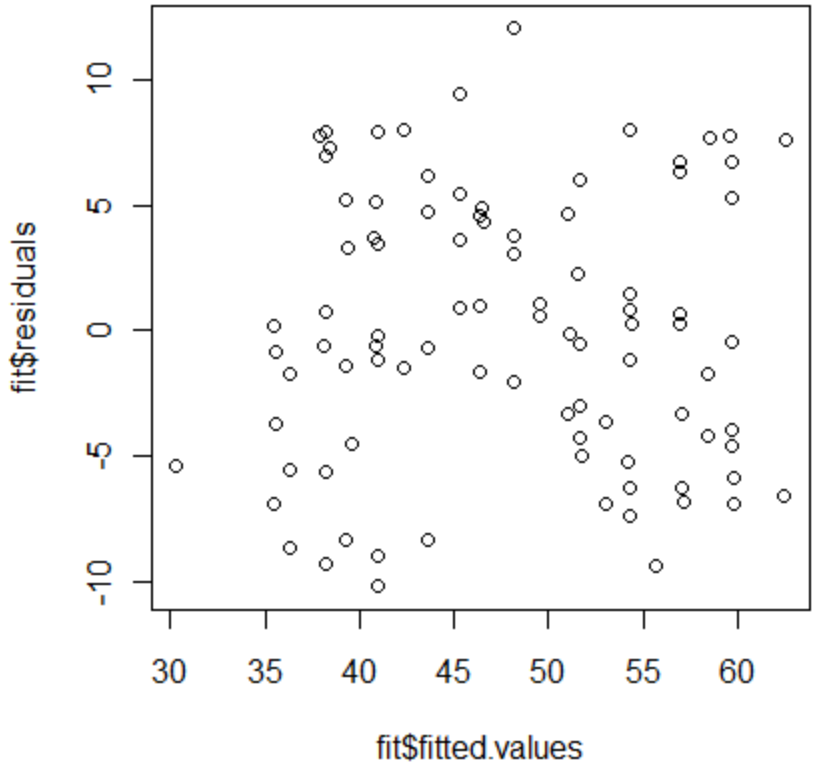
OK



Residuenanalyse: Beep-Test



OK



OK