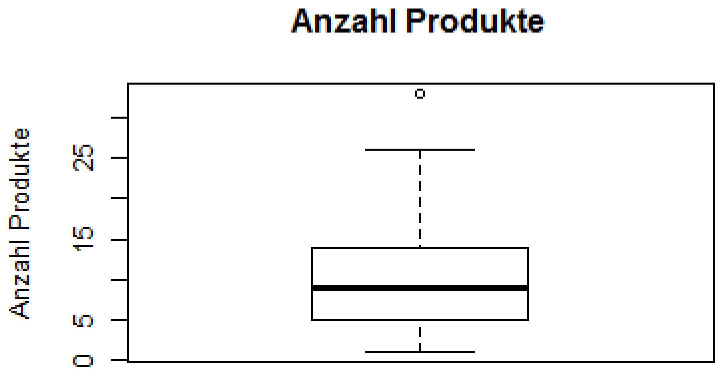
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

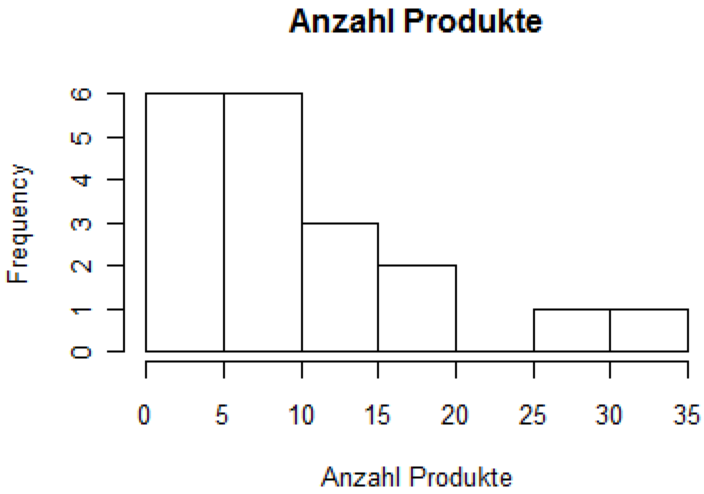
Einfache lineare Regression

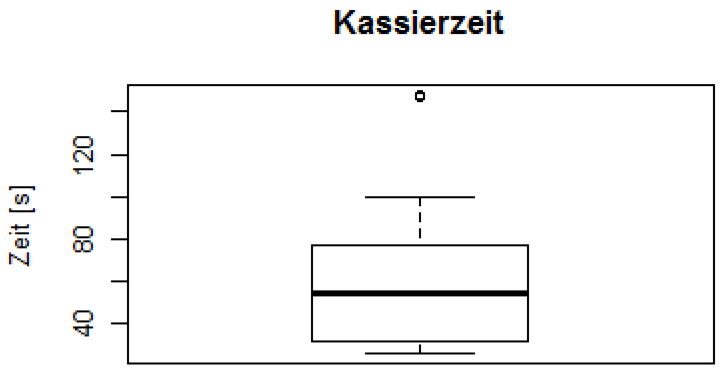
Statistik (Biol./Pharm./HST) – FS 2017

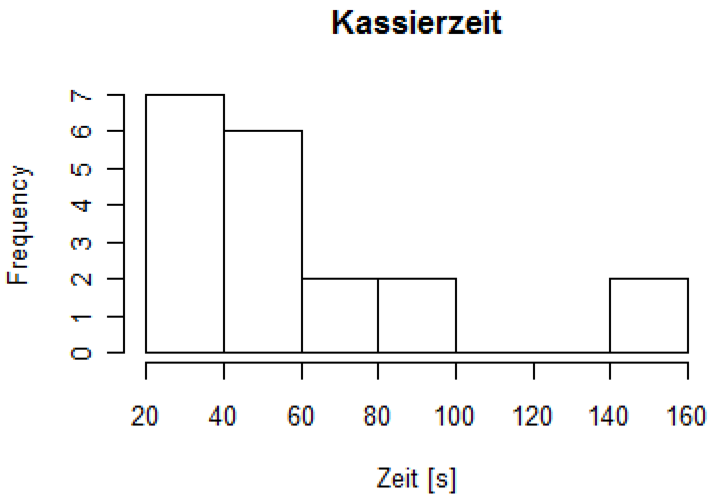
Welche Schlange?

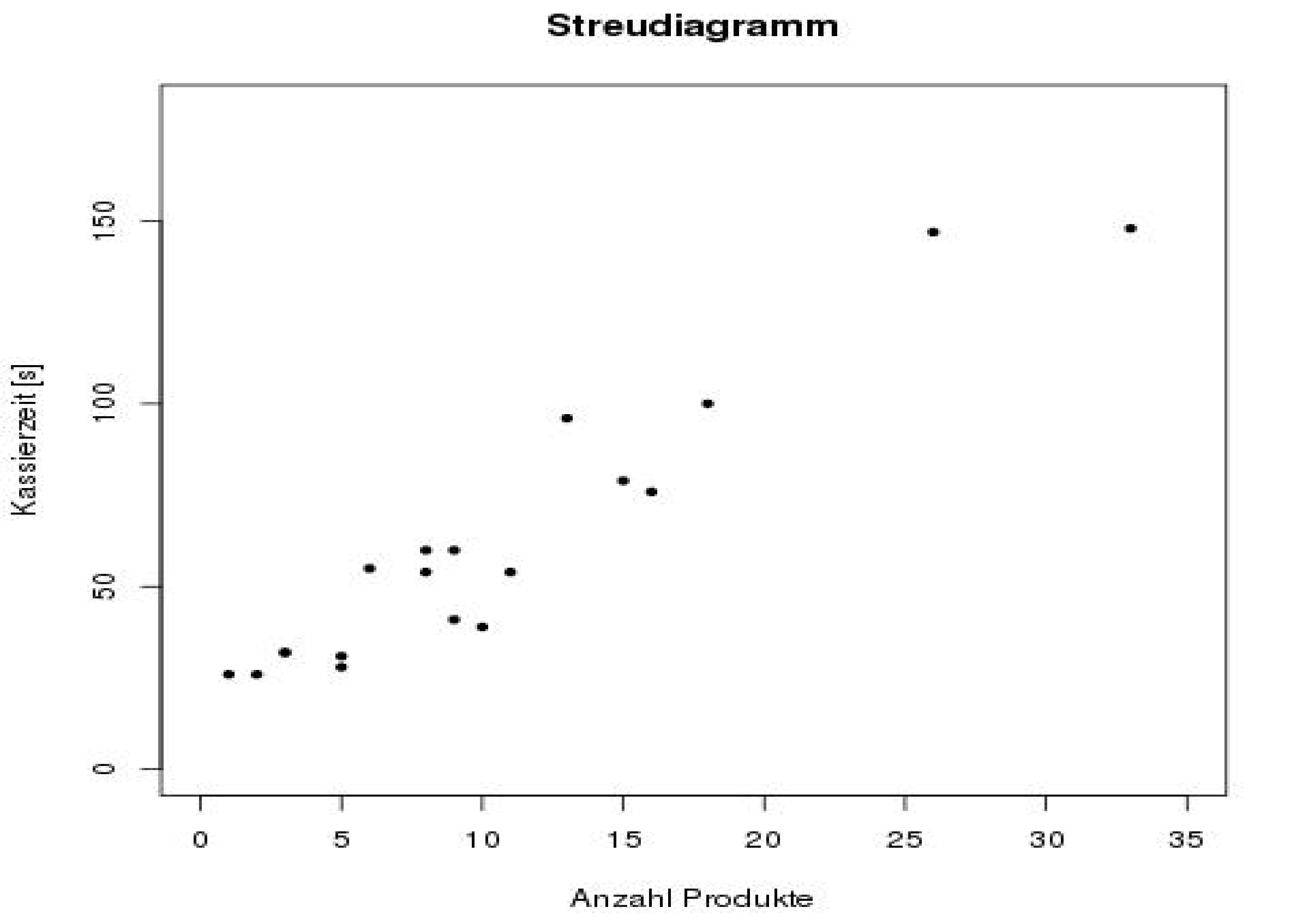
Coop Hauptbahnhof
Di, 22.11.2011, 17:40 – 18:00
(eine Kassiererin)



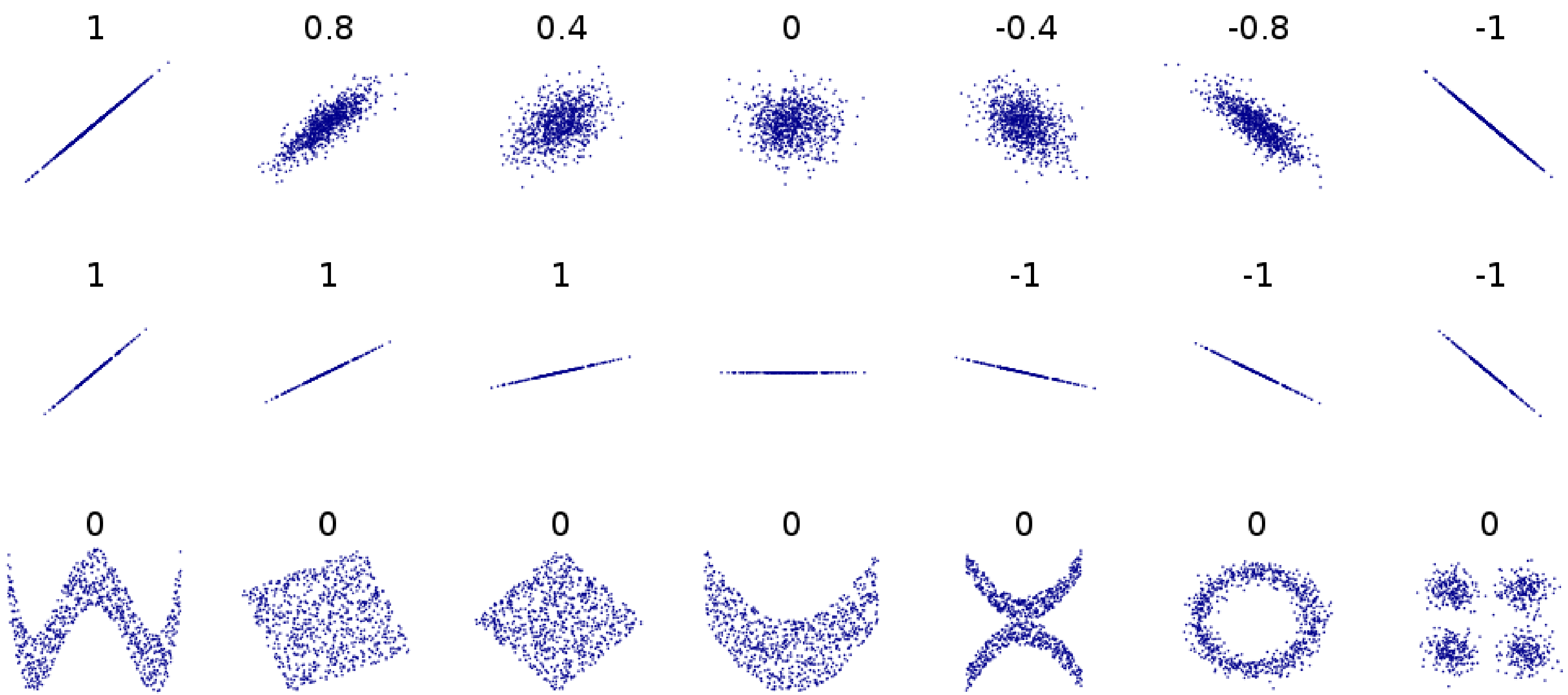


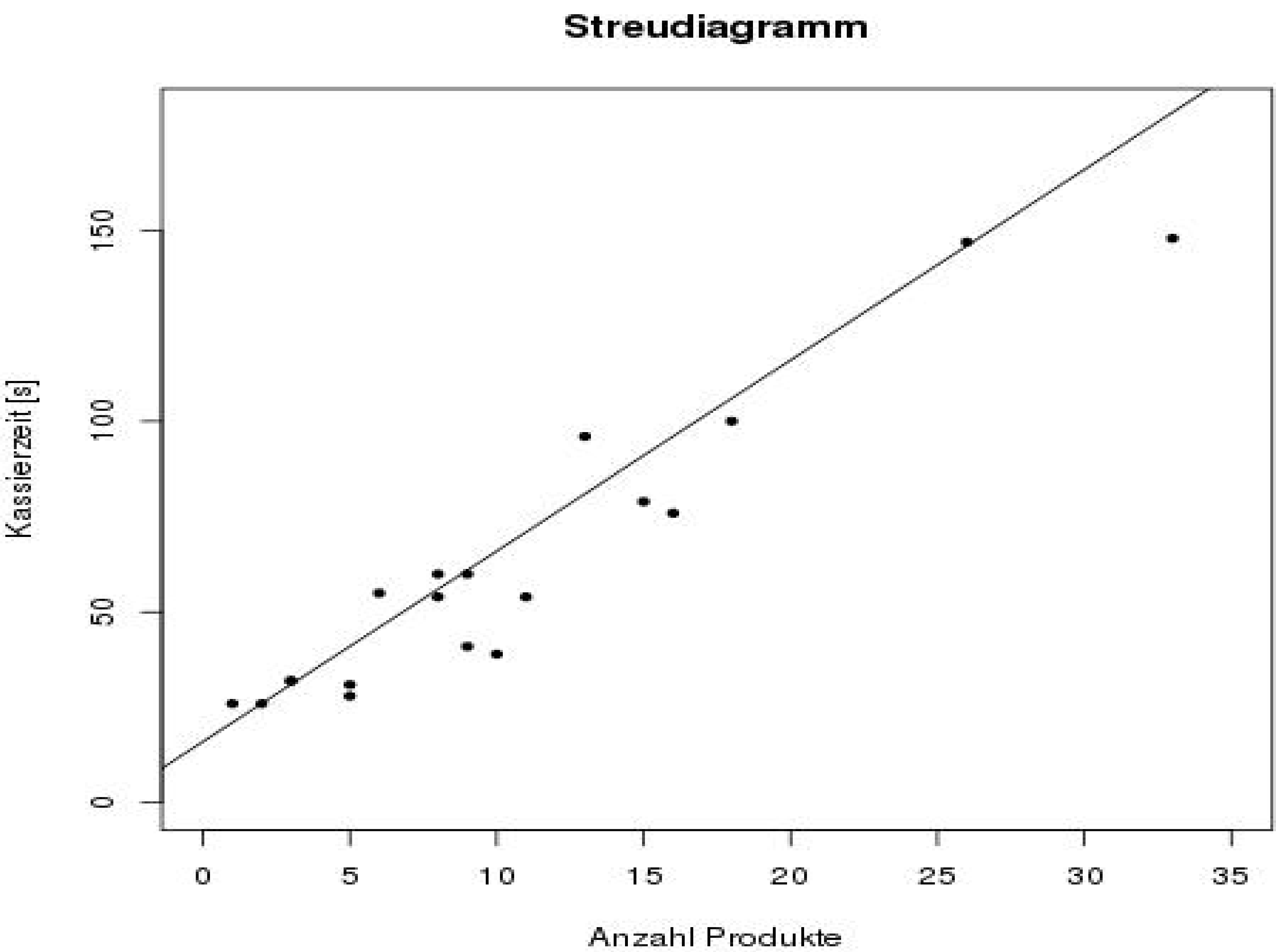


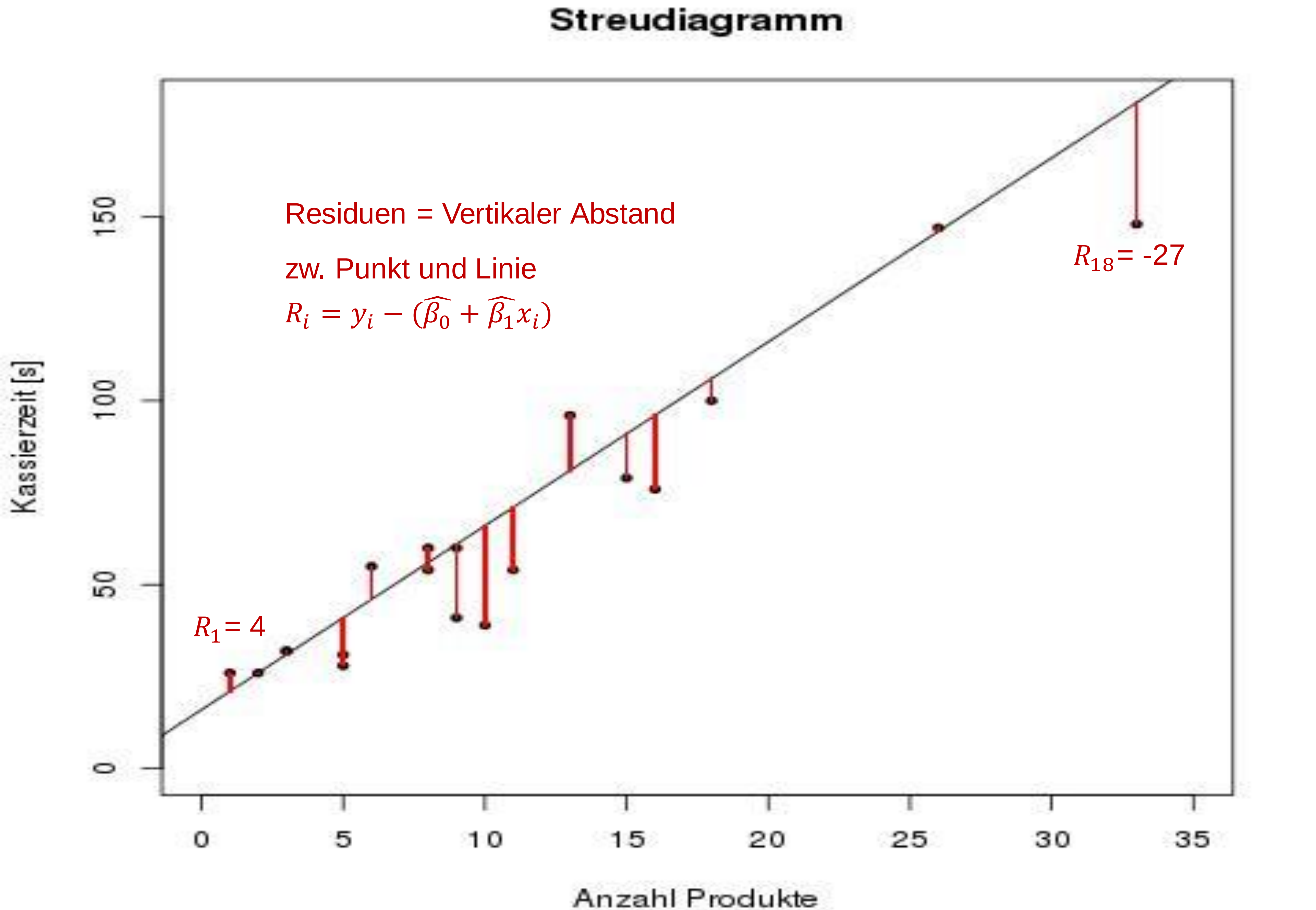



Streudiagramm
Kassierzeit [s]
0
50
100
150
0
5
10
15
20
25
30
35
Anzahl Produkte

Wdh: Korrelation




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_{18}= -27
R_1= 4
Kassierzeit [s]
0
50
100
150
0
5
10
15
20
25
30
35
Anzahl Produkte

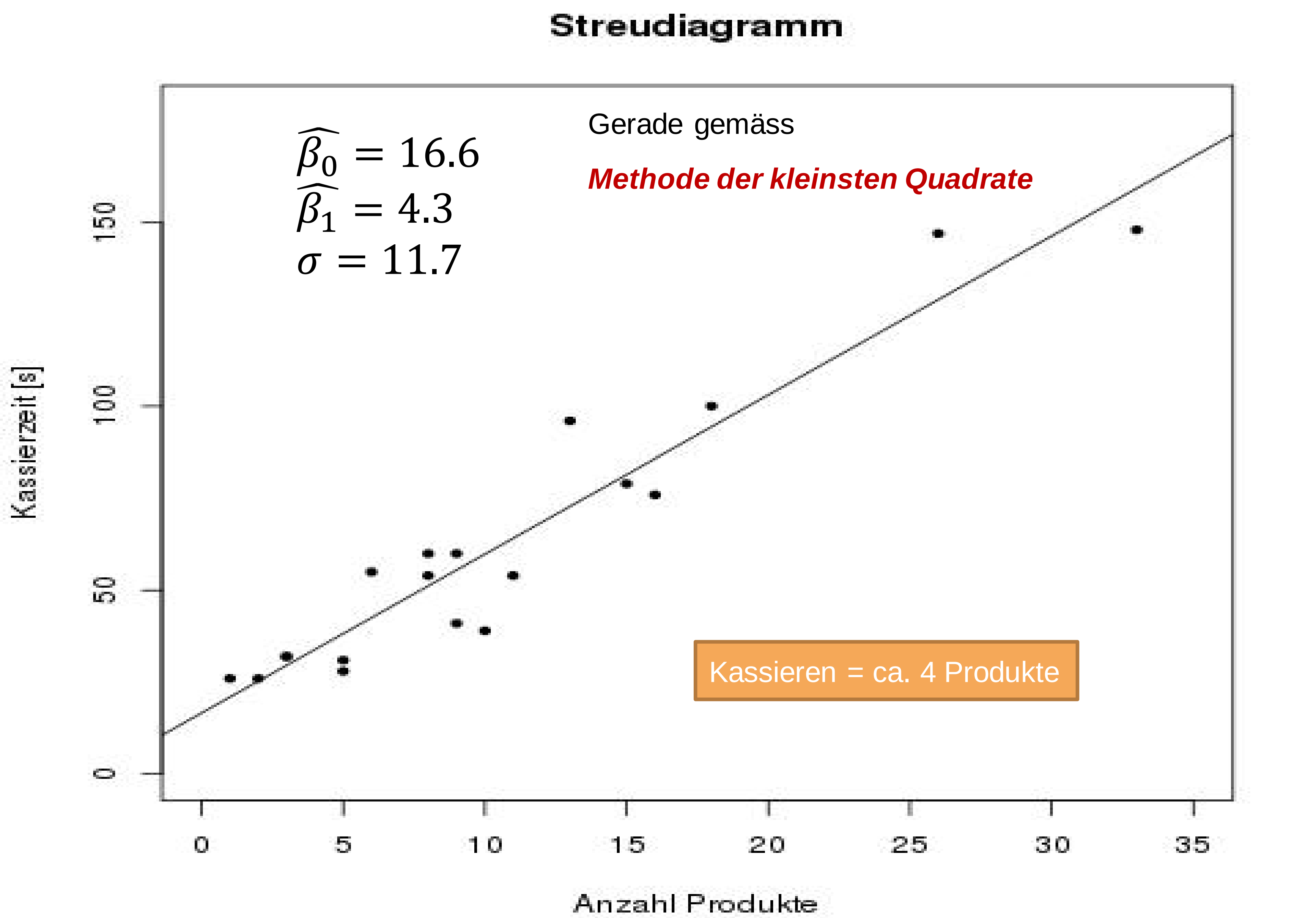
Parameterschätzung: Methode der kleinsten Quadrate ("Least Squares", LS)

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

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

- Lösung mit Analysis:

$$\hat{\beta}_1 = \left(\sum_{i=1}^{n} (y_i - \overline{y}_n)(x_i - \overline{x}_n) \ / \sum_{i=1}^{n} (x_i - \overline{x}_n)^2 \right)$$
$$\hat{\beta}_0 = \overline{y}_n - \hat{\beta}_1 \overline{x}_n$$



Streudiagramm
$\widehat{\beta_0} = 16.6$
$\widehat{\beta_1} = 4.3$
$\sigma = 11.7$
Gerade gemäss
Methode der kleinsten Quadrate
Kassieren = ca. 4 Produkte
Kassierzeit [s]
Anzahl Produkte
0
50
100
150
0
5
10
15
20
25
30
35

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*

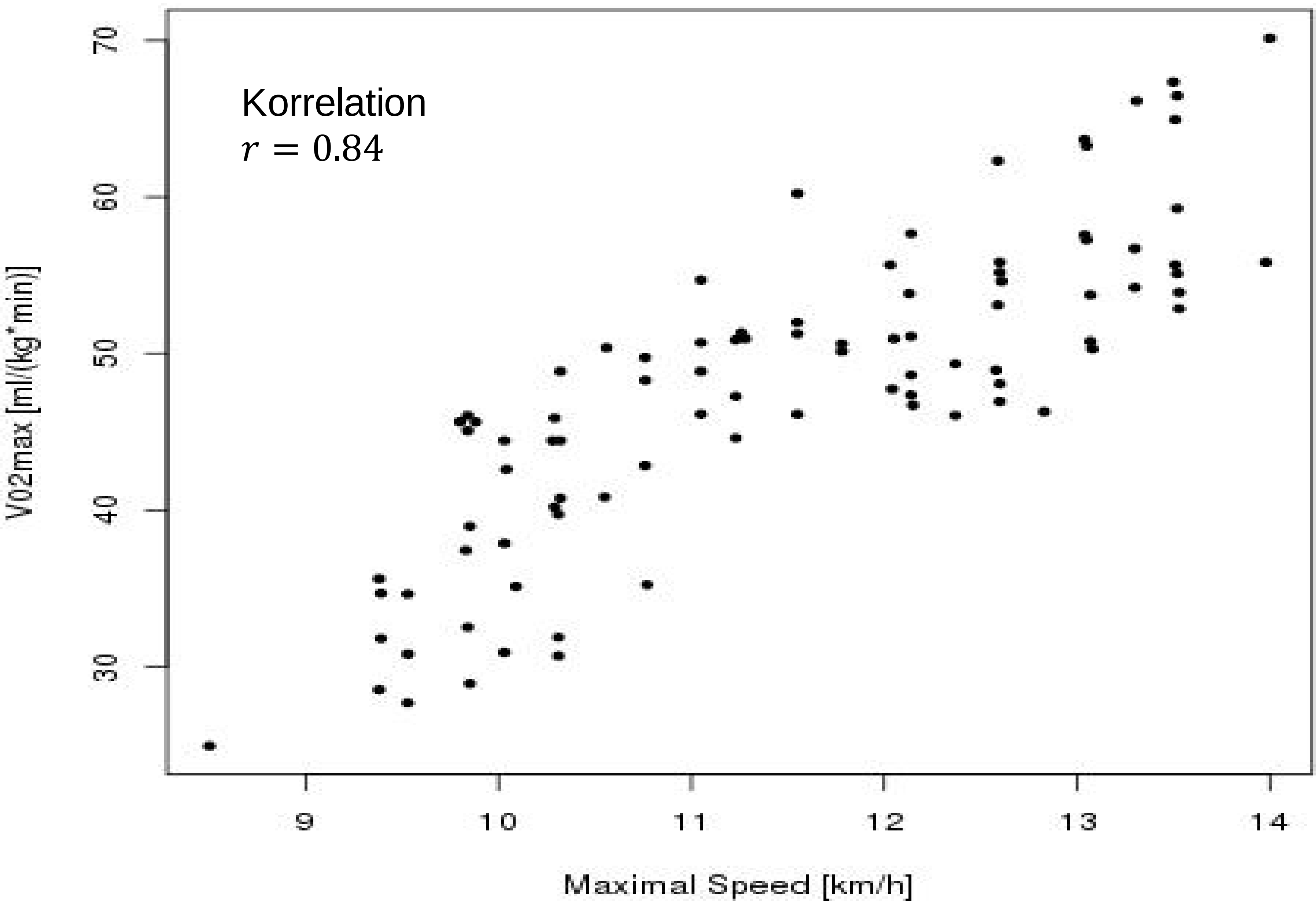
Luc A. Léger[1] and J. Lambert[2]

[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

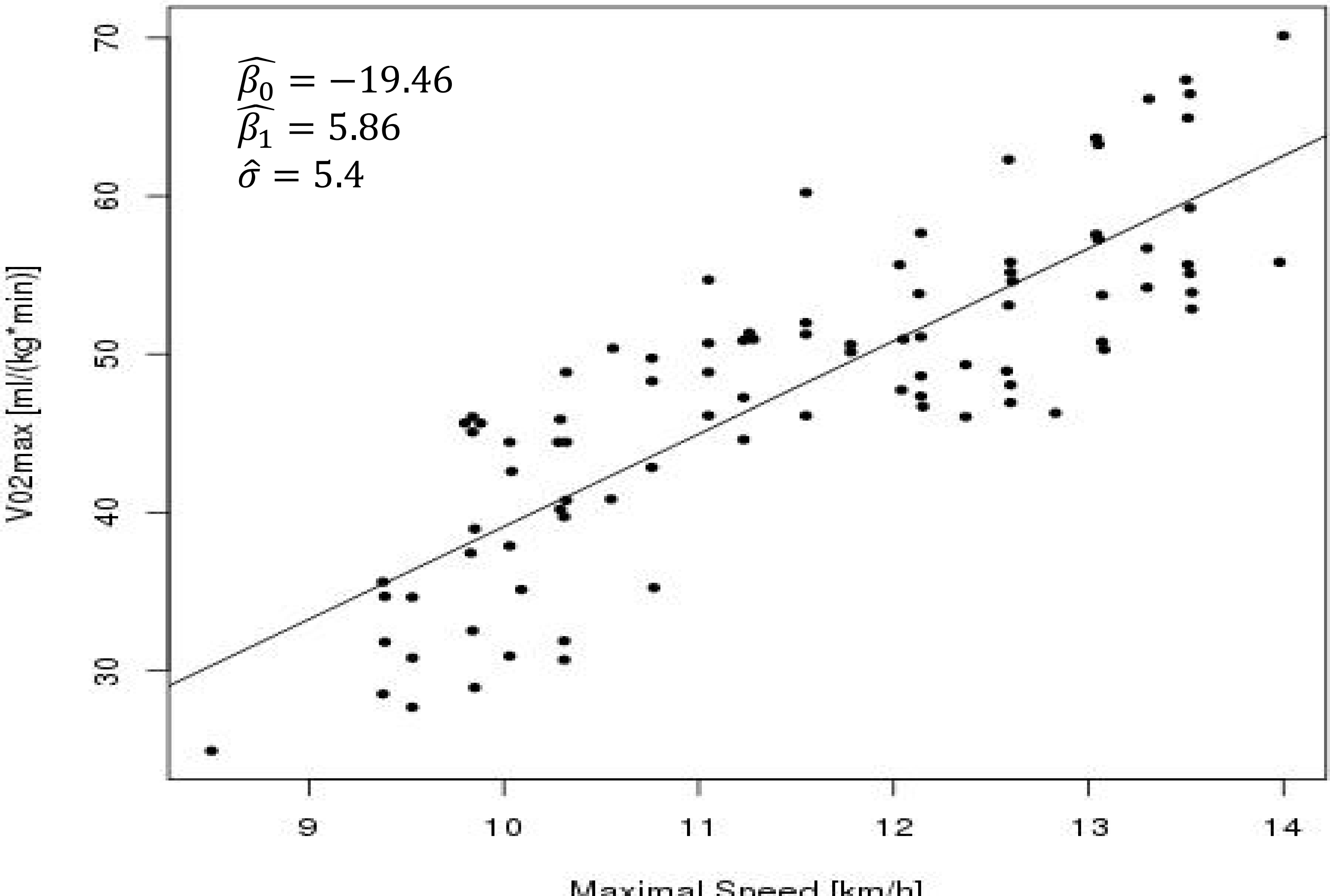
Ersatz: Cooper & Shuttle

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

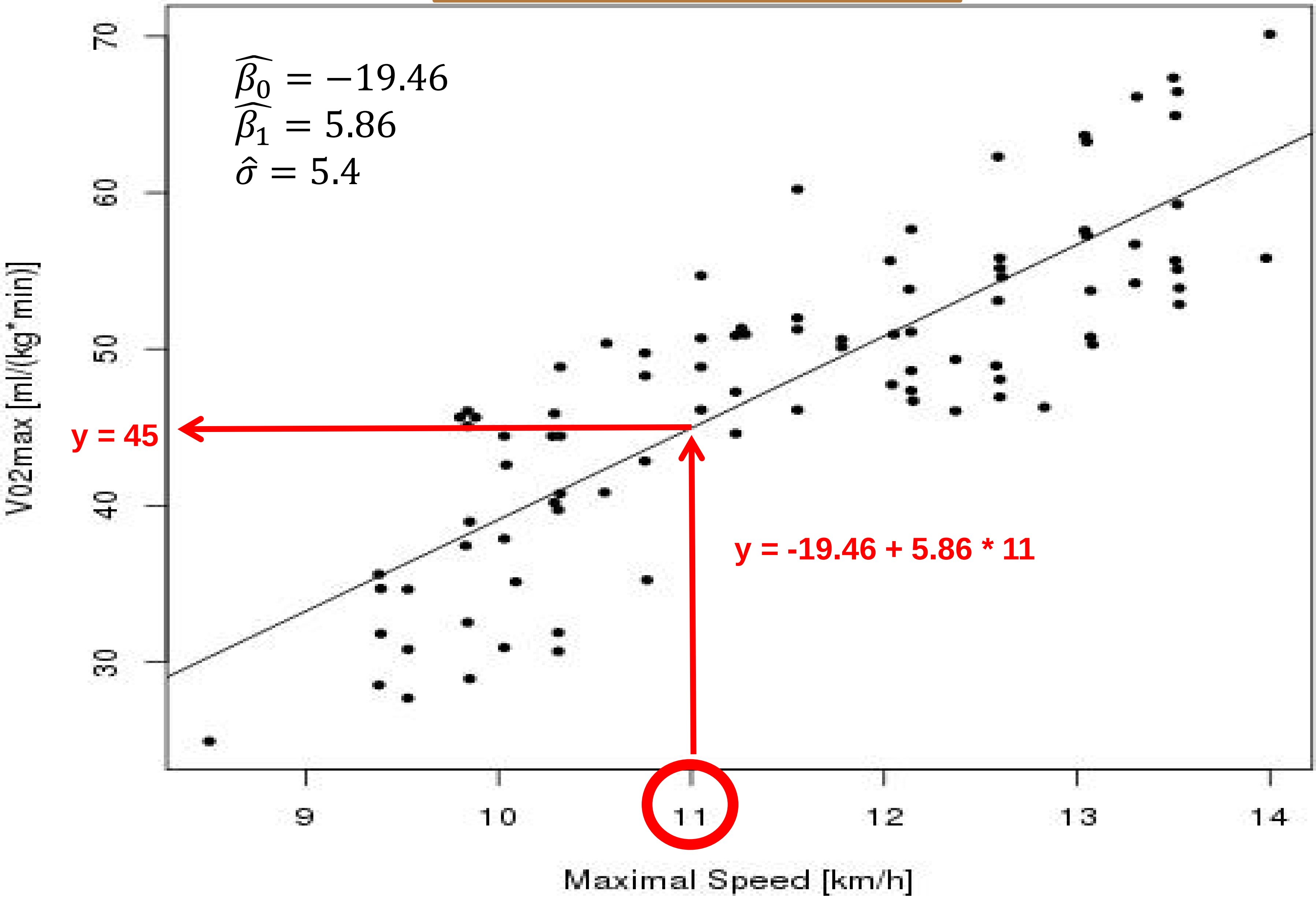
- Kann Shuttle-Test den VO2max-Wert vorhersagen?
- Falls ja: Einfache Testmöglichkeit für breite Bevölkerung


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

Methode der kleinsten Quadrate



- Wie genau stimmen Parameter?
- Wie genau stimmt Vorhersage?



t-Test in der Linearen Regression

1. Modell: $Y_i = \beta_0 + \beta_1 x_i + E_i, E_1, \ldots, E_n \;\; iid \;\; N(0, \sigma^2)$
2. Nullhypothese: $H_0: \beta_1 = 0$
   Alternative: $H_A: \beta_1 \neq 0$ (es wird normalerweise ein zweiseitiger Test durchgeführt)
3. Teststatistik:
   $$T = \frac{beobachtet - erwartet}{geschätzter\ Standardfehler} = \frac{\widehat{\beta_1} - 0}{\widehat{s.e.}\left(\widehat{\beta_1}\right)}$$
   Dabei ist $s.e.\left(\widehat{\beta_1}\right) = \sqrt{Var\left(\widehat{\beta_1}\right)}$ der “Standard Error” von $\widehat{\beta_1}$
   Verteilung der Teststatistik unter $H_0: T \sim t_{n-2}$
4. Signifikanzniveau: α
5. Verwerfungsbereich der Teststatistik:
   $$K = \left(-\infty, -t_{n-2;1-\frac{\alpha}{2}}\right] \cup \left[t_{n-2;1-\frac{\alpha}{2}}, \infty\right)$$
6. Testentscheid: Überprüfe, ob der beobachtete Wert der Teststatistik im Verwerfungsbereich liegt.

Lineare Regression in R

Modell: $Y_i = \beta_0 + \beta_1 x_i + E_i,\ E_i \sim N(0, \sigma^2)\ i.i.d$

Modell: $Y_i = -19.46 + 5.86 x_i + E_i,\ E_i \sim N(0, 5.43^2)\ i.i.d$

Standardfehler von $\widehat{\beta_1}$ $(= \hat{\sigma}_{\widehat{\beta 1}})$

Approx. 95%-VI:

$5.86 \pm 2 * 0.41$

Exaktes 95%-VI:

$5.86 \pm 1.99 * 0.41$

$t_{89;0.975}$

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

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
```

Beobachtete Teststatistik

im Test $H_0: \beta_1 = 0$ vs.

$H_A: \beta_1 \neq 0$

P-Wert:

Angenommen $\beta_1 = 0$;

wie wa. ist Beobachtung

oder etwas extremeres?

Freiheitsgrade: n – (Anz. β’s) = 91 – 2 = 89

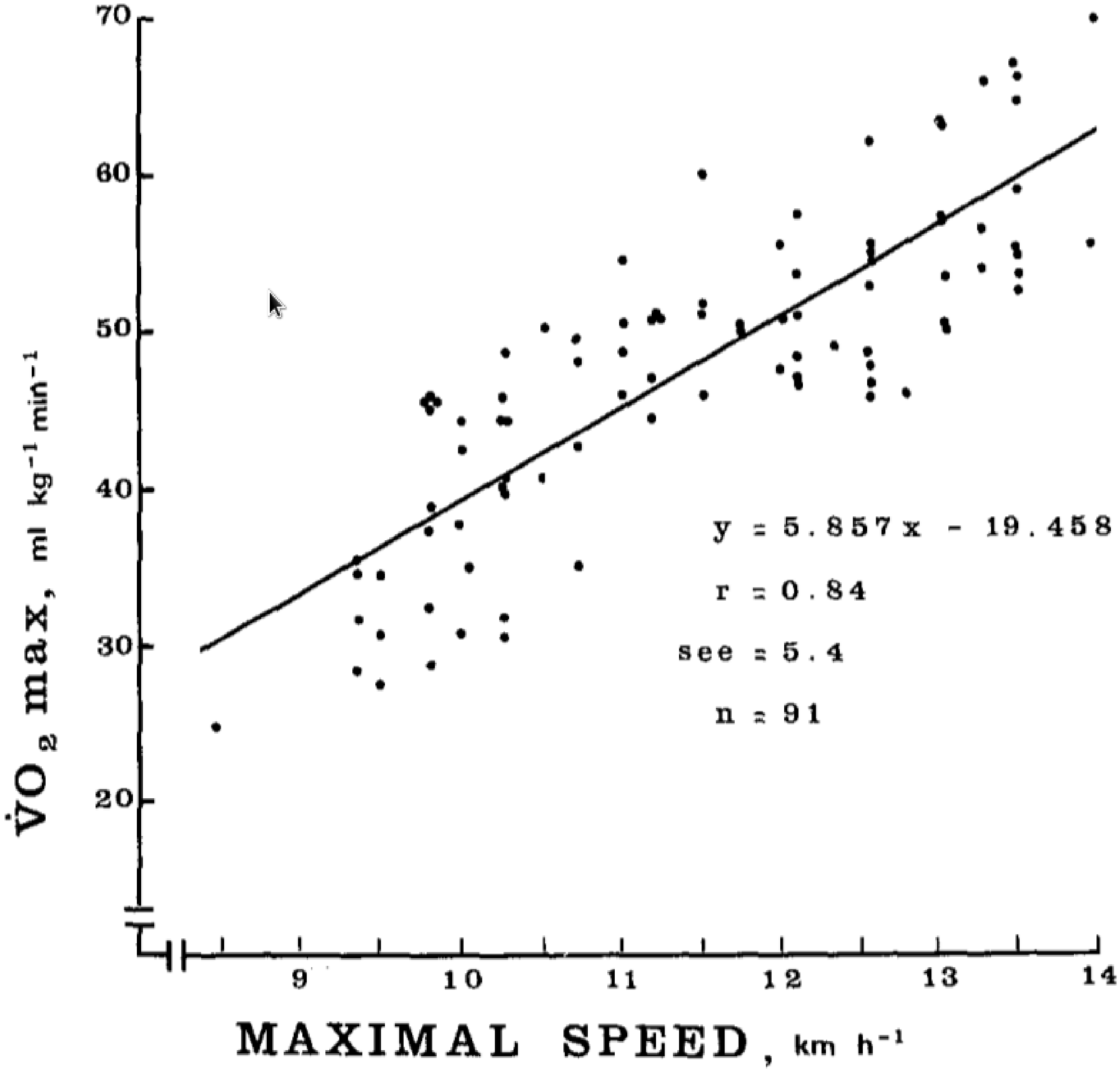


Fig. 2. $\dot{V}O_2$ max as a function of the maximal speed achieved in the 20-m shuttle run test for a total sample of 91 adult subjects. Each point in this figure represents maximal effort

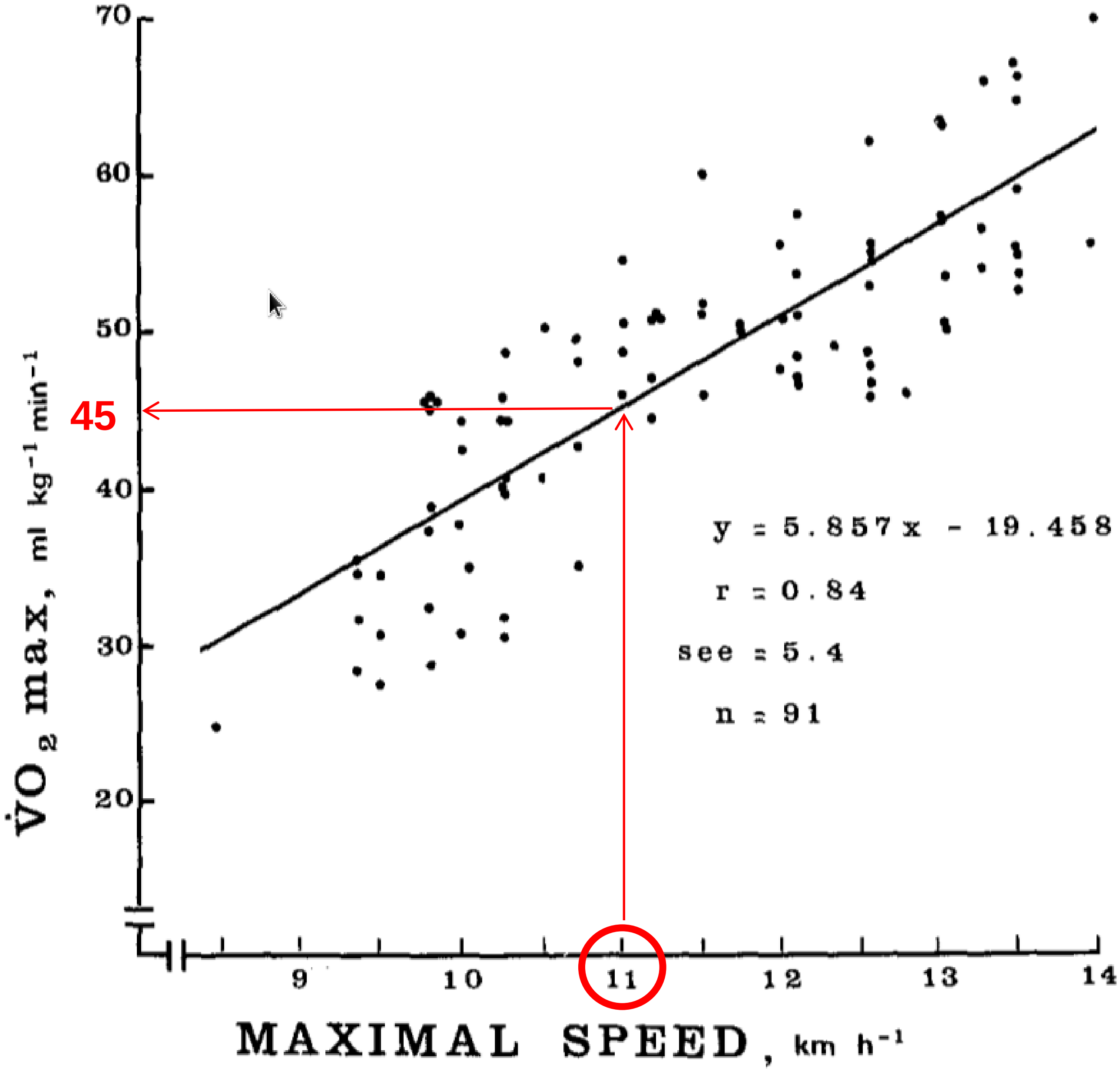


Fig. 2. $\dot{V}O_2$ max as a function of the maximal speed achieved in the 20-m shuttle run test for a total sample of 91 adult subjects. Each point in this figure represents maximal effort

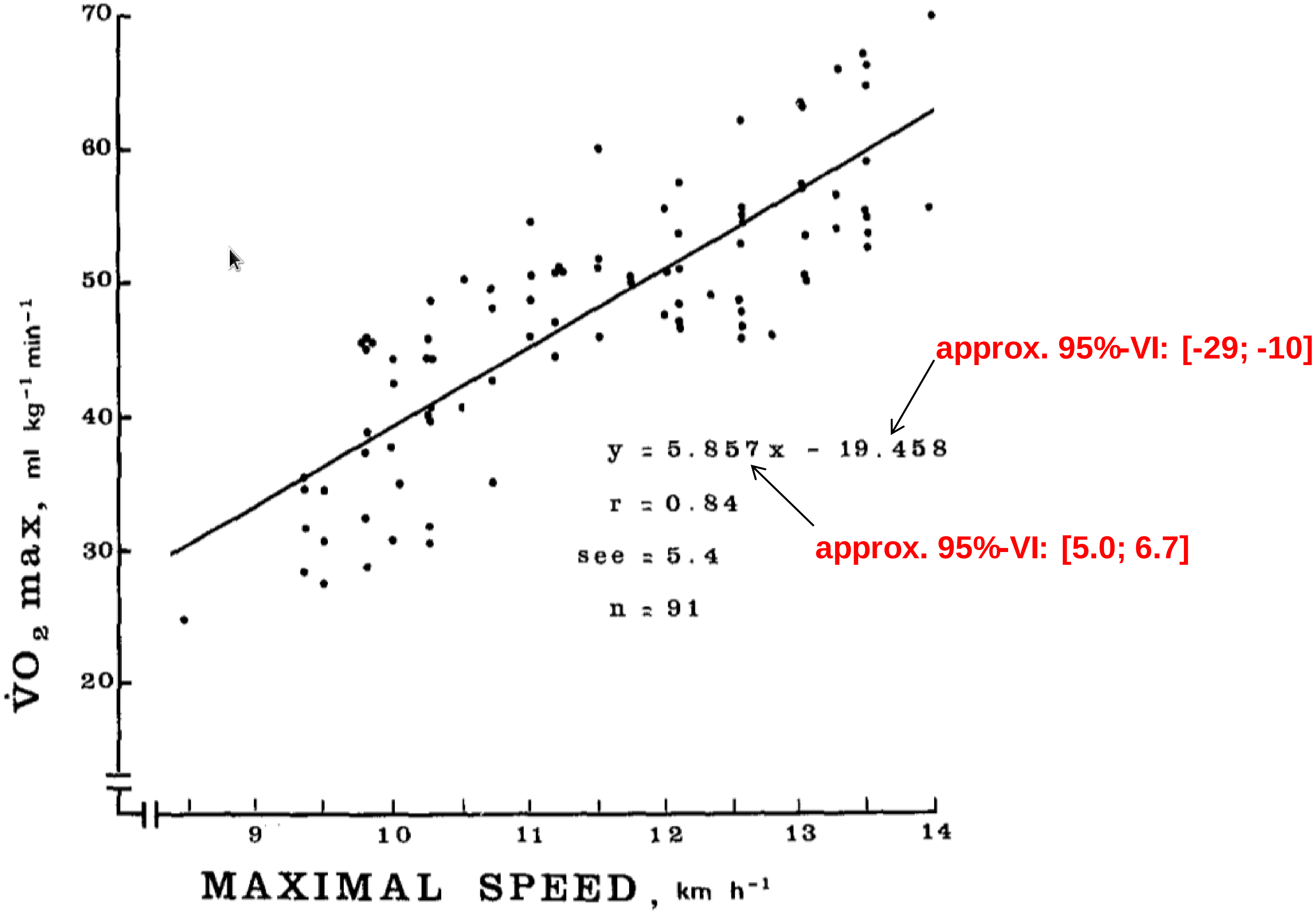


Fig. 2. $\dot{V}O_2$ max as a function of the maximal speed achieved in the 20-m shuttle run test for a total sample of 91 adult subjects. Each point in this figure represents maximal effort

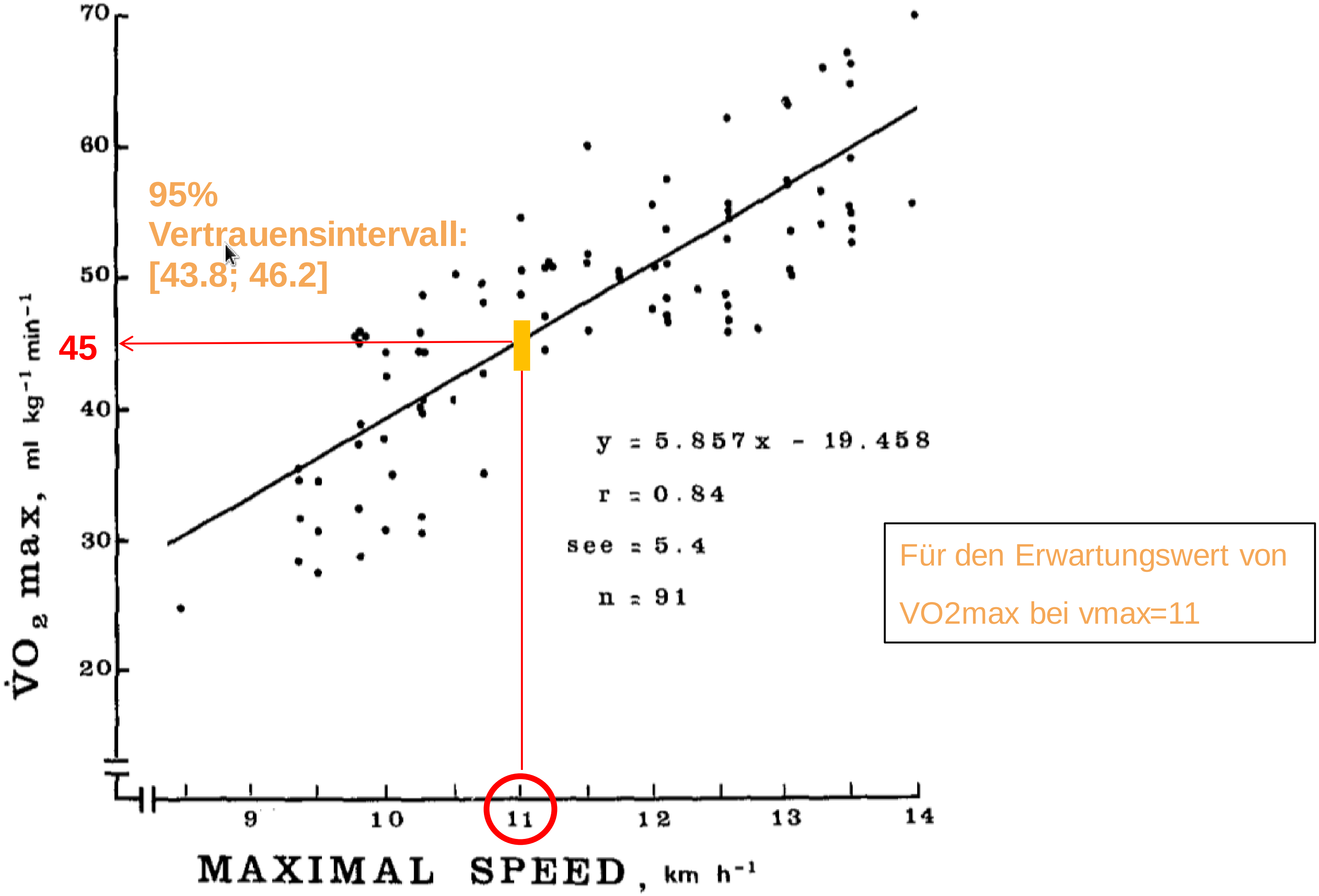


Fig. 2. $\dot{V}O_2$ max as a function of the maximal speed achieved in the 20-m shuttle run test for a total sample of 91 adult subjects. Each point in this figure represents maximal effort

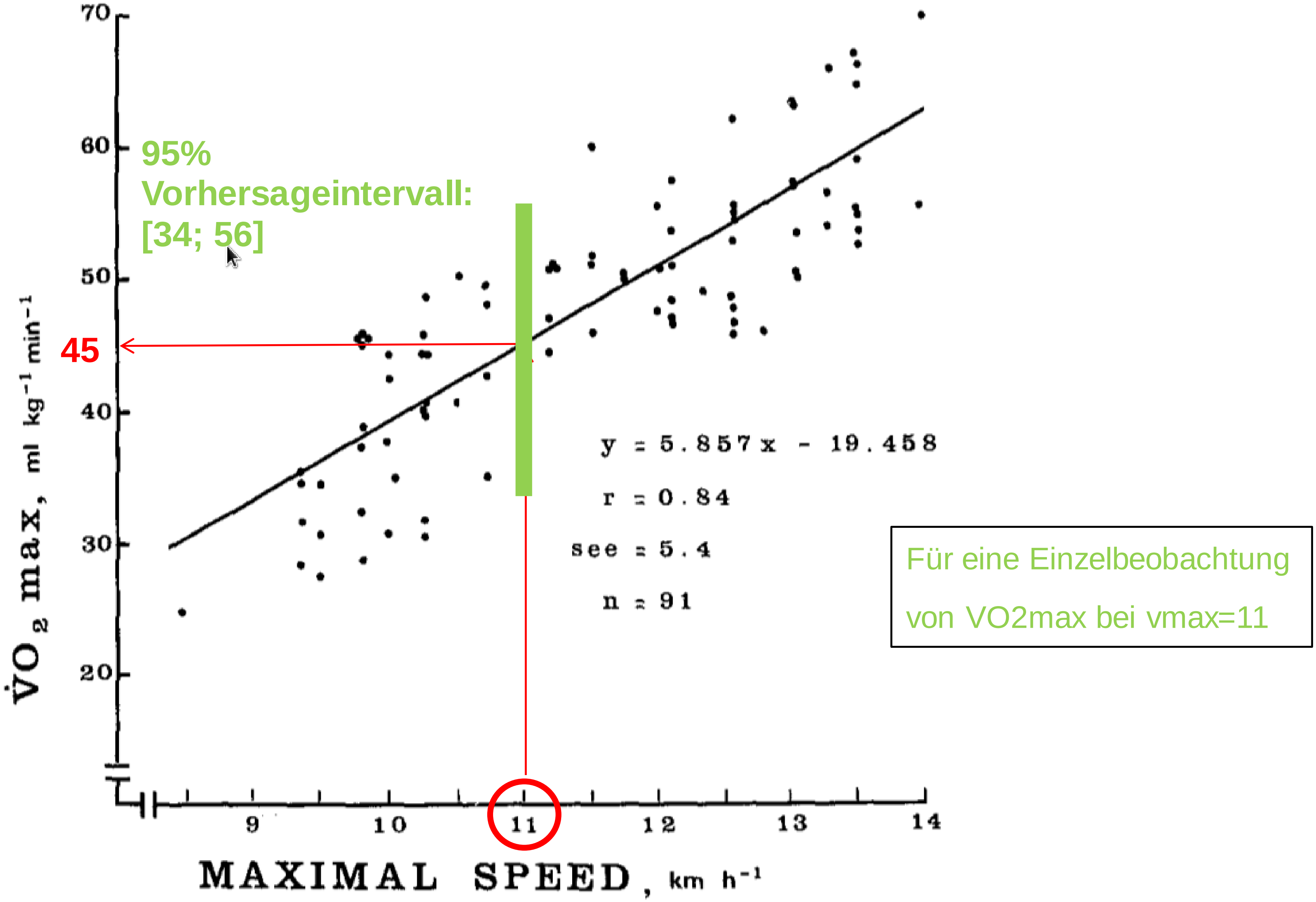
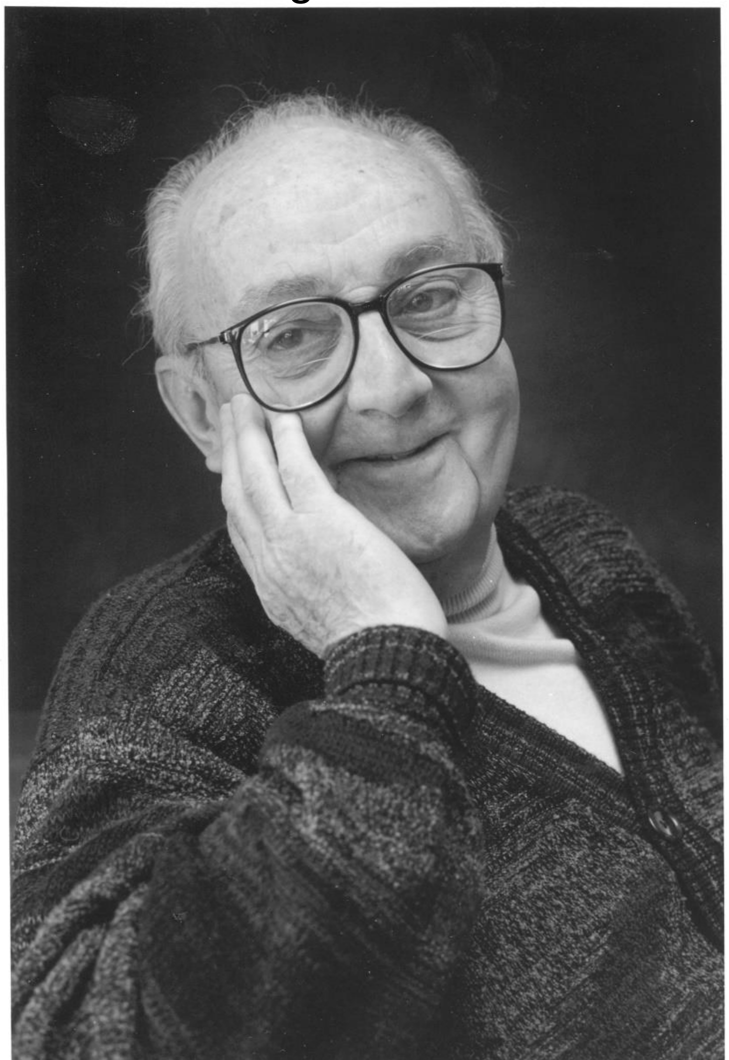


Fig. 2. $\dot{V}O_2$ max as a function of the maximal speed achieved in the 20-m shuttle run test for a total sample of 91 adult subjects. Each point in this figure represents maximal effort

George E.P. Box



“Essentially,
all models are
wrong,
but some are
useful.“

Residuenanalyse: Wie gut stimmt das Modell ?

$$\underline{Y_i = \beta_0 + \beta_1 x_i} + \varepsilon_i \;\; ; \varepsilon_i \sim \underline{N}(0, \underline{\sigma^2}) \quad iid$$

- Form des funktionellen Zusammenhangs
- Varianz der Fehler ist konstant
- Fehler sind normalverteilt

Einfache Regression:

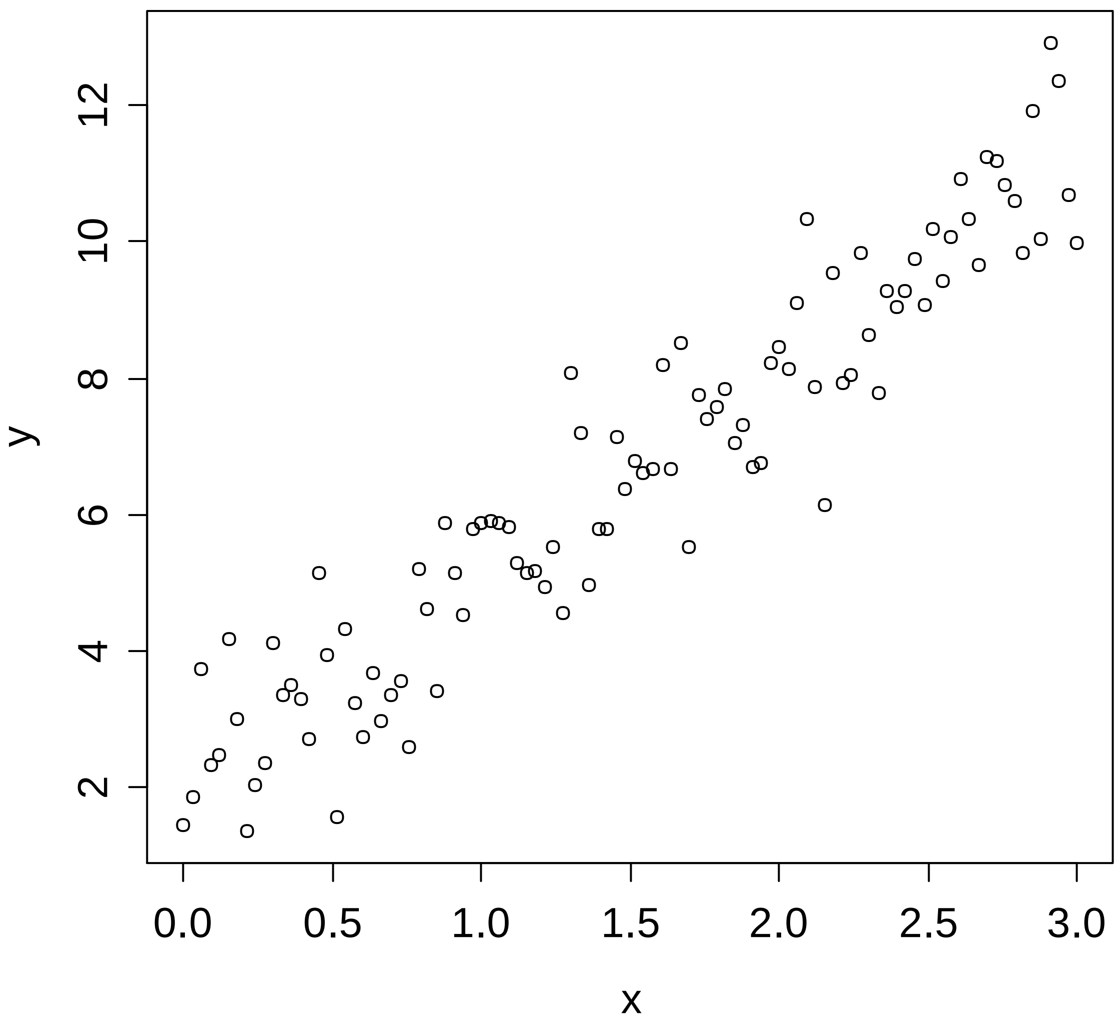
Streudiagramm

Multiple Regression:

Tukey-Anscombe Plot

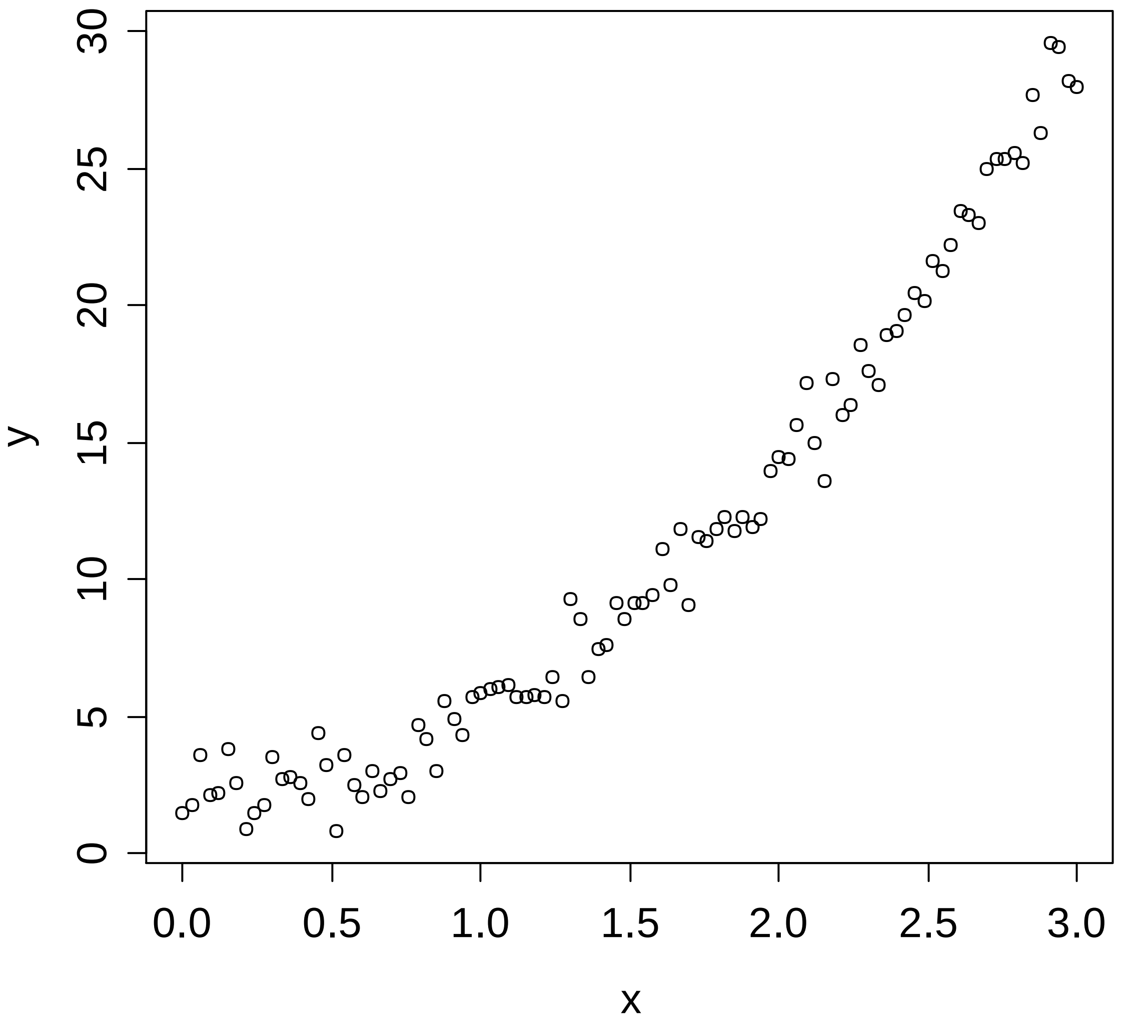
QQ-Plot der Residuen

Streudiagramm bei einfacher linearer Regression



OK

Streudiagramm bei einfacher linearer Regression

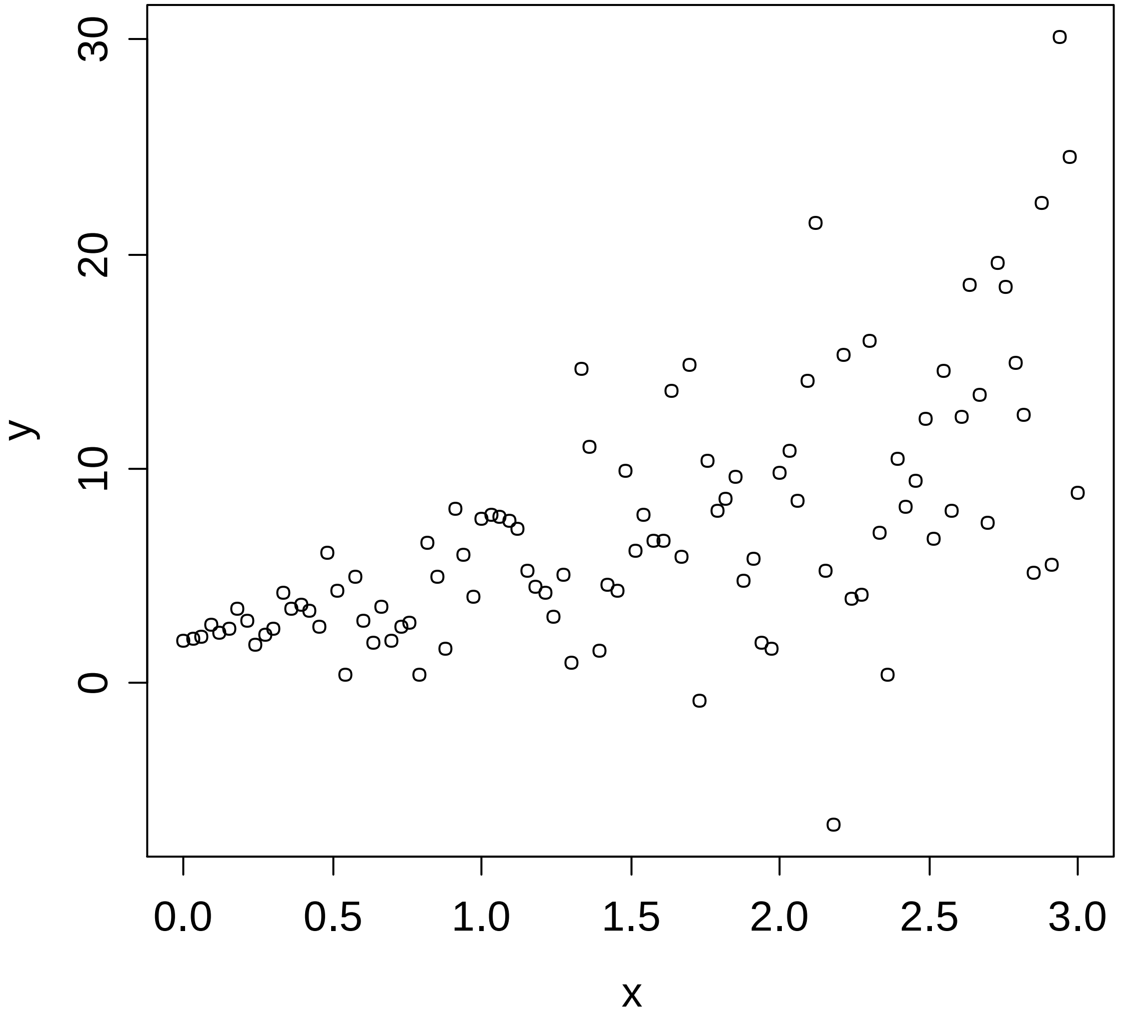


Systematischer Fehler

Krümmung:

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

Streudiagramm bei einfacher linearer Regression



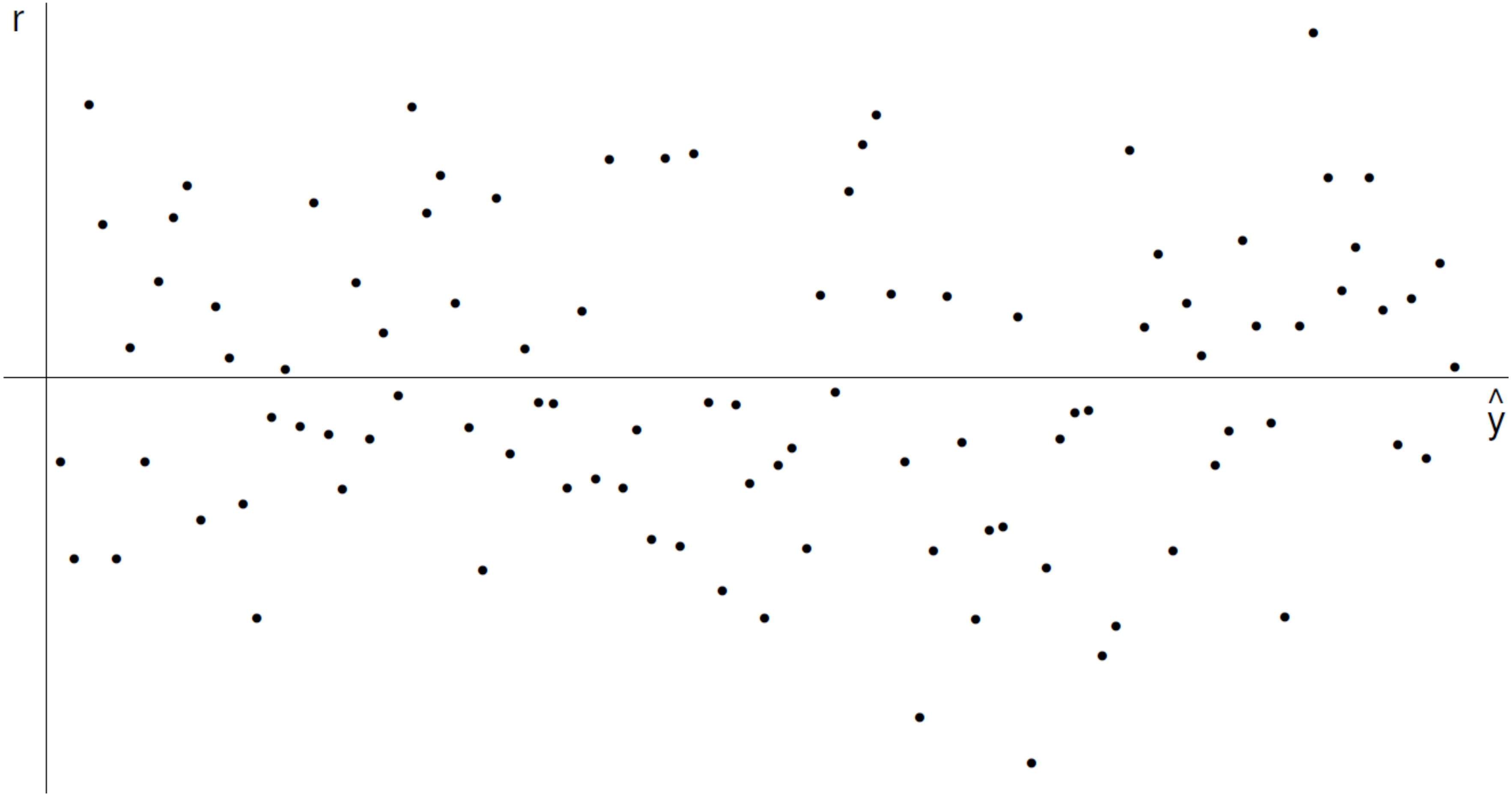
Fehlervarianz nicht konstant

Tukey-Anscombe Plot

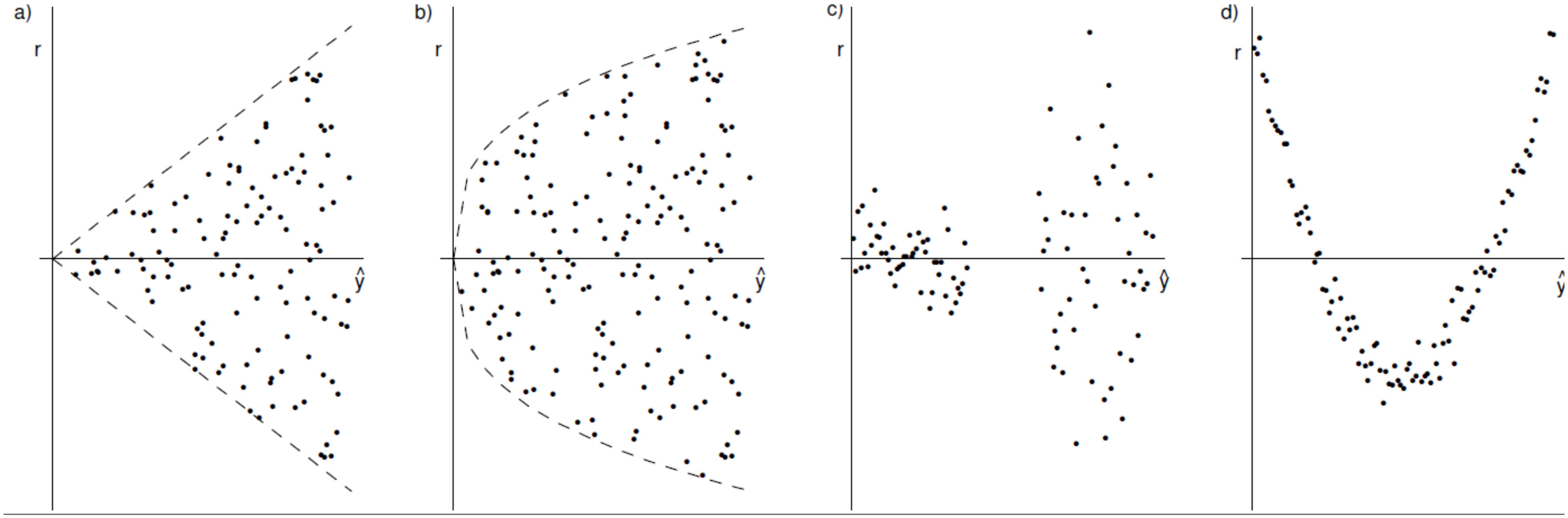
- Alternative zum Streudiagramm
  Vor allem sinnvoll bei multipler Regression
- Standard-Output in Residuenanalyse von Statistikprogrammen
- x-Achse: Vom Modell vorhergesagte Werte ($\hat{y}$)
- y-Achse: Residuen
- Das Modell sortiert die Datenpunkte also gemäss der vorhergesagten y-Werte

- Idealer Tukey-Anscombe Plot:
  Band mit konstanter Breite und uniformer Streuung der Punkte innerhalb des Bandes

Beispiel für guten Tukey-Anscombe Plot



Beispiele für schlechte Tukey-Anscombe Plots



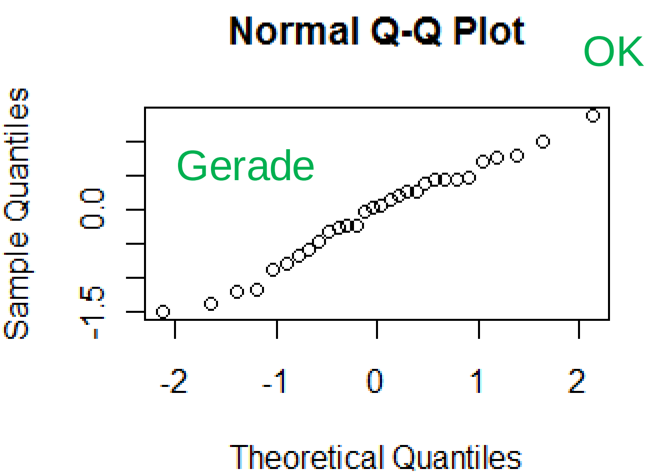
Fehlervarianz nicht konstant

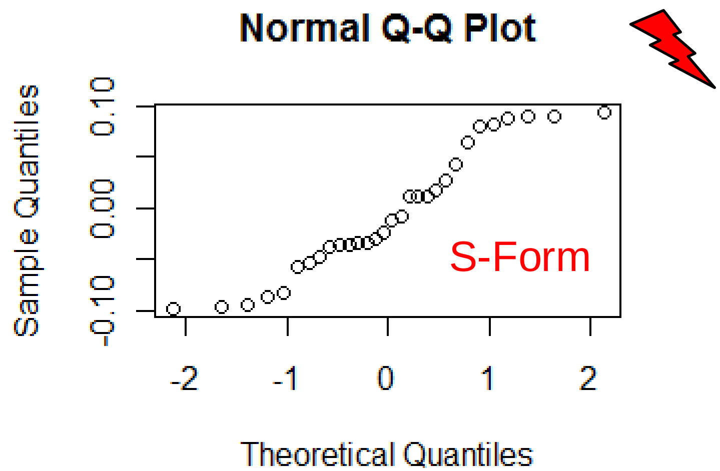
Systematischer Fehler

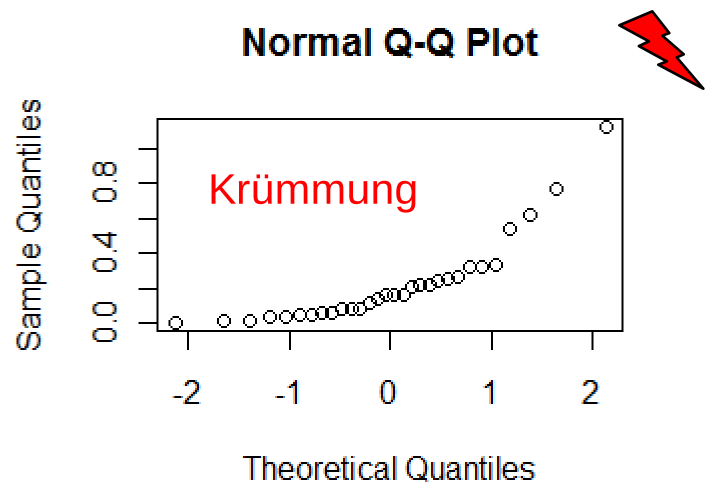
Residuenanalyse: QQ-Plot

Gerade = “gut”

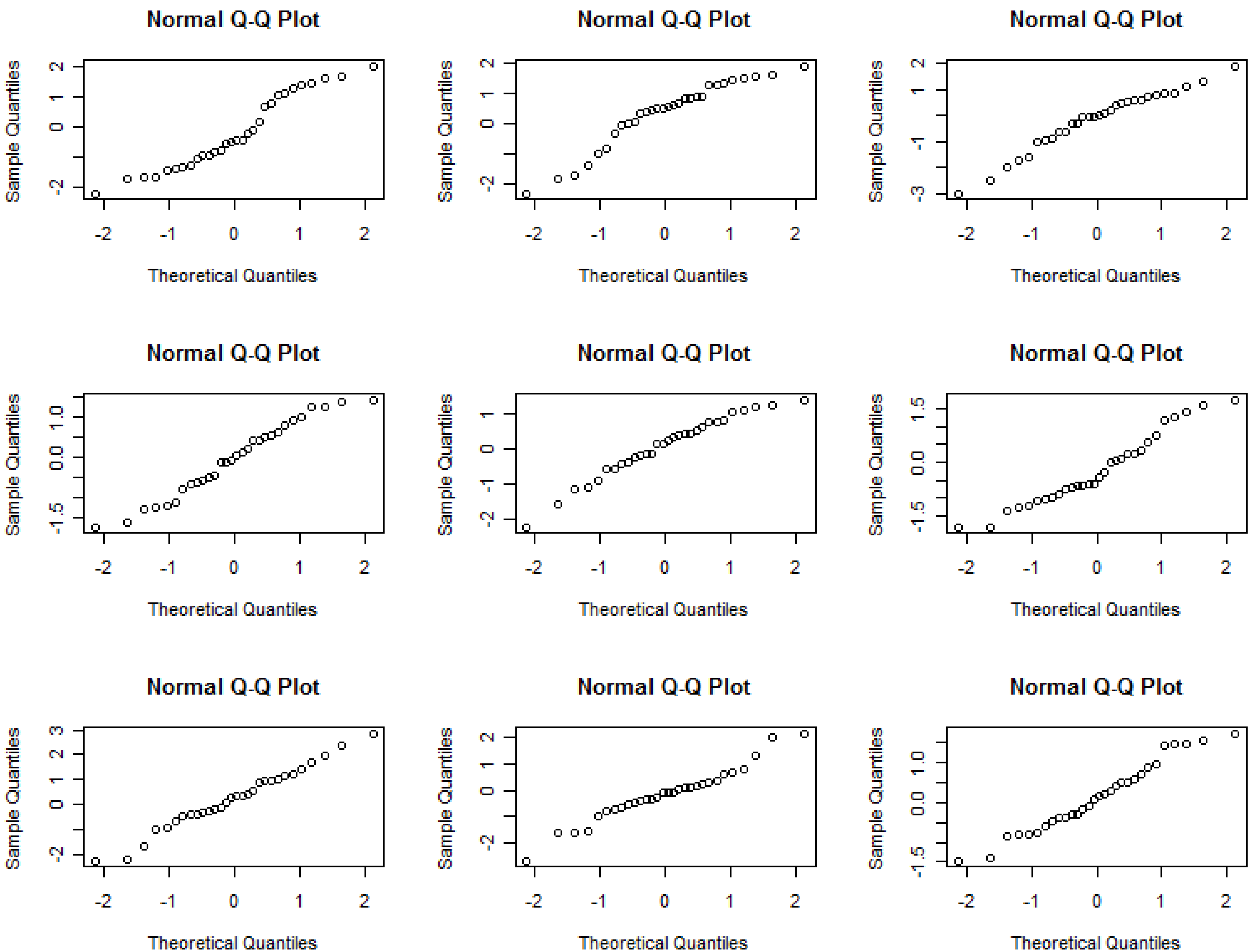
Krümmung = “schlecht”







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



Falls Residuenplots schlecht

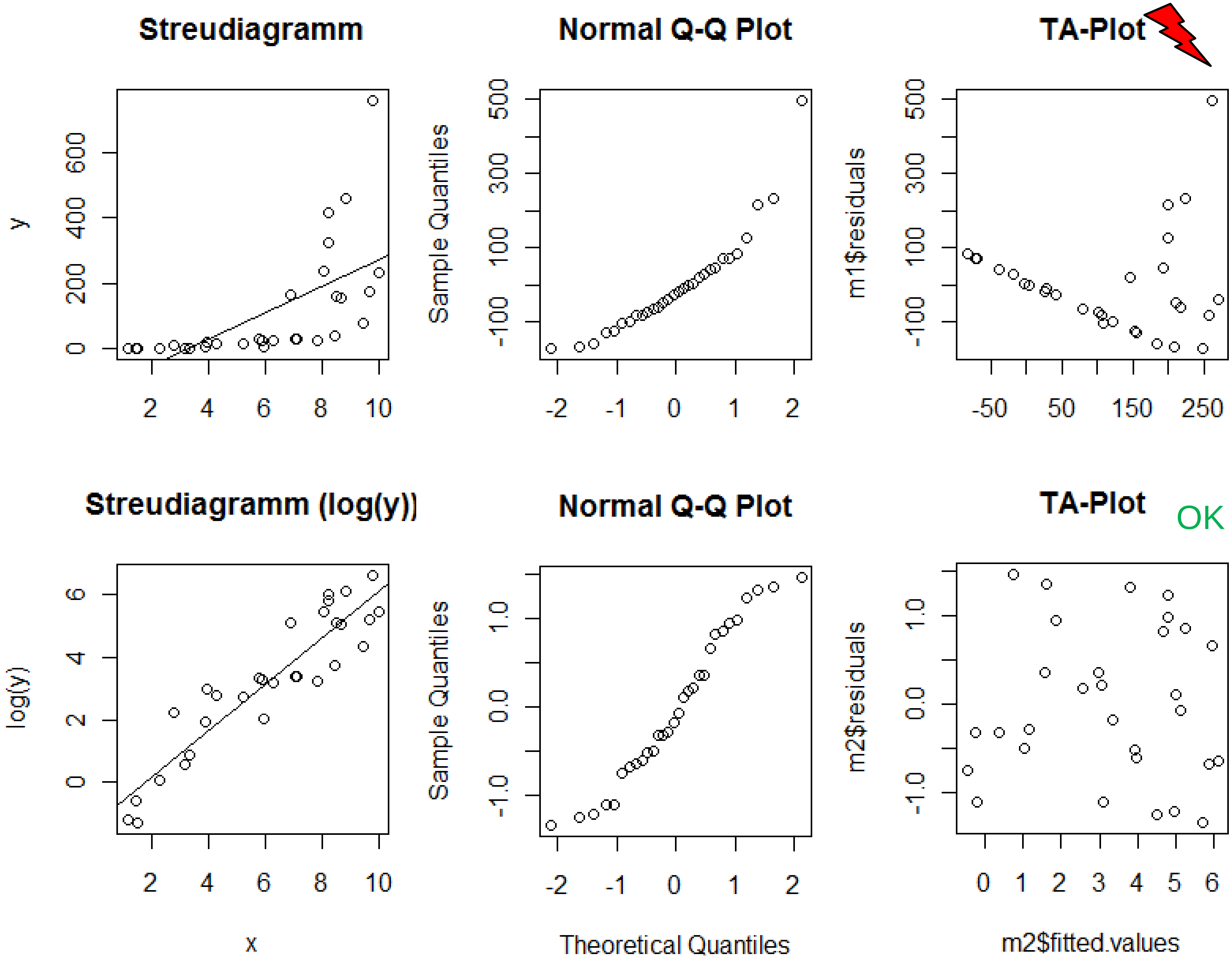
- Oft helfen Transformationen von x oder y
- Achtung: Vorsicht beim Interpretieren der neuen Parameter

- 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 ersetzt wird, ändert sich Y “im Mittel” zu $Y * \exp(\beta_1)$

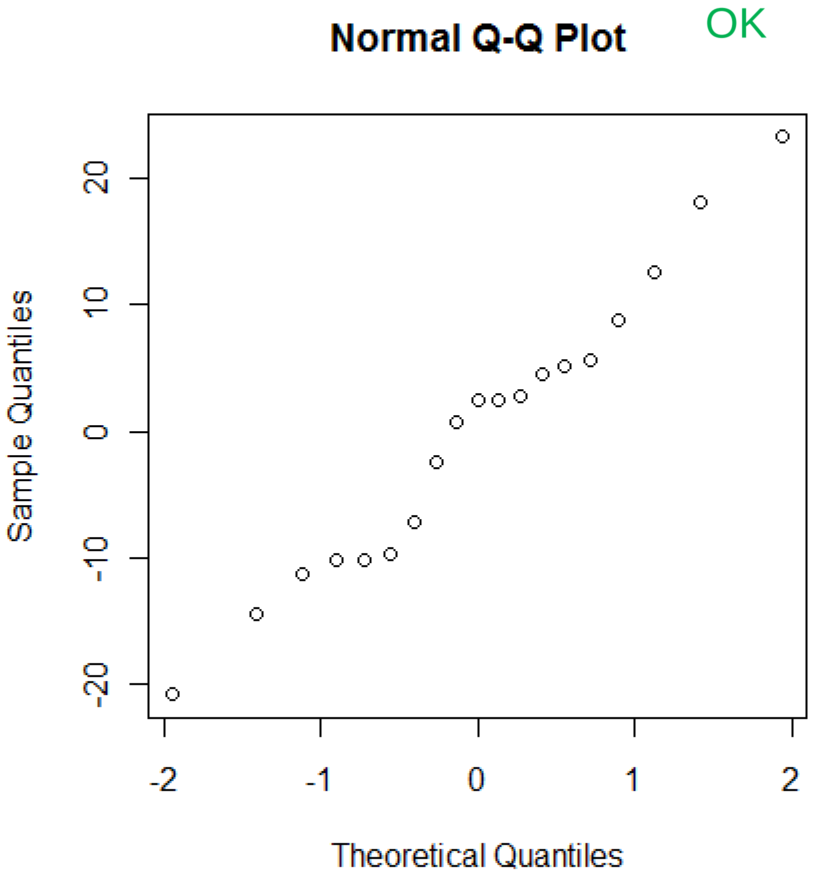
Bsp: Ohne Log-Transformation

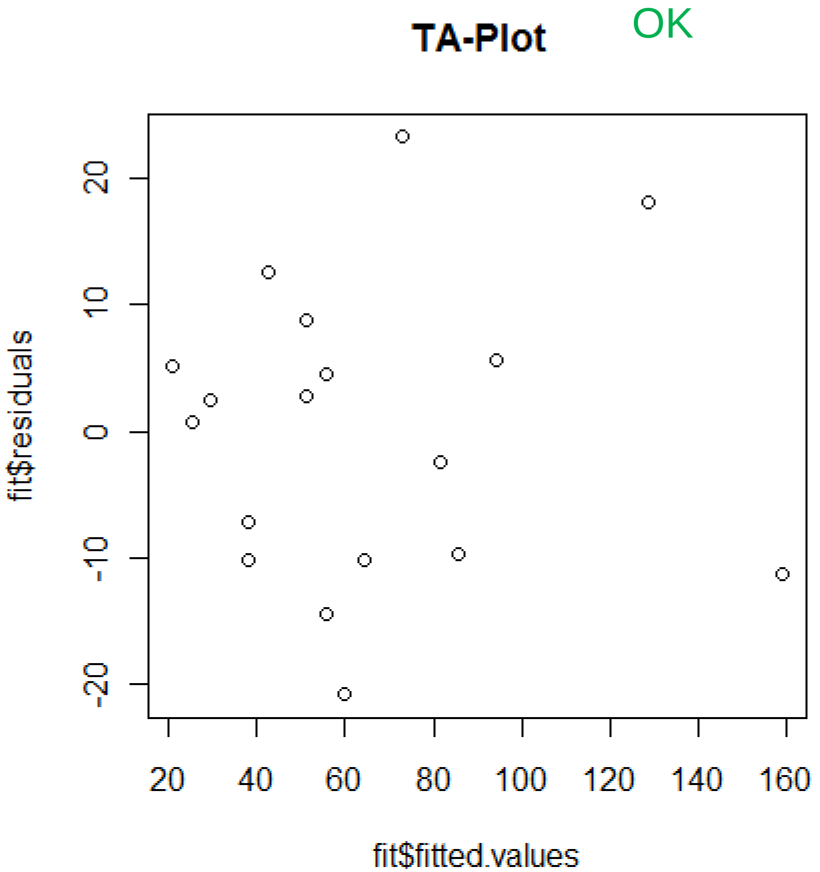


y

$\log(y)$

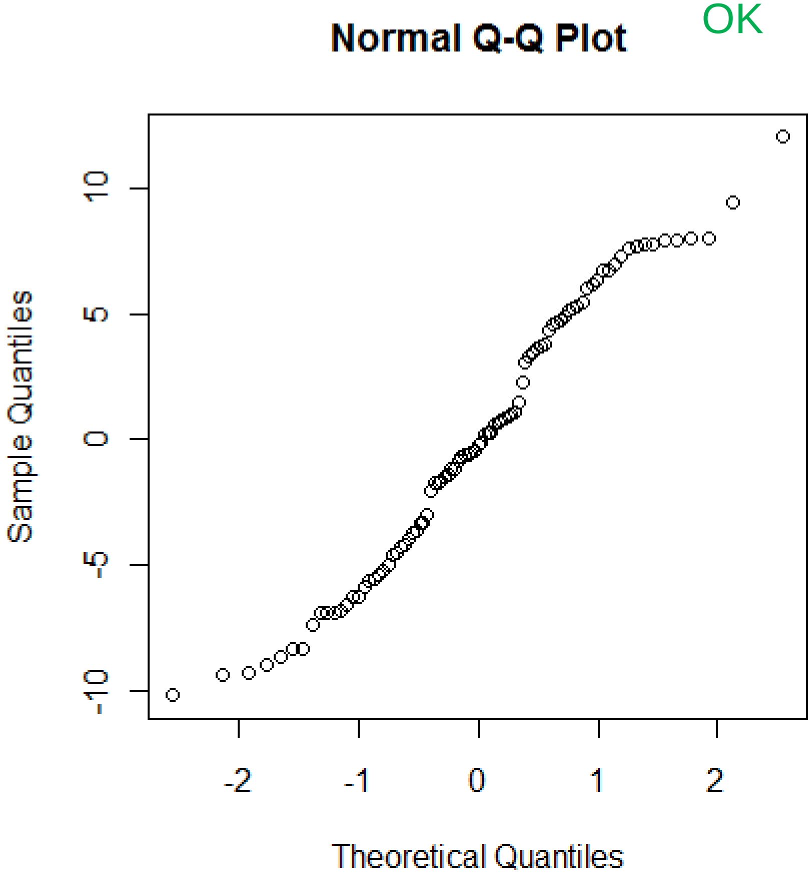
Residuenanalyse: Supermarkt





Residuenanalyse: Beep-Test

OK



OK