ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

DMATH
Department of Mathematics

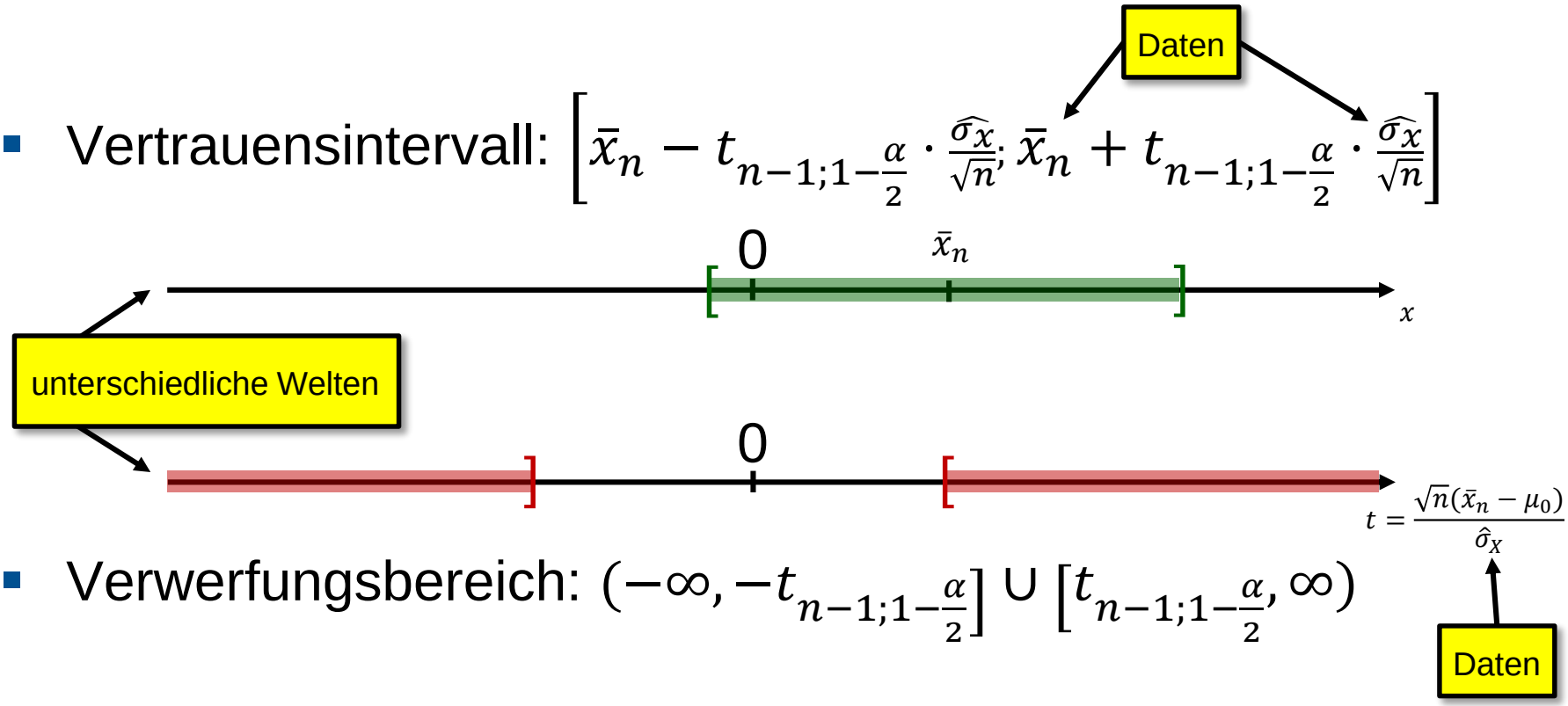
Einfache lineare Regression

für D-UWIS, D-ERDW, D-USYS und D-HEST – SS15

Vertrauensintervall versus Verwerfungsbereich

- Es ist verlockend das Vertrauensintervall als Komplement des Verwerfungsbereiches zu verstehen – trotzdem **falsch**!

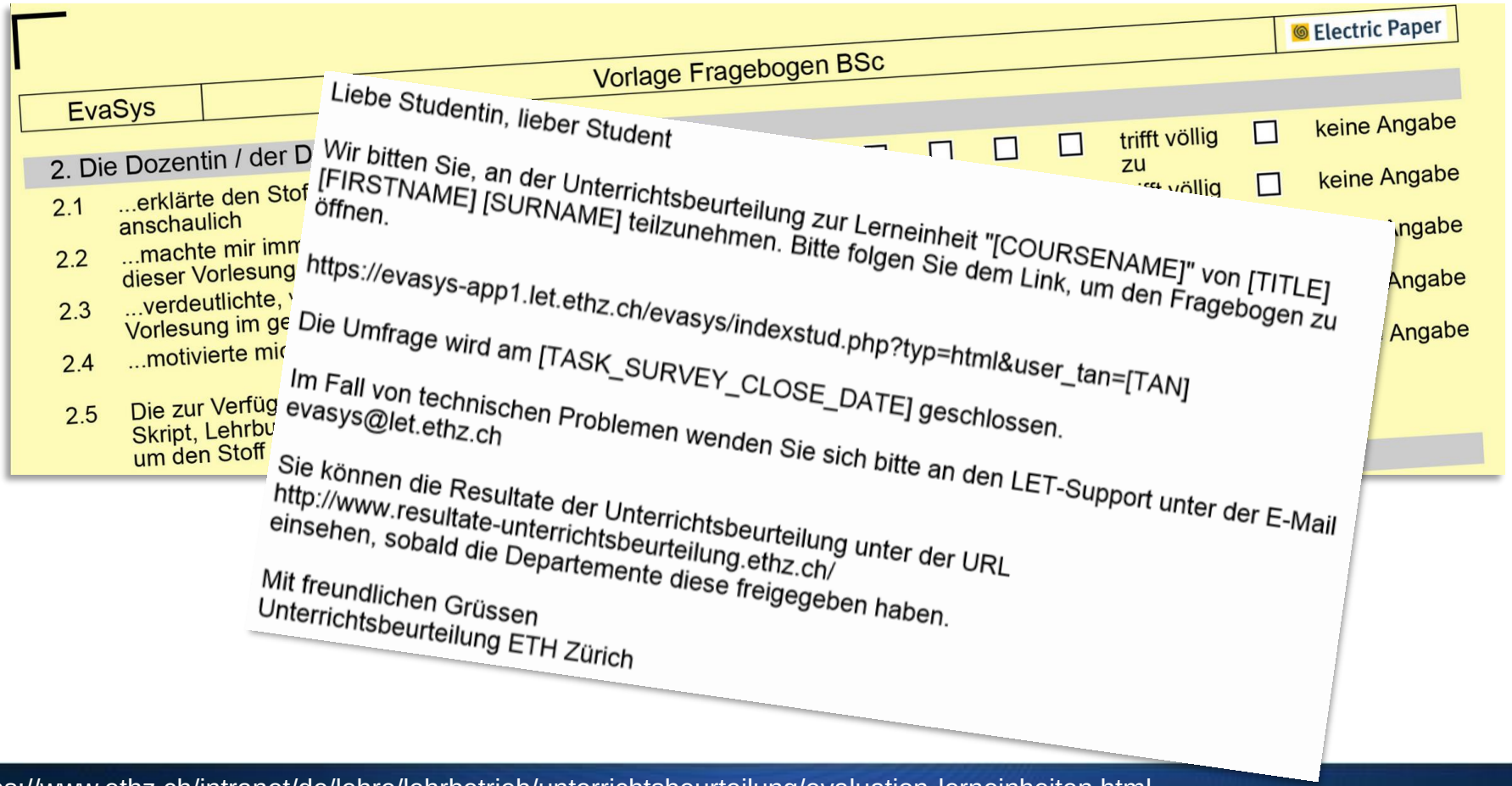
- Vertrauensintervall: $\left[\bar{x}_n - t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{\widehat{\sigma_X}}{\sqrt{n}}; \bar{x}_n + t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{\widehat{\sigma_X}}{\sqrt{n}}\right]$



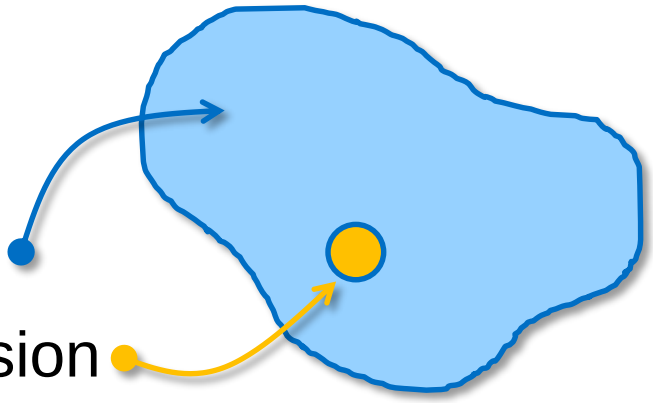
- Verwerfungsbereich: $(-\infty, -t_{n-1;1-\frac{\alpha}{2}}] \cup [t_{n-1;1-\frac{\alpha}{2}}, \infty)$

Unterrichtsbeurteilung

- ETH Zürich evasys@let.ethz.ch



Lernziele heute

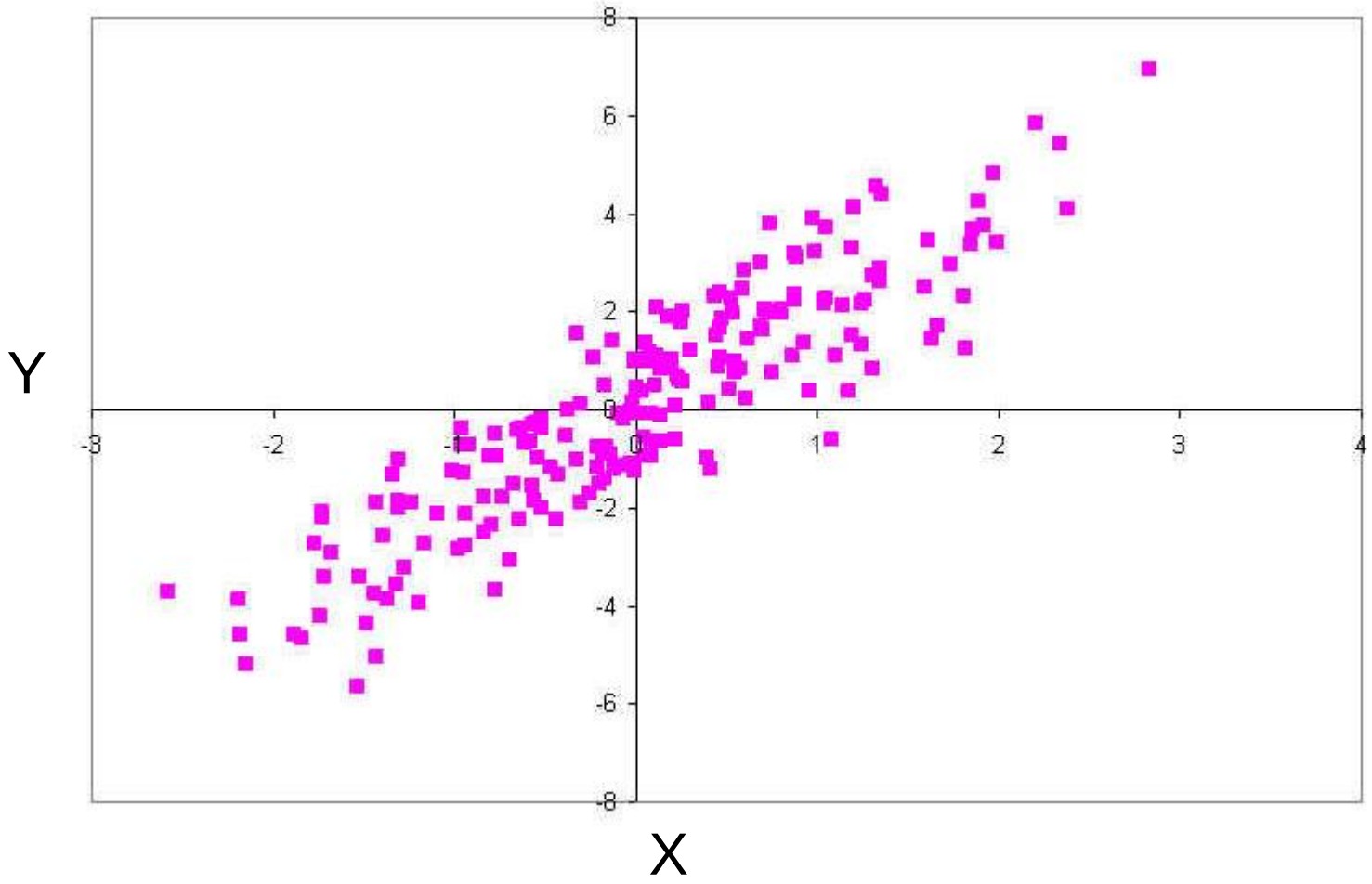


- Idee: Generalized Linear Model
- Details: Einfach lineare Regression

Hausaufgaben

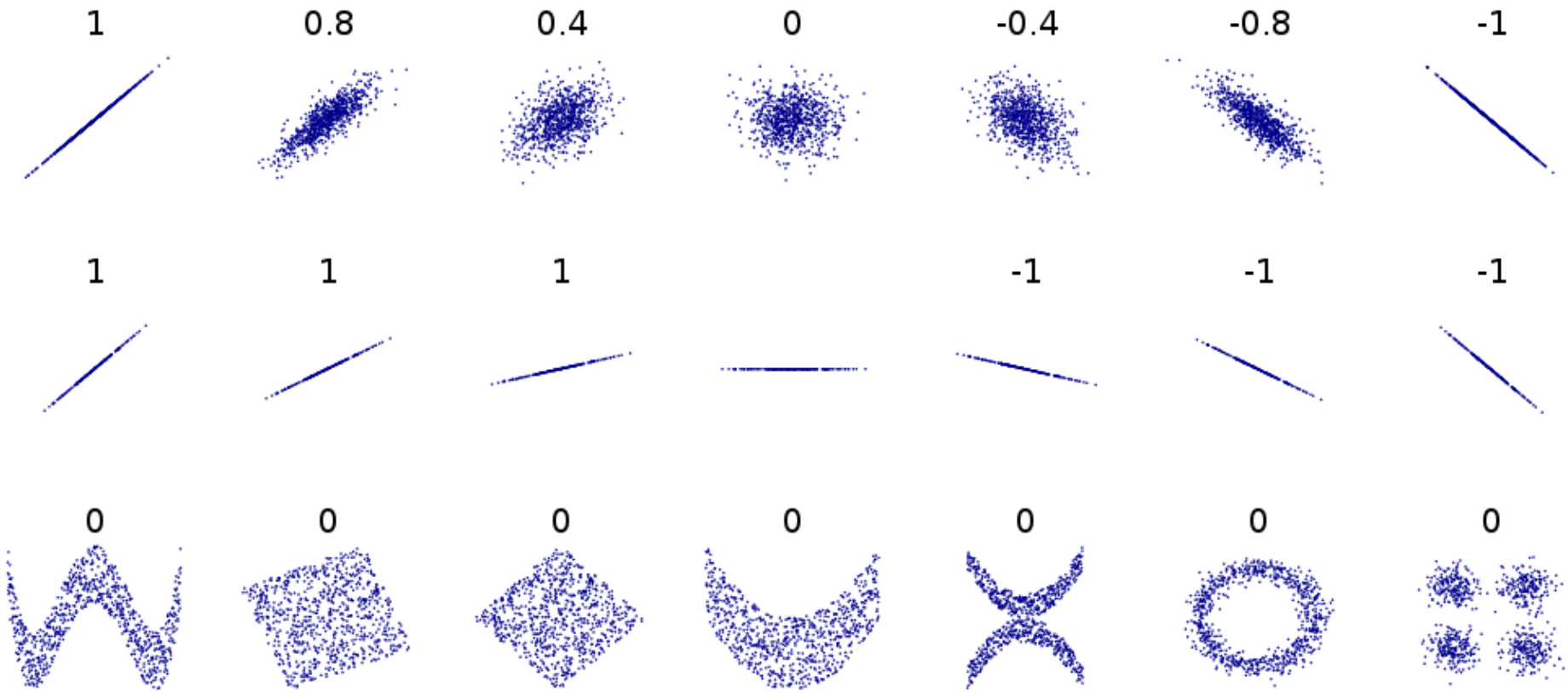
- Skript: Kapitel 5.1, 5.2 lesen
- Serie 11 lösen
- Quiz 11 bearbeiten
- etutoR Lektion 9 (**!!!**)

Zusammenhang zwischen 2 Zufallsvariablen (ZV)



Korrelation: Linearer Zusammenhang, in [-1, 1]

Repetition: Korrelation



Heute:

- Möglichst präziser Zusammenhang zwischen ZV und erklärenden Variablen (nicht zufällig)
- ➢ zweistufige, stochastische Modelle (*generalized linear models*)

1. Zielvariable hat Verteilung: $Y \sim \mathcal{F}(\theta)$
2. Parameter der Verteilung ist eine Funktion:

$$\theta = f(x_1, x_2, \ldots, x_p)$$

Bisher war nur θ nur irgendein Wert, z.B. $n = 10, p = 0.3$

Big Picture: Generalized Linear Models (GLMs)

- **bisher:** Population wird mit einer Verteilung beschrieben
  - Bsp: Medikament wirkt mit 30% Wahrscheinlichkeit.
  - ➢ Wie wahrscheinlich ist es, dass von 10 mind. 5 Pat. gesund werden?
- **neu:** Population wird mit einer Verteilung beschrieben, die von einem (oder mehreren) Parametern abhängt
  - Wirkwahrscheinlichkeit hängt von Dosis ab.
  - ➢ Bei welcher Dosis werden im Mittel 90% der Pat. gesund?

Generalized Linear Models:

- Zusammenhang zwischen erklärenden Variablen (z.B. Dosis) und Parametern einer Verteilung (z.B. Erfolgsw'keit p der Binomialverteilung).

Beispiel 1: Wirkung eines Medikaments

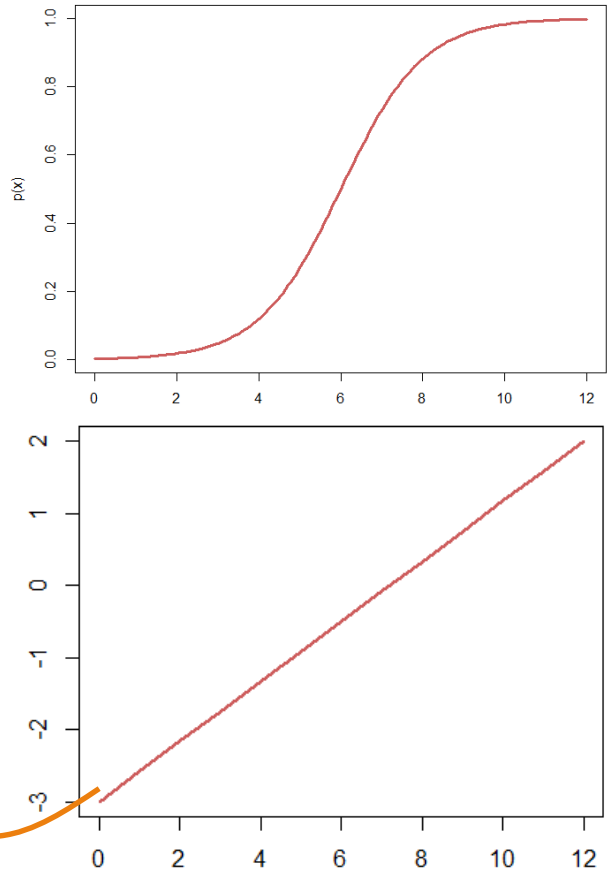
- X: Dosis des Wirkstoffs,
  n: # Patienten,
  p: Genesungsw'keit,
  Y: # gesunder Patienten nach Behandlung
- $Y \sim \mathrm{Bin}\big(n, p(x)\big)$
- Zusammenhang von p und x:

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

- …kann man umformen zu:

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

Linear in den β's



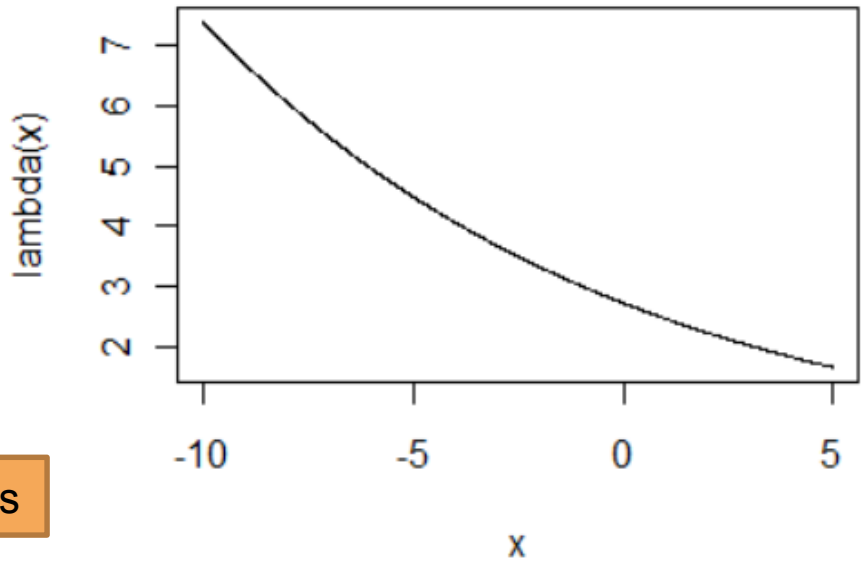
- **"Logistische Regression"** (Binomialregression)

"Bei welcher Dosis ist die Genesungsw'keit 80%"

Beispiel 2: Anzahl Autounfälle im Winter

- X: Temperatur in Grad Celsius
  Y: # Autounfälle pro Tag in ZH
- $Y \sim Pois\big(\lambda(x)\big)$
- Zusammenhang λ und x:
  $\lambda(x) = \exp(\beta_0 + \beta_1 x)$
- …kann man umformen zu:
  $\log\big(\lambda(x)\big) = \beta_0 + \beta_1 x$

Linear in den β's



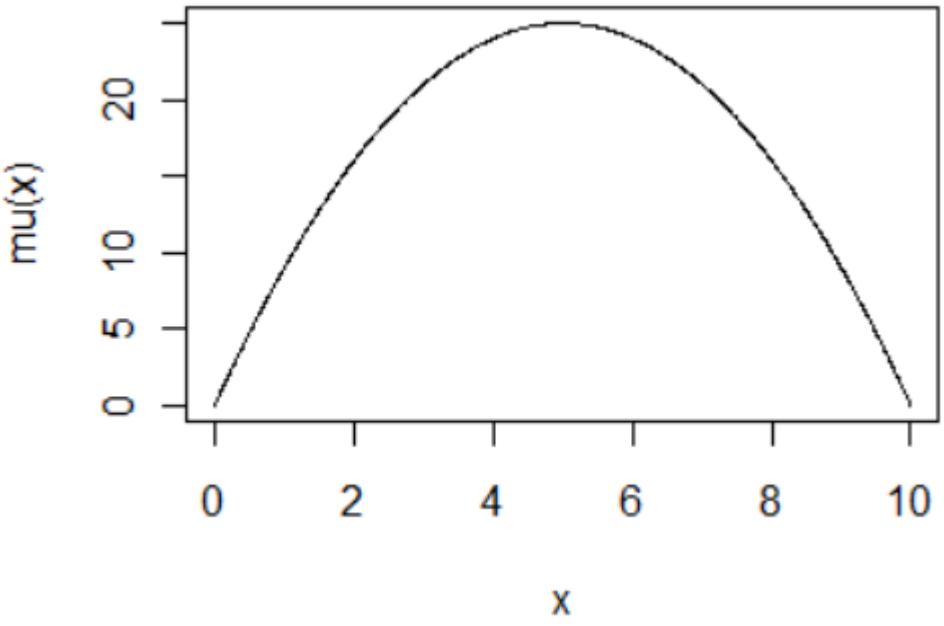
- **"Poissonregression"**

"Morgen wird es -5°C. Was ist das 95%-Quantil der Unfälle morgen?"

Beispiel 3: Kraftzuwachs bei Training

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

Linear in den β's



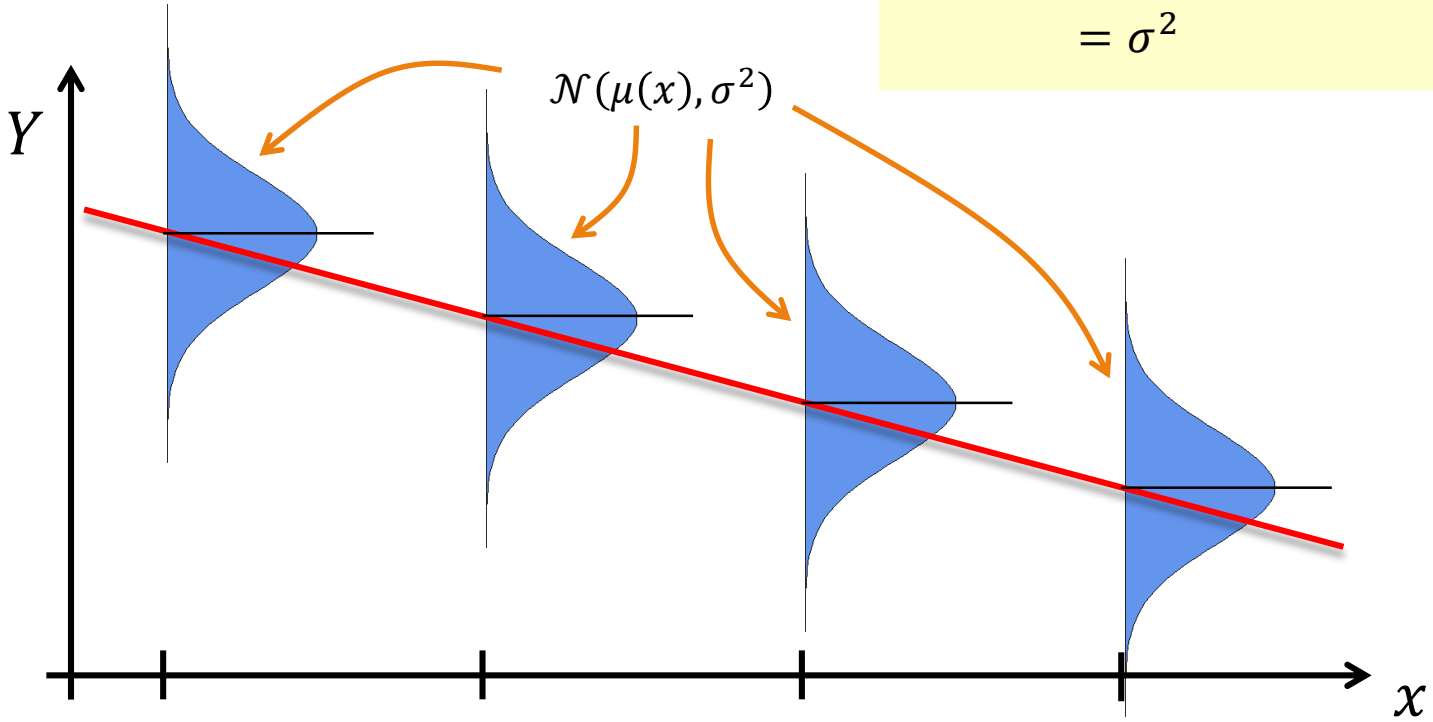
- **"Lineare Regression"**
  - Einfache Lineare Regression
    - $\mu(x) = \beta_0 + \beta_1 x$: eine Erklärende
  - Multiple Lineare Regression
    - $\mu(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$

"Welche Trainingsdauer pro Woche bringt optimalen Kraftzuwachs?"

Lineare Regression: Zwei Definitionen

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

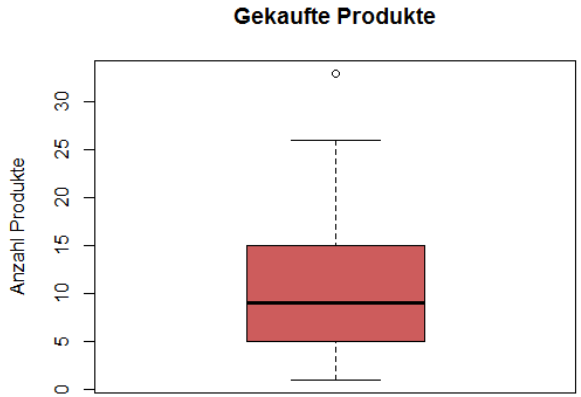
$$
\begin{aligned}
E(Y) &= E(\beta_0 + \beta_1 x + \varepsilon) \\
&= \beta_0 + \beta_1 x + E(\varepsilon) \\
&= \beta_0 + \beta_1 x \\
Var(Y) &= Var(\beta_0 + \beta_1 x + \varepsilon) \\
&= Var(\varepsilon) \\
&= \sigma^2
\end{aligned}
$$

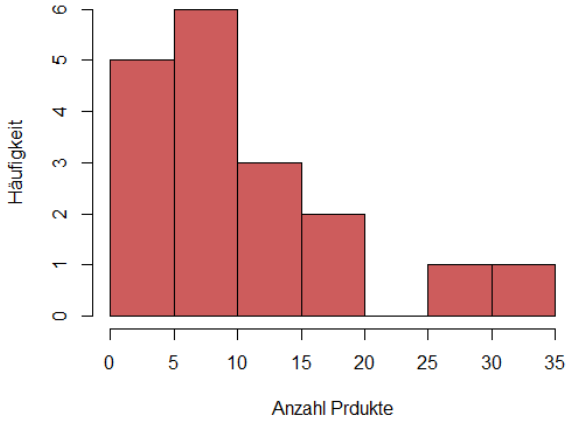


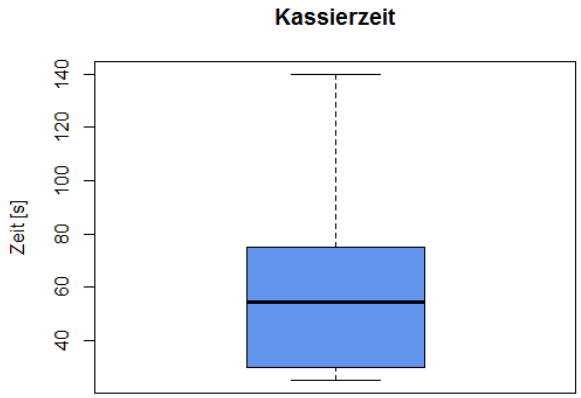

LIFE

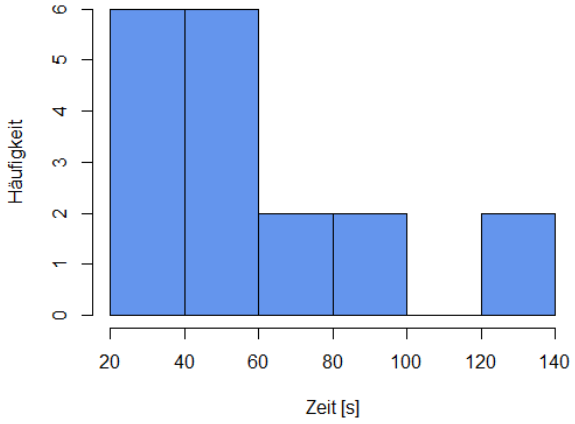
Wo anstehen?

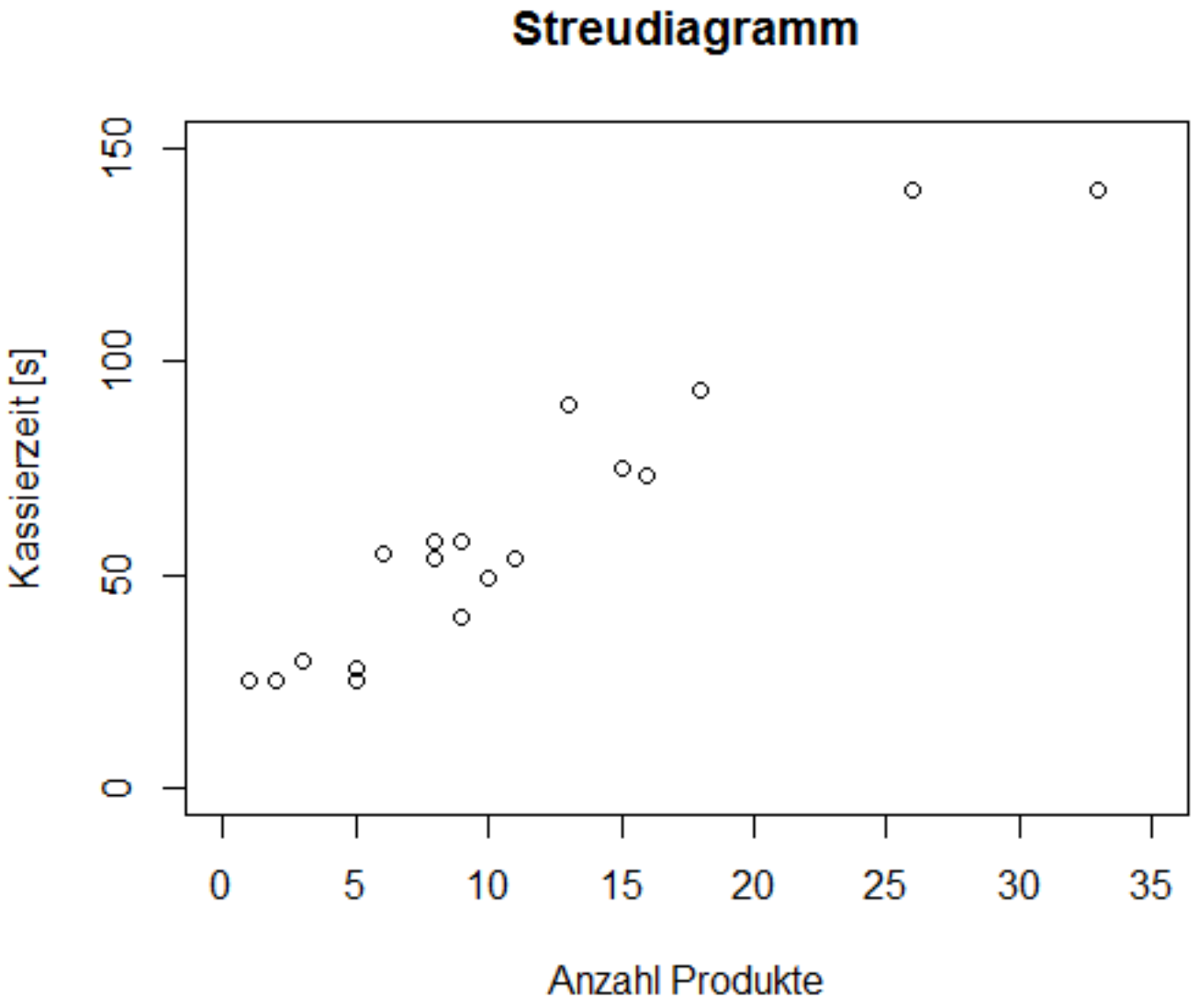
Coop Zürich HB
Eine Kassiererin, 17:40 – 18:00

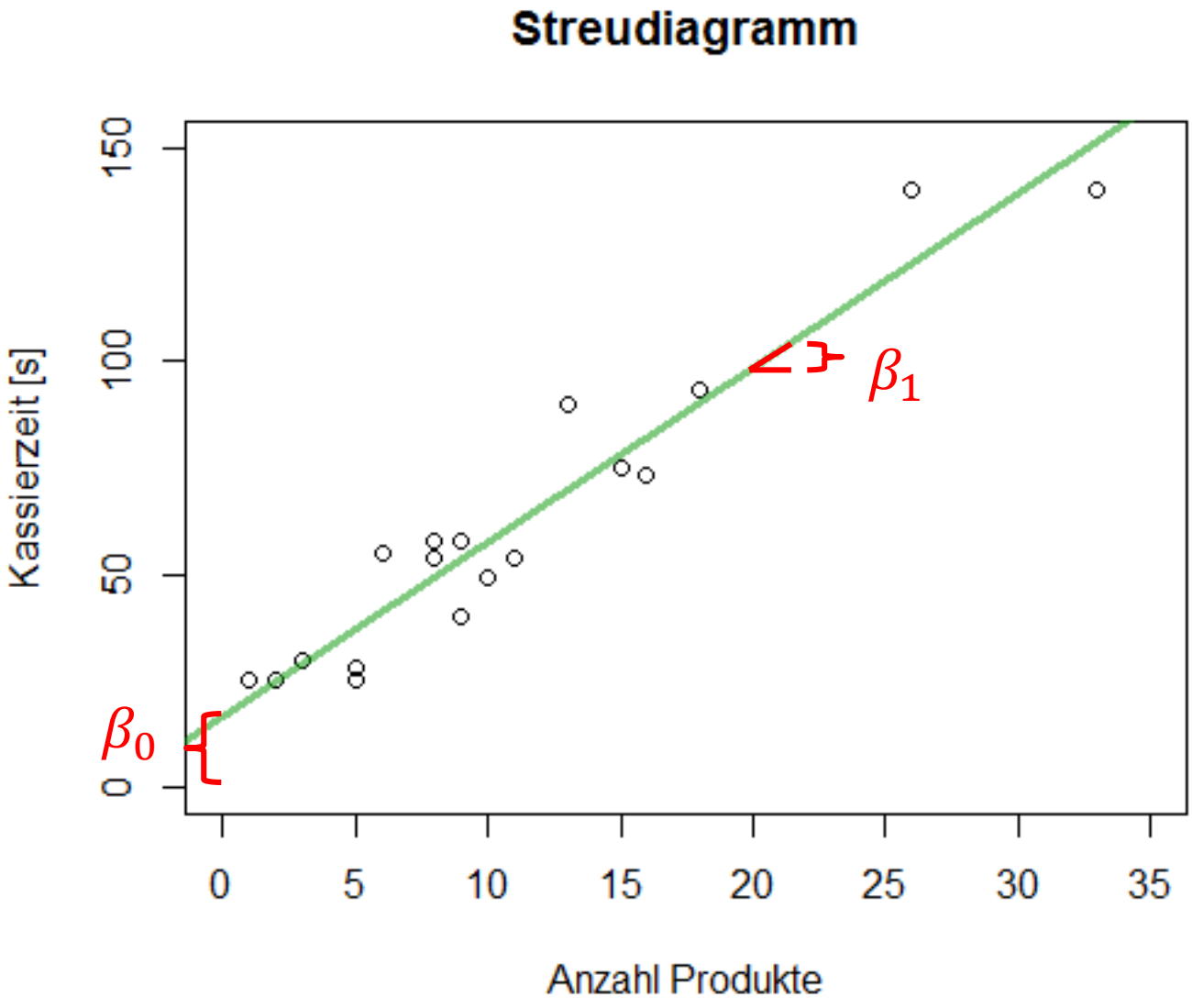


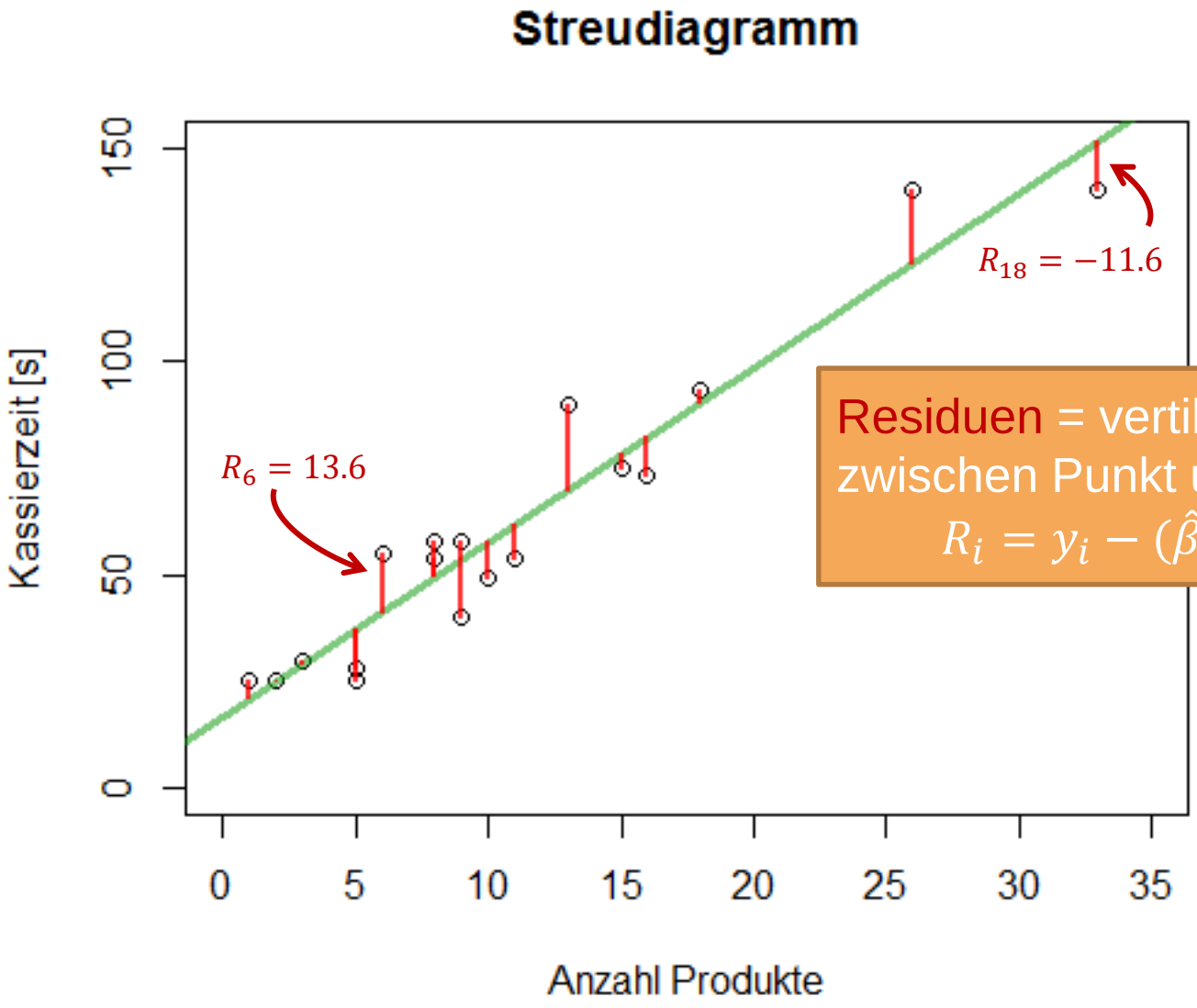







Streudiagramm
Kassierzeit [s]
0
50
100
150
0
5
10
15
20
25
30
35
Anzahl Produkte


Streudiagramm
Kassierzeit [s]
0
50
100
150
0
5
10
15
20
25
30
35
Anzahl Produkte
β_0
β_1



Residuen = vertikaler Abstand zwischen Punkt und Linie

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

Parameterschätzung – Variante 1
Methode der kleinsten Quadrate ("Least Squares")

- Welche Gerade passt am besten zu den Punkten?
- Wähle $\hat{\beta}_0, \hat{\beta}_1$ so, dass die Summe der quadrierten Residuen minimal ist:

$$\hat{\beta}_0, \hat{\beta}_1 \text{ minimieren } \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

- Lösung im Skript, p. 85
- wir schauen uns gleich ein grafisches Beispiel an, aber zuerst noch …

Parameterschätzung: Variante 2
Maximum Likelihood Methode (ML)

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

$$\ell(\beta_0, \beta_1) = \log\big(\mathcal{L}(\beta_0, \beta_1)\big) =$$

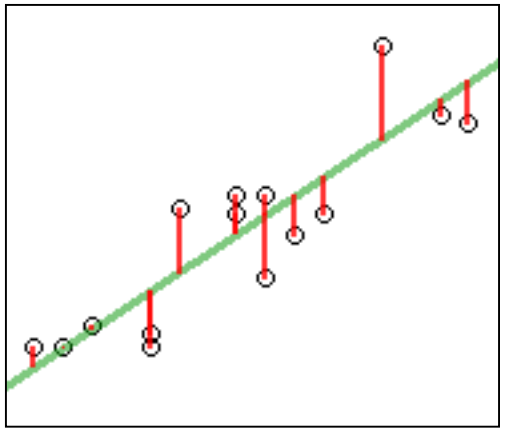
$$= -n\pi\sigma^2 - \frac{1}{2}\frac{\left(\sum_{i=1}^{n}\big(y_i - \mu(x_i)\big)^2\right)}{\sigma^2}$$

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

- log-Likelihood ist maximal, wenn $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$ minimal ist
- ML Methode ist äquivalent zu der Methode der kleinsten Quadrate

http://www.shodor.org/interactivate/activities/Regression/

Parameterschätzung, σ^2

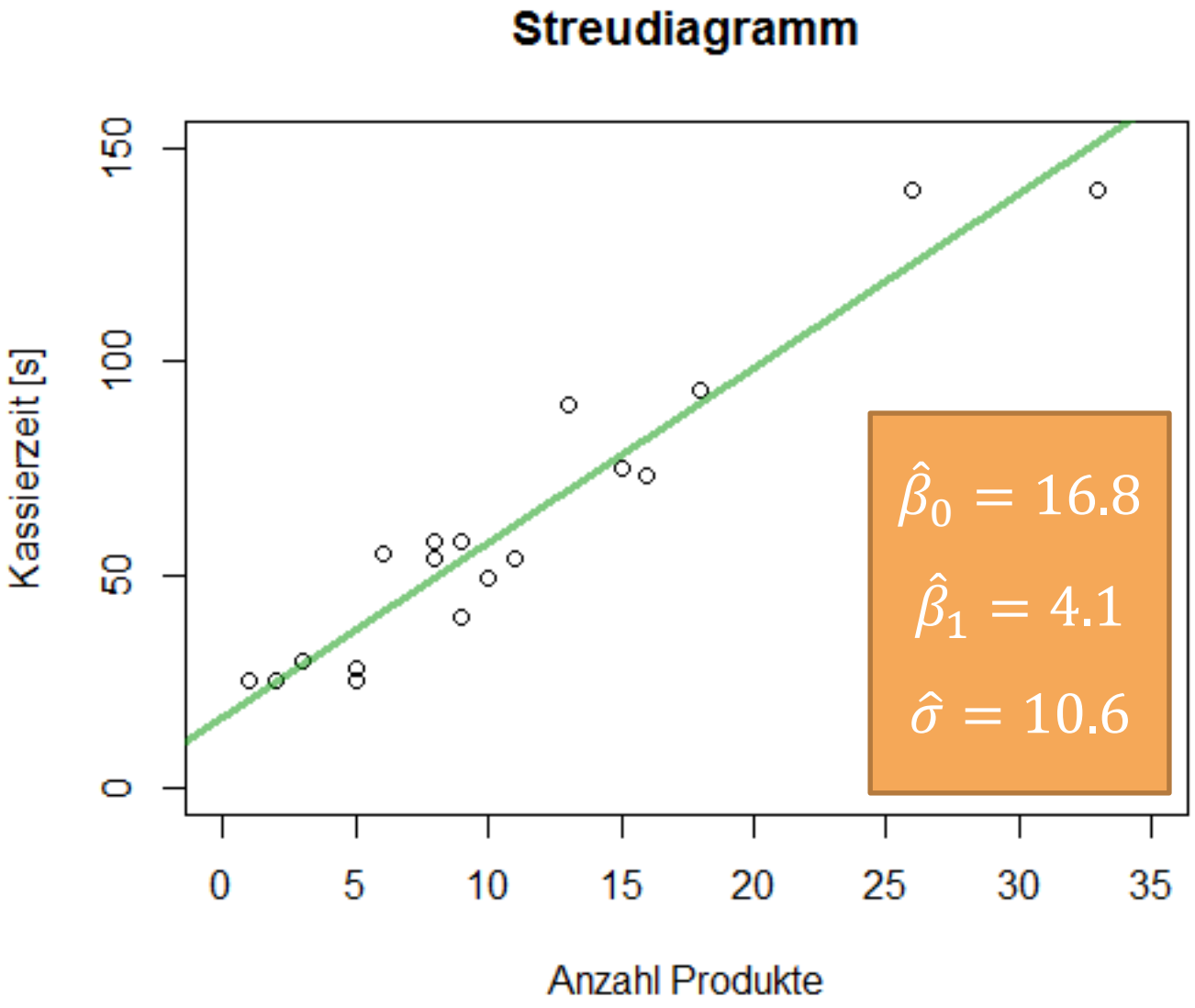


- Die Residuen erhalten wir mit:

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

- Und die Schätzung für die Varianz mit:

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

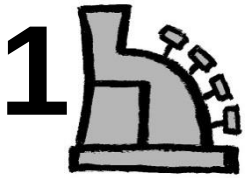

Streudiagramm
Kassierzeit [s]
0
50
100
150
0
5
10
15
20
25
30
35
Anzahl Produkte
$\hat{\beta}_0 = 16.8$
$\hat{\beta}_1 = 4.1$
$\hat{\sigma} = 10.6$

Wo anstehen?

$\hat{\beta}_0 = 16.8$

$\hat{\beta}_1 = 4.1$

$\hat{\sigma} = 10.6$

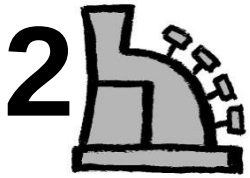


16.8 + **3**·4.1= 29.1 s

16.8 + **2**·4.1= 25 s

16.8 + **3**·4.1= 29.1 s

83.2



16.8 + **19**·4.1= 94.7 s

94.7

Test für β_0 und β_1

- X, Y sind ZV & $\hat{\beta}_0$, $\hat{\beta}_1$ sind Funktionen von X und Y
  $\Rightarrow$ $\hat{\beta}_0, \hat{\beta}_1$ sind auch ZV

Animation Lineare Regression:
http://stat.ethz.ch/~meier/teaching/animations/linreg/

Man kann zeigen, dass $\hat{\beta}_i \sim \mathcal{N}(\beta_i, \sigma^2_{\hat{\beta}_i})$, mit…

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

…und auch

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} \sim t_{n-2}$$

Aerobe Leistungsfähigkeit

- VO_2max: Menge Sauerstoff, die der Körper pro kg maximal pro Minute verwerten kann

- Test ist teuer und aufwändig
- nicht für breite Masse geeignet

- Alternative?

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

https://www.youtube.com/watch?v=Np2cnn3oOdE

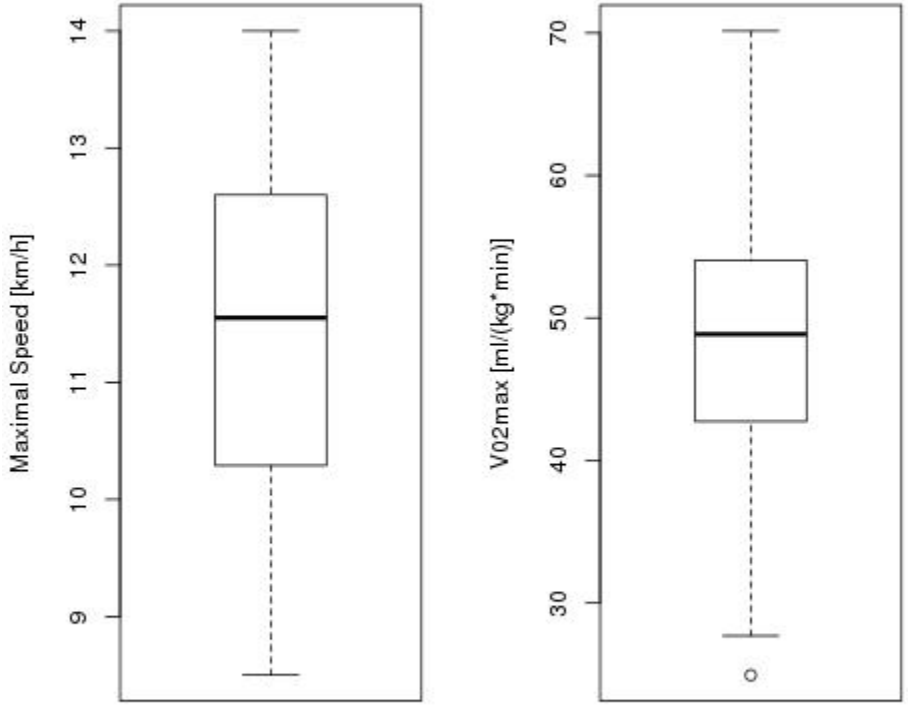
Ersatz: Cooper & Shuttle

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

- Kann Shuttle-Test den VO_2max-Wert vorhersagen?
- Falls ja: Einfache Testmöglichkeit für breite Bevölkerung

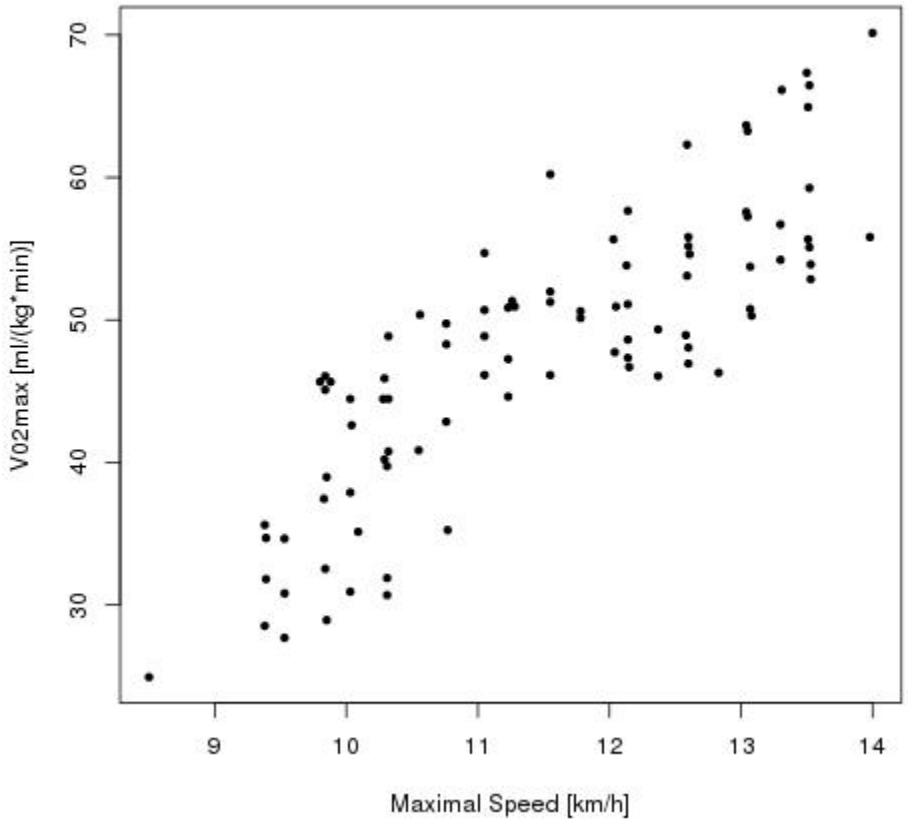
Léger et al., 1983

- 91 Personen, Shuttle-Test und VO_2max Messung



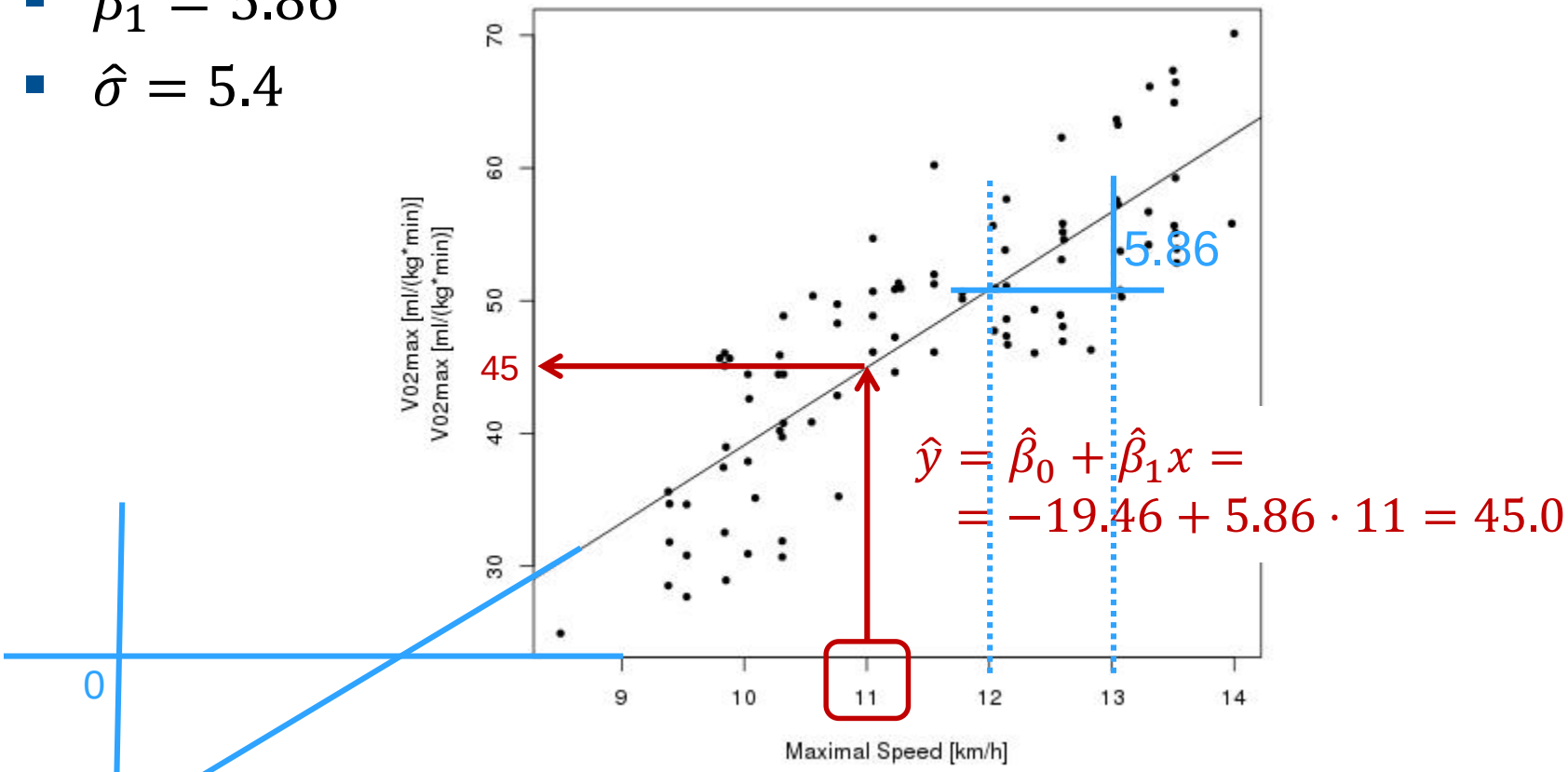
Streudiagramm VO_2max vs v_{max}

- Korrelation $r = 0.84$



Lineare Regression

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



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \\ = -19.46 + 5.86 \cdot 11 = 45.0$$

Lineare Regression in R

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

```
> 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
```

Standardfehler von $\widehat{\beta}_1$

approx. 95%-VI:

$$5.86 \pm 2 \cdot 0.41$$

exaktes 95%-VI:

$$5.86 \pm 1.99 \cdot 0.41$$

$t_{89}; 0.975$

Beobachtete Teststatistik t im Test:

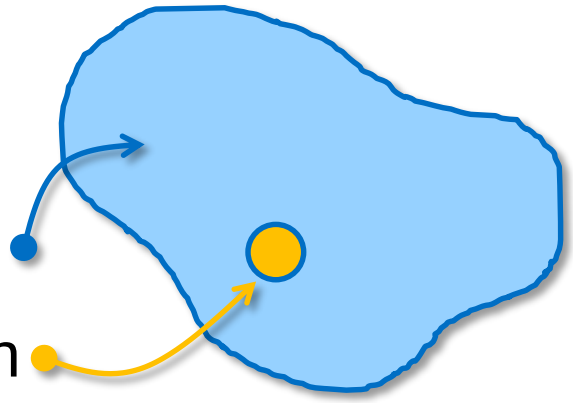
$\mathcal{H}_0: \beta_1 = 0$ vs $\mathcal{H}_A: \beta_1 \neq 0$

P-Wert:

Angenommen $\beta_1 = 0$; wie wahrscheinlich ist t oder etwas extremeres?

Freiheitsgrade: $n - (\text{Anzahl } \beta's) = 91 - 2 = 89$

Zusammenfassung



- Idee: GLM – logistisch Regression
- Details: Einfach lineare Regression
  VO_2max vorhersagen

Hausaufgaben

- Skript: Kapitel 5.1, 5.2 lesen
- Serie 11 lösen
- Quiz 11 bearbeiten
- etutoR Lektion 9 (**!!!**)