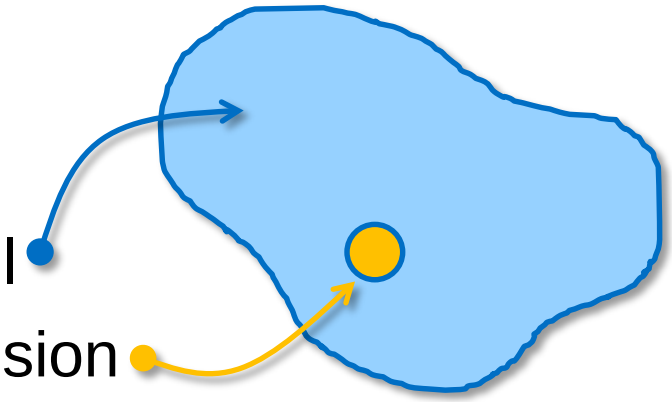
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

DMATH
Department of Mathematics

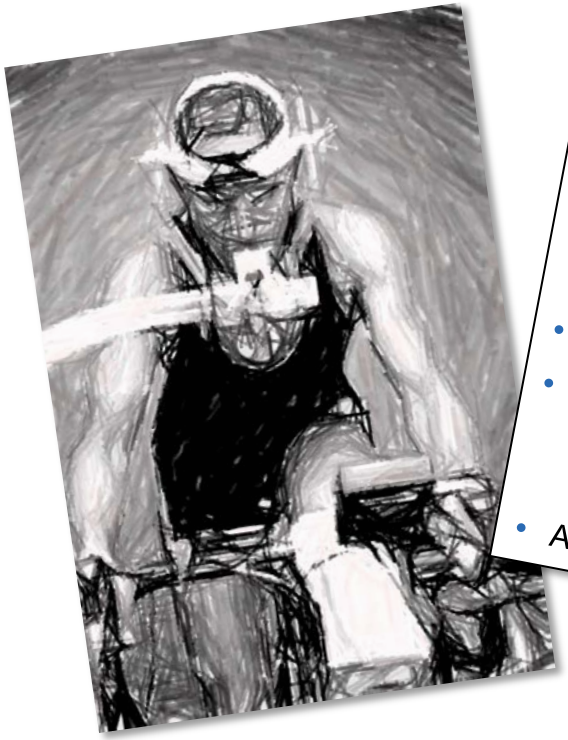
Einfache lineare Regression: Tests und Residuenanalyse

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

Lernziele vorletzte Woche



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



- 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?

Einfache lineare Regression

- $Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \quad iid, i = 1, \dots, n$

$$\hat{\beta}_0, \hat{\beta}_1 \text{ minimieren } \sum_{i=1}^{n} \underbrace{\left(Y_i - (\beta_0 + \beta_1 x_i)\right)}_{\textbf{Residuen } \boldsymbol{R_i}}^2$$

: "*Methode der kleinsten Quadrate*"

- $\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$: Steigung der Regressionsgeraden
- $\widehat{\beta_0} = \bar{y}_n - \widehat{\beta_1}\bar{x}_n$: y-Achsenabschnitt (*engl.* intercept)
- $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n} R_i^2$: Residual Standard Error

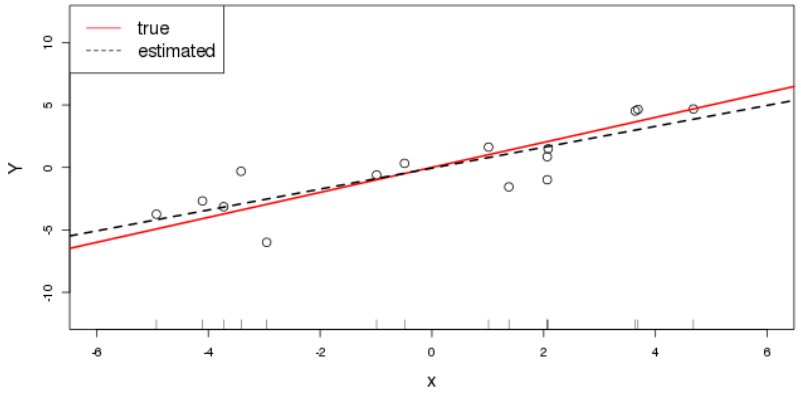
$n - p$ Feiheitsgrade (*engl*. degree of freedom), $p \mathrel{\widehat{=}} \#\beta$'s (hier: $p = 2$)

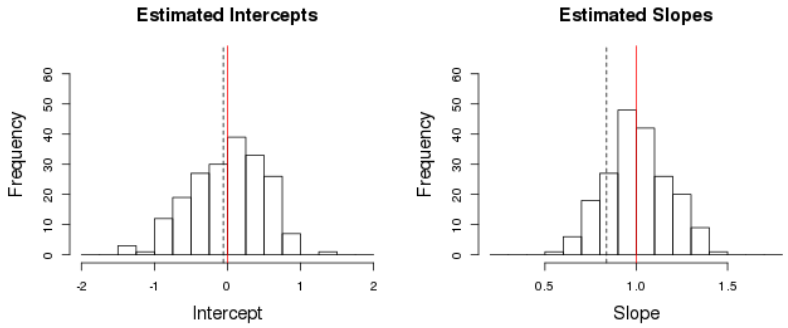
β_j's sind Zufallsvariablen

- $\widehat{\beta_j} \sim \mathcal{N}\left(\beta_j, s.e.(\widehat{\beta_j})^2\right)$
  - $E(\hat{\beta}_0) = \beta_0$
  - $E(\hat{\beta}_1) = \beta_1$
  - Standardfehler:
    - $s.e.(\hat{\beta}_0) = \sqrt{\frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x})^2}}$
    - $s.e.(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{\sum (x_i - \bar{x})^2}}$
  - und man kann zeigen ($k = 0, 1$):
    - $\frac{\widehat{\beta_k} - \beta_k}{\widehat{s.e.}(\widehat{\beta_k})} \sim t_{n-2}$

Standardabweichung der Schätzung

Feiheitsgrade von vorhin...





Beispiel: 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

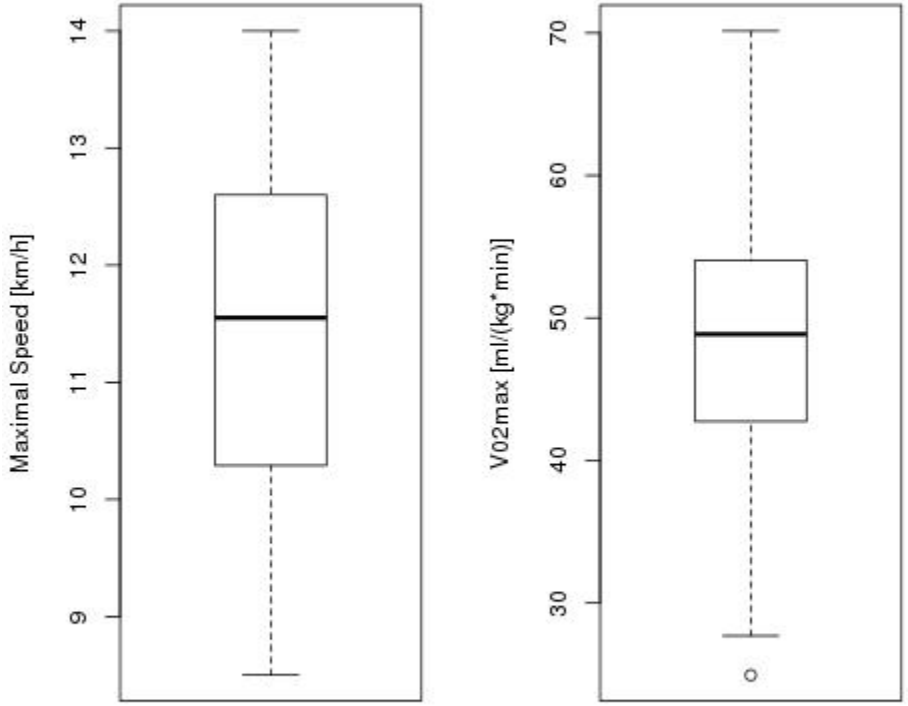
Ersatz: Cooper & Shuttle

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

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

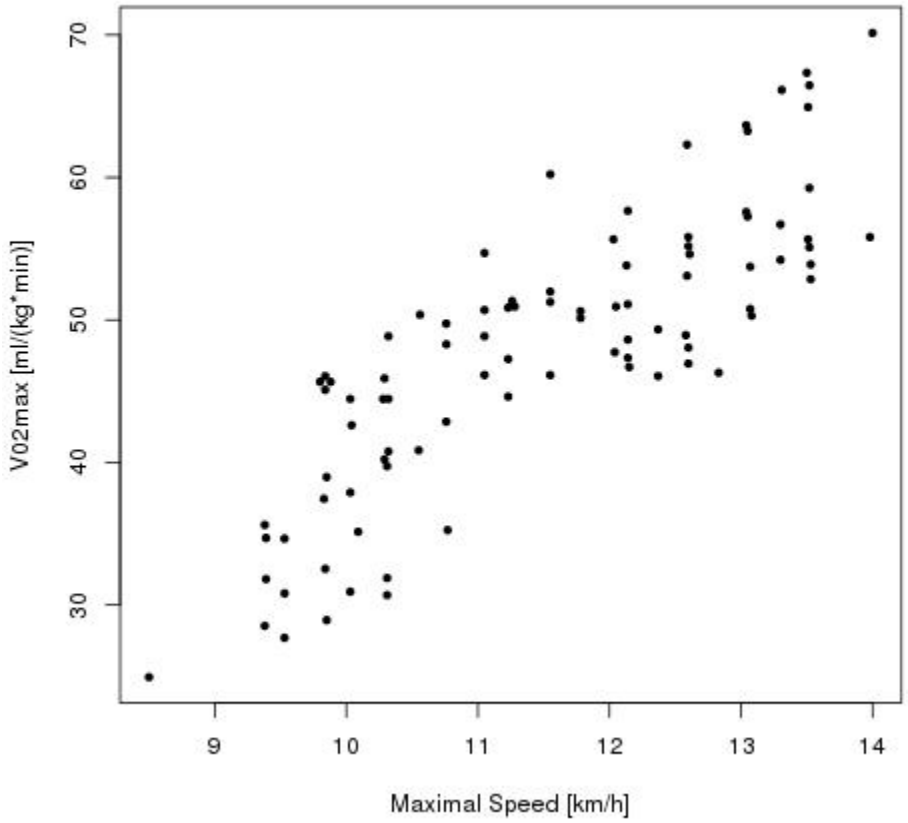
Léger et al., 1983

- 91 Personen, Shuttle-Test und VO_2max Messung

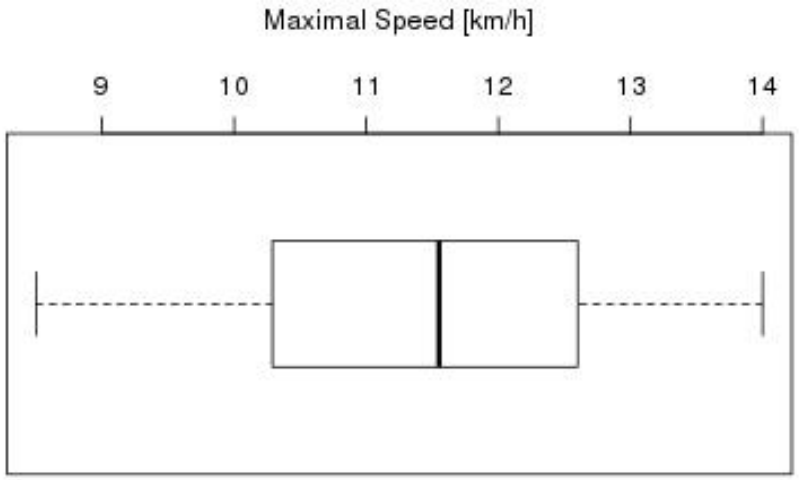


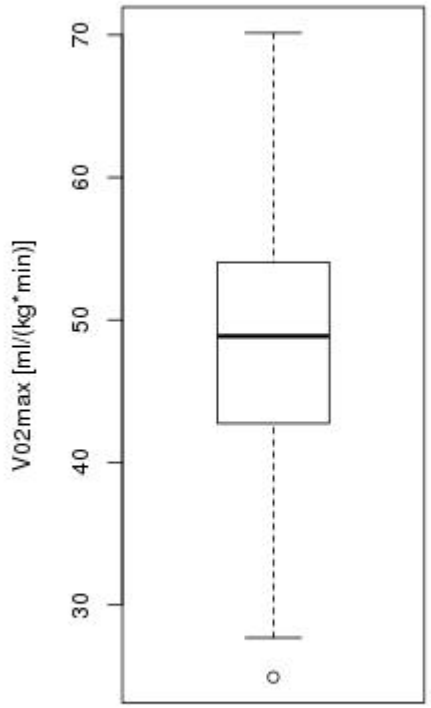
Streudiagramm VO_2max vs v_{max}

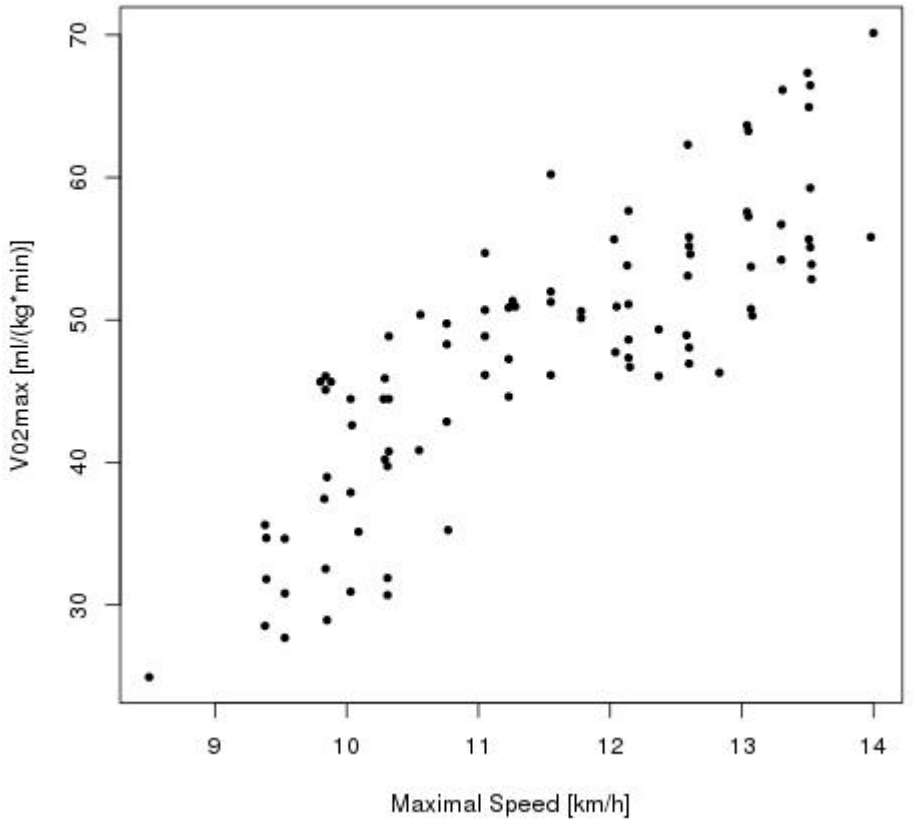
- Korrelation $r = 0.84$



Zusammenhang von Streudiagramm und Boxplots

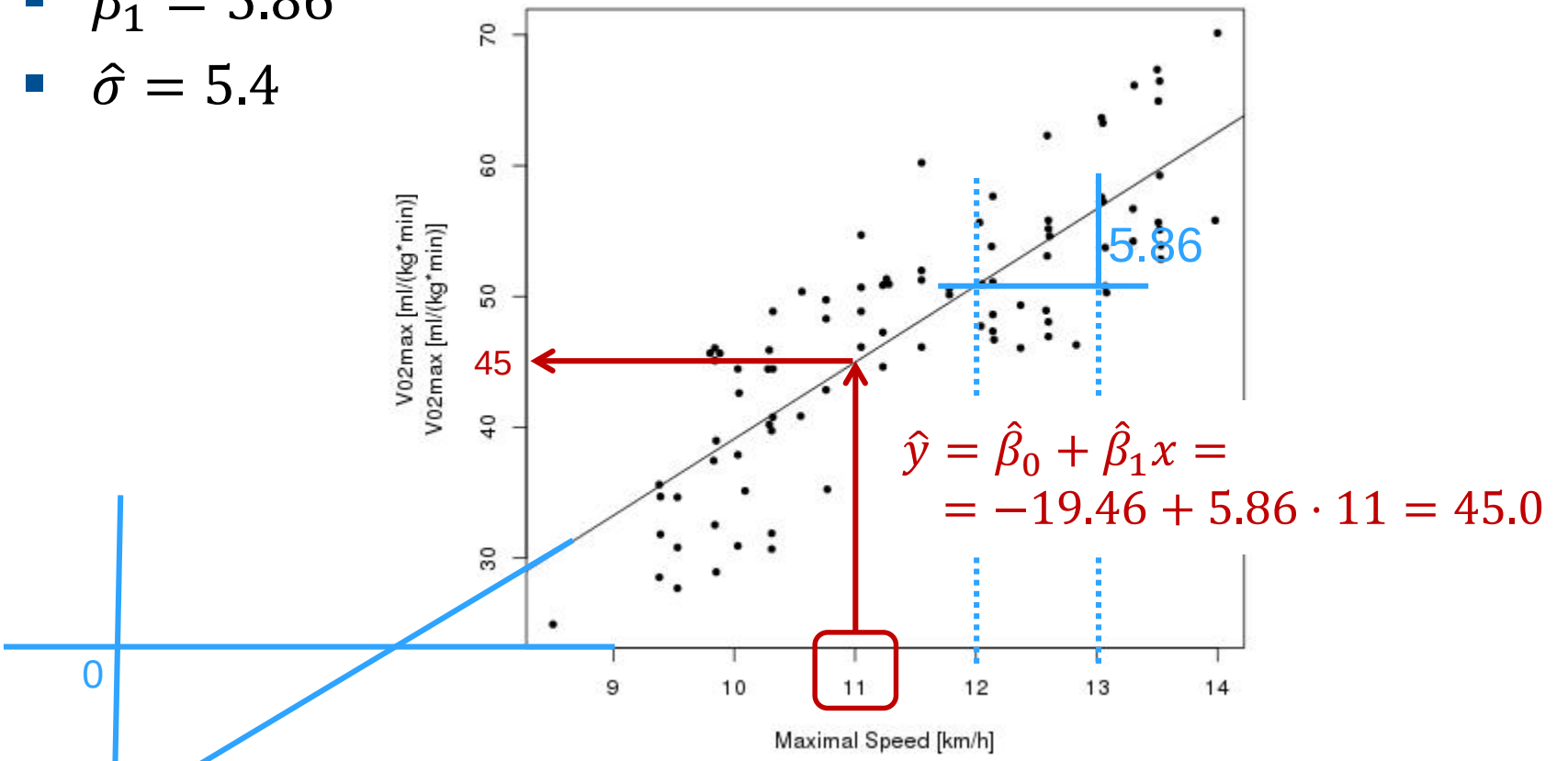






Lineare Regression

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



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

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

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

Lernziele heute

- Tests/Intervalle für die β_j's
- Intervalle für die y_i's
- Residuenanalyse (SD/TA, QQ)

Hausaufgaben

- Skript: Kapitel 5.2 fertig lesen
- Serie 12 lösen
- Quiz 12 bearbeiten
- etutoR Lektion 9 (vorletzte Woche...)

t-Test in der linearen Regression (1/2)

1. **Modell**:

$$Y_i = \beta_0 + \beta_1 x_i + E_i \; mit \; E_1, , \ldots, E_n \; i.i.d. \mathcal{N}(0 \; \sigma^2)$$

2. **Nullhypothese**: $\mathcal{H}_0 : \beta_1 = 0$

   Alternative: $\mathcal{H}_A : \beta_1 \neq 0$ (in der Regel zweiseitig)

3. **Teststatistik**:

$$T = \frac{beobachtet - erwartet}{geschätzter \; Standardfehler} = \frac{\hat{\beta}_1 - 0}{\widehat{s.e.}(\hat{\beta}_1)}$$

Dabei ist der geschätzte Standardfehler:

$$\widehat{s.e.}(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}$$

Verteilung der Teststatistik unter $\mathcal{H}_0$: $T \sim t_{n-2}$

t-Test in der linearen Regression (2/2)

4. **Signifikanzniveau**: α
5. **Verwerfungsbereich**:

$$\mathrm{K} = \left(-\infty, -t_{n-2;1-\frac{\alpha}{2}}\right] \cup \left[t_{n-2;1-\frac{\alpha}{2}}, \infty\right)$$

6. **Testentscheid**: Liegt der beobachtete Wert t der Teststatistik T in K?

Konfidenzintervall für β_j's

- Exaktes 95%-CI für ein β_j:

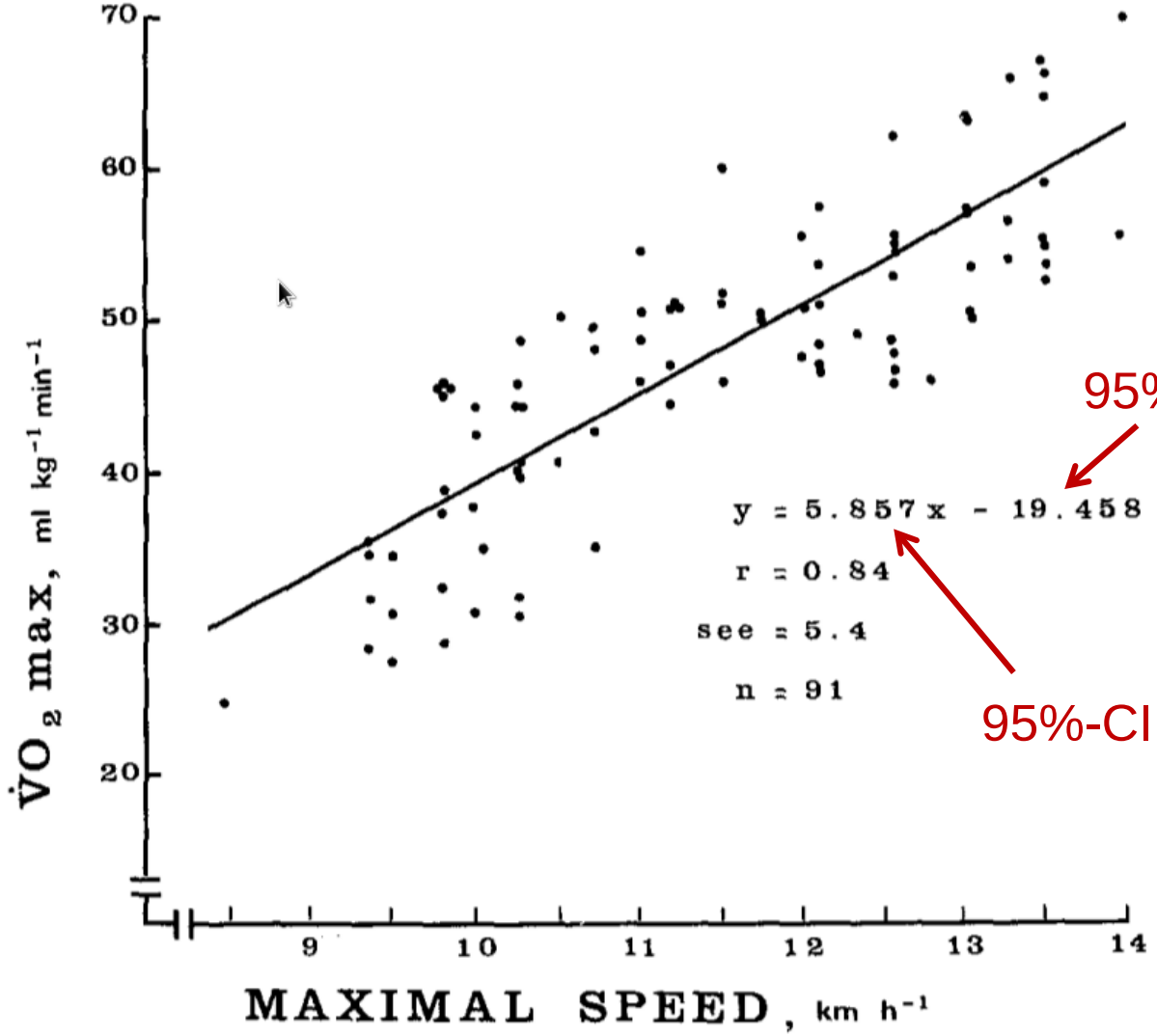
$$\left[\hat{\beta}_k - \widehat{s.e.}\,(\hat{\beta}_k) \cdot t_{n-2;1-\frac{\alpha}{2}}, \hat{\beta}_k + \widehat{s.e.}\,(\hat{\beta}_k) \cdot t_{n-2;1-\frac{\alpha}{2}}\right]$$

- Approximatives 95%-CI für ein β_j:
  - Verwende statt dem genauen Quantil der t-Verteilung einfach 2

$$\left[\hat{\beta}_k - 2 \cdot \widehat{s.e.}\,(\hat{\beta}_k), \hat{\beta}_k + 2 \cdot \widehat{s.e.}\,(\hat{\beta}_k)\right]$$

```
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 ***
```

  - $CI_{vmax} = [5.9 - 2 \cdot 0.41, 5.9 + 2 \cdot 0.41] = [5.08, 6.72]$




```
> confint(fit)
                 2.5 %     97.5 %
(Intercept) -28.844465 -10.071898
vmax          5.045479   6.667657
```

95%-CI: [-28.8, -10.1]

95%-CI: [5.0, 6.7]

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

Konfidenzintervall versus Prognoseintervall für Y_i

- Konfidenzintervall (manchmal auch Vertrauensintervall)

Plausibler Bereich für $E(Y_i)$ gegeben X_i,

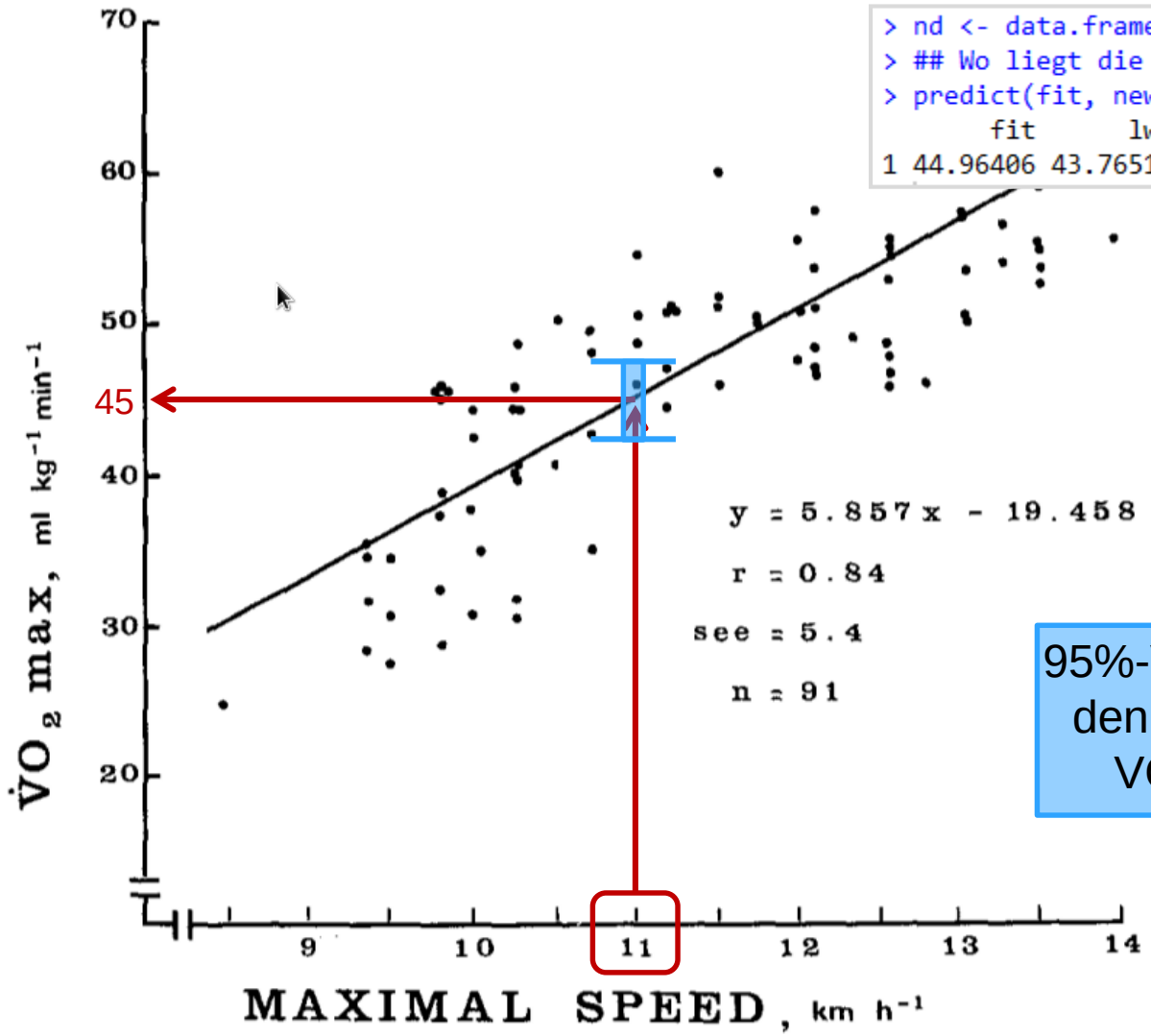
Unsicherheit wegen $\hat{\beta}_0$ und $\hat{\beta}_1$

- Prognoseintervall (manchmal auch Vorhersageintervall)

Plausibler Bereich für Y_i gegeben X_i,

Unsicherheit wegen $\hat{\beta}_0$, $\hat{\beta}_1$ und ε

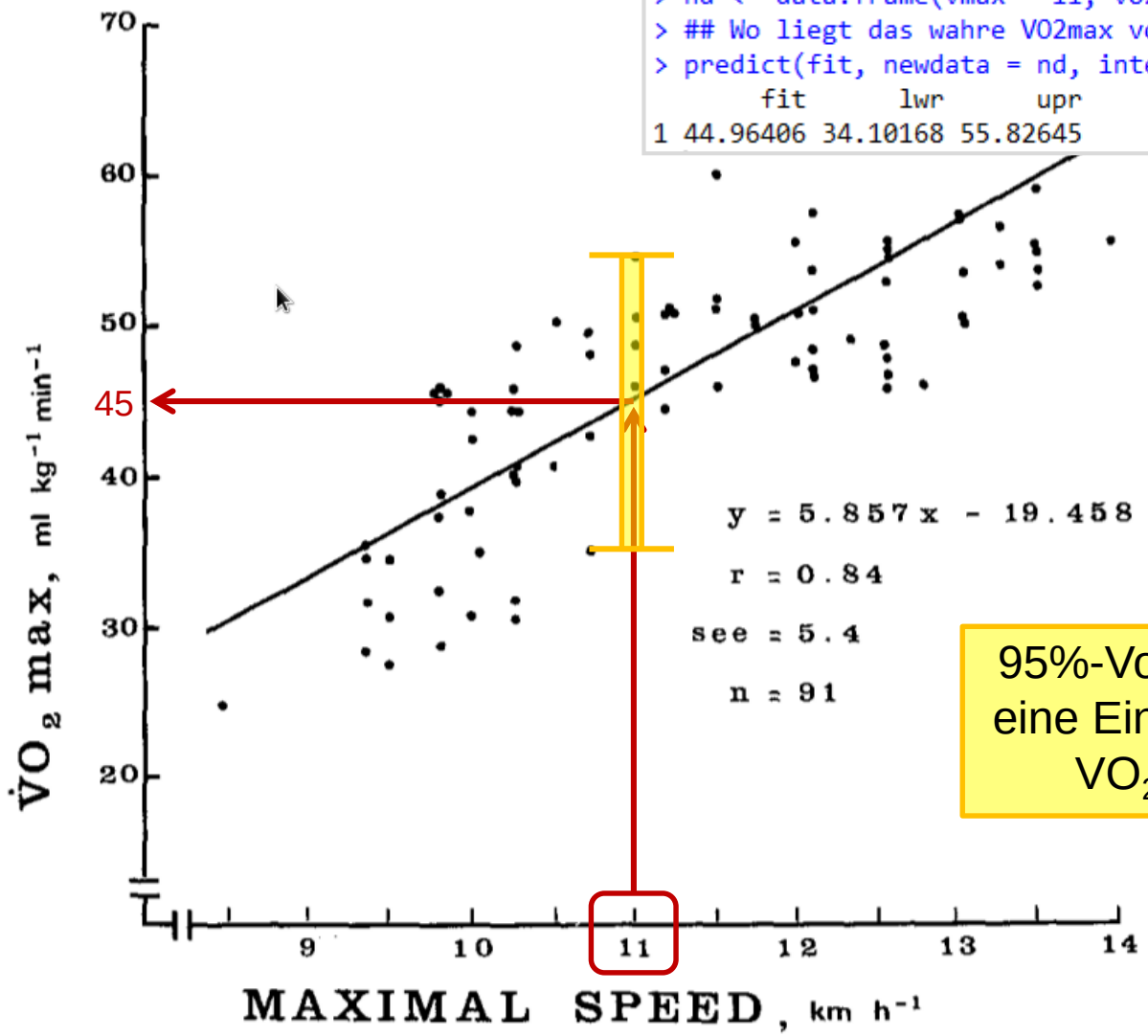
```
> nd <- data.frame(vmax = 11, vo2max = NA)
> ## Wo liegt die wahre Gerade bei x=11 mit 95% Wa.?
> predict(fit, newdata = nd, interval = "confidence")
       fit      lwr      upr
1 44.96406 43.76511 46.16302
```



95%-Vertrauensintervall für den Erwartungswert von VO_2max bei v_{max}=11

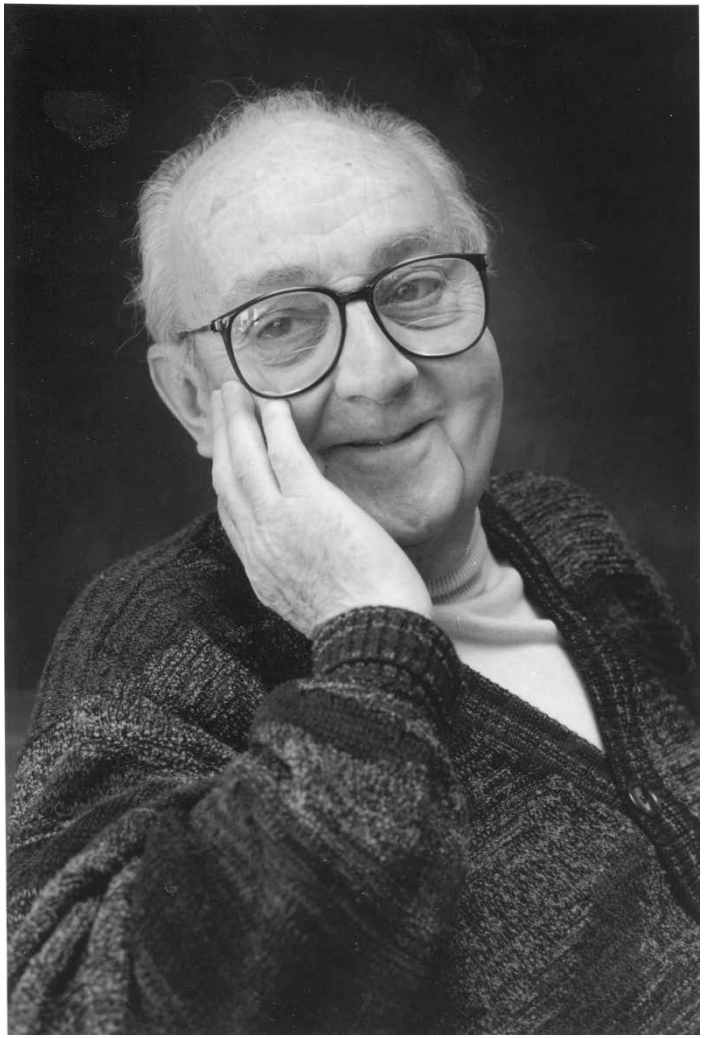
Fig. 2. $\dot{V}O_2$ max as a function of the maximal speed achieved in the 20-m shuttle run test for a total sample of 91 adult subjects. Each point in this figure represents maximal effort

```
> nd <- data.frame(vmax = 11, vo2max = NA)
> ## Wo liegt das wahre VO2max von einer Person mit x=11 mit 95% Wa.?
> predict(fit, newdata = nd, interval = "prediction")
       fit      lwr      upr
1 44.96406 34.10168 55.82645
```



95%-Vorhersageintervall für eine Einzelbeobachtung von VO_2max bei $v_{max}=11$

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



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

- George E.P. Box

Wie gut stimmt das Modell? - Residuenanalyse

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

- Form des funktionellen Zusammenhangs
- Varianz der Fehler ist konstant

- Fehler sind normalverteilt

Einfache Regression:

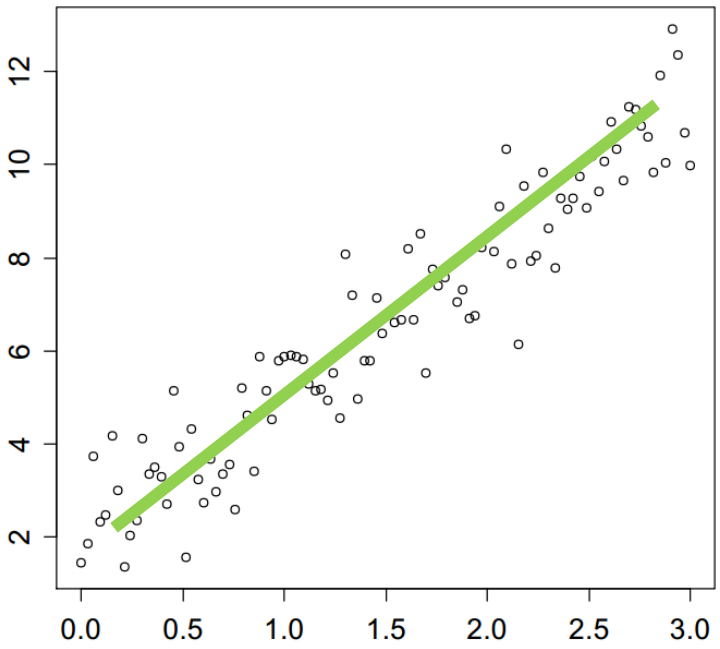
Streudiagramm (SD)

Multiple Regression:

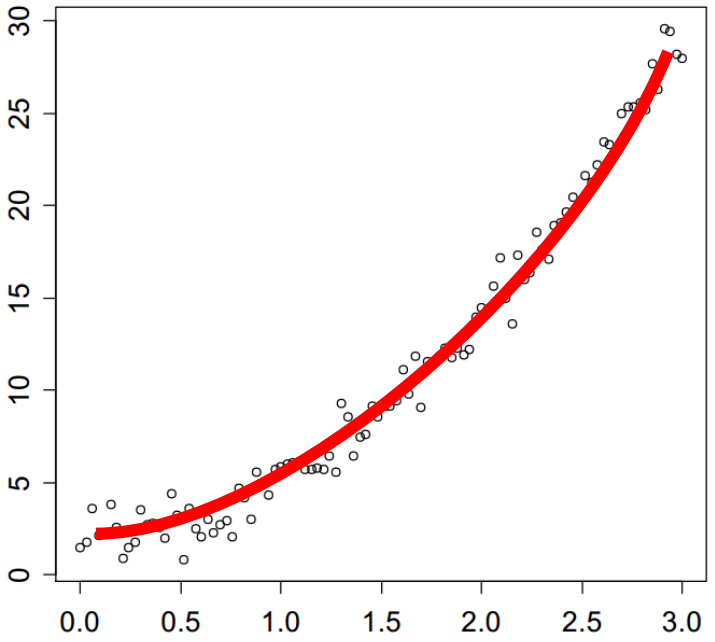
Tukey-Anscombe Plot (TA)

QQ-Plot der Residuen

Streudiagramm bei einfacher linearer Regression



OK

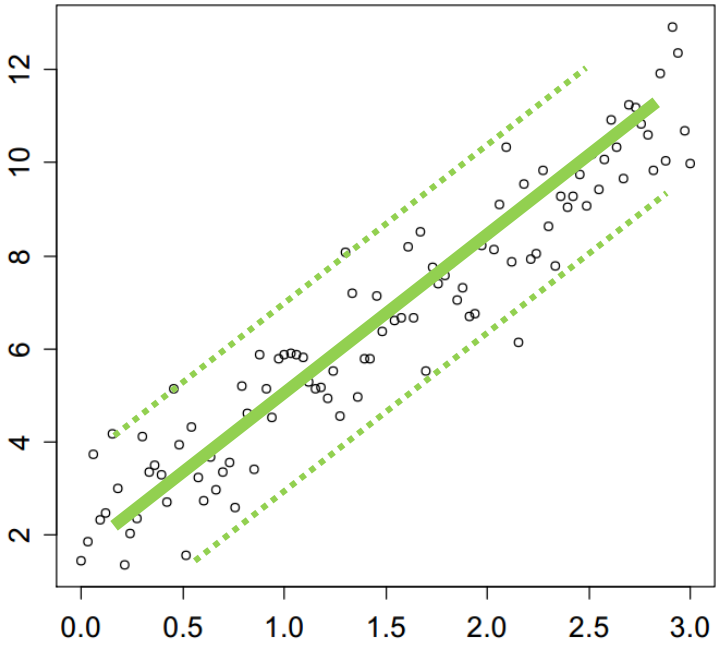


Systematischer Fehler

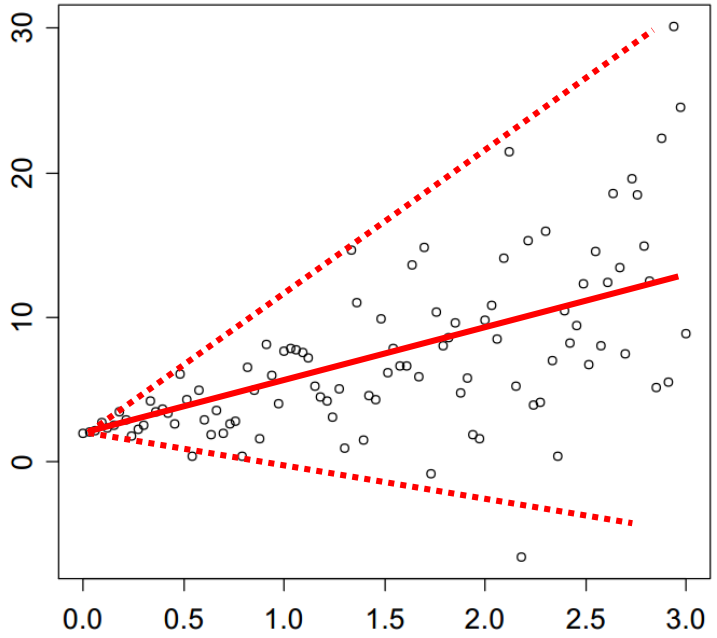
Krümmung:

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

Streudiagramm bei einfacher linearer Regression

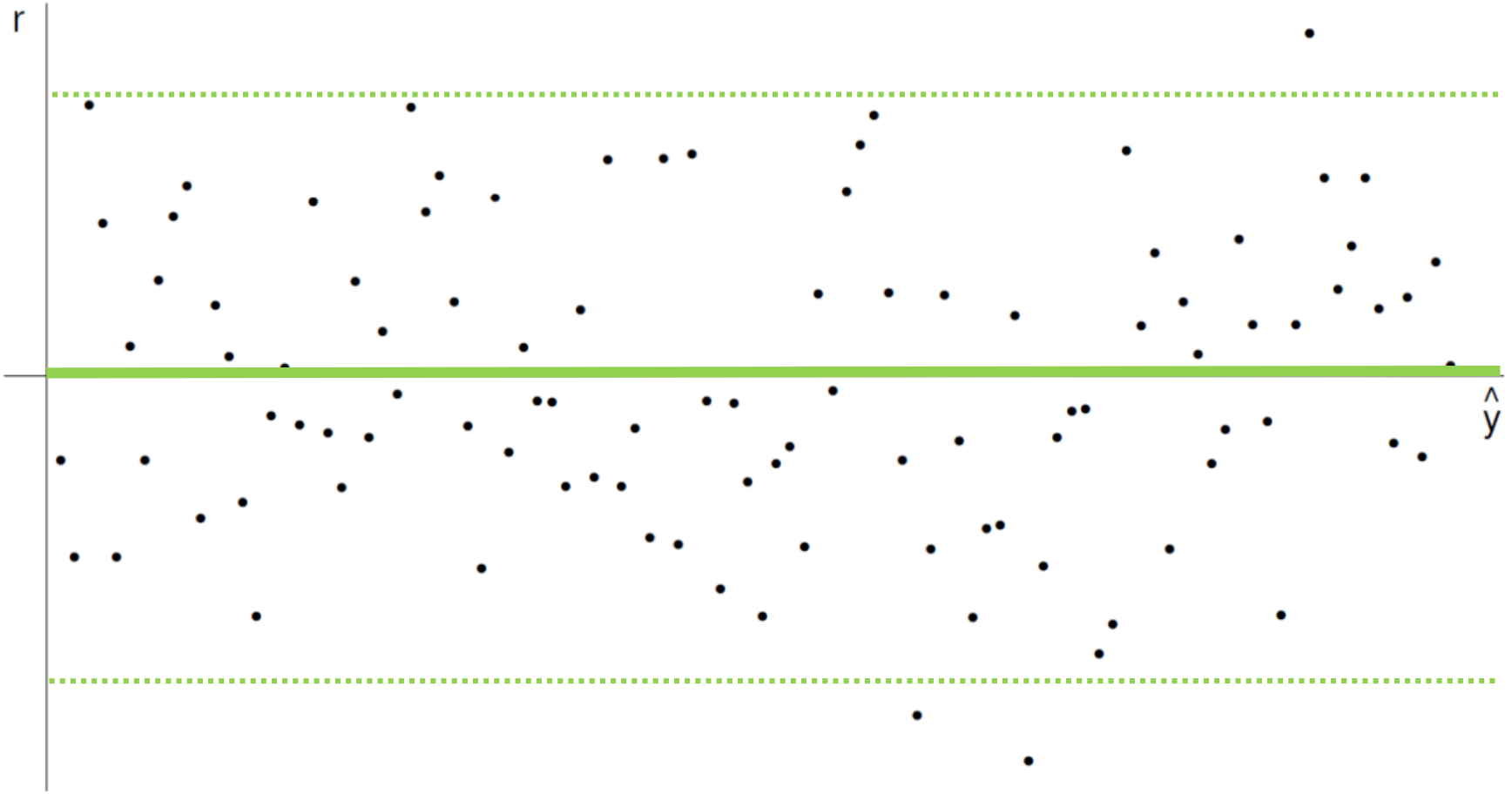


OK

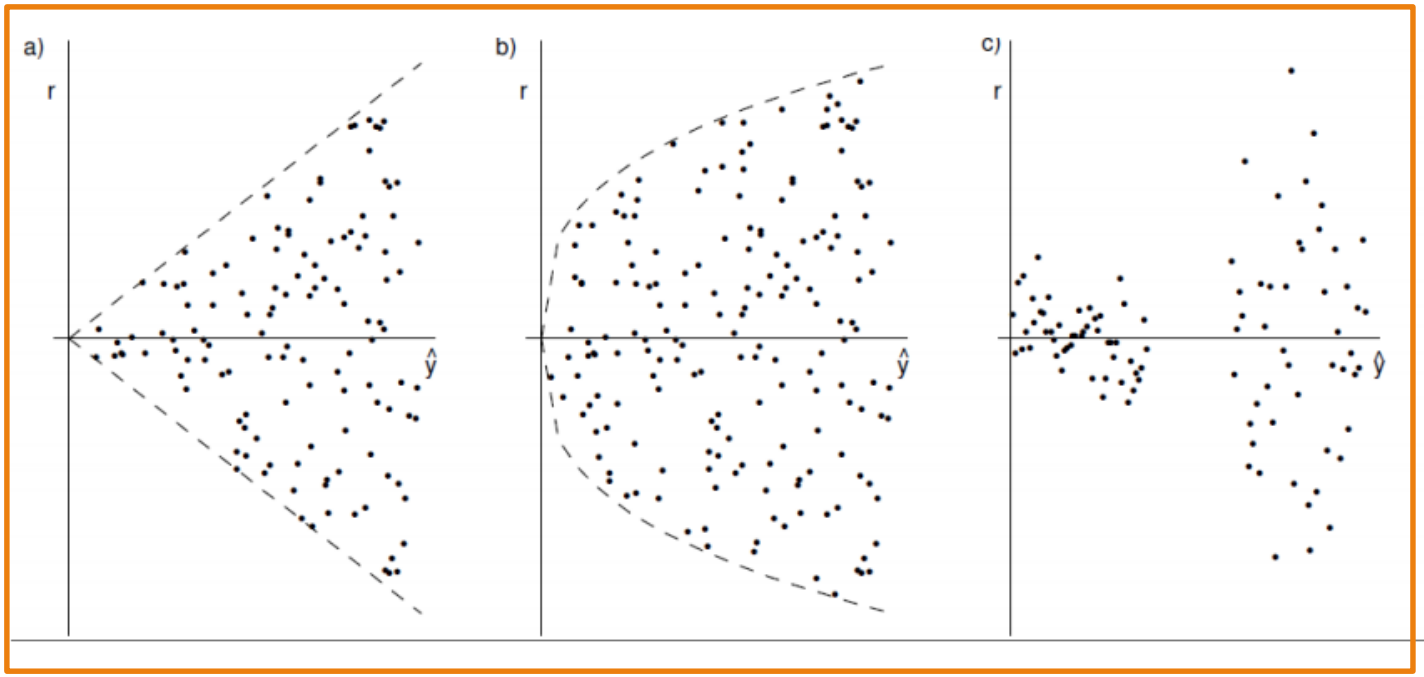


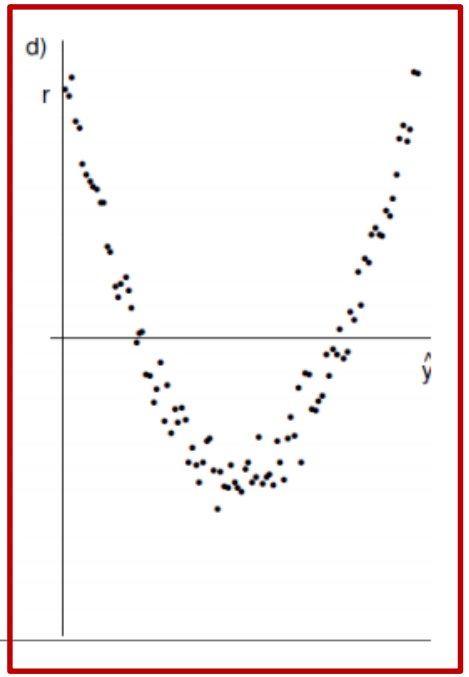
Fehlervarianz nicht konstant

Beispiel für guten Tukey-Anscombe Plot



Beispiele für schlechte Tukey-Anscombe Plots



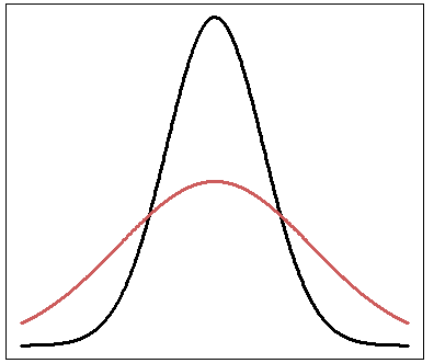


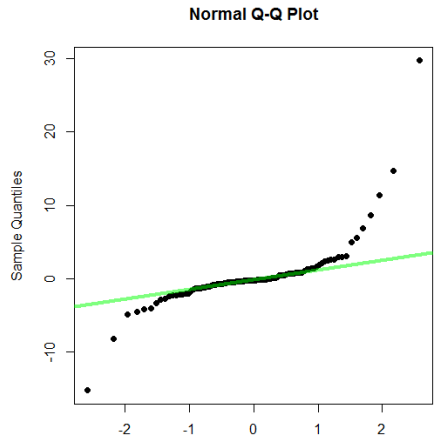
Fehlervarianz nicht konstant

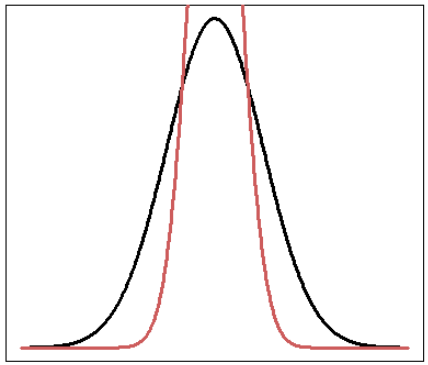
Systematischer Fehler

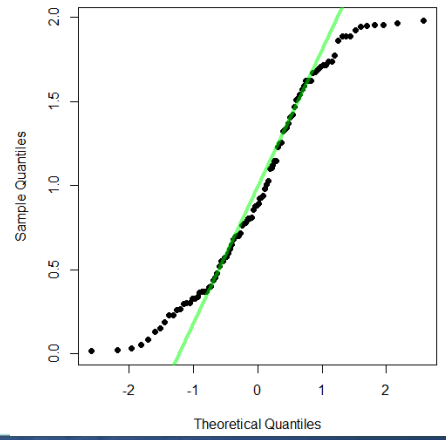
Residuenanalyse: QQ-Plot

- QQ steht für *quantile-quantile*

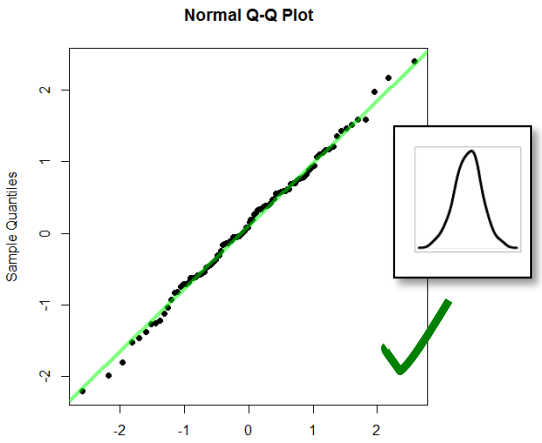




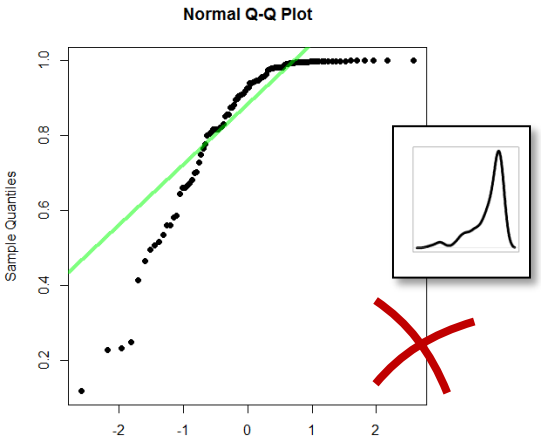




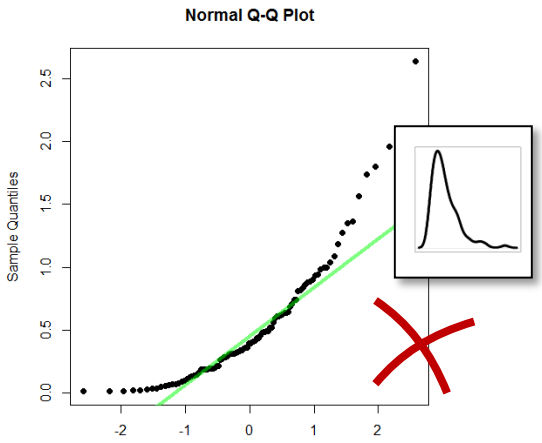
Residuenanalyse: QQ-Plot



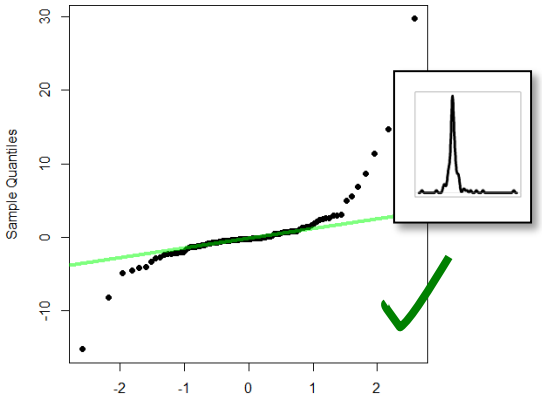
normal



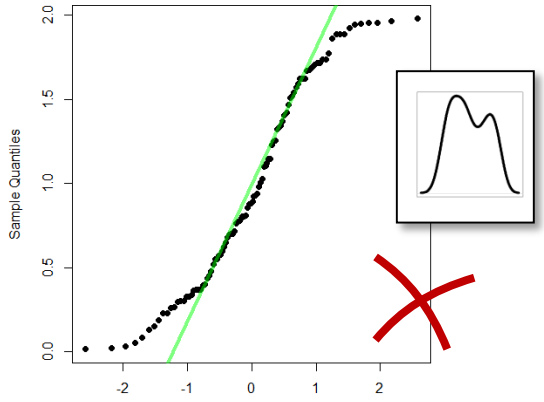
linksschief



rechtsschief

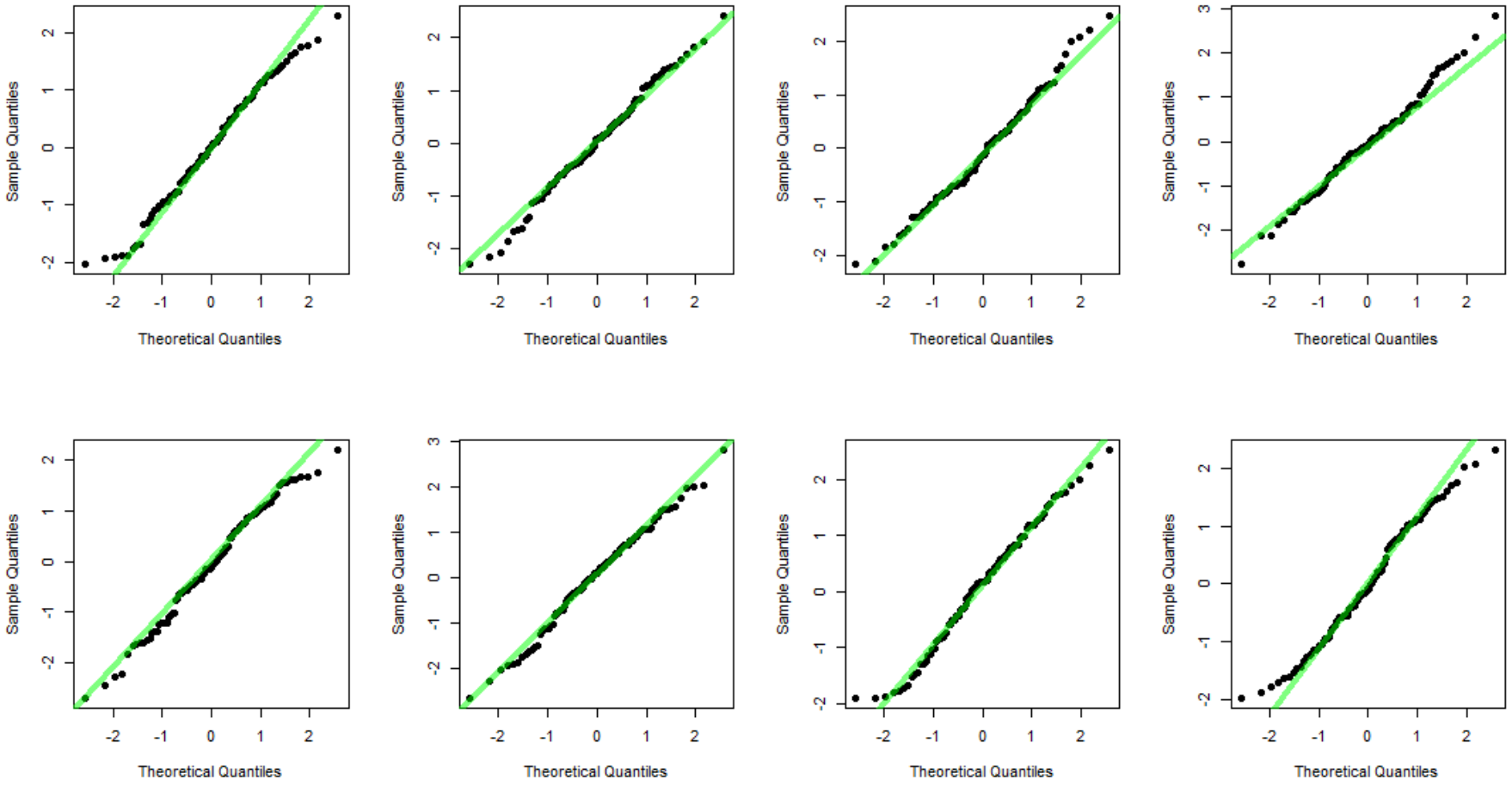


langschwänzig



kurzschwänzig

QQ-Plots: Streuung von «guten» QQ-Plots ($n = 100, R_i \sim \mathcal{N}(0,1)$)



Zusammenfassung

- Tests/Intervalle für die β_j's → macht der Koeffizient Sinn?
- Intervalle für die y_i's → Wo könnte ein VO_2max liegen?
- Residuenanalyse → Stimmt mein Modell?

Hausaufgaben

- Skript: Kapitel 5.2 fertig lesen
- Serie 12 lösen
- Quiz 12 bearbeiten
- etutoR Lektion 9 (vorletzte Woche...)