Biostatistische Studienplanung II

Dr. Matthias Kohl
SIRS-Lab GmbH

Inhalt

Lineare Modelle:

- **Definition und Beispiele**
- **KQ- und robuste Schätzer**
- **Diagnostik**
- **Ausblick: Mixed-Effects**

Definition des linearen Modells

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon$$

y: Beobachtungen (abhängige Variable)

$x_1, \ldots, x_m$: Regressoren (unabh. Variablen)

$\beta_1, \ldots, \beta_m$: unbekannte Parameter

ε: zufälliger Fehler

Beispiele für lineare Modelle

- $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$
- $y = \beta_0 + \beta_1 \log(x) + \varepsilon$
- $y = \beta_0 + \beta_1 \sin(2x) + \beta_2 \cos(x^2) + \varepsilon$

Linear bedeutet linear in den Parametern!

Kleinste-Quadrate Schätzer I

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{mi} + \varepsilon_i$$

Voraussetzungen an ε_i:

- Erwartungswert $E(\varepsilon_i) = 0$
- Homoskedastizität: $Var(\varepsilon_i) = \sigma^2 < \infty$
- Unkorreliertheit: $Cov(\varepsilon_i, \varepsilon_j) = 0 \ (i \neq j)$

Dann: KQ-Schätzer BLUE!

Kleinste-Quadrate Schätzer II

Zusätzlich:

- ε_i stochastische unabhängig
- ε_i normalverteilt

Dann: KQ-Schätzer auch ML-Schätzer!

Außerdem: Umfangreiche Testtheorie steht zur Verfügung!

Vorteile Linearer Modelle

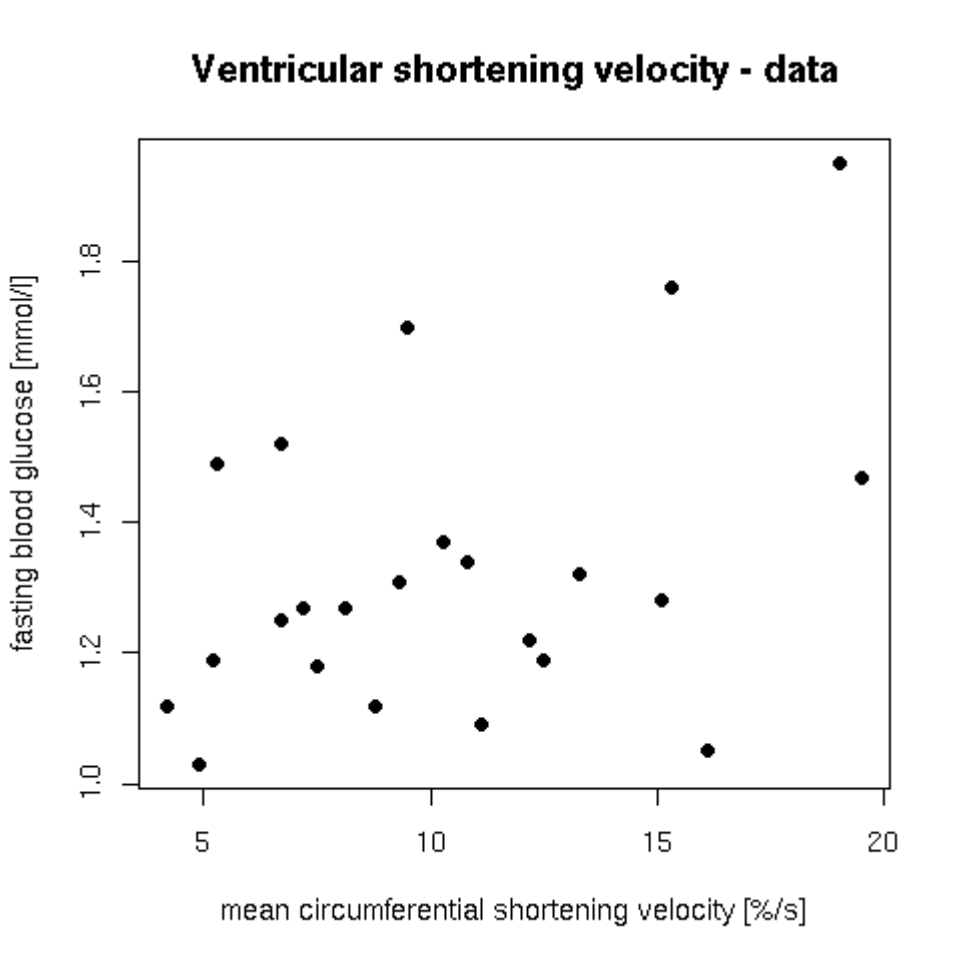
- Einfache, klare Struktur
- Nichtlineare Zusammenhänge sind lokal durch lineare Funktionen approximierbar
- Optimale Schätzer für große Klasse von Schätzern
- Theorie (nahezu) vollständig bekannt

Nachteile Linearer Modelle

- Schlechte Extrapolationseigenschaften (nur lokale Gültigkeit bzw. innerhalb der Daten)
- Große Anfälligkeit gegen Abweichungen von den Modellannahmen bzw. Ausreißern

Wichtig: Validierung des Modells!

Beispiel 1*



* D.G. Altman (1991), Practical Statistics for Medical Research, Table 11.6, Chapman & Hall.

Beispiel 1 – KQ-Schätzer I

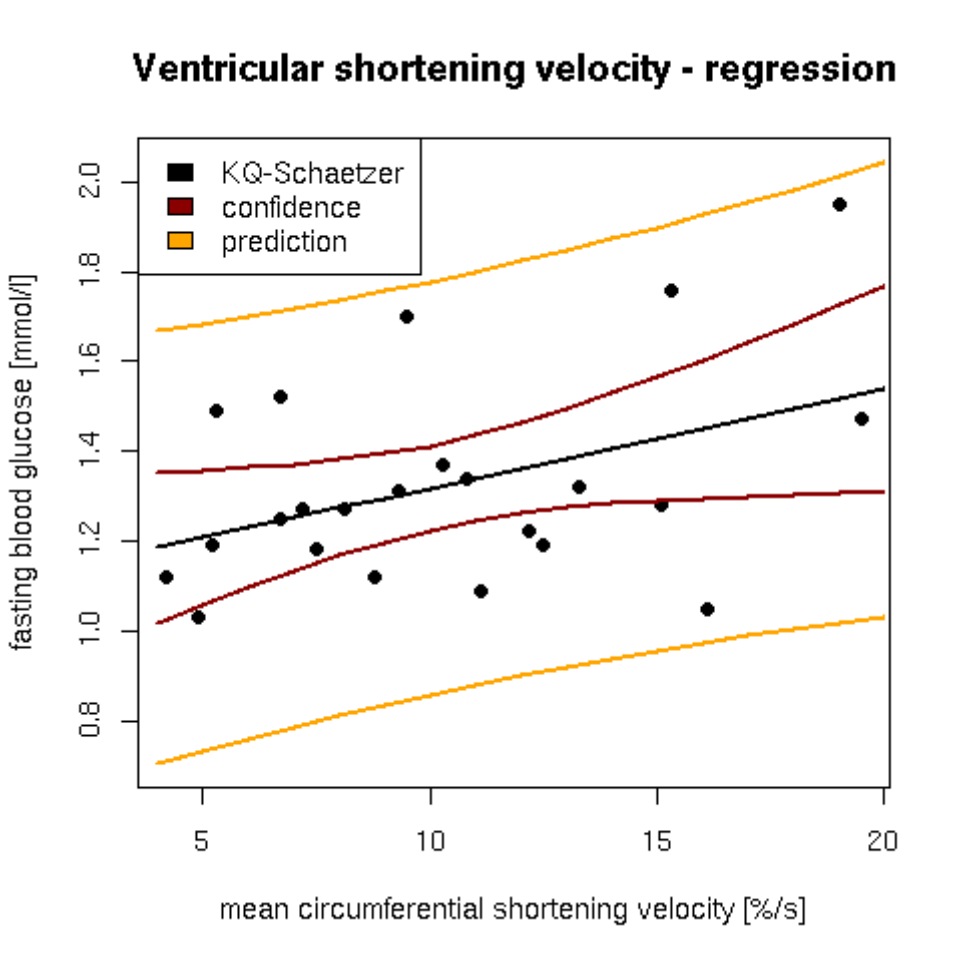
```
Coefficients:
                 Estimate  Std. Error t value Pr(>|t|)
(Intercept)      1.09781    0.11748   9.345 6.26e-09 ***
blood.glucose    0.02196    0.01045   2.101   0.0479 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2167 on 21 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-Squared: 0.1737,     Adjusted R-squared: 0.1343
F-statistic: 4.414 on 1 and 21 DF,  p-value: 0.0479
```

Einschub: R^2 und Korrelation

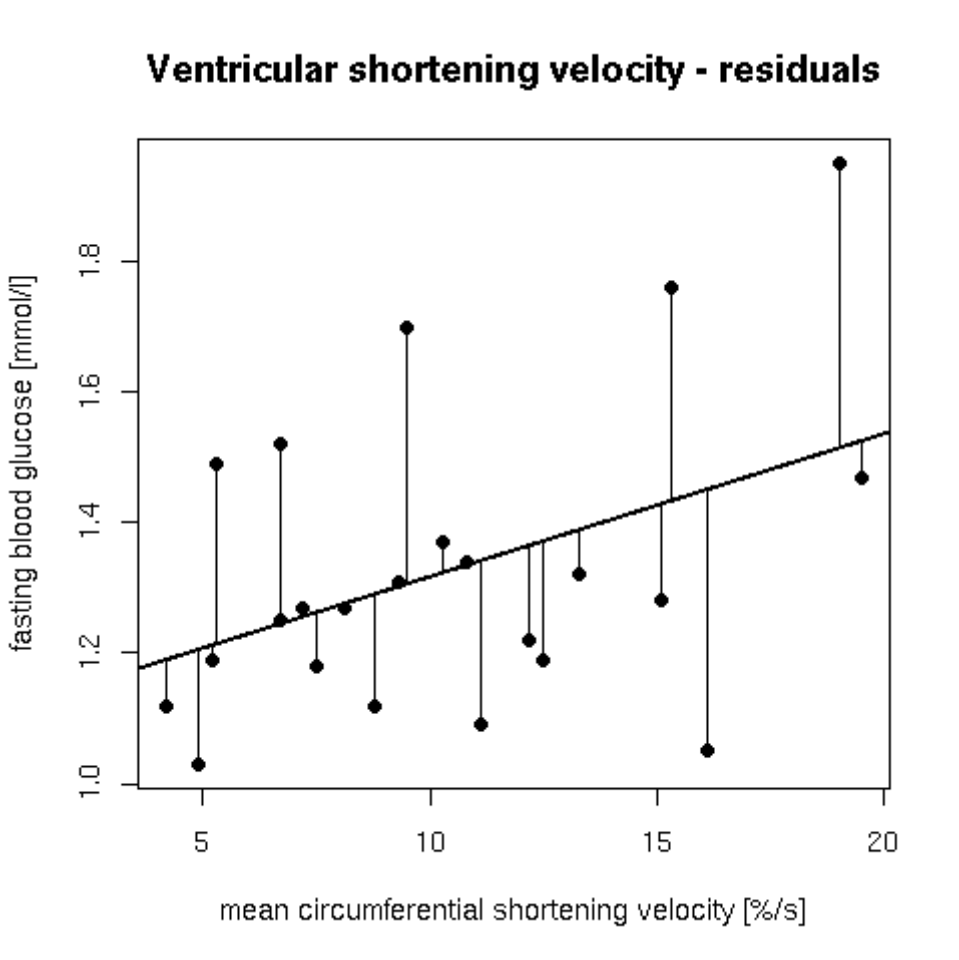
Pearson's product-moment correlation

data: blood.glucose and short.velocity
t = 2.101, df = 21, p-value = 0.0479
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.005496682 0.707429479
sample estimates:
      cor
0.4167546

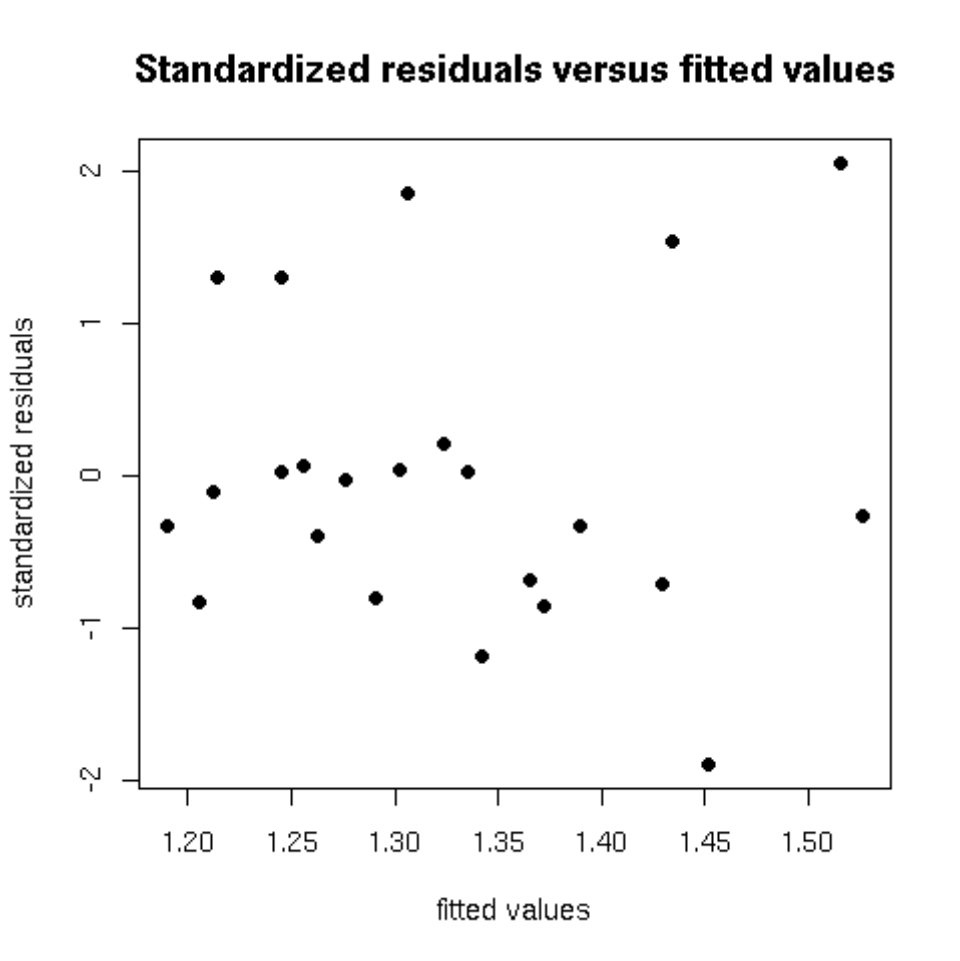
Beispiel 1 – KQ-Schätzer II



Beispiel 1 – Residuen



Beispiel 1 – Linearität und Homoskedastizität



Beispiel 1 – Heteroskedastizität

- **studentized Breusch-Pagan** test

  BP = 3.8641, df = 1, p-value = 0.04933

- **Goldfeld-Quandt** test

  GQ = 2.3156, df1 = 10, df2 = 9, p-value = 0.111

- **Harrison-McCabe** test

  HMC = 0.2812, p-value = 0.092

Beispiel 1 – Linearität

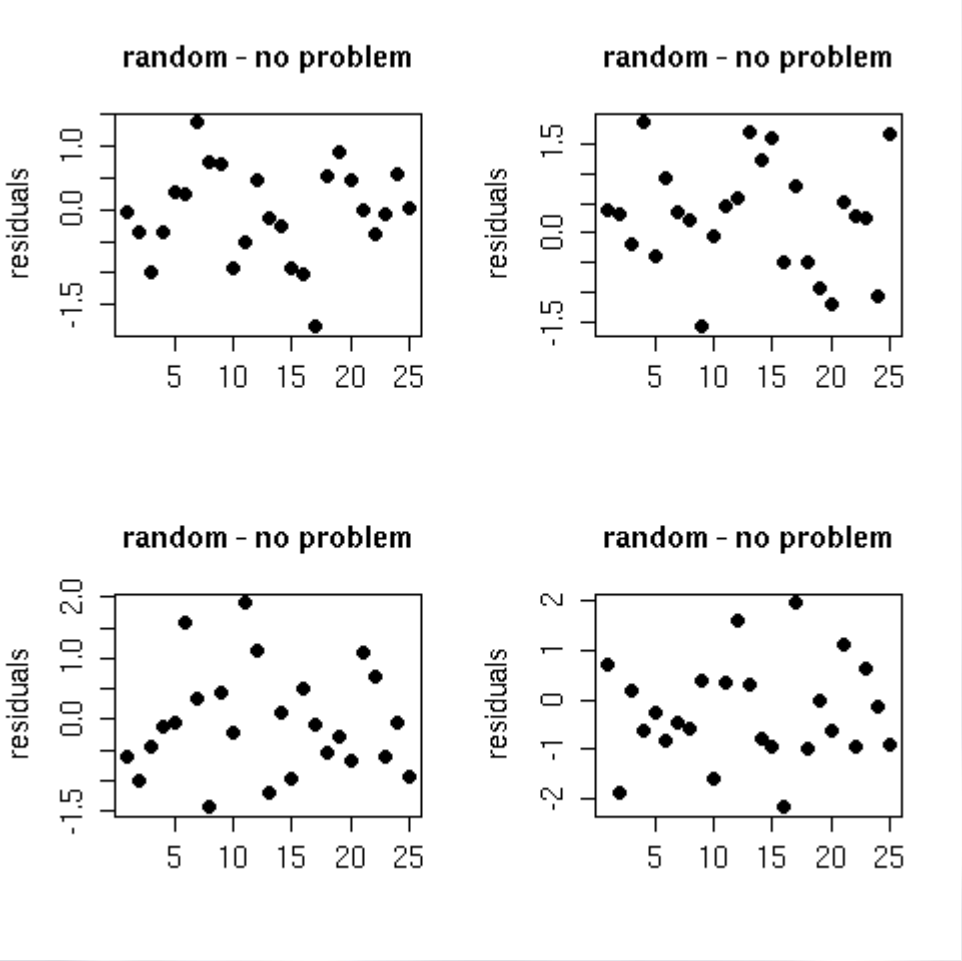
- **Harvey-Collier** test

  HC = 1.0598, df = 20, p-value = 0.3019

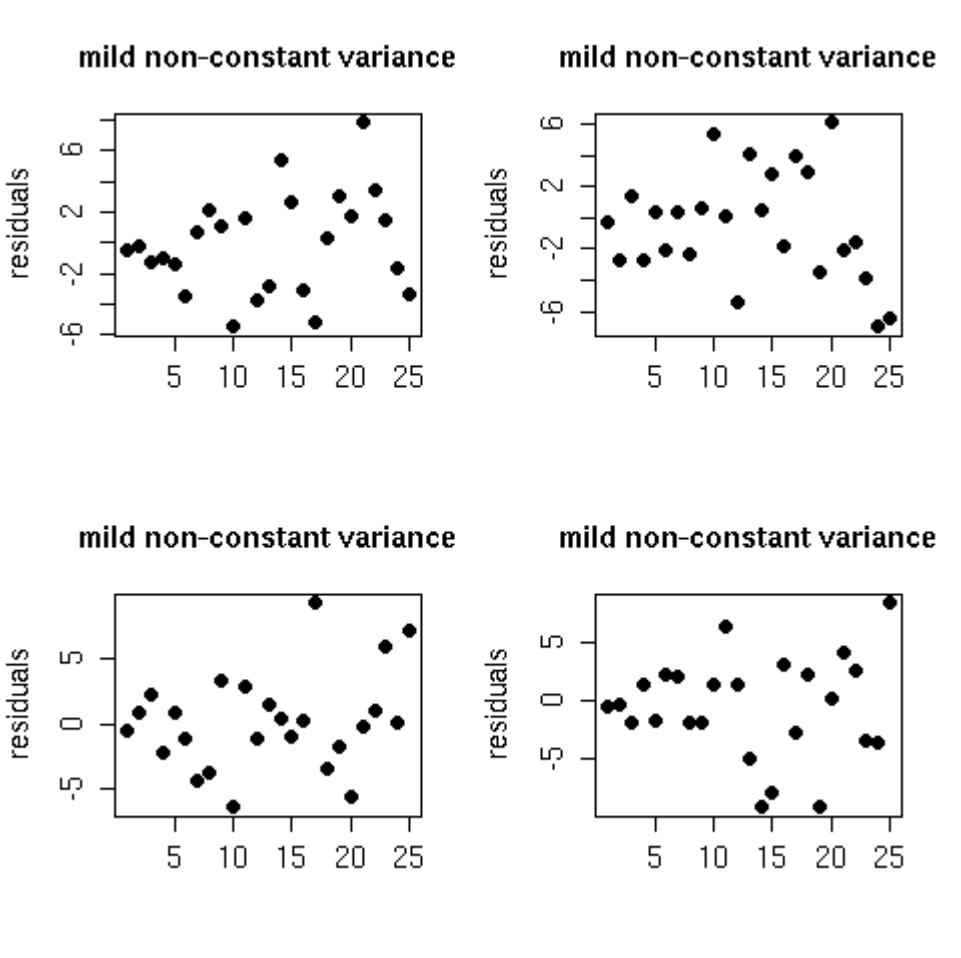
- **Rainbow** test

  Rain = 1.3188, df1 = 12, df2 = 9, p-value = 0.3447

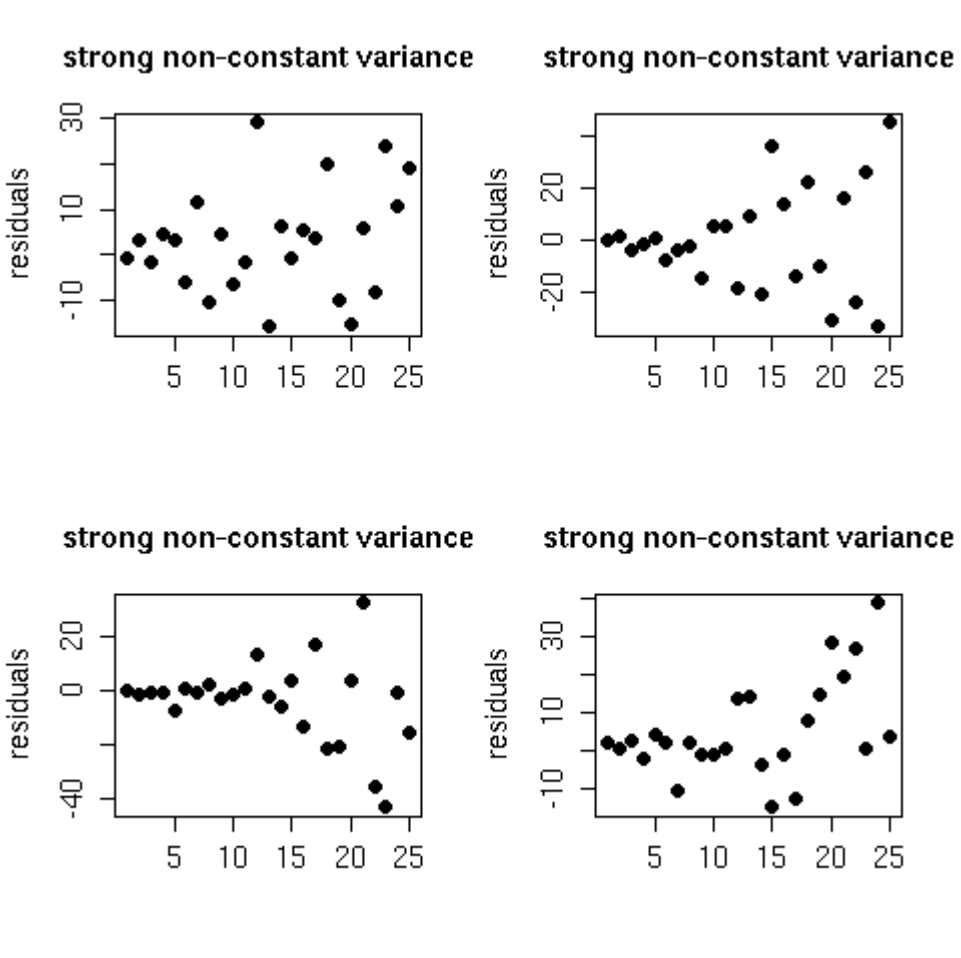
Zufällige Residuen



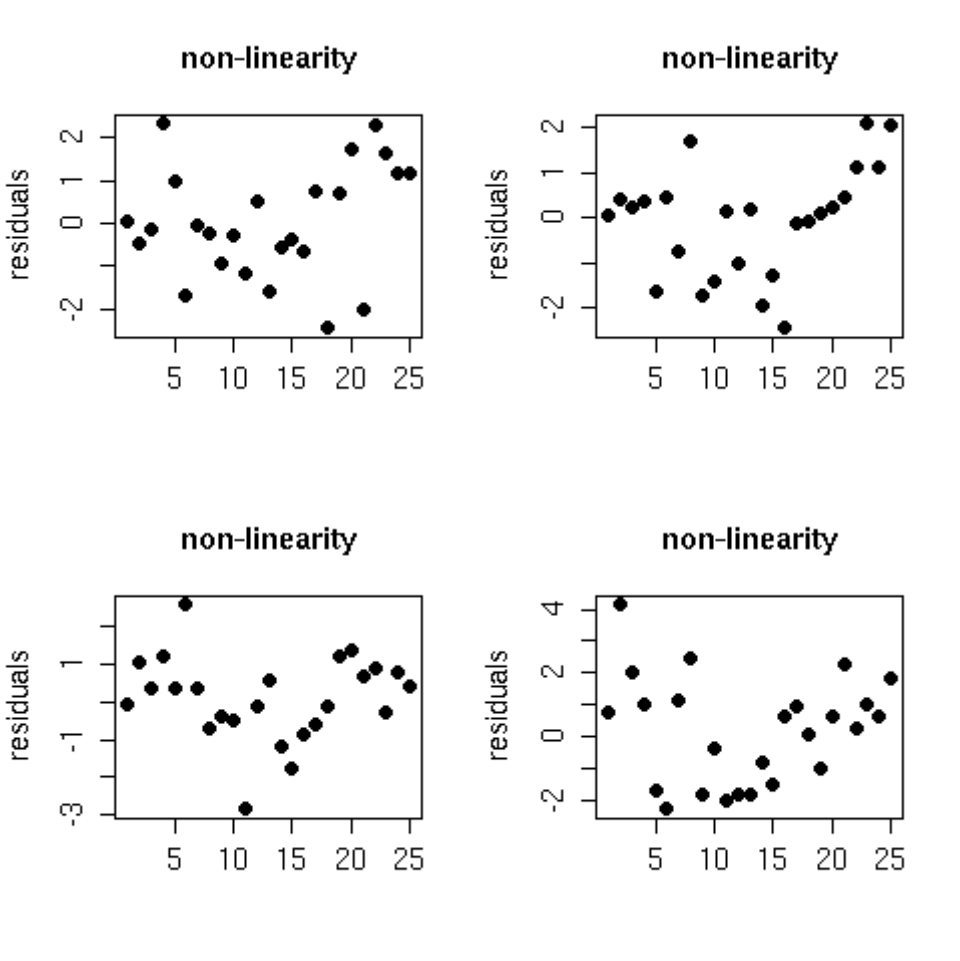
Heteroskedastische Residuen I



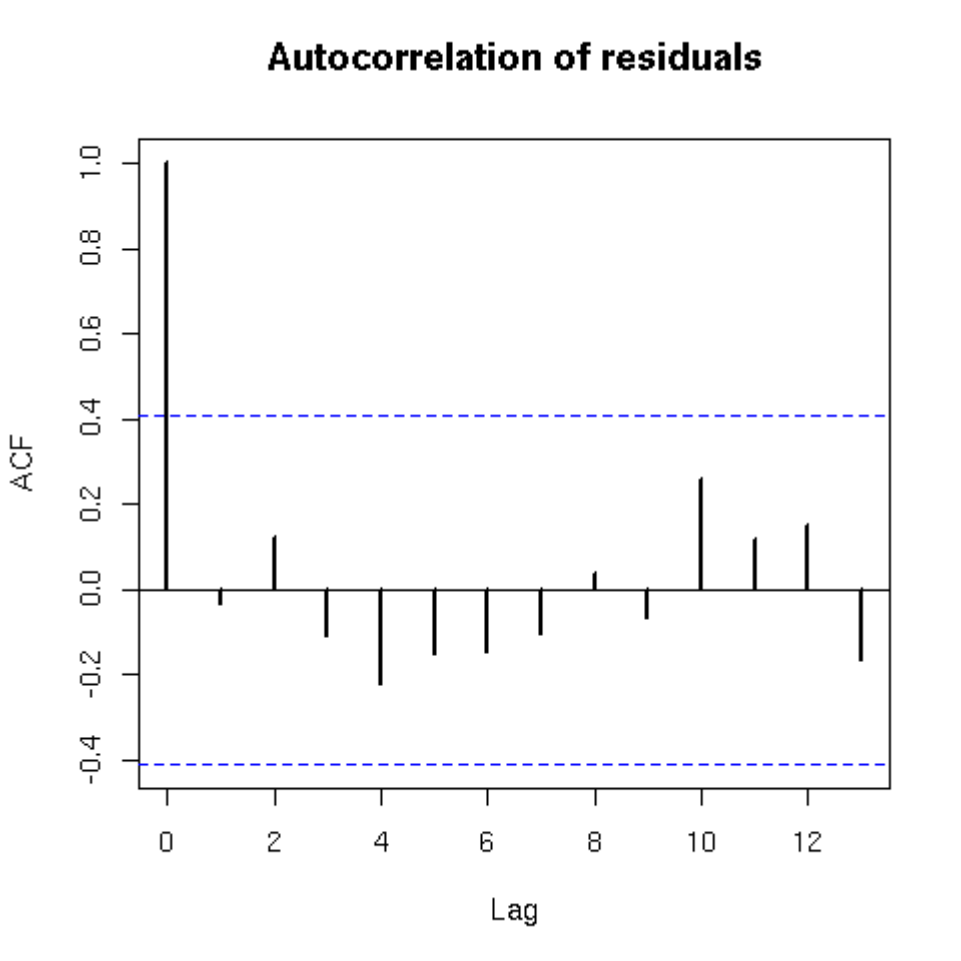
Heteroskedastische Residuen II



Residuen: Nicht-Linearität



Beispiel 1 – Korrelation I



Beispiel 1 – Korrelation II

- **Breusch-Godfrey** test for serial correlation of order 1

  LM test = 0.0312, df = 1, p-value = 0.8599

- **Breusch-Godfrey** test for serial correlation of order 3

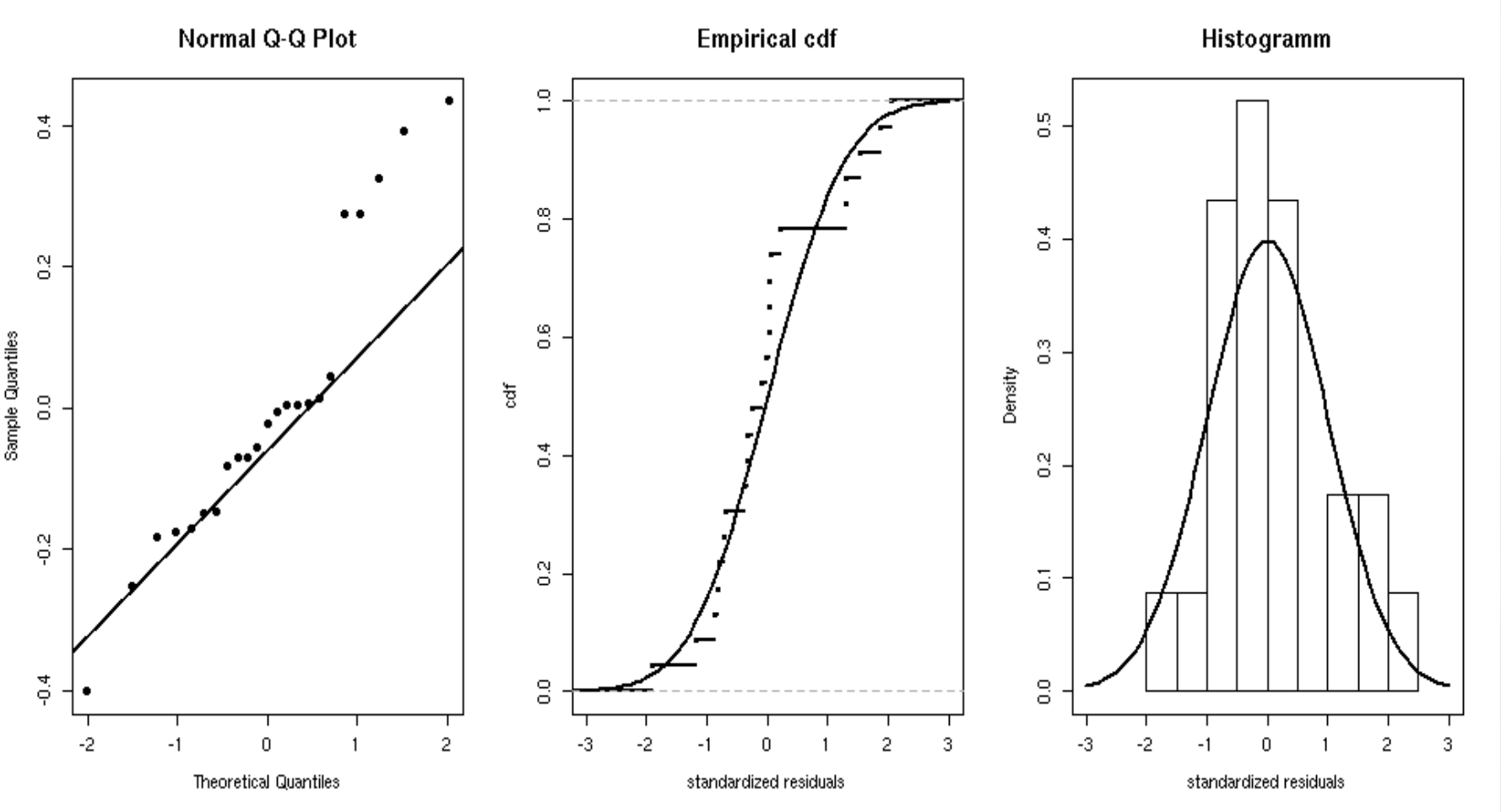
  LM test = 0.7757, df = 3, p-value = 0.8553

- **Durbin-Watson** test

  DW = 1.8024, p-value = 0.3153

  (alternative hypothesis: true autocorrelation is greater than 0)

Beispiel 1 – Normalität I



SIRS Lab

Beispiel 1 – Normalität II

Auswahl wichtiger Tests

- **Anderson-Darling** test: A = 0.8498, p-value = 0.02418
- **Cramer-von Mises** test: W = 0.1578, p-value = 0.01696
- **Lilliefors (Kolmogorov-Smirnov)** test:

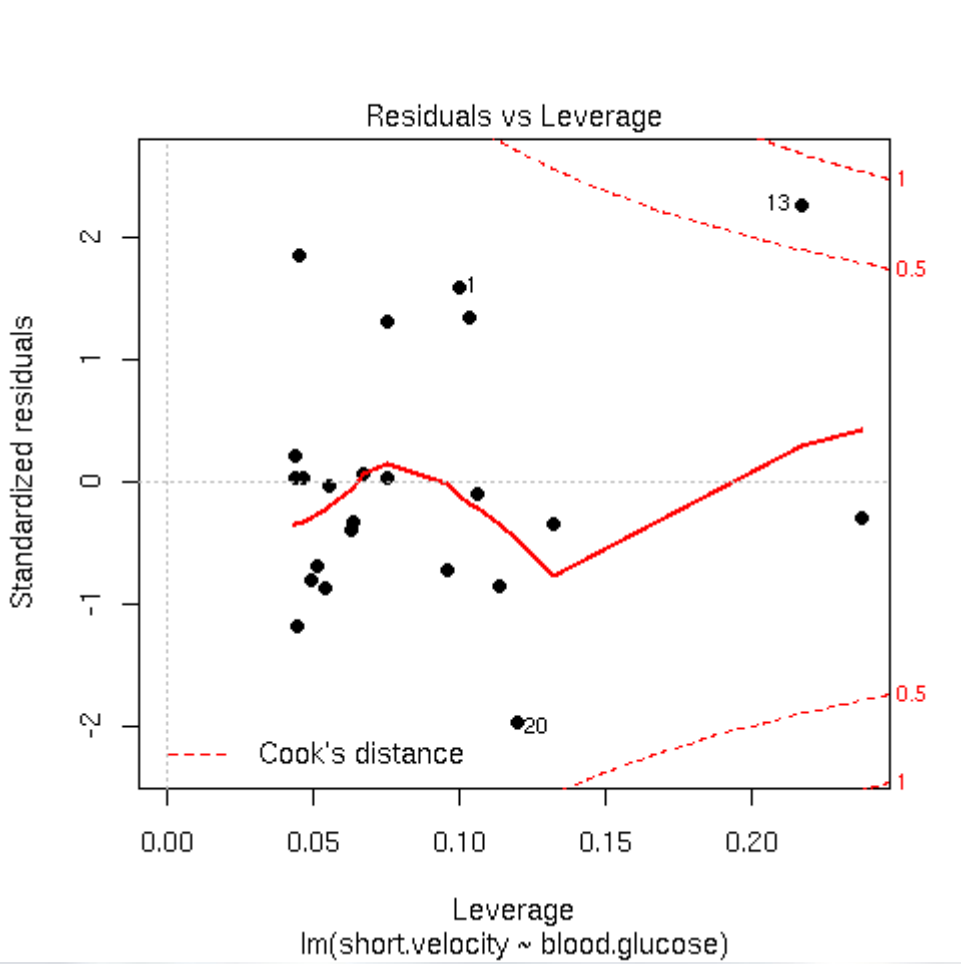
  D = 0.2127, p-value = 0.008335
- **Pearson chi-square** test: P = 10.7391, p-value = 0.0568
- **Shapiro-Francia** test: W = 0.9242, p-value = 0.07715

Transformationen

Hilfreich, falls Voraussetzungen verletzt:

- Box-Cox Potenztransformation
- Varianzstabilisierende Transformation (falls Verteilung der y_i bekannt)
- Generalisierte kleinste Quadrate Methode (falls Fehler ε_i nicht homoskedastisch bzw. nicht unkorreliert)

Beispiel 1 – Ausreißer I



Beispiel 1 – Ausreißer II

- **Dixon test for outliers:** Q.13 = 0.1762, p-value = 0.7945

  (alternative hypothesis: highest value 0.435 is an outlier)

- **Grubbs test for one outlier:**

  G.13 = 2.0542, U = 0.7995, p-value = 0.3696

  (alternative hypothesis: highest value 0.435 is an outlier)

- **Grubbs test for two opposite outliers:**

  G.13 = 3.9502, U = 0.6447, p-value = 0.6874

  (alternative hypothesis: -0.401 and 0.435 are outliers)

Robuste Schätzer

- Huber (1981): M-Schätzer
- Hampel et al (1986): M-Schätzer
- Yohai (1987): MM-Schätzer
- Rousseeuw and Leroy (1987): LMS- und LTS-Schätzer
- Rieder (1994): AL-Schätzer

Vorteile Robuster Schätzer

- Robust gegen Modellabweichungen verschiedener Art

  (d.h. Ideales Modell + Umgebung)

- M- bzw. AL-Schätzer asymptotisch optimal innerhalb großer Klasse von Schätzern

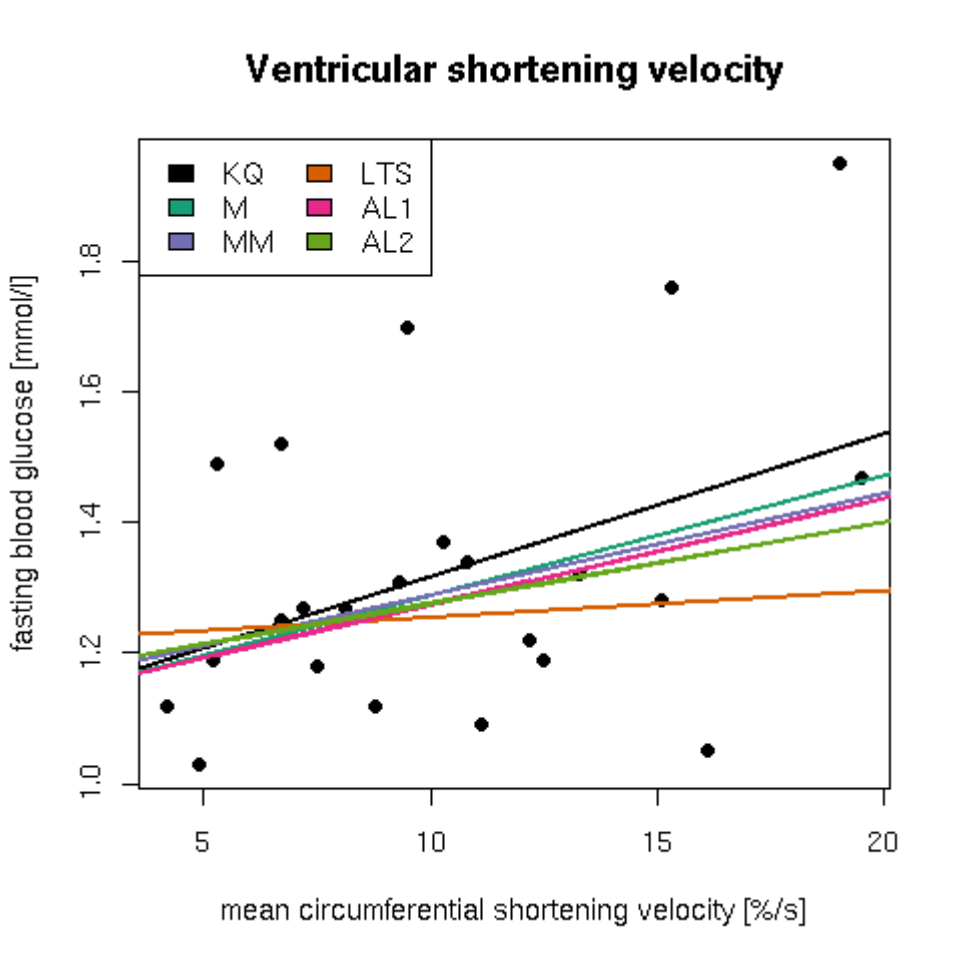
Nachteile Robuster Schätzer

- Komplizierte Theorie

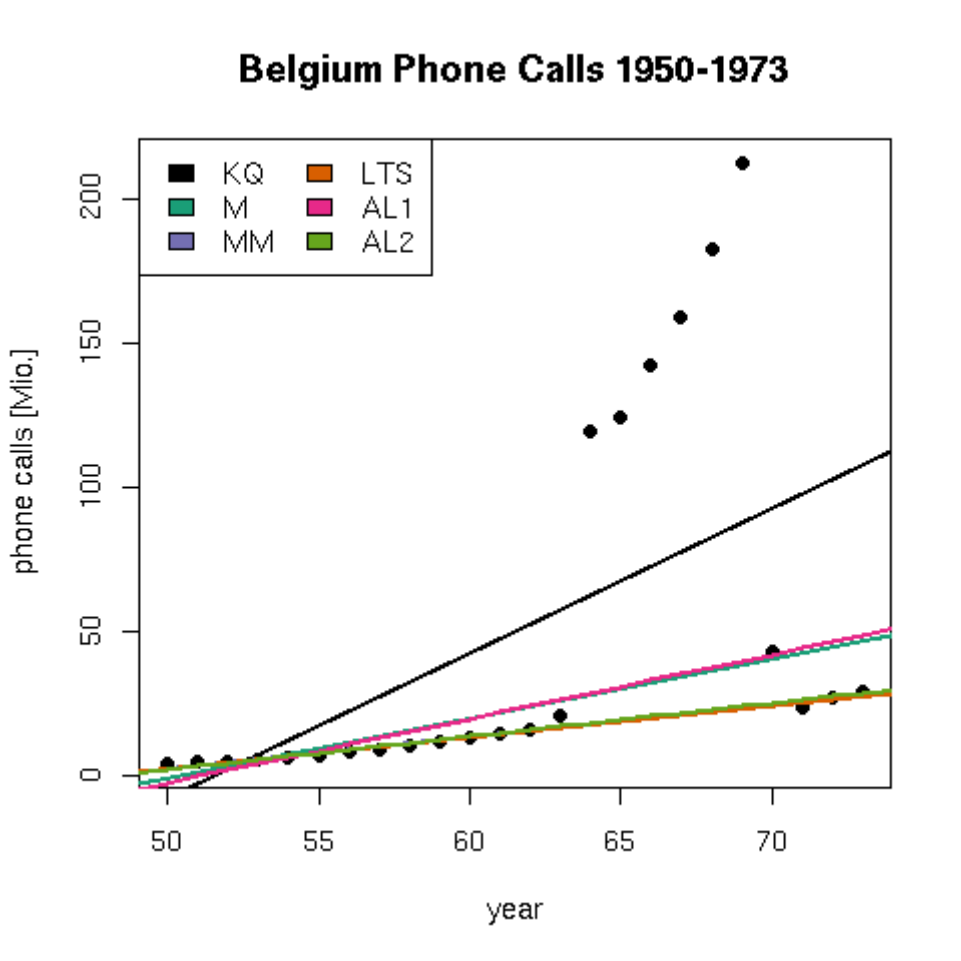
  I.d.R. nur asymptotische Aussagen; d.h., asymptotisch optimaler Schätzer, asymptotische Tests, usw.

- Höherer Rechenaufwand

Beispiel 1 – Robuste Schätzer



Beispiel 2*



* P.J. Rousseeuw and A.M. Leroy (1987), Robust Regression and Outlier Detection, Wiley.

Variablenselektion - Methoden

- **forward selection:** Füge Variablen hinzu bis eine gewisse Vorgabe erreicht ist.
- **backward elimination:** Entferne Variablen bis eine gewisse Vorgabe erreicht ist.
- **all subsets:** Betrachte alle möglichen Teilmodelle und wähle das “beste”.

Variablenselektion - Kriterien

- **p-Werte:** t- oder F-Test
- **adjusted R^2:** Wähle Modell mit größtem angepasstem R^2
- **Mallows' C_p:** Wähle Modell mit $C_p \approx p$
- **Akaike information criterion (AIC):** proportional zu C_p

Ausblick: Linear Mixed-Effect Models

$$y_{ij} = \beta_1 x_{1ij} + \cdots + \beta_p x_{pij} + b_{i1} z_{1ij} + \cdots + b_{iq} z_{qij} + \varepsilon_{ij}$$

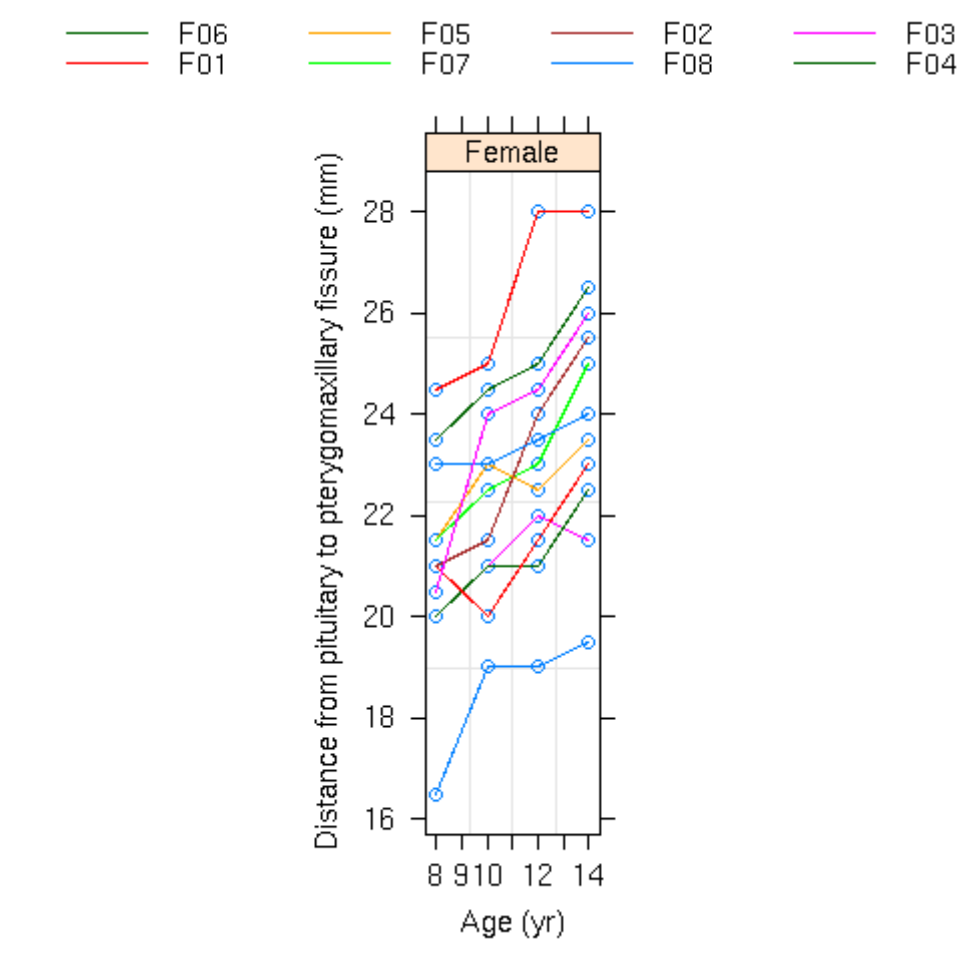
$$b_{ik} \sim N(0, \psi_k^2),\ \mathrm{Cov}(b_k, b_{k'}) = \psi_{kk'}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2 \lambda_{ijj}),\ \mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ij'}) = \sigma^2 \lambda_{ijj'}$$

Einsatzbereiche von Linear Mixed-Effect Models

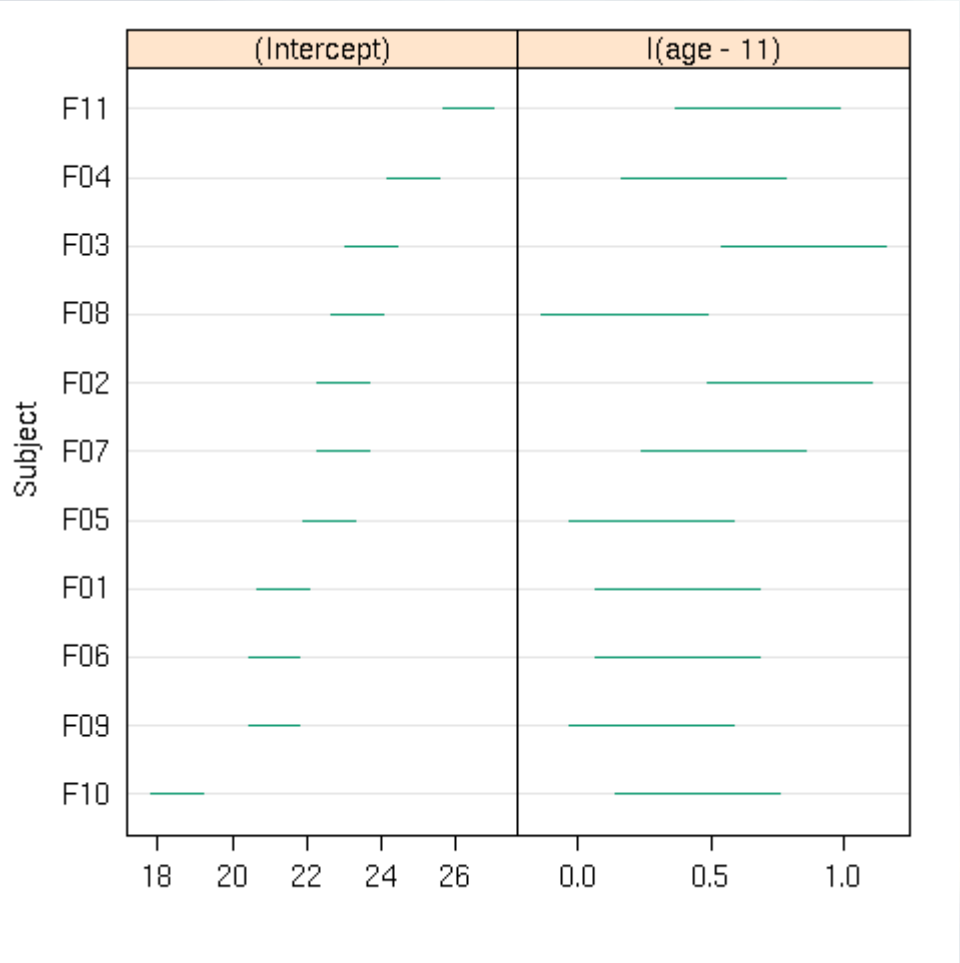
- generell: **abhängige Daten** (z.B. repeated measures oder longitudinal data)
- **Blockdesigns** (auch unbalanziert!)
- **verschachtelte Designs**
- **Split-Plot Experimente**

Beispiel 3*



* Potthoff, R. F. and Roy, S. N. (1964), A generalized multivariate analysis of variance model useful especially for growth curve problems', Biometrika, *51*, 313-326.

Beispiel 3 – Getrennte Analyse



SIRS Lab

Beispiel 3 – Linear Mixed-Effect Model I

Betrachte das Modell:

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + b_i + \varepsilon_{ij}$$

$$b_i \sim N(0, \psi^2),\ \varepsilon_{ij} \sim N(0, \sigma^2)$$

Beispiel 3 – Linear Mixed-Effect Model II

Random effects:

Formula: ~1 | Subject

	(Intercept)	Residual
StdDev:	2.06847	0.7800331

Fixed effects: distance ~ I(age - 11)

	Value	Std.Error	DF	t-value	p-value
(Intercept)	22.648	0.635	32	35.6850	0
I(age - 11)	0.480	0.053	32	9.1186	0

Software I

- **P. Dalgaard (2005).** ISwR: Introductory Statistics with R. R package version 1.0-6.
- **J. Fox (2006).** car: Companion to Applied Regression. R package version 1.1-2. http://socserv.socsci.mcmaster.ca/jfox/
- **J. Gross (2006).** nortest: Tests for Normality. R package version 1.0.
- **M. Kohl (2006).** RobRex: Optimally robust influence curves for regression and scale. R package version 0.4-2. http://www.stamats.de/RobASt.htm
- **L. Komsta (2006).** outliers: Tests for outliers. R package version 0.13. http://www.r-project.org, http://www.komsta.net/
- **J. Pinheiro, D. Bates, S. DebRoy and and D. Sarkar (2006).** nlme: Linear and nonlinear mixed effects models. R package version 3.1-76.

Software II

- **R Development Core Team (2006). R: A Language and Environment for Statistical Computing.** R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
- **V. Todorov, A. Ruckstuhl, M. Salibian-Barrera and M. Maechler (2006).** robustbase: Basic Robust Statistics. R package version 0.2-6.
- **W.N. Venables and B.D. Ripley (2002).** Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
- **D. Wuertz, many others and see the SOURCE file (2006).** fMultivar: Rmetrics - Multivariate Market Analysis. R package version 240.10068. http://www.rmetrics.org

Literatur I

- **R. Christensen (2002). Plane Answers to Complex Questions.** The Theory of Linear Models. Third Edition. Springer. New York.
- **P. Dalgaard (2002). Introductory Statistics with R.** Springer. New York.
- **J. Faraway (2002). Practical Regression and ANOVA in R.** http://www.stat.lsa.umich.edu/~faraway/book/
- **R. Heiberger and B. Holland (2004). Statistical Analysis and Data Display.** An Intermediate Course with Examples in S-Plus, R and SAS. Springer. New York.
- **J. Pinheiro and D. Bates (2000). Mixed-Effects Models in S and S-Plus.** Springer. New York.

Literatur II

- **H. J. Trampisch und J. Windeler (Hrsg.) (2000): Medizinische Statistik.** Springer Verlag. 2. Auflage.
- **W. Venables and B. Ripley (2002). Modern Applied Statistics with S.** Fourth Edition. Springer. New York.