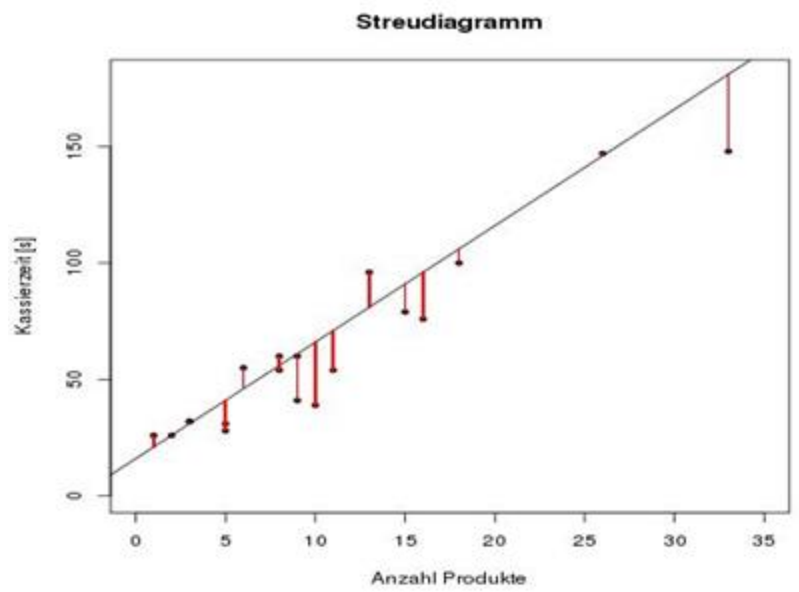
ETH zürich

Multiple Regression

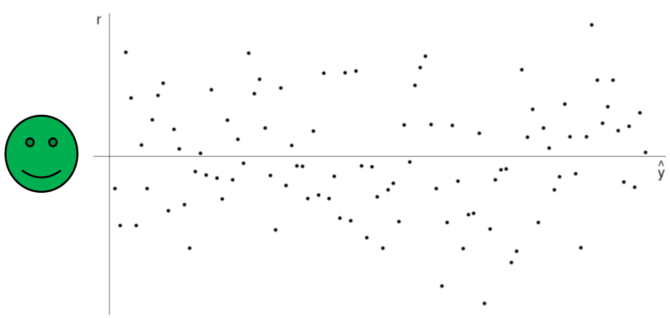
Wdh: Einfache lineare Regression

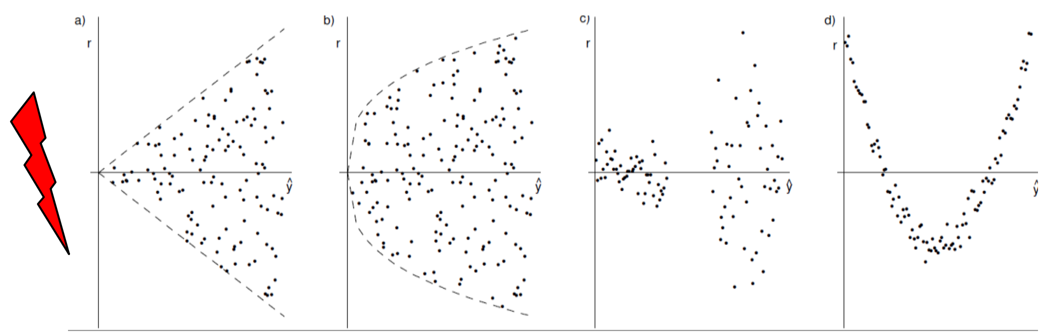
- Modell: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)\ i.i.d$
- Finde $\widehat{\beta_0}, \widehat{\beta_1}$: Methode der kleinsten Quadrate
  $\widehat{\sigma^2}$ ist geschätzte Varianz der Residuen
- $\frac{\widehat{\beta}_k - \beta_k}{\widehat{s.e.}(\widehat{\beta}_k)} \sim t_{n-2}$ → t-Test: $H_0: \beta_k = 0, \ H_A: \beta_k \neq 0$
- R: Funktion ‘lm’



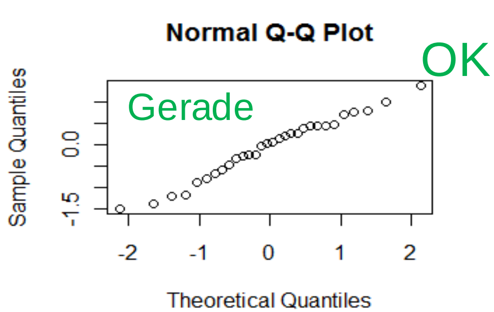
Wdh: Residuenanalyse
Sind Modellannahmen erfüllt?

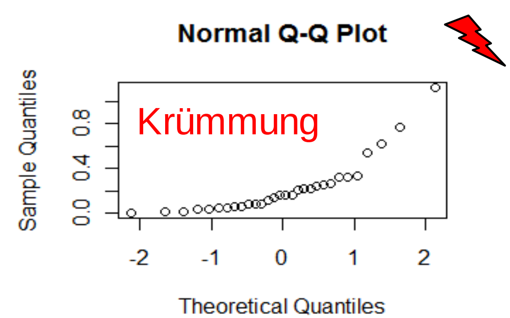
- Tukey-Anscombe Plot: Modellwert vs. Residuen (Fehlervarianz konstant, systematische Fehler)



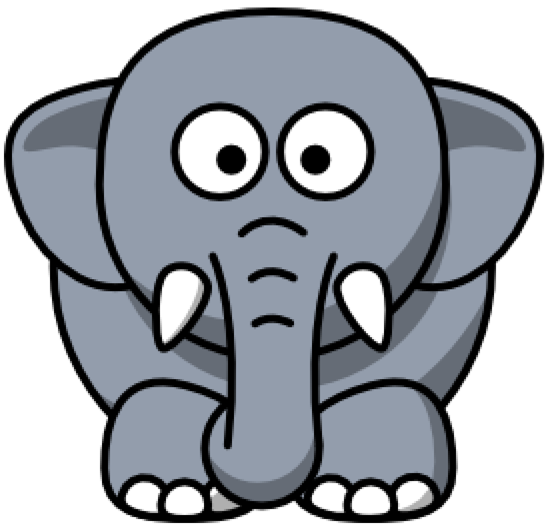


- QQ-Plot: Empirische Quantile vs. theoretische Quantile (Residuen normalverteilt)

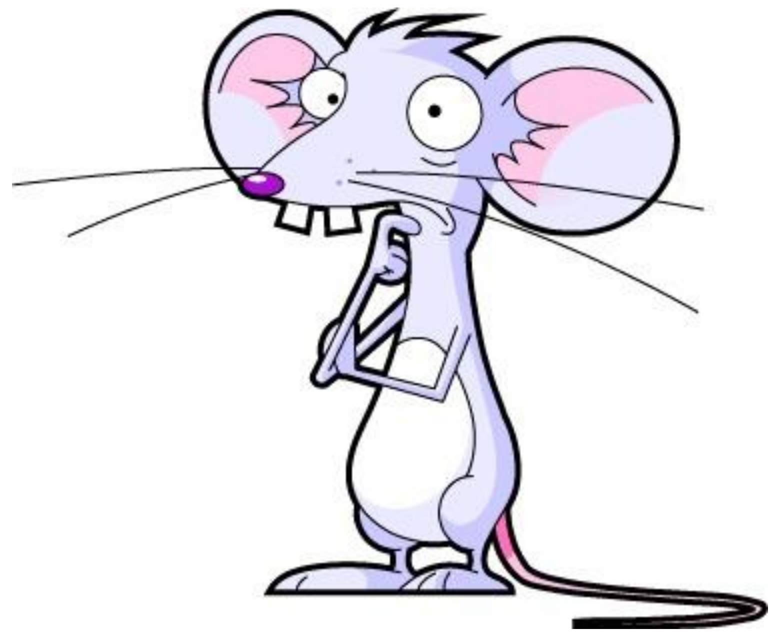




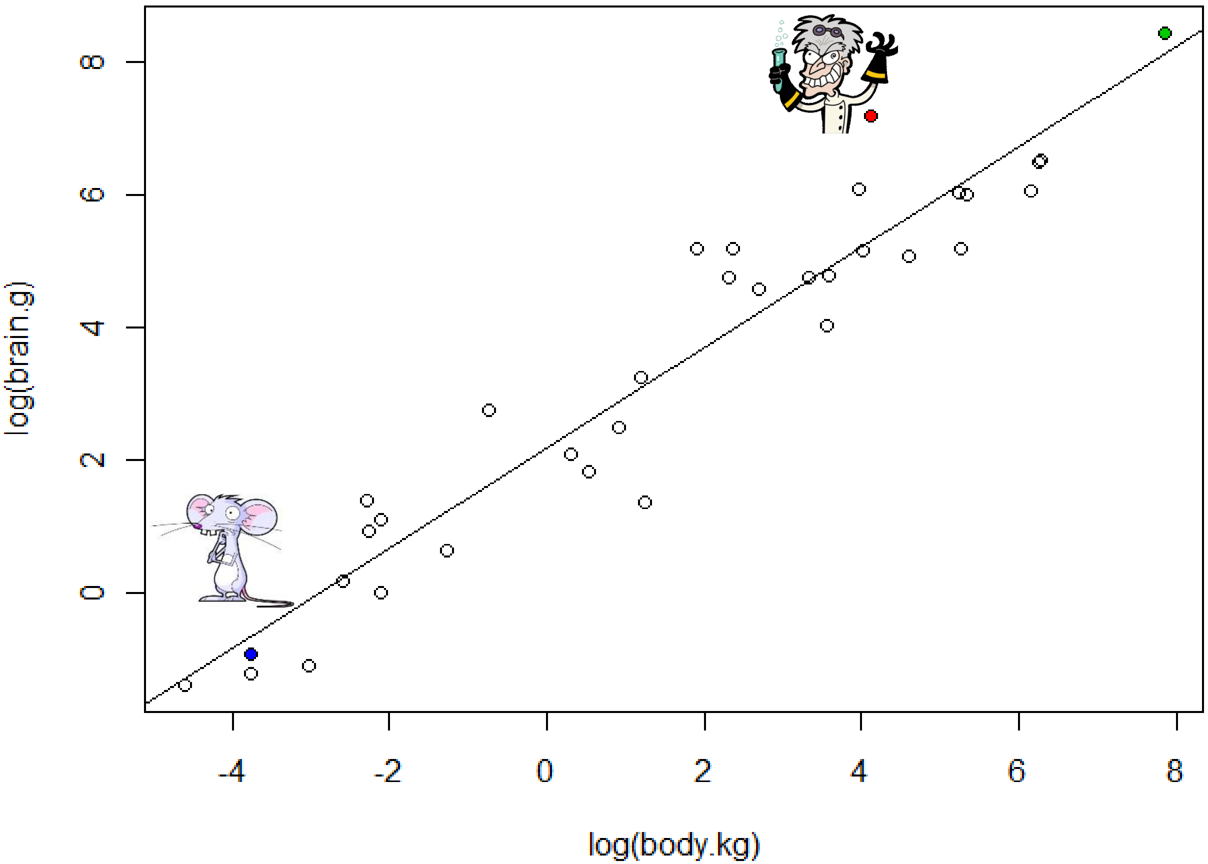
Falls Residuenanalyse schlecht: Transformationen



Zusammenhang:
Hirnmasse
und
Körpermasse



Bsp: log(Hirnmasse) vs. log(Körpermasse)



$$\log(H) = \widehat{\beta_0} + \widehat{\beta_1} * \log(K)$$

$$H = \exp\left(\widehat{\beta_0} + \widehat{\beta_1} * \log(K)\right)$$
$$\rightarrow H = \hat{a} * K^{\hat{b}}$$

$\widehat{\beta_0} = 2.19$ (95%-VI: [1.89; 2.49]); $\widehat{\beta_1} = 0.75$ (95%-VI: [0.67; 0.83])

$\hat{a} = \exp\left(\widehat{\beta_0}\right) = 8.94$ (95%-VI: $[\exp(1.89)\,;\exp(2.49)]$ = [6.60; 12.02])
$\hat{b} = \widehat{\beta_1}$ (95%-VI: [0.67; 0.83])

Übersicht über nützliche Transformationen

- Linearer Zusammenhang:
  $y = \beta_0 + \beta_1 x + \epsilon$ (keine Transformation nötig)
- Exponentieller Zusammenhang:
  $\log(y) = \beta_0 + \beta_1 x + \epsilon \rightarrow y = \exp(\beta_0) \cdot \exp(\beta_1 x) \cdot \exp(\epsilon)$
  $= a \cdot b^x \cdot \exp(\epsilon)$
  $mit\ a = \exp(\beta_0)\ \ und\ b = \exp(\beta_1)$
- Polynomieller Zusammenhang:
  $\log(y) = \beta_0 + \beta_1 \cdot \log(x) + \epsilon$
  $\rightarrow \mathrm{y} = \exp(\beta_0 + \beta_1 \cdot \log(x) + \epsilon)$
  $\rightarrow y = a \cdot x^b \cdot \exp(\epsilon)\ \ mit\ a = \exp(\beta_0)\ \ und\ b = \beta_1$
  multiplikation mit fehler

Beispiele für nicht-linearisierbare Funktionen

- Bsp 1: $y = a \cdot b^x + \epsilon$: Fehlerterm müsste multiplikativ sein, damit linearisiert werden kann

man kann nicht einfach logarithmieren für linearisieren

- Bsp 2: Logistisches Wachstum:

$$y = \frac{a}{1 + \exp\left(-\frac{x-b}{c}\right)}$$

- Nicht-lineare Funktionen haben oft eine Begründung aus dem Kontext (Physik, Chemie,...)
  → Nichtlineare Regression
  wird nicht vertieft

Multiple Lineare Regression

Wie hängt Energie von Eiweiss, Kohlenhydraten und Fett ab?

Multiple Lineare Regression - Modell

- $Y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_{p-1} x_{i,p-1} + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2)\ i.i.d.$

- Schätzer $\widehat{\beta_i}$ für β_i minimieren Residuenquadratsumme (RSS):

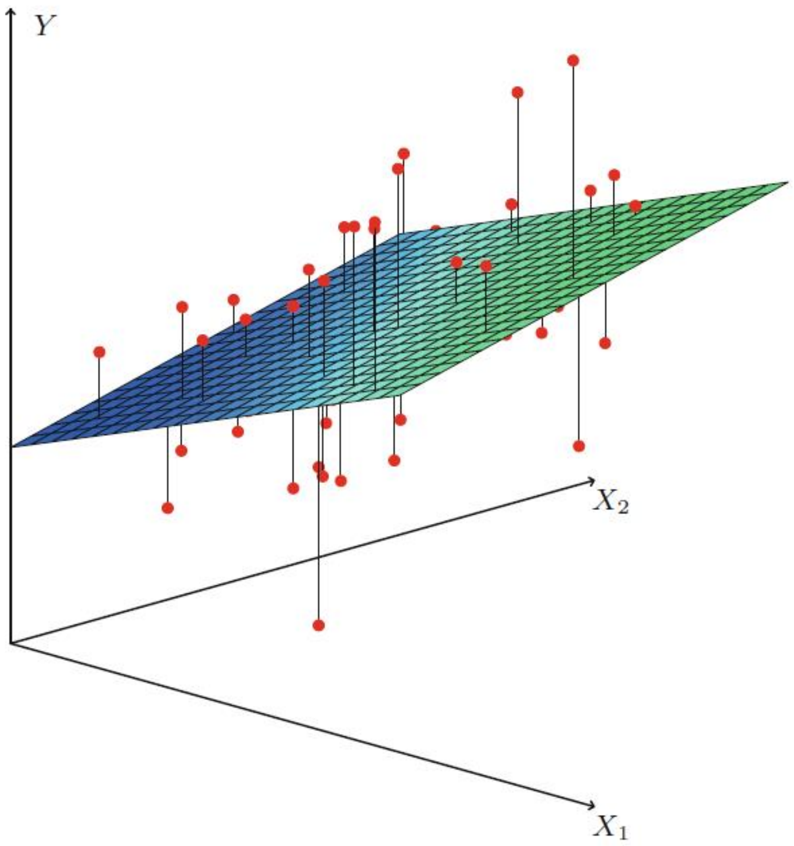
$$\widehat{\beta_i}\ minimieren \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_{i,1} + \dots + \beta_{p-1} x_{i,p-1}\right)\right)^2$$

- Unter obigen Annahmen:

$$\mathrm{t} = \frac{\widehat{\beta_i} - 0}{SE\left(\widehat{\beta_i}\right)} \sim t_{n-p}$$

→ t-Test in der Linearen Regression

Intuition: Multiple Lineare Regression



Einfache oder Multiple Regression

(Gilt für alle GLMs; hier am Bsp der linearen Regression)

- Einfache Regression:
  "Totaler Effekt"
  y ~ x → "Wenn sich x um eine Einheit erhöht, erhöht sich y um β_1"
- Multiple Regression
  "Bereinigter Effekt"
  y ~ x1 + x2 → "Wenn sich x1 um eine Einheit erhöht und x2 konstant bleibt, erhöht sich y um β_1.
- Kein "richtig" oder "falsch"; eher zwei verschiedene Sichtweisen auf das gleiche Problem

Vorteil von Multipler Regression

- Andere Einflüsse werden ausgeschaltet

Bsp: Diskriminierung

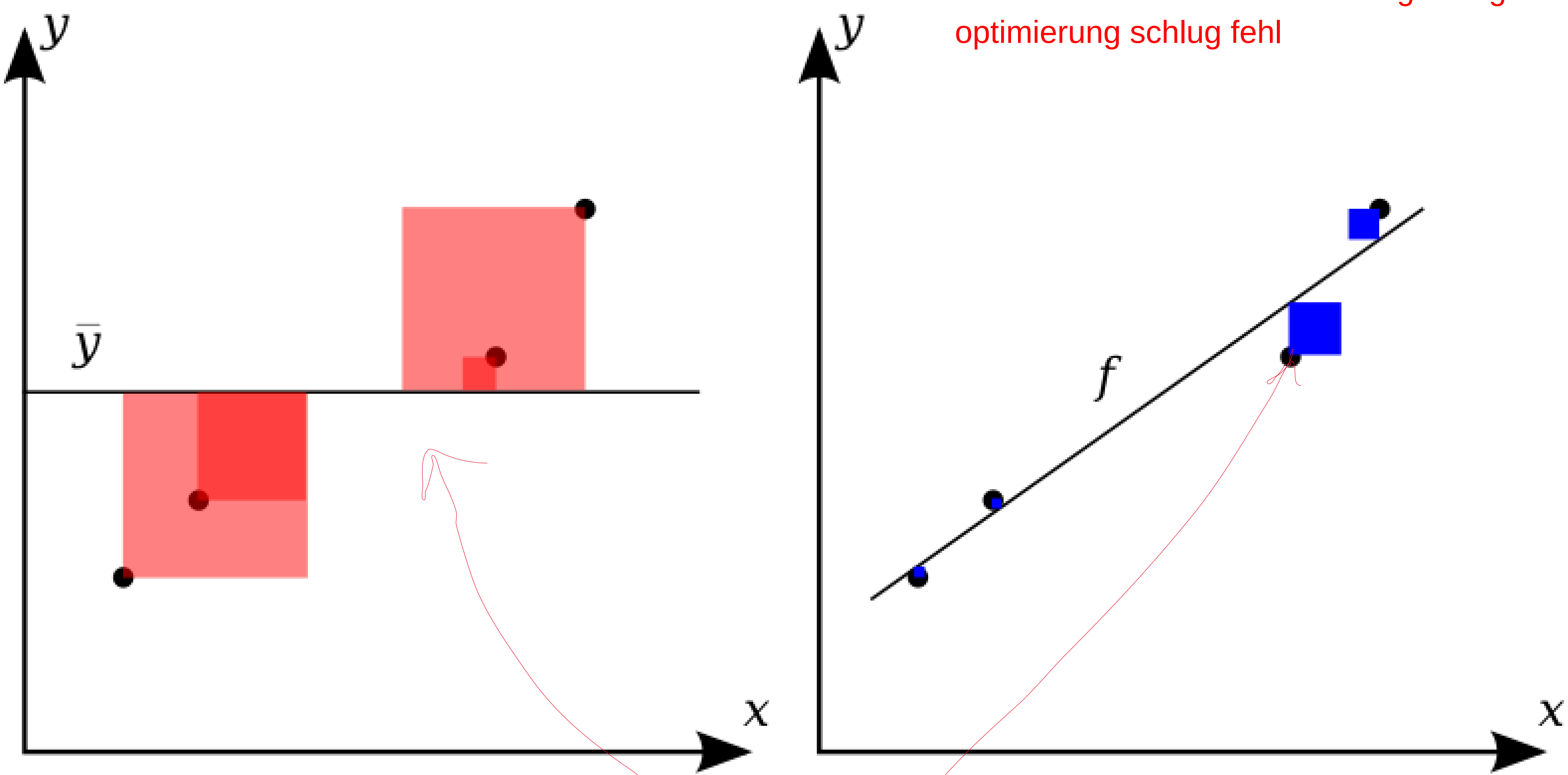
- Einfache Regression:
  Zulassung ~ Geschlecht
- Multiple Regression:
  Zulassung ~ Geschlecht + Ausbildung + etc.

  Berühmtes Beispiel: Simpson's Paradox

in multiplte regression: nicht gerade sondern hyperebene

Kompaktes Gütemass: R^2

R**2 nahe an 0 => blaue punkte im prinzip auf der gerade
R**2 nahe an 1 => flächen fast gleich gross =>
optimierung schlug fehl

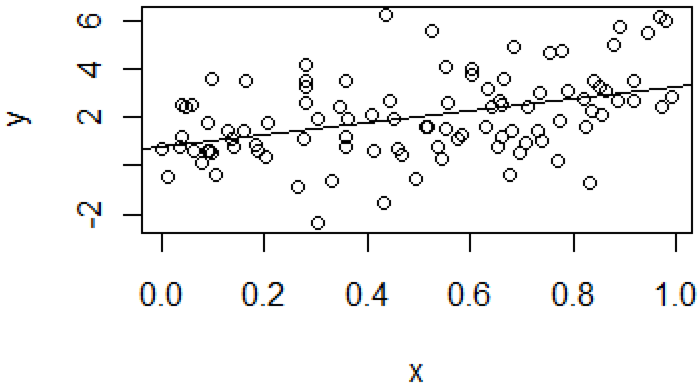


$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}$$

In R: “Multiple R-squared”

R^2: “Wie nahe liegen Punkte an der Gerade / Ebene?”
(im Vergleich zur ursprünglichen Streuung der y-Werte)

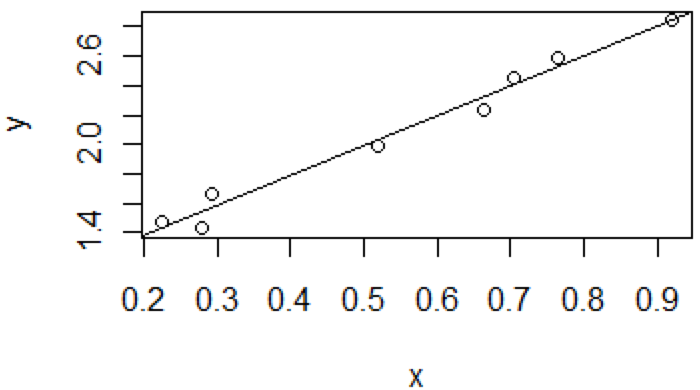
Signifikanz vs. Relevanz



Signifikant, aber evtl. nicht relevant

$H_0: \beta_1 = 0 \ \rightarrow p = 0.00008$

$R^2 = 0.15 \ oder \ |\widehat{\beta_1}| \ sehr \ "klein"$



Signifikant und wohl auch relevant (?)

$H_0: \beta_1 = 0 \ \rightarrow p = 0.00002$

$R^2 = 0.98 \ oder \ |\widehat{\beta_1}| \ "gross"$

Statistik: Entscheidet **Signifikanz**
Wissenschaft: Entscheidet **Relevanz**
(je nach Fach: Unterschiedliche Werte von R^2 gefordert)

Energiegehalt von 20 Lebensmitteln

Daten (pro 100 g)

Name	kcal	gE	gK	gF
Butter	729	0.5	0.5	82.0
Laetta	370	0.0	4.0	39.0
Mozzarella	257	19.0	1.0	20.0
Cantadou	323	7.0	3.0	32.0
Lc1	105	3.5	15.5	3.0
Emmi	130	4.0	16.0	5.5
Quark	65	12.0	2.5	0.1
LightKaese	249	29.0	2.0	14.0
Banane	93	1.0	22.0	0.0
Zucchini	19	1.6	3.3	0.4
Tomate	17	1.0	2.6	0.2
Kartoffel	86	2.0	19.0	0.1
Brot	282	11.0	53.0	1.5
CremeSchnitte	311	4.5	48.0	11.0
Pizza	227	13.0	31.0	5.0
Schoko	569	7.0	46.0	40.0
Chips	517	7.0	51.0	32.0
Spaghetti	350	12.0	72.2	1.5
Reis	358	5.0	83.0	0.5
Stocki	320	9.0	70.0	1.0

Multiple Lineare Regression: Nährwert

```
lm(formula = kcal ~ gE + gK + gF, data = dat)
```

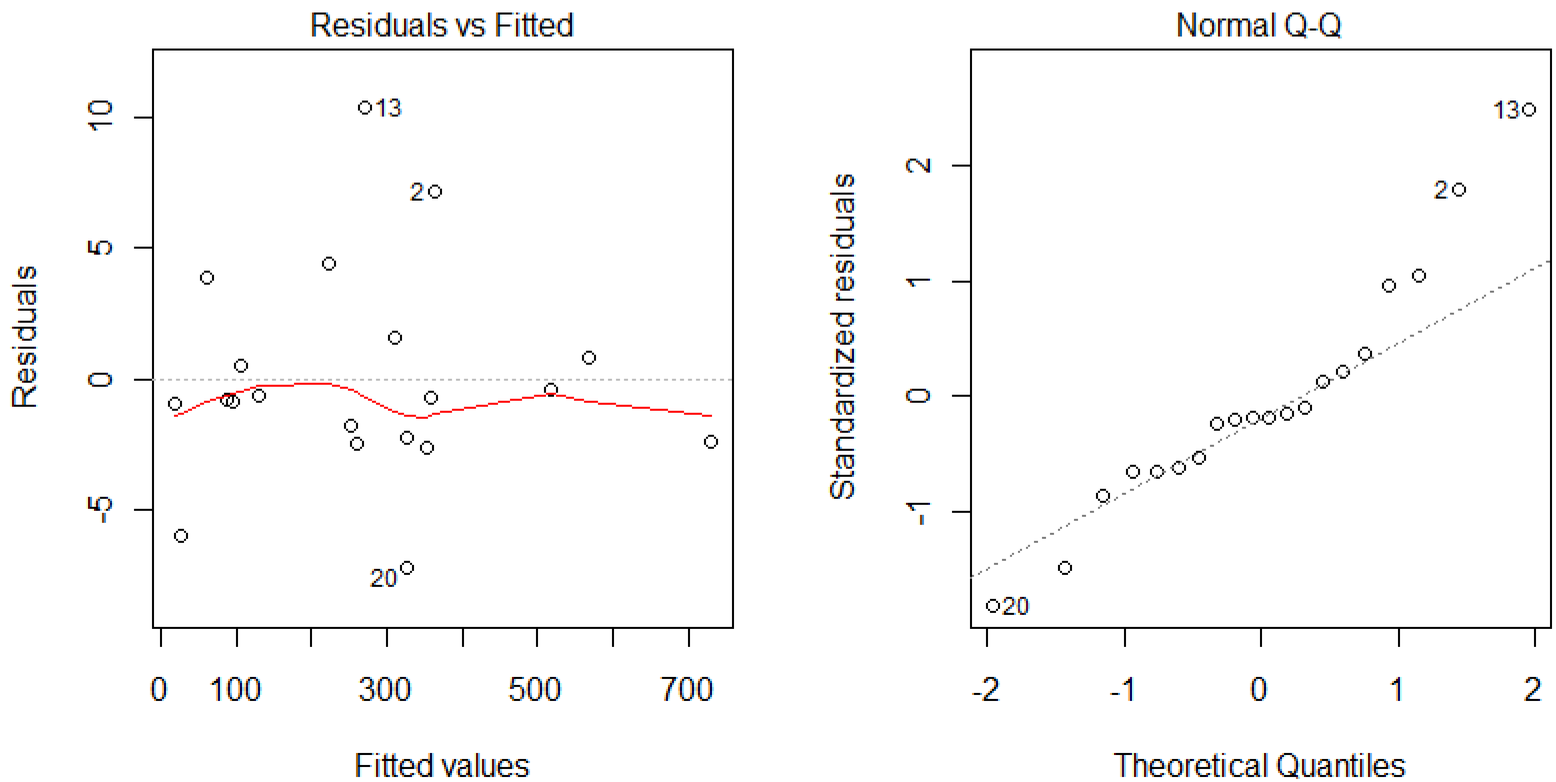
Ein Lebensmittel, das ein Gramm mehr Fett ***aber gleich viel Eiweis und Kohlenhydrate enthält***, enthält im Schnitt 8.8 kcal (95%-VI: [7.8; 9.8]) mehr Energie.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.70736    2.10299   0.812    0.429
gE           4.04087    0.14280  28.298  4.3e-15
gK           4.00415    0.03838 104.330  < 2e-16
gF           8.84937    0.05025 176.115  < 2e-16
```

```
Multiple R-squared: 0.9995
```

Die Punkte liegen äusserst genau auf der geschätzten Geraden. (verglichen mit der ursprünglichen Streuung der Energiewerte)

Residuenanalyse



Im Allgemeinen sind die Modellannahmen erfüllt. Allerdings fallen Beobachtungen 2 (Lätta) und 13 (Brot) etwas aus dem Rahmen (5-10 kcal mehr als vorhergesagt).

F-Test: Gibt es mind. einen signifikanten Einfluss?

alle steigungen haben keinen zusammenhang mit er zielgrösse, also nur beta0 ist von bedeutung

- $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ (ohne Achsenabschnitt β_0)
- H_A: Mind. ein $\beta_i \neq 0$
- "Globaltest": Wenn der F-Test signifikant ist, kann man bei den t-Tests die signifikanten Variablen suchen gehen

```
Residual standard error: 4.422 on 16 degrees of freedom
Multiple R-squared:  0.9995, Adjusted R-squared:  0.9995
F-statistic: 1.152e+04 on 3 and 16 DF,  p-value: < 2.2e-16
```

- Probleme, wenn erklärende Variablen stark korreliert sind ("Kollinearität") ...

Multiple Regression beim Autokauf

Seat Alhambra

Gebrauchtwagen: Was ist ein gutes Angebot?

Residuenanalyse beim Autokauf

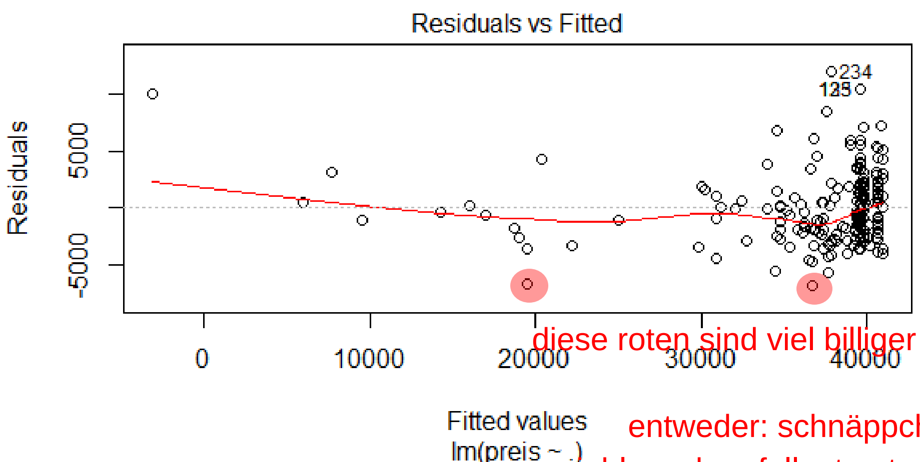
AUTO SCOUT 24

239
Seat
Alhambra

bj	km	preis	diesel	getriebe	vrad	ausstattung
3.2015	50	37890	j	a	n	style
12.2013	13500	33450	j	m	j	style
6.2014	18700	31950	j	m	n	ref
3.2015	60	40800	j	m	j	style
7.2014	20400	34500	j	m	j	style
3.2015	70	37800	j	m	j	style

⋮

Multiple Lineare Regression
Residuenplot



diese roten sind viel billiger wie vorhergesagt

Gute
Angebote ?

entweder: schnäppchen oder nicht genügend erklärende variablen, zb unfallauto etc.

Probleme beim Autokauf…

- 30 (simulierte) Autos: Preis ~ KM + Baujahr

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 40190.450   3599.965  11.164 1.27e-11 ***
km             -6.156      3.268  -1.884   0.0704 .  
alter       90137.595  48869.929   1.844   0.0761 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8537 on 27 degrees of freedom
Multiple R-squared:  0.3724, Adjusted R-squared:  0.3259
F-statistic:  8.01 on 2 and 27 DF,  p-value: 0.001857
```

not significant

signifikant

F-Test und t-Tests widersprechen sich

problem: kollinerität

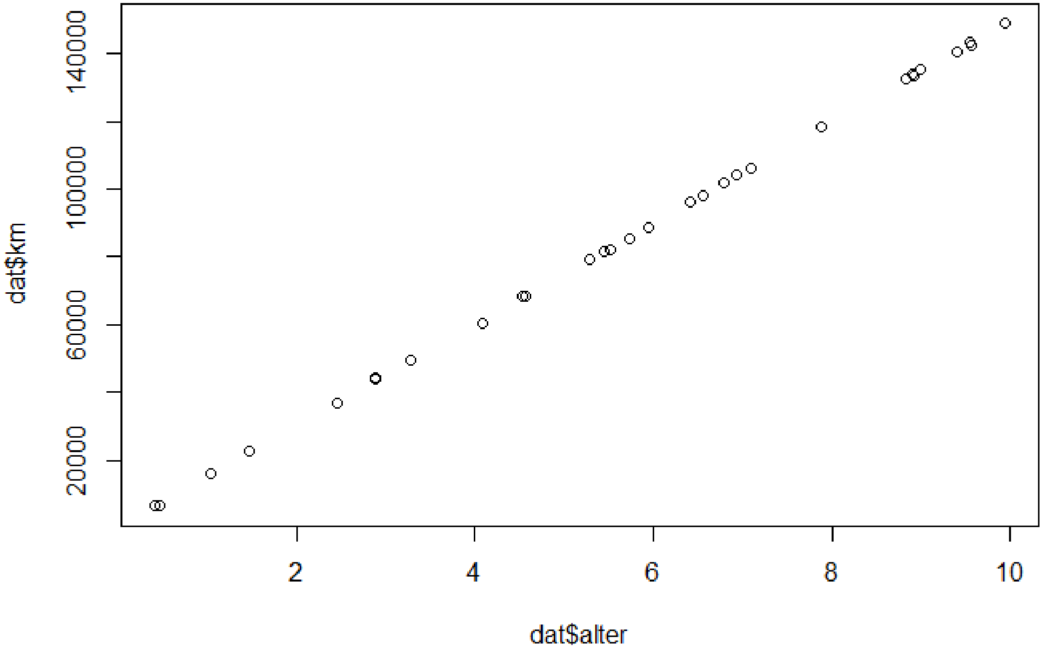
wenn F und t-test streiten, gewinnt der F-test jeweils

Interpretation der Parameter sehr schwierig

Kollinearität

- Kollinearität: Zwei erklärende Variablen sind stark korreliert
- Problem 1: Interpretation der Parameter schwierig
- Problem 2: Es kann sein, dass t-Tests keine Signifikanz mehr finden können
  (F-Test schon noch)
- Einfachste Lösung: Eine der beiden Variablen weglassen
- (Komplexere Lösung: Orthogonalisieren; nicht in dieser Vorlesung)

Probleme wegen Kolinearität: km und alter sind stark korreliert



Cor(km, alter) = 0.9999

Lösung: Lasse eine der beiden Variablen weg

Preis ~ km

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.858e+04  3.639e+03  10.602 2.61e-11 ***
km          -1.293e-01  3.792e-02  -3.409    0.002 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 8896 on 28 degrees of freedom
Multiple R-squared:  0.2933, Adjusted R-squared:  0.2681
F-statistic: 11.62 on 1 and 28 DF,  p-value: 0.001995
```

Pro km wird das Auto ca. 10 Rappen billiger

Preis ~ alter

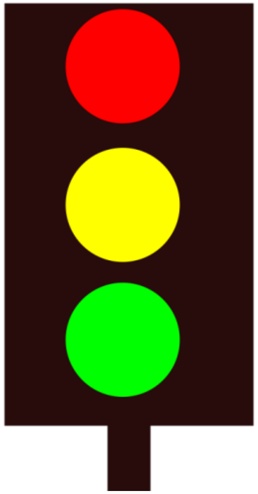
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  38477.5     3638.3  10.576 2.76e-11 ***
alter        -1921.7      568.4  -3.381  0.00215 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 8918 on 28 degrees of freedom
Multiple R-squared:  0.2899, Adjusted R-squared:  0.2645
F-statistic: 11.43 on 1 and 28 DF,  p-value: 0.002145
```

Pro Jahr wird das Auto ca. 2000 SFr billiger vif > 10

4> vif > 10

vif < 10



Weiterführend

- Kollinearität erkennen:
  Variance inflation factor **vif** (s. Wikipedia)
  Funktion vif() in Paket car
  Faustregel: S. Ampel
  ISLR Kap 3.3.3, v.a. Abschnitt 6

- Kolliniearität beheben:
  Orthogonalisieren, z.B. PCA (s. VL 13)