Woche 10: Lineare Regression

Patric Müller <patric.mueller@stat.math.ethz.ch>

ETHZ

WBL 23/25, 26.06.2023

Teil XII

Einfache Lineare Regression

Lernziele

Sie können...

- ...ein lineares Regressionsmodell aufschreiben und dessen Komponenten erläutern.
- ...eine lineare Regression in R durchführen.
- ...prüfen, ob ein Datensatz die Modellannahmen der linearen Regression erfüllt.

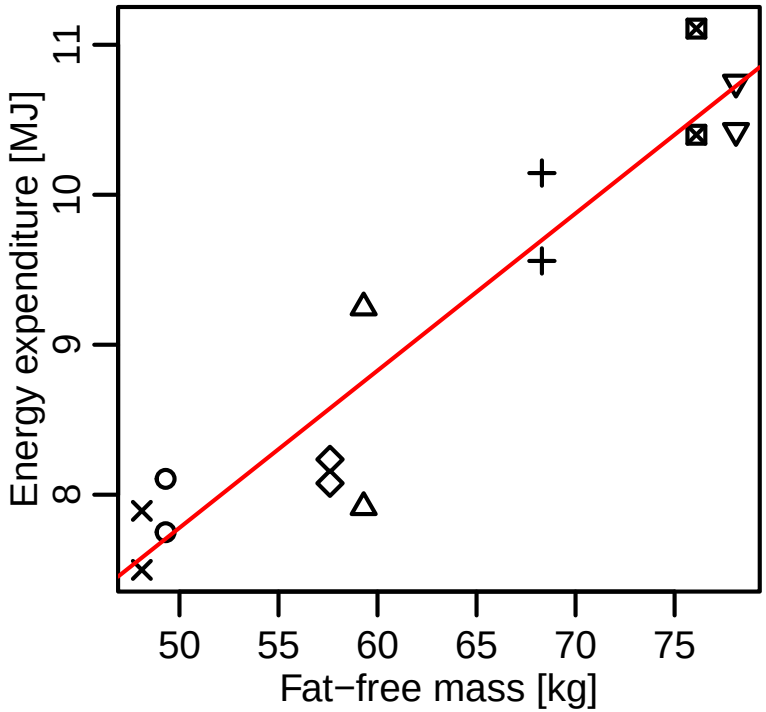
Vorlesung basiert auf Kapitel 5.2 des Skripts.

Zusammenfassung

- Korrelation
  - $Corr(X, Y)$ erklärt, wie gut die Daten auf einer Geraden liegen.
  - X_i und Y_i sind "gepaart".
- Tests
  - Ein Test ist eine Entscheidungsregel. Das Testergebnis entscheidet, ob die zwei Stichproben unterschiedliche Verteilungen haben.
  - X_i und Y_i sind gepaart oder ungepaart.
- Lineare Regression
  - Mit der linearen Regression schätzt man die (lineare) Beziehung zwischen X und Y.
  - Das heisst, man bestimmt die bestmögliche Gerade, die zu den Daten passt.
  - X_i und Y_i sind "gepaart".
  - Zusätzlich beeinflusst der Prädiktor X_i die Zielvariable Y_i.

Beispiel: Körpermasse und Energieumsatz

- In einer Studie wurde die fettfreie Masse und der Energieumsatz während 24h von 7 Probanden gemessen (2 Messungen pro Proband) (Webb, 1981).
- Der tägliche Energieumsatz scheint linear mit der fettfreien Masse zu wachsen.



- Der Zusammenhang zwischen der fettfreien Masse und dem Energieumsatz liesse sich z.B. durch Korrelation ausdrücken. Die Korrelation alleine ist aber nicht hilfreich, um für eine bekannte fettfreie Masse den dazugehörigen Energieumsatz *vorherzusagen*!
- *Ziel: Eine Formel für den Energieumsatz als Funktion der Masse bestimmen.*

Einfache lineare Regression

Modell für einfache lineare Regression:

$$Y_i = \beta_0 + \beta_1 x_i + E_i, \ \ i = 1, \ldots, n,$$
$$\text{wobei} \quad E_1, \ldots, E_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

Var.	Bezeichnung	Bedeutung	Beispiel
Y_i	**Zielvariable**	Variable, die wir vorhersagen wollen	Energieumsatz
x_i	**erklärende Variable**, **Co-Variable**	bekannte oder einfach zu messende Variable	fettfreie Masse
E_i	**Fehlervariable** oder **Rauschterm**	Abweichung von perfekter Geraden	

Fehlervariablen modellieren

- (nicht erklärbare) individuelle Abweichungen von einem Mittelwert
- nicht gemessene Einflüsse auf die Zielvariable: z.B. sportliche Aktivität, Personenunterschied,...

Einfache lineare Regression: Namensbedeutung

Modell für einfache lineare Regression:

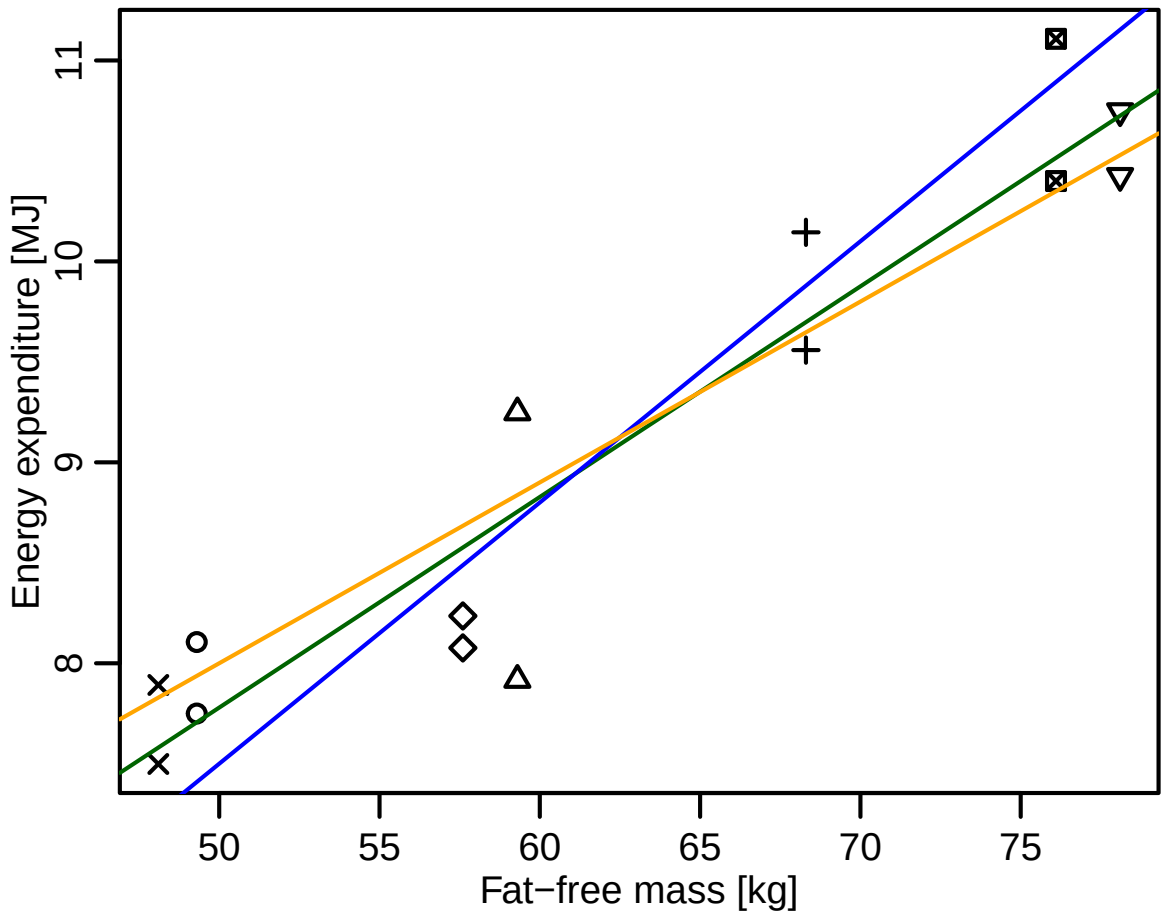
$$Y_i = \beta_0 + \beta_1 x_i + E_i, \quad i = 1, \ldots, n,$$
$$E_1, \ldots, E_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

Bezeichnung des Modells:

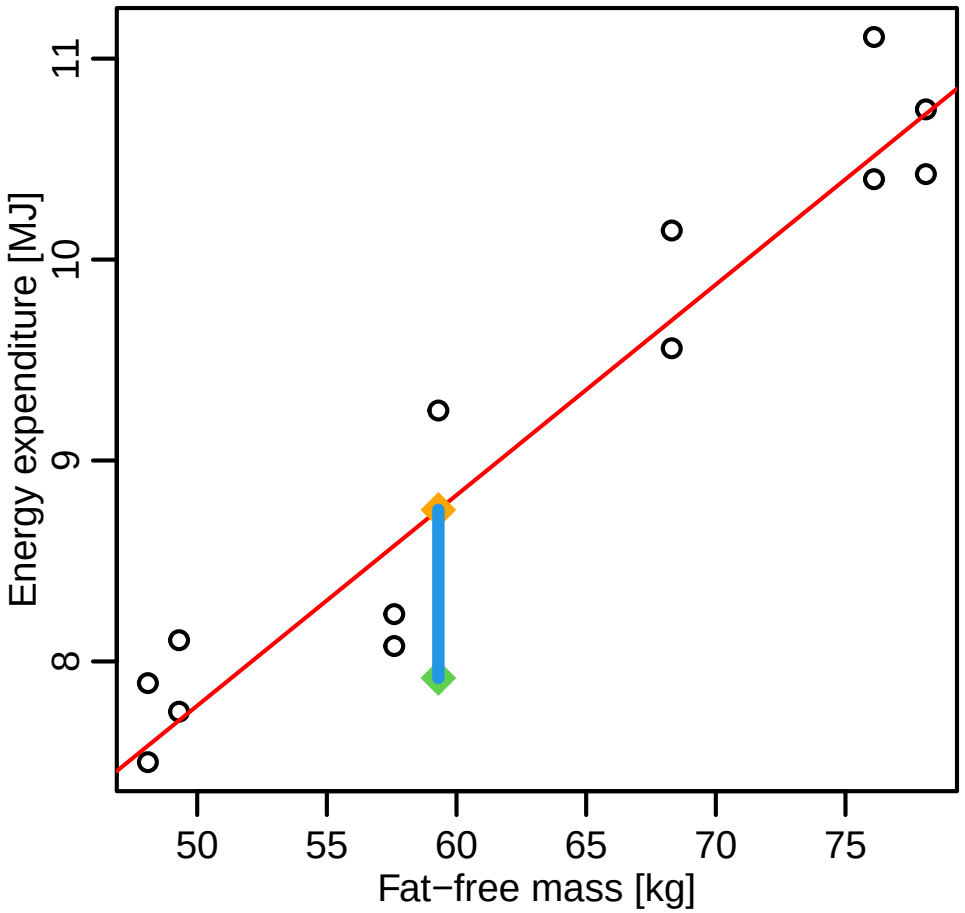
- „einfach“: nur eine erklärende Variable (andernfalls: „multiple“ lineare Regression, s. Kurs „Regression“)
- „linear“: Zielvariable ist *lineare* Funktion der Koeffizienten β_0, β_1
- **Bemerke:** Eine “versteckte Annahme” der linearen Regression ist, dass die Prädiktoren x_i exakt (ohne Fehler) gemessen wurden.

Regressionsgerade schätzen

Wie finden wir eine Gerade, die gut auf die Daten passt? Welche der gezeigten Geraden passt am besten? Was bedeutet mathematisch "am besten passen"?



Regressionsmodell graphisch



- Regressionsgerade:
  $y = \hat{\beta}_0 + \hat{\beta}_1 x,$
  $y = 2.538 + 0.105x$
- Beobachtung: (x_i, Y_i)
- Vorhersagewert: $(x_i, \hat{Y}_i)$
- Residuum: $Y_i - \hat{Y}_i$

Regressionsmodell schätzen: Überblick

Modell:

$$Y_i = \beta_0 + \beta_1 x_i + E_i, \ i = 1, \ldots, n,$$
$$\text{wobei} \quad E_1, \ldots, E_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

- Ziel: Parameter β_0, β_1 (und σ^2) schätzen.
- Mit $\hat{\beta}_0$ und $\hat{\beta}_1$ bezeichnen wir die geschätzten Parameter.
- x_i und Y_i sind gemessene Daten.
- E_i sind die (zufälligen) Fehler. Ohne Fehlerterm würden die Daten perfekt auf der Geraden $Y = \beta_0 + \beta_1 x$ liegen.

Methode der kleinsten Quadrate

- β_0 und β_1 werden mit der **Methode der kleinsten Quadrate** geschätzt.
- Das heisst durch Minimierung der **Summe der Fehlerquadrate** ("residual sum of squares")

$$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

$$(\hat{\beta}_0; \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \left(\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \right)$$

- σ^2 schätzen durch Varianzschätzung der **Residuen** $R_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

Residuen

- **Residuum**: Differenz zwischen gemessenem und angepasstem Wert der Zielvariablen.
- Gemessener Wert: Y_i (im Beispiel: gemessener Energieumsatz)
- Angepasster Wert für die gegebenen Parameter β_0^* und β_1^*: $Y^* = \beta_0^* + \beta_1^* x_i$.
  - In der Regel sind die "gegebenen Parameter" die geschätzten Parameter, somit: $\hat{\beta}_0 + \hat{\beta}_1 x_i =: \hat{Y}_i$

Definition (Residuen)

Das i-te **Residuum** ($i = 1, \ldots, n$) ist definiert als $R_i = Y_i - \beta_0^* - \beta_1^* x_i$.
Die **Summe der Fehlerquadrate** ist definiert als $\text{RSS} = \sum_{i=1}^{n} R_i^2$.

Eigenschaften der Schätzer

Theorem

$\hat{\beta}_0$ und $\hat{\beta}_1$ sind erwartungstreue Schätzer für die wahren Koeffizienten β_0 und β_1.

- Fehlervariablen E_i sind nicht direkt messbar: wir kennen bloss die Werte von x_i und Y_i
- Mit Hilfe der Residuen konstruiert man einen **Schätzer für die Fehlervarianz**:

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

Lineare Regression in R

```
> energymass <- read.table("../Daten/energy.csv", sep = ",", header = TRUE)
> energymass$energy <- 4.1868e-3*energymass$energy # Einheiten umrechnen...
> energy.fit <- lm(energy ~ mass, data = energymass)
> summary(energy.fit)

Call:
lm(formula = energy ~ mass, data = energymass)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83689 -0.25948 -0.02941  0.37778  0.59247

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.53831    0.65519   3.874  0.00221 **
mass         0.10482    0.01033  10.143 3.07e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.433 on 12 degrees of freedom
Multiple R-squared:  0.8955, Adjusted R-squared:  0.8868
F-statistic: 102.9 on 1 and 12 DF,  p-value: 3.073e-07
```

Viel Output...

```
Call:
lm(formula = energy ~ mass, data = energymass)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83689 -0.25948 -0.02941  0.37778  0.59247

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.53831    0.65519   3.874  0.00221 **
mass         0.10482    0.01033  10.143 3.07e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.433 on 12 degrees of freedom
Multiple R-squared:  0.8955, Adjusted R-squared:  0.8868
F-statistic: 102.9 on 1 and 12 DF,  p-value: 3.073e-07
```

- Koeffizienten:
  $\hat{\beta}_0 = 2.53831,$
  $\hat{\beta}_1 = 0.10482$
- Standardabweichung der Fehlervariablen:
  $\hat{\sigma} = 0.433$
- Messpunkte in der Studie: $n =$ Freiheitsgrade $+$ Anzahl Koeffizienten;
  $n = 12 + 2 = 14$

Interpretation des Outputs I

Beantworten Sie die folgenden Fragen basierend auf dem Output auf der vorangehenden Folie:

- Angenommen, Sie wiegen 4kg mehr als Ihr Bruder; um wie viel ist dann der Erwartungswert Ihres täglichen Energieumsatzes höher als der Ihres Bruders?
- Wie gross ist der erwartete Energieumsatz einer Person mit einer fettfreien Masse von 65kg?
- Geben Sie ein 95%-Vertrauensintervall für den täglichen Energieumsatz einer Person mit einer fettfreien Masse von 65kg an.

Signifikanz des ganzen Modells: F-Test

- p-Wert der „F-Statistik“: Ergebnis eines Tests zur Nullhypothese „alle Koeffizienten ausser β_0 sind 0“.
- Relevant bei multipler linearer Regression; bei der einfachen linearen Regression bezieht sich der Test nur auf β_1.

```
Call:
lm(formula = energy ~ mass, data = energymass)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83689 -0.25948 -0.02941  0.37778  0.59247

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.53831    0.65519   3.874  0.00221 **
mass         0.10482    0.01033  10.143 3.07e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.433 on 12 degrees of freedom
Multiple R-squared:  0.8955, Adjusted R-squared:  0.8868
F-statistic: 102.9 on 1 and 12 DF,  p-value: 3.073e-07
```

Signifikanz einzelner erklärender Variablen: t-Test

- *Gibt es einen signifikanten Zusammenhang zwischen fettfreier Masse und Energieumsatz?*
- Das lässt sich mit einem t-Test zur Nullhypothese $\beta_1 = 0$ herausfinden.
- Teststatistik: $T = \dfrac{\text{calc. coeff.} - \text{exp. coeff.}}{\text{standard error}} = \dfrac{\hat{\beta}_1 - 0}{\widehat{\text{se}}(\hat{\beta}_1)}$

  Unter der Nullhypothese ist die Teststatistik t-verteilt mit $n - 2$ Freiheitsgraden
- Werte der Teststatistik und zugehörige p-Werte lassen sich aus dem R-Output ablesen
- Einfache lineare Regression: Der t-Test zur Steigung liefert den gleichen p-Wert wie der F-Test zum Modell.

Signifikanz und Vertrauensintervall für erklärende Variable

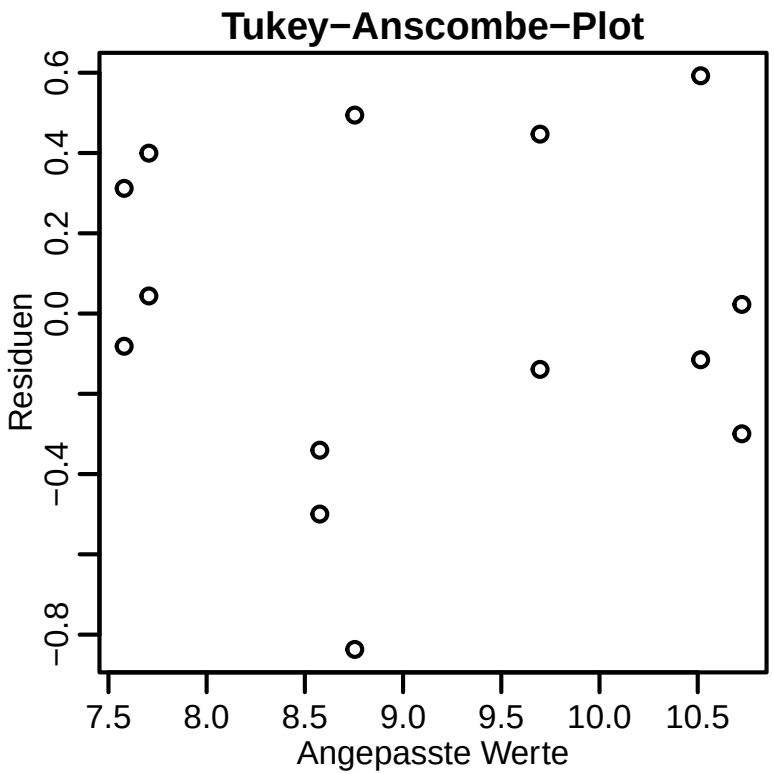
```
Call:
lm(formula = energy ~ mass, data = energymass)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83689 -0.25948 -0.02941  0.37778  0.59247

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.53831    0.65519   3.874  0.00221 **
mass         0.10482    0.01033  10.148 3.07e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.433 on 12 degrees of freedom
Multiple R-squared:  0.8955, Adjusted R-squared:  0.8868
F-statistic: 102.9 on 1 and 12 DF,  p-value: 3.073e-07
```

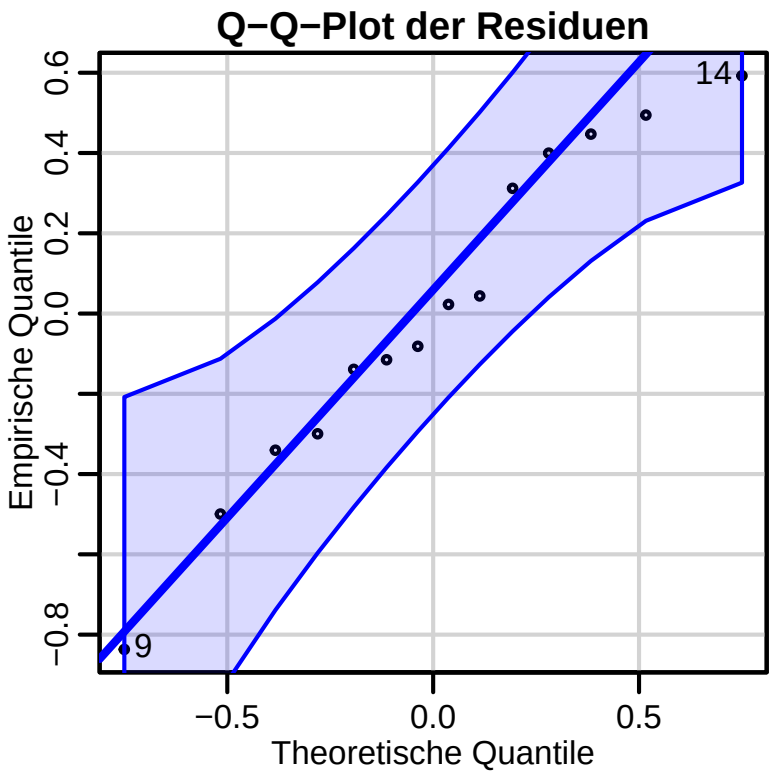
- Konfidenzintervall für β_1:
  $I = \left[\hat{\beta}_1 - \widehat{\mathrm{se}}(\hat{\beta}_1) t_{n-2,1-\frac{\alpha}{2}}, \hat{\beta}_1 + \widehat{\mathrm{se}}(\hat{\beta}_1) t_{n-2,1-\frac{\alpha}{2}}\right]$
- Hier: $I = [0.0823, 0.1273]$

Residuenanalyse: Überprüfung der Modellannahmen

Modellannahmen können im Wesentlichen mit zwei Plots überprüft werden:

- **Linearität** und **i.i.d.-Annahme** der Fehler mit dem **Tukey-Anscombe-Plot**: (Residuen gegen gefittete Werte).
- **Normalverteilungs-Annahme** der Fehler mit einem **Q-Q-Plot**



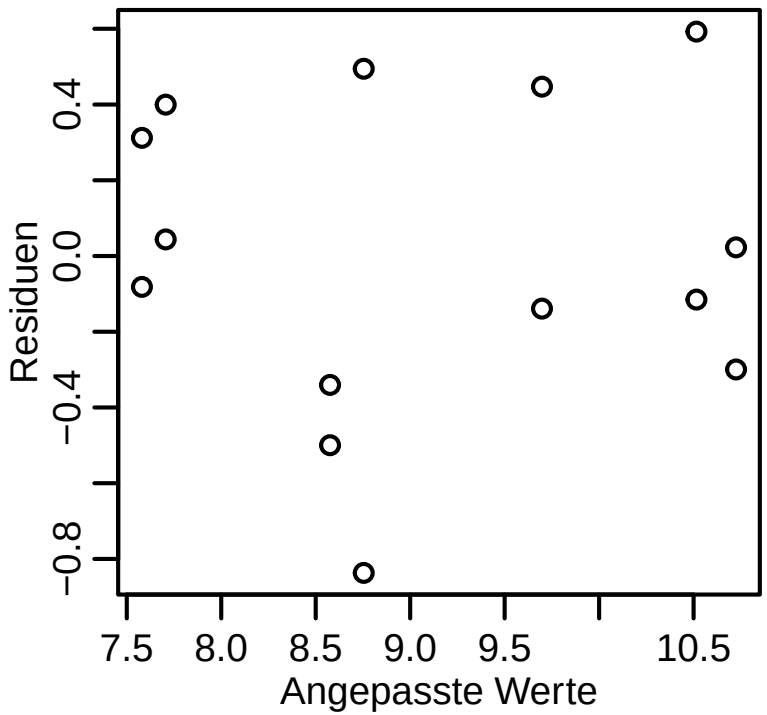


Tukey-Anscombe-Plot

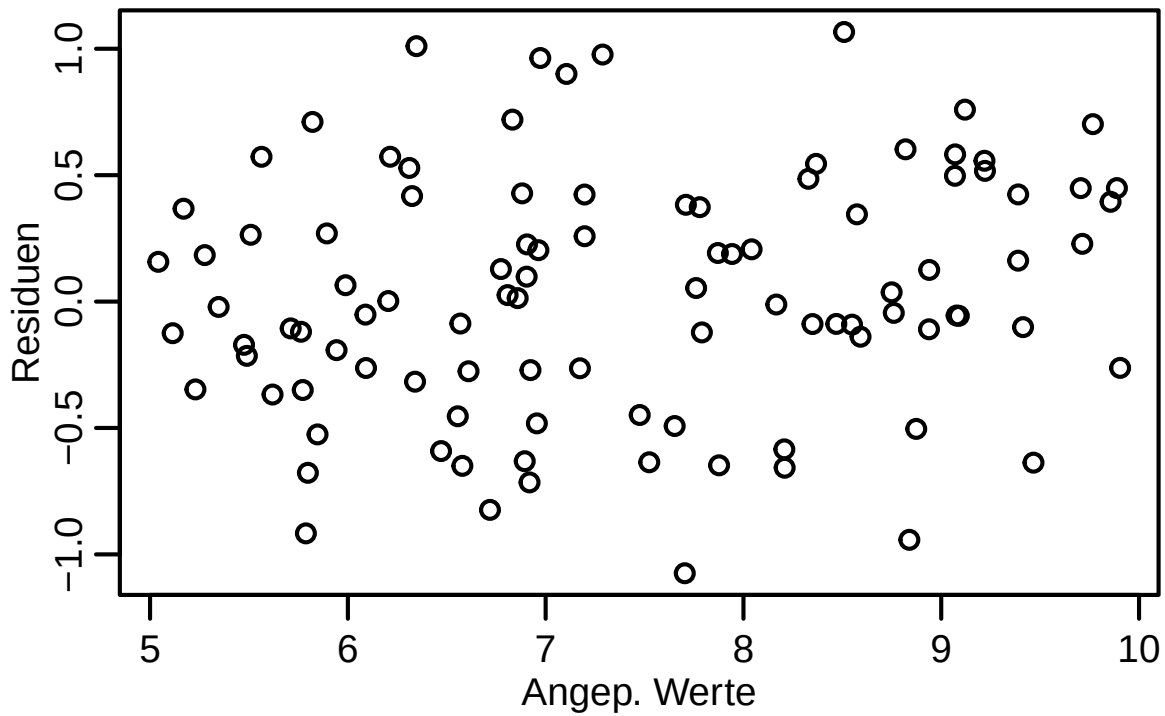
Linearität und i.i.d.-Annahme der Fehlervariablen prüfen:

- Residuen R_i gegen angepasste Werte der Zielvariable ($\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$) plotten (**Tukey-Anscombe plot**)
- Falls i.i.d.-Annahme zutrifft, sollten Punkte im Tukey-Anscombe-Plot zufällig um eine horizontale Gerade fluktuieren, ohne sichtbares Muster
- Tukey-Anscombe-Plot in R:

```
> plot(fitted(energy.fit), resid(energy.fit),
   xlab = "Angepasste Werte", ylab = "Residuen")
```

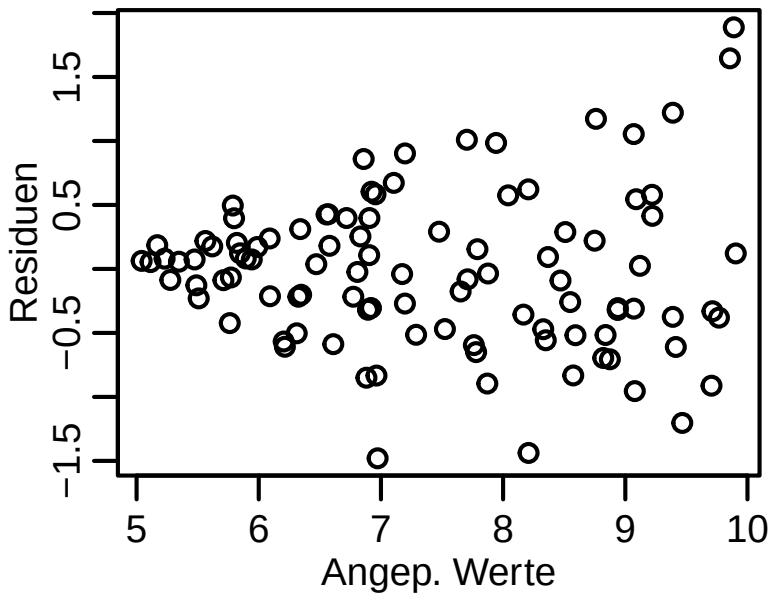


Guter Tukey-Anscombe-Plot



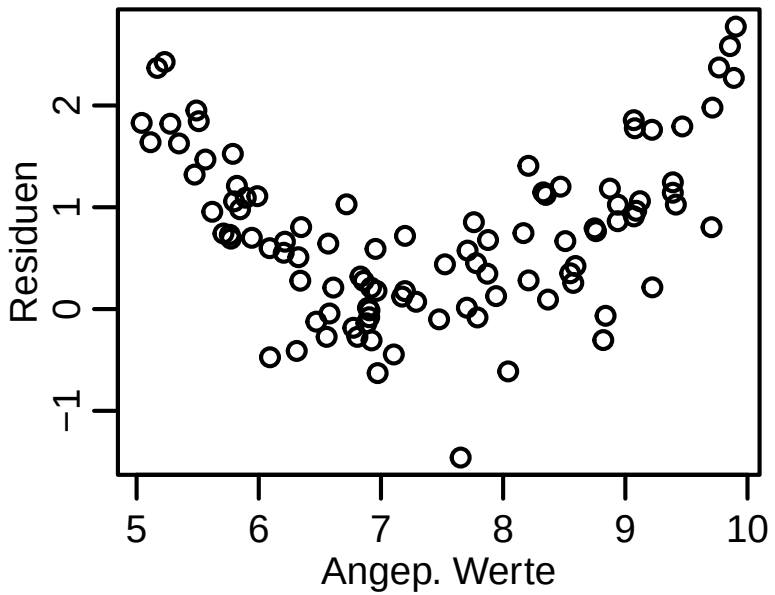
Sichtbare Verletzungen der Annahmen im Tukey-Anscombe-Plot

Unterschiedliche Fehlervarianzen:



Kegel; kann evtl. mit Hilfe einer log-Transformation behoben werden $Y_i \mapsto \log(Y_i)$

Nicht-linearer Trend:



Quadratischer Trend; kann evtl. durch Hinzufügen eines quadratischen Terms (x_i^2) behoben werden

Bestimmtheitsmass R^2

- Das **Bestimmtheitsmass** R^2 gibt an, wie gut die Datenpunkte auf einer Geraden liegen ("goodness of fit")
- Definition: Bestimmtheitsmass = quadrierter Korrelationskoeffizient zwischen gemessenen (Y_i) und angepassten ($\widehat{Y}_i$) Werten der Zielvariable, das heisst:

$$R^2 = \left(\frac{s_{\hat{y}y}}{s_{\hat{y}} s_y} \right)^2$$

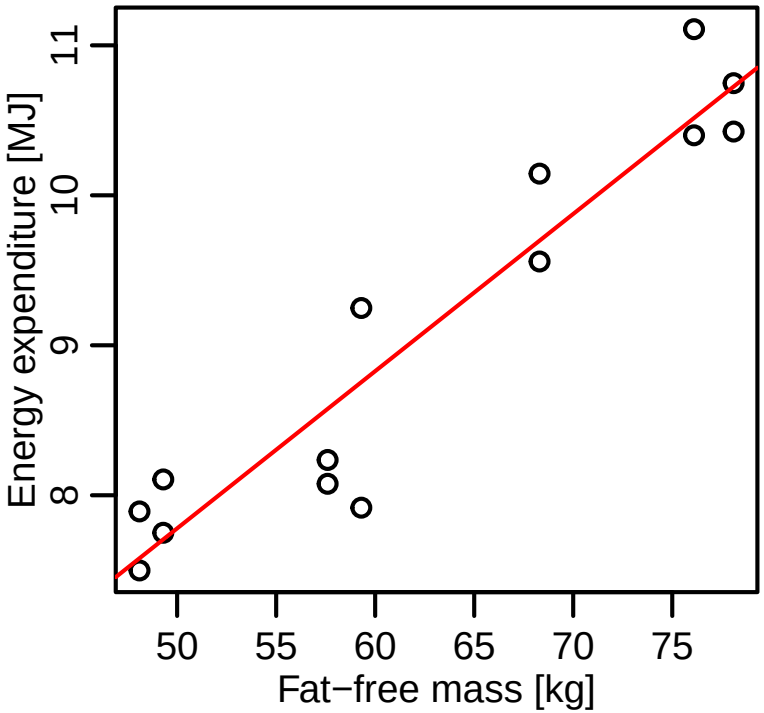
Bestimmtheitsmass im R-Output

```
Call:
lm(formula = energy ~ mass, data = energymass)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83689 -0.25948 -0.02941  0.37778  0.59247

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.53831    0.65519   3.874  0.00221 **
mass         0.10482    0.01033  10.143 3.07e-07 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.433 on 12 degrees of freedom
Multiple R-squared:  0.8955, Adjusted R-squared:  0.8868
F-statistic: 102.9 on 1 and 12 DF,  p-value: 3.073e-07
```

Manuelle Berechnung:

```
> cor(energymass$energy, fitted(energy.fit))^2
[1] 0.8955353
```

Einfache lineare Regression: Zusammenfassung



- Modell der einfachen linearen Regression:

$$Y_i = \beta_0 + \beta_1 x_i + E_i, \ \ i = 1, \ldots, n,$$

$$E_1, \ldots, E_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$$

- Zu schätzende Parameter: β_0, β_1 und σ^2

- β_0 und β_1 werden mit der „Methode der kleinsten Fehlerquadrate“ geschätzt; sie minimieren RSS $= \sum_{i=1}^{n} R_i^2$.
- Die Schätzung von σ^2 wird durch die Varianzschätzung der **Residuen** $R_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ berechent.

Literatur

Paul Webb. Energy expenditure and fat-free mass in men and women. *The American journal of clinical nutrition*, 34(9):1816–1826, 1981.