Regressionsanalyse

1. Einfache Lineare Regression
2. Inferenz in Regressionsmodellen
3. Diagnostische Aspekte
4. Simultane Inferenz
5. Matrix Algebra (Wiederholung)
6. Multiple Lineare Regression
7. Extra Quadratsummen
8. Qualitative Prädiktoren
9. Diagnostics/Residuenanalyse
10. Nichtparametrische (glatte) Modelle
11. Variablenselektion

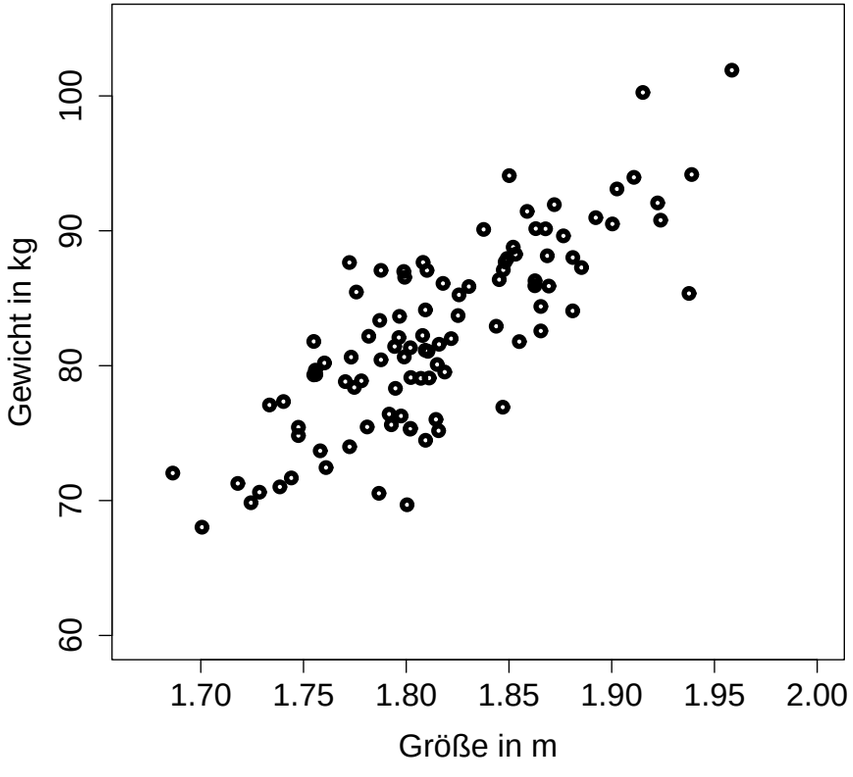
1. Einfache Lineare Regression

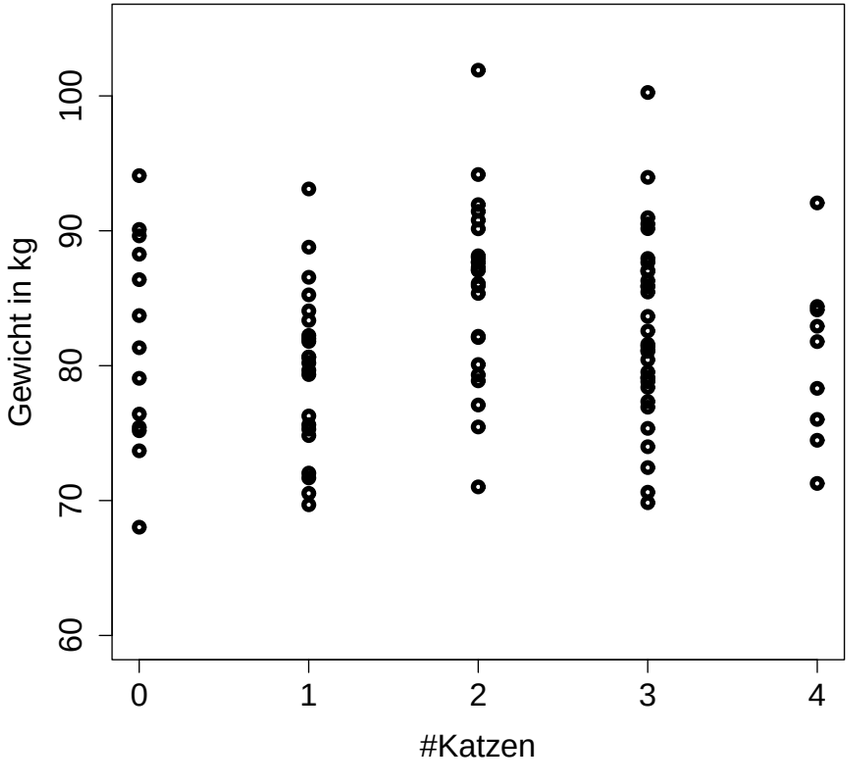
Angenommen, wir interessieren uns für das durchschnittliche Körpergewicht männlicher Bachelor Studenten an der TU Graz. Wir geben dazu die Namen all dieser Studenten (**Population**) in eine Urne und ziehen aus dieser zufällig 100 (**Stichprobe**). Hier sind sie: $Y_1, Y_2, \ldots, Y_{100}$.

Angenommen, wir messen zusätzlich auch deren Körpergrößen und die Anzahl der Katzen ihrer Eltern. Hier sind sie: $G_1, G_2, \ldots, G_{100}$ und $K_1, K_2, \ldots, K_{100}$.

Fragen: Wie würde man diese Daten verwenden, um das Durchschnittsgewicht

1. aller männlichen Studenten zu schätzen?
2. aller männlichen Studenten zu schätzen, die zwischen 1.70 und 1.75 m groß sind?
3. aller männlichen Studenten zu schätzen, deren Eltern 3 Katzen haben?


Gewicht in kg
60
70
80
90
100
1.70
1.75
1.80
1.85
1.90
1.95
2.00
Größe in m


Gewicht in kg
60
70
80
90
100
0
1
2
3
4
#Katzen

Antworten:

1. $\bar{Y} = \frac{1}{100} \sum_{i=1}^{100} Y_i$, das Stichprobenmittel.
2. Mittle die Y_i's all jener, deren G_i's zwischen 1.70 und 1.75 m sind.
3. Mittle die Y_i's all jener, deren K_i's genau 3 sind? **Nein!**
   Wie in 1., da das Gewicht sicherlich nicht von den elterlichen Katzen abhängt.

Intuitive Beschreibung von Regression:
(Gewicht) Y = interessierende Variable = Response Variable = abhängige Variable
(Größe) x = erklärende Variable = Prädiktorvariable = unabhängige Variable

Fundamentale Annahmen in der Regression:

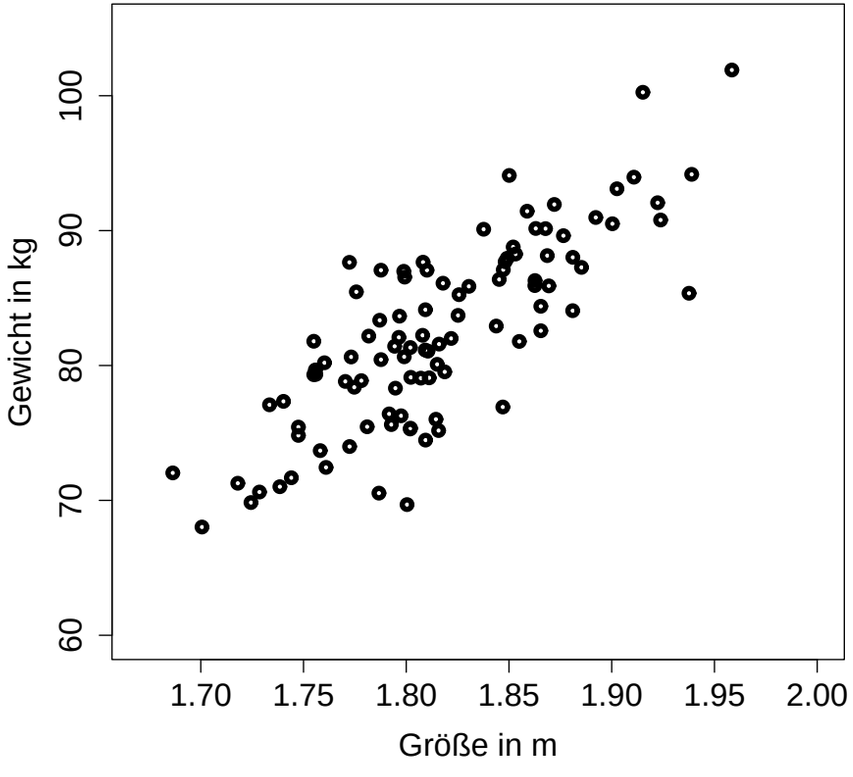
1. Für jeden einzelnen Wert x der Prädiktorvariablen, ist die Response Variable Y eine Zufallsvariable, deren Erwartungswert von x abhängt.
2. Der Erwartungswert von Y, $\mathrm{E}(Y)$, lässt sich als deterministische Funktion in x schreiben.

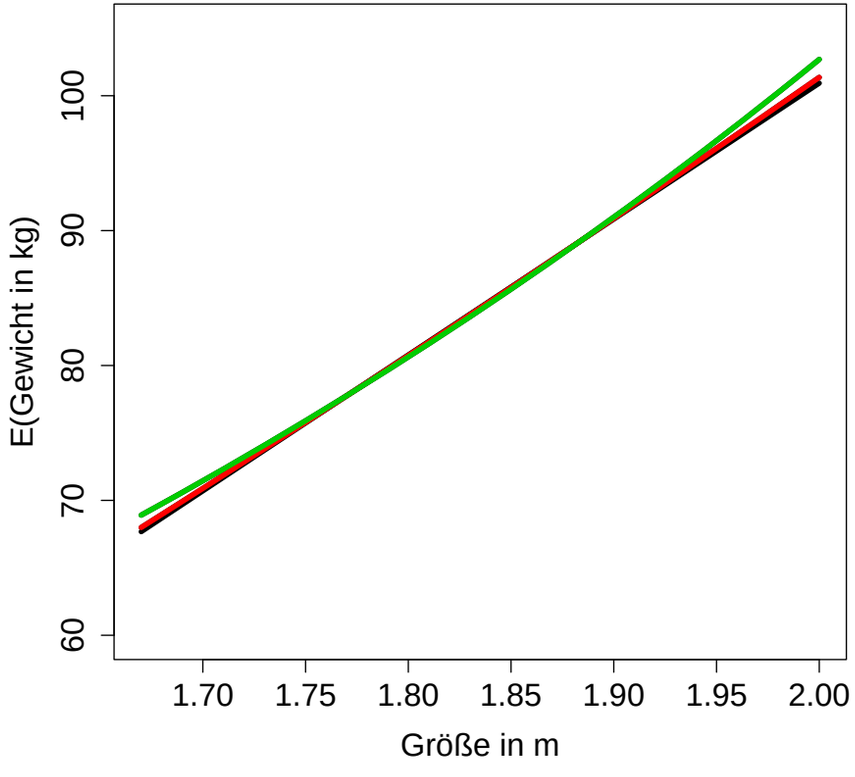
Beispiel: $\mathsf{E}(\text{Gewicht}_i) = f(\text{Größe}_i)$

$$\mathsf{E}(\text{Gewicht}_i) = \begin{cases} \beta_0 + \beta_1 \cdot \text{Größe}_i \\ \beta_0 + \beta_1 \cdot \text{Größe}_i + \beta_2 \cdot \text{Größe}_i^2 \\ \beta_0 \exp[\beta_1 \cdot \text{Größe}_i] \,, \end{cases}$$

wobei β_0, β_1, und β_2 **unbekannte Parameter** sind!

Scatterplot Größe gegen Gewicht (links) und (rechts) Größe gegen E(Gewicht):

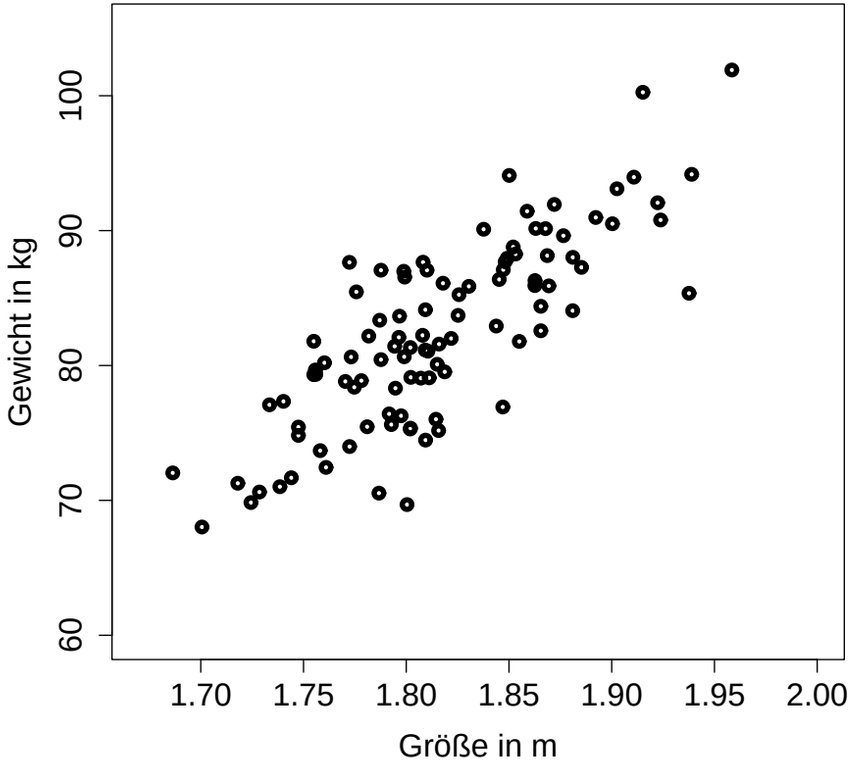


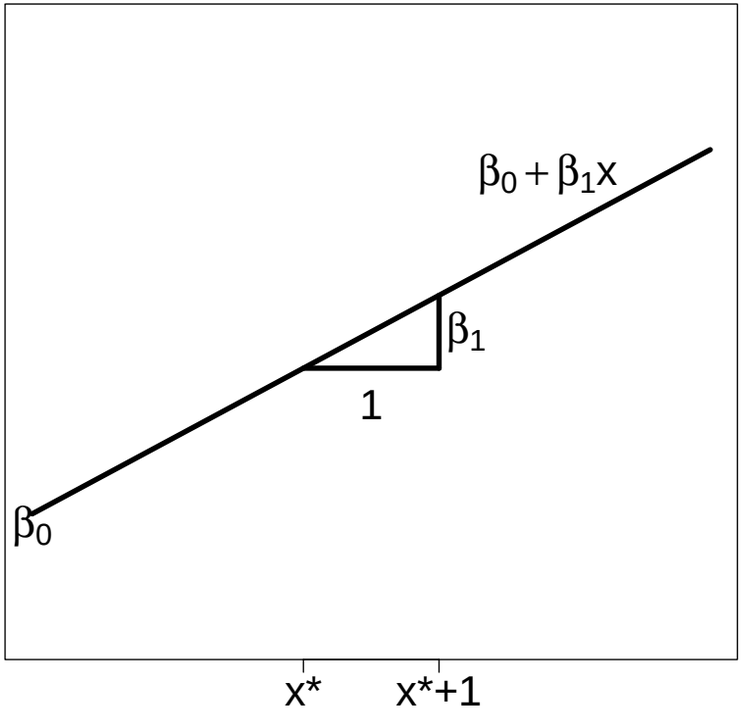


Einfache Lineare Regression (SLR)

Ein Scatterplot von 100 (x_i, Y_i) Paaren (Größe, Gewicht) weist darauf hin, dass es einen **linearen Trend** gibt.

Gleichung einer Geraden: $y = \beta_0 + \beta_1 \cdot x$ (**Konstante/Intercept** β_0 und **Steigung/Slope** β_1)





An der Stelle x^*: $y = \beta_0 + \beta_1 x^*$
An der Stelle $x^* + 1$: $y = \beta_0 + \beta_1(x^* + 1)$
Differenz beträgt: $(\beta_0 + \beta_1(x^* + 1)) - (\beta_0 + \beta_1 x^*) = \beta_1$

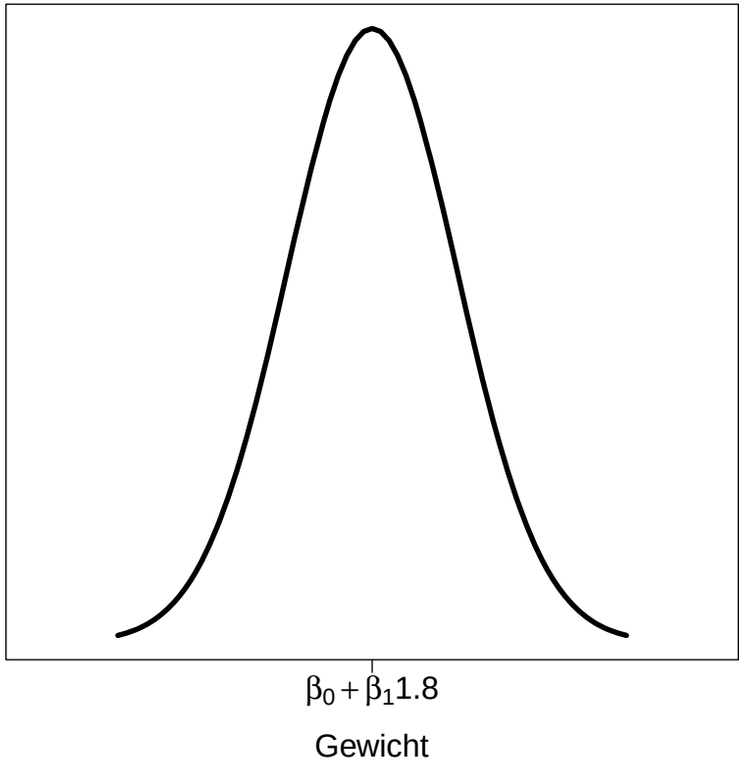
Gilt: Gewicht $= \beta_0 + \beta_1 \cdot$ Größe ? (**funktionale Beziehung**)

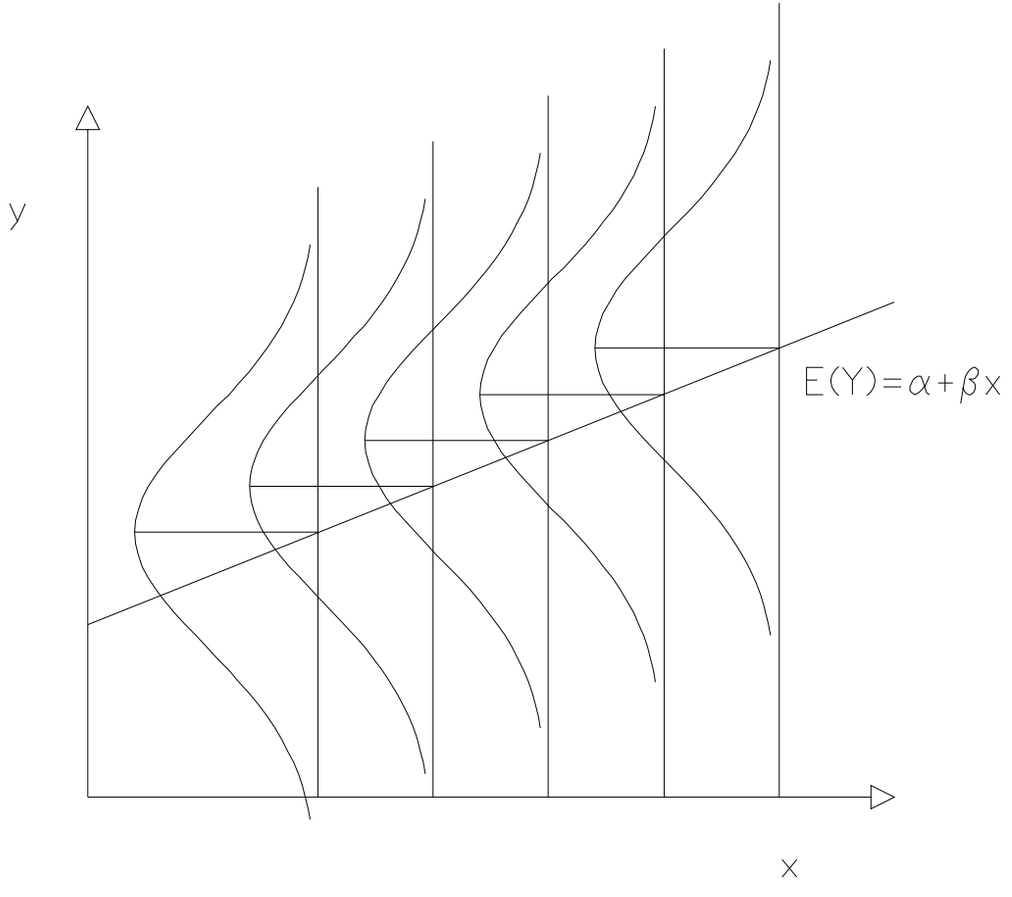
Nein! Dies ist eine **statistische Beziehung** und bei Weitem nicht perfekt!

Wir können aber sagen, dass: E(Gewicht) $= \beta_0 + \beta_1 \cdot$ Größe

Dies heißt: **Gewicht ist eine Zufallsvariable** und der **Erwartungswert von Gewicht ist eine lineare Funktion in Größe**.

Wie sieht beispielsweise die **Verteilung** des Gewichts einer Person aus, die 1.80 m groß ist, d.h. E(Gewicht) $= \beta_0 + \beta_1 \cdot 1.80$.


$\beta_0 + \beta_1 1.8$
Gewicht


y
$E(Y) = \alpha + \beta x$
x

Formale Definition des SLR Modells

Daten: $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$

Gleichung:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i\,, \qquad i = 1, 2, \ldots, n$$

Annahme:

- Y_i ist die **Response Variable** im i-ten Versuch,
- die x_i's sind **feste, bekannte Konstanten**,
- die ϵ_i's sind unabhängige und identisch verteilte Zufallsfehler, sogenannte nicht beobachtbare **statistische Fehler**, mit $\mathrm{E}(\epsilon_i) = 0$ und $\mathrm{var}(\epsilon_i) = \sigma^2$,
- β_0, β_1 und σ^2 sind **unbekannte Parameter** (Konstanten).

Konsequenzen des SLR Modells

- Die Response Y_i ist die Summe des konstanten Terms $\beta_0 + \beta_1 x_i$ und des zufälligen Terms ϵ_i. Daher ist Y_i eine Zufallsvariable.

- Die ϵ_i's sind unabhängig und da jedes Y_i nur ein ϵ_i beinhaltet, sind auch die Y_i's unabhängig.

- $\mathsf{E}(Y_i) = \mathsf{E}(\beta_0 + \beta_1 x_i + \epsilon_i) = \beta_0 + \beta_1 x_i = \mu(x_i)$.
  Regressionsfunktion (bringt den Erwartungswert von Y in Beziehung mit x) ist
  $$\mathsf{E}(Y) = \mu(x) = \beta_0 + \beta_1 x\,.$$

- $\mathsf{var}(Y_i) = \mathsf{var}(\beta_0 + \beta_1 x_i + \epsilon_i) = \mathsf{var}(\epsilon_i) = \sigma^2$.
  Daher gilt: $\mathsf{var}(Y_i) = \sigma^2$ (**gleiche**, konstante Varianz für **alle** Y_i's).

Warum nennt man dieses Modell *SLR*?

Simple/einfach: nur ein Prädiktor x_i,

Linear: Regressionsfunktion $\mathsf{E}(Y) = \beta_0 + \beta_1 x$ ist linear in den Parametern.

Warum *interessiert* uns ein Regressionsmodell?

Falls das Modell realistisch ist und falls wir glaubwürdige Schätzer der beiden Parameter β_0 und β_1 haben, dann:

1. können wir ein neues Y_i an einem neuen x_i vorhersagen, und

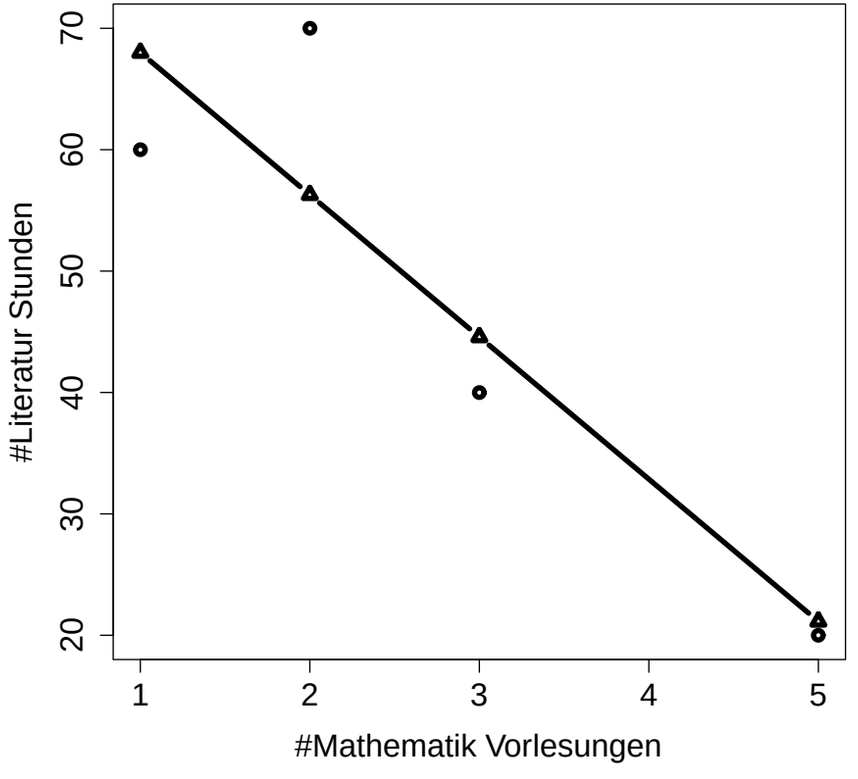
2. haben ein besseres Verständnis darüber, wie sich der Erwartungswert von Y_i, also $\mathsf{E}(Y_i) = \mu(x_i)$, mit x_i ändert.

Kleinste Quadrate (Least Squares) Schätzung von β_0 und β_1

$x_i = \#$Mathematik-Vorlesungen, die der i-te Student belegt
$Y_i = \#$Stunden, welche der i-te Student mit Literatur verbringt

Zufällig gewählte 4 Studenten:

$(x_1, y_1) = (1, 60)$,
$(x_2, y_2) = (2, 70)$,
$(x_3, y_3) = (3, 40)$,
$(x_4, y_4) = (5, 20)$



Nehmen wir für diese Daten ein SLR an, dann nehmen wir dadurch an, dass es in jedem x eine Verteilung der Literaturstunden gibt und dass die Erwartungswerte aller Response Variablen auf einer Geraden liegen.

Wir brauchen **Schätzer der unbekannten Parameter** β_0, β_1 und σ^2. Konzentrieren wir uns zuerst einmal auf β_0 und β_1.

Jedes Paar (β_0, β_1) definiert eine Gerade $\beta_0 + \beta_1 x$. Das **Kleinste Quadrate Kriterium** fordert, jene Gerade zu nehmen, die die Summe der quadrierten vertikalen Distanzen der Punkte (x_i, Y_i) zur Geraden $(x_i, \beta_0 + \beta_1 x_i)$ **minimiert**.

Formell minimieren die Kleinsten Quadrate Schätzer $\hat{\beta}_0$ und $\hat{\beta}_1$ das Kriterium

$$\mathsf{SSE}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \mu(x_i))^2 = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i))^2 \,,$$

das die Summe aller quadrierten vertikalen Distanzen von den Punkten zur Geraden darstellt (Fehlerquadratsumme oder **S**um of **S**quared **E**rrors).

Anstatt SSE für jede mögliche Gerade $\beta_0 + \beta_1 x$ auszuwerten, berechnen wir das optimale β_0 und β_1. Wir minimieren die Funktion SSE bezüglich β_0 und β_1

$$\begin{aligned}
\frac{\partial \mathsf{SSE}(\beta_0, \beta_1)}{\partial \beta_0} &= \sum_{i=1}^{n} 2(Y_i - (\beta_0 + \beta_1 x_i))(-1) \\
\frac{\partial \mathsf{SSE}(\beta_0, \beta_1)}{\partial \beta_1} &= \sum_{i=1}^{n} 2(Y_i - (\beta_0 + \beta_1 x_i))(-x_i)\,.
\end{aligned}$$

Nullsetzen liefert die beiden **Normalgleichungen** (sehr wichtig!)

$$\begin{aligned}
\sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) &= 0 \\
\sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) x_i &= 0\,.
\end{aligned}$$

Vereinfacht ergibt dies

$$\begin{aligned} \hat{\beta}_0 n + \hat{\beta}_1 n\bar{x} &= n\bar{Y} \\ \hat{\beta}_0 n\bar{x} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i Y_i \end{aligned}$$

und somit

$$\begin{aligned} \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i Y_i - n\bar{x}\ \bar{Y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{xY}^2}{s_x^2} \end{aligned}$$

mit

$$\begin{aligned} s_x^2 &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \\ s_{xY}^2 &= \sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})\,. \end{aligned}$$

Dieses Ergebnis ist **sogar noch wichtiger**! Wir verwenden die zweiten Ableitungen um zu zeigen, dass wir damit ein Minimum erhalten haben.

Sei

$$s_Y^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2\,.$$

Die minimale Fehlerquadratsumme ist (da $s_{xY}^2 = \hat{\beta}_1 s_x^2$ gilt)

$$\begin{aligned}\mathsf{SSE}(\hat{\beta}_0, \hat{\beta}_1) &= \sum_{i=1}^{n} \Big(Y_i - (\bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i)\Big)^2 = \sum_{i=1}^{n} \Big(Y_i - \bar{Y} - \hat{\beta}_1 (x_i - \bar{x})\Big)^2 \\ &= s_Y^2 - 2\hat{\beta}_1 s_{xY}^2 + \hat{\beta}_1^2 s_x^2 = s_Y^2 - \hat{\beta}_1^2 s_x^2 = s_Y^2 - s_{xY}^4 / s_x^2 \,.\end{aligned}$$

Alle Datenpunkte (x_i, y_i) liegen genau dann auf der geschätzten Regressionsgeraden $\hat{\mu}(x)$, wenn $\mathsf{SSE}(\hat{\beta}_0, \hat{\beta}_1) = 0$ und somit $s_{xY}^4 = s_x^2 s_Y^2$ gilt, also wenn

$$\left(\frac{s_{xY}^2}{\sqrt{s_x^2 s_Y^2}}\right)^2 = \widehat{\mathsf{cor}}^2(x, Y) = 1 \,,$$

d.h., falls perfekte (negative oder positive) Korrelation zwischen den x Werten und den Response Variablen Y vorliegt.

Beispiel: Wir berechnen die Schätzer der Parameter und erhalten mit

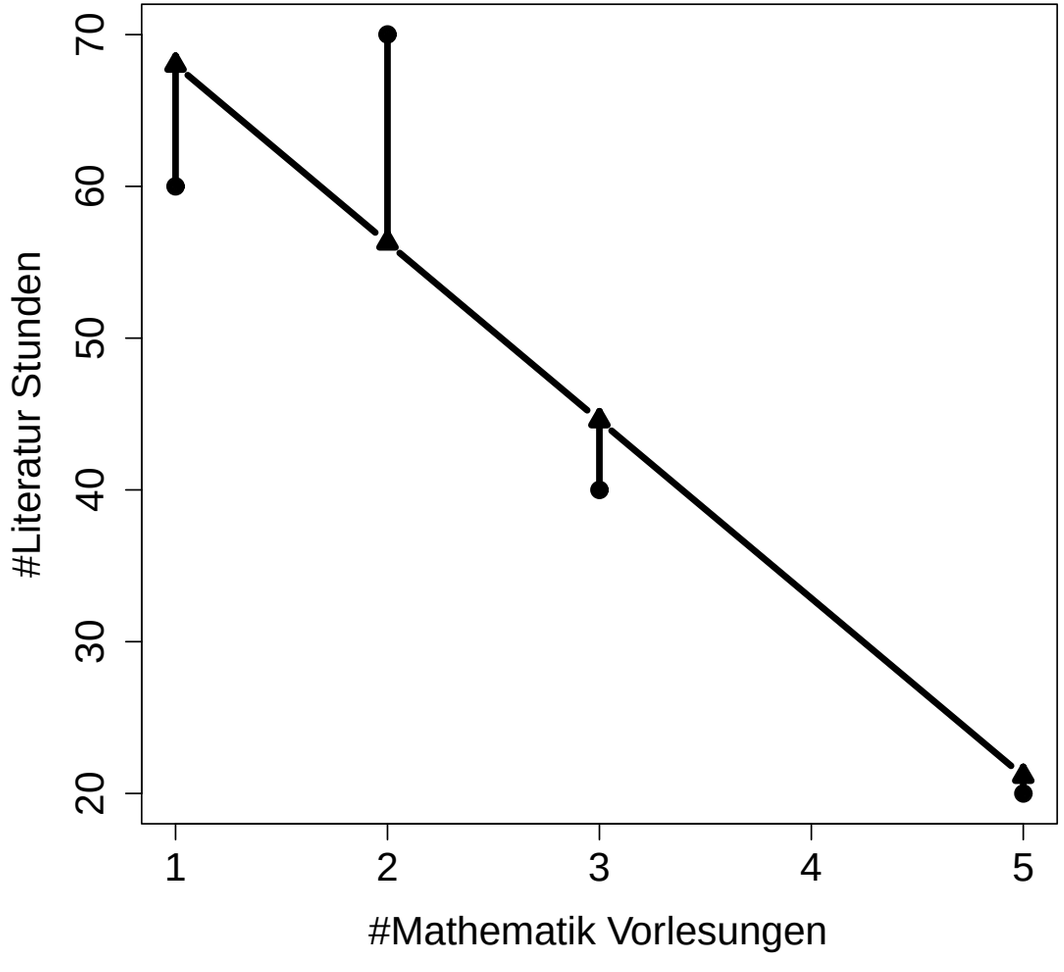
$\sum_i x_i y_i = 420$, sowie $\sum_i x_i = 11$, $\sum_i y_i = 190$, $\sum_i x_i^2 = 39$

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\big(\sum_{i=1}^n x_i\big)\big(\sum_{i=1}^n y_i\big)}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\big(\sum_{i=1}^n x_i\big)^2} = -11.7 \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 80.0 \end{aligned}$$

Die geschätzte Regressionsfunktion lautet somit

$$\widehat{\mathsf{E}(Y)} = 80 - 11.7x$$

An der Stelle $x = 1$: $\widehat{\mathsf{E}(Y)} = 80 - 11.7 \cdot 1 = 68.3$
An der Stelle $x = 5$: $\widehat{\mathsf{E}(Y)} = 80 - 11.7 \cdot 5 = 21.5$


#Literatur Stunden
20
30
40
50
60
70
1
2
3
4
5
#Mathematik Vorlesungen

Eigenschaften des Kleinsten Quadrate Schätzers

Ein wichtiger Satz, das sogenannte *Gauß Markov Theorem*, sagt aus, dass der Kleinste Quadrate Schätzer unverzerrt ist und minimale Varianz unter allen unverzerrten, linearen Schätzern hat.

Punktschätzer des Erwartungswerts:
Unter dem SLR Modell lautet die Regressionsfunktion

$$\mathsf{E}(Y) = \beta_0 + \beta_1 x\,.$$

Verwende die Schätzer von β_0 und β_1, um damit die geschätzte Regressionsfunktion zu konstruieren, d.h.

$$\widehat{\mathsf{E}(Y)} = \hat{\beta}_0 + \hat{\beta}_1 x\,.$$

Prognose- oder Vorhersagewert (Fitted Value): Schätzer für den Erwartungswert $\mu(x_i)$ von Y_i unter dem Modell

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \bar{Y} + \hat{\beta}_1 (x_i - \bar{x}) \, .$$

Residuum: beobachtbarer Fehler

$$r_i = Y_i - \hat{\mu}_i \, .$$

Beachte: das Residuum r_i ist keineswegs identisch mit dem nicht beobachtbaren, statistischen Fehler ϵ_i. Vergleiche dazu

$$\begin{aligned} r_i &= Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \\ \epsilon_i &= Y_i - \beta_0 - \beta_1 x_i \, . \end{aligned}$$

Daher verhält sich r_i so etwa wie $\hat{\epsilon}_i$, aber ϵ_i ist **kein** Parameter!

Eigenschaften der geschätzten Regressionsgeraden

Wiederholung: Mit $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ gilt $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, sowie

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n}(x_i - \bar{x})x_i = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \,.$$

- Die Summe der Residuen ist Null, d.h.

$$\sum_{i=1}^{n} r_i = 0 \,.$$

- Die Summe der quadrierten Residuen ist minimal.
- Die geschätzte Regressionsgerade geht immer durch den Punkt $(\bar{x}, \bar{Y})$.

- Die Summe der Responses entspricht der Summe ihrer Prognosen, d.h.

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{\mu}_i \,.$$

- Die Summe der mit x_i gewichteten Residuen ist Null, d.h.

$$\sum_{i=1}^{n} x_i r_i = 0 \,.$$

- Die Summe der mit $\hat{\mu}_i$ gewichteten Residuen ist Null, d.h.

$$\sum_{i=1}^{n} \hat{\mu}_i r_i = 0 \,.$$

Schätzung von σ^2 unter dem SLR:

Motivation vom iid-Fall (unabhängig und identisch verteilt):
• Sei $Y_1, \ldots, Y_n$ eine Zufallsstichprobe mit $\mathsf{E}(Y_i) = \mu$ und $\mathsf{var}(Y_i) = \sigma^2$.

Stichprobenvarianz (2 Schritte):

1. Betrachte
$$\sum_{i=1}^{n} (Y_i - \widehat{\mathsf{E}(Y_i)})^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \,,$$
quadriere Differenzen zwischen Responses und geschätzten Erwartungswerten.

2. Dividiere durch Freiheitsgrade (degrees of freedom)
$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \,.$$
Hier geht 1 Freiheitsgrad verloren, da wir 1 Parameter μ schätzen.

Betrachte nun ein SLR Modell mit $\mathsf{E}(Y_i) = \beta_0 + \beta_1 x_i$ und $\mathsf{var}(Y_i) = \sigma^2$, wobei die Responses zwar unabhängig aber natürlich nicht identisch verteilt sind.

Die entsprechenden 2 Schritte lauten:

1. Betrachte

$$\sum_{i=1}^{n} (Y_i - \widehat{\mathsf{E}(Y_i)})^2 = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \mathsf{SSE}(\hat{\beta}_0, \hat{\beta}_1)\,,$$

   quadriere Differenzen zwischen Responses und geschätzten Erwartungswerten.

2. Dividiere durch Freiheitsgrade (degrees of freedom)

$$S^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2 = \frac{1}{n-2} \mathsf{SSE}(\hat{\beta}_0, \hat{\beta}_1) =: \mathsf{MSE}(\hat{\beta}_0, \hat{\beta}_1)\,.$$

   Hier gehen 2 Freiheitsgrade verloren, da wir 2 Parameter β_0 und β_1 schätzen.

Eigenschaften des Schätzers für σ^2:

Der MSE (mittlere quadratische Fehler) ist ein **unverzerrter Schätzer** von σ^2, d.h.

$$\mathsf{E}(\mathsf{MSE}(\hat{\beta}_0, \hat{\beta}_1)) = \sigma^2$$

(Beweis später für den multiplen Fall).

SLR: Regressionsmodell mit normalverteilten Responses

Unabhängig von der angenommenen Verteilung der Fehlerterme ϵ_i liefert die **Kleinste Quadrate** Methode **unverzerrte** Punktschätzer für β_0 und β_1, welche noch dazu **minimale Varianz** unter allen unverzerrten, linearen Schätzern aufweisen.

Um jedoch Konfidenzintervalle zu konstruieren und statistische Hypothesentests durchführen zu können, müssen wir zusätzlich auch eine Annahmen über die Verteilung der ϵ_i treffen.

Das **Regressionsmodell mit normalverteilten Responses** ist definiert als:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \,, \qquad i = 1, \ldots, n \,.$$

Annahmen:

- Y_i ist die **Response** im i-ten Versuch,
- die x_i's sind **feste, bekannte Konstanten**,
- die ϵ_i's sind unabhängig Normal$(0, \sigma^2)$ verteilte **statistische Zufallsfehler**,
- β_0, β_1 und σ^2 sind konstante, **unbekannte Parameter**.

Dies impliziert, dass die Responses unabhängige Zufallsvariablen sind, mit

$$Y_i \overset{ind}{\sim} \text{Normal}(\beta_0 + \beta_1 x_i, \sigma^2) \,.$$

Motivation zur Inferenz in SLR Modellen

Sei x_i die Anzahl von Geschwistern und Y_i die Anzahl von Stunden, die man mit Literatur verbringt.

Daten $(1, 20), (2, 50), (3, 30), (5, 30)$ resultieren im geschätzten SLR

$$\widehat{\mathsf{E}(Y)} = 33 + 0.3 \cdot x$$

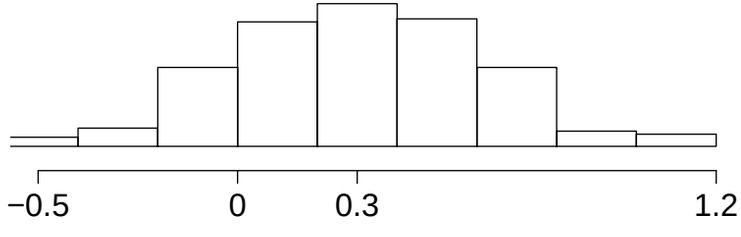
Folgerung: Da $\hat{\beta}_1$ ungleich Null ist, hängt deshalb die zu erwartende Stundenanzahl linear von der Geschwisteranzahl ab? Stimmt das?

Nein, das ist falsch!

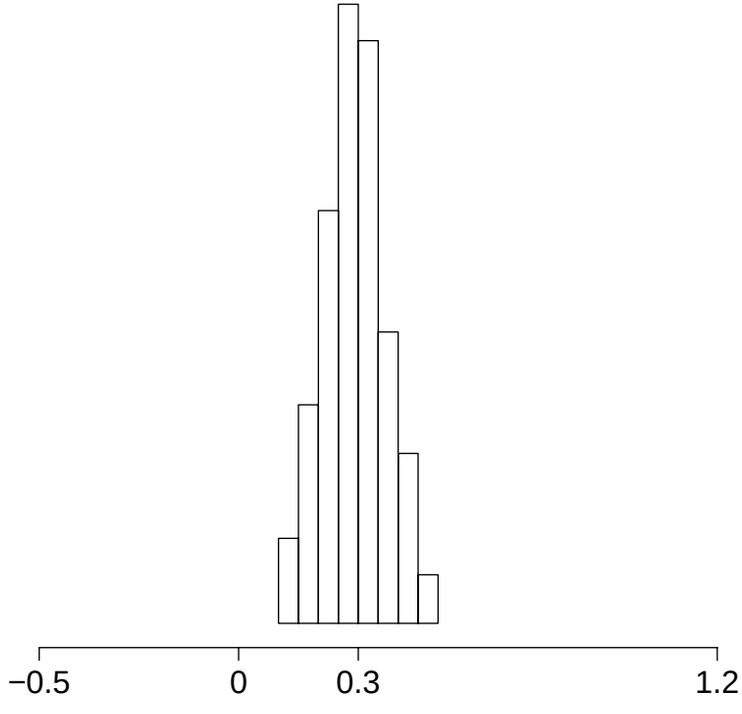
$\hat{\beta}_1$ ist auch eine Zufallsvariable, weil dieser Schätzer von den Y_i's abhängt.

Denke an eine nacheinander folgende Datensammlung und berechne jedesmal $\hat{\beta}_1$ für jeden Datensatz. Wir zeichnen ein Histogramm all dieser $\hat{\beta}_1$'s:

Szenario 1: stark variierend



Szenario 2: stark konzentriert



Betrachte $H_0 : \beta_1 = 0$

Ist H_0 falsch? Unter Szenario 1: nicht sicher,

Unter Szenario 2: mit großer Sicherheit!

Kennen wir die exakte Verteilung von $\hat{\beta}_1$, dann können wir formal entscheiden, ob H_0 wahr ist. Wir benötigen einen formalen statistischen Test von:

$H_0 : \beta_1 = 0$ (keine Abhängigkeit)

$H_1 : \beta_1 \neq 0$ (es gibt eine lineare Beziehung zwischen E(Y) und x)

2. Inferenz im Regressionsmodell

Falls $Y_i \overset{ind}{\sim} \text{Normal}(\mu_i, \sigma_i^2)$, und $a_1, \ldots, a_n$ feste Konstanten sind, dann folgt

$$\sum_{i=1}^{n} a_i Y_i \sim \text{Normal}\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right) .$$

Somit ist eine Linearkombination von unabhängigen, normalverteilten Zufallsvariablen auch selbst wiederum eine normalverteilte Zufallsvariable.

Die Kleinsten Quadrate Schätzer $\hat{\beta}_0$ und $\hat{\beta}_1$ im SLR sind Linearkombinationen der normalverteilten Responses Y_i's, denn es gilt

$$\begin{aligned}
\hat{\beta}_1 &= \frac{1}{s_x^2}\sum_{i=1}^{n}(x_i-\bar{x})(Y_i-\bar{Y}) = \sum_{i=1}^{n}\frac{x_i-\bar{x}}{s_x^2}Y_i = \sum_{i=1}^{n}a_iY_i \\
\hat{\beta}_0 &= \frac{1}{n}\sum_{i=1}^{n}Y_i - \bar{x}\sum_{i=1}^{n}a_iY_i = \sum_{i=1}^{n}\left(\frac{1}{n}-a_i\bar{x}\right)Y_i = \sum_{i=1}^{n}b_iY_i
\end{aligned}$$

mit den Konstanten

$$a_i = \frac{x_i-\bar{x}}{s_x^2}, \qquad b_i = \frac{1}{n} - \bar{x}\frac{x_i-\bar{x}}{s_x^2}.$$

Wegen

$$\begin{aligned}
\sum a_i &= \frac{1}{s_x^2}\sum(x_i-\bar{x}) = 0 \\
\sum a_i x_i &= \frac{1}{s_x^2}\sum(x_i-\bar{x})x_i = 1 \\
\sum a_i^2 &= \frac{1}{s_x^4}\sum(x_i-\bar{x})^2 = \frac{1}{s_x^2}
\end{aligned}$$

folgt

$$\begin{aligned}
\mathsf{E}(\hat{\beta}_1) &= \sum a_i\,\mathsf{E}(Y_i) = \sum a_i(\beta_0+\beta_1 x_i) = \beta_1 \\
\mathsf{var}(\hat{\beta}_1) &= \sum a_i^2 \mathsf{var}(Y_i) = \frac{\sigma^2}{s_x^2} = \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\,.
\end{aligned}$$

Wegen

$$\begin{aligned}
\sum b_i &= 1 - \frac{\bar{x}}{s_x^2}\sum(x_i - \bar{x}) = 1 \\
\sum b_i x_i &= \bar{x} - \frac{\bar{x}}{s_x^2}\sum(x_i - \bar{x})x_i = 0 \\
\sum b_i^2 &= \frac{1}{n} + \bar{x}^2 \sum \frac{(x_i - \bar{x})^2}{s_x^4} - 2 \cdot 0 = \frac{1}{n} + \frac{\bar{x}^2}{s_x^2}
\end{aligned}$$

folgt

$$\begin{aligned}
\mathsf{E}(\hat{\beta}_0) &= \sum b_i \,\mathsf{E}(Y_i) = \sum b_i(\beta_0 + \beta_1 x_i) = \beta_0 \\
\mathsf{var}(\hat{\beta}_0) &= \sum b_i^2 \mathsf{var}(Y_i) = \sigma^2 \frac{1}{n} + \sigma^2 \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\,.
\end{aligned}$$

Somit erhalten wir unter der Annahme

$$Y_i \overset{ind}{\sim} \text{Normal}(\beta_0 + \beta_1 x_i, \sigma^2)$$

für die beiden Schätzer

$$\begin{aligned}
\hat{\beta}_0 &\sim \text{Normal}\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right)\right) \\
\hat{\beta}_1 &\sim \text{Normal}\left(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})^2}\right) .
\end{aligned}$$

Frage: Welche Schätzer sind unabhängig?

Für zwei beliebige lineare Formen $\mathbf{a}^t\mathbf{y}$ und $\mathbf{b}^t\mathbf{y}$ mit $\mathbf{a} = (a_1, \ldots, a_n)^t$, $\mathbf{b} = (b_1, \ldots, b_n)^t$, und $\mathbf{y} = (y_1, \ldots, y_n)^t \sim \text{Normal}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$, mit Erwartungsvektor $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^t$ und Varianz/Kovarianzmatrix proportional zur $n \times n$ Einheitsmatrix $\mathbf{I}_n$, gilt

$$\begin{aligned} \text{cov}(\mathbf{a}^t\mathbf{y}, \mathbf{b}^t\mathbf{y}) &= \text{E}\left(\mathbf{a}^t(\mathbf{y}-\boldsymbol{\mu})\mathbf{b}^t(\mathbf{y}-\boldsymbol{\mu})\right) = \text{E}\left(\mathbf{a}^t(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^t\mathbf{b}\right) \\ &= \mathbf{a}^t\text{var}(\mathbf{y})\mathbf{b} = \sigma^2\mathbf{a}^t\mathbf{b}\,. \end{aligned}$$

Somit sind $\mathbf{a}^t\mathbf{y}$ und $\mathbf{b}^t\mathbf{y}$ genau dann unabhängig, wenn das Skalarprodukt der Koeffizientenvektoren verschwindet, also wenn

$$\mathbf{a}^t\mathbf{b} = 0\,.$$

Betrachte $\bar{y} = \mathbf{a}^t\mathbf{y}$ und $\hat{\beta}_1 = \mathbf{b}^t\mathbf{y}$. Beides sind lineare Formen in $\mathbf{y}$ mit

$$\begin{aligned} \mathbf{a} &= \frac{1}{n}(1,\ldots,1)^t \\ \mathbf{b} &= \frac{1}{s_x^2}(x_1-\bar{x},\ldots,x_n-\bar{x})^t\,. \end{aligned}$$

Da

$$\begin{aligned} \mathbf{a}^t\mathbf{b} &= \frac{1}{n}(1,\ldots,1)\frac{1}{s_x^2}\begin{pmatrix} x_1-\bar{x} \\ \vdots \\ x_n-\bar{x} \end{pmatrix} \\ &= \frac{1}{ns_x^2}\sum_{i=1}^{n}(x_i-\bar{x}) = 0 \end{aligned}$$

sind $\bar{y}$ und $\hat{\beta}_1$ stochastisch unabhängig. Bemerke jedoch, dass zwischen $\bar{y}$ und $\hat{\beta}_0$ keine Unabhängigkeit besteht.

Beispiel: Von 93 Häusern in Gainesville/Florida die im Dezember 1995 verkauft wurden, kennt man die Preise. Wir haben:

$Y =$ Preis (in 1000\$), $x =$ Wohnfläche (in 1000 square feet).

Wir nehmen an, dass dafür ein SLR hält mit

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 x_i \,.$$

Die Kleinsten Quadrate Schätzer realisieren in $\hat{\beta}_0 = -25.2$ und $\hat{\beta}_1 = 75.6$.

Wir interessieren uns für einen Test der Hypothesen

$H_0 : \beta_1 = 0$ (keine Beziehung zwischen Fläche und Preis) gegen $H_1 : \beta_1 \neq 0$.

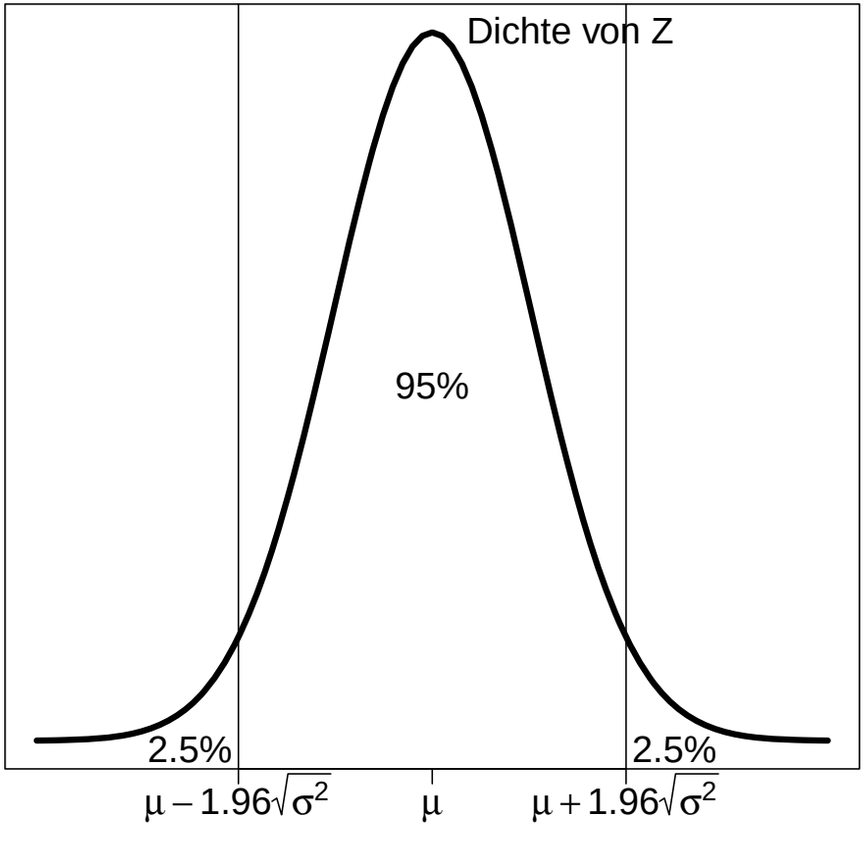
Da $75.6 \neq 0$, können wir somit schließen, dass nicht H_0 sondern H_1 wahr ist?

Zur Erinnerung: $\hat{\beta}_1 \sim$ Normal $\left(\beta_1, \frac{\sigma^2}{s_x^2}\right)$, hier mit $s_x^2 = \sum_i (x_i - \bar{x})^2 = 25.38$.

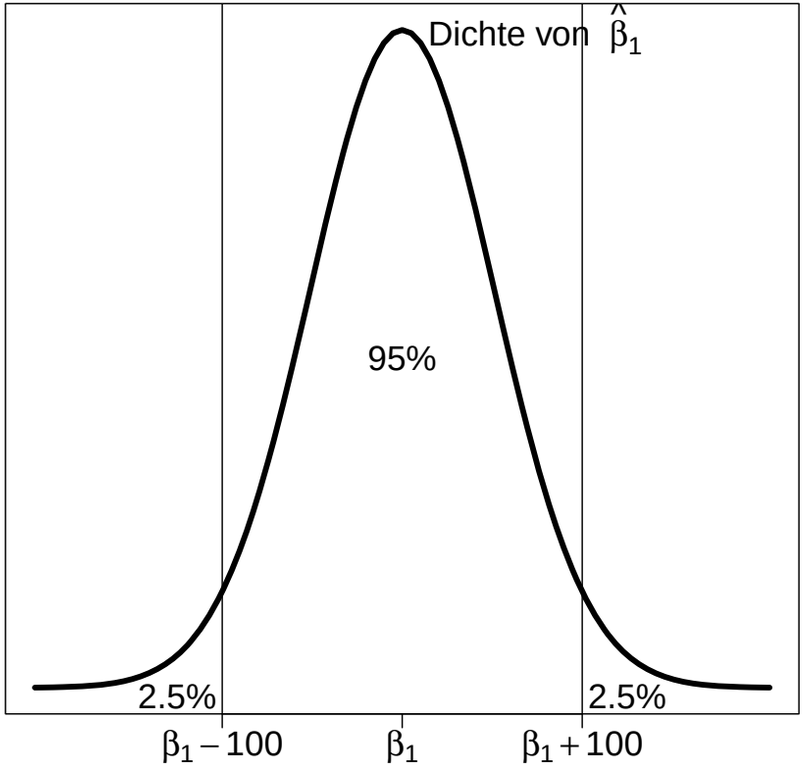
Betrachte die beiden Szenarios:

Szenario 1: $\sigma^2/s_x^2 = 2500 \Rightarrow \sqrt{\sigma^2/s_x^2} = 50$
Szenario 2: $\sigma^2/s_x^2 = \ \ 100 \Rightarrow \sqrt{\sigma^2/s_x^2} = 10$

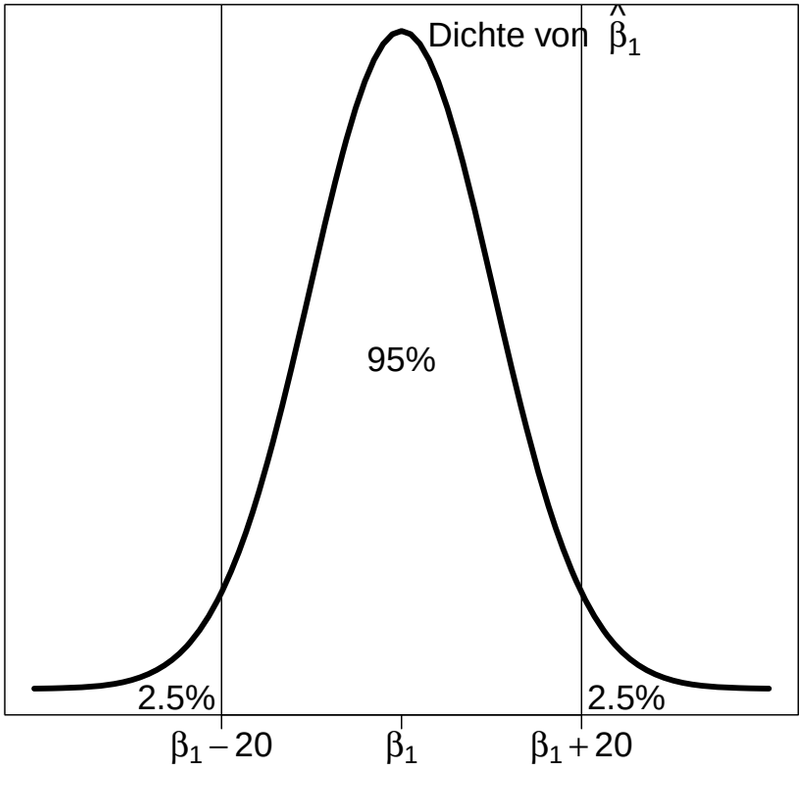
Zur Erinnerung, falls $Z \sim \text{Normal}(\mu, \sigma^2)$, dann



Scenario 1: $\sqrt{\sigma^2/s_x^2} = 50$



Scenario 2: $\sqrt{\sigma^2/s_x^2} = 10$



Szenario 1: Falls $\beta_1 = 0$ (H_0 wahr), dann besteht eine 95% Chance, dass $\hat{\beta}_1$ zwischen -100 und 100 liegt.

$\hat{\beta}_1 = 75.6$ ist somit konsistent mit $H_0 : \beta_1 = 0$.

Szenario 2: Falls $\beta_1 = 0$ (H_0 wahr), dann besteht eine 95% Chance, dass $\hat{\beta}_1$ zwischen -20 und 20 liegt.

$\hat{\beta}_1 = 75.6$ legt somit nahe, dass $H_0 : \beta_1 = 0$ falsch ist.

Fazit: Kennen wir $\sqrt{\sigma^2/s_x^2}$, dann wissen wir wie wahrscheinlich der Wert $\hat{\beta}_1 = 75.6$ unter H_0 ist und wir können entscheiden, ob $\hat{\beta}_1 = 75.6$ eher konsistent mit $H_0 : \beta_1 = 0$ oder mit $H_1 : \beta_1 \neq 0$ ist.

Wir haben bereits gezeigt, dass

$$\hat{\beta}_1 \sim \text{Normal}(\beta_1, \sigma^2/s_x^2) \qquad \Rightarrow \qquad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/s_x^2}} \sim \text{Normal}(0,1)\,.$$

Damit folgt

$$\begin{aligned} \Pr\left(-1.96 \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/s_x^2}} \leq 1.96\right) &= 0.95 \\ \Pr\left(\hat{\beta}_1 - 1.96\sqrt{\sigma^2/s_x^2} \leq \beta_1 \leq \hat{\beta}_1 + 1.96\sqrt{\sigma^2/s_x^2}\right) &= 0.95\,. \end{aligned}$$

Somit ist

$$\hat{\beta}_1 \pm 1.96\sqrt{\sigma^2/s_x^2}$$

ein **95% Konfidenzintervall** für β_1. Ist dies ein nützliches Intervall? **Nein!**

Wir müssen σ^2 unter dem SLR Modell schätzen. Zur Erinnerung ist der mittlere quadratische Fehler

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \mathsf{MSE}(\hat{\beta}_0, \hat{\beta}_1)$$

ein unverzerrter Schätzer für σ^2. Damit haben wir alles was notwendig ist!

Was folgt nun?

1. Tests und Konfidenzintervalle für β_1,
2. Konfidenzintervalle für den Erwartungswert von Y an einer beliebigen Stelle von x, z.B. x^*, also für
$$\mu(x^*) = \beta_0 + \beta_1 x^*\,,$$
3. Prädiktionsintervalle für weitere Responsevariablen beobachtbar in $x = x^*$.

Konfidenzintervalle und Tests für β_1

Der Schlüssel ist: $\hat{\beta}_1 \sim \text{Normal}(\beta_1, \sigma^2/s_x^2)$. Daher gilt

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2/s_x^2}} \sim \text{Normal}(0, 1)\,.$$

Aber dies ist nicht nützlich, weil wir den Wert von σ^2 nicht kennen.

Ersetzen wir σ^2 durch seinen Schätzer $S^2 = \text{MSE}$, so erhalten wir

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{MSE}/s_x^2}} \sim t_{n-2}\,.$$

Alles beruht auf diesem Ergebnis (Beweis später)!

Im Folgenden

- bezeichnet α die Type 1 Error Wahrscheinlichkeit, also $\Pr(\text{verwerfe } H_0 | H_0 \text{ ist wahr})$,
- ist α immer zwischen 0 und 1 (es ist eine Wahrscheinlichkeit),
- ist α gewöhnlich auf Werte wie 0.01, 0.05 oder 0.10 gesetzt.

Konfidenzintervalle für β_1

Mit Wahrscheinlichkeit $1-\alpha$ ist

$$-t_{1-\alpha/2;n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\mathsf{MSE}/s_x^2}} \leq t_{1-\alpha/2;n-2}\,.$$

Daher ist

$$\hat{\beta}_1 \pm t_{1-\alpha/2;n-2}\sqrt{\mathsf{MSE}/s_x^2}$$

ein $(1-\alpha)$ Konfidenzintervall für β_1.

Nicht zu verwechseln sind hierbei:

- t_{n-2}: bezeichnet den Typ der Verteilung (t) und ihren Parameter $(n-2)$.
- $t_{1-\alpha/2;n-2}$: bezeichnet das $1-\alpha/2$ Perzentil der t_{n-2} Verteilung.

Level α Tests für β_1

A Zweiseitiger Test $H_0 : \beta_1 = c$, $H_1 : \beta_1 \neq c$

B Einseitiger Test $H_0 : \beta_1 \geq c$, $H_1 : \beta_1 < c$

C Einseitiger Test $H_0 : \beta_1 \leq c$, $H_1 : \beta_1 > c$

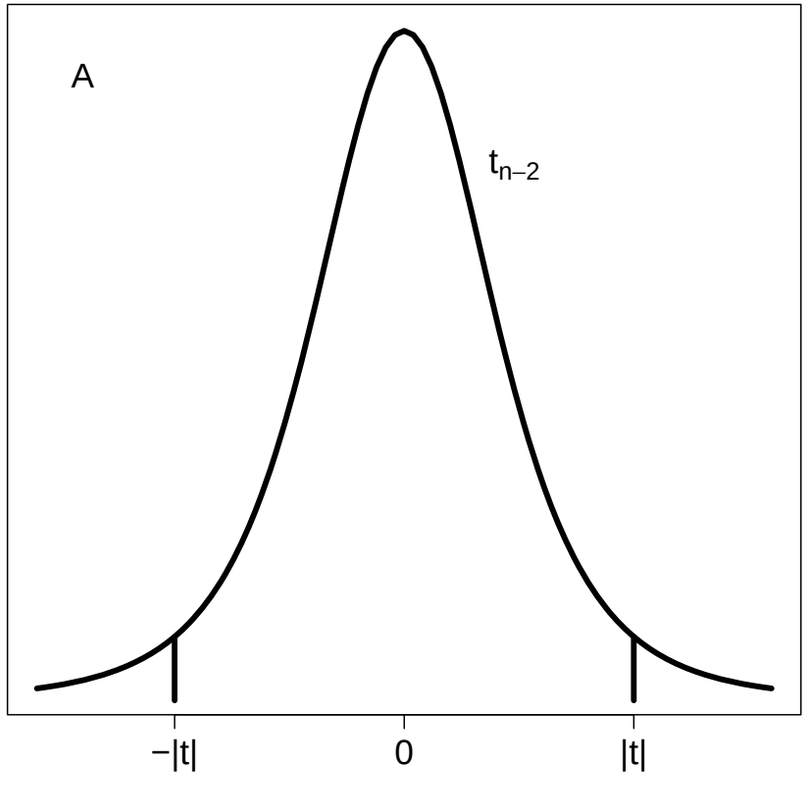
Teststatistik:

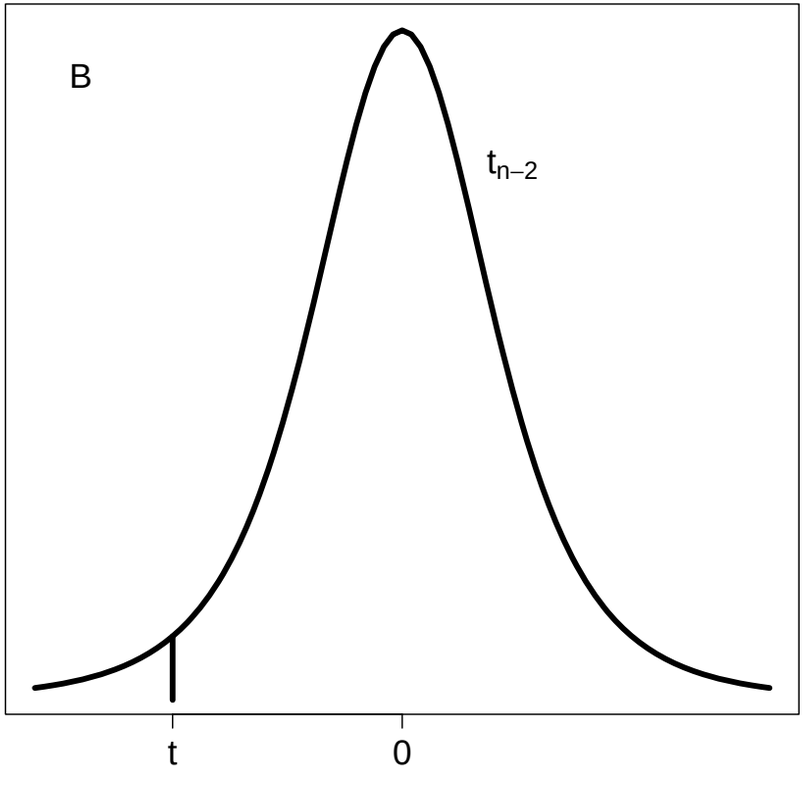
$$T = \frac{\hat{\beta}_1 - c}{\sqrt{\mathsf{MSE}/s_x^2}}$$

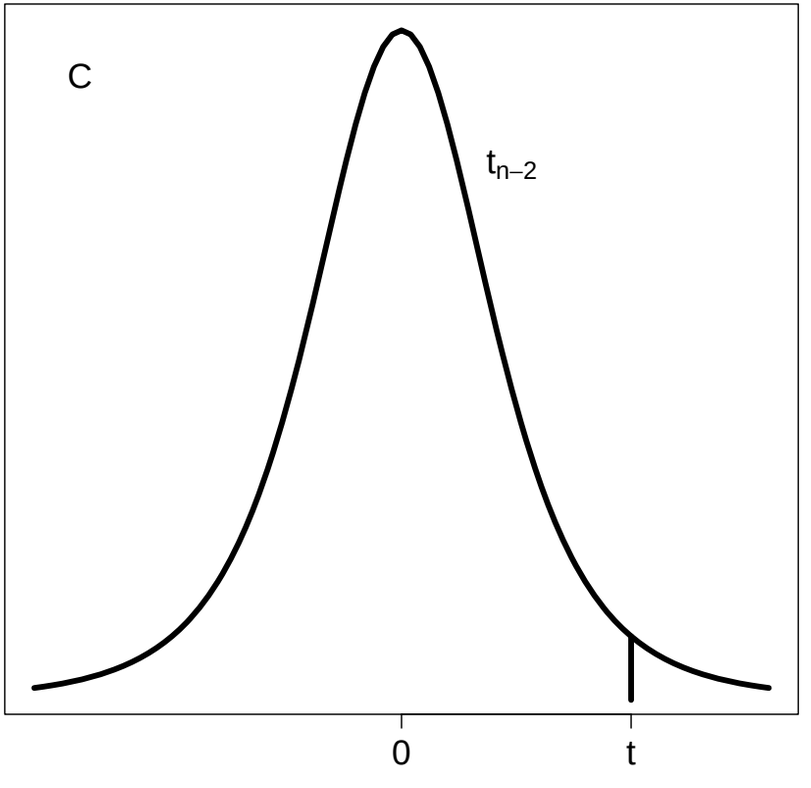
Verwerfungsregeln:

A Verwirf H_0, falls $|T| > t_{1-\alpha/2;n-2}$
B Verwirf H_0, falls $T < -t_{1-\alpha;n-2}$
C Verwirf H_0, falls $T > +t_{1-\alpha;n-2}$

p-Wert: Wahrscheinlichkeit eines $extremeren$ Wertes von T als der, den wir haben, gegeben H_0 ist wahr.




B
t_{n-2}
t
0


C
t_{n-2}
0
t

Beispiel für einen Hypothesentest

Frage: Teste $H_0 : \beta_1 = 0$ gegen $H_1 : \beta_1 \neq 0$ mit $\alpha = 0.05$ im SLR der Hauspreise. Wie groß ist der p-Wert?

$\hat{\beta}_1 = 75.6$, $s_x^2 = 25.38$, MSE $= 379.21$

Falls H_0 wahr ist, gibt es keine lineare Beziehung zwischen E(Y) und Wohnfläche.

Antwort: $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$, $\alpha = 0.05$

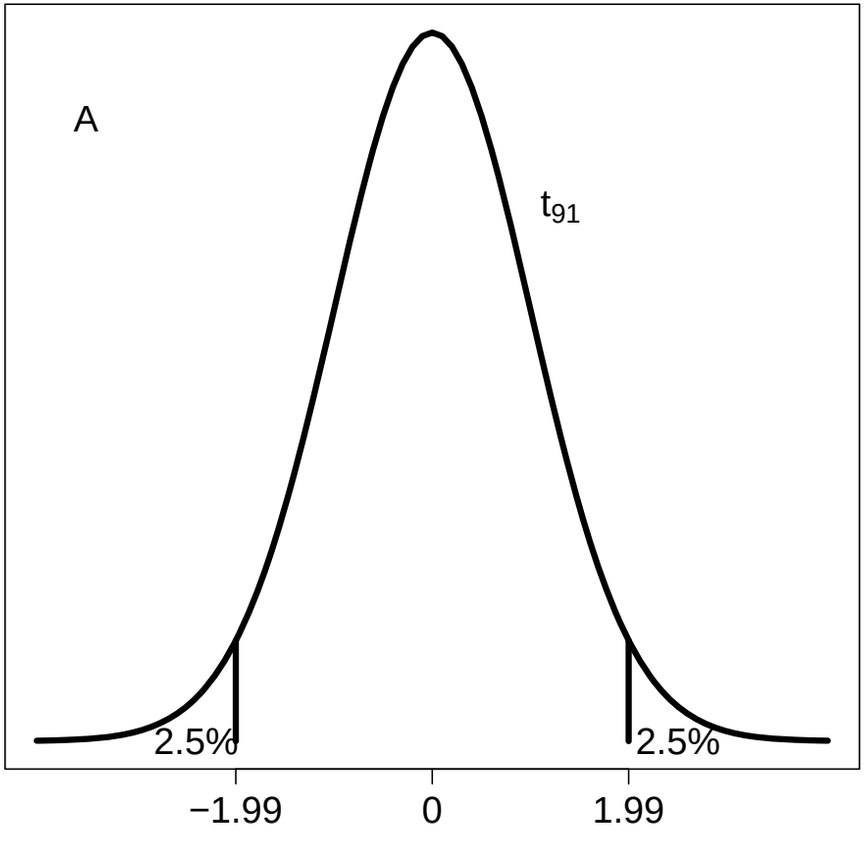
Teststatistik:

$$T = \frac{\hat{\beta}_1 - 0}{\sqrt{\text{MSE}/s_x^2}} \quad \Rightarrow \quad t = \frac{75.6}{\sqrt{379.21/25.38}} = 19.56$$

Verwerfungsregel: Verwirf H_0 falls $|t| > t_{1-\alpha/2;n-2} = t_{0.975;91} = 1.99$.

Schlussfolgerung: Verwirf H_0 da $19.56 = |t| > t_{0.975;91} = 1.99$. Somit besteht ein signifikanter linearer Zusammenhang zwischen mittlerem Preis und Fläche.

Beispiel fortgesetzt: Wie sieht das Bild dazu aus?



Wir erinnern uns an die Verwerfungsregel:

$$\begin{aligned} \Pr(\text{verwirf } H_0 | H_0 \text{ ist wahr}) &= \Pr(|T| > 1.99 | H_0 \text{ ist wahr}) \\ &= 1 - 0.95 = \alpha \,. \end{aligned}$$

Wo ist t im vorigen Bild?

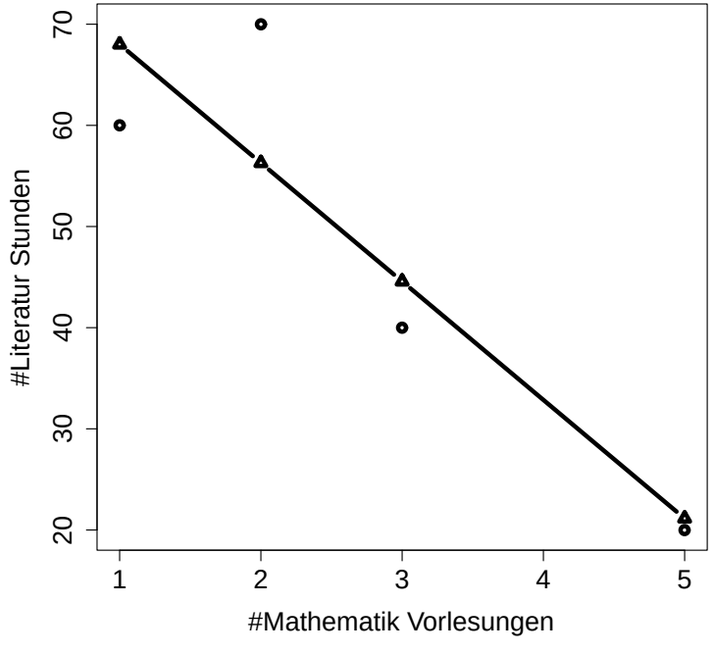
Ich würde H_0 verwerfen **für jedes** $|t| > 1.99$!

p-Wert: Wahrscheinlichkeit eines noch extremeren t (als unseres) ist fast Null.

Extrapolation ist schlecht!

Verwende niemals die geschätzte Regressionsfunktion $\widehat{\mathsf{E}}(Y) = \hat{\beta}_0 + \hat{\beta}_1 x$ außerhalb des Bereichs der x Werte der Daten!

Beispiel: Anzahl Mathematik-Vorlesungen und Anzahl Literatur-Stunden.



Mein Freund besucht 7 Mathematik-Vorlesungen im nächsten Semester. Schätze, wie viele Stunden er dann für Literatur aufbringen kann!

$80 - 11.7 \cdot 7 = -1.9 \quad \Rightarrow \quad$ Nettes Konzept, aber leider unbrauchbar!

Konfidenzintervalle für die zu erwartende Response

Sei x_h ein Wert von x für den wir $\mathrm{E}(Y_h) = \mu_h = \beta_0 + \beta_1 x_h$ schätzen wollen.

x_h kann ein Wert aus der Stichprobe sein, oder ein anderer (neuer) Wert innerhalb des betrachteten Bereichs der x.

Der Punktschätzer $\hat{\mu}_h$ von $\mathrm{E}(Y_h)$ ist

$$\hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h \,.$$

Bemerke, dass wegen $\hat{\beta}_0 = \sum_i b_i Y_i$ und $\hat{\beta}_1 = \sum_i a_i Y_i$ auch folgt, dass

$$\hat{\mu}_h = \sum_{i=1}^{n} b_i Y_i + x_h \sum_{i=1}^{n} a_i Y_i = \sum_{i=1}^{n} \Big(b_i + x_h a_i \Big) Y_i \,.$$

Somit ist auch $\hat{\mu}_h$ normalverteilt und als Erwartungswert und Varianz folgt

$$\begin{aligned} \mathsf{E}(\hat{\mu}_h) &= \beta_0 + \beta_1 x_h \\ \mathsf{var}(\hat{\mu}_h) &= \sigma^2 \left\{ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_x^2} \right\} . \end{aligned}$$

Zusammen haben wir

$$\hat{\mu}_h \sim \mathsf{Normal}\left(\beta_0 + \beta_1 x_h, \sigma^2 \left\{ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_x^2} \right\} \right)$$

oder

$$\frac{\hat{\mu}_h - (\beta_0 + \beta_1 x_h)}{\sqrt{\sigma^2 \left\{ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_x^2} \right\}}} \sim \mathsf{Normal}(0, 1) .$$

Ersetzen des unbekannten σ^2 durch den MSE liefert

$$\frac{\hat{\mu}_h - (\beta_0 + \beta_1 x_h)}{\sqrt{\mathsf{MSE}\left\{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_x^2}\right\}}} \sim t_{n-2}\,.$$

So wie schon für β_1, erhält man hierfür als $(1-\alpha)$ Konfidenzintervall für den Erwartungswert $\mu_h = \beta_0 + \beta_1 x_h$ das Intervall

$$\hat{\mu}_h \pm t_{1-\alpha/2;n-2}\sqrt{\mathsf{MSE}\left\{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_x^2}\right\}}\,.$$

Beispiel: Für das SLR der Hauspreise resultierte

$$\widehat{\mathsf{E}}(price) = \hat{\mu}(area) = -25.2 + 75.6 \cdot area$$

Weiters ist $s_x^2 = 25.38$, MSE $= 379.21$, $\bar{X} = 1.65$.

Angenommen, wir planen, einige Häuser in Gainesville mit jeweils 2000 sq.ft. zu bauen und wollen wissen, um wieviel wir diese verkaufen können.

Der Punktschätzer ist $\hat{\mu}(2) = -25.2 + 75.6 \cdot 2 = 126$

Das 95% Konfidenzintervall für $\mu(2) = \beta_0 + \beta_1 \cdot 2$ ist

$$126 \pm t_{0.975;91} \sqrt{379.21 \left\{ \frac{1}{93} + \frac{(2-1.65)^2}{25.38} \right\}} = 126 \pm 4.86 \approx (121, 131)\,.$$

Wir sind zu 95% sicher, dass der mittlere Preis dieser Häuser zwischen 121.000 und 131.000 $ liegt. (Das Konfidenzintervall für μ_h ist in $x_h = \bar{x}$ am schmälsten.)

Prädiktions-/Vorhersageintervall für $Y_{h(new)}$

Nach Erhebung der Daten wollen wir eine neue Beobachtung vorhersagen, deren x Wert x_h ist.

Zuvor schätzten wir den Erwartungswert der Verteilung von Y. Jetzt sagen wir ein spezielles Ergebnis beim Ziehen aus dieser Verteilung von Y voraus.

Beispiel: Es steht ein 2000 sq.ft. Haus zum Verkauf. Dessen Preis ist eine Zufallsvariable $Y_{h(new)}$ und $x_h = 2$.

Nehmen wir an, dass β_0 und β_1 beide bekannt sind.

Frage: Was erwarten wir für $Y_{h(new)}$?
Antwort: $Y_{h(new)} = \beta_0 + \beta_1 x_h + \epsilon_{h(new)}$

Also ist $\mathsf{E}(Y_{h(new)}) = \beta_0 + \beta_1 x_h$, $\mathsf{var}(Y_{h(new)}) = \sigma^2$ und

$$Y_{h(new)} \sim \mathsf{Normal}(\beta_0 + \beta_1 x_h, \sigma^2)\,.$$

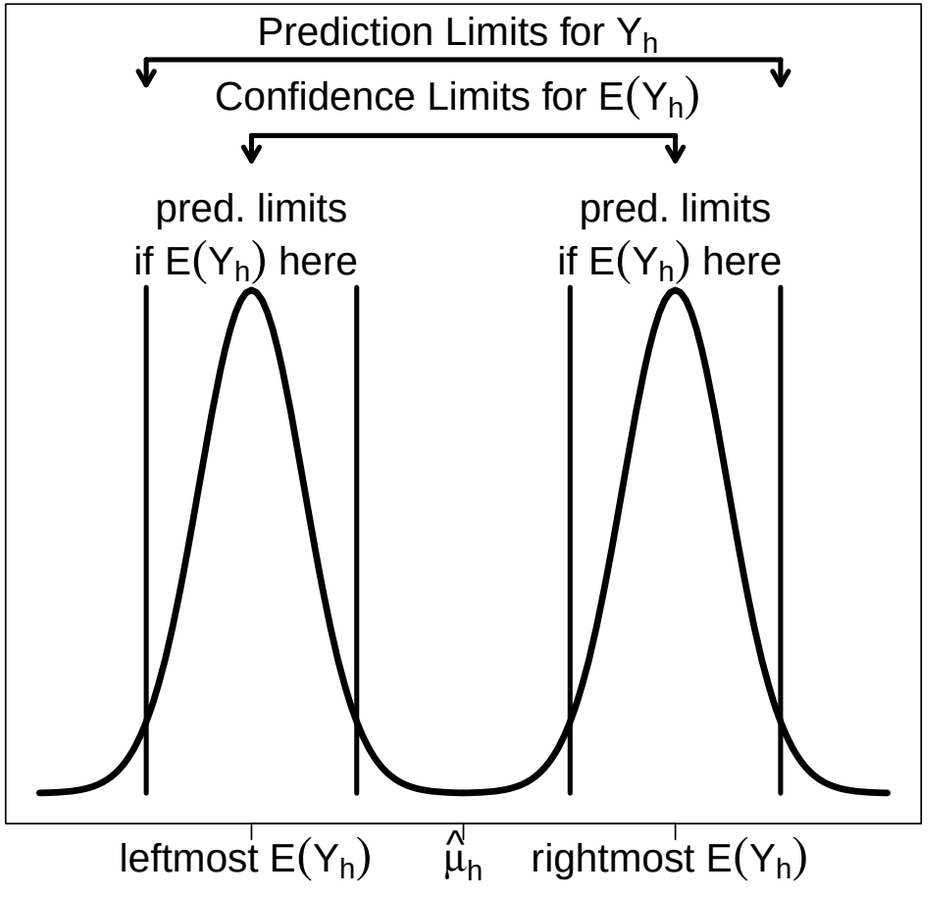
Somit resultieren als Grenzen eines $1-\alpha$ Vorhersageintervalls für $Y_{h(new)}$:

$$\mathsf{E}(Y_{h(new)}) \pm z_{1-\alpha/2} \cdot \sigma\,.$$

Aber wir kennen die Parameter nicht, haben jedoch ein $(1-\alpha)$ Konfidenzintervall für $\mu_h = \beta_0 + \beta_1 x_h$:

$$(\hat{\beta}_0 + \hat{\beta}_1 x_h) \pm t_{1-\alpha/2;n-2} \sqrt{\mathsf{MSE}\left\{\frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_x^2}\right\}}$$

Verteilungen von $Y_{h(new)}$ an der oberen und unteren Grenze dieses Intervalls.



Das $(1-\alpha)$ Prädiktionsintervall für $Y_{h(new)}$ ist etwas weiter als das $(1-\alpha)$ Konfidenzintervall für $\mu_h = \beta_0 + \beta_1 x_h$.

Betrachte die Differenz

$$Y_{h(new)} - \hat{\mu}_h = Y_{h(new)} - \sum_{i=1}^{n} (b_i + x_h a_i) Y_i \,,$$

wobei $\hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h$ unabhängig ist von $Y_{h(new)}$. Diese Differenz ist auch eine Linearkombination in den $n+1$ Responses $Y_1, \ldots, Y_n, Y_{h(new)}$, und somit auch normalverteilt mit

$$\mathsf{E}(Y_{h(new)} - \hat{\mu}_h) = \mathsf{E}(Y_{h(new)}) - \mathsf{E}(\hat{\mu}_h) = 0$$

und

$$\begin{aligned}
\mathsf{var}(Y_{h(new)} - \hat{\mu}_h) &= \mathsf{var}(Y_{h(new)}) + \mathsf{var}(\hat{\mu}_h) = \sigma^2 + \sigma^2 \left\{ \frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_x^2} \right\} \\
&= \sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_x^2} \right\} .
\end{aligned}$$

Somit folgt

$$\frac{Y_{h(new)} - \hat{\mu}_h}{\sqrt{\mathsf{var}(Y_{h(new)} - \hat{\mu}_h)}} \sim \mathsf{Normal}(0,1) \quad \Rightarrow \quad \frac{Y_{h(new)} - \hat{\mu}_h}{\sqrt{\mathsf{MSE}\left\{ 1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_x^2} \right\}}} \sim t_{n-2}$$

und ein $(1-\alpha)$ Prädiktionsintervall für $Y_{h(new)}$ ist gegeben durch:

$$\hat{\mu}_h \pm t_{1-\alpha/2;n-2} \sqrt{\mathsf{MSE}\left\{ 1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_x^2} \right\}} .$$

Beispiel: Ein 95% Prädiktionsintervall für $Y_{h(new)}$, den Preis eines 2000 sq.ft. Hauses, ist

$$126 \pm t_{0.975;91}\sqrt{379.21\left\{1+\frac{1}{93}+\frac{(2-1.65)^2}{25.38}\right\}} = 126 \pm 38.5 \approx (87.5, 164.5)\,.$$

Somit wird mit einer 95% Sicherheit der Preis dieses Hauses zwischen 87.500 und 164.500 $ liegen.

ANalysis Of VAriance: ANOVA

Nichts Neues, nur alternative Möglichkeit der Interpretation!

Wesentlich in der Regression ist die Zerlegung der **Totalen Quadratsumme**

$$\mathsf{SST} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \,.$$

Natürlich gilt

$$\sum_{i=1}^{n} (Y_i - \bar{Y}) = \sum_{i=1}^{n} (\hat{\mu}_i - \bar{Y}) + \sum_{i=1}^{n} (Y_i - \hat{\mu}_i) \,.$$

Diese Zerlegung hält aber auch im quadratischen Sinn!

Es gilt

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 &= \sum_{i=1}^{n}(\hat{\mu}_i-\bar{Y})^2+\sum_{i=1}^{n}(Y_i-\hat{\mu}_i)^2\\ \mathsf{SST} &= \mathsf{SSR}(\hat{\beta}_0,\hat{\beta}_1) + \mathsf{SSE}(\hat{\beta}_0,\hat{\beta}_1)\,.\end{aligned}$$

SST hängt nicht vom Modell ab.

Die **Regressions-Quadratsumme**

$$\mathsf{SSR}(\hat{\beta}_0,\hat{\beta}_1)=\sum_{i=1}^{n}(\hat{\mu}_i-\bar{Y})^2$$

bewertet den Unterschied zwischen dem Regressionsmodell, und einem Modell ohne x.

Je größer $\mathsf{SSR}(\hat{\beta}_0,\hat{\beta}_1)$, desto wesentlicher ist das Regressionsmodell.

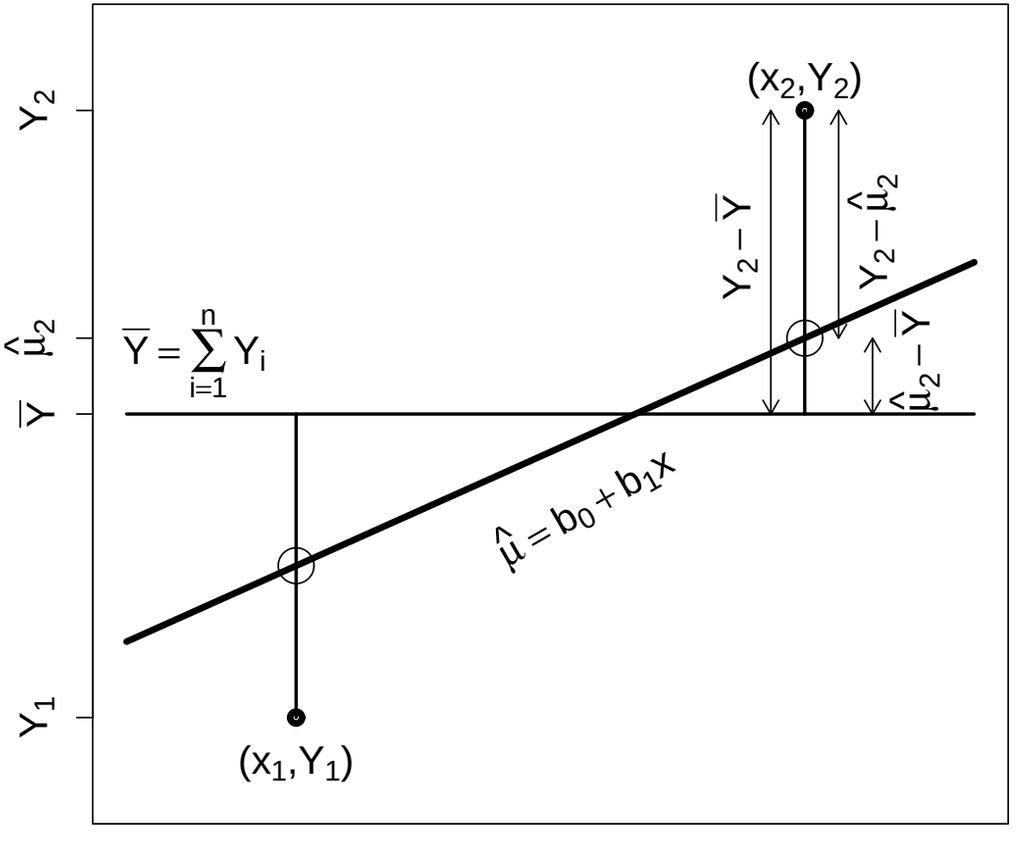
Die **Fehler-Quadratsumme**

$$\mathsf{SSE}(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_i)^2$$

ist minimal im Kleinsten Quadrate Schätzer $(\hat{\beta}_0, \hat{\beta}_1)$.

Diese Partitionierung kann auch als Variationszerlegung interpretiert werden:

$$\left(\begin{array}{c}\text{totale Variabilität}\\ \text{in } Y\end{array}\right) = \left(\begin{array}{c}\text{durch das Modell}\\ \text{erklärte Variabilität}\end{array}\right) + \left(\begin{array}{c}\text{durch das Modell } \mathbf{nicht}\\ \text{erklärte Variabilität}\end{array}\right).$$


Y_2
$\hat{\mu}_2$
$\bar{Y}$
Y_1
(x_2, Y_2)
(x_1, Y_1)
$\bar{Y} = \sum_{i=1}^{n} Y_i$
$Y_2 - \bar{Y}$
$Y_2 - \hat{\mu}_2$
$\hat{\mu}_2 - \bar{Y}$
$\hat{\mu} = b_0 + b_1 x$

Bei ANOVA Methoden zerlegt man SST in mehrere Quadratsummen mit dazugehörenden Freiheitsgraden (**degrees of freedom (df)**).

ANOVA Tafel für ein SLR:

Variations-Ursache	Quadratsumme (SS)	df	mittlere SS
Regression	$\mathsf{SSR} = \sum_i (\hat{\mu}_i - \bar{Y})^2$	1	$\mathsf{MSR} = \mathsf{SSR}/1$
Error	$\mathsf{SSE} = \sum_i (Y_i - \hat{\mu}_i)^2$	$n-2$	$\mathsf{MSE} = \mathsf{SSE}/(n-2)$
Total	$\mathsf{SST} = \sum_i (Y_i - \bar{Y})^2$	$n-1$	

Alternative, um $H_0 : \beta_1 = 0$ gegen $H_1 : \beta_1 \neq 0$ zu testen.
Teststatistik:

$$F = \frac{\mathsf{MSR}}{\mathsf{MSE}} = \frac{\mathsf{SSR}/1}{\mathsf{SSE}/(n-2)} \overset{H_0}{\sim} F_{1,n-2}\,.$$

Verwerfungsregel: verwirf H_0, falls $F > F_{1,n-2;1-\alpha}$.

Bemerke: F-Test und t-Test sind äquivalent; d.h. der F-Test verwirft genau dann, wenn der t-Test verwirft.

Mit $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$ folgt

$$\mathsf{SSR} = \sum_{i=1}^{n} (\hat{\mu}_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{Y})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = \hat{\beta}_1^2 s_x^2 \,.$$

Also gilt

$$F = \frac{\hat{\beta}_1^2 s_x^2}{\mathsf{MSE}} = \frac{\hat{\beta}_1^2}{\mathsf{MSE}/s_x^2} = \left(\frac{\hat{\beta}_1}{\sqrt{\mathsf{MSE}/s_x^2}} \right)^2 = T^2$$

Wir wissen, dass $T \sim t_{n-2}$ äquivalent ist mit $T^2 \sim F_{1,n-2}$.

Bestimmtheitsmaß R^2

Frage: Wie stark ist die **lineare** Beziehung zwischen Y und x?

SSE$(\hat{\beta}_0, \hat{\beta}_1)$ sollte im Vergleich zu SSR$(\hat{\beta}_0, \hat{\beta}_1)$ möglichst klein sein.

SSE$(\hat{\beta}_0, \hat{\beta}_1)$ ist unter dem Modell bereits minimal.

SST hängt nur von den Y_i ab (darin ist keine Information über das SLR enthalten).

$$\underbrace{\mathsf{SST}}_{(fest)} = \underbrace{\mathsf{SSR}(\hat{\beta}_0, \hat{\beta}_1)}_{(maximal)} + \underbrace{\mathsf{SSE}(\hat{\beta}_0, \hat{\beta}_1)}_{(minimal)} .$$

Zur Beurteilung der Güte der Anpassung verwendet man das Bestimmtheitsmaß

$$R^2 = \frac{\mathsf{SSR}(\hat{\beta}_0, \hat{\beta}_1)}{\mathsf{SST}} = 1 - \frac{\mathsf{SSE}(\hat{\beta}_0, \hat{\beta}_1)}{\mathsf{SST}}, \qquad 0 \leq R^2 \leq 1 .$$

R^2 gibt den relativen Variationsanteil an, der durch das SLR erklärt wird.

Für das SLR gilt:

$$\begin{aligned} R^2 &= \frac{\sum_i (\hat{\mu}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} = \frac{\sum_i (\hat{\beta}_1 (x_i - \bar{x}))^2}{\sum_i (Y_i - \bar{Y})^2} = \hat{\beta}_1^2 \frac{s_x^2}{s_Y^2} = \frac{s_{xY}^4}{s_x^4} \frac{s_x^2}{s_Y^2} \\ &= \left(\frac{s_{xY}^2}{s_x s_Y} \right)^2 = \widehat{\mathsf{cor}}^2(x, Y) \,. \end{aligned}$$

Also entspricht R^2 dem Quadrat des empirischen Korrelationskoeffizienten zwischen Y_i und x_i. Je größer R^2, desto stärker die lineare Beziehung!

Aber: $R^2 \approx 0$ heißt nicht immer, dass es **überhaupt keine** Beziehung zwischen Y und x gibt! Es heißt nur, dass die Beziehung **nicht linear** ist.

Extreme Fälle:

- $\hat{\mu}_i = Y_i$: dann ist SSE $= 0 \Rightarrow R^2 = 1$
- $\hat{\beta}_1 = 0 \Rightarrow \hat{\mu}_i = \bar{Y}$: dann ist SSR $= 0 \Rightarrow R^2 = 0$.

3. Diagnostische Aspekte

Bis jetzt betrachteten wir Daten (x_i, Y_i) und **nahmen an**, dass

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

mit

- $\epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$,
- β_0, β_1 und σ^2 sind unbekannte Parameter,
- x_i's sind feste Konstanten.

Frage:
Was sind mögliche **Fehler oder Verletzungen** dieser Annahmen?

1. Regressionsfunktion ist nicht linear ($\mathrm{E}(Y) \neq \beta_0 + \beta_1 x$)

2. Error Terme haben keine konstante Varianz ($\mathrm{var}(\epsilon_i) \neq \sigma^2$, $i = 1, \ldots, n$)

3. Error Terme sind nicht unabhängig

4. Modell passt überall bis auf wenige Ausnahmen

5. Error Terme sind nicht normalverteilt

6. SLR ist unglaubwürdig (Modell sollte mehrere Prädiktoren beinhalten)

Verwende **Residuen-Plots**, um Probleme zu diagnostizieren.

Residuen: $r_i = Y_i - \hat{\mu}_i$ mit

empirischem Mittel $\bar{r} = \frac{1}{n}\sum_i r_i = 0$ und

empirischer Varianz $\frac{1}{n-1}\sum_i (r_i - \bar{r})^2 = \frac{1}{n-1}\sum_i r_i^2 \approx$ MSE.

Semi-studentisierte Residuen (*fast* standardisierte Residuen):

$$r_i^* = \frac{r_i - \bar{r}}{\sqrt{\text{MSE}}} = \frac{r_i}{\sqrt{\text{MSE}}}$$

Bemerkung: Wäre der MSE ein Schätzer der Varianz des Residuums r_i, dann würden wir r_i^* studentisiertes (oder standardisiertes) Residuum nennen. Die Standardabweichung eines Residuums ist jedoch viel komplizierter und variiert für verschiedene Residuen. Der MSE ist nur eine einfache Approximation. Daher nennen wir r_i^* ein semi-studentisiertes Residuum.

Nichtlinearität der Regressionsfunktion (1.)

Betrachte Residuen-Plot gegen den **Prädiktor** x,

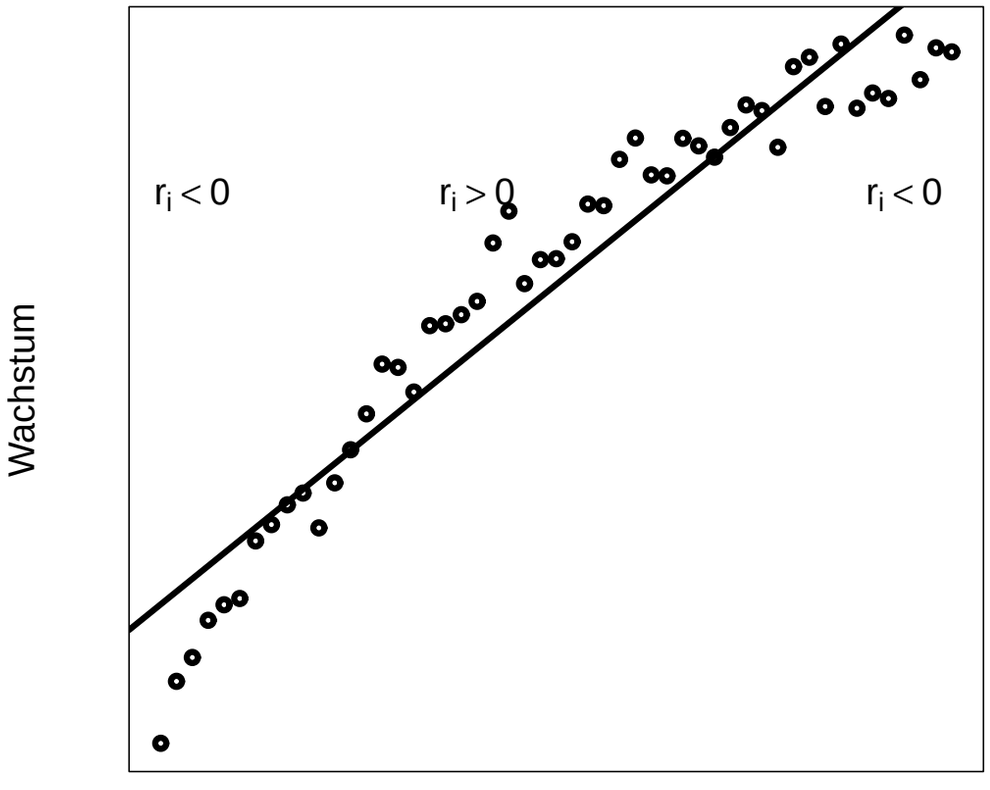
oder den Residuen-Plot gegen die **geschätzten Erwartungswerte** $\hat{\mu}$.

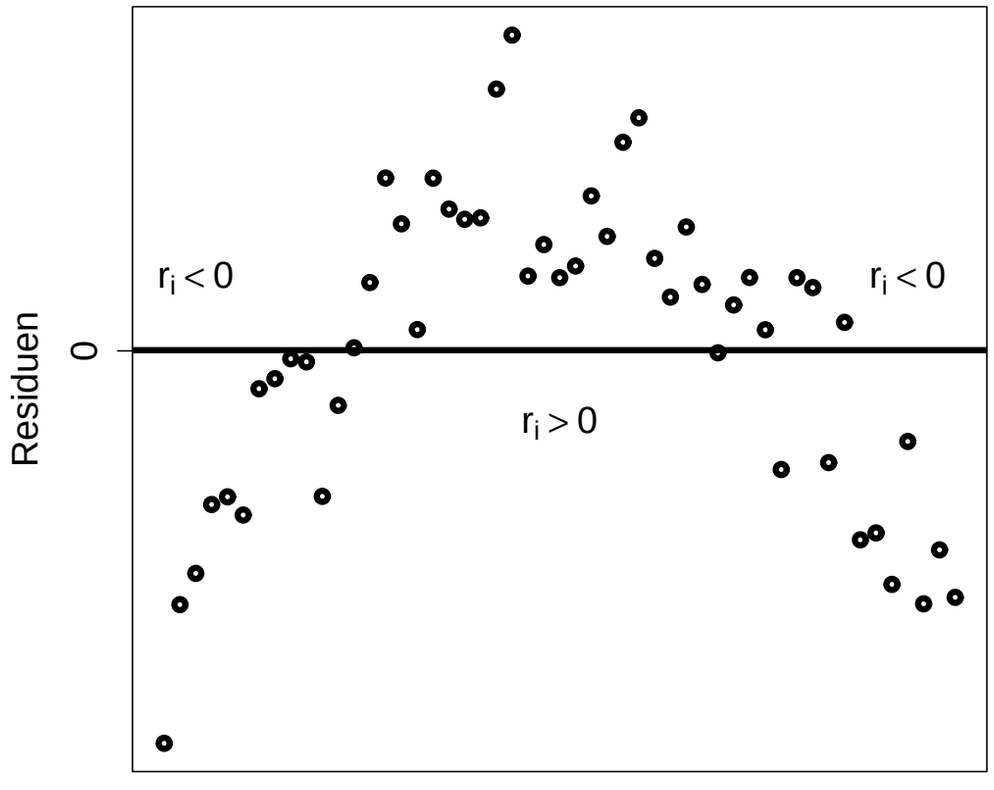
Schau auf systematische Abweichungen!

Beispiel: Blumengießen

$x_i =$ Menge an Wasser/Woche

$Y_i =$ Pflanzenwachstum in den ersten beiden Monaten.


Wachstum
$r_i < 0$
$r_i > 0$
$r_i < 0$
Wasser/Woche


Residuen
0
$r_i < 0$
$r_i > 0$
$r_i < 0$
Wasser/Woche

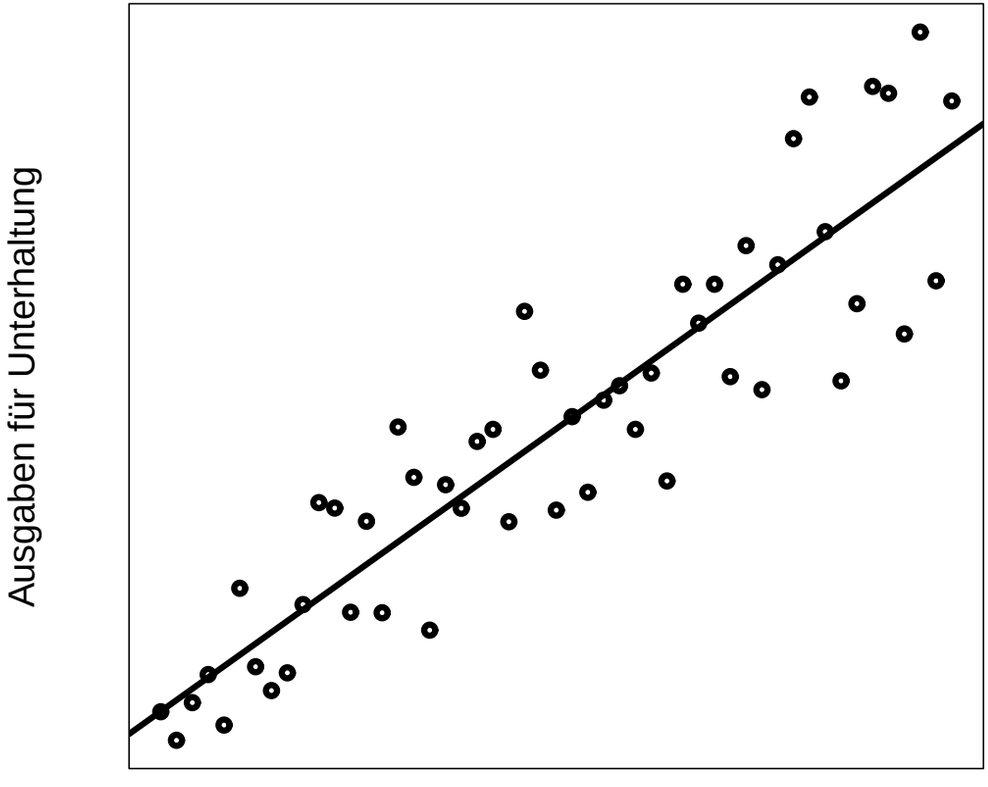
Nichtkonstanz der Error Varianz (2.)

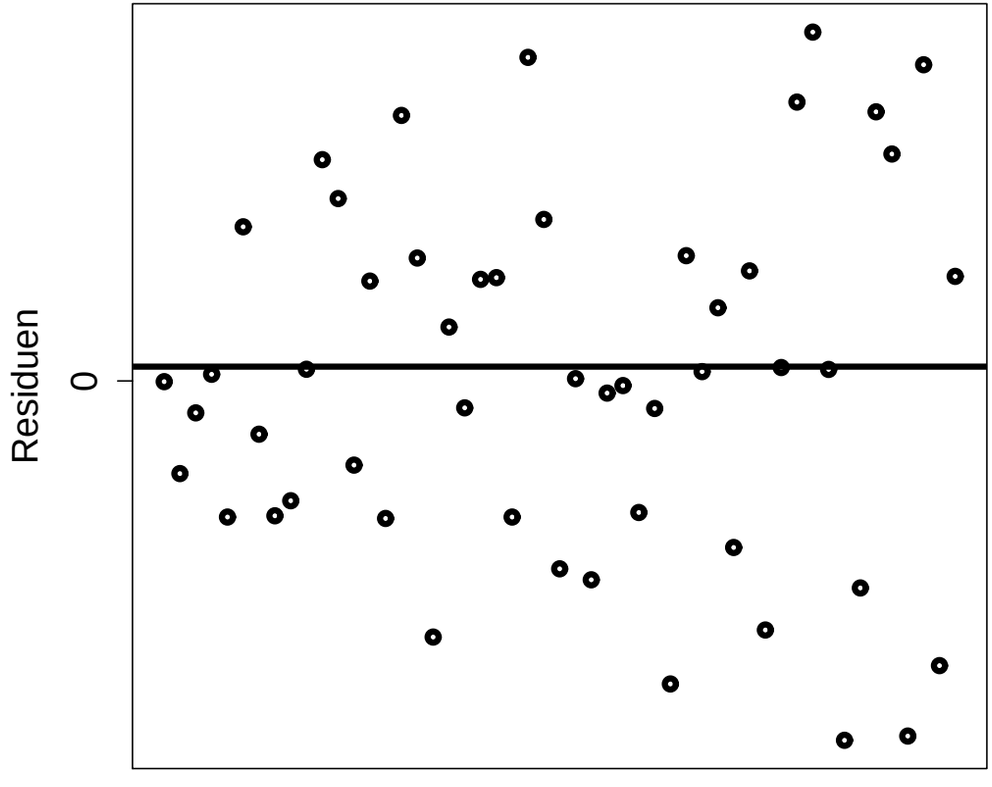
Nichtkonstante Error Varianz diagnostiziert man mittels Residuen-Plot gegen x, in dem Struktur entdeckt wird.

Beispiel: Einkommen

$x_i =$ Einkommen

$Y_i =$ Ausgaben für Unterhaltung


Ausgaben für Unterhaltung
Einkommen


Residuen
0
Einkommen

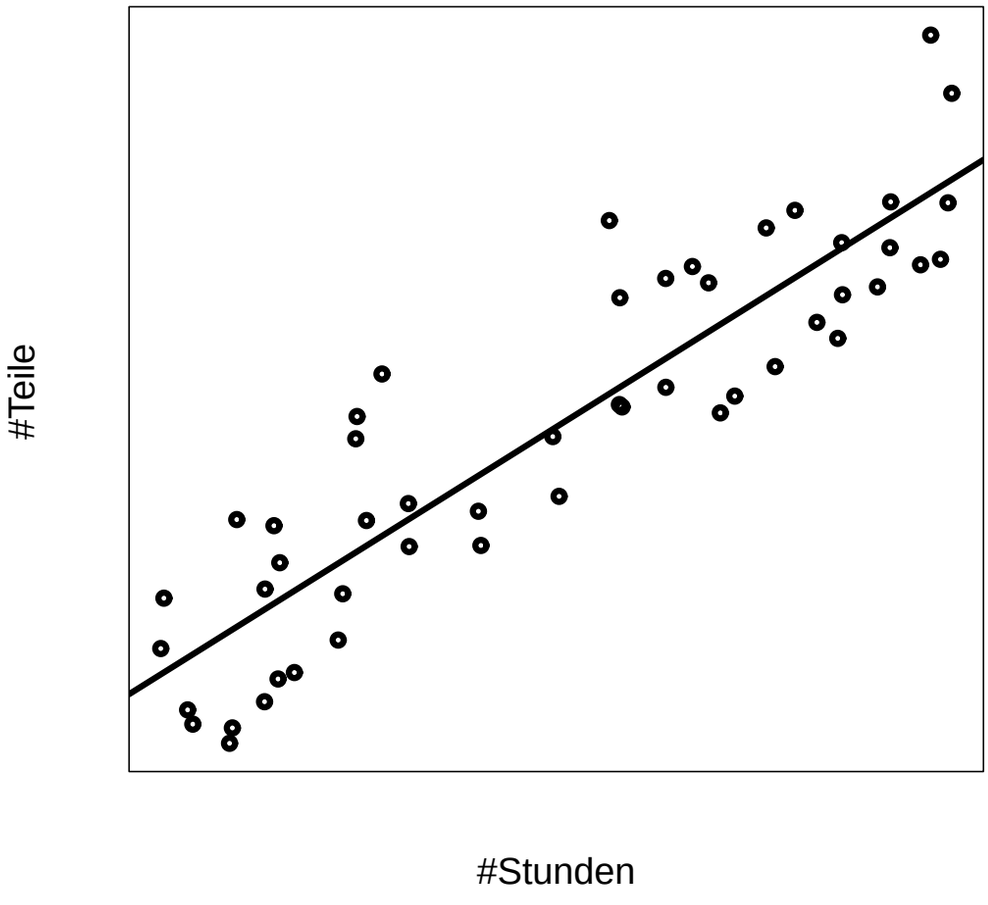
Nichtunabhängigkeit der Error Terme (3.)

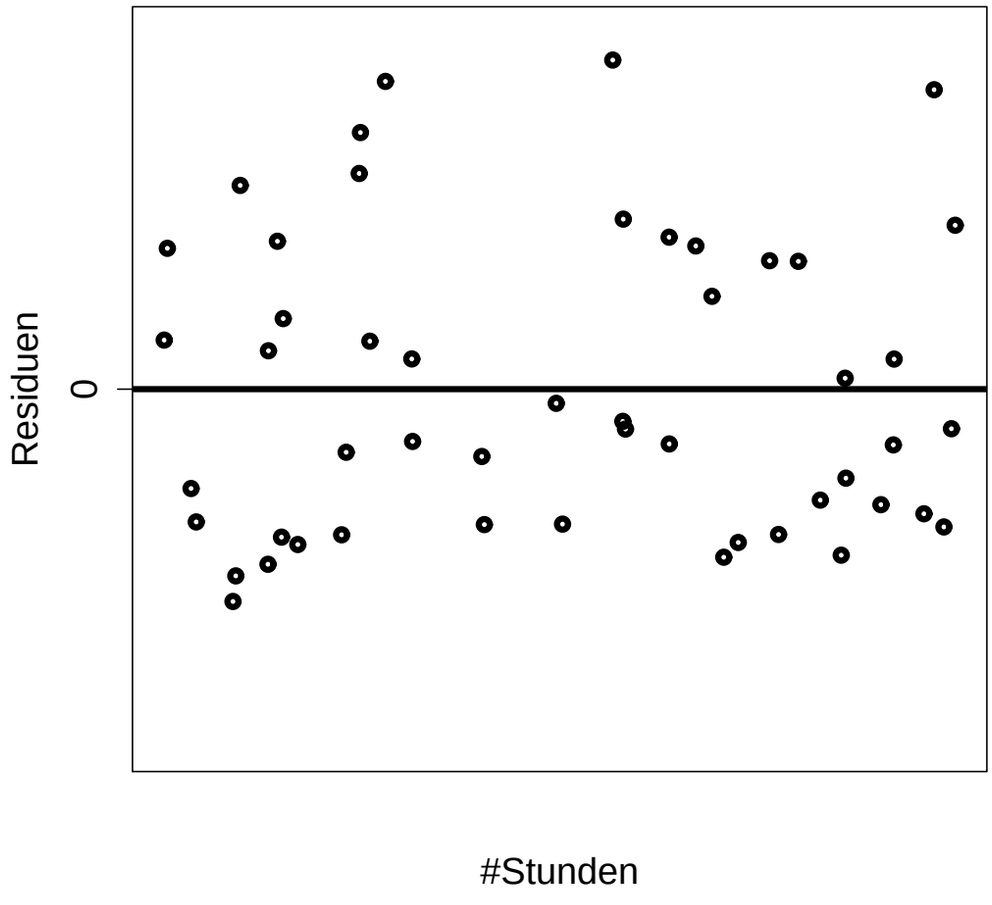
Wir erkennen Nichtunabhängigkeit der Error Terme **über die Zeit** oder **in einer Sequenz** durch einen Residuen-Plot gegen die Zeit (oder die Sequenz), und suchen nach Struktur.

Beispiel: Arbeitsleistung

$x_i = \#$Arbeitsstunden

$Y_i = \#$gefertigte Teile

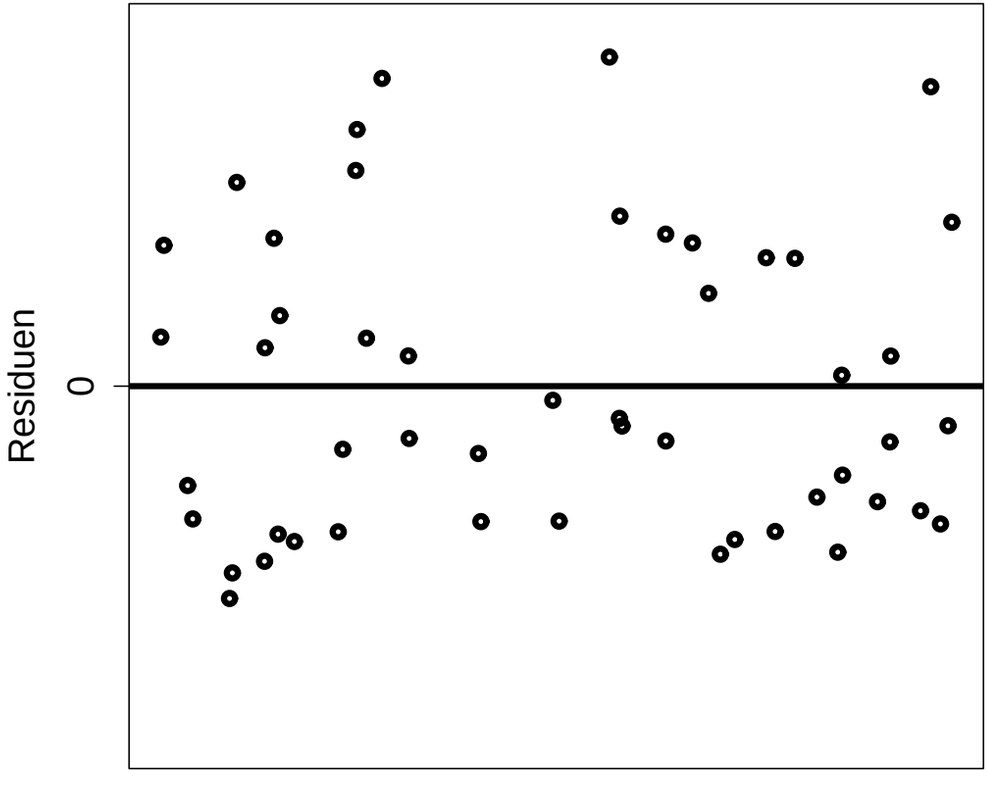

#Teile
#Stunden

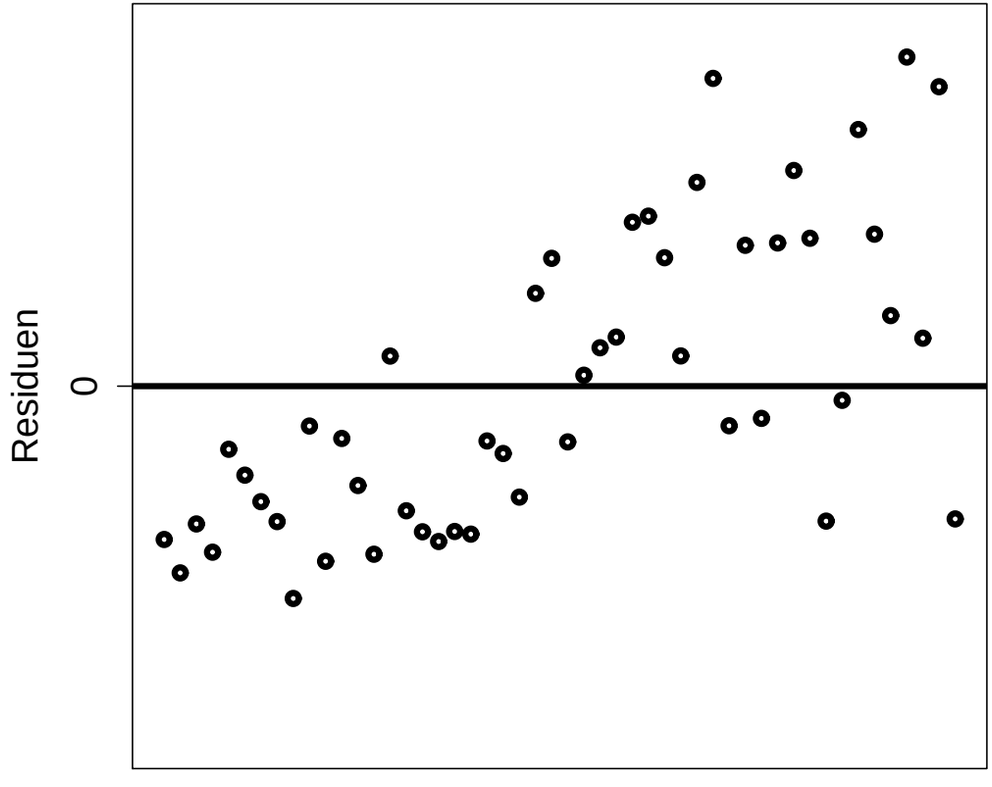

Residuen
0
#Stunden

Aber, liegen Daten vor wie

Tag 1: (x_1, Y_1)
Tag 2: (x_2, Y_2)
$\vdots$
Tag n: (x_n, Y_n)

dann können wir einen **Lerneffekt** beobachten.


Residuen
0
#Stunden


Residuen
0
Tag

Modell passt fast überall (4.)

Beispiel: LS Schätzer mit 2 Ausreißern (durchgehend) und ohne diese (strichliert).

Faustregel: Falls $|r_i^*| > 3$, dann prüfe diesen Datenpunkt (versichere dich, dass es sich um keinen falschen Wert handelt)!

Wirf keine Punkte hinaus, weil diese bzgl. des angenommenen SLR Ausreißer sind!

Ausreißer erkennt man durch einen Plot der r_i^* gegen x_i.

y
x

$r_i/\sqrt{MSE}$
x

Error Terme sind nicht normalverteilt (5.)

Wir haben angenommen, dass $\epsilon_1, \ldots, \epsilon_n$ iid Normal$(0, \sigma^2)$, aber wir können diese Terme nicht beobachten.

Diese Annahme wäre glaubwürdig, falls $r_1, \ldots, r_n$ aus Normal$(0, \text{MSE})$ zu kommen scheinen.

Tatsache: Falls $r_1/\sqrt{\text{MSE}}, \ldots, r_n/\sqrt{\text{MSE}}$ aus Normal$(0, 1)$, dann können wir zeigen, dass für den geschätzten Erwartungswert des i-t kleinsten Terms, $r_{(1)} \leq \cdots \leq r_{(n)}$, gilt:

$$\widehat{\mathsf{E}}\left(\frac{r_{(i)}}{\sqrt{\text{MSE}}}\right) = \Phi^{-1}\left(\frac{i}{n+1}\right) .$$

Beweis: Ordnungsstatistiken von Zufallsstichproben aus stetigen Population

Sei $X_1, \ldots, X_n$ Zufallsstichprobe mit stetiger Verteilungsfunktion $F(x)$. Ordnen der Stichprobe liefert $X_{(1)} < \cdots < X_{(n)}$. $X_{(i)}$ nennt man $i$te Ordnungsstatistik. Da $F(x)$ stetig, ist die Wahrscheinlichkeit einer **Bindung** Null, d.h. $X_{(i)} < X_{(i+1)}$.

Betrachte die Zufallsvariable $Y = \sum_{i=1}^{n} I(X_i \leq x)$, mit der Indikatorfunktion $I(\cdot)$. Nun ist $\Pr(X_i \leq x) = F(x)$ und wegen der Unabhängigkeit all dieser Ereignisse folgt $Y \sim$ Binomial$(n, F(x))$. Nun besteht die Beziehung: $(X_{(k)} \leq x) = (Y \geq k)$. Deshalb gilt

$$F_{X_{(k)}}(x) = \Pr(X_{(k)} \leq x) = \Pr(Y \geq k) = \sum_{i=k}^{n} \binom{n}{i} F(x)^i (1 - F(x))^{n-i}$$

mit stetiger Dichtefunktion

$$f_{X_{(k)}}(x) = k \binom{n}{k} F(x)^{k-1} (1 - F(x))^{n-k} f(x)\,.$$

Herleitung der Dichtefunktion: Differenzieren liefert unmittelbar

$$\begin{aligned} f_{X_{(k)}}(x) &= \frac{\partial}{\partial x} F_{X_{(k)}}(x) \\ &= F'(x) \sum_{i=k}^{n} \binom{n}{i} \left[i F(x)^{i-1} (1-F(x))^{n-i} - (n-i) F(x)^{i} (1-F(x))^{n-i-1} \right] . \end{aligned}$$

Nun gilt $F'(x) = f(x)$ sowie

$$i \binom{n}{i} = n \binom{n-1}{i-1} \qquad \text{und} \qquad (n-i) \binom{n}{i} = n \binom{n-1}{i} .$$

Damit ergibt sich die Teleskopsumme

$$f_{X_{(k)}}(x) = n f(x) \sum_{i=k}^{n} \left[\binom{n-1}{i-1} F(x)^{i-1} (1-F(x))^{n-i} - \binom{n-1}{i} F(x)^{i} (1-F(x))^{n-(i+1)} \right] .$$

Die Summanden heben sich bis auf den ersten Term für $i = k$ auf und es folgt

$$\begin{aligned} f_{X_{(k)}}(x) &= nf(x)\binom{n-1}{k-1}F(x)^{k-1}(1-F(x))^{n-k} \\ &= kf(x)\binom{n}{k}F(x)^{k-1}(1-F(x))^{n-k}\,. \end{aligned}$$

Bemerkung: Dies lässt sich auch als Multinomial-Wahrscheinlichkeit schreiben:

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!1!(n-k)!}F(x)^{k-1}f(x)(1-F(x))^{n-k}\,.$$

Der erste Term beschreibt die Wahrscheinlichkeit, dass $k-1$ Stichprobenelemente unter x liegen, der Term in der Mitte zeigt an, dass 1 Element gerade um x liegt, und der dritte Term deutet an, dass $n-k$ Terme über x liegen.

Ordungsstatistik aus der Gleichverteilung:

Angenommen, die Zufallsstichprobe $X_1, \ldots, X_n$ stammt aus der stetigen Uniform$(0,1)$ Verteilung. Dafür ist $F(x) = x$ (sowie $f(x) = 1$), für $0 < x < 1$, und als Dichte der $k$ten Ordnungsstatistik ergibt sich

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}\,.$$

Dies entspricht einer Beta-Dichtefunktion

$$f_Y(y) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}\,, \qquad 0 < y < 1\,.$$

Wegen $\Gamma(n+1) = n!$ sind hierfür $a = k$ und $b = n - k + 1$ und es resultiert

$$\mathsf{E}(X_{(k)}) = \mathsf{E}(Y) = \frac{a}{a+b} = \frac{k}{n+1}\,.$$

Ordungsstatistik aus beliebiger Verteilung:

Sei nun $X_1, \ldots, X_n$ eine Zufallsstichprobe aus einer beliebigen stetigen Verteilung $F(x)$ mit Ordnungsstatistiken

$$X_{(1)} < X_{(2)} < \cdots < X_{(n)} \,.$$

Betrachte nun $W_i = F(X_{(i)})$, $i = 1, \ldots, n$. Da $F(\cdot)$ eine monoton wachsende Funktion ist, gilt dafür auch

$$W_1 < W_2 < \cdots < W_n \,.$$

Wir wissen, dass $F(X_i) \overset{iid}{\sim}$ Uniform$(0, 1)$. Nun sind die Ordnungsstatistiken von der Zufallsstichprobe $F(X_1), \ldots, F(X_n)$ gerade die $W_1, \ldots, W_n$, da es egal ist, ob zuerst die X_i sortiert werden und darauf $F(\cdot)$ angewandt wird, oder zuerst $F(X_i)$ gebildet wird und diese dann sortiert werden. Somit gilt $W_i \sim$ Beta mit $\mathsf{E}(W_i) = \mathsf{E}(F(X_{(i)})) = i/(n+1)$.

Mittels Momenten-Methode ergibt sich

$$\widehat{\mathsf{E}}(W_i) = \widehat{\mathsf{E}}(F(X_{(i)})) = W_i = F(X_{(i)})\,,$$

da dies der Erwartungswert einer Zufallsstichprobe mit Umfang 1 darstellt.

Interpretation: Bemerke, dass $F(X_{(i)})$ die zufällige Fläche unter der Dichte $f(x)$ links von $X_{(i)}$ bezeichnet. Die Größe $F(X_{(i)})$ beschreibt somit eine zufällige Fläche und $\mathsf{E}(F(X_{(i)}))$ ist die erwartete Fläche links von $X_{(i)}$.

Sei der Stichprobenumfang n ungerade, also $n = 2m+1$. Dann ist der empirische Median gerade $X_{(m+1)}$. Bemerke, dass

$$\mathsf{E}(W_{m+1}) = \frac{m+1}{n+1} = \frac{m+1}{2m+1+1} = \frac{1}{2}\,.$$

Nimmt man $X_{(m+1)}$ als Punktschätzer für den Populationsmedian, dann erwartet man von $X_{(m+1)}$, dass im Mittel 50% der Population unter diesem Wert liegen.

Nun ist die erwartete Fläche unter der Dichte zwischen $X_{(i)}$ und $X_{(i-1)}$ gerade

$$\mathsf{E}(W_i - W_{i-1}) = \mathsf{E}(F(X_{(i)}) - F(X_{(i-1)})) = \frac{i}{n+1} - \frac{i-1}{n+1} = \frac{1}{n+1}\,.$$

Außerdem ist die erwartete Fläche unter der Dichte über dem Maximum $X_{(n)}$

$$\mathsf{E}(1 - F(X_{(n)})) = 1 - \frac{n}{n+1} = \frac{1}{n+1}\,.$$

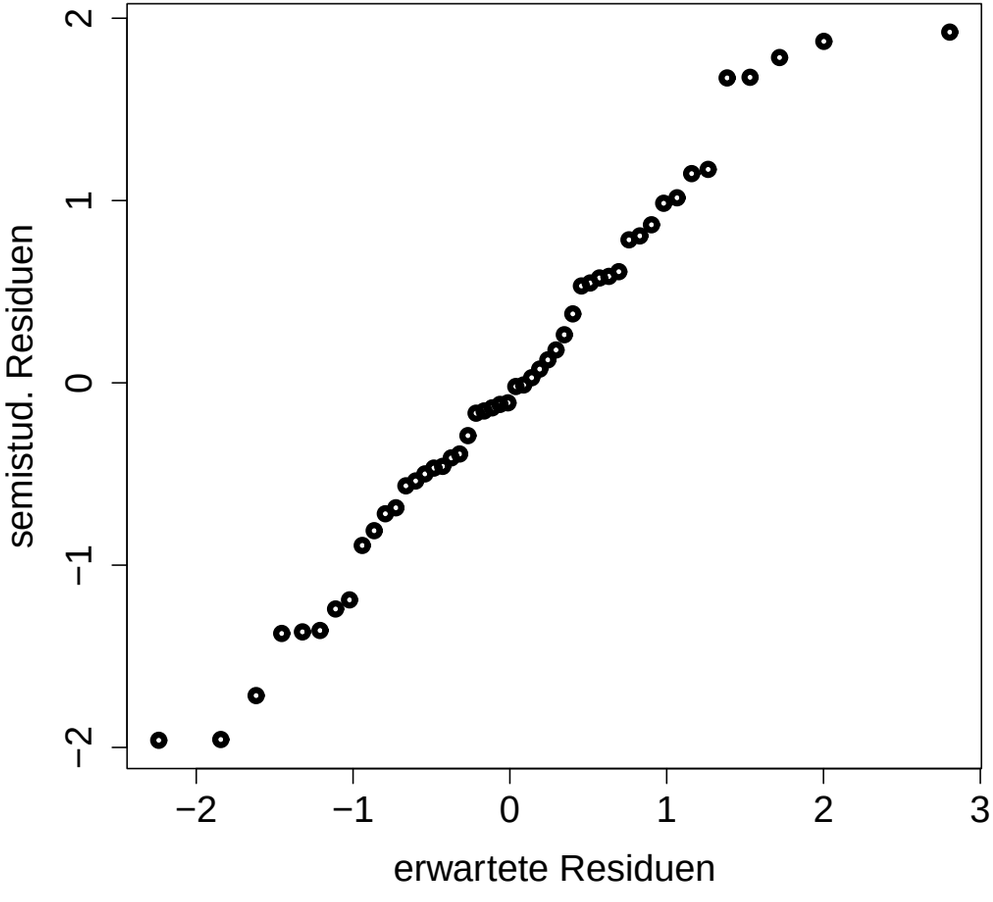
Die Ordnungsstatistiken $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ teilen die Fläche unter der Dichtekurve $f(x)$ in $n+1$ Teile, welche alle im Mittel gleich $1/(n+1)$ sind.

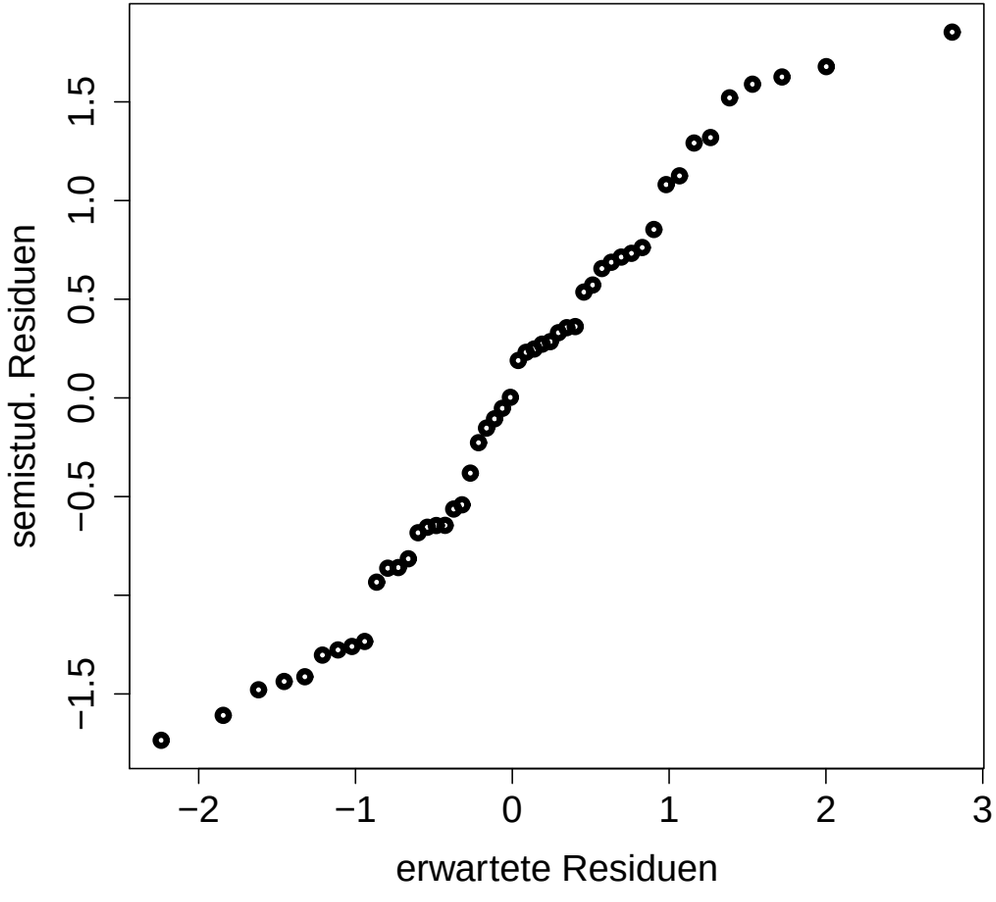
Es macht Sinn, Ordnungsstatistiken als Schätzer für Perzentile zu verwenden. So verwenden wir $X_{(i)}$ als $100{\cdot}p$-tes Perzentil der Stichprobe mit $p = i/(n+1)$. $X_{(i)}$ ist ein Schätzer für das Populationsperzentil x_p, wobei die Fläche unter der Dichte $f(x)$ links von x_p genau p ist. Falls aber $(n+1)p$ kein Integer ist, dann interpolieren wir zwischen den beiden Ordnungsstatistiken.

Beim Normal QQ-Plot betrachtet man den Punktverlauf der geordneten Residuen gegen (korrigierte) theoretische Quantile

$$\left(\Phi^{-1} \left(\frac{i - 3/8}{n + 1/4} \right), \frac{r_{(i)}}{\sqrt{\mathsf{MSE}}} \right) , \qquad i = 1, \ldots, n \, .$$

Wir plotten die beobachteten geordneten Residuen r_i^* gegen deren erwartete Werte (**Normal Probability Plot**), um damit Abweichungen von der Normalverteilung zu entdecken.


semistud. Residuen
erwartete Residuen
−2
−1
0
1
2
3


semistud. Residuen
erwartete Residuen
−1.5
−0.5
0.0
0.5
1.0
1.5
−2
−1
0
1
2
3

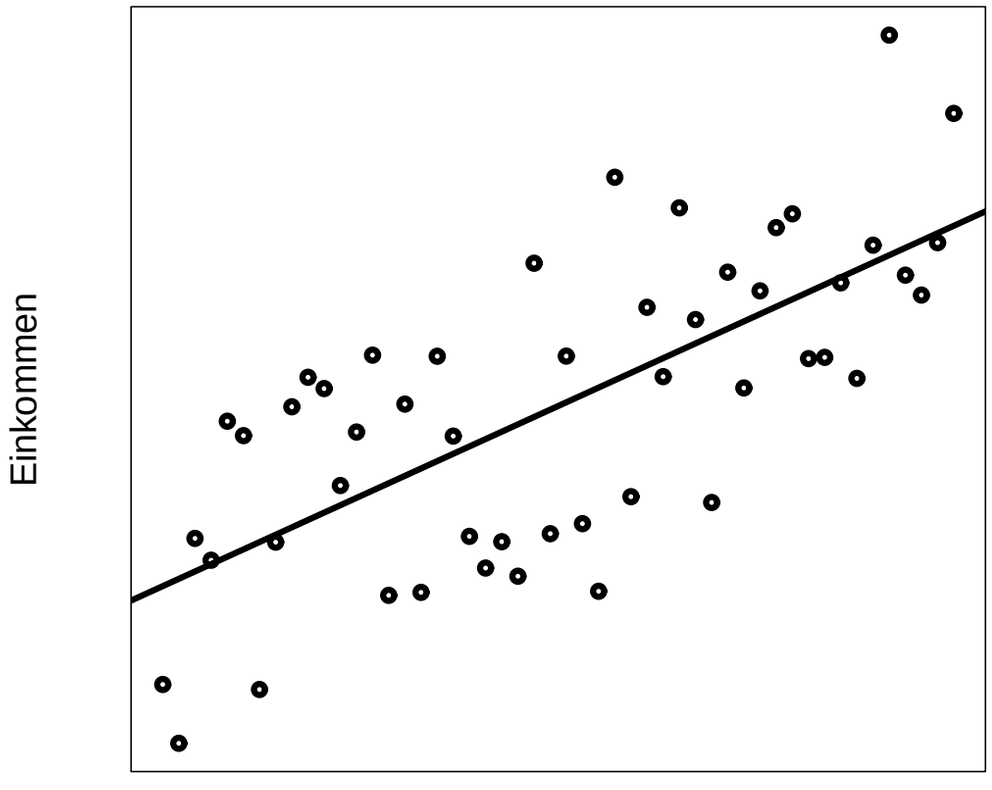
Nichtberücksichtigung wichtiger Prädiktoren (6.)

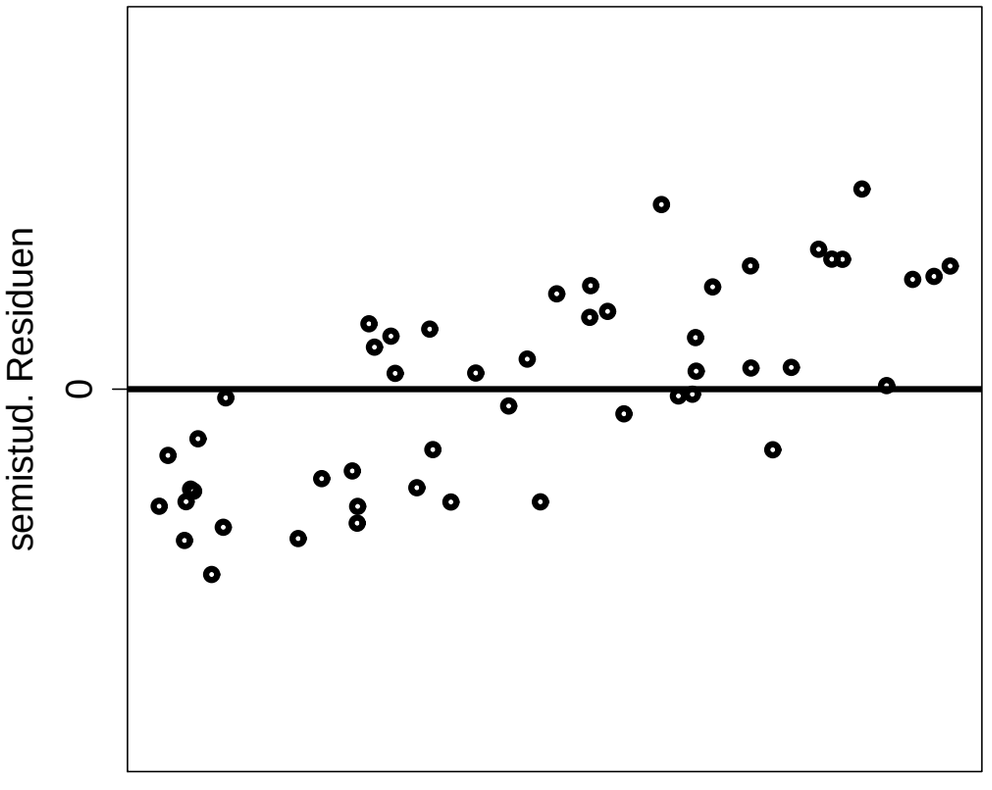
Beispiel: Einkommen

$x_i = \#$Ausbildungsjahre

$Y_i =$ Einkommen

Angenommen wir haben zusätzlich noch: $z_i = \#$Jahre in diesem Job


Einkommen
#Ausbildungsjahre


semistud. Residuen
0
#Jahre in diesem Job

Dies ist ein Indikator dafür, dass ein besseres Modell daher folgende Form hat:

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i$$

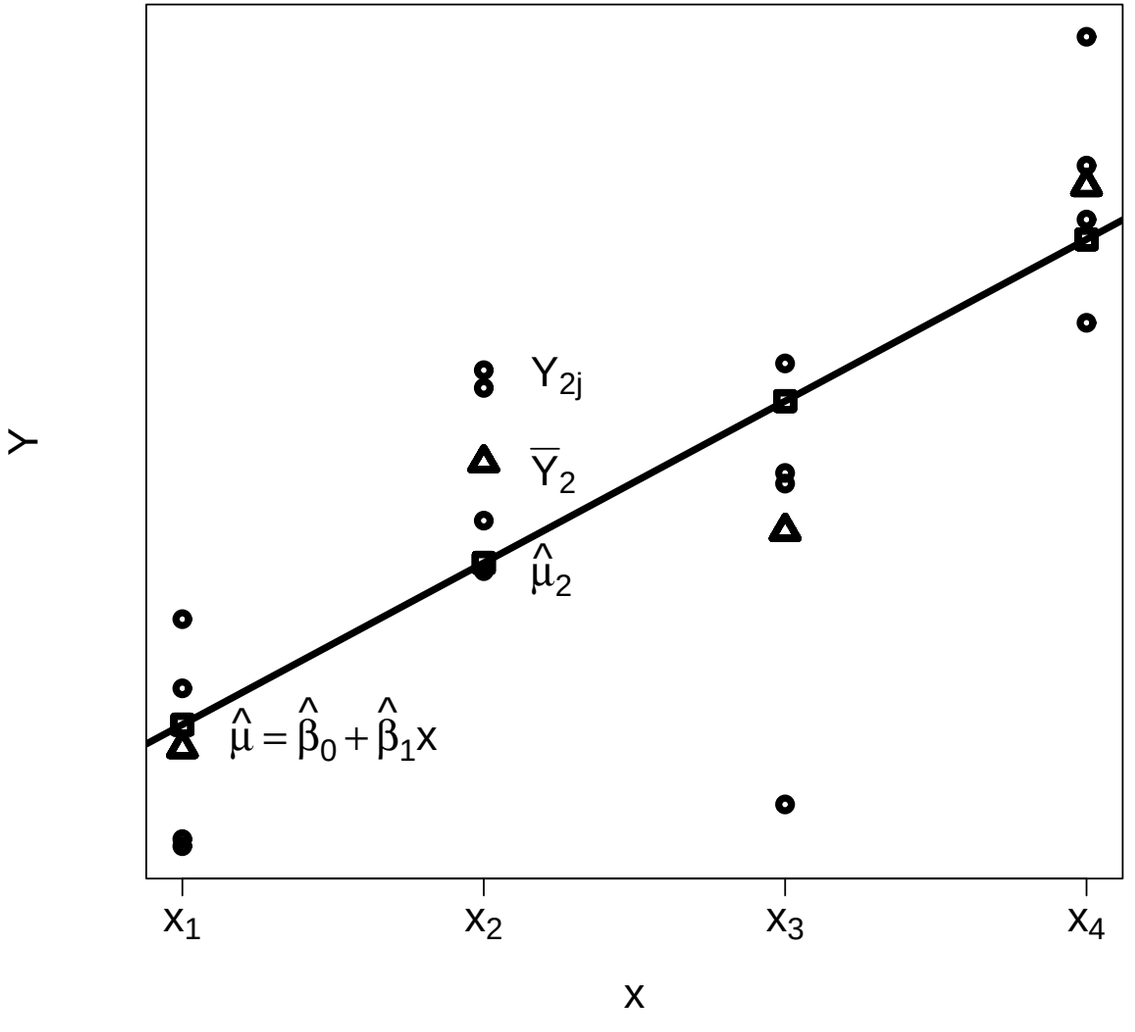
(**Multiples Regressionsmodel**)

Lack of Fit Test

Formaler Test für: $H_0 : \mathsf{E}(Y) = \beta_0 + \beta_1 x$ gegen H_1 : nicht H_0

Wir können den folgenden Test nur dann verwenden, wenn wir mehrere Responses in zumindest einem x beobachtet haben.

Motivation: SLR nimmt an, dass die Erwartungswerte alle auf einer Geraden sind! Wie besser wäre ein Modell **ohne** diese Restriktion?


Y
Y_{2j}
$\overline{Y}_2$
$\hat{\mu}_2$
$\hat{\mu} = \hat{\beta}_0 + \hat{\beta}_1 x$
x_1
x_2
x_3
x_4
x

Ein Modell mit weniger starken Restriktionen kommt **gänzlich ohne Struktur** für die Erwartungswerte in jedem x aus.

Neue Notation: Y's sind an c unterschiedlichen Stellen von x beobachtet. Wir bezeichnen diese mit $x_1, x_2, \ldots, x_c$.

Gerade n_i dieser Y's, $Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$, sind an der Stelle x_i beobachtet worden, $i = 1, 2, \ldots, c$, $n_i \geq 1$.

Sei $\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ das Mittel aller Y's in x_i and sei dort $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ der geschätzte Erwartungswert unter dem SLR.

Neue Datenstruktur:

$$\begin{array}{lllll}
\text{in } x_1: & (Y_{11}, x_1), (Y_{12}, x_1), & \ldots & , (Y_{1n_1}, x_1) & \Rightarrow \bar{Y}_1 \\
\text{in } x_2: & (Y_{21}, x_2), (Y_{22}, x_2), & \ldots & , (Y_{2n_2}, x_2) & \Rightarrow \bar{Y}_2 \\
 & & \vdots & & \\
\text{in } x_c: & (Y_{c1}, x_c), (Y_{c2}, x_c), & \ldots & , (Y_{cn_c}, x_c) & \Rightarrow \bar{Y}_c
\end{array}$$

Bemerke, dass

$$Y_{ij} - \hat{\mu}_i = (Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \hat{\mu}_i)\,.$$

Das Residuum zerfällt in einen Teil, der den reinen Fehler (**pure error**) beschreibt, und einen anderen Teil, der auf schlechte Modellanpassung (**lack of fit**) zurückzuführen ist.

Partitioniere daher den SSE in die folgenden 2 Teile:

$$\mathsf{SSE}(\hat{\beta}_0, \hat{\beta}_1) = \mathsf{SSPE} + \mathsf{SSLF}(\hat{\beta}_0, \hat{\beta}_1)\,,$$

wobei

$$\sum_{i=1}^{c}\sum_{j=1}^{n_i}(Y_{ij} - \hat{\mu}_i)^2 = \sum_{i=1}^{c}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{c}\sum_{j=1}^{n_i}(\bar{Y}_i - \hat{\mu}_i)^2\,.$$

• SSPE $\approx$ SSE bedeutet, die Mittel ($\triangle$) sind nahe den geschätzten Erwartungswerten ($\square$). Also können wir sogar mit dem weniger restriktiven Modell den Anteil nicht-erklärter Variabilität nicht reduzieren.

• SSLF $\approx$ SSE bedeutet, die Mittel ($\triangle$) sind von den geschätzten Erwartungswerten ($\square$) weit weg und die (lineare) Restriktion scheint damit unglaubwürdig.

Somit haben wir

$$\mathsf{SST} = \mathsf{SSE} + \mathsf{SSR} = \mathsf{SSLF} + \mathsf{SSPE} + \mathsf{SSR}\,.$$

Formaler Test für: $H_0 : \mathsf{E}(Y) = \beta_0 + \beta_1 x$ gegen $H_1 : \mathsf{E}(Y) \neq \beta_0 + \beta_1 x$

Definiere

$$\mathsf{MSLF} := \frac{\mathsf{SSLF}}{c-2} \quad \text{und} \quad \mathsf{MSPE} := \frac{\mathsf{SSPE}}{n-c}\,.$$

Teststatistik: $F = \dfrac{\mathsf{MSLF}}{\mathsf{MSPE}}$

Verwerfungsregel: Verwirf H_0, falls $F > F_{c-2,n-c;1-\alpha}$.

Dies passt gut in das Konzept der **ANOVA Tabelle**:

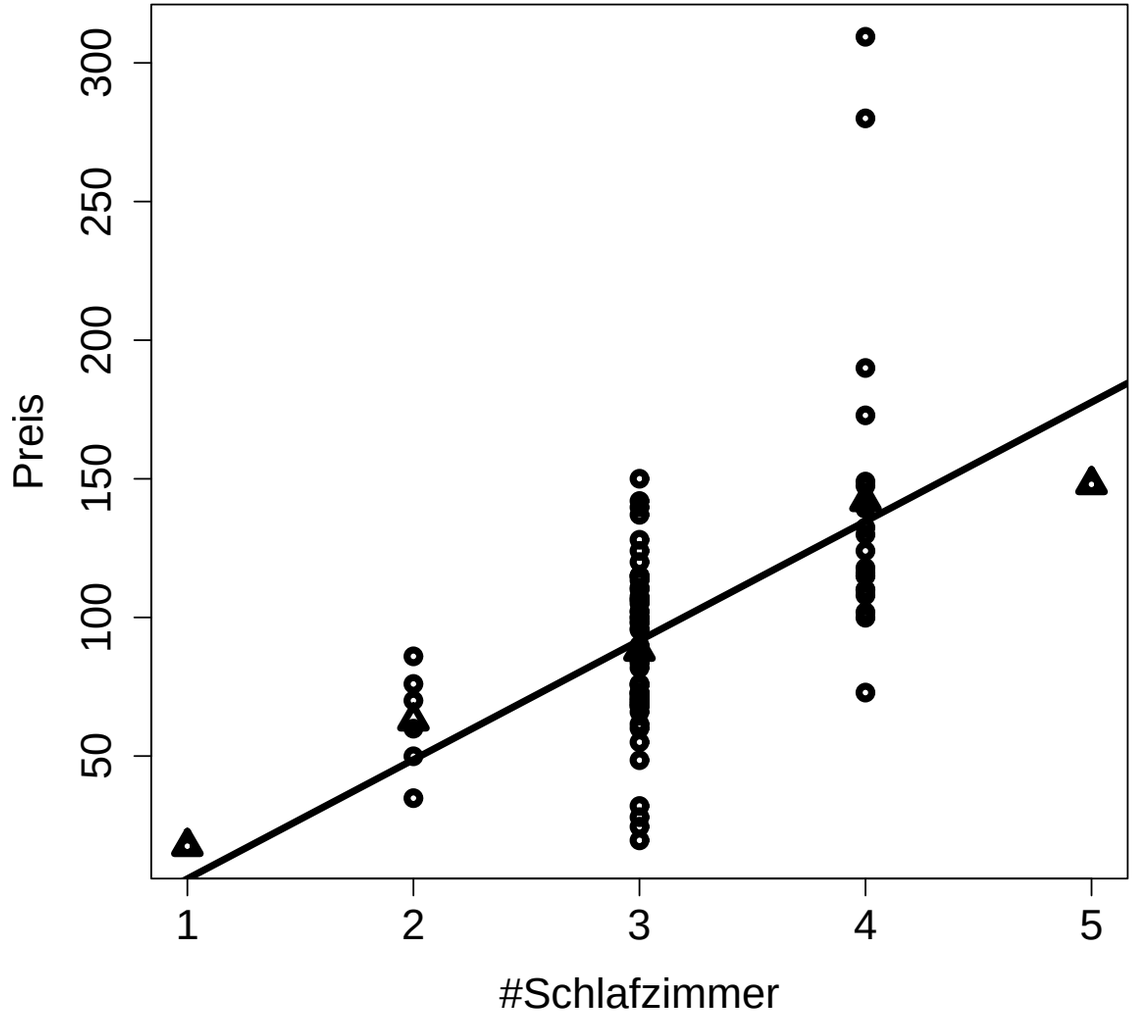
Ursache der Variation	SS	df	MSS
Regression	SSR	1	MSR
Error	SSE	$n-2$	MSE
Lack of Fit	SSLF	$c-2$	MSLF
Pure Error	SSPE	$n-c$	MSPE
Total	SST	$n-1$	

Beispiel: Angenommen, die Hauspreise folgen einem SLR mit erklärender Variablen x =#Schlafzimmer. Als geschätzte Regressionsfunktion erhalten wir

$$\widehat{\mathsf{E}}(\text{Preis}/1000) = -37.2 + 43.0 \cdot \#\text{Schlafzimmer}\,.$$

Variation	SS	df	MS
Regression	62578	1	62578
Error	117028	91	1286
Lack of Fit	4295	3	1432
Pure Error	112733	88	1281
Total	179606	92	

Wegen MSLF/MSPE $= 1432/1281 = 1.12 < F_{3,88;0.95} = 2.71$ können wir H_0 **nicht verwerfen**.


Preis
50
100
150
200
250
300
1
2
3
4
5
#Schlafzimmer

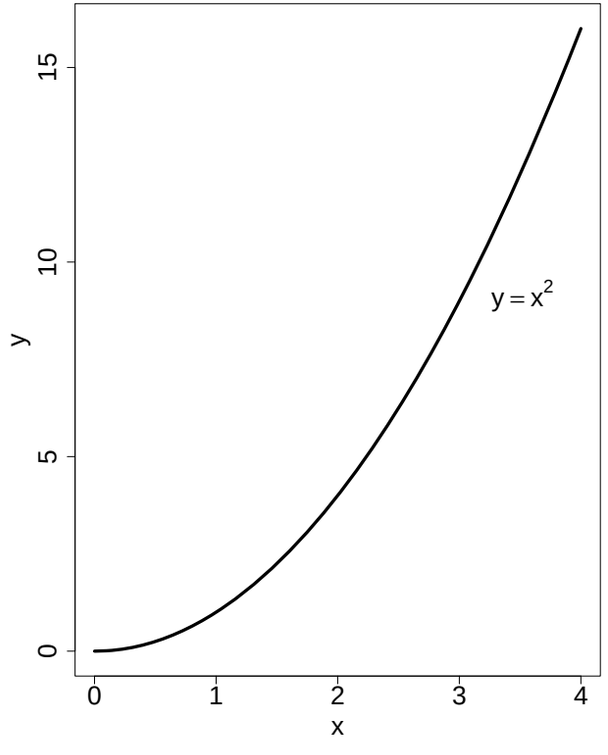
Lösungsvorschläge zu den Problemen

Viele der zweckmäßigen Maßnahmen beruhen auf Material, das erst später behandelt werden wird.

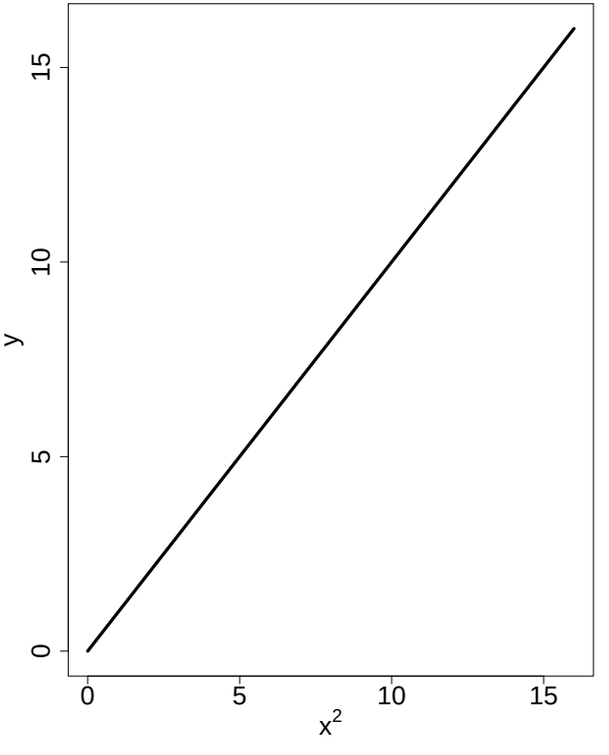
Transformationen sind eine Möglichkeit, um das Problem 1. (Nichtlineare Regressionsfunktion) und eine Kombination aus den Problemen 1. und 2. (nichtkonstante Fehlervarianzen) zu lösen.

Motivation: Betrachte die Funktion $y = x^2$

x	y
0	0
1	1
2	4
3	9
4	16



x^2	y
0	0
1	1
4	4
9	9
16	16



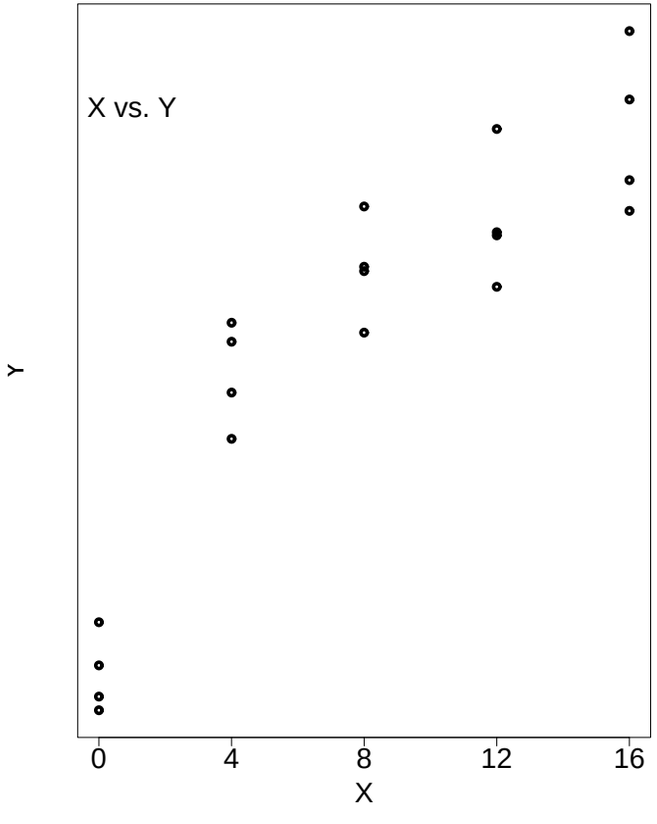
Hat man $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ und weiß man $y = f(x)$, dann liegen $(f(x_1), y_1), (f(x_2), y_2), \ldots, (f(x_n), y_n)$ auf einer **Geraden**.

Zwei Situationen in denen Transformationen hilfreich sein können:

Situation 1: Nichtlineare Regressionsfunktion mit konstanter Fehlervarianz (1.)

Bemerke, dass hier $\mathrm{E}(Y)$ nicht linear in x zu sein scheint, da die Punkte nicht um eine Gerade zu liegen scheinen.

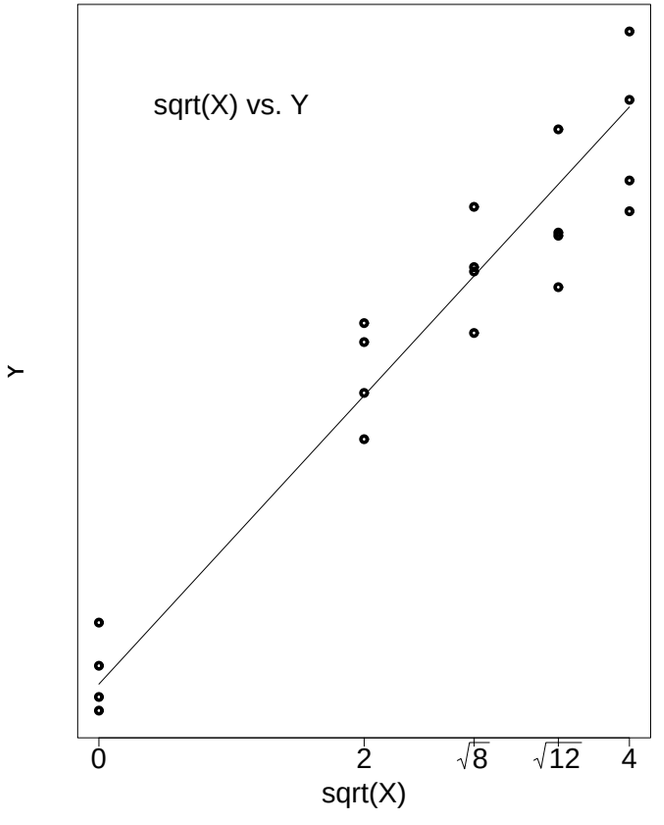
Die Streuungen der Y's auf jeder Stufe von x scheinen jedoch ziemlich vergleichbar (konstant) zu sein.



Maßnahme – Transformiere nur x

Betrachte $\sqrt{x}$

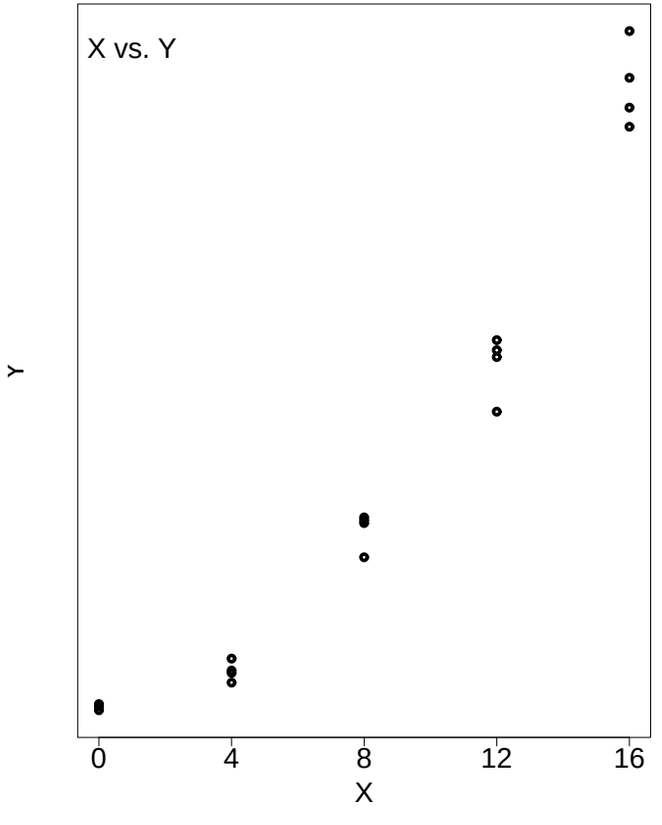
Transformieren nicht Y, weil dies die Konstantheit der Streuungen der Y's auf allen Stufen von x verändern würde.



Situation 2: Nichtlineare Regressionsfunktion mit nichtkonstanten Fehlervarianzen (1. mit 2.)

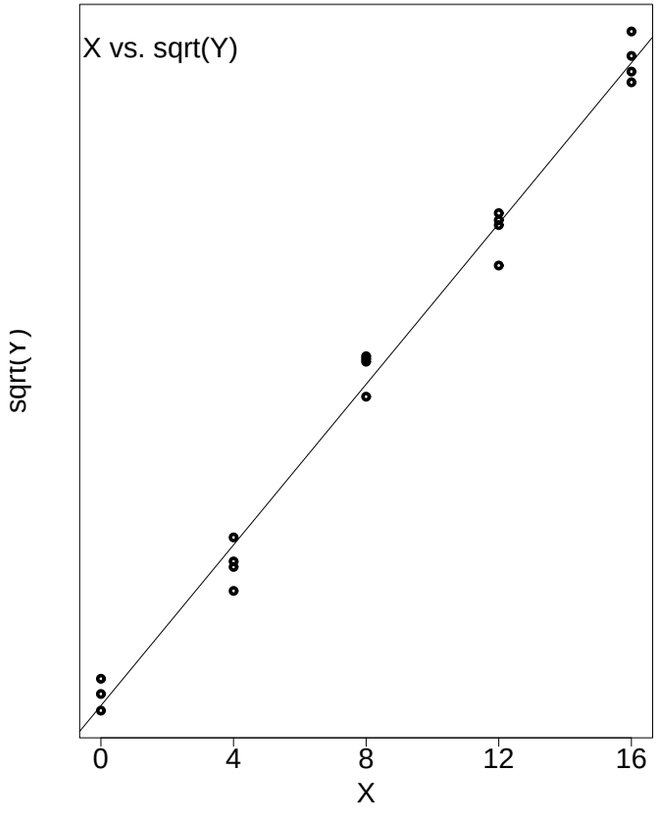
Bemerke, dass hier $\mathrm{E}(Y)$ eine nichtlineare Funktion in x ist.

Die Varianzen der Y's auf allen Stufen von x scheint weiters mit x zuzunehmen.

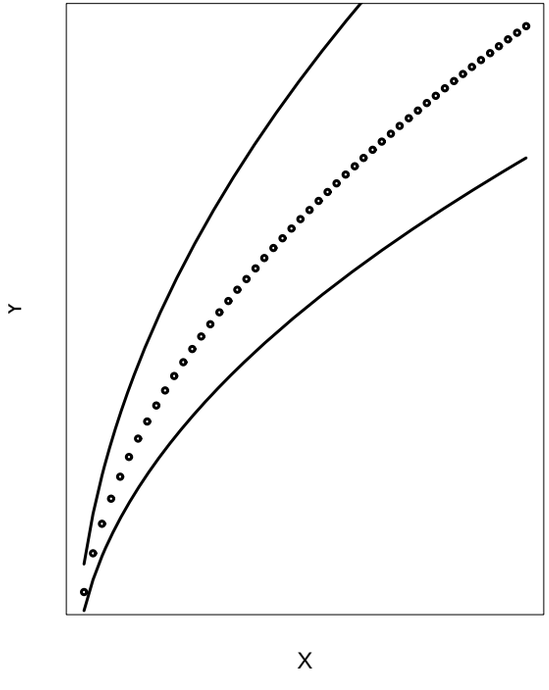


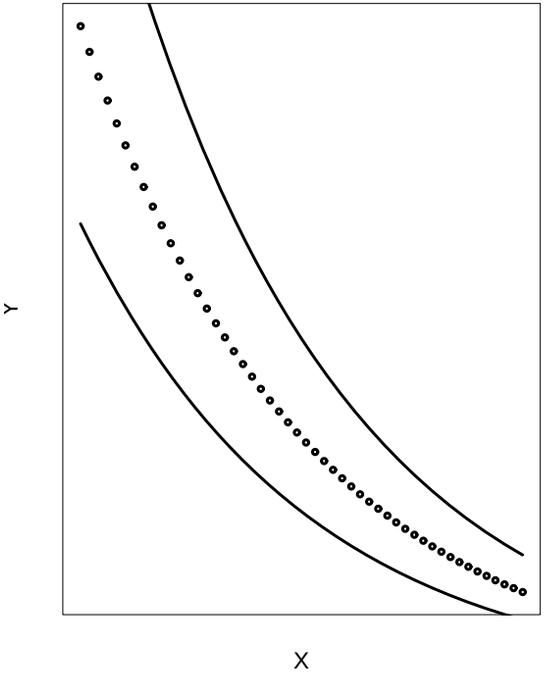
Maßnahme – Transformiere Y
(oder vielleicht besser x **und** Y)

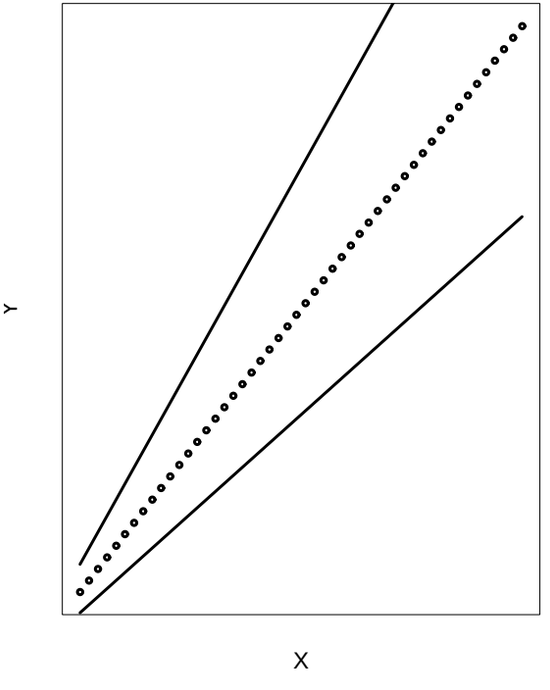
Wir betrachten $\sqrt{Y}$ und hoffen, dass dadurch beide Probleme gelöst werden.



Prototypen von Transformationen von Y

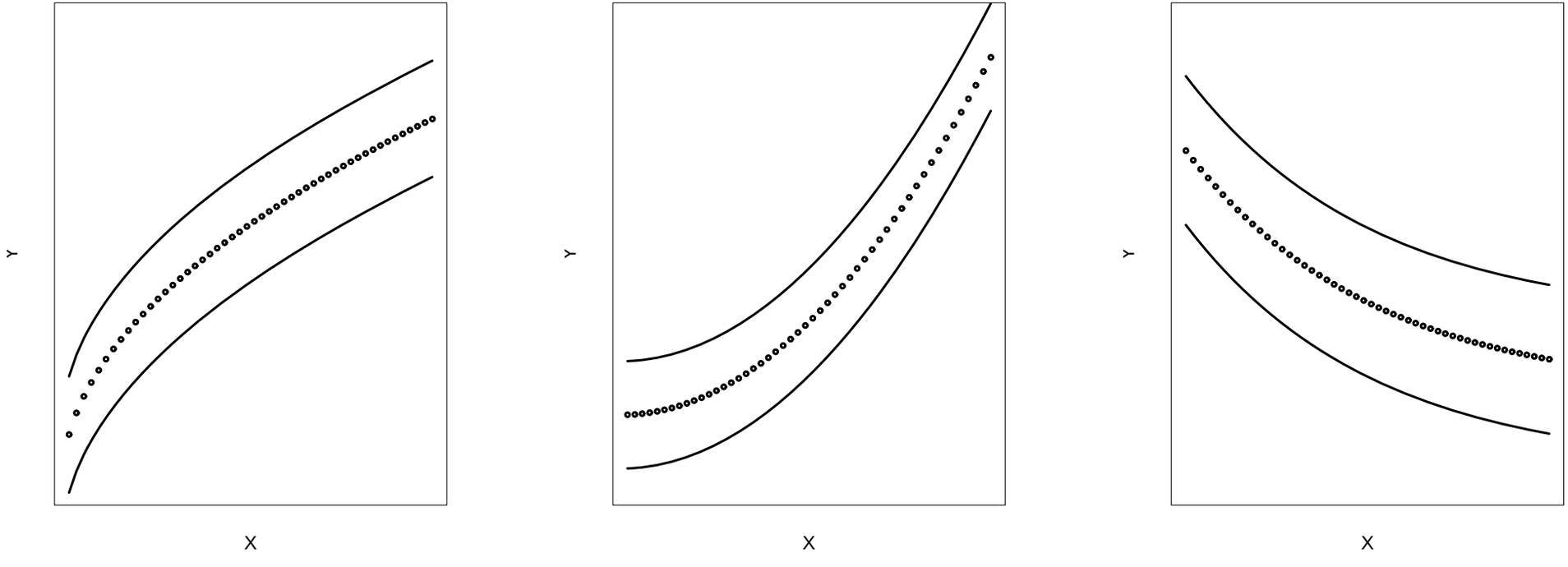






Versuche $\sqrt{Y}$, $\log_{10} Y$, oder $1/Y$.

Prototypen von Transformationen von x

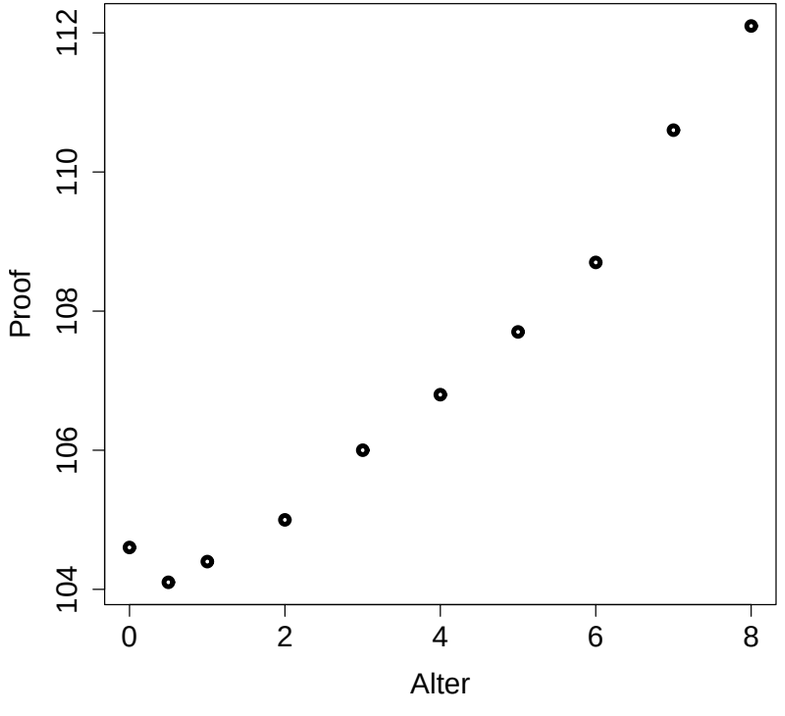


Versuche $\sqrt{x}$ oder $\log_{10} x$ (links); x^2 oder $\exp(x)$ (Mitte); $1/x$ oder $\exp(-x)$ (rechts).

4. Simultane Inferenz

Daten: $x_i =$ Alter von Whiskey, $Y_i =$ Proof (57% Volumenalkohol $= 100°$ Proof)

x	Y
0	104.6
0.5	104.1
1	104.4
2	105
3	106
4	106.8
5	107.7
6	108.7
7	110.6
8	112.1



Unter der Annahme eines SLR resultiert: $\widehat{\mathsf{E}}(Y) = 103.5 + 0.955 \cdot x$.

Weiters folgt $r^2 = 0.9487$, $\sqrt{\mathsf{MSE}} = 0.6617$, $\bar{x} = 3.65$, $s_x^2 = 71.025$.

Wir wollen 2 Typen von Whiskey verkaufen:
- 2 Jahre alt
- 5 Jahre alt

Die Behörden verlangen Konfidenzintervalle für die mittleren Proofs mit **simultaner/gemeinsamer** Überdeckungswahrscheinlichkeit von 95%.

- Proof nach 2 Jahren: $\mathsf{E}(Y_2) = \beta_0 + \beta_1 \cdot 2$
- Proof nach 5 Jahren: $\mathsf{E}(Y_5) = \beta_0 + \beta_1 \cdot 5$

- 95% Konfidenzintervall für $\mathsf{E}(Y_2)$ ist

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot 2 \pm t_{0.975;8} \sqrt{\mathsf{MSE}\left(\frac{1}{10} + \frac{(2-\bar{x})^2}{s_x^2}\right)} = (104.86, 105.99)$$

- 95% Konfidenzintervall für $\mathrm{E}(Y_5)$ ist

$$\hat{\beta}_0 + \hat{\beta}_1 \cdot 5 \pm t_{0.975;8} \sqrt{\mathrm{MSE}\left(\frac{1}{10} + \frac{(5-\bar{x})^2}{s_x^2}\right)} = (107.75, 108.83)$$

Man sagt, ein Konfidenzintervall **überdeckt**, falls es den wahren Parameter enthält.

Definiere A_2 als Ereignis, dass das Konfidenzintervall für $\mathrm{E}(Y_2)$ überdeckt. Definiere A_5 als Ereignis, dass das Konfidenzintervall für $\mathrm{E}(Y_5)$ überdeckt.

Frage: Gilt dann

$$\Pr(A_2) = 0.95\,, \quad \Pr(A_5) = 0.95\,, \quad \Pr(A_2 \cap A_5) \stackrel{?}{=} 0.95$$

Nein! Machen wir einige Experimente:

Exp't #	A_2	A_5	$A_2 \cap A_5$
1	ja	ja	ja
2	ja	nein	nein
3	ja	ja	ja
4	nein	ja	nein
5	ja	ja	ja
6	ja	ja	ja
⋮	⋮	⋮	⋮

Es werden um die 5% „nein“ in der A_2 Spalte und etwa 5% „nein“ in der A_5 Spalte sein. Aber es werden mehr als 5% (jedoch weniger als 10%) „nein“ in der $A_2 \cap A_5$ Spalte sein.

$$\begin{aligned}
\Pr(A_2 \cap A_5) &= \Pr(A_2) + \Pr(A_5) - \Pr(A_2 \cup A_5) \\
&= 1 - \Pr(\overline{A}_2) + 1 - \Pr(\overline{A}_5) - \Pr(A_2 \cup A_5) \\
&= 1 - \Pr(\overline{A}_2) - \Pr(\overline{A}_5) + \Pr(\overline{A_2 \cup A_5}) \\
&\geq 1 - \Big(\Pr(\overline{A}_2) + \Pr(\overline{A}_5) \Big) .
\end{aligned}$$

$\overline{A}_2$ ist das Ereignis, dass das Konfidenzintervall für $\mathrm{E}(Y_2)$ **nicht überdeckt**, wobei $\Pr(\overline{A}_2) = \alpha$.

Folgerung: Um $\Pr(A_2 \cap A_5) \geq 0.95$ zu sichern, muss $\Pr(\overline{A}_2) + \Pr(\overline{A}_5) \leq 0.05$.

Bonferroni Ungleichung: für M derartige Ereignisse A_m gilt

$$\Pr(A_1 \cap A_2 \cap \cdots \cap A_M) \geq 1 - \sum_{m=1}^{M} \Pr(\overline{A}_m) .$$

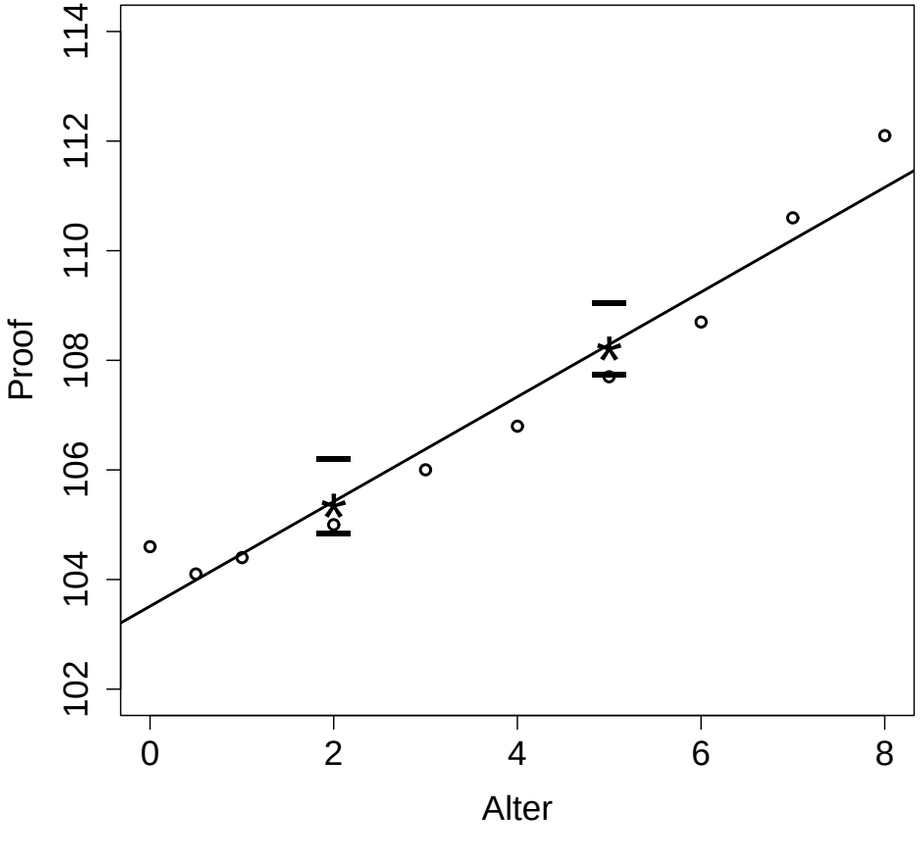
Für die Wahl von $\Pr(\overline{A}_m) = \alpha/M$ folgt

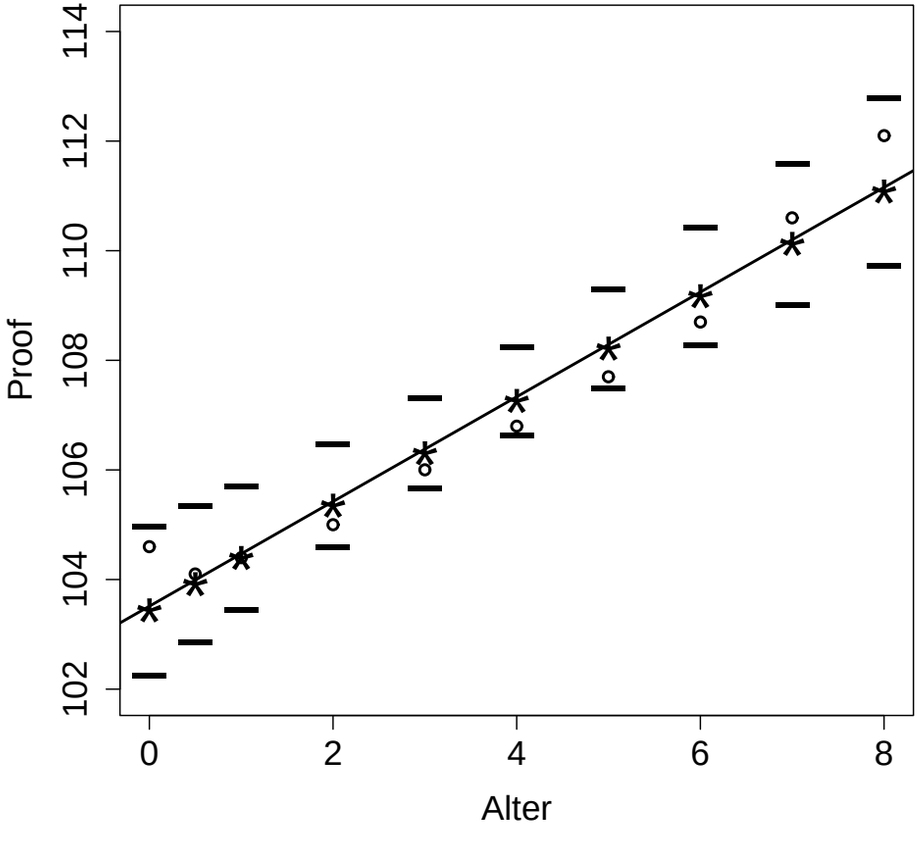
$$\Pr(A_1 \cap A_2 \cap \cdots \cap A_M) \geq 1 - \sum_{m=1}^{M} \frac{\alpha}{M} = 1 - \alpha\,.$$

Damit gilt

$$\begin{aligned}
\hat{\mu}_1 &\pm t_{n-2;1-\alpha/(2M)} \sqrt{\mathsf{MSE}\left(\frac{1}{n} + \frac{(x_1 - \bar{x})^2}{s_x^2}\right)} \\
\hat{\mu}_2 &\pm t_{n-2;1-\alpha/(2M)} \sqrt{\mathsf{MSE}\left(\frac{1}{n} + \frac{(x_2 - \bar{x})^2}{s_x^2}\right)} \\
&\vdots \\
\hat{\mu}_M &\pm t_{n-2;1-\alpha/(2M)} \sqrt{\mathsf{MSE}\left(\frac{1}{n} + \frac{(x_M - \bar{x})^2}{s_x^2}\right)}\,.
\end{aligned}$$

Diese Konfidenzintervalle überdecken mit **gemeinsamer** Überdeckungswahrscheinlichkeit $1 - \alpha$.

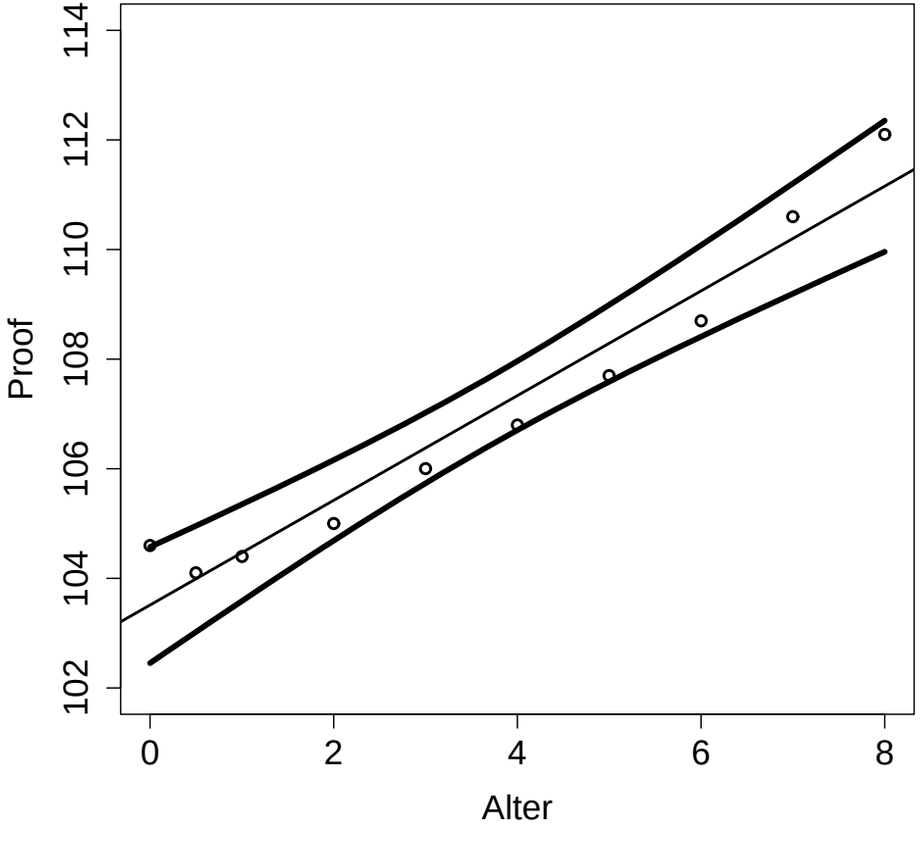




Working-Hotelling Prozedur: Working und Hotelling (1929) konstruierten ein Konfidenzband für die gesamte (wahre) Regressionsgerade, $\beta_0 + \beta_1 x$, im gesamten Bereich der Daten.

Dieses Band ist punktweise konstruiert in x_a

$$\hat{\mu}_a \pm \sqrt{2F_{2,n-2;1-\alpha}} \sqrt{\mathsf{MSE}\left(\frac{1}{n} + \frac{(x_a - \bar{x})^2}{s_x^2}\right)}.$$



Intervall ist am engsten in $\bar{x}$.

Wir sind zu $(1-\alpha)$ sicher, dass die wahre Regressionsfunktion innerhalb dieses Bandes liegt.

Da für das gesamte Band eine Sicherheit von $(1-\alpha)$ hält, können wir so viele Konfidenzintervalle für $\mathrm{E}(Y)$ betrachten wie wir wollen und die **gemeinsame** Sicherheit beträgt zumindest $(1-\alpha)$.

Vergleich der beiden Konfidenzintervalle: ist einfach und basiert auf Vergleich der Werte von

$$t_{n-2;1-\alpha/(2M)} \qquad \text{mit} \qquad \sqrt{2F_{2,n-2;1-\alpha}}\,.$$

Beispiel: Wir wollen 2 Konfidenzintervalle für den mittleren Proof von Whiskey, wenn das Alter 2 und 5 Jahre ist, welche gemeinsam mit 95% überdecken.

Bonferroni: verwende 2 Konfidenzintervalle jedes mit Niveau $\alpha/M = \alpha/2 = 0.025$

$$t_{8;1-0.0125} = 2.751\,.$$

Working-Hotelling:

$$\sqrt{2F_{2,8;0.95}} = 2.986\,.$$

Alter	Individuelle Konfidenzintervalle	Bonferroni simultan	Working-Hotelling
2	(104.86, 105.99)	(104.75, 106.10)	(104.69, 106.16)
5	(107.75, 108.83)	(107.64, 108.94)	(107.59, 108.99)

5. Matrix Algebra: Ein Auftakt zur Multiplen Regression

Wir hatten $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ und formulierten das SLR als

$$Y_i = \mathsf{E}(Y_i) + \epsilon_i\,, \qquad i = 1, 2, \ldots, n\,.$$

Mittels Matrizen und Vektoren kann dies formuliert werden als

$$\mathbf{Y}_{n\times 1} = \mathsf{E}(\mathbf{Y}_{n\times 1}) + \boldsymbol{\epsilon}_{n\times 1}$$

mit den $n \times 1$ Spaltenvektoren

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathsf{E}(\mathbf{Y}) = \begin{bmatrix} \mathsf{E}(Y_1) \\ \mathsf{E}(Y_2) \\ \vdots \\ \mathsf{E}(Y_n) \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Alle Erwartungswerte bilden die Gerade

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 x_i \,, \qquad i = 1, 2, \ldots, n \,.$$

Mit

$$\mathbf{X}_{n\times 2} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{\beta}_{2\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

erhalten wir

$$\mathsf{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} .$$

Somit schreiben wir das SLR als

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \,.$$

Wichtige Matrizen in der Regression:

$$\mathbf{Y'Y} = \begin{bmatrix} Y_1 & Y_2 & \dots & Y_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \sum_{i=1}^{n} Y_i^2$$

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}$$

$$\mathbf{X'Y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \sum_i Y_i \\ \sum_i x_i Y_i \end{bmatrix}$$

Wichtigste inverse Matrix in der Regression ist die Inverse von

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix}.$$

Ihre Determinante ist

$$D \quad = \quad n\sum_i x_i^2 - \Big(\sum_i x_i\Big)^2 = n\left(\sum_i x_i^2 - n\bar{x}^2\right) = ns_x^2 \neq 0\,.$$

Somit folgt

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{ns_x^2}\begin{bmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix} = \frac{1}{s_x^2}\begin{bmatrix} \frac{1}{n}\sum_i x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}.$$

Einige wichtige Eigenschaften von Matrizen:

1. $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
2. $\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB}$
3. $(\mathbf{A}')' = \mathbf{A}$
4. $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$
5. $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$
6. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
7. $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
8. $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$
9. $(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$

Zufallsvektoren und Matrizen

Ein Zufallsvektor ist ein Vektor von Zufallsvariablen, z.B. $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$.

Der **Erwartungswert** von $\mathbf{Y}$ ist der Vektor $\mathsf{E}(\mathbf{Y}) = (\mathsf{E}(Y_1), \mathsf{E}(Y_2), \ldots, \mathsf{E}(Y_n))'$.

So gilt im SLR beispielsweise

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \qquad \text{mit} \qquad \mathsf{E}(\boldsymbol{\epsilon}) = \begin{bmatrix} \mathsf{E}(\epsilon_1) \\ \mathsf{E}(\epsilon_2) \\ \vdots \\ \mathsf{E}(\epsilon_n) \end{bmatrix} = \mathbf{0}_{n\times 1}\,.$$

Natürlich gelten weiterhin die Regeln für Erwartungswerte:

Angenommen $\mathbf{V}$ und $\mathbf{W}$ sind Zufallsvektoren und $\mathbf{A}$, $\mathbf{B}$ und $\mathbf{C}$ konstante Matrizen. Damit folgt

$$\mathrm{E}(\mathbf{AV} + \mathbf{BW} + \mathbf{C}) = \mathbf{A}\,\mathrm{E}(\mathbf{V}) + \mathbf{B}\,\mathrm{E}(\mathbf{W}) + \mathbf{C}\,.$$

Anwendung: Für $\mathrm{E}(\mathbf{Y}) = \mathrm{E}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})$ resultiert

$$\mathrm{E}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathrm{E}(\mathbf{X}\boldsymbol{\beta}) + \mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{0} = \mathbf{X}\boldsymbol{\beta}\,.$$

Varianz/Kovarianz Matrix eines Zufallsvektors: Für den Zufallsvektor $\mathbf{Z}_{n\times 1}$ ist

$$\mathsf{var}(\mathbf{Z}) = \begin{bmatrix} \mathsf{var}(Z_1) & \mathsf{cov}(Z_1, Z_2) & \dots & \mathsf{cov}(Z_1, Z_n) \\ \mathsf{cov}(Z_2, Z_1) & \mathsf{var}(Z_2) & \dots & \mathsf{cov}(Z_2, Z_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{cov}(Z_n, Z_1) & \mathsf{cov}(Z_n, Z_2) & \dots & \mathsf{var}(Z_n) \end{bmatrix},$$

wobei $\mathsf{cov}(Z_i, Z_j) = \mathsf{E}\Big[(Z_i - \mathsf{E}(Z_i))(Z_j - \mathsf{E}(Z_j))\Big] = \mathsf{cov}(Z_j, Z_i)$ gilt. Dies ist eine symmetrische $(n \times n)$ Matrix.

Für unabhängige Z_i und Z_j ist $\mathsf{cov}(Z_i, Z_j) = 0$. Im SLR sind n unabhängige, zufällige Fehler ϵ_i, jeder mit gleicher Varianz σ^2. Dafür ist

$$\mathsf{var}(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_{n\times n}\,.$$

Regeln für Varianz/Kovarianz Matrizen:

Zur Erinnerung: falls V eine Zufallsvariable ist und a, b Konstanten, dann ist

$$\text{var}(aV + b) = \text{var}(aV) = a^2\text{var}(V)\,.$$

Sei nun $\mathbf{V}$ ein Zufallsvektor und $\mathbf{A}$, $\mathbf{B}$ konstante Matrizen, dann ist

$$\text{var}(\mathbf{AV} + \mathbf{B}) = \text{var}(\mathbf{AV}) = \mathbf{A}\text{var}(\mathbf{V})\mathbf{A}'\,.$$

Damit resultiert unter dem SLR

$$\text{var}(\mathbf{Y}) = \text{var}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \text{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}_{n\times n}\,.$$

Alle Nebendiagonalelemente sind Null, da die ϵ_i's, und somit die Y_i's, unabhängig sind.

SLR in Matrixschreibweise

Wir schreiben das SLR mittels Matrizen kompakt als

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

und wir nehmen weiters an, dass

- $\boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, \sigma^2\mathbf{I})$
- $\boldsymbol{\beta}$ und σ^2 sind unbekannte Parameter
- $\mathbf{X}$ ist eine konstante Matrix.

Konsequenzen: $\text{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ and $\text{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$.

Wir definieren nun die Kleinsten Quadrate Schätzer $(\hat{\beta}_0, \hat{\beta}_1)$ in Matrixnotation.

Normal-Gleichungen: Kleinste Quadrate Kriterium

$$\mathrm{SSE}(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i))^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Nimmt man die Ableitungen von $\mathrm{SSE}(\beta_0, \beta_1)$ nach β_0 und β_1 und setzt beide Null, dann liefert dies die Normal-Gleichungen

$$\begin{aligned} n\hat{\beta}_0 + \quad n\bar{x}\hat{\beta}_1 &= n\bar{Y} \\ n\bar{x}\hat{\beta}_0 + \sum_i x_i^2 \hat{\beta}_1 &= \sum_i x_i Y_i \,. \end{aligned}$$

Wir schreiben diese Gleichungen in Matrixform als

$$\begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum_i x_i^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} n\bar{Y} \\ \sum_i x_i Y_i \end{bmatrix} .$$

Aber dies ist mit $\hat{\boldsymbol{\beta}}_{2\times 1} = (\hat{\beta}_0, \hat{\beta}_1)'$ genau äquivalent mit

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{Y})\,.$$

Multiplikation dieser Gleichung mit der Inversen $(\mathbf{X}'\mathbf{X})^{-1}$ (falls diese existiert) von links liefert eine explizite Form der Kleinsten Quadrate Schätzer

$$\begin{aligned}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \\ \hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\,.\end{aligned}$$

Fitted Values und Residuen

Zur Erinnerung ist $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Wegen

$$\begin{bmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_1 x_1 \\ \vdots \\ \hat{\beta}_0 + \hat{\beta}_1 x_n \end{bmatrix}$$

kann man den Vektor der Fitted Values schreiben als

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}\,.$$

Mit $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ bekommt man

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\,.$$

Alternative Schreibweise der Fitted Values unter Verwendung der $(n \times n)$ **Hat Matrix**

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

($\mathbf{H}$ setzt den Hut auf $\boldsymbol{\mu}$) ergibt

$$\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{Y}\,.$$

Die Matrix $\mathbf{H}$ ist **symmetrisch** $(\mathbf{H} = \mathbf{H}')$ und **idempotent** $(\mathbf{H}\mathbf{H} = \mathbf{H})$

Symmetrisch:

$$\begin{aligned}\mathbf{H}' &= \Big(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Big)' \overset{9.}{=} \mathbf{X}\Big((\mathbf{X}'\mathbf{X})^{-1}\Big)'\mathbf{X}' \\ &\overset{8.}{=} \mathbf{X}\Big((\mathbf{X}'\mathbf{X})'\Big)^{-1}\mathbf{X}' \overset{5.}{=} \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}\,.\end{aligned}$$

Idempotent: Wegen $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$ haben wir

$$\begin{aligned}\mathbf{HH} &= \Big(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Big)\Big(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Big) \\ &= \mathbf{XI}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}\,.\end{aligned}$$

Mit diesen Resultaten bekommen wir sofort $\mathbf{HX} = \mathbf{X}$, sowie damit

$$\begin{aligned}\mathsf{E}(\hat{\boldsymbol{\mu}}) &= \mathsf{E}(\mathbf{HY}) = \mathbf{HX}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}\,, \\ \mathsf{var}(\hat{\boldsymbol{\mu}}) &= \mathsf{var}(\mathbf{HY}) = \mathbf{H}\ \sigma^2\mathbf{I}\ \mathbf{H} = \sigma^2\mathbf{H}.\end{aligned}$$

Die Fitted Values $\hat{\boldsymbol{\mu}}$ sind somit erwartungstreue Schätzer der unbekannten Erwartungswerte der Responsevariablen. Weiters sind sie nicht unabhängig und haben individuelle Varianzen.

Residuen: $\mathbf{r} = \mathbf{Y} - \hat{\boldsymbol{\mu}} = \mathbf{I}\mathbf{Y} - \mathbf{H}\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$.

Wie $\mathbf{H}$ ist auch $\mathbf{I} - \mathbf{H}$ symmetrisch und idempotent. Damit erhält man

$$\begin{aligned} \mathsf{E}(\mathbf{r}) &= (\mathbf{I} - \mathbf{H})\,\mathsf{E}(\mathbf{Y}) = (\mathbf{I} - \mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} = \mathbf{0} \\ \mathsf{var}(\mathbf{r}) &= \mathsf{var}((\mathbf{I} - \mathbf{H})\mathbf{Y}) = \sigma^2(\mathbf{I} - \mathbf{H})\,. \end{aligned}$$

Alle Residuen haben Erwartung Null, sind aber nicht unabhängig und haben individuelle Varianzen.

Inferenz in Regressionsmodellen

Zur Verteilung der Kleinsten Quadrate Schätzer:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{C}\mathbf{Y} = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ c_{21} & \cdots & c_{2n} \end{bmatrix} \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},$$

wobei $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ eine $2 \times n$ Matrix von Konstanten ist. Somit ist jedes Element von $\hat{\boldsymbol{\beta}}$ eine Linearkombination von unabhängigen, normalverteilten Zufallsvariablen Y_i und daher eine bivariat normalverteilte Zufallsvariable. Also gilt

$$\hat{\boldsymbol{\beta}} \sim \mathsf{Normal}_2(\mathsf{E}(\hat{\boldsymbol{\beta}}), \mathsf{var}(\hat{\boldsymbol{\beta}}))\,.$$

Als Momente folgen

$$\mathsf{E}(\hat{\boldsymbol{\beta}}) = \mathsf{E}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathsf{E}(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}\,,$$

sowie

$$\begin{aligned}\mathsf{var}(\hat{\boldsymbol{\beta}}) &= \mathsf{var}\Big((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\Big) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ \mathsf{var}(\mathbf{Y})\ \Big((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Big)' \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ \sigma^2\mathbf{I}\ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\,.\end{aligned}$$

Als Schätzer für diese Matrix verwenden wir

$$\widehat{\mathsf{var}}(\hat{\boldsymbol{\beta}}) = \mathsf{MSE}\cdot(\mathbf{X}'\mathbf{X})^{-1} = \frac{\mathsf{MSE}}{s_x^2}\begin{bmatrix} \frac{1}{n}\sum_i x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}.$$

Als Kovarianz/Korrelation zwischen $\hat{\beta}_0$ und $\hat{\beta}_1$ bekommen wir

$$\begin{aligned} \mathsf{cov}(\hat{\beta}_0, \hat{\beta}_1) &= -\frac{\sigma^2}{s_x^2}\,\bar{x} \\ \mathsf{cor}(\hat{\beta}_0, \hat{\beta}_1) &= \frac{\mathsf{cov}(\hat{\beta}_0, \hat{\beta}_1)}{\sqrt{\mathsf{var}(\hat{\beta}_0)\mathsf{var}(\hat{\beta}_1)}} = \frac{-\bar{x}}{\sqrt{\frac{1}{n}\sum_i x_i^2}}\,. \end{aligned}$$

Die Schätzer $\hat{\beta}_0$ und $\hat{\beta}_1$ sind nicht unabhängig! Somit gilt zusammenfassend

$$\hat{\boldsymbol{\beta}} \sim \mathsf{Normal}\Big(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\Big)\,.$$

Dies werden wir verwenden, um Konfidenzintervalle und Tests über $\boldsymbol{\beta}$ herzuleiten.

Schätze den Erwartungswert der Response in x_h:

Zur Erinnerung ist unser Schätzer für $\mathsf{E}(Y_h) = \beta_0 + \beta_1 x_h$ gleich

$$\hat{\mu}_h = \hat{\beta}_0 + \hat{\beta}_1 x_h = \mathbf{X}_h' \hat{\boldsymbol{\beta}}$$

mit $\mathbf{X}_h' = (1, x_h)$. Der Fitted Value ist eine normalverteilte Zufallsvariable mit

$$\begin{aligned} \mathsf{E}(\hat{\mu}_h) &= \mathsf{E}(\mathbf{X}_h' \hat{\boldsymbol{\beta}}) = \mathbf{X}_h' \, \mathsf{E}(\hat{\boldsymbol{\beta}}) = \mathbf{X}_h' \boldsymbol{\beta} \\ \mathsf{var}(\hat{\mu}_h) &= \mathsf{var}(\mathbf{X}_h' \hat{\boldsymbol{\beta}}) = \mathbf{X}_h' \mathsf{var}(\hat{\boldsymbol{\beta}}) \mathbf{X}_h = \sigma^2 \mathbf{X}_h' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_h \, . \end{aligned}$$

Damit folgt

$$\frac{\hat{\mu}_h - \mathbf{X}_h' \boldsymbol{\beta}}{\sqrt{\sigma^2 \cdot \mathbf{X}_h' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_h}} \sim \mathsf{Normal}(0,1) \quad \Rightarrow \quad \frac{\hat{\mu}_h - \mathbf{X}_h' \boldsymbol{\beta}}{\sqrt{\mathsf{MSE} \cdot \mathbf{X}_h' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_h}} \sim t_{n-2} \, .$$

Was ist hierbei im SLR das Skalar $\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h$?

$$
\begin{aligned}
\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h &= \begin{bmatrix} 1 & x_h \end{bmatrix} \frac{1}{s_x^2} \begin{bmatrix} \frac{1}{n}\sum_i x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} 1 \\ x_h \end{bmatrix} \\
&= \frac{1}{s_x^2} \begin{bmatrix} \frac{1}{n}\sum_i x_i^2 - \bar{x}x_h & -\bar{x} + x_h \end{bmatrix} \begin{bmatrix} 1 \\ x_h \end{bmatrix} \\
&= \frac{1}{s_x^2} \left(\frac{1}{n} \sum_i x_i^2 - \bar{x}x_h - \bar{x}x_h + x_h^2 \right) \\
&= \frac{1}{s_x^2} \left(\frac{1}{n} \Big(s_x^2 + n\bar{x}^2 \Big) - 2\bar{x}x_h + x_h^2 \right) \\
&= \frac{1}{n} + \frac{(x_h - \bar{x})^2}{s_x^2}
\end{aligned}
$$

wegen $s_x^2 = \sum_i x_i^2 - n\bar{x}^2$. var$(\hat{\mu}_h)$ ist umso größer, je weiter x_h von $\bar{x}$ weg ist.

Matrix Algebra mit R: Whiskey Beispiel

```
> one <- rep(1, 10);  age <- c(0, 0.5, 1, 2, 3, 4, 5, 6, 7, 8)
> y <- c(104.6, 104.1, 104.4, 105.0, 106.0, 106.8, 107.7, 108.7, 110.6, 112.1)
> X <- matrix(c(one, age), ncol=2)
> (XtX <- t(X) %*% X)
     [,1]   [,2]
[1,] 10.0  36.50
[2,] 36.5 204.25

> solve(XtX)
            [,1]        [,2]
[1,]  0.28757480 -0.05139036
[2,] -0.05139036  0.01407955

> (b <- solve(XtX) %*% t(X)%*%y)
            [,1]
[1,] 103.5131644
[2,]   0.9552974
```

```
> H <- X %*% solve(XtX) %*% t(X)
> r <- y - H %*% y;  (SSE <- t(r) %*% r)
         [,1]
[1,] 3.503069

> as.numeric(SSE/8) * solve(XtX)
             [,1]          [,2]
[1,]  0.12592431 -0.022502997
[2,] -0.02250300  0.006165205


> summary(lm(y ~ age))

Coefficients:
             Estimate Std.Error t value Pr(>|t|)
(Intercept)  103.51316  0.35486 291.70  < 2e-16 ***
age            0.95530  0.07852  12.17 1.93e-06 ***
```

6. Multiple Lineare Regression

SLR: 1 Prädiktor x, **MLR:** mehr als 1 Prädiktor

Beispiel:
Y_i = #Punkte erzielt vom UF Football Team im Spiel i
x_{i1} = #gewonnene Spiele des Gegners in dessen letzten 10 Spielen
x_{i2} = #gesunde Starter (Stammspieler) für UF (von 22) im Spiel i

i	Punkte	x_{i1}	x_{i2}
1	47	6	18
2	24	9	16
3	60	3	19
⋮	⋮	⋮	⋮

Einfachstes multiples lineares Regressionsmodel (MLR):

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i\,, \qquad i = 1, 2, \ldots, n$$

- $\epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2)$
- β_0, β_1, β_2 und σ^2 sind unbekannte Parameter
- x_{ij}'s sind bekannte Konstanten.

SLR: $\text{E}(Y) = \beta_0 + \beta_1 x$
β_1 ist die Änderung in $\text{E}(Y)$ bei einer Zunahme von einer Einheit von x.

MLR: $\text{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
Haben wir mehr als 1 Prädiktor, dann müssen wir darüber nachdenken, wie sie sich gegenseitig beeinflussen.

Angenommen wir fixieren $x_{i1} = 5$ (Spiele gewonnen vom $i$ten Gegner):

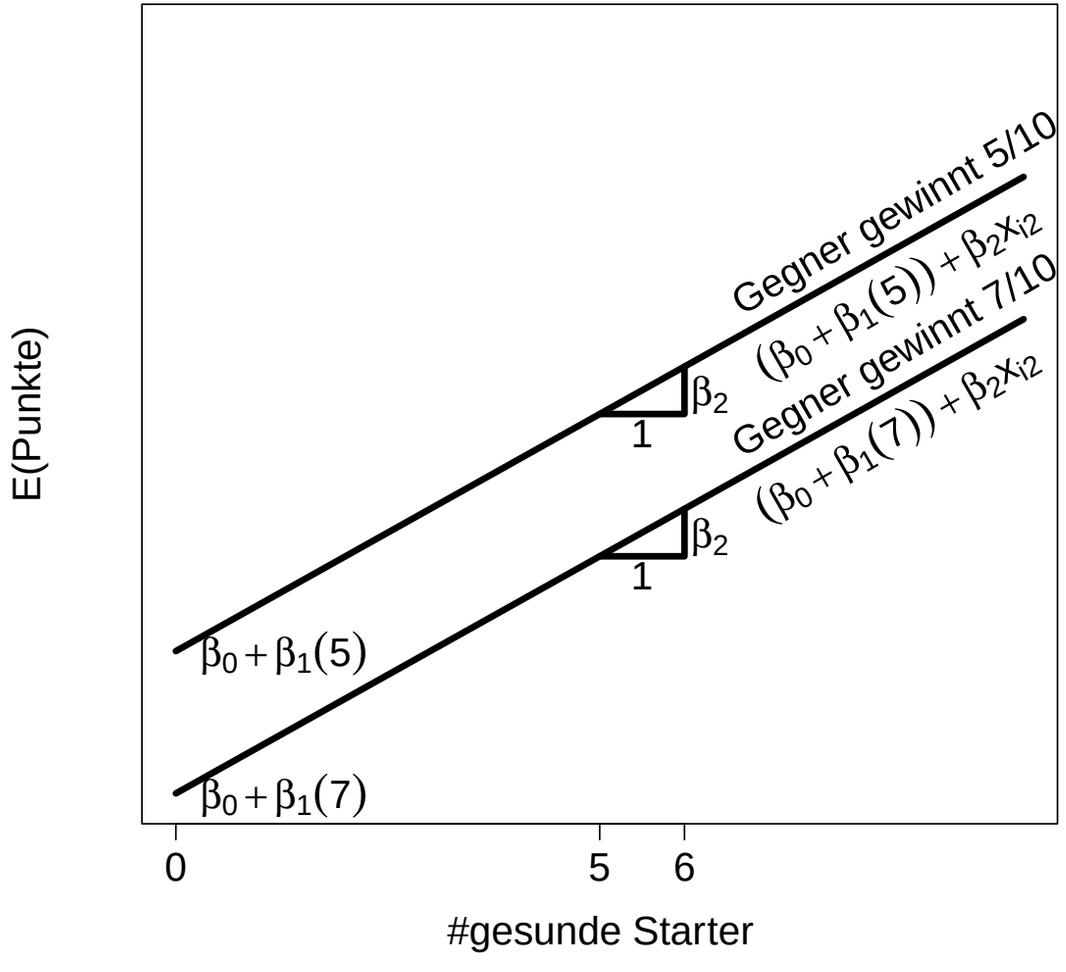
$$\mathsf{E}(Y_i) \quad = \quad (\beta_0 + \beta_1 \cdot 5) + \beta_2 x_{i2}$$

Angenommen wir fixieren $x_{i1} = 7$:

$$\mathsf{E}(Y_i) \quad = \quad (\beta_0 + \beta_1 \cdot 7) + \beta_2 x_{i2}\,.$$

Wir erhalten dadurch SLR Modelle mit unterschiedlichen Intercepts, aber gleichen Steigungen.

Plot von $\mathsf{E}(Y)$ gegen x_2 für feste Werte von x_1 könnte folgendermaßen aussehen:



Unter diesem Modell nehmen wir an, dass für einen beliebigen Wert von x_{i1} (gegnerische Siege), die Änderung in E(Y) bzgl. der Hinzunahme eines weiteren gesunden Starters gleich β_2 für alle Spiele ist.

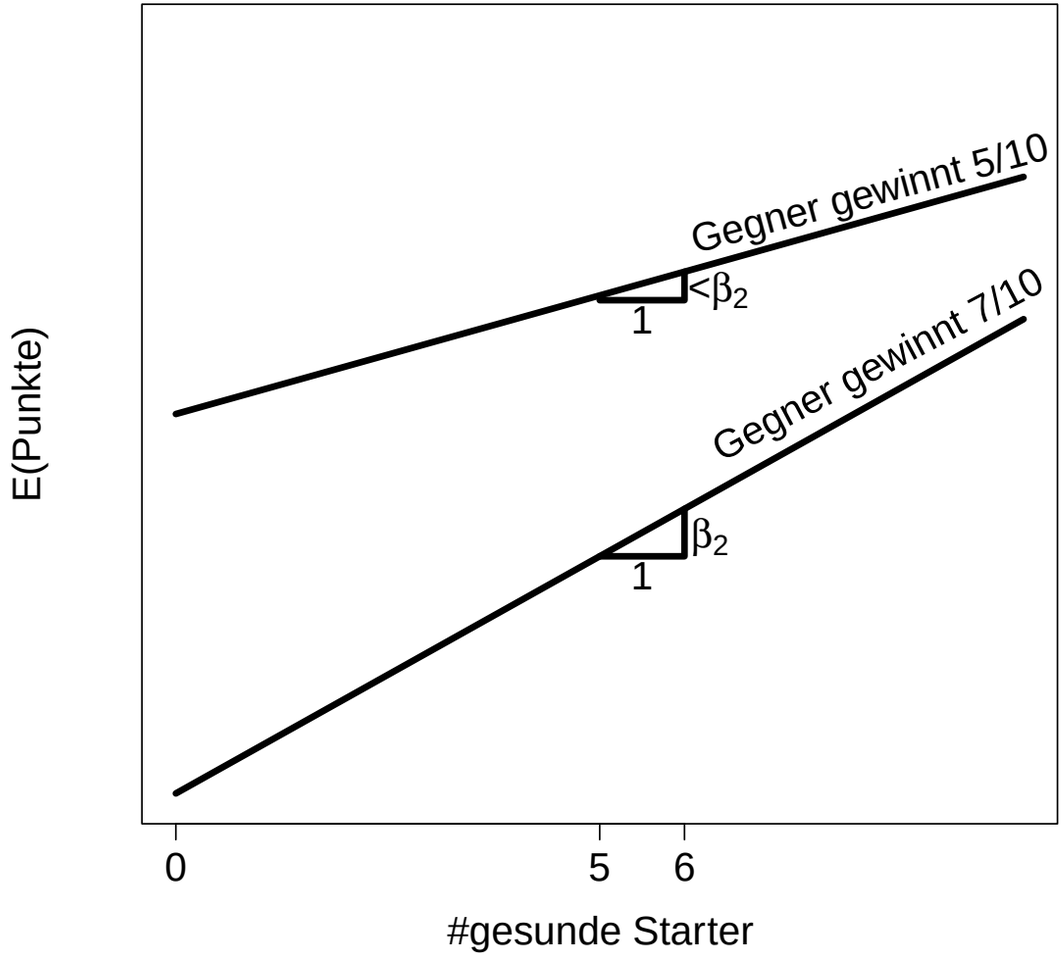
Ist dies glaubwürdig?

Angenommen, AU ist sieglos in den letzten 10 Spielen. Unser Modell nimmt an, dass wir bei Hinzunahme eines weiteren gesunden Starters erwarten, dass UF dadurch zusätzliche β_2 Punkte macht.

Angenommen, BU hat die letzten 10 Spiele gewonnen. Wiederum, wenn wir einen weiteren gesunden Starter bringen, dann erwarten wir, dass UF deswegen zusätzliche β_2 Punkte macht.

Starter werden voraussichtlich gar nicht gegen AU spielen. Daher erwarten wir, gar nichts dazu zu gewinnen falls ein Starter gesund wird und eingesetzt werden könnte.

Vielleicht sollte für diese Situation der Plot folgendermaßen aussehen:



Geringere Steigung, da Starter gegen schlechtere Teams selten eingesetzt werden und daher weniger wichtig sind.

Q: Wie kann man das Modell verändern um ein derartiges Verhalten zu ermöglichen?

A: Hinzunahme einer Interaktion!

Multiples Lineares Regressionsmodell (MLR)

$$\mathsf{E}(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}\,.$$

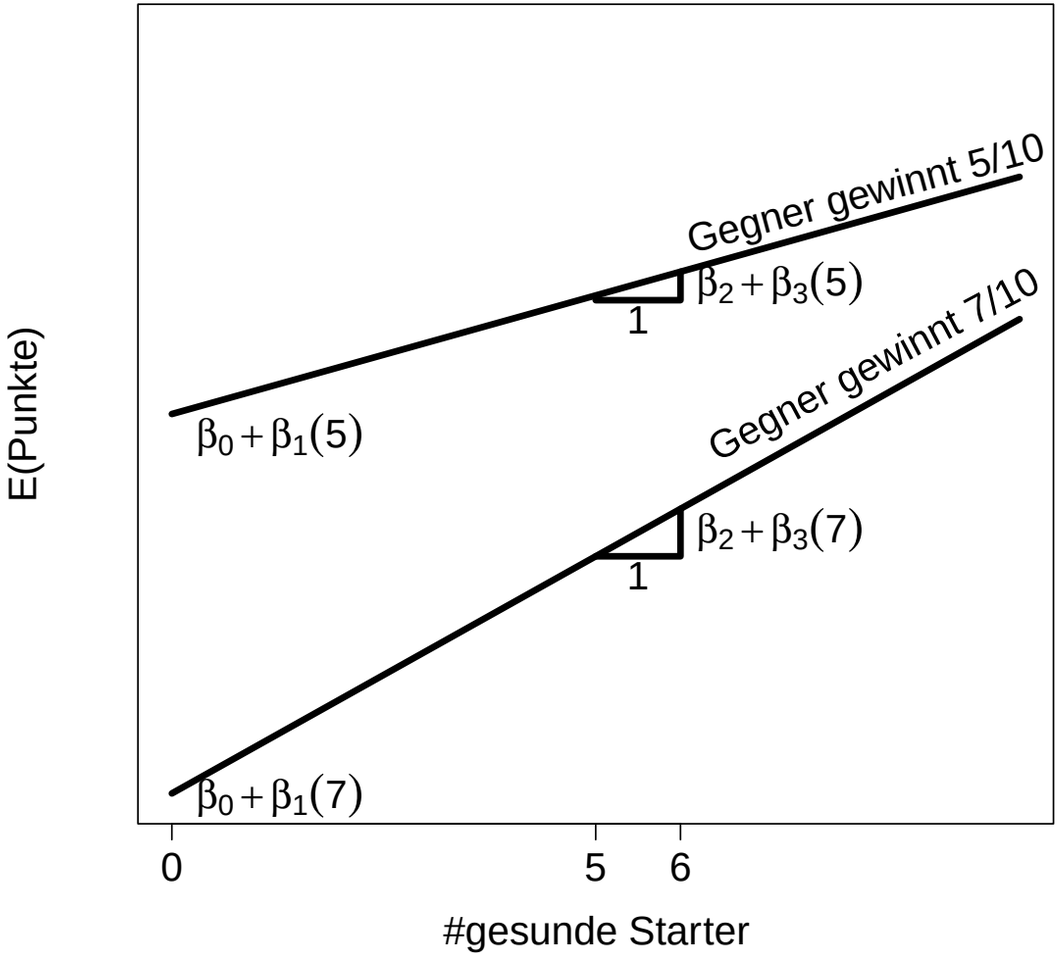
Diese Funktion ist nicht länger eine einfache Ebene!

Für $x_{i1} = 5$:

$$\mathsf{E}(Y_i) = (\beta_0 + \beta_1 \cdot 5) + (\beta_2 + \beta_3 \cdot 5) x_{i2}$$

Für $x_{i1} = 7$:

$$\mathsf{E}(Y_i) = (\beta_0 + \beta_1 \cdot 7) + (\beta_2 + \beta_3 \cdot 7) x_{i2}$$



Nun hängt der Zuwachs an erwarteten Punkten, der durch Hinzunahme eines weiteren gesunden Starters begründet ist, genau so von x_{i1} ab wie er sollte.

$\beta_1 < 0$,
$\beta_2 > 0$, $\beta_3 > 0$

Allgemeines Lineares Regressionsmodell

Daten $(x_{i1}, x_{i2}, \ldots, x_{i,p-1}, Y_i)$, $i = 1, 2, \ldots, n$

Modellgleichung und Annahmen

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

- $\epsilon_i \overset{iid}{\sim}$ Normal$(0, \sigma^2)$
- $\beta_0, \beta_1, \ldots, \beta_{p-1}$ und σ^2 sind unbekannte Parameter
- x_{ij}'s sind bekannte Konstanten.

Zwei Fälle:

1. $p-1$ unterschiedliche Prädiktoren
2. einige der Prädiktoren sind Funktionen von den anderen
   (a) Polynomiale Regression:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

   Definiere $z_{i1} = x_i$ und $z_{i2} = x_i^2$, dann

$$Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \epsilon_i$$

   (b) Interaktionseffekte:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2} + \epsilon_i$$

   Definiere $x_{i3} = x_{i1} x_{i2}$. Damit haben wir wieder ein allgemeines lineares Regressionsmodell.

(c) Sowohl (a) als auch (b):

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1} x_{i2} + \epsilon_i$$

Mit $z_{i1} = x_{i1}$, $z_{i2} = x_{i2}$, $z_{i3} = x_{i1}^2$, $z_{i4} = x_{i2}^2$, $z_{i5} = x_{i1} x_{i2}$ wird dies wieder ein allgemeines lineares Regressionsmodell

$$Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4} + \beta_5 z_{i5} + \epsilon_i \,.$$

Multiple lineare Regression

Betrachte ein lineares Modell mit mehreren $(p-1)$ erklärenden Variablen

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

mit Responsevektor $\mathbf{y} = (y_1, \ldots, y_n)'$, Parametervektor $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})'$ und Fehlervektor $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$, sowie der $n \times p$ Designmatrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \ldots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{n,p-1} \end{pmatrix} .$$

Wie zuvor nehmen wir an, dass

$$\mathbf{y} \sim \text{Normal}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n) \qquad \Longleftrightarrow \qquad \boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, \sigma^2\mathbf{I}_n) .$$

Schätzer für die Parameter

Der **Kleinste Quadrate Schätzer (LSE)** $\hat{\boldsymbol{\beta}}$ minimiert

$$\mathrm{SSE}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\,.$$

Als Ableitungsvektor erhalten wir

$$\frac{\partial}{\partial \beta}\mathrm{SSE}(\beta) = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\,.$$

Das Minimum ist somit definiert als Lösung von

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}\,.$$

Falls $\mathbf{X}'\mathbf{X}$ regulär, existiert die Inverse und wir erhalten explizit

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}\,.$$

Ist dies ein Minimum (ist die Matrix der zweiten Ableitung positiv semidefinit)?

$$\frac{\partial^2}{\partial\beta\partial\beta'}\mathsf{SSE}(\boldsymbol{\beta}) = 2\mathbf{X}'\mathbf{X} > 0\,.$$

Der **Maximum-Likelihood Schätzer** maximiert die (Log-) Likelihood-Funktion

$$\begin{aligned}\log f(\mathbf{y}|\boldsymbol{\beta},\sigma^2) &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i(\boldsymbol{\beta}))^2 \\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\mathsf{SSE}(\boldsymbol{\beta}))\,.\end{aligned}$$

Die Maximierung von $\log f(\mathbf{y}|\boldsymbol{\beta},\sigma^2)$ in $\boldsymbol{\beta}$ ist somit unabhängig vom Wert von σ^2 und äquivalent mit der Minimierung von $\mathsf{SSE}(\boldsymbol{\beta})$. Daher entspricht der MLE $\hat{\boldsymbol{\beta}}$ dem LSE $\hat{\boldsymbol{\beta}}$.

Jede Komponente des Vektors $\hat{\boldsymbol{\beta}}$ ist eine Linearkombination der normalverteilten Responses, weshalb

$$\hat{\boldsymbol{\beta}} \sim \text{Normal}\left(\text{E}(\hat{\boldsymbol{\beta}}), \text{var}(\hat{\boldsymbol{\beta}})\right)$$

gilt mit

$$\text{E}(\hat{\boldsymbol{\beta}}) = \text{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\text{E}(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

und

$$\text{var}(\hat{\boldsymbol{\beta}}) = \text{var}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{var}(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\,.$$

Als **Prognosevektor (Fitted Values)** erhalten wir

$$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

mit der symmetrischen $n \times n$ Hat-Matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,.$$

$\mathbf{H}$ ist auch idempotent, da

$$\mathbf{H}\mathbf{H}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}\,.$$

Der Vektor der **Residuen** ist

$$\mathbf{r} = \mathbf{y} - \hat{\boldsymbol{\mu}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y}\,.$$

Auch $\mathbf{I} - \mathbf{H}$ ist symmetrisch und idempotent, da

$$(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' = \mathbf{I} - 2\mathbf{H} + \mathbf{H} = \mathbf{I} - \mathbf{H}\,.$$

Bemerke, dass die Elemente in $\hat{\boldsymbol{\mu}}$ und $\mathbf{r}$ Linearkombinationen der normalverteilten Responses sind und damit auch gilt

$$\hat{\boldsymbol{\mu}} \sim \text{Normal}\left(\text{E}(\hat{\boldsymbol{\mu}}), \text{var}(\hat{\boldsymbol{\mu}})\right)\,, \qquad \mathbf{r} \sim \text{Normal}\left(\text{E}(\mathbf{r}), \text{var}(\mathbf{r})\right)\,.$$

Für die Momente erhält man

$$\begin{aligned} \mathsf{E}(\hat{\boldsymbol{\mu}}) &= \mathbf{X}\,\mathsf{E}(\hat{\boldsymbol{\beta}}) = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu} \\ \mathsf{var}(\hat{\boldsymbol{\mu}}) &= \mathbf{X}\mathsf{var}(\hat{\boldsymbol{\beta}})\mathbf{X}' = \sigma^2\mathbf{H}\,, \end{aligned}$$

sowie

$$\begin{aligned} \mathsf{E}(\mathbf{r}) &= (\mathbf{I}-\mathbf{H})\,\mathsf{E}(\mathbf{y}) = (\mathbf{I}-\mathbf{H})\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \\ \mathsf{var}(\mathbf{r}) &= (\mathbf{I}-\mathbf{H})\mathsf{var}(\mathbf{y})(\mathbf{I}-\mathbf{H}) = \sigma^2(\mathbf{I}-\mathbf{H})\,. \end{aligned}$$

Schätzer der Responsevarianz

Für die ML Schätzung betrachten wir die beiden Systeme von Score-Funktionen

$$\begin{aligned}
\frac{\partial}{\partial \beta_j} \log f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij}(y_i - \mu_i)\,, \qquad j = 0, \ldots, p-1 \\
\frac{\partial}{\partial \sigma^2} \log f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{\sigma^4} \sum_{i=1}^{n} (y_i - \mu_i)^2\,.
\end{aligned}$$

Simultanes Nullsetzen liefert das System der Normalgleichungen und als Lösung den MLE $\hat{\boldsymbol{\beta}}$ (unabhängig von σ^2), sowie den MLE für die Varianz

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2 = \frac{1}{n} \mathsf{SSE}(\hat{\boldsymbol{\beta}})\,.$$

(1) Zeige, $\hat{\boldsymbol{\beta}}$ und $\mathbf{r}$, also $\hat{\boldsymbol{\beta}}$ und $\mathrm{SSE}(\hat{\boldsymbol{\beta}}) = \mathbf{r}'\mathbf{r}$, sind unabhängig, d.h. $\mathrm{cov}(\hat{\boldsymbol{\beta}}, \mathbf{r}) = \mathbf{0}$.

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\
\mathbf{r} &= (\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}\,.
\end{aligned}$$

Dafür folgt

$$\begin{aligned}
\mathrm{cov}(\hat{\boldsymbol{\beta}}, \mathbf{r}) &= \mathrm{cov}\left(\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}, (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}\right) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \underbrace{\mathrm{cov}\left(\boldsymbol{\epsilon}, \boldsymbol{\epsilon}\right)}_{=\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{I}} (\mathbf{I} - \mathbf{H}) \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{H}) \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{0}\,.
\end{aligned}$$

Also sind $\hat{\boldsymbol{\beta}}$ und $\mathrm{SSE}(\hat{\boldsymbol{\beta}}) = \mathbf{r}'\mathbf{r}$ unter Normalverteilungsannahme unabhängig.

(2) Zeige

$$\mathrm{SSE}(\hat{\boldsymbol{\beta}})/\sigma^2 \sim \chi^2_{n-p}$$

Betrachte zuerst

$$\begin{aligned}
\boldsymbol{\epsilon}'\boldsymbol{\epsilon} &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\right)' \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\right) \\
&= \left(\mathbf{r} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\right)' \left(\mathbf{r} + \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}\right) \\
&= \left(\mathbf{r} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)' \left(\mathbf{r} + \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) .
\end{aligned}$$

Wegen $\mathbf{HX} = \mathbf{X}$ verschwindet der gemischte Term, denn damit ist

$$\mathbf{r}'\mathbf{X} = ((\mathbf{I} - \mathbf{H})\mathbf{y})'\,\mathbf{X} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{X} = \mathbf{0}\,.$$

Somit gilt

$$\boldsymbol{\epsilon}'\boldsymbol{\epsilon}/\sigma^2 \quad = \quad \mathbf{r}'\mathbf{r}/\sigma^2 + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \,.$$

Weiters wissen wir, dass

$$\boldsymbol{\epsilon}'\boldsymbol{\epsilon}/\sigma^2 = \sum_{i=1}^{n} (\epsilon_i/\sigma)^2 \sim \chi^2_n \,.$$

Aus der Normalverteilungseigenschaft des MLE $\hat{\boldsymbol{\beta}}$ resultiert unmittelbar

$$\hat{\boldsymbol{\beta}} \sim \mathsf{Normal}_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}) \quad \Longrightarrow \quad (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})^{1/2}/\sigma \sim \mathsf{Normal}_p(\mathbf{0}, \mathbf{I}_p)$$

womit außerdem folgt

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\sigma^2 \sim \chi^2_p \,.$$

Die momentenerzeugende Funktion einer χ^2_n-verteilten Zufallsvariablen ist

$$M(t) = (1-2t)^{-n/2}\,, \quad t < 1/2\,.$$

Auf unsere Zerlegung angewandt liefert dies (da $\mathbf{r}$ und $\hat{\boldsymbol{\beta}}$ unabhängig sind)

$$(1-2t)^{-n/2} = \mathsf{E}(\exp(t \cdot \mathbf{r}'\mathbf{r}/\sigma^2))(1-2t)^{-p/2}\,.$$

Daraus folgt nun direkt

$$\mathsf{E}(\exp(t \cdot \mathbf{r}'\mathbf{r}/\sigma^2)) = (1-2t)^{-(n-p)/2}\,,$$

also

$$\mathbf{r}'\mathbf{r}/\sigma^2 = \frac{1}{\sigma^2}\mathsf{SSE}(\hat{\boldsymbol{\beta}}) \sim \chi^2_{n-p}\,.$$

Somit ist

$$\mathsf{E}(\mathsf{SSE}(\hat{\boldsymbol{\beta}})) = \sigma^2(n-p)\,, \qquad \mathsf{var}(\mathsf{SSE}(\hat{\boldsymbol{\beta}})) = 2\sigma^4(n-p)\,.$$

Daher modifizieren wir den MLE $\hat{\sigma}^2$, damit dieser unverzerrt wird. Erwartungstreuer Schätzer für die Varianz σ^2 unter dem multiplen linearen Regressionsmodell ist somit

$$S^2 = \frac{1}{n-p}\mathsf{SSE}(\hat{\boldsymbol{\beta}})\,.$$

Gerade wurde der minimale SSE($\hat{\boldsymbol{\beta}}$) für normalverteilte Responses diskutiert. Wir werden nun zeigen, dass E(SSE($\hat{\boldsymbol{\beta}}$)) $= \sigma^2(n-p)$ auch ohne expliziter Annahme der Normalverteilung resultiert!

Schreibe

$$\text{SSE}(\hat{\boldsymbol{\beta}}) = \mathbf{r}'\mathbf{r} = \mathbf{y}'(\mathbf{I}-\mathbf{H})'(\mathbf{I}-\mathbf{H})\mathbf{y} = \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}\,.$$

Unter der Annahme $\text{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}$ folgt dafür ganz allgemein

$$\begin{aligned}
\text{E}(\text{SSE}(\hat{\boldsymbol{\beta}})) &= \text{E}(\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}) \\
&= \text{E}\left((\mathbf{y}-\boldsymbol{\mu})'(\mathbf{I}-\mathbf{H})(\mathbf{y}-\boldsymbol{\mu})\right) \\
&\quad + \text{E}\left(\boldsymbol{\mu}'(\mathbf{I}-\mathbf{H})\mathbf{y} + \mathbf{y}'(\mathbf{I}-\mathbf{H})\boldsymbol{\mu} - \boldsymbol{\mu}'(\mathbf{I}-\mathbf{H})\boldsymbol{\mu}\right)
\end{aligned}$$

mit den Skalaren $\mathbf{y}'(\mathbf{I}-\mathbf{H})\boldsymbol{\mu} = \boldsymbol{\mu}'(\mathbf{I}-\mathbf{H})\mathbf{y}$ mit $\text{E}(\mathbf{y}'(\mathbf{I}-\mathbf{H})\boldsymbol{\mu}) = \boldsymbol{\mu}'(\mathbf{I}-\mathbf{H})\boldsymbol{\mu}$.

Wir erhalten somit

$$\mathrm{E}(\mathrm{SSE}(\hat{\boldsymbol{\beta}})) = \mathrm{E}\left((\mathbf{y}-\boldsymbol{\mu})'(\mathbf{I}-\mathbf{H})(\mathbf{y}-\boldsymbol{\mu})\right) + \boldsymbol{\mu}'(\mathbf{I}-\mathbf{H})\boldsymbol{\mu}\,.$$

Nimmt man weiters an, dass $\mathrm{var}(\mathbf{y}) = \sigma^2\mathbf{I}$ hält, dann resultiert das erste Skalar

$$\begin{aligned}
\mathrm{E}\left((\mathbf{y}-\boldsymbol{\mu})'(\mathbf{I}-\mathbf{H})(\mathbf{y}-\boldsymbol{\mu})\right) &= \mathrm{E}\left(\mathrm{trace}\Big((\mathbf{y}-\boldsymbol{\mu})'(\mathbf{I}-\mathbf{H})(\mathbf{y}-\boldsymbol{\mu})\Big)\right) \\
&= \mathrm{trace}\Big(\mathrm{E}\left((\mathbf{I}-\mathbf{H})(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})'\right)\Big) \\
&= \sigma^2(n-p)\,,
\end{aligned}$$

während das zweite Skalar generell verschwindet, denn es ist

$$\begin{aligned}
\boldsymbol{\mu}'(\mathbf{I}-\mathbf{H})\boldsymbol{\mu} &= (\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}-\mathbf{H})(\mathbf{X}\boldsymbol{\beta}) \\
&= \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{0}\,.
\end{aligned}$$

Zusammengefasst gilt daher unter den Momentenannahmen $\mathsf{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} = \boldsymbol{\mu}$ und $\mathsf{var}(\mathbf{y}) = \sigma^2\mathbf{I}$

$$\mathsf{E}(\mathsf{SSE}(\hat{\boldsymbol{\beta}})) = \sigma^2(n-p)\,.$$

Im nächsten Schritt wollen wir nun explizit untersuchen, ob die beiden MLE's $(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2)$ die Cramér-Rao Schranke für die Varianzen erreichen.

Likelihood Terme: Fisherinformation liefert Varianzschranke für die MLE's

$$
\begin{aligned}
\frac{\partial}{\partial \beta_j} \log f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij}(y_i - \mu_i)\,, \qquad j = 0, \ldots, p-1 \\
\frac{\partial}{\partial \sigma^2} \log f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{\sigma^4} \sum_{i=1}^{n} (y_i - \mu_i)^2 \\
-\frac{\partial^2}{\partial \beta_j \partial \beta_k} \log f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij} x_{ik}\,, \qquad j, k = 0, \ldots, p-1 \\
-\frac{\partial^2}{\partial \beta_j \partial \sigma^2} \log f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) &= \frac{1}{\sigma^4} \sum_{i=1}^{n} x_{ij}(y_i - \mu_i)\,, \qquad j = 0, \ldots, p-1 \\
-\frac{\partial^2}{\partial (\sigma^2)^2} \log f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) &= \frac{1}{2}\left(-\frac{n}{\sigma^4} + \frac{2}{\sigma^6} \sum_{i=1}^{n} (y_i - \mu_i)^2 \right)\,.
\end{aligned}
$$

Die Fisherinformation ist nun der Erwartungswert der Matrix dieser negativen 2. Ableitungen

$$\begin{aligned}
\mathsf{E}\left(-\frac{\partial^2}{\partial\beta_j\partial\beta_k}\log f(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\right) &= \frac{1}{\sigma^2}x_j'x_k\,, \qquad j,k=0,\ldots,p-1\\
\mathsf{E}\left(-\frac{\partial^2}{\partial\beta_j\partial\sigma^2}\log f(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\right) &= 0\,, \qquad j=0,\ldots,p-1\\
\mathsf{E}\left(-\frac{\partial^2}{\partial(\sigma^2)^2}\log f(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\right) &= \frac{n}{2}\frac{1}{\sigma^4}\,.
\end{aligned}$$

Die erwartete Informationsmatrix und deren Inverse sind

$$I(\boldsymbol{\beta},\sigma^2)=\begin{pmatrix}\sigma^{-2}\mathbf{X}'\mathbf{X} & \mathbf{0}\\ \mathbf{0}' & n\sigma^{-4}/2\end{pmatrix}\,,\quad I^{-1}(\boldsymbol{\beta},\sigma^2)=\begin{pmatrix}\sigma^2(\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0}\\ \mathbf{0}' & 2\sigma^4/n\end{pmatrix}$$

Die Null in der Nebendiagonalen weist darauf hin, dass $\hat{\boldsymbol{\beta}}$ und $\hat{\sigma}^2$ asymptotisch unkorreliert sind.

$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ stellt die Varianz/Kovarianzmatrix von $\hat{\boldsymbol{\beta}}$ dar, weshalb der für $\boldsymbol{\beta}$ unverzerrte MLE $\hat{\boldsymbol{\beta}}$ diese Varianzschranke erreicht.

Alternativ dazu könnte man auch eine Faktorisierung der Scorefunktion suchen, d.h. man betrachtet

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}} \log f(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) &= \frac{1}{\sigma^2}\mathbf{X}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma^2}\left(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\boldsymbol{\beta}\right) \\
&= \frac{1}{\sigma^2}(\mathbf{X}'\mathbf{X})\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \boldsymbol{\beta}\right) \\
&= \frac{1}{\sigma^2}(\mathbf{X}'\mathbf{X})\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right) .
\end{aligned}$$

Aus $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ folgt wiederum, dass die Varianz des MLE $\hat{\beta}$ die Cramér-Rao Schranke erreicht.

Wie sieht dies für einen Varianzschätzer aus?

Die entsprechende Scorefunktion lässt sich hierfür nun schreiben als

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\log f(\mathbf{y}|\boldsymbol{\beta},\sigma^2) &= -\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2}\frac{1}{\sigma^4}\sum_{i=1}^{n}(y_i-\mu_i)^2 \\ &= \frac{n}{2\sigma^4}\left(\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_i)^2-\sigma^2\right),\end{aligned}$$

wobei $n^{-1}\text{SSE}(\boldsymbol{\beta})$ auch erwartungstreu für σ^2 ist. Also erreicht $n^{-1}\text{SSE}(\boldsymbol{\beta})$, mit $\boldsymbol{\beta}$ fest, die Cramér-Rao Schranke. Dies ist aber kein zulässiger Schätzer!

Gezeigt, dass auch $S^2=(n-p)^{-1}\sum_i(y_i-\hat{\mu}_i)^2$ erwartungstreu für σ^2 ist mit $(n-p)S^2/\sigma^2\sim\chi^2_{n-p}$, also

$$\text{var}(S^2)=2\sigma^4/(n-p)$$

was natürlich größer als $2\sigma^4/n$ ist.

Konfidenz- und Vorhersageintervalle

Ist auch wirklich jedes $\mathbf{x}_j$ in der Designmatrix $\mathbf{X}$ für das Modell relevant? Wir betrachten also Hypothesen der Form

$$H_0 : \beta_j = 0, \qquad j = 1, \ldots, p-1\,.$$

Es ist $\hat{\beta}_j \sim \text{Normal}(\beta_j, \sigma^2 v_{jj})$, mit v_{jj} dem $(j+1)$-ten Diagonalelement von $(\mathbf{X}'\mathbf{X})^{-1}$. Somit

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 v_{jj}}} \sim \text{Normal}(0, 1)\,.$$

Weiters sind $\hat{\boldsymbol{\beta}}$ und $\text{SSE}(\hat{\boldsymbol{\beta}})$ unter Normalverteilungsannahme unabhängig, und es gilt $(n-p)S^2/\sigma^2 \sim \chi^2_{n-p}$.

Somit definiert man

$$T = \frac{\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 v_{jj}}}}{\sqrt{\frac{n-p}{\sigma^2} S^2/(n-p)}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{S^2 v_{jj}}} \sim t_{n-p}\,.$$

Möchte man die Relevanz von $\mathbf{x}_j$ testen, so verwendet man dazu

$$T = \frac{\hat{\beta}_j}{\sqrt{S^2 v_{jj}}}$$

und verwirft H_0, falls $|T| > t_{n-p,1-\alpha/2}$.

Diese Überlegungen liefern auch ein zweiseitiges Konfidenzintervall für β_j:

$$KIV(\beta_j) = \hat{\beta}_j \pm S\; v_{jj}^{1/2} t_{n-p,1-\alpha/2}\,.$$

Bezeichnen wir mit $\mathbf{x}_+ = (1, x_{+1}, \ldots, x_{+p-1})'$ einen neuen Vektor und interessieren uns dort für eine Schätzung des Erwartungswertes $\mu_+ = \mathbf{x}'_+\boldsymbol{\beta}$. Für den MLE $\hat{\mu}_+ = \mathbf{x}'_+\hat{\boldsymbol{\beta}}$ gilt

$$\begin{aligned} \mathsf{E}(\hat{\mu}_+) &= \mathbf{x}'_+\boldsymbol{\beta} \\ \mathsf{var}(\hat{\mu}_+) &= \mathbf{x}'_+\mathsf{var}(\hat{\boldsymbol{\beta}})\mathbf{x}_+ = \sigma^2\mathbf{x}'_+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_+ \,. \end{aligned}$$

$\hat{\mu}_+$ ist eine Linearkombination von $\hat{\boldsymbol{\beta}} \sim$ Normal und somit auch normalverteilt. S^2, $\hat{\boldsymbol{\beta}}$ sind unabhängig womit folgt

$$\frac{\mathbf{x}'_+\hat{\boldsymbol{\beta}} - \mathbf{x}'_+\boldsymbol{\beta}}{\sqrt{S^2\mathbf{x}'_+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_+}} \sim t_{n-p} \,.$$

Vergleiche dies mit der Varianz eines *vorhandenen* $\hat{\mu}_i = x'_i\hat{\beta}$, also mit

$$\mathsf{var}(\hat{\mu}_i) = \sigma^2\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i = \sigma^2 h_{ii}.$$

Vorhersage-/Prädiktionsintervall für neue Beobachtung $y_+ = \mathbf{x}'_+\boldsymbol{\beta} + \epsilon_+$, mit $\epsilon_+ \sim \text{Normal}(0, \sigma^2)$. Ein Vorhersageintervall überdeckt mit Wahrscheinlichkeit $(1-\alpha)$ das y_+.

Wir betrachtenn $y_+ - \hat{\text{E}}(y_+) = y_+ - \mathbf{x}'_+\hat{\boldsymbol{\beta}}$, wofür gilt

$$\begin{aligned} \text{E}(y_+ - \mathbf{x}'_+\hat{\boldsymbol{\beta}}) &= \mathbf{x}'_+\boldsymbol{\beta} - \mathbf{x}'_+\boldsymbol{\beta} = 0 \\ \text{var}(y_+ - \mathbf{x}'_+\hat{\boldsymbol{\beta}}) &= \sigma^2(1 + \mathbf{x}'_+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_+)\,. \end{aligned}$$

Sämtliche Inferenz bzgl. der neue Response y_+ basiert auf der Statistik

$$\frac{y_+ - \mathbf{x}'_+\hat{\boldsymbol{\beta}}}{\sqrt{S^2(1 + \mathbf{x}'_+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_+)}} \sim t_{n-p}\,.$$

Beispiel: 2 Gruppen normalverteilter Responses mit

$$\begin{aligned} y_{0i} &= \beta_0 + \epsilon_{0i}\,, \qquad i = 1, \ldots, n_0 \\ y_{1i} &= \beta_0 + \beta_1 + \epsilon_{1i}\,, \qquad i = 1, \ldots, n_1\,. \end{aligned}$$

Die Fehlerterme ϵ_{gi}, $g \in \{0, 1\}$, seien alle unabhängig mit Varianz σ^2.

Die Matrixform dieses Modells lautet

$$\begin{pmatrix} y_{01} \\ \vdots \\ y_{0n_0} \\ y_{11} \\ \vdots \\ y_{1n_1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \epsilon_{01} \\ \vdots \\ \epsilon_{0n_0} \\ \epsilon_{11} \\ \vdots \\ \epsilon_{1n_1} \end{pmatrix}\,.$$

Mit $\bar{y}_g = n_g^{-1} \sum_i y_{gi}$ ergibt sich als MLE

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} n_0 + n_1 & n_1 \\ n_1 & n_1 \end{pmatrix}^{-1} \begin{pmatrix} n_0 \bar{y}_0 + n_1 \bar{y}_1 \\ n_1 \bar{y}_1 \end{pmatrix}.$$

Nun gilt

$$\begin{aligned} \begin{pmatrix} n_0 + n_1 & n_1 \\ n_1 & n_1 \end{pmatrix}^{-1} &= \frac{1}{(n_0 + n_1)n_1 - n_1^2} \begin{pmatrix} n_1 & -n_1 \\ -n_1 & n_0 + n_1 \end{pmatrix} \\ &= \frac{1}{n_0 n_1} \begin{pmatrix} n_1 & -n_1 \\ -n_1 & n_0 + n_1 \end{pmatrix} = \begin{pmatrix} n_0^{-1} & -n_0^{-1} \\ -n_0^{-1} & n_0^{-1} + n_1^{-1} \end{pmatrix}, \end{aligned}$$

womit folgt

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} n_0^{-1} & -n_0^{-1} \\ -n_0^{-1} & n_0^{-1} + n_1^{-1} \end{pmatrix} \begin{pmatrix} n_0 \bar{y}_0 + n_1 \bar{y}_1 \\ n_1 \bar{y}_1 \end{pmatrix} = \begin{pmatrix} \overline{y}_0 \\ \overline{y}_1 - \bar{y}_0 \end{pmatrix}.$$

$\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ beinhaltet die Varianzen und die Kovarianz der MLE's $\hat{\beta}_0$ und $\hat{\beta}_1$, also $\mathsf{var}(\hat{\beta}_0) = \mathsf{var}(\bar{y}_0) = \sigma^2/n_0$ oder $\mathsf{var}(\hat{\beta}_1) = \mathsf{var}(\bar{y}_1 - \bar{y}_0) = \sigma^2(1/n_1 + 1/n_0)$.

Unter diesem Modell erhält man $\widehat{\mathsf{E}}(y_{0i}) = \hat{\beta}_0 = \bar{y}_0$ sowie $\widehat{\mathsf{E}}(y_{1i}) = \hat{\beta}_0 + \hat{\beta}_1 = \bar{y}_1$.

$H_0 : \mathsf{E}(y_{0i}) = \mathsf{E}(y_{1i})$ ist somit äquivalent zu $H_0 : \beta_1 = 0$, mit der Teststatistik

$$T = \frac{\hat{\beta}_1 - 0}{\sqrt{\widehat{\mathsf{var}}(\hat{\beta}_1)}} = \frac{\bar{y}_1 - \overline{y}_0}{\sqrt{S^2(1/n_0 + 1/n_1)}} \sim t_{n-2}\,.$$

Dies ist die bekannte t-Test Statistik, falls die Varianzen in beiden Gruppen gleich sind.

Multiples Bestimmtheitsmass

Auch für ein MLR ist das Bestimmtheitsmass definiert als

$$R^2 = \frac{\mathsf{SSR}(\hat{\boldsymbol{\beta}})}{\mathsf{SST}} = 1 - \frac{\mathsf{SSE}(\hat{\boldsymbol{\beta}})}{\mathsf{SST}}\,, \qquad 0 \leq R^2 \leq 1\,.$$

Extreme Situationen liefern:
- $R^2 = 1$: perfekte Anpassung, d.h. $y_i = \hat{\mu}_i \Rightarrow \mathsf{SSE}(\hat{\boldsymbol{\beta}}) = 0$
- $R^2 = 0$: keine lineare Abhängigkeit von $\mathbf{x}_1, \ldots, \mathbf{x}_{p-1} \Rightarrow \beta_1 = \cdots = \beta_{p-1} = 0 \Rightarrow \hat{\mu}_i = \bar{y} \Rightarrow \mathsf{SSR}(\hat{\boldsymbol{\beta}}) = 0$.

Nachteil: für $p \to \infty$ geht $R^2 \to 1$. Aus diesem Grund wird es adjustiert

$$R^2_{adj} = 1 - \frac{\mathsf{SSE}(\hat{\boldsymbol{\beta}})/(n-p)}{\mathsf{SST}/(n-1)}\,.$$

Kann leider bei sehr schlecht passenden Modellen auch negative Werte haben.

Varianzanalyse (ANalysis Of VAriance – ANOVA)

Zur Erinnerung: wir hatten bereits im SLR

$$\begin{aligned} \text{SST} &= \text{SSR}(\hat{\boldsymbol{\beta}}) + \text{SSE}(\hat{\boldsymbol{\beta}}) \\ \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 \end{aligned}$$

Für ein MLR ändern sich nur die Freiheitsgrade (df):

- SST hat noch immer $n - 1$ df
- SSE hat $n - p$ df
- SSR hat deshalb $p - 1$ df, da p Parameter in $\hat{\mu}_i$

ANOVA Tabelle für ein MLR:

Ursache der Variation	Quadratsumme (SS)	df	Mittlere SS
Regression	SSR $= \sum_i (\hat{\mu}_i - \bar{y})^2$	$p-1$	SSR$(\hat{\boldsymbol{\beta}})/(p-1)$
Fehler	SSE $= \sum_i (y_i - \hat{\mu}_i)^2$	$n-p$	SSE$(\hat{\boldsymbol{\beta}})/(n-p)$
Total	SST $= \sum_i (y_i - \bar{y})^2$	$n-1$	

Overall F-Test auf sämtliche Prädiktoren im Regressionsmodell

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$

H_1 : nicht alle β_j $(j = 1, \ldots, p-1)$ sind gleich Null.

H_0 behauptet, dass alle Prädiktoren $\mathbf{x}_1, \ldots, \mathbf{x}_{p-1}$ nutzlos sind (keine Beziehung zwischen der Response und dem Satz der Prädiktoren besteht), während wir unter H_1 glauben, dass darunter zumindest ein Prädiktor nützlich ist.

Teststatistik

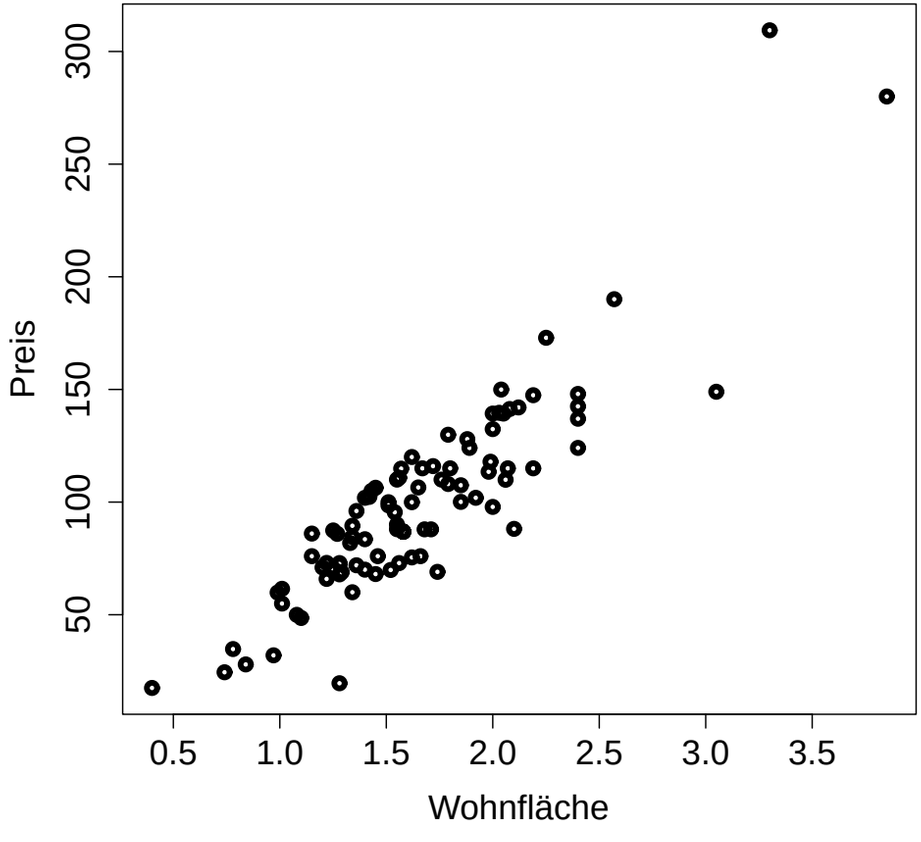
$$F = \frac{\mathsf{MSR}(\hat{\boldsymbol{\beta}})}{\mathsf{MSE}(\hat{\boldsymbol{\beta}})}$$

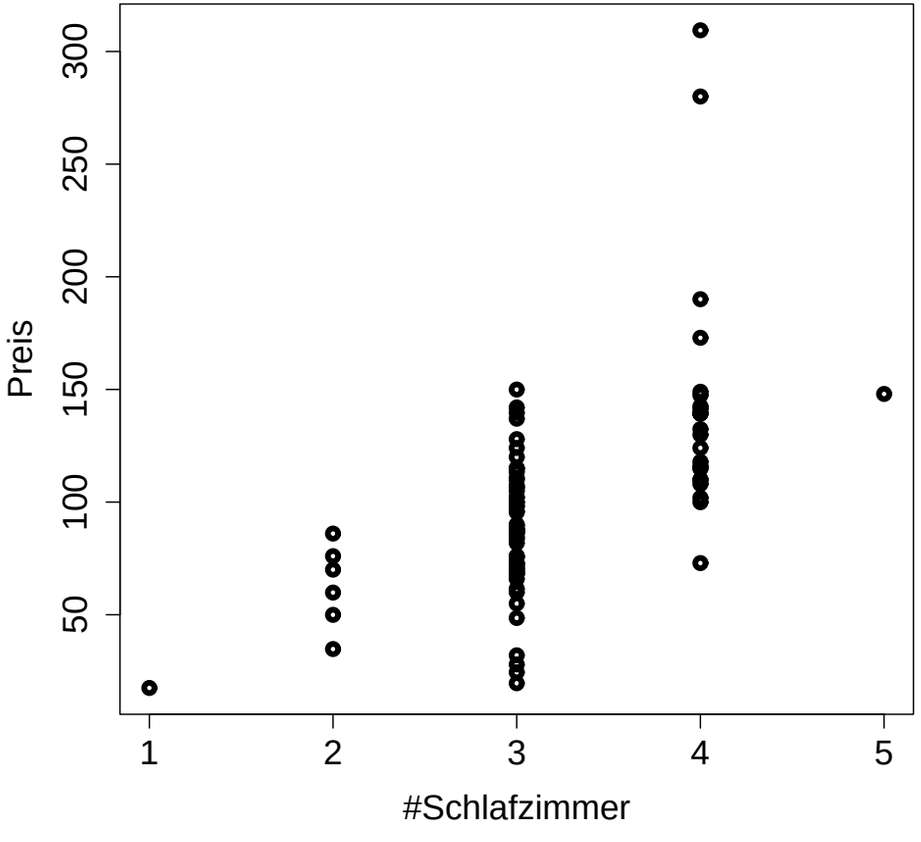
Verwerfungsregel: Verwirf H_0, falls $F > F_{p-1,n-p;1-\alpha}$ gilt.

Bemerke: Falls $p - 1 = 1$, so ist dies der F-Test für $H_0 : \beta_1 = 0$ im SLR.

Beispiel: Hauspreise mit R

```
> houses <- read.table("houses.dat",col.names=c("price","area","bed","bath","new"))
> attach(houses)
> plot(area, price); plot(bed, price)
```

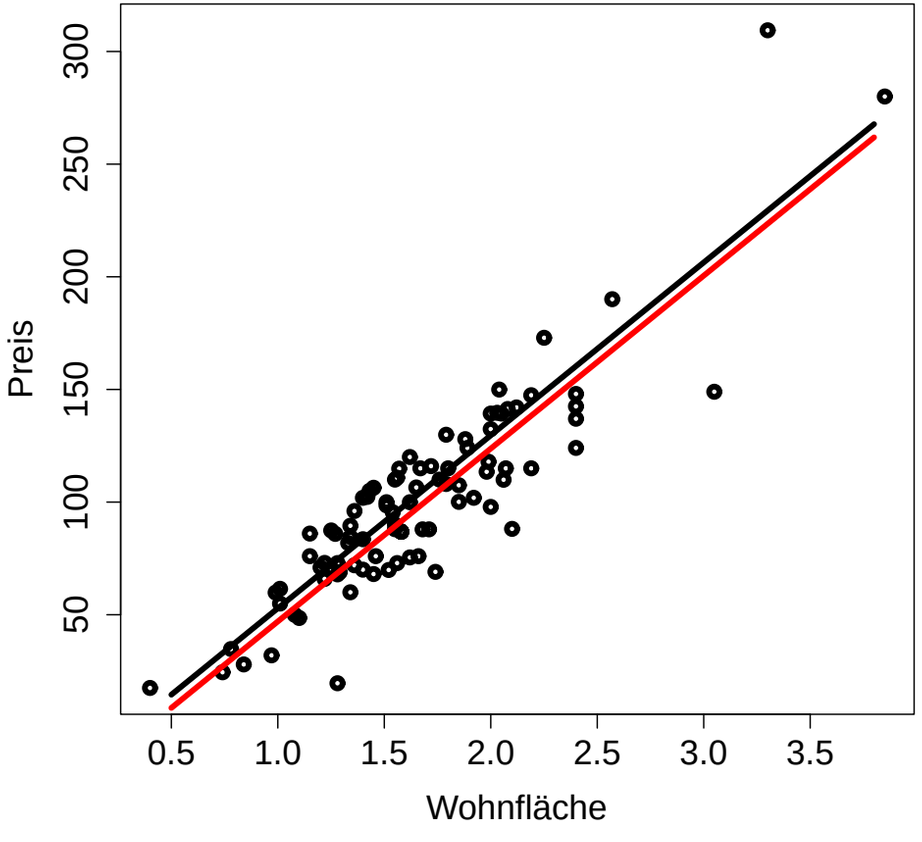

Preis
50
100
150
200
250
300
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Wohnfläche

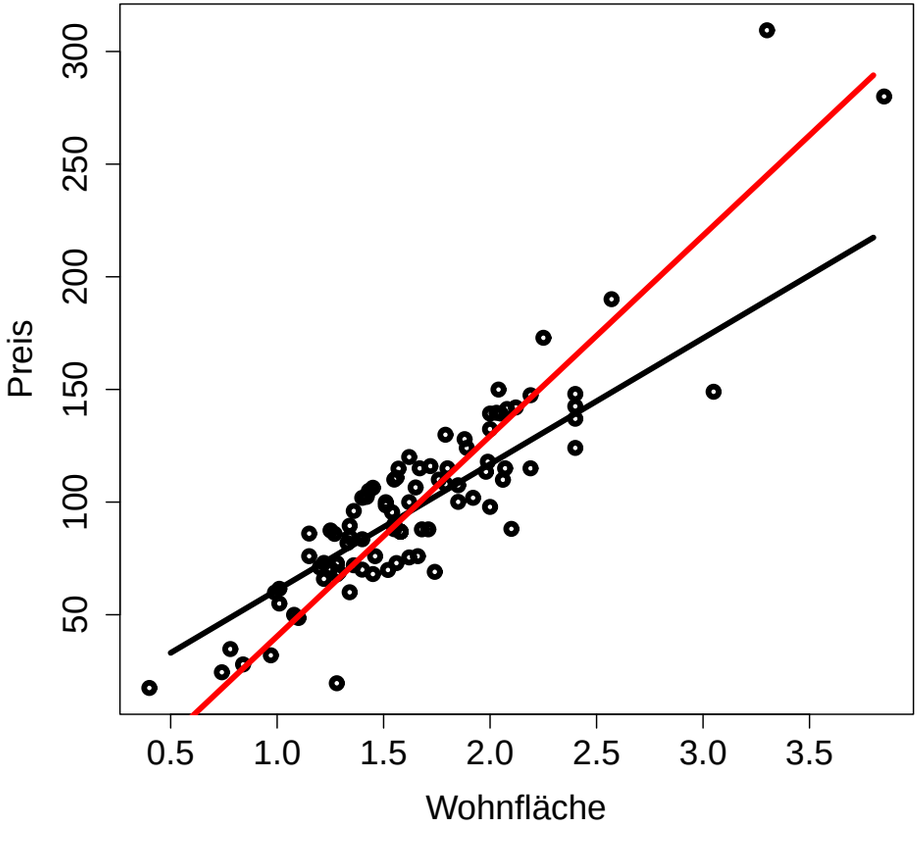

Preis
50
100
150
200
250
300
1
2
3
4
5
#Schlafzimmer

```
> (model <- lm(price ~ area + bed))

Call:
lm(formula = price ~ area + bed)

Coefficients:
(Intercept)         area          bed
    -22.393       76.742       -1.468

> attributes(model)
$names
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"

$class
[1] "lm"
```

```
> summary(model)
Call:
lm(formula = price ~ area + bed)

Residuals:
    Min      1Q  Median      3Q     Max
-56.797 -11.751   2.859  10.817  84.417

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -22.393     10.937  -2.048   0.0435 *
area          76.742      5.227  14.682   <2e-16 ***
bed           -1.468      4.523  -0.325   0.7462
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.57 on 90 degrees of freedom
Multiple R-squared: 0.8081,     Adjusted R-squared: 0.8038
F-statistic: 189.5 on 2 and 90 DF,  p-value: < 2.2e-16
```

```
> summary(model.i <- lm(price ~ 1+area+bed+area:bed)) # in short price~area*bed
Call:
lm(formula = price ~ area + bed + area * bed)
Residuals:
   Min     1Q Median     3Q    Max
-61.94 -11.51   1.92  12.27  78.29

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   18.549     26.577   0.698  0.48704
area          47.595     18.037   2.639  0.00982 **
bed          -13.416      8.379  -1.601  0.11292
area:bed       8.270      4.903   1.687  0.09515 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.37 on 89 degrees of freedom
Multiple R-squared: 0.814,      Adjusted R-squared: 0.8078
F-statistic: 129.9 on 3 and 89 DF,  p-value: < 2.2e-16
```


Preis
50
100
150
200
250
300
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Wohnfläche


Preis
50
100
150
200
250
300
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Wohnfläche

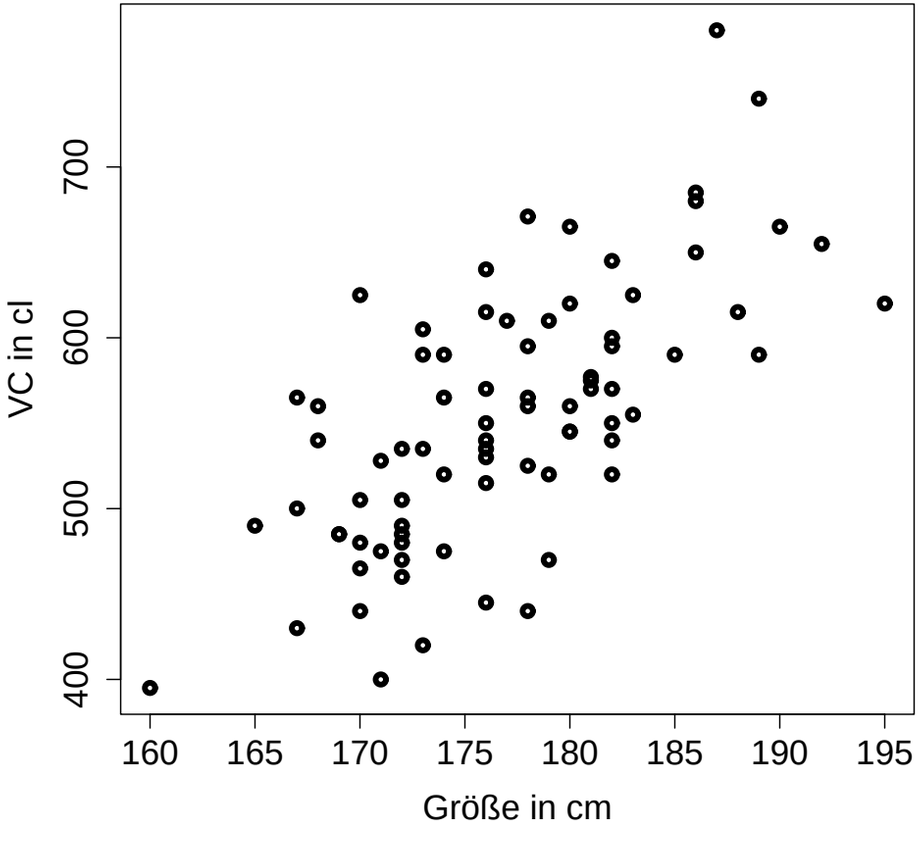
```
> # left plot (main effects model)
> plot(area, price, xlab="Wohnfläche", ylab="Preis", lwd=4)
> (newdat <- expand.grid(area=c(0.5,3.8), bed=c(1,5)))
  area bed
1  0.5   1
2  3.8   1
3  0.5   5
4  3.8   5
> p <- predict(model, newdat)
> lines(c(0.5, 3.8), p[1:2], lwd=4, col=1) # black = 1 bed
> lines(c(0.5, 3.8), p[3:4], lwd=4, col=2) # red = 5 beds

> # right plot (model with interaction term)
> plot(area, price, xlab="Wohnfläche", ylab="Preis", lwd=4)
> p.i <- predict(model.i, newdat)
> lines(c(0.5, 3.8), p.i[1:2], lwd=4, col=1) # black = 1 bed
> lines(c(0.5, 3.8), p.i[3:4], lwd=4, col=2) # red = 5 beds
```

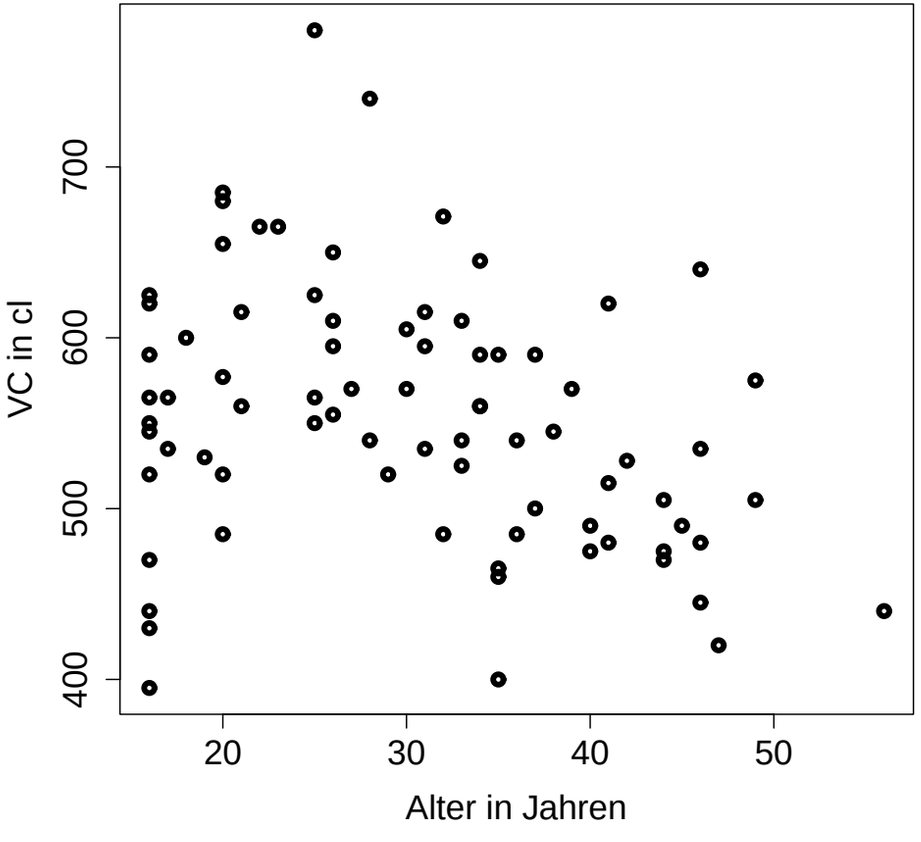
```
> anova(model.i)
Analysis of Variance Table

Response: price
          Df Sum Sq Mean Sq  F value  Pr(>F)
area       1 145097  145097 386.6340 < 2e-16 ***
bed        1     40      40   0.1076 0.74371
area:bed   1   1068    1068   2.8453 0.09515 .
Residuals 89  33400     375
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Beispiel: Vitalkapazitäten mit R

Von $n = 79$ männlichen Nichtrauchern aus Murau und dem Aichfeld liegen Messungen ihrer Vitalkapazität in Zentiliter vor. Gesucht ist ein Regressionsmodel, das den Zusammenhang zwischen der Vitalkapazität und der Körpergröße und/oder des Alters beschreibt.

```
> aimu <- read.table("aimu.dat", header=T)
> names(aimu) <- c("nr", "year", "age", "height", "weight", "vc", ..., "region")
> attach(aimu)

> plot(height, vc, xlab="Größe in cm", ylab="VC in cl", lwd=4)
> plot(age, vc, xlab="Alter in Jahren", ylab="VC in cl", lwd=4)
```


VC in cl
400
500
600
700
160
165
170
175
180
185
190
195
Größe in cm


VC in cl
400
500
600
700
20
30
40
50
Alter in Jahren

Einfache Lineare Modelle:

```
> summary(lm(vc ~ height))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -812.3122   166.5833  -4.876 5.69e-06 ***
height         7.7197     0.9409   8.205 4.10e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 56.07 on 77 degrees of freedom
Multiple R-squared: 0.4665,     Adjusted R-squared: 0.4595
F-statistic: 67.32 on 1 and 77 DF,  p-value: 4.099e-12
```

Falls `vc` durch `height` erklärt wird, ist dies signifikant. Ein cm mehr an Größe bedeutet im Mittel 0.077 Liter mehr an `vc`.

```
> summary(lm(vc ~ age))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 617.7413    25.4151  24.306  < 2e-16 ***
age          -2.1219     0.7938  -2.673  0.00917 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 73.43 on 77 degrees of freedom
Multiple R-squared: 0.08492,    Adjusted R-squared: 0.07303
F-statistic: 7.146 on 1 and 77 DF,  p-value: 0.009171
```

Ein zusätzliches Jahr lässt vc im Mittel um 0.021 Liter abnehmen.

Kombiniert man beide einfachen Modelle, so resultiert

```
> summary(lm(vc ~ height + age))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -717.3919   175.2551  -4.093 0.000105 ***
height         7.3525     0.9593   7.664 4.82e-11 ***
age           -0.9892     0.6180  -1.601 0.113584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55.51 on 76 degrees of freedom
Multiple R-squared: 0.4839,     Adjusted R-squared: 0.4703
F-statistic: 35.62 on 2 and 76 DF,  p-value: 1.216e-11
```

Das zusätzliche Alter zum Modell, das bereits die Größe inkludiert, liefert keine Signifikanz (p-Wert 11%). Jedoch ist bekannt, dass kein linearer Zusammenhang zwischen vc und age besteht sondern eher ein quadratischer oder kubischer.

```
> summary(lm(vc ~ height + age + I(age^2) + I(age^3)))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.106e+03  2.055e+02  -5.381 8.33e-07 ***
height       6.660e+00  9.106e-01   7.314 2.55e-10 ***
age          4.888e+01  1.575e+01   3.105  0.00270 **
I(age^2)    -1.462e+00  4.979e-01  -2.935  0.00443 **
I(age^3)     1.318e-02  4.945e-03   2.665  0.00944 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 51.59 on 74 degrees of freedom
Multiple R-squared: 0.566,      Adjusted R-squared: 0.5426
F-statistic: 24.13 on 4 and 74 DF,  p-value: 8.468e-13
```

Einfachere Spezifikation durch `lm(vc~height + poly(age,degree=3,raw=TRUE))`

Wie sieht die Designmatrix zu diesem komplexen Modell aus?

Diese erhält man sehr einfach durch

```
> (X <- lm(vc ~ height + age + I(age^2) + I(age^3), x=TRUE)$x)
   (Intercept) height age I(age^2) I(age^3)
1            1    171  42     1764    74088
2            1    178  32     1024    32768
3            1    176  46     2116    97336
:            :      :   :        :        :
78           1    176  41     1681    68921
79           1    180  41     1681    68921
attr(,"assign")
[1] 0 1 2 3 4
```

Damit kann man jetzt beispielsweise direkt die Standardfehler von $\hat{\boldsymbol{\beta}}$ berechnen. Die Wurzel der Diagonalelemente in $S^2(\mathbf{X}'\mathbf{X})^{-1}$ erhält man mittels

```
> s <- summary(mod <- lm(vc ~ height + age + I(age^2) + I(age^3)))$sigma
> sqrt(diag(solve(t(X) %*% X))) * s
  (Intercept)       height          age     I(age^2)     I(age^3)
 2.054779e+02 9.106415e-01 1.574705e+01 4.979678e-01 4.945972e-03
```

Wir entscheiden uns für dieses MLR. Definiere eine Person x_+ mit Alter 23 Jahre und einer Größe von 184 cm. Welchen vorhergesagten mittleren VC-Wert hat diese Person?

```
> h <- 184; a <- 23; x.plus <- matrix(c(1, h, a, a^2, a^3)) # column vector
> (hatmu.plus <- as.numeric(t(x.plus) %*%  matrix(coef(mod))))
[1] 631.1984
```

Berechne auch ein 95% Konfidenzintervall für den Parameter μ_+:

```
> alpha <- 0.05; df <- summary(mod)$df[2]
> var.hatmu.plus <- s^2*as.numeric(t(x.plus) %*% solve(t(X) %*% X) %*% x.plus)
> c(hatmu.plus - sqrt(var.hatmu.plus)* qt(1-alpha/2, df),
+   hatmu.plus + sqrt(var.hatmu.plus)* qt(1-alpha/2, df))
[1] 610.0327 652.3641
```

Viel einfacher mittels:

```
> predict(mod, new=data.frame(age=a, height=h), interval="confidence")
       fit      lwr      upr
1 631.1984 610.0327 652.3641
```

Vorhersageintervall für eine neue Responsebeobachtung in x_+

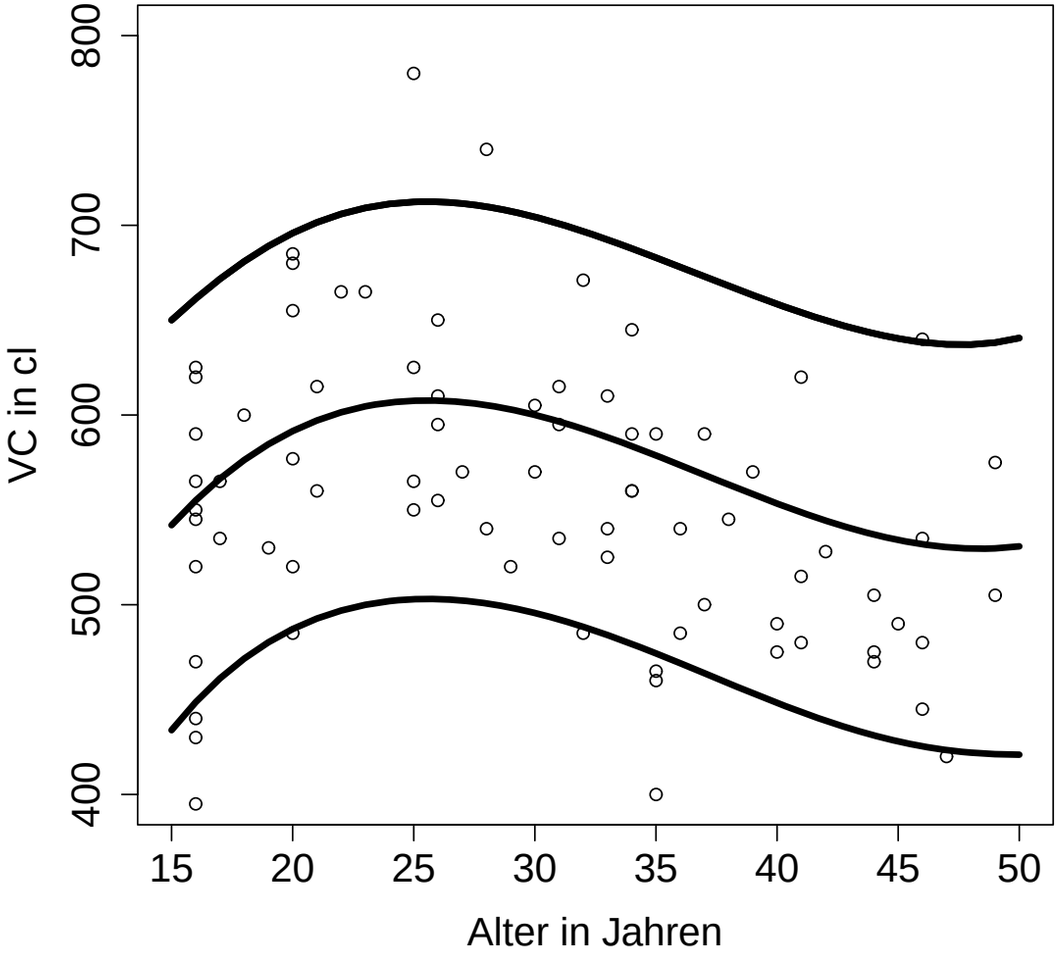
```
> var.haty.plus <- s^2*as.numeric((1 + t(x.plus) %*% solve(t(X) %*% X) %*% x.plus))
> c(hatmu.plus - sqrt(var.haty.plus)* qt(1-alpha/2, df),
    hatmu.plus + sqrt(var.haty.plus)* qt(1-alpha/2, df))
[1] 526.2563 736.1405
```

oder wiederum viel einfacher durch

```
> predict(mod, new=data.frame(age=a, height=h), interval="predict")
       fit      lwr      upr
1 631.1984 526.2563 736.1405
```

Zeichne dieses Modell mit punktweisen 95% Vorhersageintervallen für Personen mit `height`=180.

```
> a <- 15:50; h <- rep(180, length(a))
> p <- predict(mod, new=data.frame(age=a, height=h), interval="predict")
> plot (a, p[ , "fit"], ..., ylim=c(400,800), type="l", lwd=4)
> lines(a, p[ , "lwr"], lwd=4); lines(a, p[ , "upr"], lwd=4)
> points(age,vc)
```


VC in cl
800
700
600
500
400
15
20
25
30
35
40
45
50
Alter in Jahren

ANOVA – Einige offengebliebene Fragen

Geometrische Interpretation der Schätzer

Der LSE $\hat{\boldsymbol{\beta}}$ minimiert die Distanz $\text{SSE}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Diese ist minimal, falls $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ orthogonal zu dem von den Spalten in $\mathbf{X}$ aufgespannten Raum ist. Damit folgt für jede Spalte $\mathbf{x}$ in $\mathbf{X}$

$$\mathbf{x}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}\,,$$

und die Scoregleichungen $\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$ halten.

Der Punkt $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$ ist die orthogonale Projektion von $\mathbf{y}$ auf die durch die Spalten von $\mathbf{X}$ aufgespannte Ebene. Bemerke, $\hat{\boldsymbol{\mu}}$ ist eindeutig unabhängig davon, ob $\mathbf{X}'\mathbf{X}$ invertierbar ist.

Residuen $\mathbf{r} = \mathbf{y} - \hat{\boldsymbol{\mu}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$ und Fitted Values $\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{y}$ sind orthogonale Vektoren. Um dies algebraisch zu erkennen, prüfen wir

$$\hat{\boldsymbol{\mu}}'\mathbf{r} = \mathbf{y}'\mathbf{H}'(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}'(\mathbf{H} - \mathbf{H})\mathbf{y} = 0\,.$$

Wichtig: Beziehung zwischen Orthogonalität und Unabhängigkeit normalverteilter Vektoren.

$u = \mathbf{a}'\mathbf{y}$ und $v = \mathbf{b}'\mathbf{y}$, $\mathbf{y} \sim$ Normal$(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, sind unabhängig, falls cov$(u, v) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{b} = 0$.

$\mathbf{u} = \mathbf{A}'\mathbf{y}$ und $\mathbf{v} = \mathbf{B}'\mathbf{y}$ daher unabhängig, falls cov$(\mathbf{u}, \mathbf{v}) = \mathbf{A}'\boldsymbol{\Sigma}\mathbf{B} = \mathbf{0}$.

Im Regressionsfall ist $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. Die Unabhängigkeit zweier Vektoren $\mathbf{u}, \mathbf{v}$ (lineare Formen in $\mathbf{y}$) ist somit gegeben, wenn cov$(\mathbf{u}, \mathbf{v}) = \sigma^2\mathbf{A}'\mathbf{I}\mathbf{B} = \mathbf{0}$. Dies ist genau dann erfüllt, wenn $\mathbf{u}, \mathbf{v}$ orthogonal sind. *Orthogonalität* ist somit bei Normalverteilung äquivalent zur *Unabhängigkeit*!

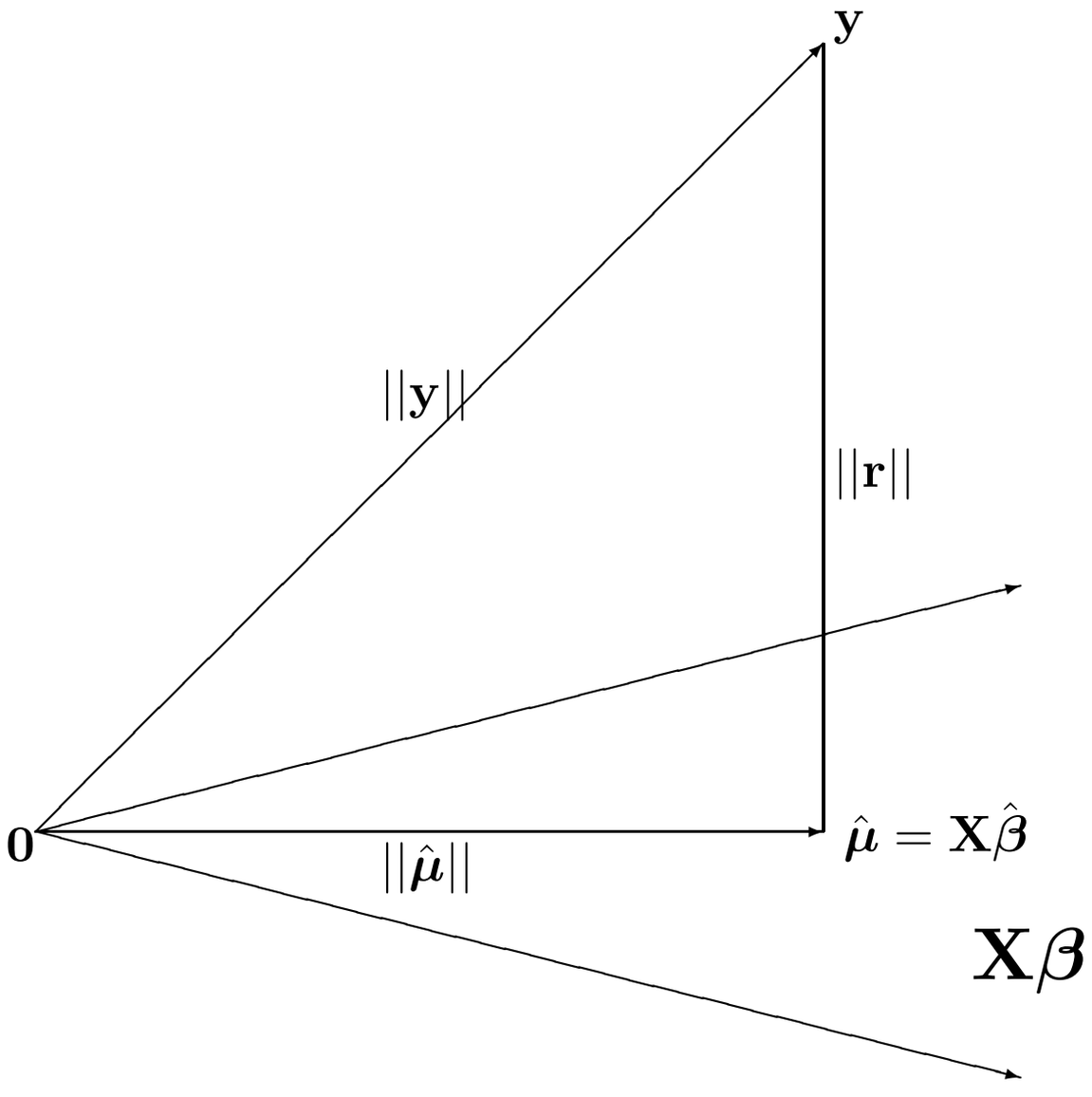
So gilt für die Overall Sum of Squares:

$$\mathbf{y}'\mathbf{y} = (\mathbf{y} - \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}})'(\mathbf{y} - \hat{\boldsymbol{\mu}} + \hat{\boldsymbol{\mu}}) = (\mathbf{r} + \hat{\boldsymbol{\mu}})'(\mathbf{r} + \hat{\boldsymbol{\mu}}) = \mathbf{r}'\mathbf{r} + \hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}} + 2\mathbf{r}'\hat{\boldsymbol{\mu}}\,.$$

Wie zuvor gezeigt sind $\mathbf{r}$ und $\hat{\boldsymbol{\mu}}$ orthogonal, also gilt $\mathbf{r}'\hat{\boldsymbol{\mu}} = 0$. Wir können daher schreiben

$$\mathbf{y}'\mathbf{y} = \mathbf{r}'\mathbf{r} + \hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\mu}}\,.$$

Dies resultiert auch unmittelbar aus dem **Satz von Pythagoras**.


y
||y||
||r||
0
$\hat{\mu} = X\hat{\beta}$
$||\hat{\mu}||$
$X\beta$

F-Statistiken

Schlüsselfrage in der Regression: Beeinflussen erklärende Variablen die Response?

Angenommen, das Modell lautet

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = (\mathbf{X}_1, \mathbf{X}_2) \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}\,.$$

$\mathbf{X}_1$ ist eine $n \times q$ Matrix, $\mathbf{X}_2$ eine $n \times (p-q)$ Matrix ($q < p$), und $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ sind Parametervektoren. Alle Variablen in $\mathbf{X}_2$ sind dann nicht notwendig, wenn $\boldsymbol{\beta}_2 = \mathbf{0}$. In diesem Fall hält das einfachere Modell $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$.

Wie erkennt man dies ?

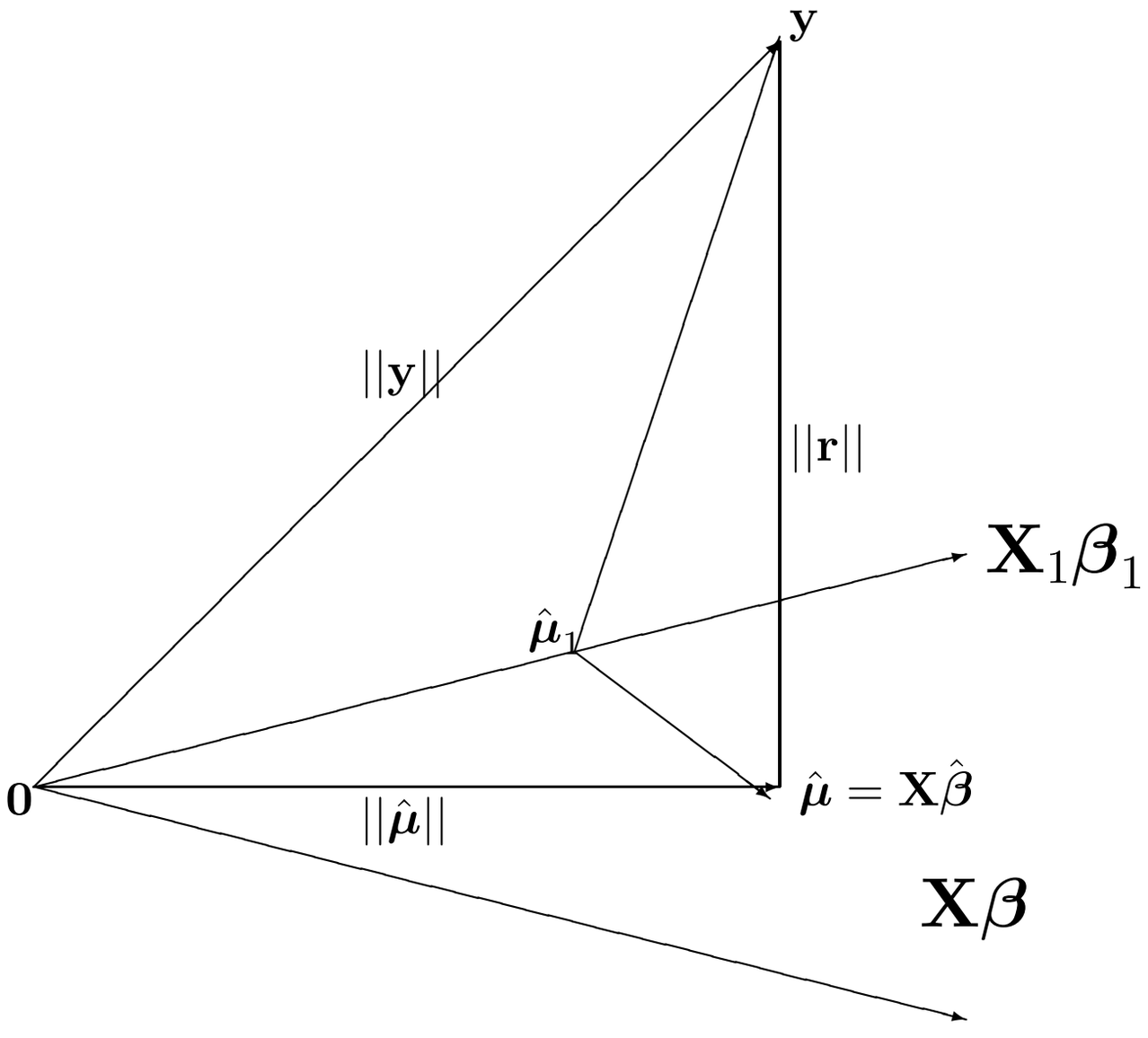

$\mathbf{y}$
$||\mathbf{y}||$
$||\mathbf{r}||$
$\mathbf{X}_1\boldsymbol{\beta}_1$
$\hat{\boldsymbol{\mu}}_1$
$\mathbf{0}$
$\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$
$||\hat{\boldsymbol{\mu}}||$
$\mathbf{X}\boldsymbol{\beta}$

Abbildung: $\hat{\boldsymbol{\mu}}_1 = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$ ist orthogonale Projektion von $\mathbf{y}$ auf $\mathbf{X}_1$. Die Residuen $\mathbf{y} - \hat{\boldsymbol{\mu}}_1 = (\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1')\mathbf{y}$ zerfallen in 2 orthogonale Vektoren

$$\mathbf{y} - \hat{\boldsymbol{\mu}}_1 = (\mathbf{y} - \hat{\boldsymbol{\mu}}) + (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_1)\,,$$

mit $(\mathbf{y} - \hat{\boldsymbol{\mu}})'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_1) = 0$.

Diese Residuen sind darstellbar durch die Residuen vom komplexeren Modell $\mathbf{y} - \hat{\boldsymbol{\mu}}$, und den Änderungen in den geschätzten Werten, falls $\mathbf{X}_2$ in die Designmatrix hinzugenommen wird, $\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_1$.

Da die beiden Vektoren $\mathbf{y} - \hat{\boldsymbol{\mu}}$ und $\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_1$ orthogonale lineare Funktionen der normalverteilten Responses $\mathbf{y}$ sind, sind sie auch unabhängig. Der Satz von Pythagoras impliziert

$$(\mathbf{y} - \hat{\boldsymbol{\mu}}_1)'(\mathbf{y} - \hat{\boldsymbol{\mu}}_1) = (\mathbf{y} - \hat{\boldsymbol{\mu}})'(\mathbf{y} - \hat{\boldsymbol{\mu}}) + (\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_1)'(\hat{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_1)\,,$$

oder äquivalent

$$\mathsf{SSE}(\hat{\boldsymbol{\beta}}_1) = \mathsf{SSE}(\hat{\boldsymbol{\beta}}) + \Big(\mathsf{SSE}(\hat{\boldsymbol{\beta}}_1) - \mathsf{SSE}(\hat{\boldsymbol{\beta}})\Big)\,.$$

Quadratsumme des einfacheren Modells als Summe zweier **unabhängiger** Teile: Quadratsumme des komplexeren Modells, $\mathsf{SSE}(\hat{\boldsymbol{\beta}})$, und die Reduktion in der Quadratsumme wenn $\mathbf{X}_2$ ins Modell genommen wird, $\mathsf{SSE}(\hat{\boldsymbol{\beta}}_1) - \mathsf{SSE}(\hat{\boldsymbol{\beta}})$.

Was kann verteilungstheoretisch ausgesagt werden?

Unter $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ ist das einfache Modell korrekt, aber auch das komplexere (setze darin $\boldsymbol{\beta}_2 = \mathbf{0}$). Es gilt somit $\mathsf{SSE}(\hat{\boldsymbol{\beta}}_1) \sim \sigma^2 \chi^2_{n-q}$, $\mathsf{SSE}(\hat{\boldsymbol{\beta}}) \sim \sigma^2 \chi^2_{n-p}$. Gerade haben wir gezeigt, dass die Differenz $\mathsf{SSE}(\hat{\boldsymbol{\beta}}_1) - \mathsf{SSE}(\hat{\boldsymbol{\beta}})$ unabhängig von $\mathsf{SSE}(\hat{\boldsymbol{\beta}})$ ist.

Falls $\boldsymbol{\beta}_2 = \mathbf{0}$, dann ist $\mathsf{SSE}(\hat{\boldsymbol{\beta}}_1) - \mathsf{SSE}(\hat{\boldsymbol{\beta}}) \sim \sigma^2 \chi^2_{p-q}$ (z.z. wie zuvor bei der Diskussion von S^2 mittels Momentenerzeugender Funktion). Somit gilt unter $\boldsymbol{\beta}_2 = \mathbf{0}$

$$F = \frac{\big(\mathsf{SSE}(\hat{\boldsymbol{\beta}}_1) - \mathsf{SSE}(\hat{\boldsymbol{\beta}})\big)/(p-q)}{\mathsf{SSE}(\hat{\boldsymbol{\beta}})/(n-p)} \sim F_{p-q,n-p}\,.$$

Für $\boldsymbol{\beta}_2 \neq \mathbf{0}$ wird die mittlere Reduktion in der Quadratsumme größer sein als unter $\boldsymbol{\beta}_2 = \mathbf{0}$ erwartet. F wird daher im Vergleich zur $F_{p-q,n-p}$-Verteilung groß sein. Wir testen $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$ mittels F und verwerfen H_0, falls $F > F_{p-q,n-p;1-\alpha}$.

Zwei Spezialfälle:

• Besteht $\mathbf{X}_2$ nur aus einer erklärenden Variablen $\mathbf{x}_2$, dann ist β_2 ein Skalar. Wir betrachten $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{x}_2\beta_2 + \boldsymbol{\epsilon}$ und berechnen

$$T = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{S^2 v_{rr}}},$$

wobei v_{rr} das zu β_2 gehörende Diagonalelement von $(\mathbf{X}'\mathbf{X})^{-1}$ ist, mit $\mathbf{X} = (\mathbf{X}_1|\mathbf{x}_2)$. Der Schätzer $S^2 = \text{SSE}(\hat{\boldsymbol{\beta}})/(n-p)$ bezieht sich auf das komplexere Modell mit $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \beta_2)'$. Wir haben bereits gezeigt, dass $T \sim t_{n-p}$ unter $H_0 : \beta_2 = 0$ gilt. Wir erinnern uns an die einfache Beziehung zwischen T und F:

$$F = T^2 = \frac{\hat{\beta}_2^2}{S^2 v_{rr}}.$$

• Beinhaltet X_1 nur den Intercept $\mathbf{x}_1 = 1$, dann wird mit der F Statistik die Hypothese getestet, dass alle $p-1$ Steigungsparameter Null sind, d.h.

$$H_0 : \beta_1 = \cdots = \beta_{p-1} = 0\,.$$

Keine der erklärenden Variable ist somit im Modell relevant.

In diesem Fall (Modell nur mit Intercept) ist $\mathsf{SSE}(\hat{\beta}_1) = \mathsf{SSE}(\bar{y}) = \mathsf{SST}$. Für die Differenz ergibt sich $\mathsf{SSE}(\hat{\beta}_1) - \mathsf{SSE}(\hat{\boldsymbol{\beta}}) = \mathsf{SST} - \mathsf{SSE}(\hat{\boldsymbol{\beta}}) = \mathsf{SSR}(\hat{\boldsymbol{\beta}})$ und diese ist unabhängig von $\mathsf{SSE}(\hat{\boldsymbol{\beta}})$. Zusammen liefert dies die Statistik zum **Overall F-Test**

$$F = \frac{\mathsf{SSR}(\hat{\boldsymbol{\beta}})/(p-1)}{\mathsf{SSE}(\hat{\boldsymbol{\beta}})/(n-p)} \sim F_{p-1,n-p}\,.$$

Quadratsummen

Interpretation von Quadratsummen, falls diese in Reduktionen zerlegbar sind, die durch sukzessives Aufnehmen von erklärenden Variablen in die Designmatrix entstehen.

Betrachte

$$\mathbf{y} = \mathbf{1}_n\beta_0 + \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \cdots + \mathbf{X}_m\boldsymbol{\beta}_m + \boldsymbol{\epsilon}\,.$$

Die Matrizen $\mathbf{1}_n, \mathbf{X}_1, \ldots, \mathbf{X}_m$ nennt man Terme.

Das einfachste Modell (Null-Model) beinhaltet nur den Intercept (entspricht der Annahme von iid Responses)

$$\mathbf{y} = \mathbf{1}_n\beta_0 + \boldsymbol{\epsilon}\,,$$

und liefert $\hat{\boldsymbol{\mu}}_0 = \mathbf{1}_n\bar{y}$ mit $\mathsf{SSE}_0 = \mathsf{SSE}(\hat{\beta}_0) = \sum_i (y_i - \bar{y})^2$ und df $\nu_0 = n - 1$.

Reduziere nun sukzessive SSE durch Aufnahme weiterer Terme in die Designmatrix. Bezeichne $\hat{\boldsymbol{\mu}}_r$ die Schätzung wenn $\mathbf{X}_1, \ldots, \mathbf{X}_r$ inkludiert sind. Schreibe

$$\mathbf{y} - \hat{\boldsymbol{\mu}}_0 = (\mathbf{y} - \hat{\boldsymbol{\mu}}_m) + (\hat{\boldsymbol{\mu}}_m - \hat{\boldsymbol{\mu}}_{m-1}) + \cdots + (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)\,.$$

Alle Klammerausdrücke sind orthogonal. Mit Pythagoras folgt

$$(\mathbf{y} - \hat{\boldsymbol{\mu}}_0)'(\mathbf{y} - \hat{\boldsymbol{\mu}}_0) = (\mathbf{y} - \hat{\boldsymbol{\mu}}_m)'(\mathbf{y} - \hat{\boldsymbol{\mu}}_m) + \cdots + (\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)'(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0)\,,$$

oder äquivalent

$$\mathsf{SSE}_0 = \mathsf{SSE}_m + \Big(\mathsf{SSE}_{m-1} - \mathsf{SSE}_m\Big) + \cdots + \Big(\mathsf{SSE}_0 - \mathsf{SSE}_1\Big)\,.$$

Alle Terme rechts sind unabhängig. Die Differenz $(\mathsf{SSE}_{r-1} - \mathsf{SSE}_r)$ ist eine Reduktion in der Fehlerquadratsumme durch Aufnahme von X_r ins Modell, das $\mathbf{1}_n, \mathbf{X}_1, \ldots, \mathbf{X}_{r-1}$ bereits beinhaltet. Gibt man immer mehr Terme in das Modell, so reduzieren sich dadurch die Freiheitsgrade (df) und es gilt $\nu_0 \geq \nu_1 \geq \cdots \geq \nu_m$.

Situation $\nu_r = \nu_{r+1}$ tritt auf, wenn die Spalten von $\mathbf{X}_{r+1}$ Linearkombinationen der Spalten von $\mathbf{1}_n, \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_r$ sind. Dann ist $\mathbf{X}_{r+1}$ für unser lineares Modell redundant und es gilt $\nu_r = \nu_{r+1}$, $\hat{\boldsymbol{\mu}}_r = \hat{\boldsymbol{\mu}}_{r+1}$, sowie $\mathsf{SSE}_r = \mathsf{SSE}_{r+1}$.

Quadratsummen zusammenfassen zur **ANOVA Tabelle:**

Terme	df	Resid. QS	Term dazu-geben	df	Reduktion in QS	mittlere QS
$\mathbf{1}_n$	$n-1$	SSE_0				
$\mathbf{1}_n, \mathbf{X}_1$	ν_1	SSE_1	$\mathbf{X}_1$	$(n-1)-\nu_1$	$\mathsf{SSE}_0 - \mathsf{SSE}_1$	$\frac{\mathsf{SSE}_0 - \mathsf{SSE}_1}{n-1-\nu_1}$
$\mathbf{1}_n, \mathbf{X}_1, \mathbf{X}_2$	ν_2	SSE_2	$\mathbf{X}_2$	$\nu_1 - \nu_2$	$\mathsf{SSE}_1 - \mathsf{SSE}_2$	$\frac{\mathsf{SSE}_1 - \mathsf{SSE}_2}{\nu_1 - \nu_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathbf{1}_n, \mathbf{X}_1, \ldots, \mathbf{X}_m$	ν_m	SSE_m	$\mathbf{X}_m$	$\nu_{m-1} - \nu_m$	$\mathsf{SSE}_{m-1} - \mathsf{SSE}_m$	$\frac{\mathsf{SSE}_{m-1} - \mathsf{SSE}_m}{\nu_{m-1} - \nu_m}$

Beispiel: MLR für `vc` – ANOVA und sequentielle F-Tests

```
> anova(mod)
Analysis of Variance Table
Response: vc
          Df Sum Sq Mean Sq F value    Pr(>F)
height     1 211652  211652 79.5378 2.372e-13 ***
age        1   7896    7896  2.9672  0.089144 .
I(age^2)   1  18376   18376  6.9057  0.010441 *
I(age^3)   1  18901   18901  7.1028  0.009443 **
Residuals 74 196915    2661
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation: Gibt man zum Nullmodell (nur Intercept) den Prädiktor `height` dazu, so stellt dies eine signifikante Verbesserung (***) dar. Beinhaltet das Modell den Intercept und `height`, so ist zusätzlich age nur mit p-Wert $= 8.9\%$ relevant. Jedoch ist die Verbesserung des Modells durch die weitere Aufnahme von `I(age^2)` (p-Wert 0.01) und `I(age^3)` (p-Wert 0.009) jeweils signifikant.

Bemerke, dass dies eine spezielle Sequenz darstellt. Interessant wäre auch, zuerst alle Altersterme aufzunehmen und dann erst den Prädiktor `height`:

```
> anova(modh <- lm(vc ~ age + I(age^2) + I(age^3) + height))
  Analysis of Variance Table

  Response: vc
            Df Sum Sq Mean Sq F value    Pr(>F)
  age        1  38531   38531  14.480 0.0002896 ***
  I(age^2)   1  39624   39624  14.890 0.0002415 ***
  I(age^3)   1  36331   36331  13.653 0.0004188 ***
  height     1 142339  142339  53.490 2.546e-10 ***
  Residuals 74 196915    2661
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation: Dies ist eine Aufnahmesequenz, in der jeder zusätzliche Prädiktor für das Modell signifikant notwendig ist.

7. Extra Quadratsummen

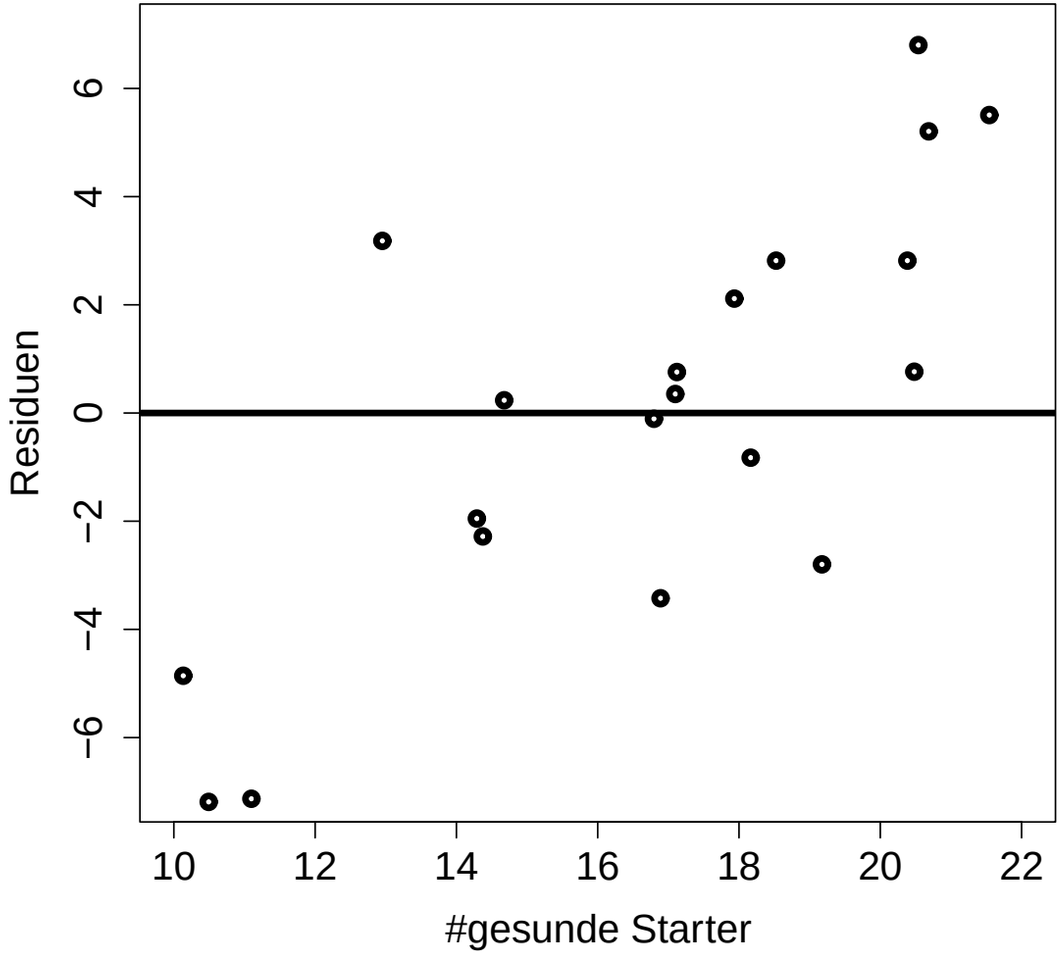
Football Beispiel:

y_i = #Punkte erzielt vom UF Football Team im Spiel i
x_{i1} = #gewonnene Spiele des Gegners in dessen letzten 10 Spielen
x_{i2} = #gesunde Starter (Stammspieler) für UF (von 22) im Spiel i

Angenommen, wir schätzen das SLR

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

und tragen die Residuen r_i gegen x_{i2} auf:



Q: Was können wir daraus schließen?

A: Residuen scheinen linear von der Anzahl gesunder Starter abzuhängen.
Daher sollte x_{i2} in das Modell zusätzlich aufgenommen werden.

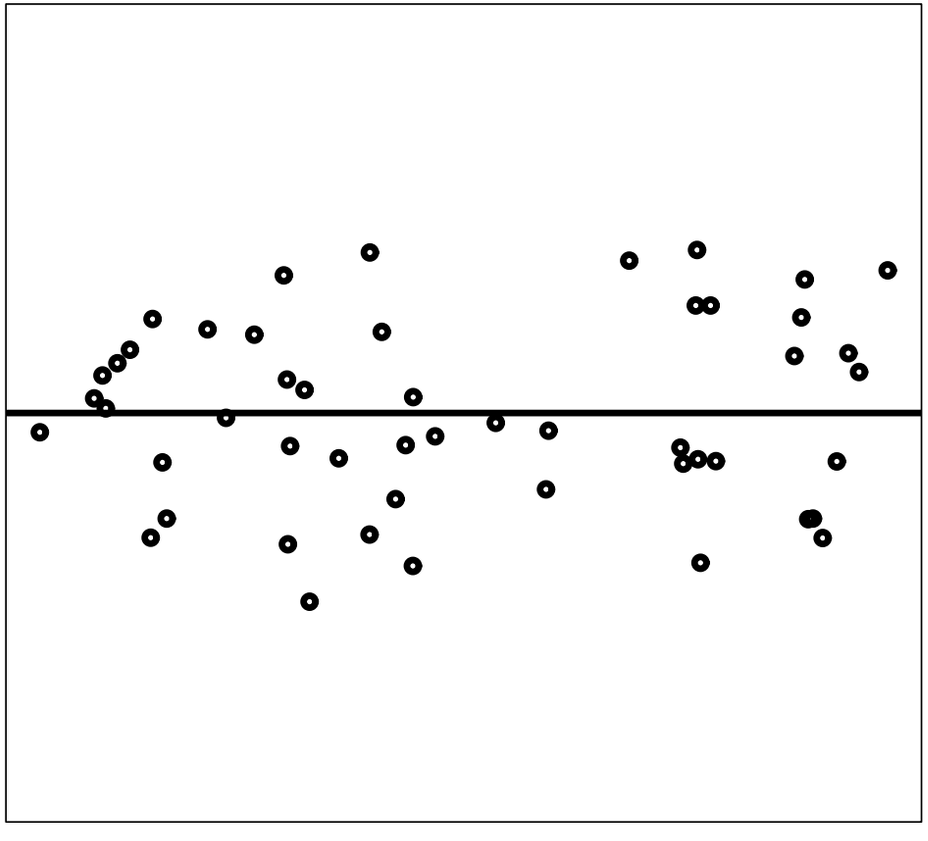
Ein anderes Beispiel:

y_i = Größe einer Person
x_{i1} = Länge des linken Fußes
x_{i2} = Länge des rechten Fußes

Angenommen wir schätzen das SLR

$$y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$$

und tragen die Residuen r_i gegen x_{i2} auf:



Q: Warum jetzt kein Muster?

A: x_{i2} beinhaltet dieselbe Information über y wie x_{i1}.
Obwohl x_{i2} ein sehr guter Prädiktor für die Körpergröße ist, ist er nicht notwendig, falls bereits x_{i1} im Modell ist.

Extra Quadratsummen liefern eine Möglichkeit, formal zu testen, ob ein Satz von Prädiktoren für das Modell notwendig ist, **gegeben** ein anderer Satz von Prädiktoren ist bereits im Modell.

Zur Erinnerung:

$$\begin{aligned} \mathsf{SST} &= \mathsf{SSR}(\hat{\boldsymbol{\beta}}) + \mathsf{SSE}(\hat{\boldsymbol{\beta}}) \\ \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 \\ R^2 &= \frac{\mathsf{SSR}(\hat{\boldsymbol{\beta}})}{\mathsf{SST}} \end{aligned}$$

Wichtige Tatsache: R^2 wird niemals kleiner werden, wenn ein Prädiktor in das Modell zusätzlich aufgenommen wird.

Betrachte 2 unterschiedliche Modelle:

$$\begin{aligned} \mathsf{E}(y_i) &= \beta_0 + \beta_1 x_{i1} \\ \mathsf{E}(y_i) &= \beta_0^* + \beta_1^* x_{i1} + \beta_2^* x_{i2} \end{aligned}$$

Q: Ist SST dieselbe für beide Modelle?

A: Ja! Daher wird SSR($\hat{\boldsymbol{\beta}}$) niemals kleiner werden wenn ein weiterer Prädiktor im Modell aufgenommen wird.

Da SSE und SSR jeweils von den Prädiktoren im Modell abhängen, verwenden wir von nun an die folgende Notation:

SSR(x_1): SSR für ein Modell, das nur x_1 enthält,

SSR(x_1, x_2): SSR für ein Modell mit x_1 und x_2,

SSE(x_1) und SSE(x_1, x_2) haben analoge Definitionen.

Bemerke

$$\begin{aligned} \text{SST} &= \text{SSR}(x_1) + \text{SSE}(x_1) \\ \text{SST} &= \text{SSR}(x_1, x_2) + \text{SSE}(x_1, x_2)\,. \end{aligned}$$

Wir wissen außerdem, dass $\text{SSR}(x_1, x_2) \geq \text{SSR}(x_1)$.

Daher gilt $\text{SSE}(x_1, x_2) \leq \text{SSE}(x_1)$.

Folgerung: SSE wird durch Aufnahme weiterer Prädiktoren nie größer werden.

Betrachtungen zum Beispiel:

y_i = Größe einer Person
x_{i1} = Länge des linken Fußes
x_{i2} = Länge des rechten Fußes

Q: Wie groß wird hierbei folgende Differenz sein?

$$\mathrm{SSR}(x_1, x_2) - \mathrm{SSR}(x_1)$$

A: Wahrscheinlich relativ klein! Kennen wir die Länge des linken Fußes, dann wird das zusätzliche Wissen der Länge des rechten Fußes nicht viel bringen.

Notation: Zusätzliche Quadratsummen

$$\mathsf{SSR}(x_2|x_1) = \mathsf{SSR}(x_1, x_2) - \mathsf{SSR}(x_1)$$

$\mathsf{SSR}(x_2|x_1)$ sagt uns, wie viel wir gewinnen, wenn wir x_2 in das Modell aufnehmen, **gegeben** dass sich bereits x_1 im Modell befindet.

Wir definieren entsprechend $\mathsf{SSR}(x_1|x_2) = \mathsf{SSR}(x_1, x_2) - \mathsf{SSR}(x_2)$.

Wir können dies mit so vielen Prädiktoren machen wie wir nur wollen. Generell gilt beispielsweise:

$$\begin{aligned}
\mathsf{SSR}(x_3, x_5|x_1, x_2, x_4) &= \mathsf{SSR}(x_1, x_2, x_3, x_4, x_5) - \mathsf{SSR}(x_1, x_2, x_4) \\
&= \mathsf{SSR}(\text{alle Prädiktoren}) - \mathsf{SSR}(\text{gegebene Prädiktoren})
\end{aligned}$$

Angenommen, unser Modell lautet:

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

Betrachte nun Tests auf die Parameter β_1, β_2 und β_3.

Ein Parameter: $H_0 : \beta_k = 0, \quad k = 1, 2$ **oder** 3
H_1 : nicht H_0

In Worten fragen wir bei diesem Test: „Brauchen wir x_k, gegeben die beiden anderen (alle übrigen) Prädiktoren sind bereits im Modell?“

Wir können dies direkt mit dem t-Test machen:

$$T = \frac{\hat{\beta}_k}{\sqrt{\mathsf{MSE} \cdot [(\mathbf{X}'\mathbf{X})^{-1}]_{k+1,k+1}}}$$

Zwei Parameter: (einige Parameter)

$H_0 : \beta_1 = \beta_2 = 0$	$H_0 : \beta_1 = \beta_3 = 0$	$H_0 : \beta_2 = \beta_3 = 0$
H_1 : nicht H_0	H_1 : nicht H_0	H_1 : nicht H_0

Zum Beispiel fragen wir beim ersten Test: „Brauchen wir x_1 und x_2, gegeben x_3 ist bereits im Modell?“

Alle Parameter: $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
H_1 : nicht H_0

Dies ist gerade der *Overall* F-Test

Wir können all diese Tests mittels der *Extra Quadratsummen* durchführen.

Hier ist die **ANOVA Tabelle** zum Modell

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

Ursache der Variation	Quadratsumme	df
Regression	$\mathsf{SSR}(x_1, x_2, x_3)$	$p - 1 = 3$
Fehler	$\mathsf{SSE}(x_1, x_2, x_3)$	$n - p = n - 4$
Total	SST	$n - 1$

Partitioniere $\mathsf{SSR}(x_1, x_2, x_3)$ in 3 *Extra Quadratsummen* mit jeweils $df = 1$. Z.B.

$$\mathsf{SSR}(x_1, x_2, x_3) = \mathsf{SSR}(x_1) + \mathsf{SSR}(x_2|x_1) + \mathsf{SSR}(x_3|x_1, x_2)\,.$$

Modifizierte ANOVA Tabelle:

Ursache der Variation	Quadratsumme	df
Regression	SSR(x_1, x_2, x_3)	3
	SSR(x_1)	1
	SSR$(x_2 \mid x_1)$	1
	SSR$(x_3 \mid x_1, x_2)$	1
Fehler	SSE(x_1, x_2, x_3)	$n-4$
Total	SST	$n-1$

Bemerke: es gibt $3! = 6$ gleichwertige Möglichkeiten, um SSR(x_1, x_2, x_3) zu partitionieren.

Drei Klassen von Tests: ($p = 4$ in unserem Beispiel)

• **Ein Parameter:** $H_0 : \beta_2 = 0$ gegen H_1 : nicht H_0

Teststatistik: $F = \dfrac{\mathsf{SSR}(x_2|x_1, x_3)/1}{\mathsf{SSE}(x_1, x_2, x_3)/(n-p)}$

Verwerfungsregel: Verwirf H_0, falls $F > F_{1,n-p;1-\alpha}$

• **Einige Parameter:** $H_0 : \beta_2 = \beta_3 = 0$ gegen H_1 : nicht H_0

Teststatistik: $F = \dfrac{\mathsf{SSR}(x_2, x_3|x_1)/2}{\mathsf{SSE}(x_1, x_2, x_3)/(n-p)}$

Verwerfungsregel: Verwirf H_0, falls $F > F_{2,n-p;1-\alpha}$

• **Alle Parameter:** $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ gegen H_1 : nicht H_0

Teststatistik: $F = \dfrac{\mathsf{SSR}(x_1, x_2, x_3)/3}{\mathsf{SSE}(x_1, x_2, x_3)/(n-p)}$

Verwerfungsregel: Verwirf H_0, falls $F > F_{p-1,n-p;1-\alpha}$

Kommen wir zum Modell

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

zurück und denken an einen Test auf

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{gegen} \quad H_1 : \text{nicht } H_0$$

Teststatistik: $F = \dfrac{\mathsf{SSR}(x_2, x_3|x_1)/2}{\mathsf{SSE}(x_1, x_2, x_3)/(n-4)}$

Wie bekommen wir $\mathsf{SSR}(x_2, x_3|x_1)$ wenn wir bereits $\mathsf{SSR}(x_1)$, $\mathsf{SSR}(x_2|x_1)$ und $\mathsf{SSR}(x_3|x_1, x_2)$ haben?

$$\mathsf{SSR}(x_2, x_3|x_1) = \mathsf{SSR}(x_2|x_1) + \mathsf{SSR}(x_3|x_1, x_2)$$

Wenn wir aber $\mathsf{SSR}(x_2)$, $\mathsf{SSR}(x_1|x_2)$ und $\mathsf{SSR}(x_3|x_1, x_2)$ hätten? **Wir stecken fest!**

Umsetzung in R: Sequenz hängt von der Spezifikation des Linearen Prädiktors durch die Modellformel ab!

`lm(y ~ x1 + x2 + x3)`	`lm(y ~ x2 + x1 + x3)`
SSR(x_1)	SSR(x_2)
SSR$(x_2 \vert x_1)$	SSR$(x_1 \vert x_2)$
SSR$(x_3 \vert x_1, x_2)$	SSR$(x_3 \vert x_1, x_2)$

Beispiel: Patientenzufriedenheit

y_i = Zufriedenheit ($n = 23$)
x_{i1} = Alter (in Jahren)
x_{i2} = Schwere der Krankheit (Index)
x_{i3} = Ängstlichkeit (Index)

Modell 1: Betrachte das Modell mit allen paarweisen Interaktionen ($p = 7$)

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3}$$

und teste (level α) auf die Notwendigkeit all dieser Interaktionen, d.h. auf

$$H_0 : \beta_4 = \beta_5 = \beta_6 = 0 \quad \text{gegen} \quad H_1 : \text{nicht } H_0.$$

Bezeichne mit i_{jk} die Interaktion $x_j x_k$. Dann ist dafür die Teststatistik:

$$F = \frac{\mathsf{SSR}(i_{12}, i_{13}, i_{23} | x_1, x_2, x_3)/3}{\mathsf{SSE}(x_1, x_2, x_3, i_{12}, i_{13}, i_{23})/(n-p)}$$

Verwerfungsregel: Verwirf H_0, falls $F > F_{3,n-p;1-\alpha}$.

Q: Wie bekommt man diese Quadratsumme?

Es gibt $6! = 720$ Partitionen von $\mathsf{SSR}(x_1, x_2, x_3, i_{12}, i_{13}, i_{23})$.

Q: Welche davon erlauben die Berechnung von F?
A: Jene mit i_{12}, i_{13} und i_{23} zuletzt.

$$\begin{aligned}\mathsf{SSR}(\cdot) &= \mathsf{SSR}(x_1) + \mathsf{SSR}(x_2|x_1) + \mathsf{SSR}(x_3|x_1, x_2) + \mathsf{SSR}(i_{12}|x_1, x_2, x_3) \\ &\quad + \mathsf{SSR}(i_{13}|x_1, x_2, x_3, i_{12}) + \mathsf{SSR}(i_{23}|x_1, x_2, x_3, i_{12}, i_{13})\,.\end{aligned}$$

Addiere die letzten 3 (die Interaktionsterme) $\Rightarrow \mathsf{SSR}(i_{23}, i_{24}, i_{34}|x_1, x_2, x_3)$

```
> anova(mod1 <- lm(sat ~ age + sev + anx + age:sev + age:anx + sev:anx))
Analysis of Variance Table
Response: sat
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1 3678.4  3678.4 32.1981 3.452e-05 ***
sev        1  402.8   402.8  3.5256   0.07878 .
anx        1   52.4    52.4  0.4588   0.50787
age:sev    1    0.0     0.0  0.0002   0.98932
age:anx    1  144.2   144.2  1.2625   0.27775
sev:anx    1   39.4    39.4  0.3451   0.56510
Residuals 16 1827.9   114.2
```

```
> anova(mod1)
Analysis of Variance Table
Response: sat
          Df  Sum Sq Mean Sq F value Pr(>F)
age        1 3678.44 3678.44 32.20  3.45e-05 ***
sev        1  402.78  402.78  3.53  0.079    .
anx        1   52.41   52.41  0.46  0.508
age:sev    1    0.02    0.02  0.00  0.989
age:anx    1   144.2   144.2  1.26  0.278
sev:anx    1    39.4    39.4  0.35  0.565
Residuals 16  1827.9   114.2
```

$$F = \frac{(0.02 + 144.2 + 39.4)/3}{114.24} = 0.54 < F_{3,16;0.95} = 3.24$$

H_0 kann nicht verworfen werden $\Rightarrow$ Alle 3 Interaktionen scheinen irrelevant!

Modell 2: Wir sind jetzt alle Interaktionsterme los und betrachten von nun an das Modell, welches nur die Haupteffekte beinhaltet (main effects model):

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}\,.$$

Benötigen wir x_2 (Schwere der Krankheit) und x_3 (Ängstlichkeit), falls x_1 (Alter) bereits im Modell ist?

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{gegen} \quad H_1 : \text{nicht } H_0$$

Teststatistik: $F = \dfrac{\mathsf{SSR}(x_2, x_3|x_1)/2}{\mathsf{SSE}(x_1, x_2, x_3)/(n-p)}$

Verwerfungsregel: Verwirf H_0, falls $F > F_{2,n-p;1-\alpha}$.

Q: Wie bekommt man diese Quadratsumme?

Beispielsweise gilt

$$\mathsf{SSR}(x_2, x_3|x_1) = \mathsf{SSR}(x_2|x_1) + \mathsf{SSR}(x_3|x_1, x_2)$$

```
> anova(mod2 <- lm(sat ~ age + sev + anx))
Analysis of Variance Table
Response: sat
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1 3678.4  3678.4 34.7439 1.124e-05 ***
sev        1  402.8   402.8  3.8044   0.06603 .
anx        1   52.4    52.4  0.4951   0.49021
Residuals 19 2011.6   105.9
```

Wir haben somit

$$F = \frac{(402.8 + 52.4)/2}{105.9} = 2.15 < F_{2,19;0.95} = 3.52\,.$$

Da $F < 3.52$ kann H_0 nicht verworfen werden (x_2 und x_3 nicht notwendig).

Modell 3: Wir sind x_2 (Schwere der Krankheit) und x_3 (Ängstlichkeit) los, und betrachten das SLR nur mit x_1 (Alter)

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1}$$

```
> anova(mod3 <- lm(sat ~ age))
Analysis of Variance Table

Response: sat
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1 3678.4  3678.4  31.315 1.489e-05 ***
Residuals 21 2466.8   117.5
```

Gesucht werden jetzt 95% Konfidenzintervalle für den Parameter β_1 und für den Erwartungswert $\mathsf{E}(y_h) = \mathbf{x}_h'\boldsymbol{\beta}$, mit $\mathbf{x}_h' = (1, 40, 50, 2)$, unter allen 3 Modellen:

```
> new.data <- data.frame(age=40, sev=50, anx=2)
```

Modell 3: $(p = 2)$

```
> predict(mod3, new.data, interval="confidence", level=0.95)
       fit     lwr      upr
1 60.75029 56.0453 65.45528
> confint(mod3)
                2.5 %      97.5 %
(Intercept) 98.868284 144.7953507
age         -2.094526  -0.9595502
```

Modell 2: $(p = 4)$

```
> predict(mod2, new.data, interval="confidence", level=0.95)
       fit      lwr      upr
1 63.94183 55.85138 72.03228
```

```
> confint(mod2)
                 2.5 %      97.5 %
(Intercept) 108.926839 216.8249581
age          -1.841264  -0.5793727
sev          -2.384272   1.0524608
anx         -34.234265  17.0082018
```

Modell 1: $(p = 7)$

```
> predict(mod1, new.data, interval="confidence", level=0.95)
       fit      lwr      upr
1 63.67873 54.93979 72.41767
> confint(mod1)
                   2.5 %      97.5 %
(Intercept) -118.6330908 601.7751751
age           -9.5863470  10.1485879
sev          -17.7867745   5.1327693
anx         -191.4690710 239.5207824
:               :           :
```

Zusammenfassung der Ergebnisse:

Modell	KIV(μ_h) Länge	KIV(β_1) Länge
1 ($p = 7$)	17.478	19.735
2 ($p = 4$)	16.181	1.262
3 ($p = 2$)	9.410	1.135

Konfidenzintervalle für μ_h und β_1 werden alle mit steigender Komplexität breiter! Am deutlichsten ist dieses Verhalten für KIV(β_1) zu beobachten.

Im Gegensatz zu den sequentiellen QS sind die **partiellen** QS definiert als

$$\mathsf{SSR}(x_j|x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_{p-1})\,.$$

Diese QS erklärt den Wert von x_j als zusätzlicher Prädiktor im Modell, das alle übrigen Variablen beinhaltet, d.h. man testet $H_0 : \beta_j = 0$ im Modell $\mathsf{E}(y) = \mathbf{x}'\boldsymbol{\beta}$.

Dafür gilt

$$F = \frac{\mathsf{SSR}(x_j|x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_{p-1})/1}{\mathsf{SSE}(x_1,\ldots,x_{p-1})/(n-p)} \overset{H_0}{\sim} F_{1,n-p}\,,$$

was äquivalent ist zum bekannten t-Test

$$T = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \overset{H_0}{\sim} t_{n-p}\,.$$

Berechnung am Beispiel des Modells für die Zufriedenheit der Patienten in Abhängigkeit aller Haupteffekte, d.h. $\mathrm{E}(\texttt{sat}) = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{sev} + \beta_3 \texttt{sev}$

Durch `anova(mod)`, wobei jedoch nur die **letzte Zeile** verwendet werden kann!

```
> anova(mod2)
Analysis of Variance Table

Response: sat
          Df Sum Sq Mean Sq F value    Pr(>F)
age        1 3678.4  3678.4 34.7439 1.124e-05 ***
sev        1  402.8   402.8  3.8044   0.06603 .
anx        1   52.4    52.4  0.4951   0.49021
Residuals 19 2011.6   105.9

> anova(lm(sat ~ sev + anx + age))[3,3] # 1706.666
> anova(lm(sat ~ age + anx + sev))[3,3] #   69.651
> anova(lm(sat ~ age + sev + anx))[3,3] #   52.414
```

Korrelation unter den Prädiktoren: Multikollinearität

Wiederholung der SLR Situation: Daten (x_i, y_i), $i = 1, \ldots, n$.

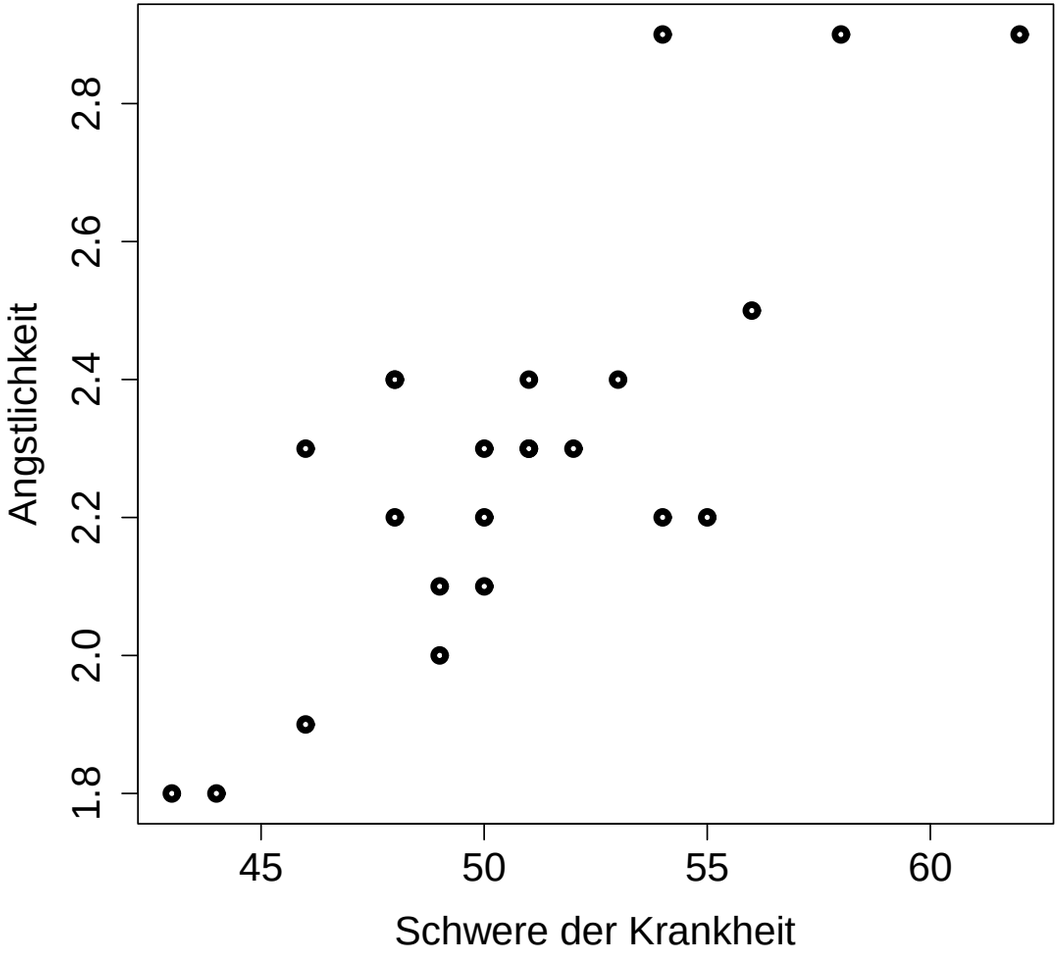
$$R^2 = \text{SSR/SST}$$

beschreibt den Anteil an der totalen Variation in den y_i's, der durch die lineare Abhängigkeit zwischen x and y erklärt ist.

Wegen SSR $= \hat{\beta}_1^2 s_x^2$, mit $\hat{\beta}_1 = s_{xY}^2/s_x^2$ und mit $s_Y^2 = \text{SST}$, ist der empirische Korrelationskoeffizient zwischen x and y gleich

$$R = \text{sign}(\hat{\beta}_1)\sqrt{R^2} = \frac{s_{xY}^2}{\sqrt{s_x^2 s_Y^2}}\,.$$

Er liefert uns Information über die Stärke der linearen Beziehung zwischen x und y, wie auch über das Vorzeichen der Steigung ($-1 \leq R \leq 1$).



Zufriedenheit der Patienten:

Korrelation zwischen

$X_{i2} =$ Schwere der Krankheit

$X_{i3} =$ Ängstlichkeit

$r_{23} = 0.7945$ (siehe unten)

Für eine MLR Datenzeile $(x_{i1}, \ldots, x_{i,p-1}, y_i)$ bezeichne

r_{jY} den empirischen Korrelationskoeffizienten zwischen x_j und y, und

r_{jk} sei der empirische Korrelationskoeffizient zwischen x_j und x_k.

• Falls $r_{jk} = 0$, dann sind x_j und x_k **unkorreliert**.

Sind die meisten r_{jk}'s in der Nähe von $+1$ oder -1, dann sprechen wir von **Multikollinearität** unter den Prädiktoren.

```
> cor(patsat)
        sat     age     sev     anx
sat  1.0000 -0.7737 -0.5874 -0.6023
age -0.7737  1.0000  0.4666  0.4977
sev -0.5874  0.4666  1.0000  0.7945
anx -0.6023  0.4977  0.7945  1.0000
```

Unkorrelierte versus korrelierte Prädiktoren

Betrachte die 3 Modelle:

(1) $\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1}$

(2) $\mathsf{E}(y_i) = \beta_0 + \qquad\quad \beta_2 x_{i2}$

(3) $\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$

und die beiden Situationen:

- x_1 und x_2 sind unkorreliert ($r_{12} \approx 0$), dann wird

  $\hat{\beta}_1$ unter den Modellen (1) und (3) ähnlich sein,

  $\hat{\beta}_2$ unter den Modellen (2) und (3) ähnlich sein,

  $\mathsf{SSR}(x_1|x_2) \approx \mathsf{SSR}(x_1)$,

  $\mathsf{SSR}(x_2|x_1) \approx \mathsf{SSR}(x_2)$.

- x_1 und x_2 sind stark korreliert ($|r_{12}| \approx 1$), dann wird

  $\hat{\beta}_1$ unter den Modellen (1) und (3) unterschiedlich sein,

  $\hat{\beta}_2$ unter den Modellen (2) und (3) unterschiedlich sein,

  $\text{SSR}(x_1|x_2) < \text{SSR}(x_1)$,

  $\text{SSR}(x_2|x_1) < \text{SSR}(x_2)$.

Ist $r_{12} \approx 0$, dann beinhalten x_1 und x_2 **keine redundante Information** über y.

Dann erklärt x_1 denselben Anteil an SST wenn x_2 im Modell ist, wie wenn x_2 nicht im Modell ist.

Übersicht: Effekt der Multikollinearität

Die Standardfehler der Parameterschätzer sind aufgebläht. Somit könnten Konfidenzintervalle für die Regressionsparameter bereits zu groß sein, um noch nützlich zu sein.

Inferenz über $\mathsf{E}(y_h) = \mathbf{X}_h'\boldsymbol{\beta}$, dem Erwartungswert einer Response in $\mathbf{X}_h'$, und $y_{h(new)}$, eine neue Zufallsvariable, die man in $\mathbf{X}_h$ beobachten wird, bleiben davon großteils unbeeinflusst.

Die Idee, x_1 zu vergrößern und dabei x_2 festzuhalten, ist nicht mehr glaubwürdig:

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\,.$$

Interpretation: β_1 bezeichnet "die Änderung im Erwartungswert von y entsprechend einer Änderung von x_1 um eine Einheit, wenn dabei x_2 festgehalten ist".

Polynomiale Regression

Angenommen, wir haben SLR Daten (x_i, y_i), $i = 1, \ldots, n$. Falls $y_i = f(x_i) + \epsilon_i$, mit $f(\cdot)$ unbekannt, approximieren wir $f(\cdot)$ durch ein Polynom (Potenzreihe), d.h.

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots .$$

Meist geht man nicht weiter als bis zum 3. Grad.

Standard Prozedur:
• Starte mit einem Polynom höherer Ordnung und versuche zu vereinfachen.
• Wird man x^k nicht los, so sollten auch alle niedrigeren Ordnungen x^{k-1}, $x^{k-2}, \ldots, x$ im Modell bleiben (Hierarchie).

Warnung:
• Das Modell $\mathsf{E}(y_i) = \beta_0 + \beta_1 x_i + \cdots + \beta_{n-1} x_i^{n-1}$ passt immer perfekt ($p = n$).
• Polynome in x sind stark korreliert.

Polynomiale Regression: Beispiel Fischdaten

Definiere $y_i = \log(\text{Artenreichtum} + 1)$ beobachtet im See i, $i = 1, \ldots, 80$, im Adirondack State Park/NY und pH_i den dazugehörenden pH-Wert.

Wir betrachten das polynomiale Modell 3. Ordnung, d.h.

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 \text{pH}_i + \beta_2 \text{pH}_i^2 + \beta_3 \text{pH}_i^3 \,.$$

```
> fish <- read.table("fish.dat",header=T) # 1166 observations
> fish80 <- fish[1:80, ]                  # take only the first 80 of these
> attach(fish80)

> lnsr <- log(rch+1) # log(species richness + 1)
> ph2 <- ph^2; ph3 <- ph*ph2
```

```
> summary(m3 <- lm(lnsr ~ ph + ph2 + ph3))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -16.82986    7.44163  -2.262   0.0266 *
ph            7.07937    3.60045   1.966   0.0529 .
ph2          -0.87458    0.56759  -1.541   0.1275
ph3           0.03505    0.02930   1.196   0.2354
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4577 on 76 degrees of freedom
Multiple R-squared: 0.4466,     Adjusted R-squared: 0.4248
F-statistic: 20.45 on 3 and 76 DF,  p-value: 8.174e-10
```

```
> anova(m3)
Analysis of Variance Table

Response: lnsr
          Df  Sum Sq Mean Sq F value    Pr(>F)
ph         1  7.9340  7.9340 37.8708 3.280e-08 ***
ph2        1  4.6180  4.6180 22.0428 1.158e-05 ***
ph3        1  0.2998  0.2998  1.4308    0.2354
Residuals 76 15.9221  0.2095
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

pH^3 wird nicht benötigt!

Vielleicht reicht sogar ein SLR. Wir testen dazu

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{gegen} \quad H_1 : \text{nicht } H_0$$

Teststatistik:

$$\begin{aligned} F &= \frac{\mathsf{SSR}(pH^2, pH^3|pH)/2}{\mathsf{SSE}(pH, pH^2, pH^3)/(n-4)} \\ &= \frac{(4.6180 + 0.2998)/2}{0.2095} = 11.74 \end{aligned}$$

Verwerfungsregel: Verwirf H_0, falls $F > F_{2,76;0.95} = 3.1$.

Also ist ein Term höherer Ordnung notwendig.

Testen wir

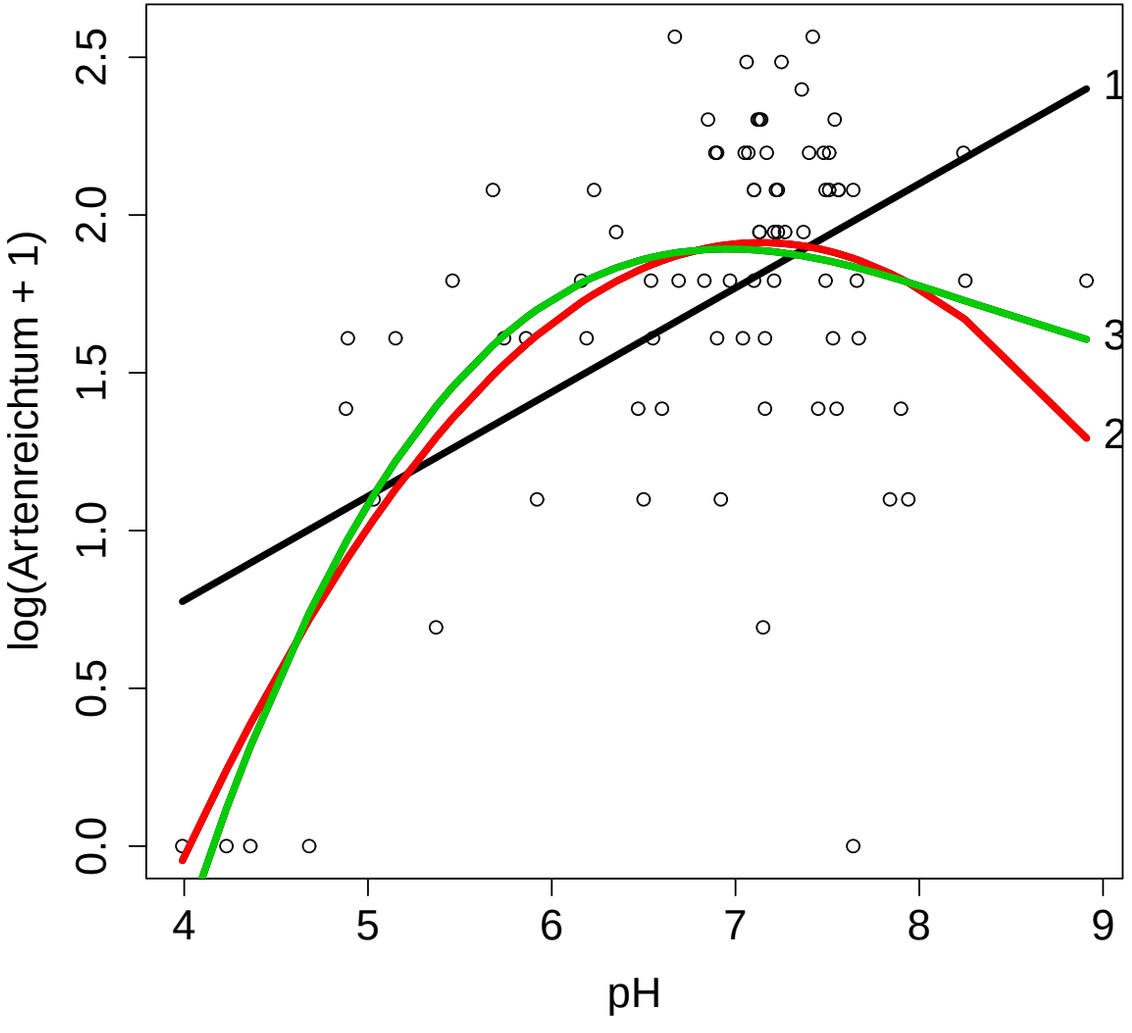
$$H_0 : \beta_3 = 0 \quad \text{gegen} \quad H_1 : \beta_3 \neq 0$$

Teststatistik: $T = 1.196$ mit p-Wert 0.2354.

Schlussfolgerung: wir werfen pH3 aus dem Modell raus und verbleiben bei

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 pH_i + \beta_2 pH_i^2$$

```
> summary(m2 <- lm(lnsr ~ ph + ph2))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.15351    1.66750  -4.890 5.40e-06 ***
ph           2.82010    0.53452   5.276 1.18e-06 ***
ph2         -0.19752    0.04219  -4.682 1.20e-05 ***
---
Residual standard error: 0.459 on 77 degrees of freedom
Multiple R-squared: 0.4362,     Adjusted R-squared: 0.4216
F-statistic: 29.79 on 2 and 77 DF,  p-value: 2.615e-10
```


log(Artenreichtum + 1)
0.0
0.5
1.0
1.5
2.0
2.5
4
5
6
7
8
9
pH
1
3
2

Q: Wozu das alles? Wir haben doch nur den Term 3. Ordnung weggebracht.
A: Angenommen wir interessieren uns für ein 95% Konfidenzintervall für β_1:

Modell	$\hat{\beta}_1$	s.e.	KIV(β_1)
3. Ordnung	7.08	3.60	$(-0.12, 14.28)$
2. Ordnung	2.82	0.53	$(+1.75, \ \ 3.89)$

Dies zeigt ganz deutlich, wie wertvoll sparsame Modelle sein können!

```
> # confint.default() would be based on asymptotic normality
> confint(m3) # based on t(0.975, 76)-quantiles
                   2.5 %      97.5 %
(Intercept) -31.65114823 -2.00856503
ph           -0.09154457 14.25029168
ph2          -2.00502704  0.25586604
ph3          -0.02330737  0.09339972
> confint(m2) # based on t(0.975, 76)-quantiles
                  2.5 %     97.5 %
(Intercept) -11.4739346 -4.8330846
ph            1.7557261  3.8844675
ph2          -0.2815279 -0.1135132
```

Wir können dies alles natürlich auch mit mehreren Prädiktoren machen.

Angenommen, wir haben MLR Daten der Form (x_{i1}, x_{i2}, y_i), $i = 1, \ldots, n$.

Modell 2. Ordnung:

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1}^2 + \beta_4 x_{i2}^2 + \beta_5 x_{i1} x_{i2}$$

Wir können testen $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ und damit hinterfragen: "Ist ein Modell 1. Ordnung ausreichend?"

Teststatistik:

$$F \quad = \quad \frac{\mathsf{SSR}(x_1^2, x_2^2, x_1 x_2 | x_1, x_2)/3}{\mathsf{SSE}(x_1, x_2, x_1^2, x_2^2, x_1 x_2)/(n-6)}$$

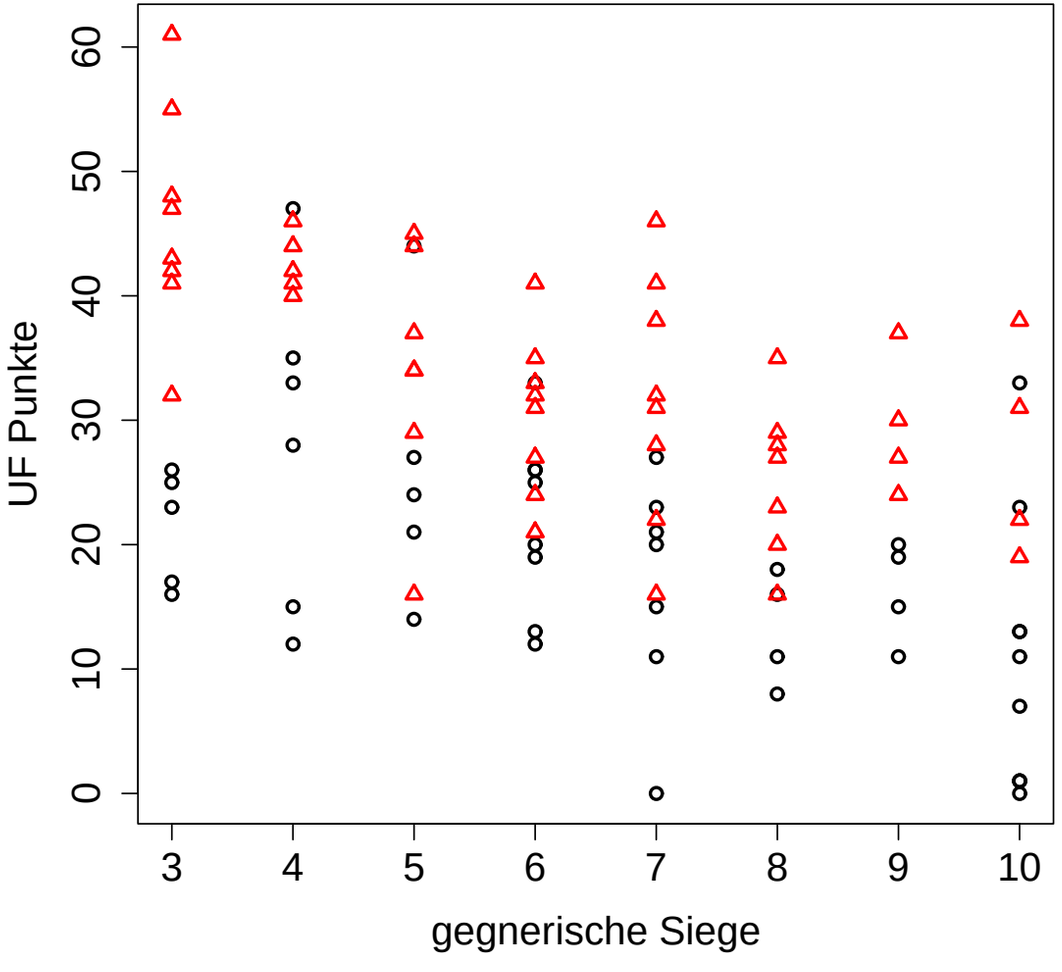
Verwerfungsregel: Verwirf H_0, falls $F > F_{3,n-6;0.95}$.

8. Qualitative Prädiktoren

Beispiel: Für die letzten 100 UF Football Spiele haben wir:

y_i = #Punkte erzielt vom UF Football Team im Spiel i
x_{i1} = #gewonnene Spiele des Gegners in dessen letzten 10 Spielen

Unterscheide jetzt auch noch zwischen **Heim-** ($\triangle$) und **Auswärtsspiel** ($\circ$).



Q: Wie können wir "Heimspiel" und "Auswärtsspiel" in das SLR einbauen?

A: Eine **Indikatorvariable**:

$$x_{i2} = \begin{cases} 1 & \text{Spiel } i \text{ ist Heimspiel} \\ 0 & \text{sonst} \end{cases}$$

Neues Modell:

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\,.$$

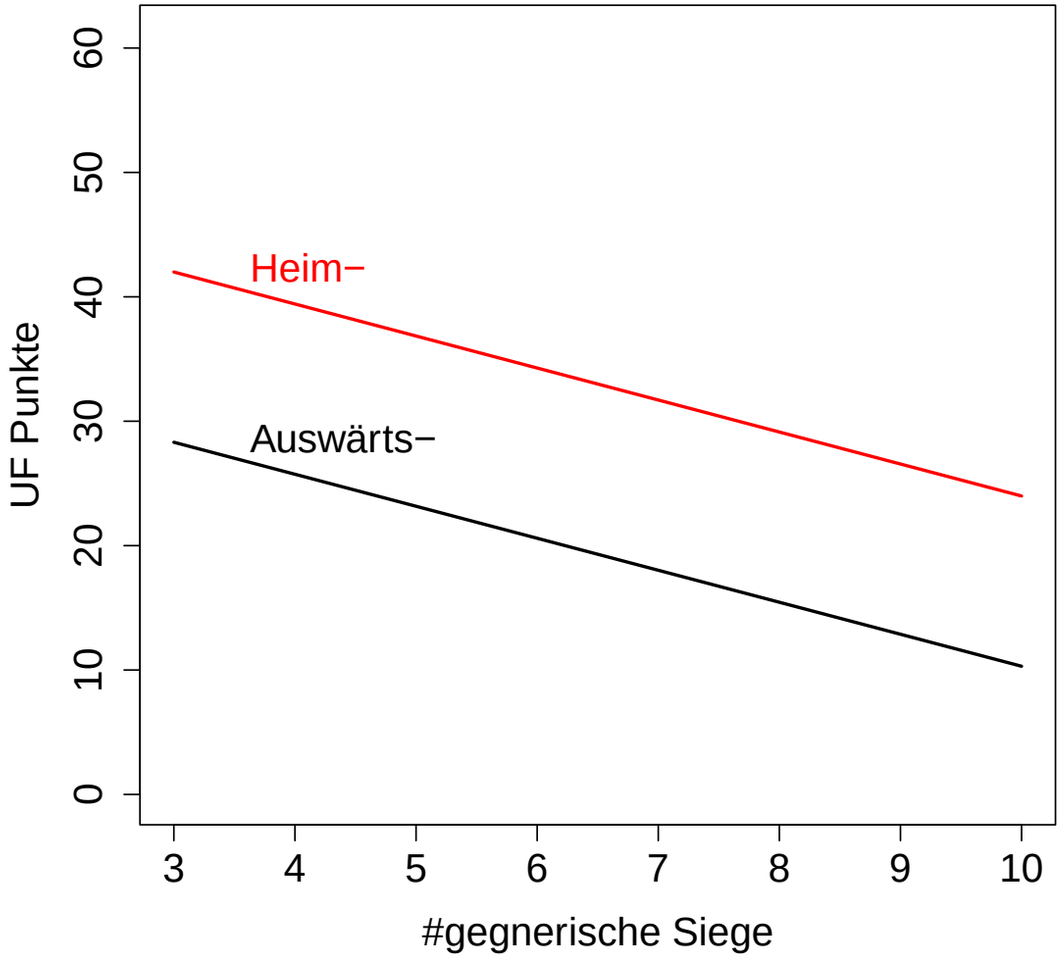
Für Heimspiele:

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 \cdot 1 = (\beta_0 + \beta_2) + \beta_1 x_{i1}\,.$$

Für Auswärtsspiele:

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 \cdot 0 = \beta_0 + \beta_1 x_{i1}\,.$$

Wie sieht dazu die Abbildung aus?



Identische Steigung β_1 aber unterschiedliche Intercepts $(\beta_0 + \beta_2)$ und β_0

Wie entscheidet man, ob verschiedene Intercepts notwendig sind?

Teste $H_0 : \beta_2 = 0$ gegen
H_1 : nicht H_0

t-Test:
$t = \hat{\beta}_2 / \sqrt{\mathsf{MSE} \cdot [(\mathbf{X}'\mathbf{X})^{-1}]_{3,3}}$
F-test:
$F = \mathsf{SSR}(x_2|x_1)/\mathsf{MSE}(x_1, x_2)$

Warum nicht 2 Indikatoren?

Definiere

$$x_{i2}^* = \begin{cases} 1 & \text{Heimspiel} \\ 0 & \text{sonst} \end{cases} \qquad x_{i3}^* = \begin{cases} 1 & \text{Auswärtsspiel} \\ 0 & \text{sonst} \end{cases}$$

und betrachte das Modell

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2^* x_{i2}^* + \beta_3^* x_{i3}^* \,.$$

Bemerke, $x_{i2}^* + x_{i3}^* = 1$, das ist der Intercept (1) in der i-ten Zeile von $\mathbf{X}$. Daher sind die Spalten von $\mathbf{X}$ nicht mehr länger linear unabhängig.

Allgemeine Regel: Eine qualitative Variable mit c Klassen wird durch $c - 1$ Indikatoren repräsentiert, wobei jeder die Werte 0 und 1 annimmt.

Frage: Wie realistisch sind parallele Geraden?

Wir hinterfragen damit, wie realistisch es ist, anzunehmen, dass "das UF Team wird um β_2 mehr Punkte bei Heimspielen als bei Auswärtsspielen machen, unabhängig von der Stärke des Gegners"?

Wie können wir das Modell diesbezüglich flexibler machen?

Antwort: Gib einen Interaktionsterm in das Modell, d.h.

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$$

Für Heimspiele: $\mathsf{E}(y_i) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1}$

Für Auswärtsspiele: $\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1}$

Q: Wie kann man die Frage beantworten "Reicht ein Gerade aus"?

A: Teste: $H_0 : \beta_2 = \beta_3 = 0$ gegen H_1 : nicht H_0

Teststatistik:

$$F = \frac{\mathsf{SSR}(x_1 x_2, x_2 | x_1)/2}{\mathsf{SSE}(x_1, x_2, x_1 x_2)/(n-p)}$$

Verwerfungsregel: Verwirf H_0, falls $F > F_{2,n-p;1-\alpha}$.

Q: Wann ist diese Extra Quadratsumme in `R` verfügbar?

A: Passe ein Modell an, in dem der Interaktionsterm an letzter Stelle ist!

Komplexere Modelle

Mehr als 2 Klassen

Beispiel: $y_i =$ Reichweite des $i$ten Fahrzeugs
$x_{i1} =$ Alter des i-ten Fahrzeugs
Typ des $i$ten Fahrzeugs: heimische und ausländische PKWs, sowie LKWs

Wiederholung der allgemeinen Regel: Die Anzahl notwendiger Indikatoren ist um eins geringer als die Anzahl der Stufen dieses Faktors. Hier benötigen wir daher zwei Indikatoren:

$$x_{i2} = \begin{cases} 1 & \text{heimischer PKW} \\ 0 & \text{sonst} \end{cases} \qquad x_{i3} = \begin{cases} 1 & \text{ausländischer PKW} \\ 0 & \text{sonst} \end{cases}$$

Modell:

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

$$x_{i2} = \begin{cases} 1 & \text{heimischer PKW} \\ 0 & \text{sonst} \end{cases} \qquad x_{i3} = \begin{cases} 1 & \text{ausländischer PKW} \\ 0 & \text{sonst} \end{cases}$$

Modell:

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

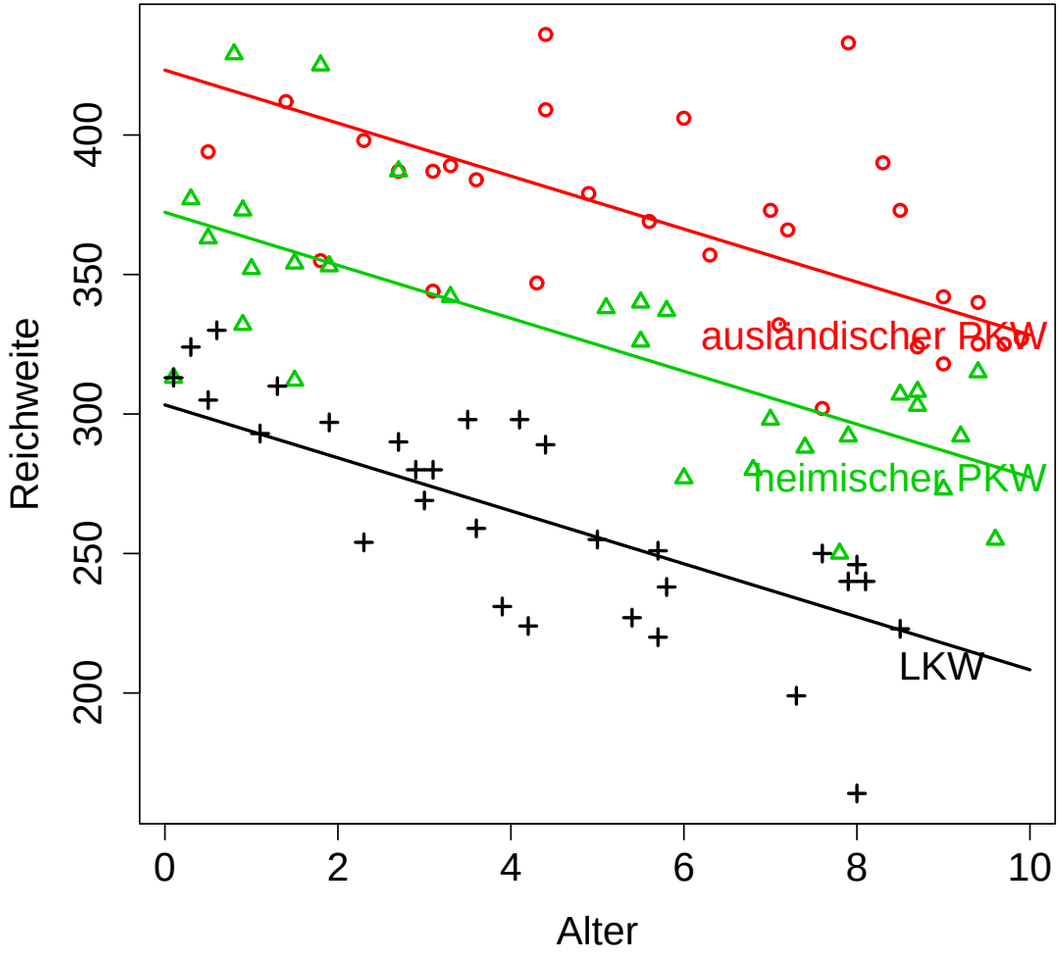
heimischer PKW:	$\mathsf{E}(y_i) = (\beta_0 + \beta_2) + \beta_1 x_{i1}$
ausländischer PKW:	$\mathsf{E}(y_i) = (\beta_0 + \beta_3) + \beta_1 x_{i1}$
LKW:	$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1}$

```
> attach(car); car
> car
   milage age     type
1     307 8.5 domestic
2     354 1.5 domestic
:
89    290 2.7    truck
90    305 0.5    truck
```

```
> x2 <- rep(0, 90) + (type=="domestic")
> x3 <- rep(0, 90) + (type=="foreign")
> lm(milage ~ age + x2 + x3, data=car)
Coefficients:
(Intercept)          age           x2           x3
    303.266       -9.494       69.020      119.992
```

Viel einfacher ist die direkte Verwendung eines **Faktors** in R. Als 1. Stufe (Referenzstufe repräsentiert durch den Intercept) wird hierbei die lexikografisch erste Stufe (`domestic` $<$ `foreign` $<$ `truck`) verwendet.

```
> is.factor(type)
[1] TRUE
> lm(milage ~ age + type, data=car)
Coefficients:
(Intercept)          age  typeforeign    typetruck
    372.286       -9.494       50.972      -69.020
```

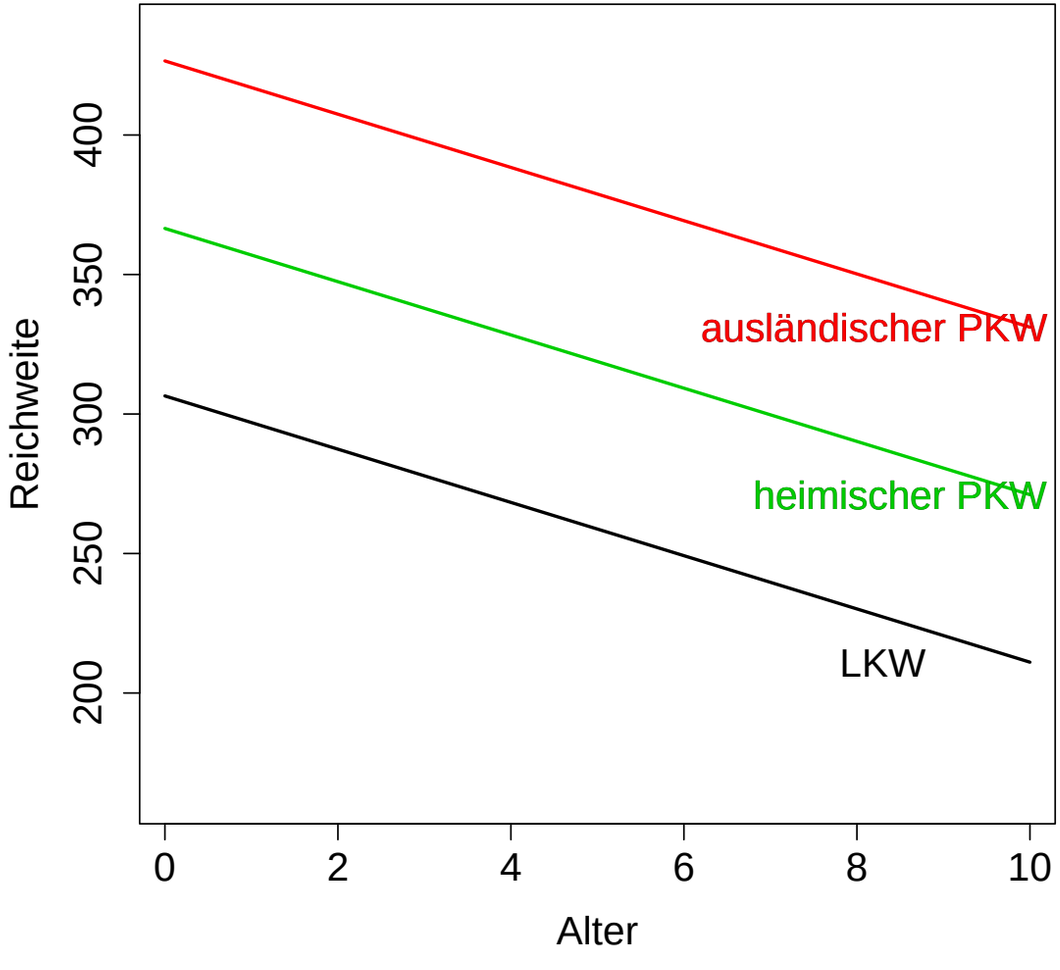

Reichweite
200
250
300
350
400
0
2
4
6
8
10
Alter
ausländischer PKW
heimischer PKW
LKW

Q: Warum kann man nicht eine Variable mit 3 Werten dafür benutzen:

$$x_{i2}^* = \begin{cases} 0 & \text{trucks} \\ 1 & \text{domestic} \\ 2 & \text{foreign} \end{cases}$$

A: Modell: $\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2^* x_{i2}^*$

```
> x2star <- x2 + 2*x3
> lm(milage ~ age + x2star, data=car)
Coefficients:
(Intercept)          age       x2star
    306.482       -9.544       60.037
```

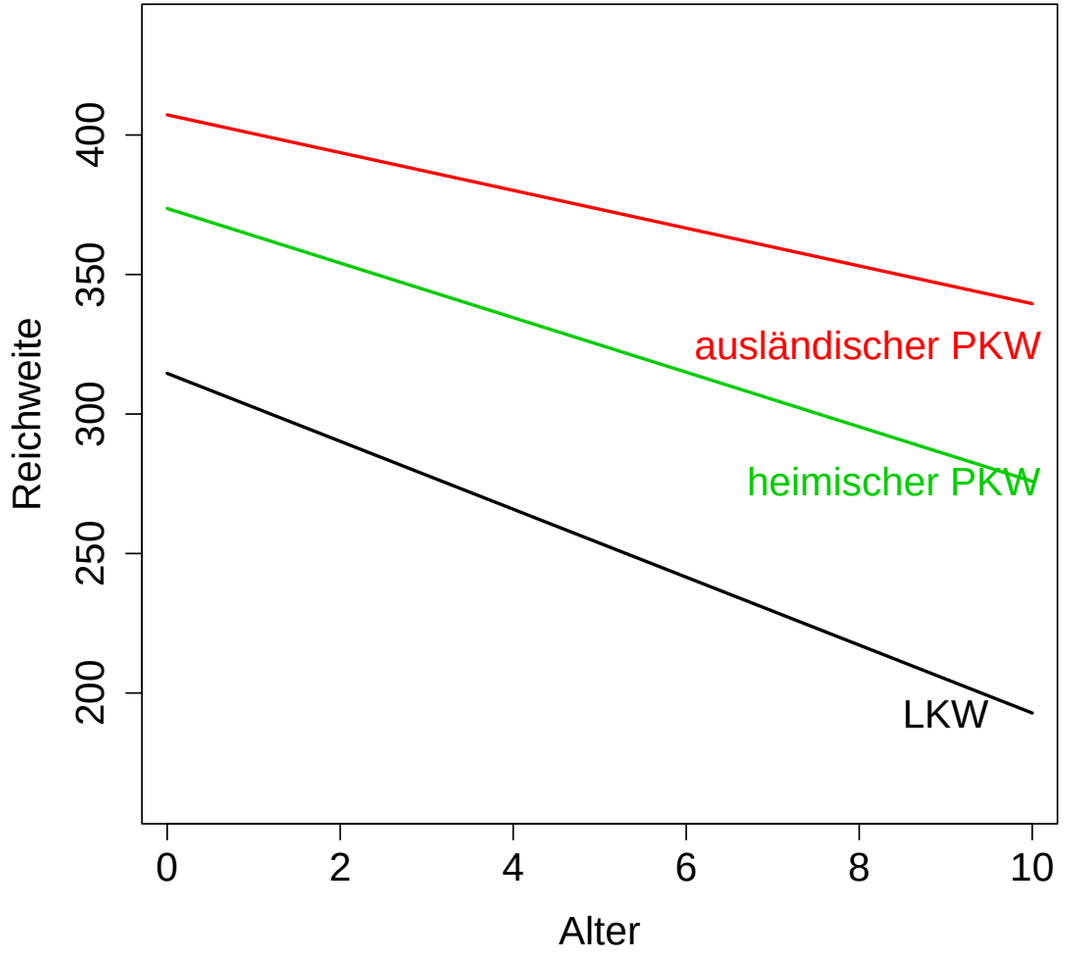

Reichweite
200
250
300
350
400
0
2
4
6
8
10
Alter
ausländischer PKW
heimischer PKW
LKW

Q: Wie könnten wir im Modell erlauben, dass jeder Fahrzeugtyp einen eigenen Intercept und Slope hat?

A: Gib Interaktionen dazu!

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3}$$

```
> lm(milage ~ age + x2 + x3 + x2:age + x3:age)
Coefficients:
(Intercept)          age           x2           x3       age:x2       age:x3
    314.566      -12.174       59.109       92.664        2.393        5.406
```



ausländisch:

$$\mathsf{E}(y_i) = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1$$

heimisch:

$$\mathsf{E}(y_i) = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1$$

LKW:

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_1$$

Mehr als 1 Qualitativer Prädiktor:

Beispiel: 100 UF Football Spiele

y_i = #Punkte erzielt vom UF Football Team in Spiel i
x_{i1} = #gewonnene Spiele des Gegners in dessen letzten 10 Spielen

Unterscheide zwischen Heim-/Auswärtsspiele und Tag-/Nachtspiele

$$x_{i2} = \begin{cases} 1 & \text{Heimspiel} \\ 0 & \text{Auswärtsspiel} \end{cases} \qquad x_{i3} = \begin{cases} 1 & \text{Tagspiel} \\ 0 & \text{Nachtspiel} \end{cases}$$

Modell: $\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$

Auswärts/Tag:	$\mathsf{E}(y_i) = (\beta_0 + \beta_3) + \beta_1 x_{i1}$
Auswärts/Nacht:	$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1}$

Man erzielt bei Auswärtsspielen um β_3 Punkte mehr am Tag als in der Nacht.

Heim/Tag: $\mathsf{E}(y_i) = (\beta_0 + \beta_2 + \beta_3) + \beta_1 x_{i1}$
Heim/Nacht: $\mathsf{E}(y_i) = (\beta_0 + \beta_2) + \beta_1 x_{i1}$

Man erzielt bei Heimspielen auch um β_3 Punkte mehr am Tag als in der Nacht.

Zusätzliche Interaktionen sind auch möglich!

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3}\,.$$

Beispiel – Häuserdaten:

y_i = Preis (in 1000 \$),
x_{i1} = Wohnfläche (in 1000 square feet)

$$x_{i2} = \begin{cases} 1 & \text{neu} \\ 0 & \text{gebraucht} \end{cases}$$

Ein Modell, das neuen und gebrauchten Häusern eigene Intercepts und Slopes erlaubt, ist

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}\,.$$

Untermodelle:

Neue Häuser: $\mathsf{E}(y_i) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{i1}$
Gebrauchte Häuser: $\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1}$

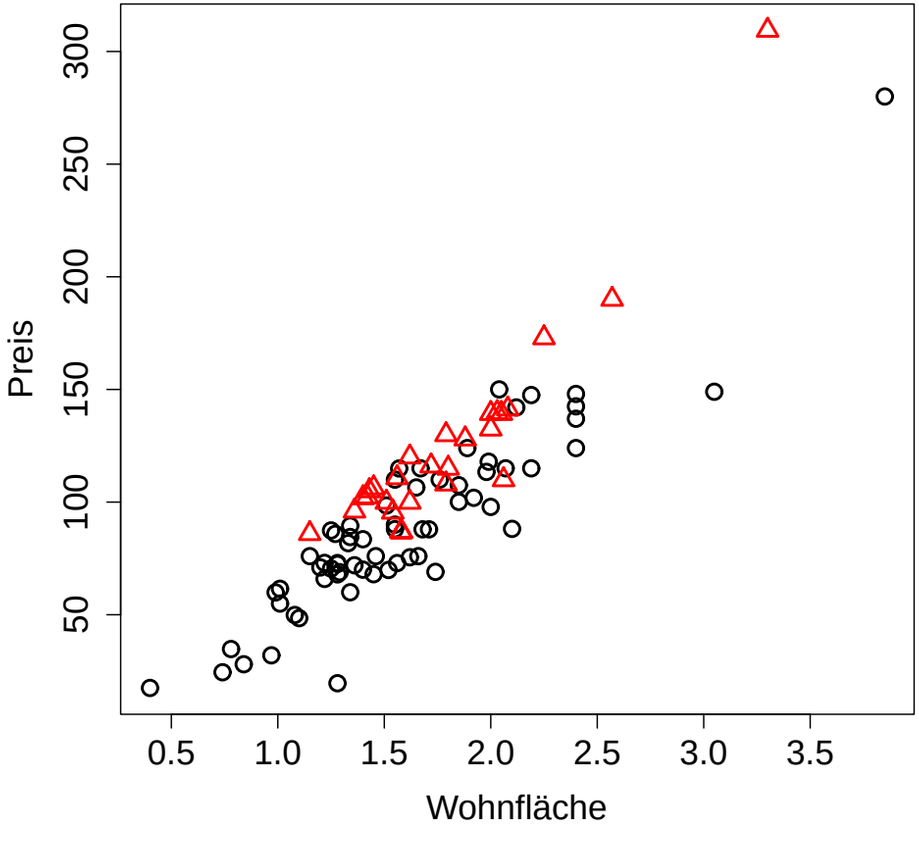
Wie kann man testen, ob die Regressionsgeraden dieselbe Steigung haben?

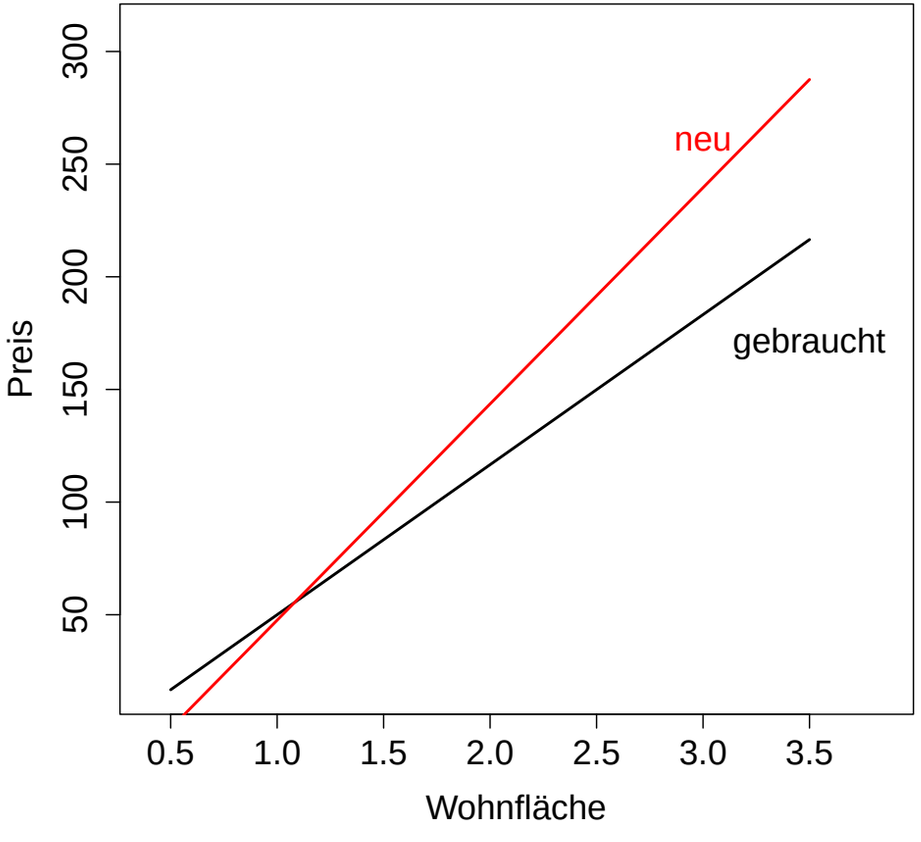
$H_0 : \beta_3 = 0$ gegen $H_1 : \beta_3 \neq 0$

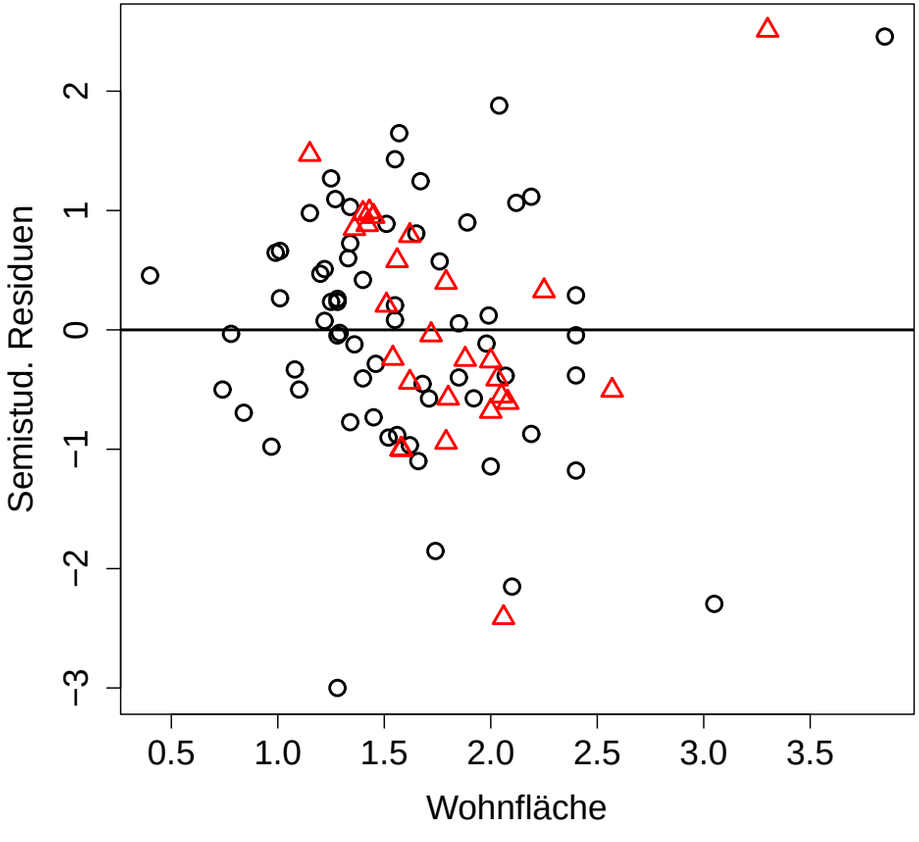
$$
\begin{aligned}
F &= \frac{\mathsf{SSR}(\texttt{area:new} | \texttt{area, new})/1}{\mathsf{SSE}(\texttt{area, new, area:new})/(n-4)} \\
T &= \frac{\hat{\beta}_3}{\sqrt{\mathsf{MSE} \cdot [(\mathbf{X}'\mathbf{X})^{-1}]_{4,4}}}
\end{aligned}
$$

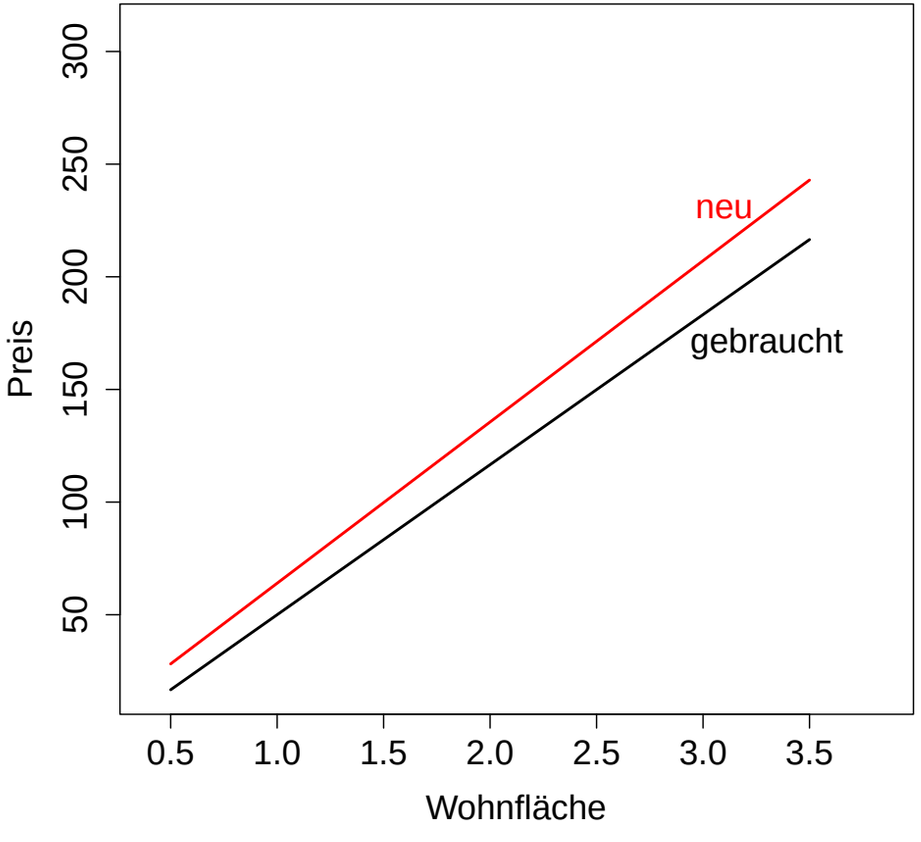
```
> attach(houses)
> summary(hm <- lm(price ~ area*new))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -16.600      6.210  -2.673 0.008944 **
area          66.604      3.694  18.033  < 2e-16 ***
new          -31.826     14.818  -2.148 0.034446 *
area:new      29.392      8.195   3.587 0.000547 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 16.35 on 89 degrees of freedom
Multiple R-squared: 0.8675,     Adjusted R-squared: 0.8631
F-statistic: 194.3 on 3 and 89 DF,  p-value: < 2.2e-16
```

```
> anova(hm)
Analysis of Variance Table

Response: price
          Df Sum Sq Mean Sq F value    Pr(>F)
area       1 145097  145097 542.722 < 2.2e-16 ***
new        1   7275    7275  27.210 1.178e-06 ***
area:new   1   3439    3439  12.865 0.0005467 ***
Residuals 89  23794     267
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```


Preis
50
100
150
200
250
300
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Wohnfläche


Preis
50
100
150
200
250
300
neu
gebraucht
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Wohnfläche


Semistud. Residuen
−3
−2
−1
0
1
2
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Wohnfläche


Preis
50
100
150
200
250
300
neu
gebraucht
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Wohnfläche

Vergleichen wir die beiden Modelle:

Modell 1: $\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i1} x_{i2}$

mit $x_{i2} = \begin{cases} 1 & \text{neu} \\ 0 & \text{gebraucht} \end{cases}$

Modell 2: $\mathsf{E}(y_i) = \beta_0^* + \beta_1^* x_{i1} + \beta_2^* x_{i2}^* + \beta_3^* x_{i1} x_{i2}^*$

mit $x_{i2}^* = \begin{cases} 1 & \text{gebraucht} \\ 0 & \text{neu} \end{cases}$

Parameter	Modell 1	Modell 2
Intercept für neu	$\beta_0 + \beta_2$	β_0^*
Intercept für gebraucht	β_0	$\beta_0^* + \beta_2^*$
Slope für neu	$\beta_1 + \beta_3$	β_1^*
Slope für gebraucht	β_1	$\beta_1^* + \beta_3^*$

Parameter	Modell 1	Modell 2
Intercept für neu	$\beta_0 + \beta_2$	β_0^*
Intercept für gebraucht	β_0	$\beta_0^* + \beta_2^*$
Slope für neu	$\beta_1 + \beta_3$	β_1^*
Slope für gebraucht	β_1	$\beta_1^* + \beta_3^*$

Somit sollte gelten:

$$\begin{aligned}
\hat{\beta}_0^* &= \hat{\beta}_0 + \hat{\beta}_2 \\
\hat{\beta}_1^* &= \hat{\beta}_1 + \hat{\beta}_3 \\
\hat{\beta}_2^* &= -\hat{\beta}_2 \\
\hat{\beta}_3^* &= -\hat{\beta}_3
\end{aligned}$$

Wir zeigen, dass dies tatsächlich der Fall ist!

$\mathbf{X}_{n\times 4}$ = design matrix for model 1
$\mathbf{X}^*_{n\times 4}$ = design matrix for model 2

Wir suchen die Matrix $\mathbf{M}_{4\times 4}$, so dass $\mathbf{X}^* = \mathbf{XM}$

$$\begin{bmatrix} 1 & x_{11} & 0 & 0 \\ 1 & x_{21} & 1 & x_{21} \\ 1 & x_{31} & 1 & x_{31} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & 1 & x_{11} \\ 1 & x_{21} & 0 & 0 \\ 1 & x_{31} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & 1 & x_{n1} \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix}$$

$$\begin{aligned} \hat{\boldsymbol{\beta}}^* &= (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{Y} = ((\mathbf{XM})'(\mathbf{XM}))^{-1}(\mathbf{XM})'\mathbf{Y} \\ &= (\mathbf{M}'\mathbf{X}'\mathbf{XM})^{-1}\mathbf{M}'\mathbf{X}'\mathbf{Y} = \left(\mathbf{M}^{-1}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{M}')^{-1}\right)\mathbf{M}'\mathbf{X}'\mathbf{Y} \\ &= \mathbf{M}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{M}^{-1}\hat{\boldsymbol{\beta}}\,. \end{aligned}$$

Es ist nun einfach zu zeigen, dass $\mathbf{M} = \mathbf{M}^{-1}$ gilt, und somit

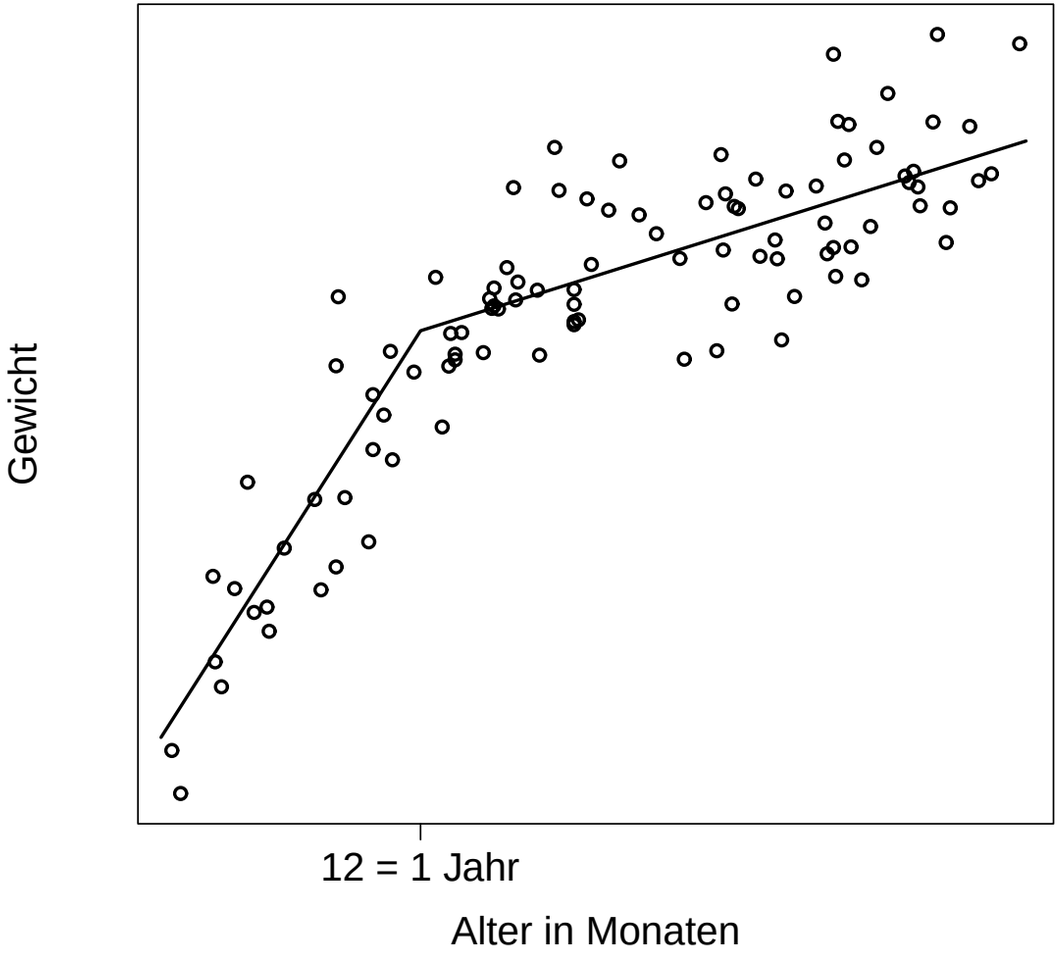
$$\begin{bmatrix} \hat{\beta}_0^* \\ \hat{\beta}_1^* \\ \hat{\beta}_2^* \\ \hat{\beta}_3^* \end{bmatrix} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_0 + \hat{\beta}_2 \\ \hat{\beta}_1 + \hat{\beta}_3 \\ -\hat{\beta}_2 \\ -\hat{\beta}_3 \end{bmatrix}.$$

Stückweise Lineare Regression

Beispiel:

y_i = Gewicht junger Hunde, $i = 1, \ldots, n$,
x_{i1} = Alter in Monaten.

Wir erwarten eine unterschiedliche Gewichtszunahme, für junge bzw. ausgewachsene Hunde. Ein Scatterplot könnte folgendermaßen aussehen:



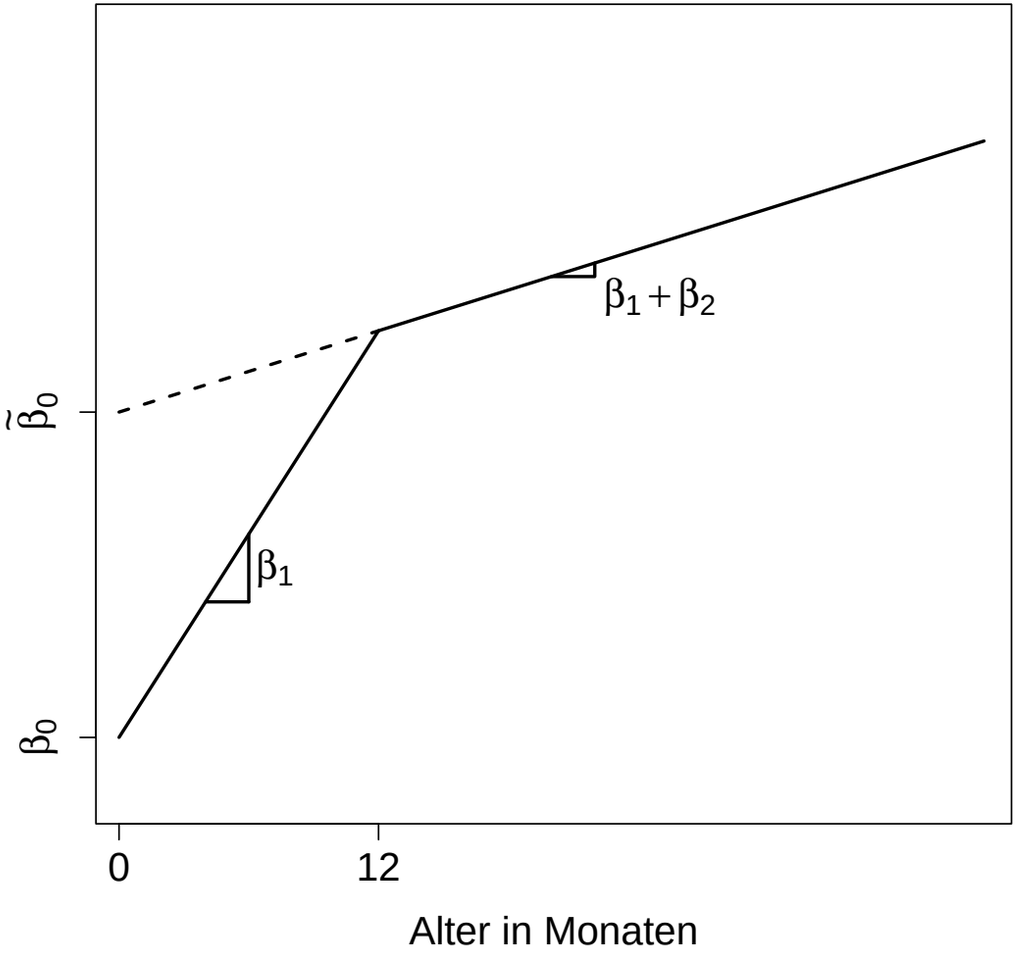
Wie würden wir diesen Typ von Daten modellieren?

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i1} - 12) x_{i2},$$

wobei

$$x_{i2} = \begin{cases} 1 & x_{i1} \geq 12, \\ 0 & x_{i1} < 12. \end{cases}$$

Das Alter von 12 Monaten wird **change point** genannt.



$x_{i1} < 12$:

$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1}$

$x_{i1} \geq 12$:

$\mathsf{E}(y_i) = \tilde{\beta}_0 + (\beta_1 + \beta_2) x_{i1}$

Aber beide Modelle müssen im change point denselben Wert ergeben, d.h.

$$\begin{aligned} \beta_0 + \beta_1 \cdot 12 &= \tilde{\beta}_0 + (\beta_1 + \beta_2) \cdot 12 \\ \tilde{\beta}_0 &= \beta_0 - 12\beta_2 \end{aligned}$$

Somit brauchen wir:

für $x_{i1} < 12$: $\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1}$

für $x_{i1} \geq 12$: $\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1} - 12\beta_2$.

Mit dem Altersindikator x_{i2} ist dies äquivalent zum Modell

$$\mathsf{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 (x_{i1} - 12) x_{i2}\,.$$

9. Diagnostics/Residuenanalyse

Ausreißer in x- und y-Richtung beeinflussen den LSE.

Hat-Matrix $\mathbf{H}$ ist zur Erkennung wichtig!

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' .$$

Geometrisch gilt $\hat{\boldsymbol{\mu}} = \mathbf{H}\mathbf{y}$. Mit

$$h_{ij} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j$$

folgt $\hat{\mu}_i = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{j\neq i} h_{ij}y_j$, also

$$\partial\hat{\mu}_i/\partial y_i = h_{ii} .$$

Die Diagonalelemente h_{ii} messen also den Einfluss von y_i auf $\hat{\mu}_i$.

Da $\mathbf{H}$ symmetrisch und idempotent, gilt

$$h_{ii} = \sum_{j=1}^{n} h_{ij}h_{ji} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$$

Dies bedeutet:

1. $0 \leq h_{ii} \leq 1$,
2. falls $h_{ii} = 0 \Rightarrow h_{ij} = 0, \quad \forall i, j$,
3. falls $h_{ii} = 1 \Rightarrow h_{ij} = 0, \quad \forall i, j, \ i \neq j$,
4. falls $h_{ii} = 0 \Rightarrow \hat{\mu}_i$ wird nicht von y_i beeinflusst,
5. falls $h_{ii} = 1 \Rightarrow \hat{\mu}_i = y_i$, d.h. das Modell liefert eine exakte Schätzung für y_i.

Da trace$(\mathbf{H}) = p$, folgt $\bar{h} = \frac{1}{n}\sum_i h_{ii} = p/n$. Falls alle h_{ii} gleich groß sind (d.h. $h_{ii} = \bar{h}$, $i = 1, \ldots, n$), spricht man von einem **D-optimalen Design**.

Für ein SLR $\mu(x) = \beta_0 + \beta_1 x$ gilt die Darstellung

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n} (x_k - \bar{x})^2}.$$

Für ein MLR gilt eine verallgemeinerte Darstellung

$$h_{ii} = \frac{1}{n} + \frac{1}{n-1}(\mathbf{x}_i^1 - \bar{\mathbf{x}}^1)'\mathbf{S}^{-1}(\mathbf{x}_i^1 - \bar{\mathbf{x}}^1)\,,$$

wobei $\mathbf{x}_i^1 = (x_{i1}, \ldots, x_{i,p-1})'$ die i-te Designzeile ohne Intercept bezeichnet.
$\mathbf{S}$ ist die empirische Varianz-Kovarianzmatrix dieser $\mathbf{x}_i^1$ Vektoren, d.h.

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i^1 - \bar{\mathbf{x}}^1)'(\mathbf{x}_i^1 - \bar{\mathbf{x}}^1)\,.$$

Somit gilt für ein Modell mit Intercept, dass $h_{ii} \geq 1/n$.

h_{ii} wächst mit der Distanz zwischen x_i und $\bar{x}$. Je größer h_{ii}, desto extremer liegt $\mathbf{x}_i$ im Designraum. Man nennt die i-te Beobachtung einen **high-leverage** Punkt (hat große Hebelwirkung), falls

$$h_{ii} > 2\bar{h} = 2\frac{p}{n}\,.$$

High-leverage Punkte sind extreme Punkte, deren $\hat{\mu}_i$ stark von y_i abhängen. Zur Identifizierung ist auch die euklidische- oder **Mahalanobisdistanz** verwendbar. Die Mahalanobisdistanz für die i-te Beobachtung ist definiert durch

$$MD_i^2 = (n-1)\left(h_{ii} - \frac{1}{n}\right)\,.$$

High-leverage Punkte müssen nicht **einflussreich** sein. Sie sind einflussreich, wenn ihre Elimination eine starke Änderung in den Schätzungen ergeben.

Auffinden solcher Beobachtungen wird erschwert, da diese sowohl einzeln als auch gemeinsam einflussreich sein können (**Masking-Effekt**).

Residuen

Gewöhnliche Residuen: r_i (beobachtbare Fehler) bei LS Schätzung

$$\mathbf{r} = \mathbf{y} - \hat{\boldsymbol{\mu}} = (\mathbf{I} - \mathbf{H})\mathbf{y}\,.$$

Für Residuenvektor $\mathbf{r}$ und nicht beobachtbaren Fehlervektor $\boldsymbol{\epsilon}$ gilt

$$\begin{aligned} \mathbf{r} &= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}\,, \\ r_i &= \epsilon_i - \sum_{j=1}^{n} h_{ij}\epsilon_j\,, \quad i = 1, \ldots, n\,. \end{aligned}$$

Die Beziehung zwischen $\mathbf{r}$ und $\boldsymbol{\epsilon}$ hängt also nur von $\mathbf{H}$ ab. Sind die h_{ij} **klein**, ist $\mathbf{r}$ *Ersatz* für $\boldsymbol{\epsilon}$.

Vergleiche

$$\begin{aligned} \mathsf{E}(\boldsymbol{\epsilon}) &= \mathbf{0} \qquad & \mathsf{var}(\boldsymbol{\epsilon}) &= \sigma^2 \mathbf{I} \\ \mathsf{E}(\mathbf{r}) &= \mathbf{0} \qquad & \mathsf{var}(\mathbf{r}) &= \sigma^2 (\mathbf{I} - \mathbf{H}) \,. \end{aligned}$$

r_i haben unterschiedliche Varianz. Genügt $\boldsymbol{\epsilon}$ einer Normalverteilung, so auch

$$\mathbf{r} \sim \mathsf{Normal}\left(\mathbf{0}, \sigma^2 (\mathbf{I} - \mathbf{H})\right) \,.$$

Für die nicht unabhängigen Residuen gilt

$$\begin{aligned} \mathsf{cov}(r_i, r_j) &= -\sigma^2 h_{ij} \\ \mathsf{cor}(r_i, r_j) &= -\frac{h_{ij}}{\sqrt{(1 - h_{ii})(1 - h_{jj})}} \,. \end{aligned}$$

Sie sind also negativ korreliert.

Standardisierte Residuen: Wegen $\mathrm{var}(r_i) = \sigma^2(1 - h_{ii})$ betrachten wir

$$r_i^* = \frac{y_i - \hat{\mu}_i}{S\sqrt{1 - h_{ii}}}$$

mit

$$\mathsf{E}(r_i^*) = 0 \quad \text{und} \quad \mathrm{var}(r_i^*) = 1\,.$$

y_i ist ein *potentieller Ausreißer*, falls

$$|r_i^*| > 2\sqrt{\mathrm{var}(r_i^*)} = 2\,.$$

Standardisierte Residuen sollten nicht mit den *studentisierten Residuen* verwechselt werden. Da r_i und S^2 nicht unabhängig sind, stammen die r_i^* nicht aus einer Student-t-Verteilung.

Deletion (Jackknife) Residuen

Frage: Wie ändert sich das Residuum von y_i, wenn die i-te Beobachtung gar nicht verwendet wird? Schließe so auf Einfluss der i-ten Beobachtung auf Schätzungen.

Falls i-te Beobachtung nicht verwendet, gilt

$$\begin{aligned}
\mathbf{X}'_{(i)}\mathbf{X}_{(i)} &= \mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}'_i \\
\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1} &= (\mathbf{X}'\mathbf{X})^{-1} + \frac{1}{1-h_{ii}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\,,
\end{aligned}$$

sowie

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{(i)} &= \left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)} = \left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1}(\mathbf{X}'\mathbf{y} - \mathbf{x}_i\mathbf{y}_i) \\
&= \hat{\boldsymbol{\beta}} - \frac{r_i}{1-h_{ii}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\,.
\end{aligned}$$

$$\begin{aligned} \mathsf{SSE} &= (n-p)S^2 = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \\ \mathsf{SSE}_{(i)} &= (n-1-p)S^2_{(i)} \\ &= \mathsf{SSE} - r_i^{*2}S^2 = S^2(n-p-r_i^{*2})\,. \end{aligned}$$

Die **Deletion-Residuen** sind definiert als

$$t_i = \frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{(i)}}{S_{(i)}\sqrt{1+\mathbf{x}_i'\left(\mathbf{X}_{(i)}'\mathbf{X}_{(i)}\right)^{-1}\mathbf{x}_i}} = \frac{r_i}{S_{(i)}\sqrt{1-h_{ii}}}\,,$$

weil $\mathbf{x}_i'\left(\mathbf{X}_{(i)}'\mathbf{X}_{(i)}\right)^{-1}\mathbf{x}_i = h_{ii}/(1-h_{ii})$ gilt.

Dies ist sinnvoll, denn

$$\mathsf{var}(y_i - \hat{\mu}_{i(i)}) = \sigma^2 \left(1 + \mathbf{x}_i' \left(\mathbf{X}_{(i)}' \mathbf{X}_{(i)}\right)^{-1} \mathbf{x}_i\right) = \frac{\sigma^2}{1 - h_{ii}}\,.$$

σ^2 wird hier durch $S^2_{(i)}$ geschätzt. Der Unterschied zwischen den standardisierten und den Deletion-Residuen besteht somit in der Schätzung von σ^2.

Betrachten wir nun die Quadratsummenzerlegung

$$\mathsf{SSE} = \mathsf{SSE}_{(i)} + \frac{r_i^2}{1 - h_{ii}}\,,$$

so folgt die Unabhängigkeit von $\mathsf{SSE}_{(i)}$ und r_i^2. Bei Normalverteilungsannahme ist weiters $\mathsf{SSE}_{(i)} \sim \sigma^2 \chi^2_{n-1-p}$. Da $\mathsf{E}\left(\frac{r_i}{\sigma\sqrt{1-h_{ii}}}\right) = 0$ und $\mathsf{var}\left(\frac{r_i}{\sigma\sqrt{1-h_{ii}}}\right) = 1$, folgt $\frac{r_i}{\sigma\sqrt{1-h_{ii}}} \sim \mathsf{Normal}(0,1)$ und damit $\frac{r_i^2}{\sigma^2(1-h_{ii})} \sim \chi^2_1$.

Mit diesen beiden χ^2-verteilten Größen wird das Deletion-Residuum gebildet, denn

$$\frac{\frac{r_i^2}{\sigma^2(1-h_{ii})}/1}{\frac{\mathsf{SSE}_{(i)}}{\sigma^2}/(n-1-p)} = \frac{r_i^2}{S_{(i)}^2(1-h_{ii})} = t_i^2\,.$$

Weiters gilt daher

$$t_i^2 \sim F_{1,n-1-p}\,, \quad \text{bzw.} \quad t_i \sim t_{n-1-p}$$

mit

$$\mathsf{E}(t_i) = 0 \quad \text{und} \quad \mathsf{var}(t_i) = \frac{n-1-p}{n-3-p}\,.$$

Die Varianz ist also konstant (unabhängig von i). Allerdings sind die Zufallsvariablen t_i und t_j mit $i \neq j$ *nicht unabhängig*.

Eine andere Schreibweise der t_i liefert Informationen über die Beziehung zwischen standardisierten und Deletion-Residuen

$$t_i = r_i^* \sqrt{\frac{n-1-p}{n-p-r_i^{*2}}} \, .$$

Daraus ist ersichtlich, dass t_i^2 eine monotone, nichtlineare Transformation der r_i^{*2} ist. Weisberg bezeichnet y_i als Ausreißer zum Niveau α, falls

$$|t_i| \geq t_{n-1-p;1-\alpha/(2n)} \, .$$

Dieses auf die Bonferroni-Ungleichung basierende Kriterium ist zwar sicherlich sehr konservativ, berücksichtigt aber das theoretisch durchzuführende n-*fache Testen*.

Distanzanalyse

Residuen (gewöhnliche, standardisierte, Deletion) dienen der Anzeige auffälliger y-Werte.

Nun: Maße für den Einfluss einzelner Beobachtungen auf die Schätzer

DFBETA

$$\text{DFBETA}_i = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{r_i}{1 - h_{ii}} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i$$

DFSIGMA

$$\text{SSE}_{(i)} = \text{SSE} - \frac{r_i^2}{1 - h_{ii}}$$

DFFIT

$$\text{DFFIT}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}} - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{(i)} = r_i \frac{h_{ii}}{1 - h_{ii}}$$

DFBETAS

$$\text{DFBETAS}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{S_{(i)}\sqrt{(\mathbf{X}'\mathbf{X})^{-1}_{jj}}}$$

DFFITS

$$\text{DFFITS}_i = \frac{r_i}{S_{(i)}\sqrt{1-h_{ii}}}\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

COVRATIO

$$\text{COVRATIO}_i = \frac{\det(S^2_{(i)}(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1})}{\det(S^2\ (\mathbf{X}'\mathbf{X})^{-1})}$$

Cook-Distanz

Der Einfluss der i-ten Beobachtung auf die Schätzung von $\boldsymbol{\beta}$ wird gemessen durch

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})/p}{\mathsf{SSE}(\hat{\boldsymbol{\beta}})/(n-p)}\,, \qquad i = 1, \ldots, n\,.$$

Ein großer Wert von D_i signalisiert starken Einfluss. Eine Beobachtung ist einflussreich, wenn $D_i > 1/2$.

Cook (1977) misst die "Distanz" zwischen $\hat{\boldsymbol{\beta}}$ und $\hat{\boldsymbol{\beta}}_{(i)}$ über die F-Statistik zur "Hypothese", dass $\beta_j = \hat{\beta}_{j(i)}$, $j = 0, \ldots, p-1$. Wegen $D_i \sim F_{p,n-p}$, vergleiche man D_i mit den Quantilen der F-Verteilung.

Substituiert man $\hat{\boldsymbol{\beta}}_{(i)}$ in D_i, so resultiert

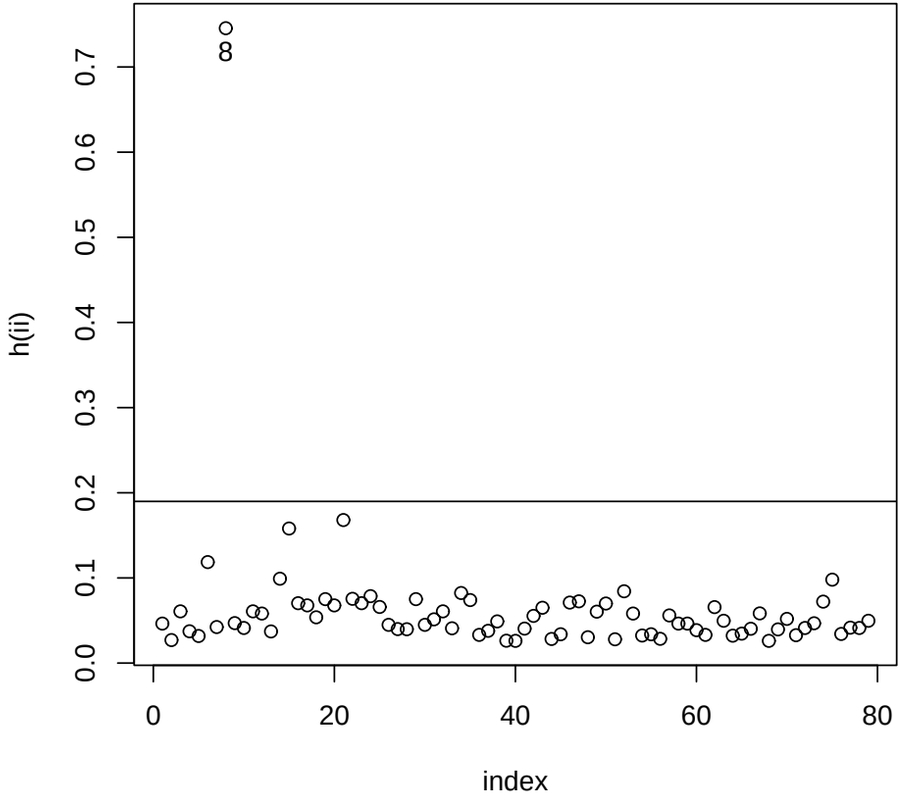
$$D_i = \frac{r_i^2 \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i}{S^2(1-h_{ii})^2 p} = \frac{r_i^2 h_{ii}}{S^2(1-h_{ii})^2 p} = \frac{r_i^{*2}}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right).$$

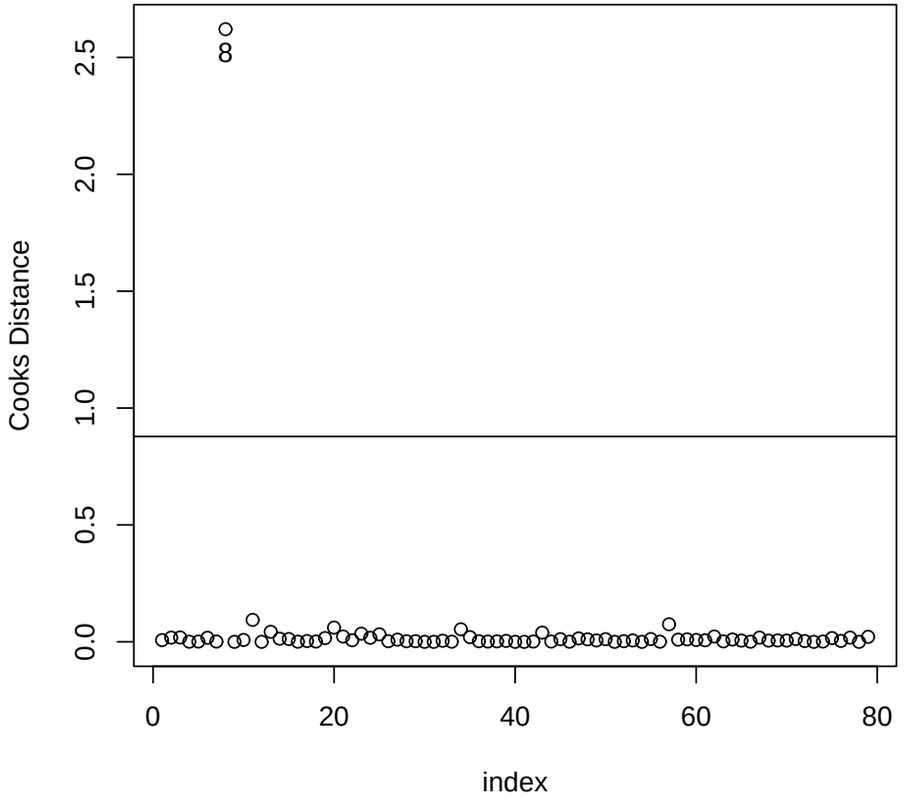
Die **Cook-Distanz** hängt nur von r_i^* und h_{ii} ab. D_i ist groß, falls i-te Beobachtung ein high-leverage point ist, oder das Residuum groß ist, oder beides vorliegt.

Angewandte Diagnostics

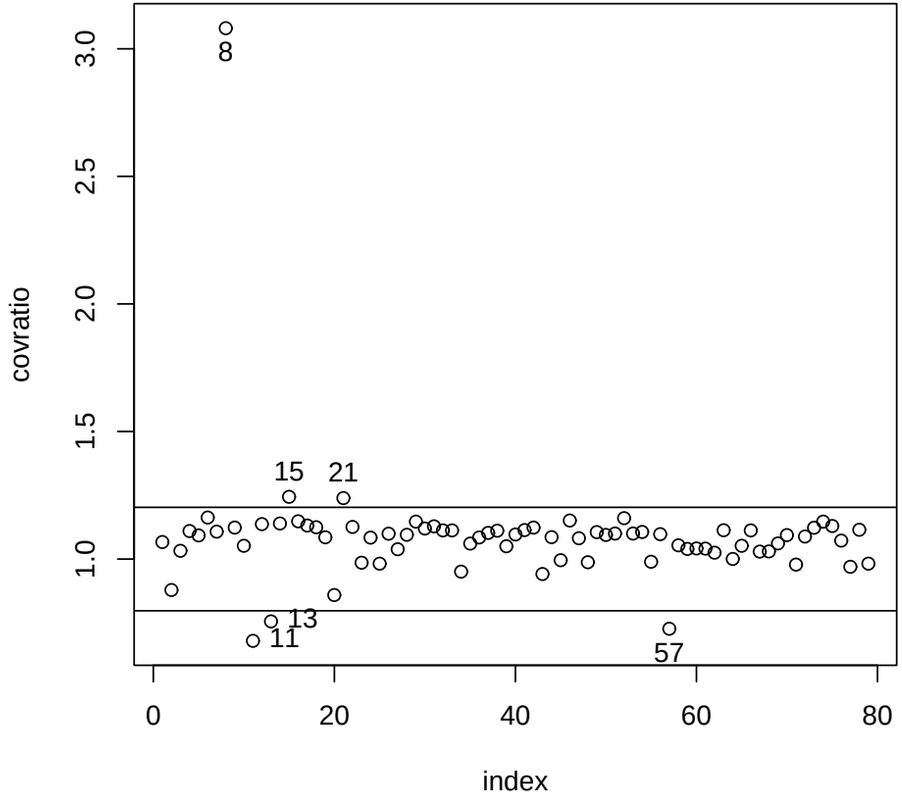
```
> summary(inflm.mod <- influence.measures(mod))
  Potentially influential observations of
          lm(formula = vc ~ height + age + I(age^2) + I(age^3)) :

    dfb.1_  dfb.hght dfb.age dfb.I(^2 dfb.I(^3 dffit   cov.r   cook.d  hat
 8   1.29_*  0.25    -2.21_*  2.45_*  -2.70_*  -3.71_*  3.08_* 2.62_*  0.75_*
11   0.09    0.15    -0.28    0.27    -0.24     0.72    0.68_* 0.09    0.06
13  -0.17    0.25    -0.02   -0.01     0.04    -0.48    0.76_* 0.04    0.04
15  -0.20    0.18     0.06   -0.05     0.04    -0.24    1.24_* 0.01    0.16
21   0.07   -0.26     0.16   -0.14     0.12    -0.33    1.24_* 0.02    0.17
57  -0.51    0.36     0.31   -0.31     0.30     0.63    0.73_* 0.07    0.06

> which(apply(inflm.mod$is.inf, 1, any))
 8 11 13 15 21 57
 8 11 13 15 21 57
```

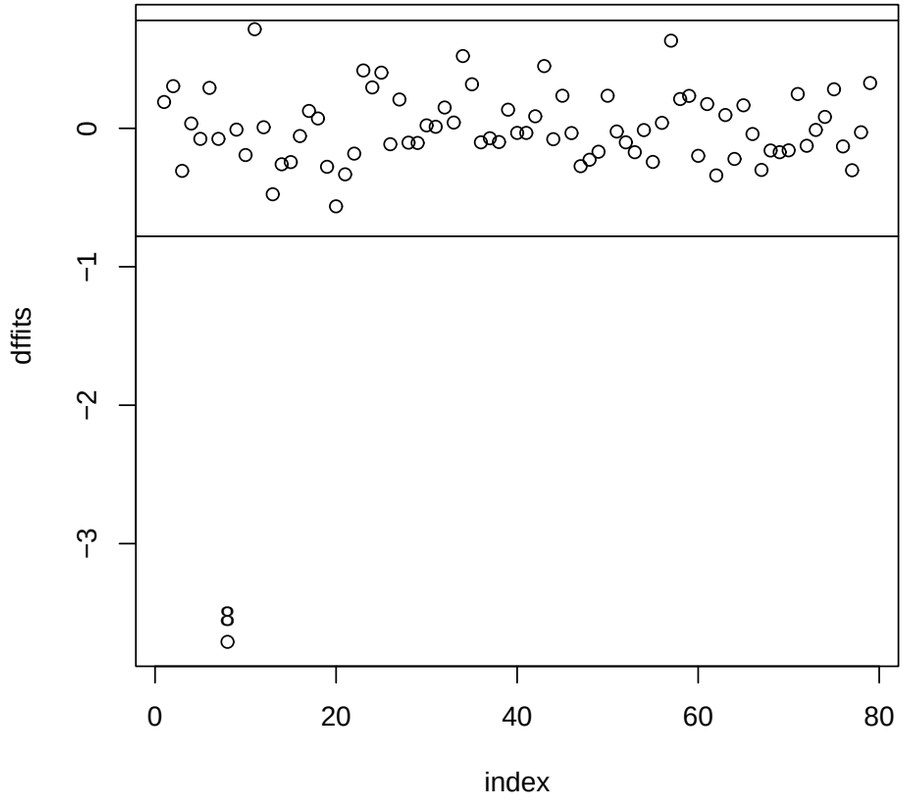
```
> h <- lm.influence(mod)$hat
> n <- length(mod$fitted)
> plot(1:n, h, xlab="index", ylab="h(ii)")
> abline(h = 3*5/79);  identify(1:n, h)

> c <- cooks.distance(mod) # pf(c, p, n - p) > 0.5
> pf(c, 5, 74) > 0.5       # obs nr 8 only
> plot(1:n, c, xlab="index", ylab="Cooks Distance")
> abline(h = qf(0.5, 5, 74)); identify(1:n, c)
```


h(ii)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
8
0
20
40
60
80
index

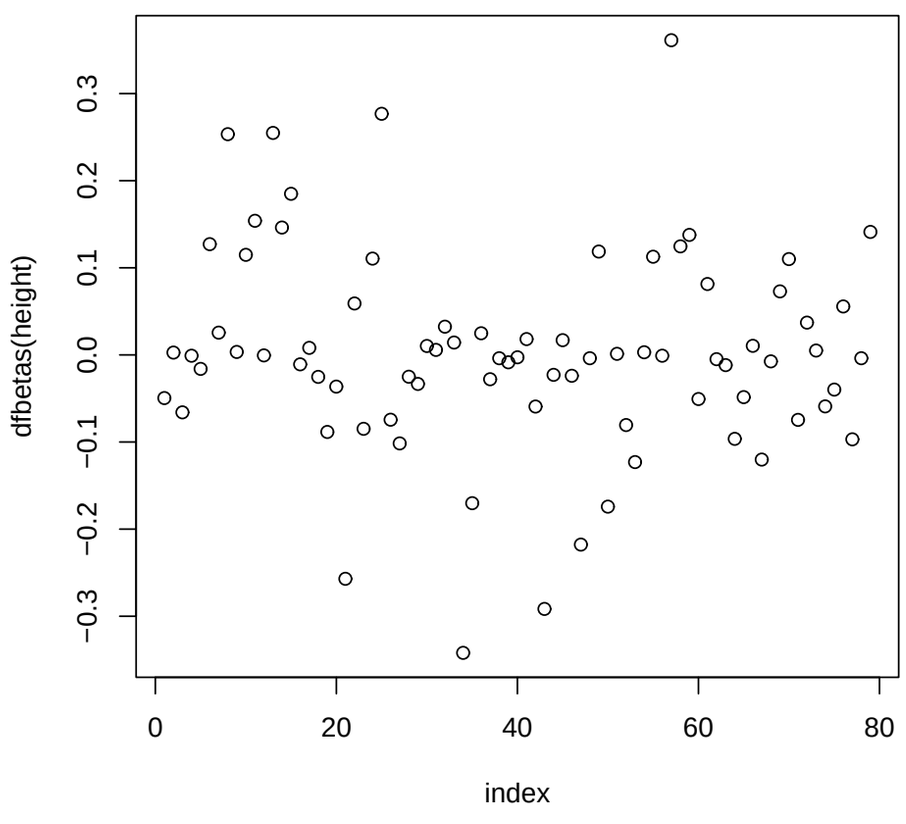

Cooks Distance
0.0
0.5
1.0
1.5
2.0
2.5
8
0
20
40
60
80
index

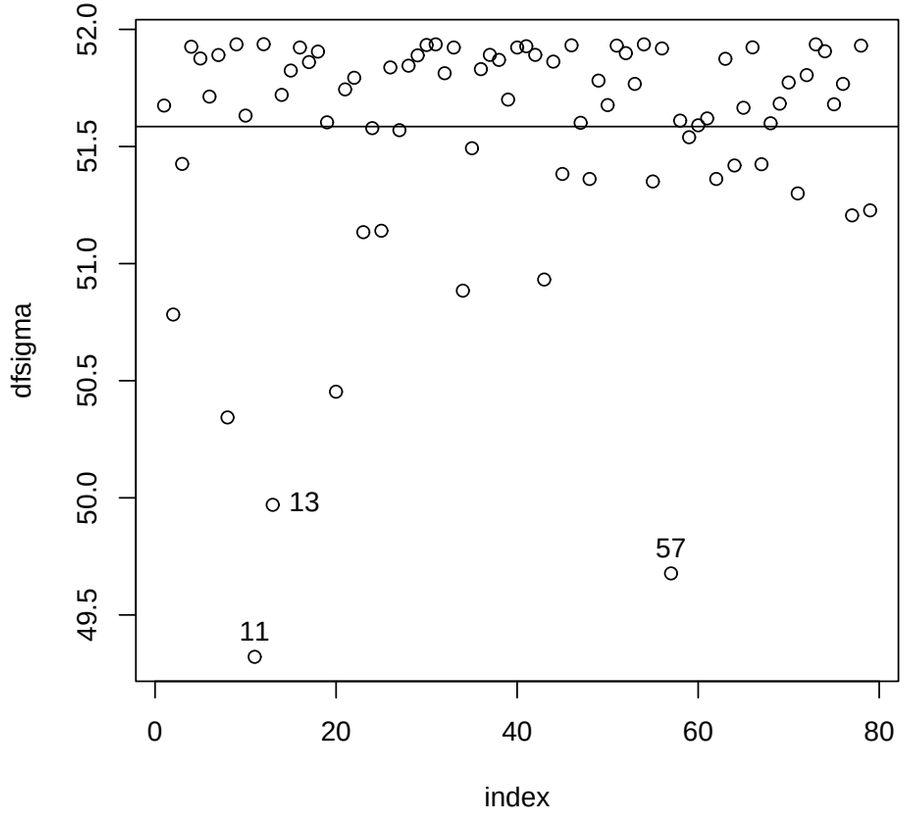
```
> cov.r <- covratio(mod) # 3p/(n - p)
> plot(1:n, cov.r, xlab="index", ylab="covratio")
> abline(h = c(1 - 3*5/74, 1 + 3*5/74)); identify(1:n, cov.r)

> dffits <- dffits(mod) # > 3*sqrt(p/(n - p))
> plot(1:n, dffits, xlab="index", ylab="dffits")
> abline(h = c(+3*sqrt(5/74), -3*sqrt(5/74))); identify(1:n, dffits)
```


covratio
3.0
2.5
2.0
1.5
1.0
8
15
21
13
11
57
0
20
40
60
80
index

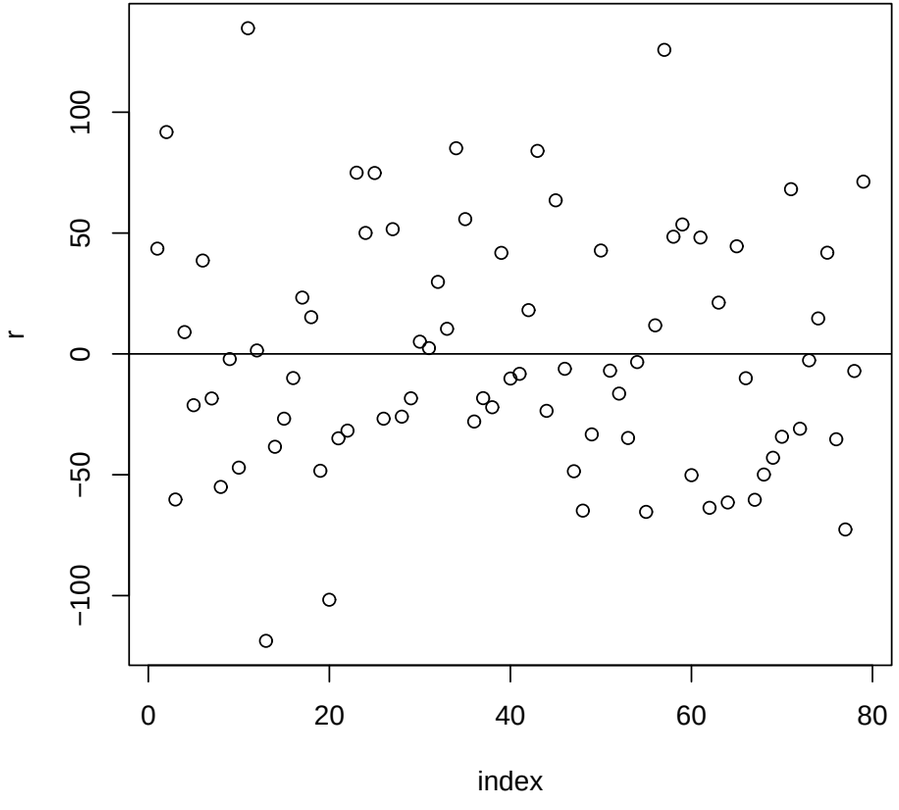

dffits
0
−1
−2
−3
8
0
20
40
60
80
index

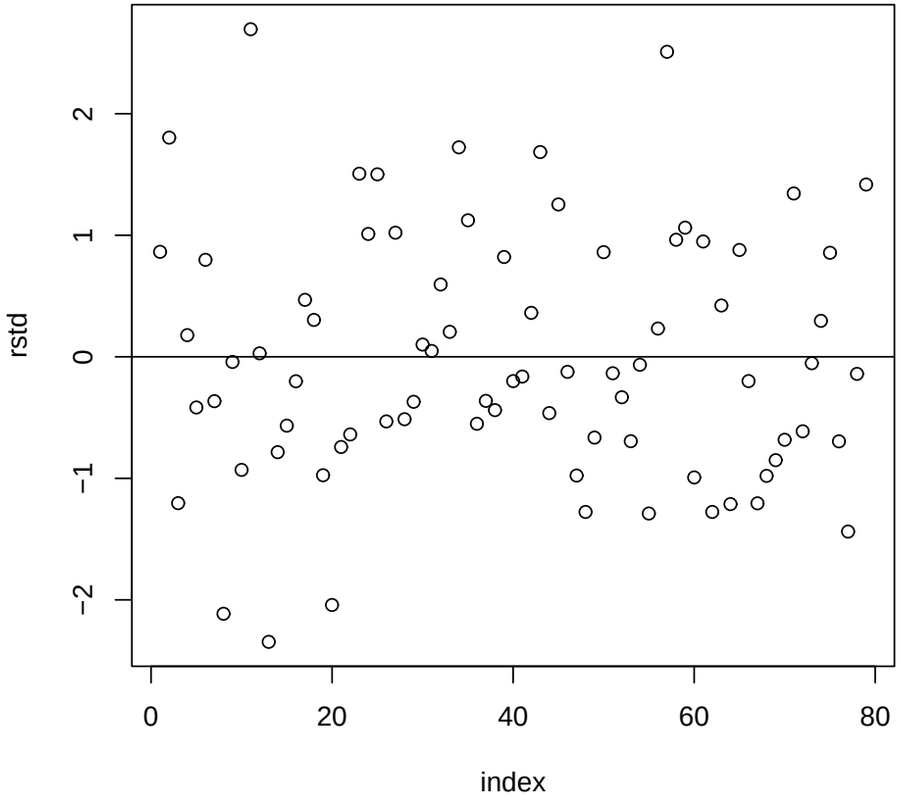
```
> dfbs <- dfbetas(mod) # > 1
> dfbs
     (Intercept)        height           age      I(age^2)      I(age^3)
 1   0.101258665 -0.0495899609 -0.090844379  0.0971073825 -0.0987564488
 2  -0.089469667  0.0026368428  0.104212527 -0.0815388812  0.0593050671
 3  -0.036833659 -0.0659490671  0.120175549 -0.1152957392  0.1011026054
 4  -0.017335527 -0.0009253076  0.024847735 -0.0239820293  0.0225516747
 :
> plot(1:n, dfbs[ , 2], xlab="index", ylab="dfbetas(height)")
> abline(h = c(1,-1)); identify(1:n, dfbs[,2])

> dfsigma <- lm.influence(mod)$sigma  # MSE_(i)
       1        2        3        4        5        6        7
51.67445 50.78282 51.42537 51.92606 51.87597 51.71320 51.89032
> plot(1:n, dfsigma, xlab="index", ylab="dfsigma")
> abline(h = summary(mod)$sigma); identify(1:n, dfsigma)
```

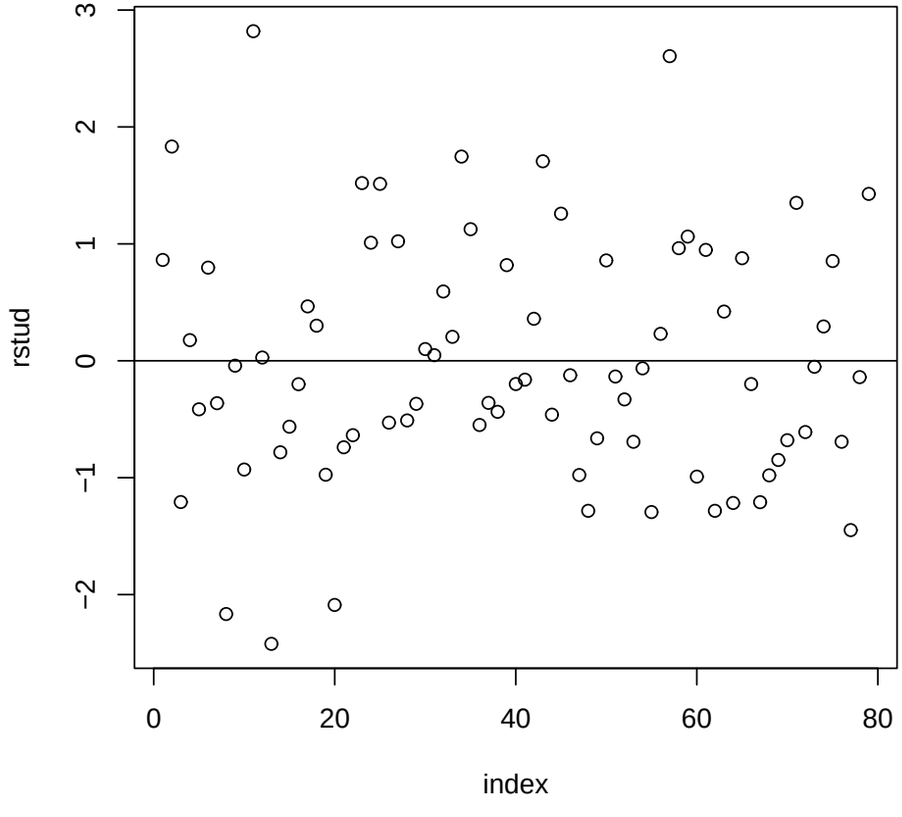

dfbetas(height)
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
0
20
40
60
80
index


dfsigma
52.0
51.5
51.0
50.5
50.0
49.5
13
57
11
0
20
40
60
80
index

```
> r <- residuals(mod)    # raw residuals
> plot(1:n, r, xlab="index", ylab="r"); abline(h = 0)

> rstd <- rstandard(mod) # stand. residuals
> plot(1:n, rstd, xlab="index", ylab="rstd")
> abline(h = c(-3, 0, +3)); identify(1:n, rstd)

> rstud <- rstudent(mod) # deletion resididuals
> plot(1:n, rstud, xlab="index", ylab="rstud")
> abline(h = c(0, qt(1-0.05/158, 73))); identify(1:n, rstud)
```


r
index
100
50
0
−50
−100
0
20
40
60
80


rstd
index
2
1
0
−1
−2
0
20
40
60
80


rstud
3
2
1
0
−1
−2
0
20
40
60
80
index

10. Nichtparametrische (glatte) Regressionsmodelle

Für Daten der Form $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ suchen wir einen nichtparametrischen Schätzer der Regressionsfunktion $g(x)$ unter dem Modell

$$y = g(x) + \epsilon \,.$$

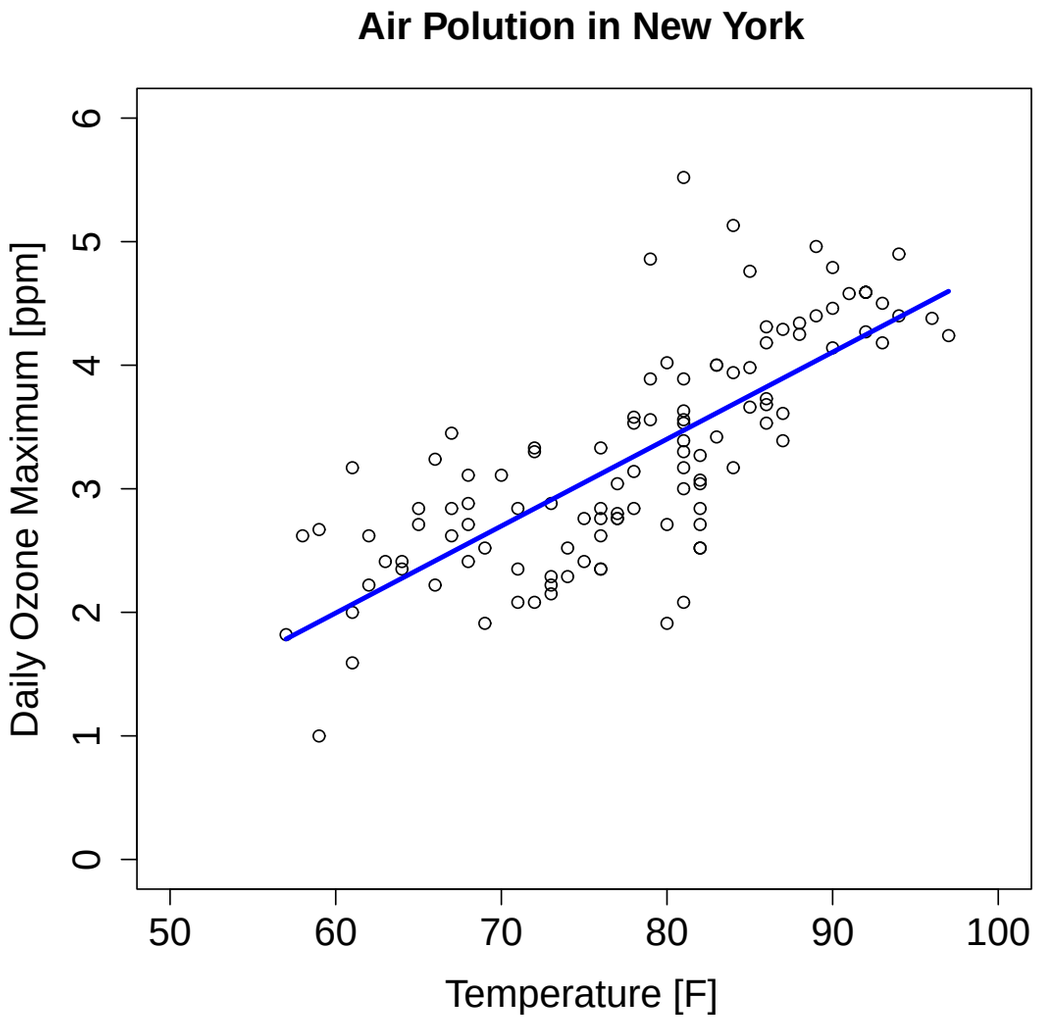
Hierbei sei $g(x)$ eine glatte Funktion in x und der Störterm ϵ genüge den gewöhnlichen Bedingungen wie im klassischen linearen Modell.

Parametrische versus Nichtparametrische Regression:

1. Natur der parametrischen Statistik: reduziere unbekannte (möglicherweise komplizierte) Funktion auf eine einfache Form mit geringer Anzahl unbekannter Parameter.

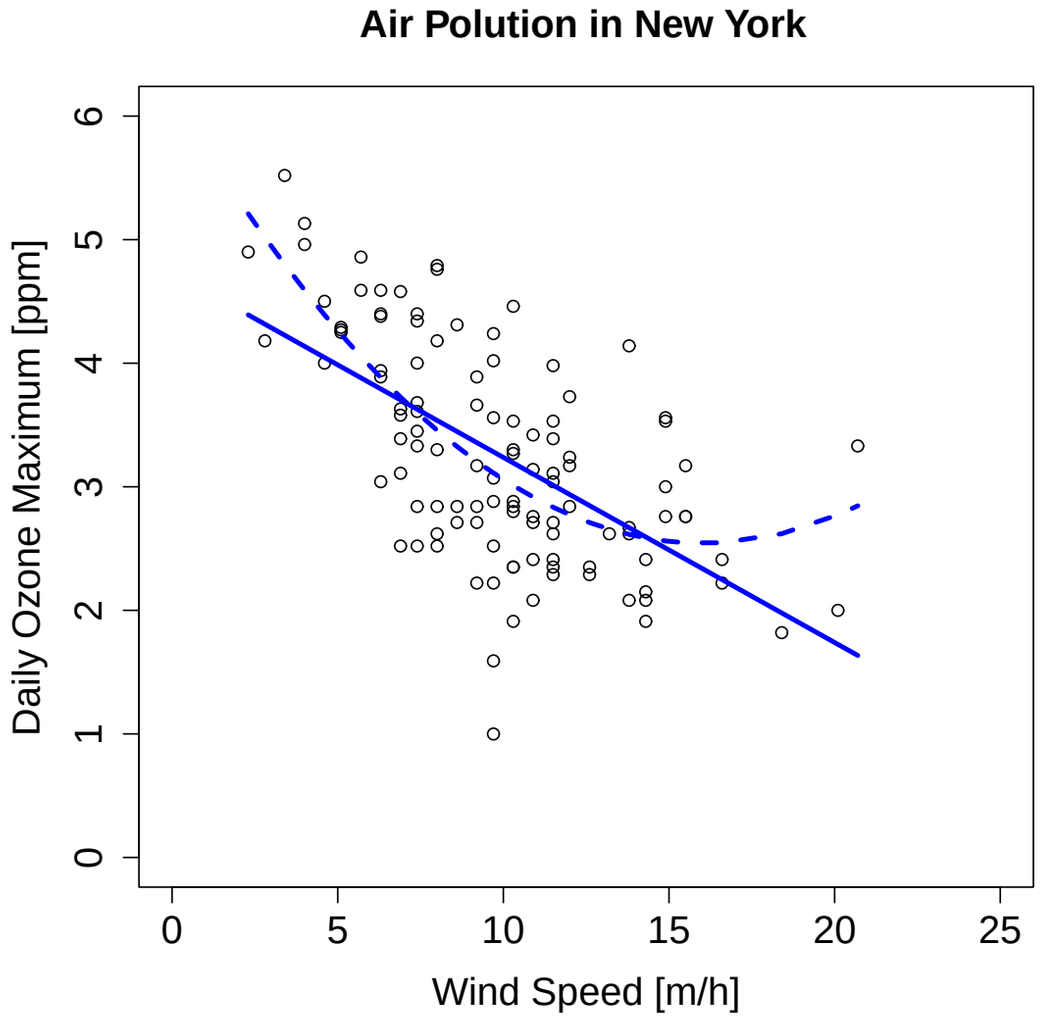
2. Im Gegensatz dazu macht man beim nichtparametrischen Ansatz so wenig wie möglich Annahmen über die Regressionsfunktion $g(\cdot)$.

3. Stattdessen werden wir die Daten so intensiv wie möglich einsetzen um über die mögliche Gestalt von $g(\cdot)$ zu lernen, und erlauben $g(\cdot)$ sehr flexibel zu sein (jedoch **glatt**).


Air Polution in New York
Daily Ozone Maximum [ppm]
0
1
2
3
4
5
6
50
60
70
80
90
100
Temperature [F]

Linearer Zusammenhang

```
> air <- read.table("air.txt", header=TRUE)
> attach(air)

> # construct matrix with columns 1 and x
> X <- outer(temperature, 0:1, "^")

> # do regression
> fit <- lsfit(X, ozone, intercept=F)
> beta <- fit$coef

> # compute fitted values
> mu <- X %*% beta
```


Air Polution in New York
Daily Ozone Maximum [ppm]
Wind Speed [m/h]
0
1
2
3
4
5
6
0
5
10
15
20
25

Gekrümmter Zusammenhang

- Linearer Fit zu einfach? Gib höhere Potenzen von x in das Modell:

$$\mu_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots = \sum_j \beta_j B_j(x)$$

- Weitere Spalten in der Matrix $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 & \dots & x_1^k \\ 1 & x_2 & x_2^2 & x_2^3 & \dots & x_2^k \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 & \dots & x_n^k \end{bmatrix}$$

- Regressions-Gleichungen: $\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^t\mathbf{y} \Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$

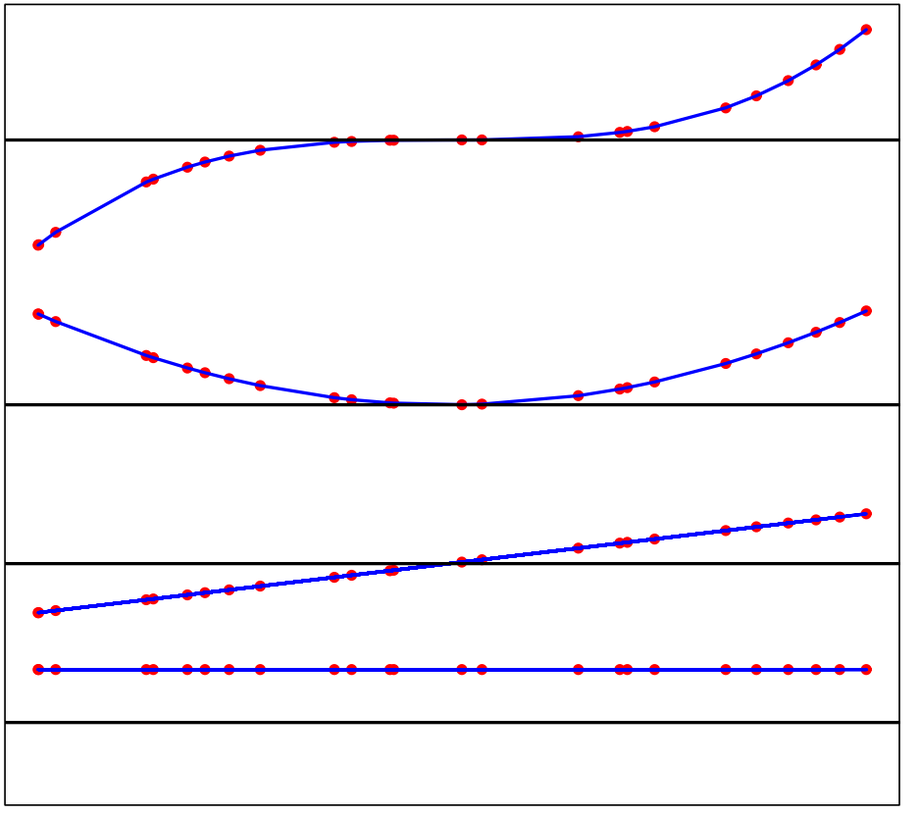
Gekrümmter Zusammenhang

```
> # construct matrix with columns of powers of x
> k <- 3
> X <- outer(wind, 0:k, "^")

> # do regression
> fit <- lsfit(X, ozone, intercept=F)
> beta <- fit$coef

> # compute fitted values
> mu <- X %*% beta
```

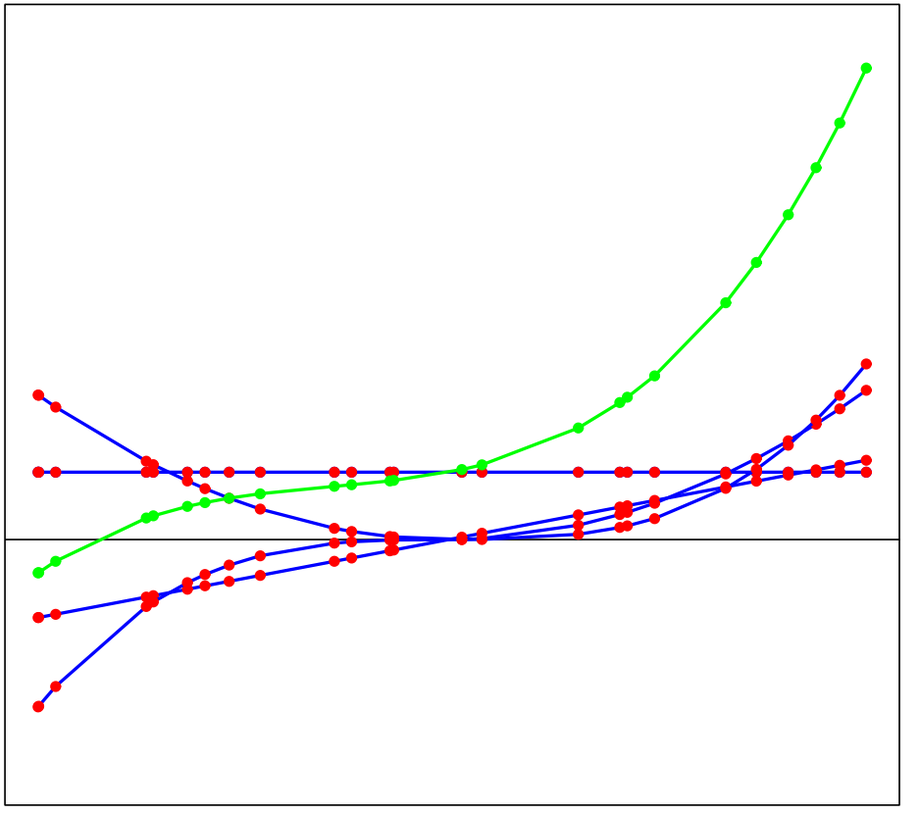
Basis Funktionen

- Regressionsmodell $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$
- Spalten von $\mathbf{X}$: Basis Funktionen (polynomiale Basis)
- Mit sortierten x schöne visuelle Representation

Cubic polynomial basis



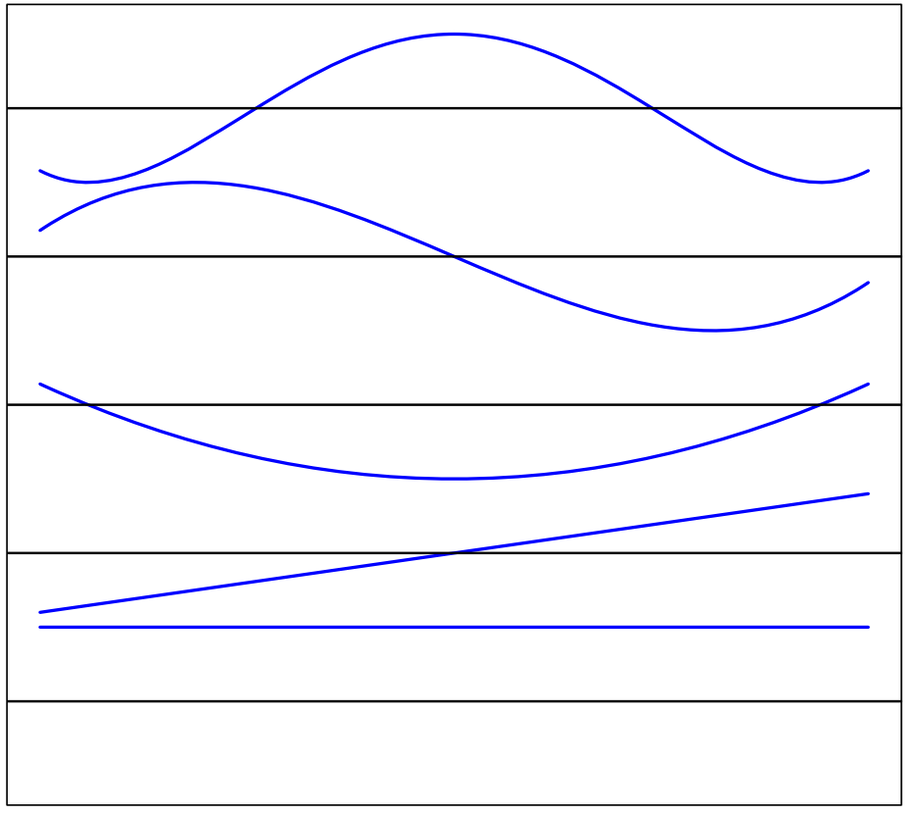
Weighted sum of cubic polynomial basis



Numerische Aspekte

- Polynome höheren Grades sind numerisch instabil
- Rundungsprobleme mit $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$
- Teillösung: Zentrieren und normalisieren von x
- Besser: verwende orthogonale Polynome
- Einfach sind Chebyshev Polynome: $C(x;k) = \cos[k \arccos(x)]$

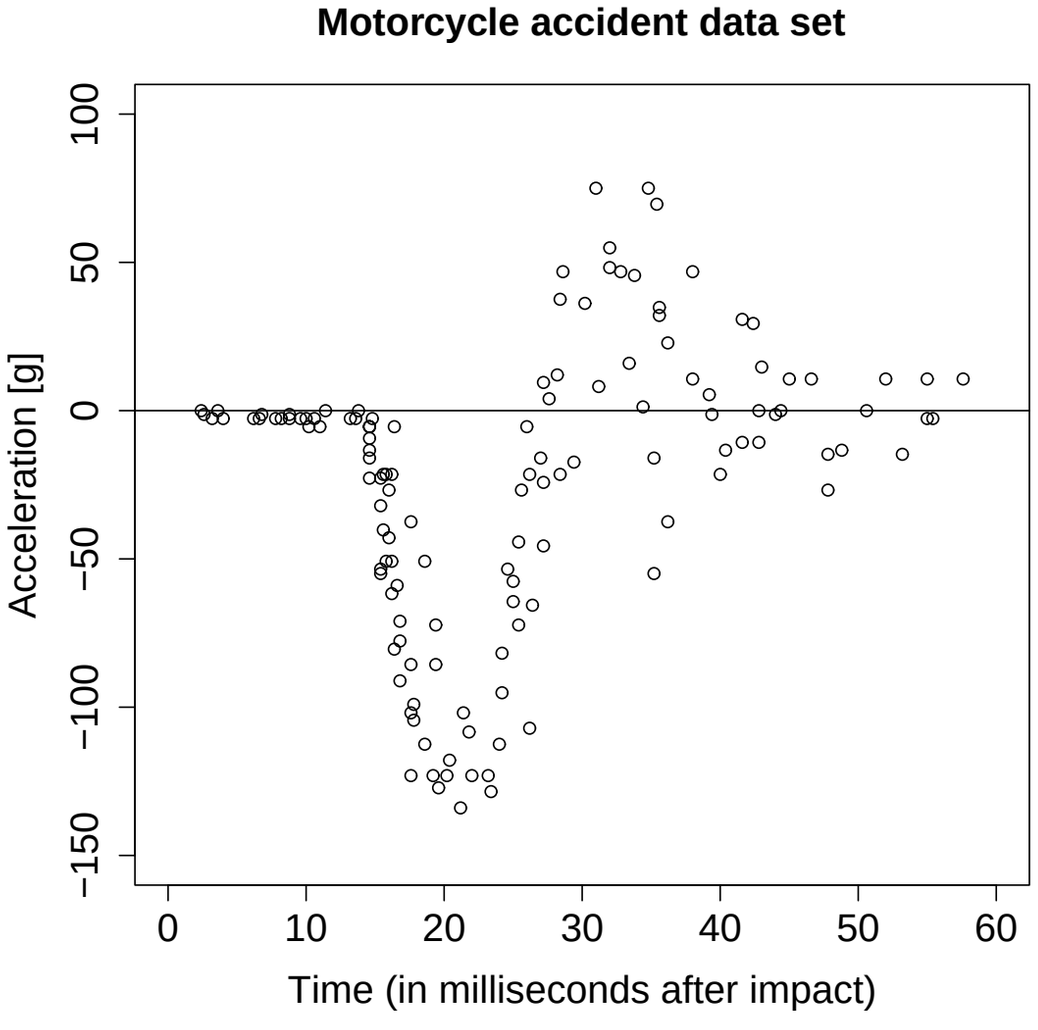
Basis of Chebychev polynomials

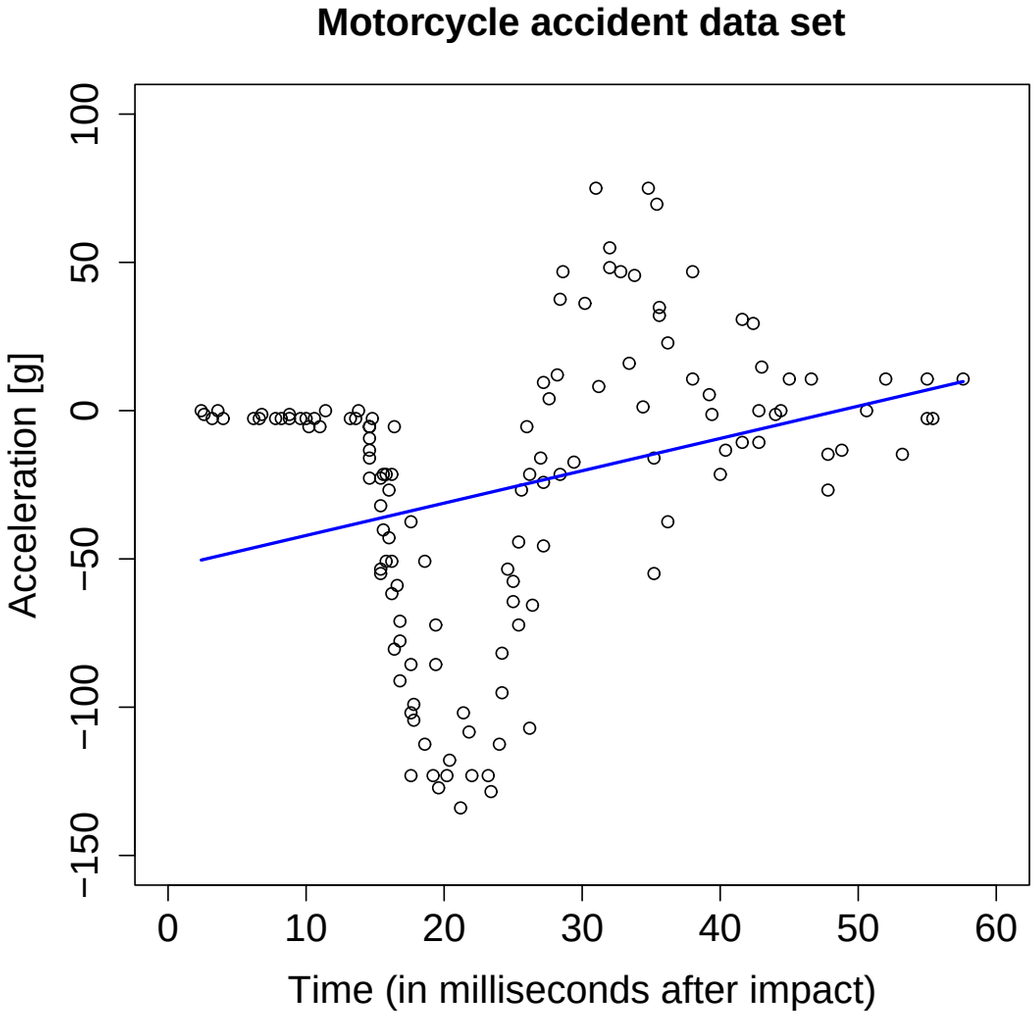


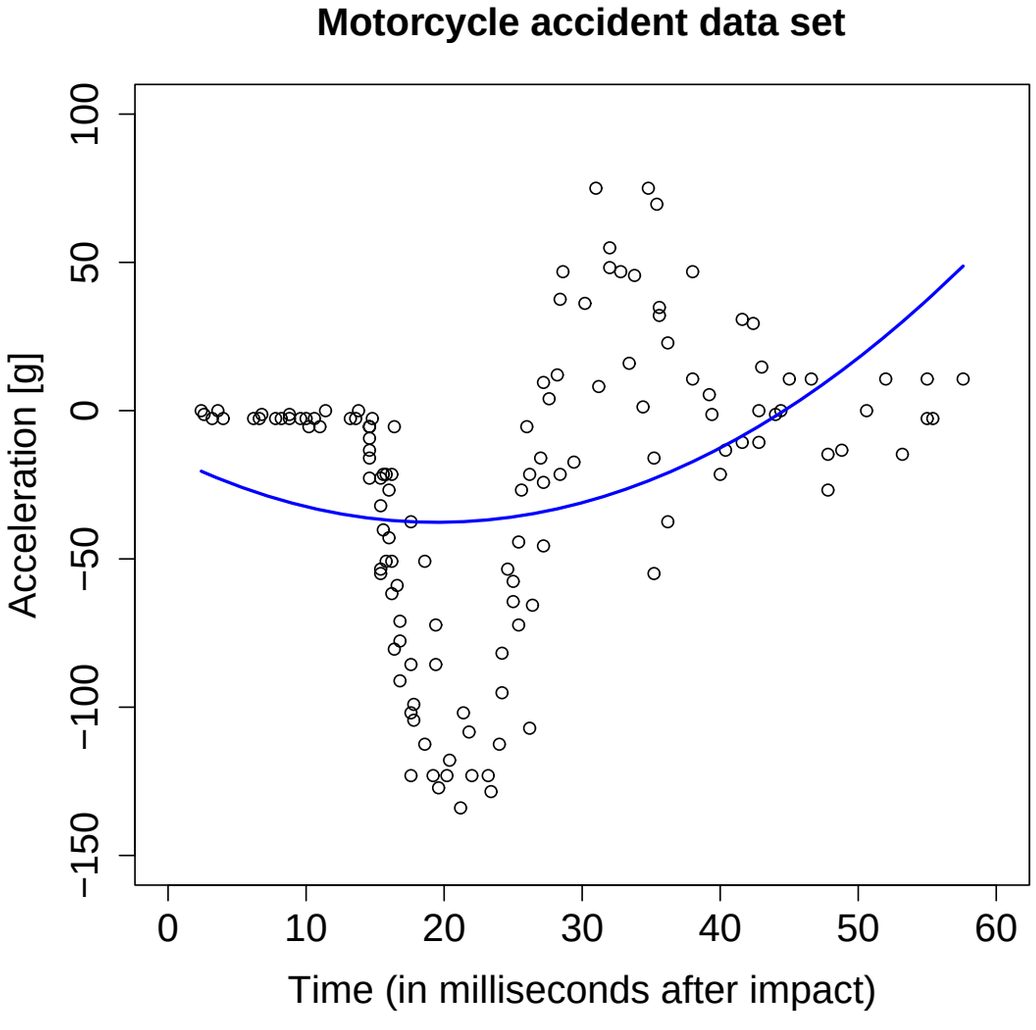
Motorradhelm Daten

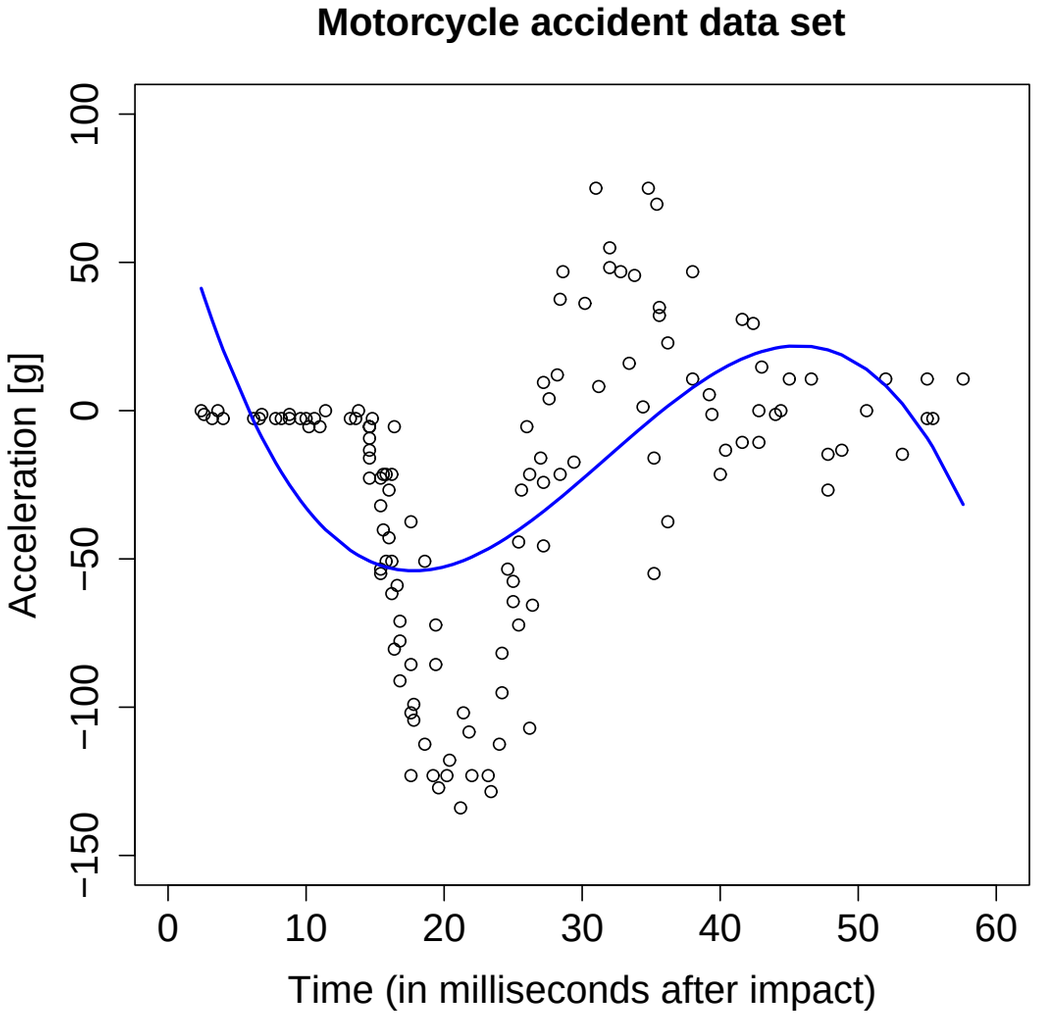
- Simuliertes Crash Experiment
- Beschleunigung von Motorradhelmen gemessen beim Unfall

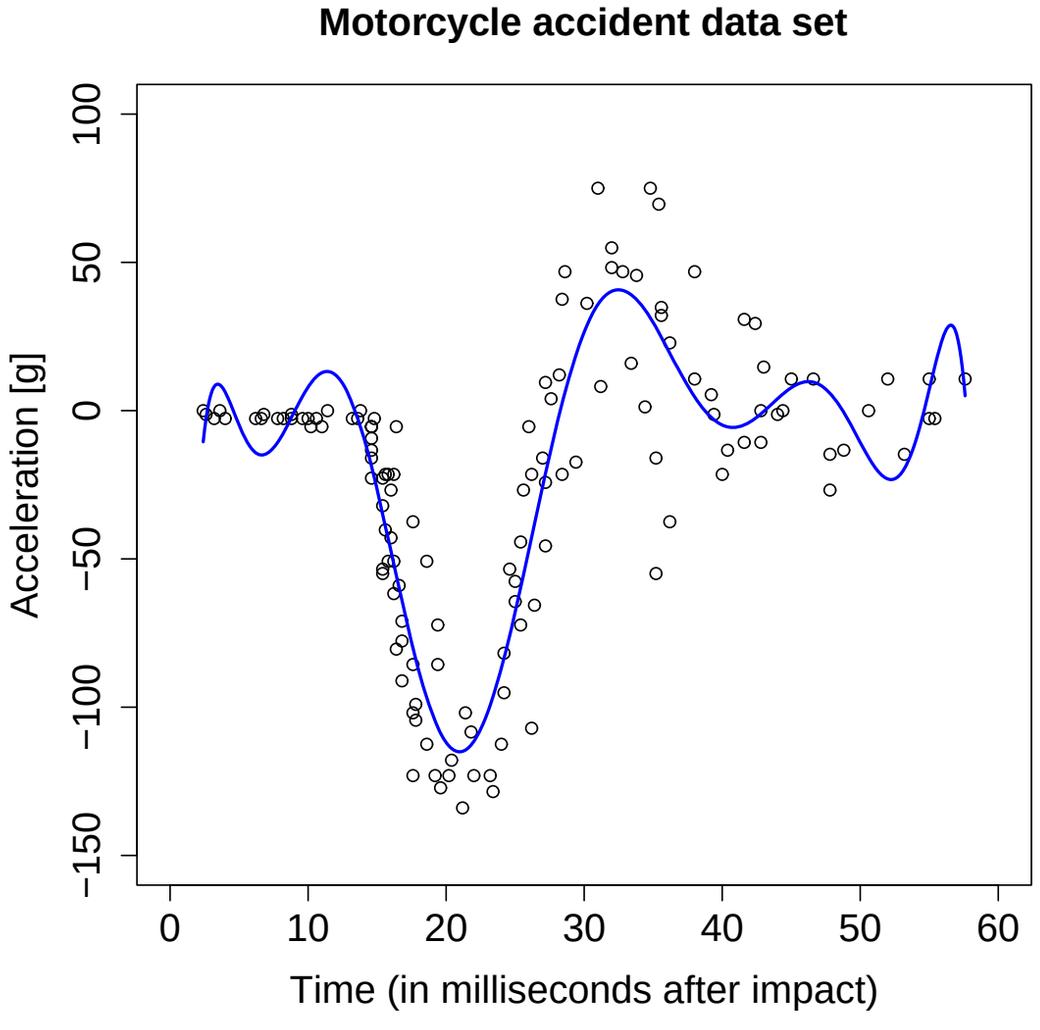
Linearer, quadratischer, kubischer und polynomialer Fit mit Grad 10


Motorcycle accident data set
Acceleration [g]
100
50
0
−50
−100
−150
0
10
20
30
40
50
60
Time (in milliseconds after impact)


Motorcycle accident data set
Acceleration [g]
100
50
0
−50
−100
−150
0
10
20
30
40
50
60
Time (in milliseconds after impact)


Motorcycle accident data set
Acceleration [g]
100
50
0
−50
−100
−150
0
10
20
30
40
50
60
Time (in milliseconds after impact)


Motorcycle accident data set
Acceleration [g]
100
50
0
−50
−100
−150
0
10
20
30
40
50
60
Time (in milliseconds after impact)

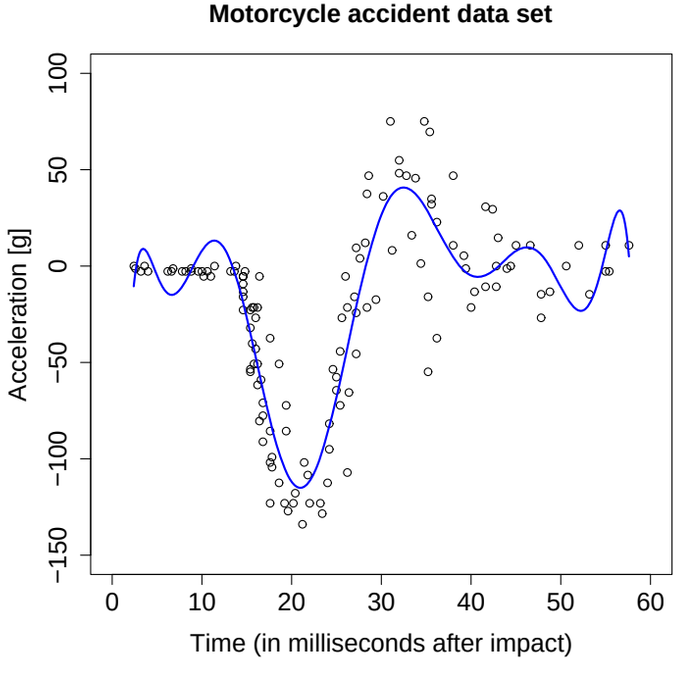

Motorcycle accident data set
Acceleration [g]
100
50
0
−50
−100
−150
0
10
20
30
40
50
60
Time (in milliseconds after impact)

Realisierung in R

```
> library(MASS); data(mcycle); attach(mcycle)

> # Construct polynomial basis
> pbase <- function(x, k) {
    P <- outer(x, 0:k, "^")
    P }

> plot(times, accel)
> k <- 10;  P <- pbase(times, k)
> fit <- lsfit(P, accel, intercept=F)
> beta <- fit$coef; mu <- P %*% beta
> to <- order(times);  lines(times[to], mu[to])

> # Smoother predictions by ...
> x <- seq(min(times), max(times), by=0.1)
> Px <- pbase(x, k); pmu <- Px %*% beta
> lines(x, pmu, col="blue", lwd=2)
```

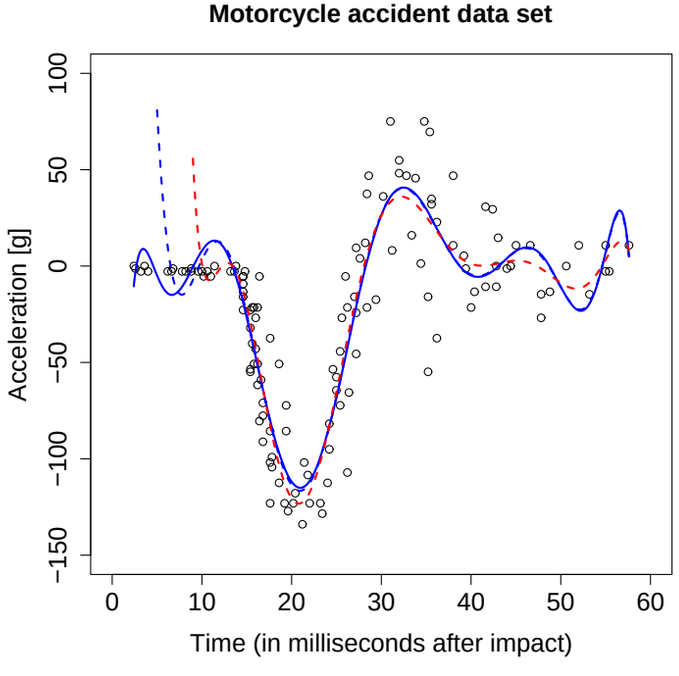
Weitere Darstellungen

- Höher Grad benötigt, um annehmbare Kurvenanpassung zu erhalten
- Schlechte numerische Eigenschaften (verwende orthogonale Polynome)



Sensitivität auf Datenänderungen

- Längerer linker Teil nahe Null
- Bemerke die Wackler

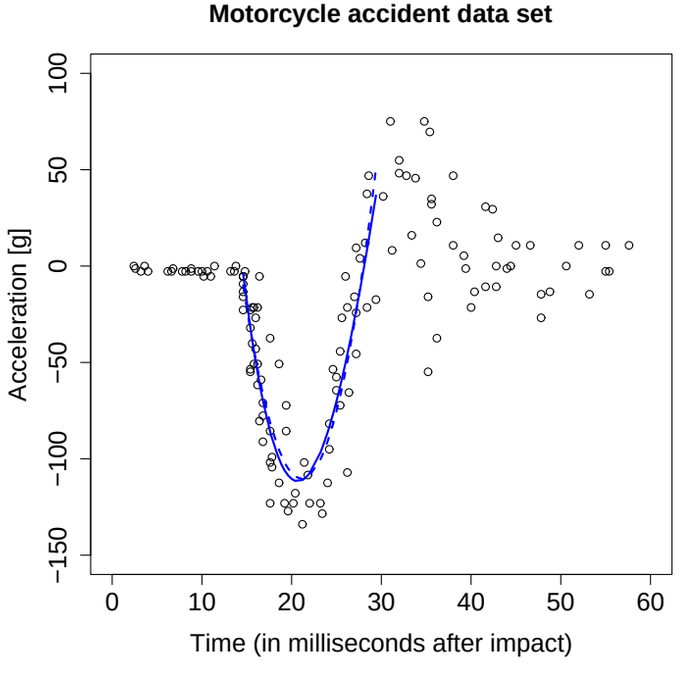


Ärger mit Polynomen

- Hoher Grad (10 oder mehr) wird möglicherweise benötigt
- Basis Funktionen (Potenzen von x) sind global
- Änderung an einem Ende (vertikal) impliziert auch Änderung am anderen Ende
- Gute Anpassung in einem Ende ruiniert diese am anderen Ende
- Unerwartete Wackler
- Je höher der Grad umso sensitiver der Fit
- Globale Polynome sind eine Sackgasse (Brian Marx)

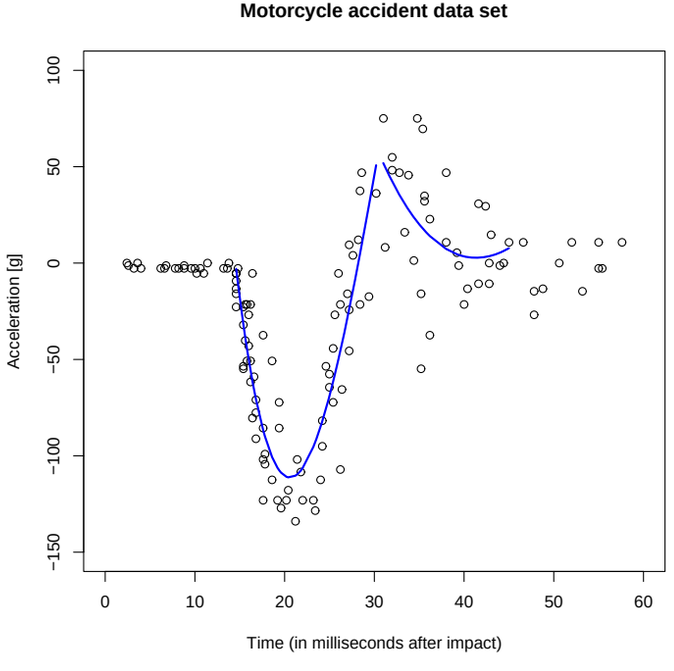
Arbeite mit Abschnitten

- Modelliere kurze Abschnitte mittels Polynome von niedrigem Grad
- Wahl der Breite der Abschnitte?



Benachbarte Abschnitte

- Keine schönen Übergänge
- Sprünge an den Rändern



Alternative: Lokale Basis Funktionen

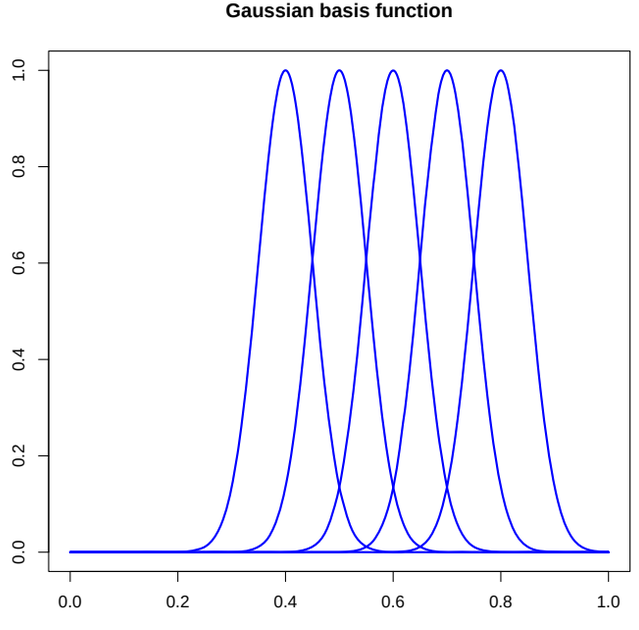
- Befreiung vom Konzept globaler Basis Funktionen
- Lokale Basis Funktionen: Nicht-Null auf limitierten Bereich
- Dort können sie sich frei ändern ohne sonstwo etwas anzustellen
- Einfaches Beispiel: Gaußkurve zu N(τ, σ^2)

$$g(x|\tau, \sigma^2) = \exp\left(-\frac{(x-\tau)^2}{2\sigma^2}\right)$$

- Eigentlich 0 für $|x - \tau| > 3\sigma$
- (Keine Division durch $\sqrt{2\pi\sigma^2}$: Maximum von g ist immer 1)

Gauß Basis

- Eine Basis von Gauß Funktionen mit gleichem σ^2 aber unterschiedlichen τ's
- Abstände der τ's: 2σ



Kurvenanpassung mit Gauß Basis Funktionen

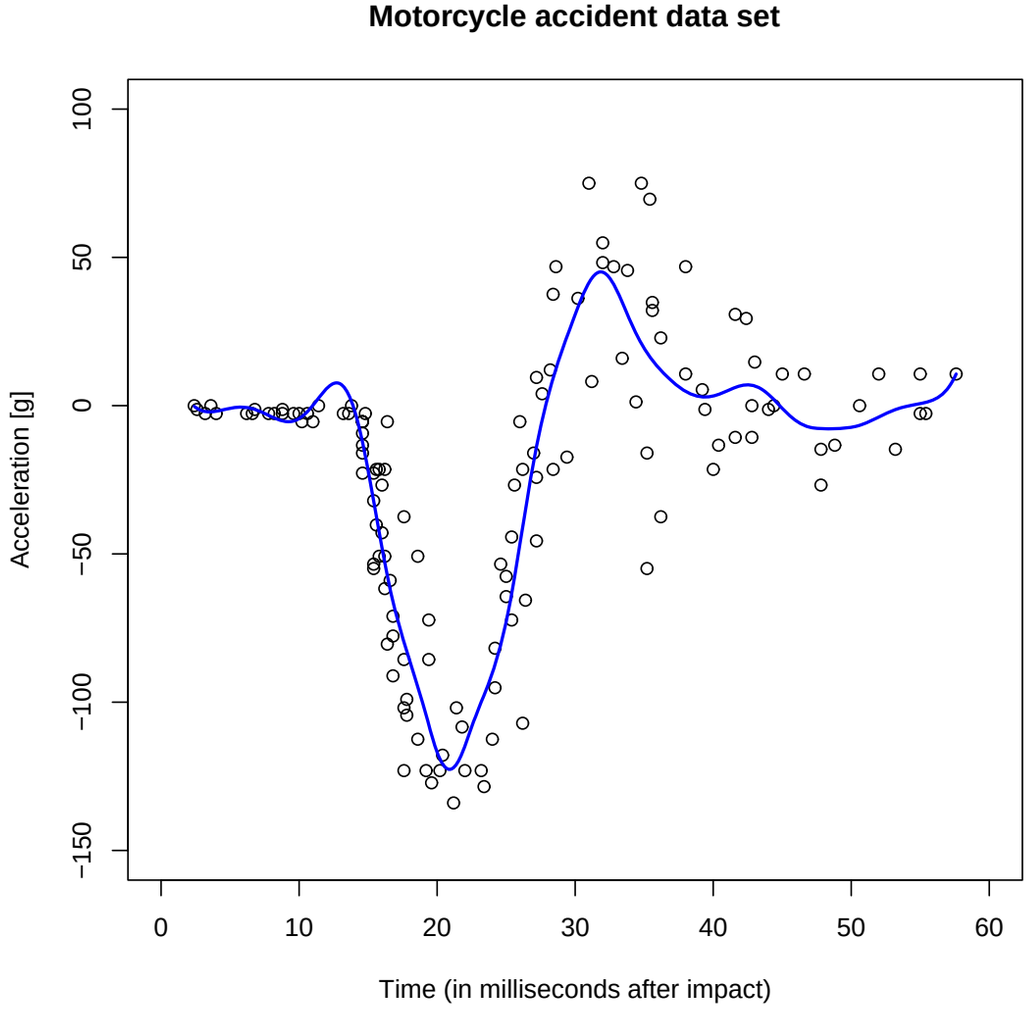
- Basis Funktionen sind die Spalten in der Matrix $\mathbf{G}$
- Eine Zeile für jedes x_i, eine Spalte für jedes τ_j

$$g_{ij} = g(x_i|\tau_j, \sigma) = \exp\left(-\frac{(x_i - \tau_j)^2}{2\sigma^2}\right), \quad j = 1, \ldots, m$$

- Modell $\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu} = \mathbf{G}\boldsymbol{\beta}$
- Lineare Regression: minimiere $\mathrm{SSE}(\boldsymbol{\beta}) = |\mathbf{y} - \mathbf{G}\boldsymbol{\beta}|^2$
- Normal Gleichungen $\mathbf{G}^t\mathbf{G}\hat{\boldsymbol{\beta}} = \mathbf{G}^t\mathbf{y}$
- Explizite Lösung $\hat{\boldsymbol{\beta}} = (\mathbf{G}^t\mathbf{G})^{-1}\mathbf{G}^t\mathbf{y}$

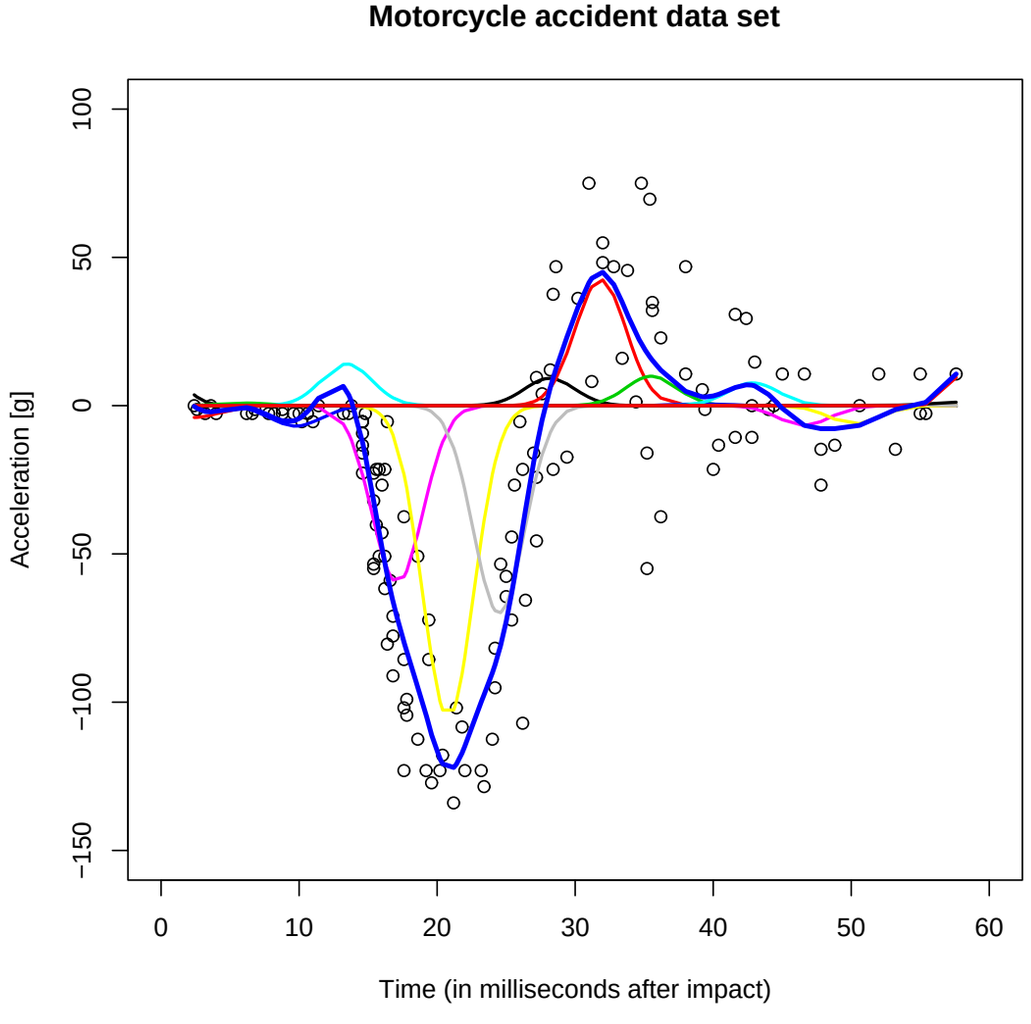
Motorradhelm Daten mit Gauß Basis

```
> gauss <- function(x, knots, sigma) {
>  exp(-(x - knots)^2/(2*sigma^2)) }

> # Construct Gaussian basis (m = nseg + 1)
> gbase <- function(x, xl=min(x), xr=max(x), nseg=10){
>  dx <- (xr - xl)/nseg
>  knots <- seq(xl, xr, by = dx)
>  sigma <- dx/2
>  G <- outer(x, knots, gauss, sigma)
>  G }

> G <- gbase(times, nseg=15)   # equals k=16
> fit <- lsfit(G, accel, intercept=F)
> beta <- fit$coef;  mu <- G %*% beta
```


Motorcycle accident data set
Acceleration [g]
100
50
0
−50
−100
−150
0
10
20
30
40
50
60
Time (in milliseconds after impact)

Komponenten dieses Modells mit Gauß Basis

```
> plot(times, accel)
> to <- order[times]

> for (j in 1:dim(G)[2]) {
>  comp <- G[ , j] * beta[j]
>  lines(times[to], comp[to], col=j)
> }

> lines(times[to], mu[to], col="blue")
```


Motorcycle accident data set
Acceleration [g]
100
50
0
−50
−100
−150
0
10
20
30
40
50
60
Time (in milliseconds after impact)

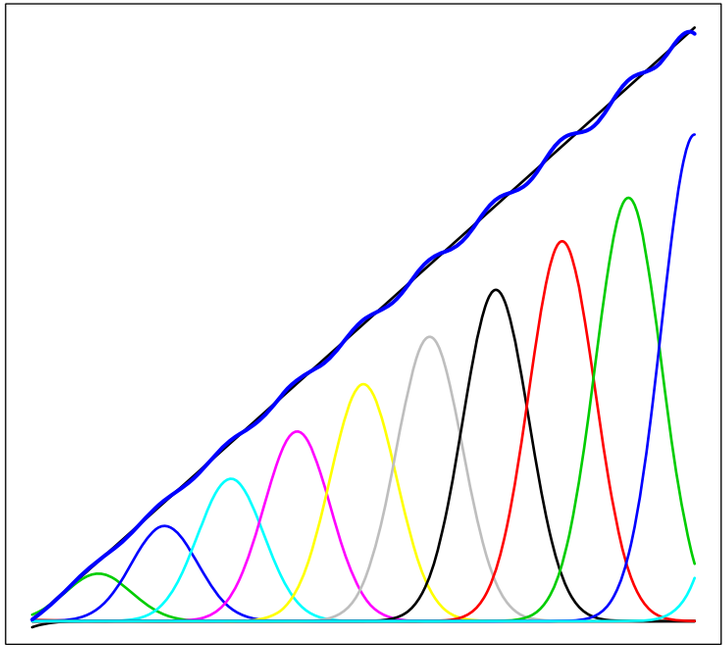
Eigenschaften einer Gauß Basis

- Gauß Basis Funktionen sind ziemlich praktikabel
- Einfach zu berechnen
- Einfach zu erklären
- Nachteil 1: nicht wirklich lokal
- Nachteil 2: keine exakte Anpassung von Geraden (Polynome)
- Alternative: B-Splines (kurz für **B**asis Spline)

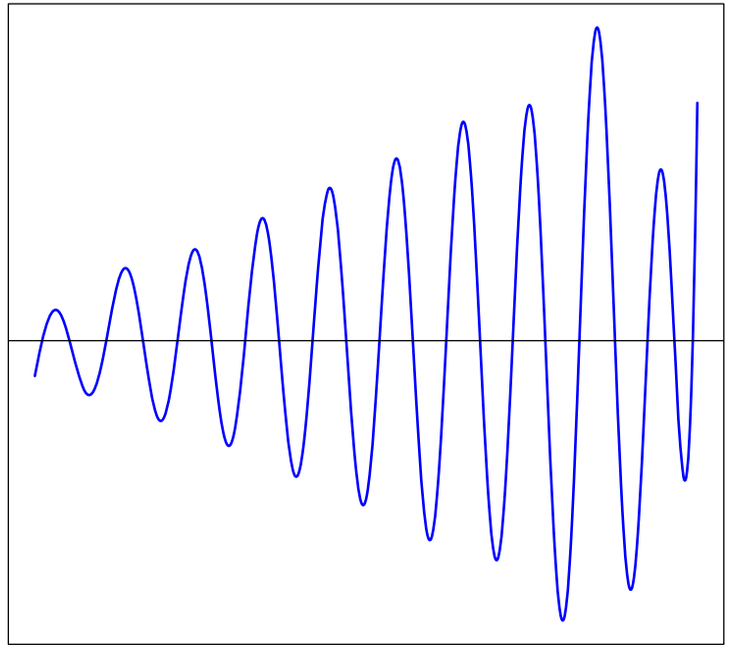
Problem Gauß Residuenwelle: fit/residuals

```
> # Gaussian Ripple
> y <- x <- (1:1000)/1000  # all y's are on a straight line
> G <- gbase(x, nseg=11)
> fit <- lsfit(G, y, intercept=F)
> beta <- fit$coef; mu <- G %*% beta

> plot(x, y)
> for (k in 1:dim(G)[2]) {
>   comp <- G[ ,k]*beta[k]
>   lines(x, comp, lwd=2, col=k)
> }
> lines(x, mu, lwd=3, col="blue")

> r <- y-mu
> plot(x, r)
> abline(0, 0)
```

Gauss fit to straight line



Residuals



(Natürliche) Polynomiale Splines

Eine stückweise Funktion $f : [a, b] \mapsto \mathbb{R}$ nennt man **polynomialer Spline** vom Grad k mit Knoten $a = \tau_1 < \cdots < \tau_m = b$, wenn

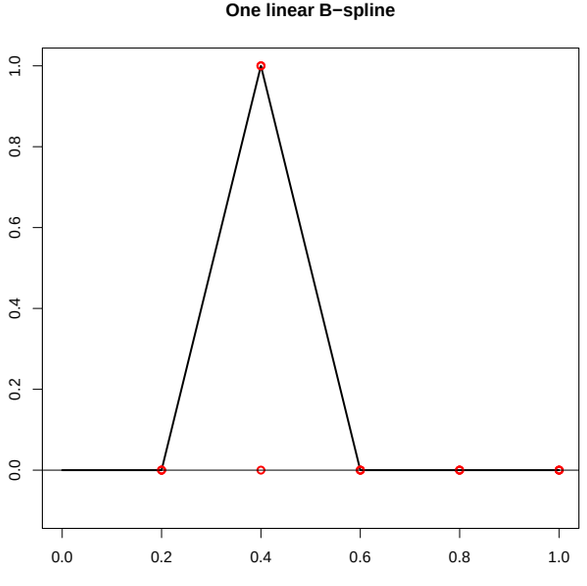
1. $f(x)$ ist $(k-1)$-mal stetig differenzierbar in den inneren Punkten τ_j, mit $\lim_{x \uparrow \tau_j} f^{(l)}(x) = \lim_{x \downarrow \tau_j} f^{(l)}(x)$, $j = 2, \ldots, m-1$, $l = 0, \ldots, k-1$, und
2. $f(x)$ ist ein Polynom vom Grad k auf den Intervallen $[\tau_j, \tau_{j+1})$ definiert durch die Knoten.

Die Funktion $f(x)$ nennt man **natürlicher polynomialer Spline** mit Knoten $a < \tau_1 < \cdots < \tau_m < b$, wenn

1. $f(x)$ ist ein polynomialer Spline für die gegebenen Knoten, und
2. $f(x)$ genügt der Randbedingung $f''(a) = f''(b) = 0$, d.h. $f(x)$ ist linear in den Intervallen $[a, \tau_1)$ und $[\tau_m, b]$.

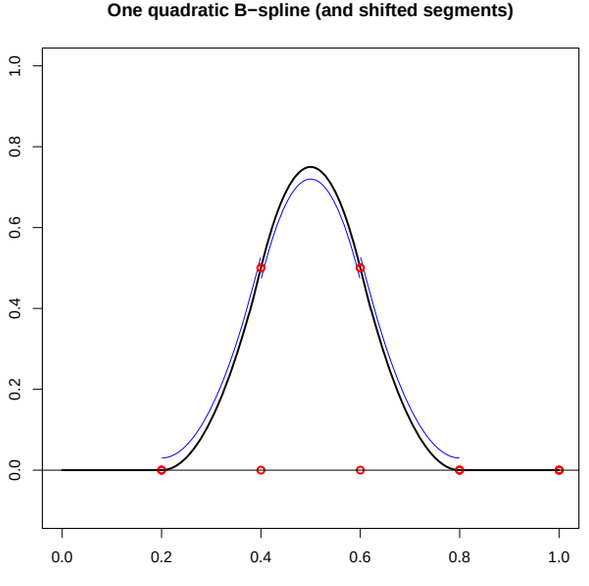
1 linearer B-Spline

- 2 Stücke, jedes eine Gerade, sonst Null
- Schöne Verbindungen in den Knoten (t_1 bis t_3): gleiche Werte
- Steigung springt in den Knoten



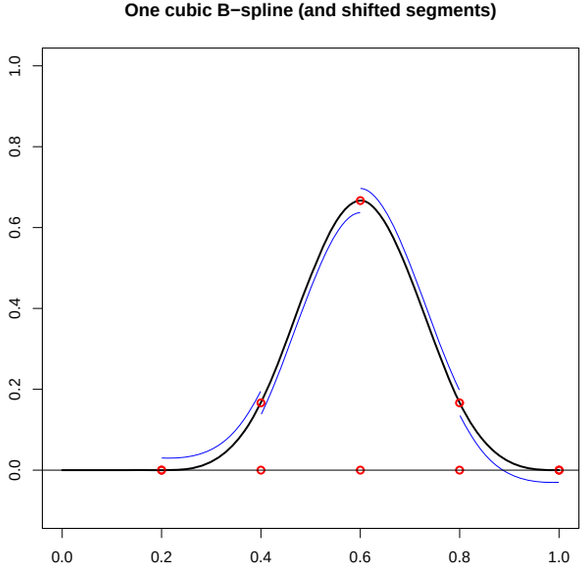
1 quadratischer B-Spline

- 3 Stücke, jedes ein quadratisches Segment, sonst Null
- Schöne Verbindungen in den Knoten (t_1 bis t_4): gleiche Werte und Steigungen
- Gestalt etwa ähnlich einer Gaußfunktion



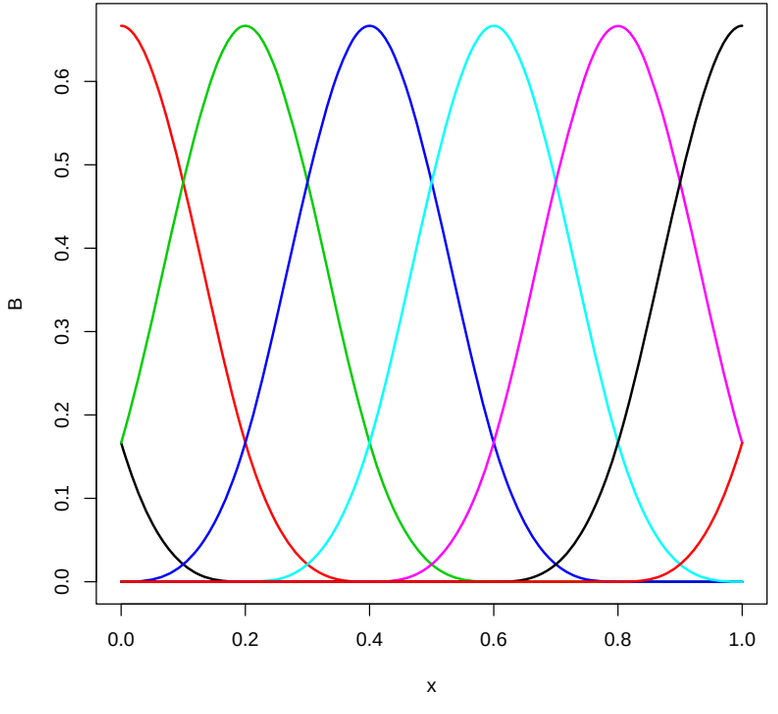
1 kubischer B-Spline

- 4 Stücke, jedes ein kubisches Segment, sonst Null
- In den Knoten (t_1 bis t_5): gleiche Werte, erste und zweite Ableitungen
- Gestalt noch ähnlicher einer Gaußfunktion



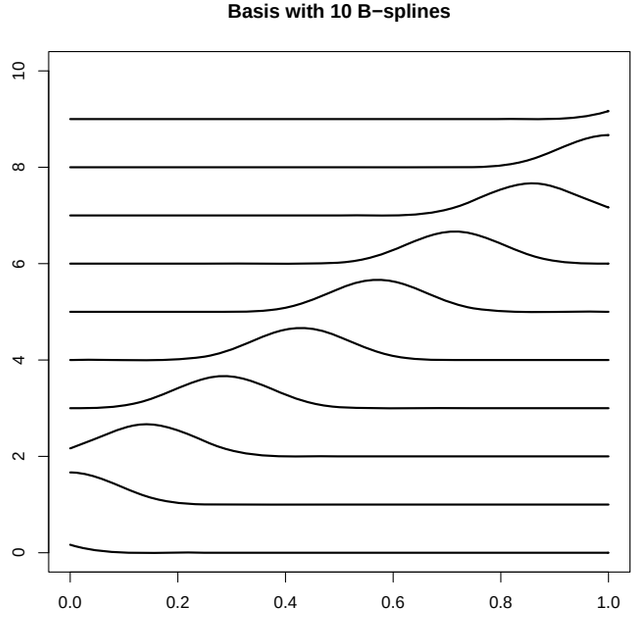
Ein ganzer Satz kubischer B-Splines

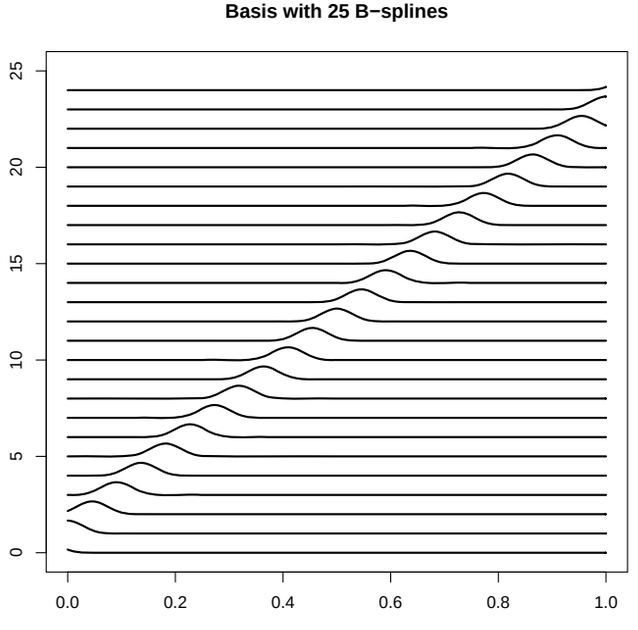
```
> x <- (0:1000)/1000
> B <- bbase(x, nseg=5, deg=3); matplot(x, B, type='l')
```



Kubische B-Splines in Perspektive

```
> B <- bbase(x, nseg=7, deg=3)
> plot(x, B[ , 1], type="l");  c <- rep(1, length(x))
> for (k in 2:dim(B)[2]) {
>  lines(x, B[ ,k]+c, lwd=2);  c <- c+1 }
```





B-Spline Basis

- Basis matrix $\mathbf{B}$

- Spalten sind B-Splines

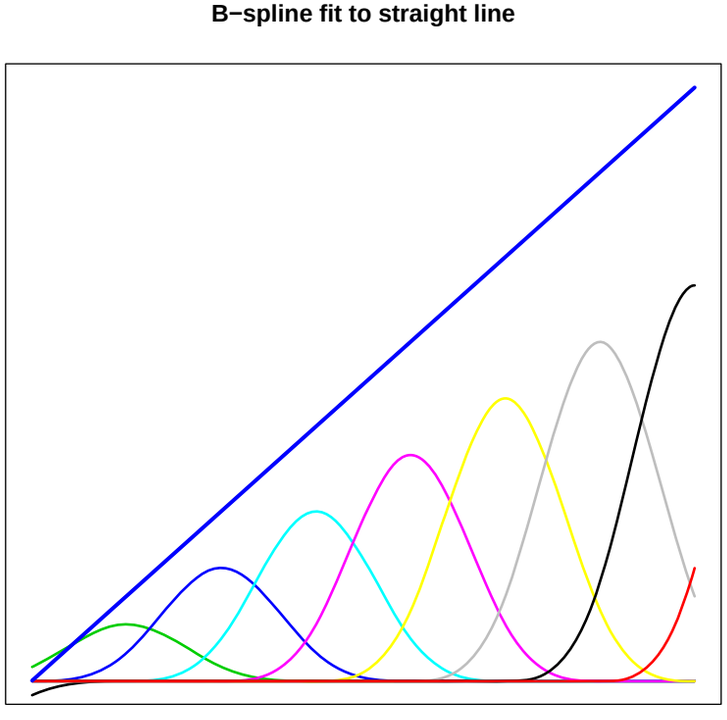
$$\begin{pmatrix} B_1(x_1) & B_2(x_1) & B_3(x_1) & \dots & B_k(x_1) \\ B_1(x_2) & B_2(x_2) & B_3(x_2) & \dots & B_k(x_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ B_1(x_n) & B_2(x_n) & B_3(x_n) & \dots & B_k(x_n) \end{pmatrix}$$

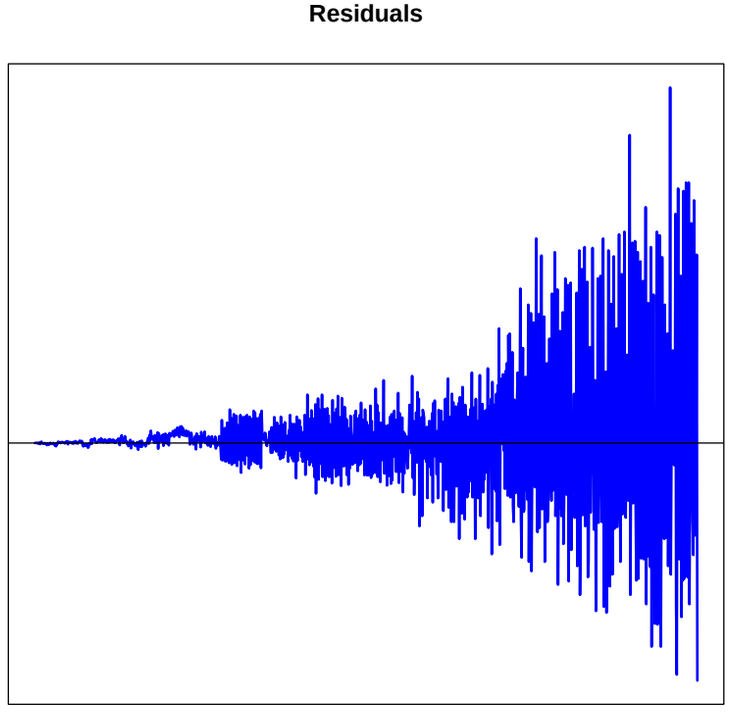
- In jeder Zeile nur einige wenige nicht-null Elemente (Grad plus 1)

- Nur wenige Basis Funktionen tragen bei zu $\mu_i = \sum_{j=1}^{k} B_j(x_i)\beta_j = \mathbf{B}_i^t \boldsymbol{\beta}$.

B-Splines haben keine Residuenwelle

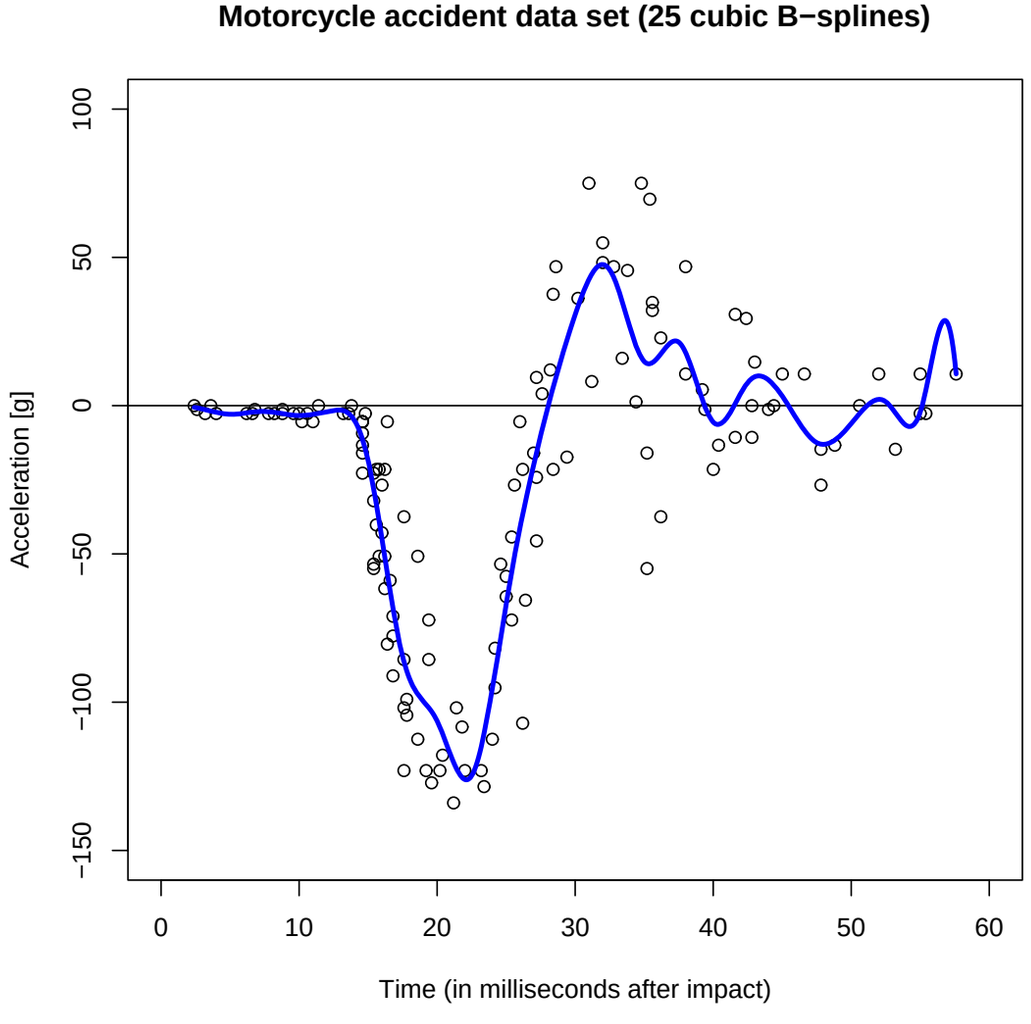
```
> # B-splines have no ripple
> y <- x <- (1:1000)/1000  # all y's are on a straight line
> B <- bbase(x, nseg=7)
> fit <- lsfit(B, y, intercept=F)
> beta <- fit$coef; mu <- B %*% beta

> plot(x, y)
> for (k in 1:dim(B)[2]) {
>   comp <- B[ ,k]*beta[k]
>   lines(x, comp, lwd=2, col=k)
> }
>lines(x, mu, lwd=3, col="blue")

> r <- y-mu
> plot(x, r)
> abline(0, 0)
```

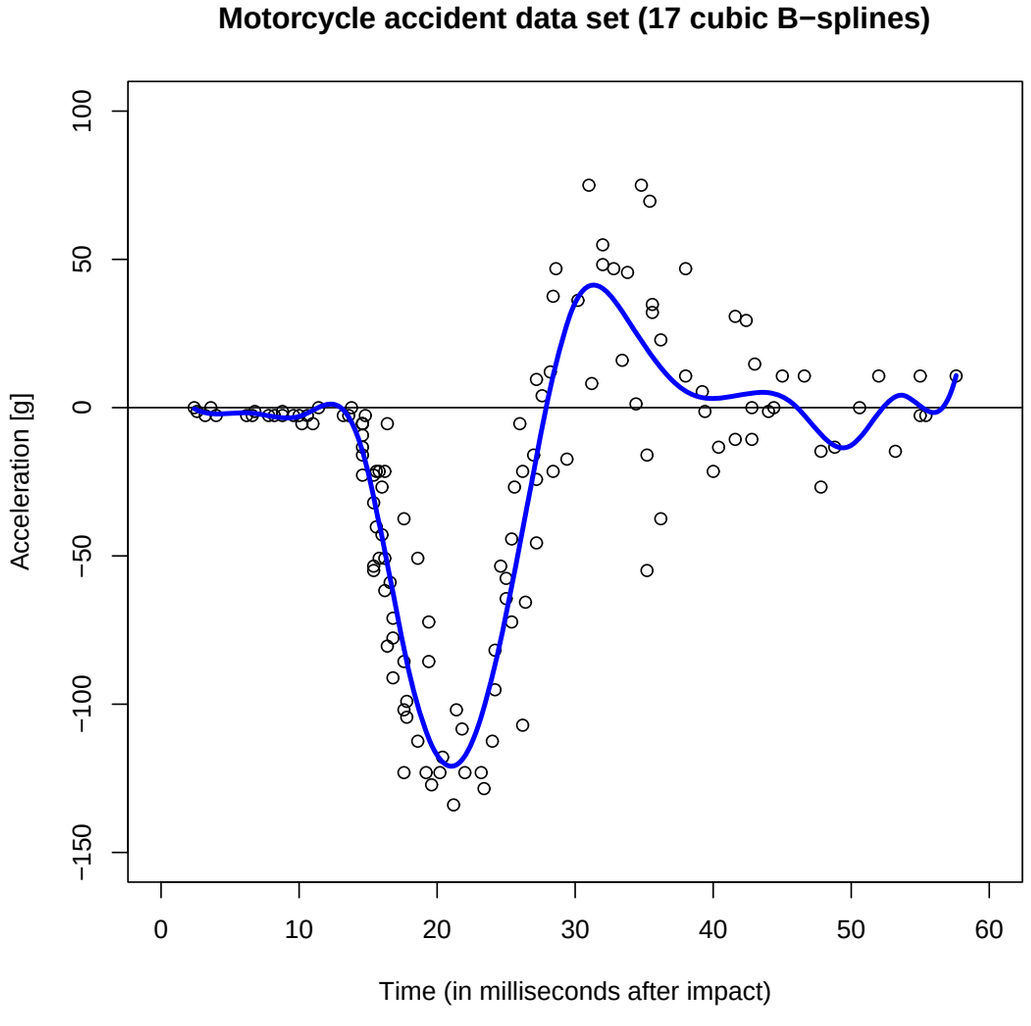

B−spline fit to straight line

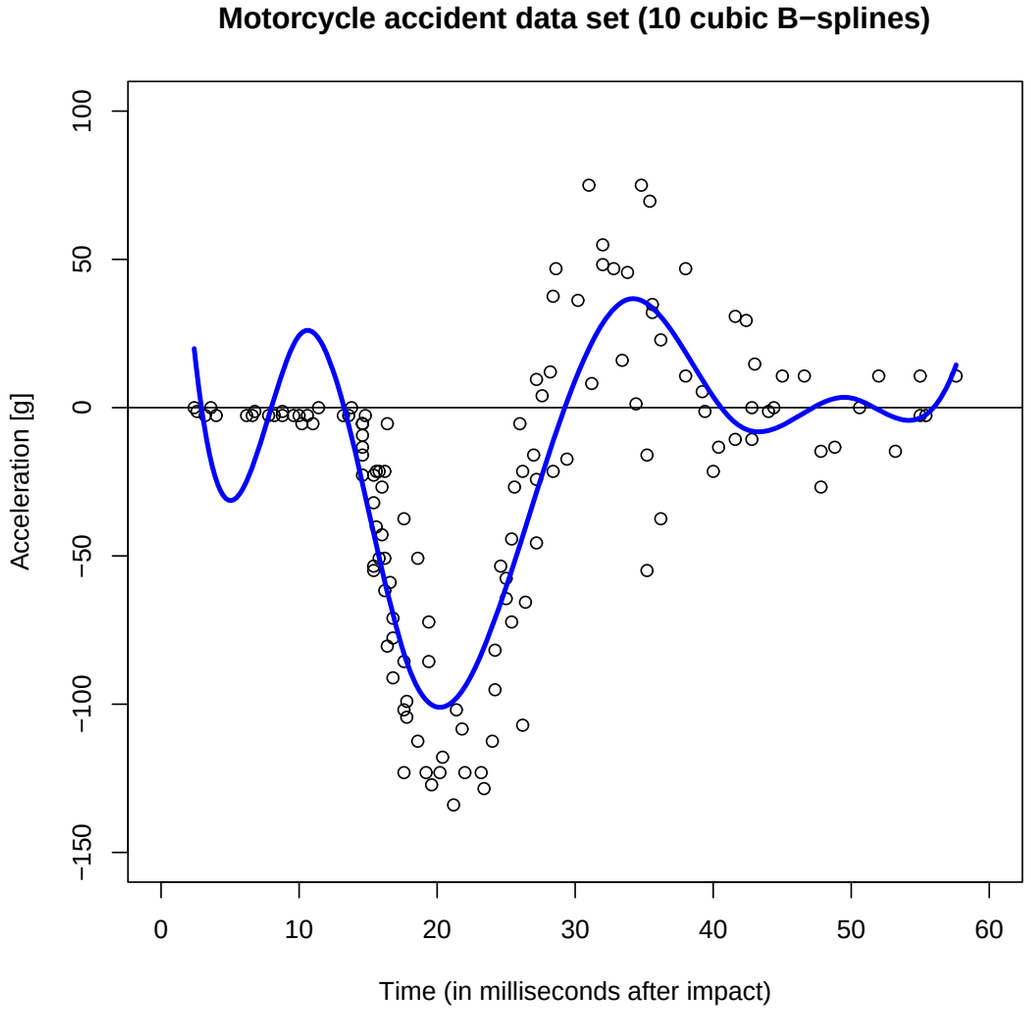

Residuals

B-Splines in Aktion

```
> plot(times, accel); abline(0, 0)

> B <- bbase(times, nseg=22, deg=3) # results in "nseg+3 = 25" functions
> beta <- solve(t(B) %*% B, t(B) %*% accel)
> mu <- B %*% beta

> lines(times, mu)
```


Motorcycle accident data set (25 cubic B−splines)
Acceleration [g]
100
50
0
−50
−100
−150
0
10
20
30
40
50
60
Time (in milliseconds after impact)


Motorcycle accident data set (17 cubic B−splines)
Acceleration [g]
100
50
0
−50
−100
−150
0
10
20
30
40
50
60
Time (in milliseconds after impact)


Motorcycle accident data set (10 cubic B−splines)
Acceleration [g]
100
50
0
−50
−100
−150
0
10
20
30
40
50
60
Time (in milliseconds after impact)

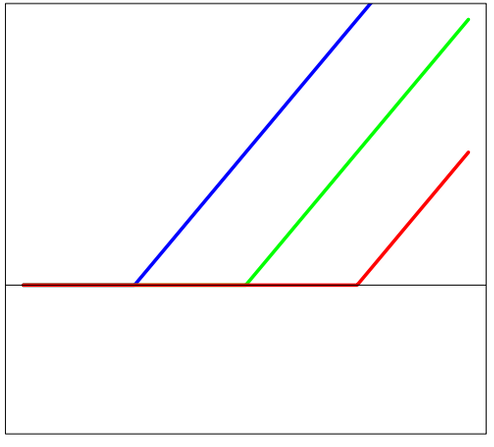
Wie berechnet man B-Splines?

- Work from first principles
- Berechne die Parameter der polynomialen Segmente
- Neun (3 mal 3) Koeffizienten, 8 Bedingungen, Höhe beliebig
- Leichter: rekursive Formel von De Boor
- Noch leichter: Differenzen von abgeschnittenen Potenz Funktionen (TPF)
- TPF: $f(x|t,p) = (x-t)_+^p = (x-t)^p I(x > t)$
- Potenz Funktion wenn $x > t$, sonst 0
- Vermeidet schlechte numerische Eigenschaften der TPF (De Boor)

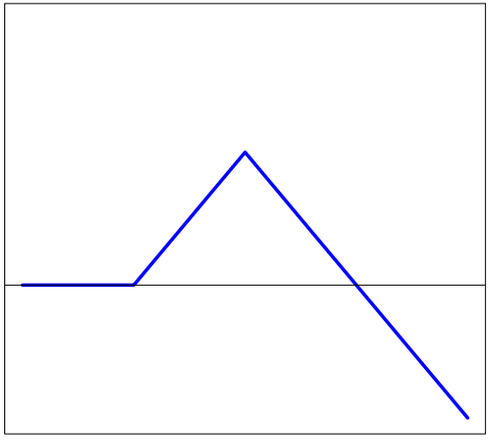
B-Splines und TPFs

```
> # Truncated p-th power function (by Eilers & Marx)
> tpower <- function(x, t, p)
>   (x - t)^p * (x>t)

> # Construct B-spline basis (by Eilers & Marx)
> bbase <- function(x, xl=min(x), xr=max(x), nseg=10, deg=3){
>   dx <- (xr-xl)/nseg
>   knots <- seq(xl-deg*dx, xr+deg* dx, by=dx)
>   P <- outer(x, knots, tpower, deg)
>   n <- dim(P)[2]
>   D <- diff(diag(n), diff=deg+1)/(gamma(deg+1)*dx^deg)
>   B <- (-1)^(deg+1)*P %*% t(D)
>   B }
```

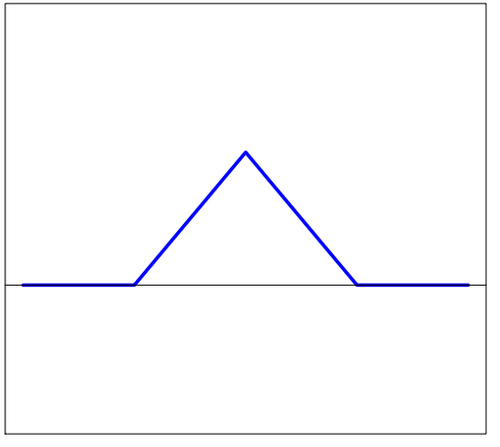
Three truncated power functions: f_1, f_2 and f_3



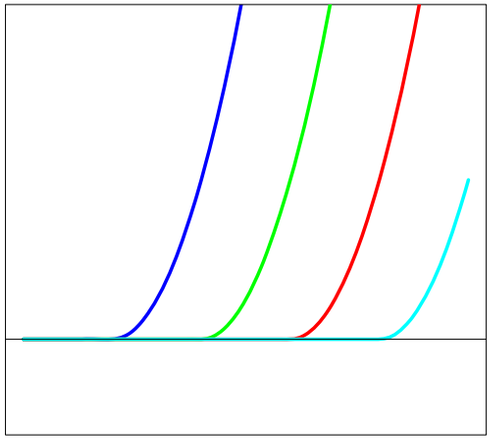
$f_1 - 2f_2$



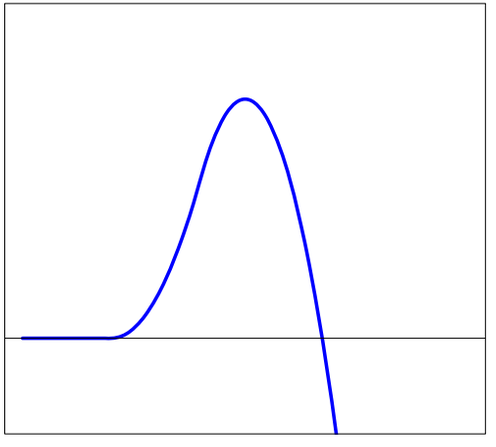
Linear B−spline: $f_1 - 2f_2 + f_3$



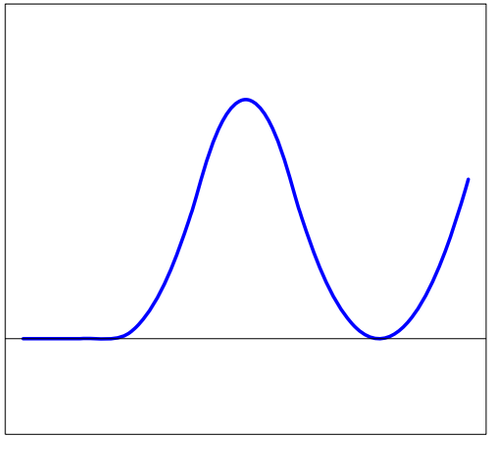
Four truncated power functions: f_1, f_2, f_3 and f_4



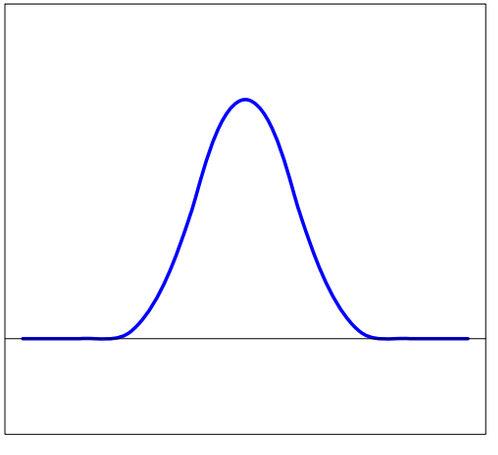
$f_1 - 3f_2$



$f_1 - 3f_2 + 3f_3$



Quadratic B−spline: $f_1 - 3f_2 + 3f_3 - f_4$



B-Spline Zusammenfassung

- B-Splines sind lokale Funktionen, schauen aus wie Gauß Funktionen
- B-Splines sind die Spalten der Basis Matrix $\mathbf{B}$
- Skalieren und Summieren ergibt die geglätteten Werte: $\boldsymbol{\mu} = \mathbf{B}\boldsymbol{\beta}$
- Die Knoten bestimmen die B-Spline Basis
- Polynomiale Stücke machen die B-Splines aus, sie sind in den Knoten verbunden
- Allgemeine Muster sind für die Wahl der Knoten möglich
- Wir betrachten nur gleiche Abstände
- Anzahl der Knoten bestimmt die Breite und die Anzahl der B-Splines

Erkenntnisse

- Polynome sind eigentlich nicht verwendbar um komplizierte Kurven zu fitten
- Lokale Basen sind besser
- Gauß'sche Glockenkurve, um die Idee zu bekommen
- B-Splines sind besser
- B-Splines sind Differenzen von abgeschnittenen Potenz Funktionen (TPFs)
- Jedoch nicht ideal bei dünn-besetzten Daten
- Nächste Idee: Penalty/Bestrafung

11. Variablenselektion

Ziel: Wähle das **beste** Modell aus einer Klasse von MLR's.

Volles Modell enthält alle m möglicherweise erklärenden Größen (Prädiktoren)

Suche nach dem besten Modell, das nur eine Teilmenge dieser Prädiktoren enthält.

Zwei Aspekte hierbei sind:

1. Evaluiere jedes dieser betrachteten Modelle.

2. Entscheide, welche Untermenge von Prädiktoren die beste ist.

Kriterien zur Evaluierung:

- Adjustiertes R^2:

Bestimmtheitsmaß ist definiert als Anteil der totalen Stichprobenvarianz in den y, der durch das Regressionsmodell erklärt ist:

$$R^2 = \frac{\mathsf{SSR}(\hat{\boldsymbol{\beta}})}{\mathsf{SST}} = 1 - \frac{\mathsf{SSE}(\hat{\boldsymbol{\beta}})}{\mathsf{SST}}\,.$$

Nimmt man weitere irrelevante Prädiktoren ins Modell auf, so hat dies oft ein Anwachsen von R^2 zur Folge. Um dies zu kompensieren, definiert man

$$R^2_{\mathsf{adj}} := 1 - \frac{\mathsf{SSE}(\hat{\boldsymbol{\beta}})/(n-p)}{\mathsf{SST}/(n-1)}\,,$$

wobei p die Anzahl der Parameter im betrachteten Modell bezeichnet.

Wir zeigen, dass die Hinzunahme eines Prädiktors nur dann zu einem größeren R^2_{adj} führt, wenn die korrespondierende partielle F-Teststatistik größer als 1 ist.

Betrachte ein Modell mit $p-1$ Prädiktoren ($\boldsymbol{\mu} = \mathbf{X}_1\boldsymbol{\beta}_1$) und gib eine weitere erklärende Variable dazu ($\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, $\mathbf{X} = (\mathbf{X}_1|\mathbf{x}_{new})$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^t, \beta_{new})^t$). So gilt

$$
\begin{aligned}
R^2_{\text{adj}}(\hat{\beta}_1) = 1 - \frac{\mathsf{SSE}(\hat{\beta}_1)/(n-(p-1))}{\mathsf{SST}/(n-1)} \quad &< \quad R^2_{\text{adj}}(\hat{\beta}) = 1 - \frac{\mathsf{SSE}(\hat{\beta})/(n-p)}{\mathsf{SST}/(n-1)} \\
\mathsf{SSE}(\hat{\beta}_1)/(n-p+1) \quad &> \quad \mathsf{SSE}(\hat{\beta})/(n-p) \\
\mathsf{SSE}(\hat{\beta}_1) \quad &> \quad \frac{n-p+1}{n-p}\mathsf{SSE}(\hat{\beta}) \\
&= \quad \left(1 + \frac{1}{n-p}\right)\mathsf{SSE}(\hat{\beta}) \\
\frac{\left(\mathsf{SSE}(\hat{\beta}_1) - \mathsf{SSE}(\hat{\beta})\right)/1}{\mathsf{SSE}(\hat{\beta})/(n-p)} = F_{1,n-p} \quad &> \quad 1\,.
\end{aligned}
$$

Oft ist es Praxis, jenen Satz an Prädiktoren zu wählen, der den größten Wert von R^2_{adj} generiert. Aber R^2_{adj} ist maximal für minimales $S^2 = \text{SSE}(\hat{\boldsymbol{\beta}})/(n-p)$. Wir können zeigen, dass dies zum Problem des **Overfitting** führt.

• AIC, Akaike's Informationskriterium:

Bewertet große Anpassungsgüte mit geringer Modellkomplexität. AIC ist derart definiert, dass gilt: je kleiner der Wert von AIC desto besser das Modell.

In diesem Sinn ist die negative Likelihood zum Modell ein Maß für die Anpassungsgüte, während p ein Maß für die Modellkomplexität ist. Wir definieren somit

$$\text{AIC} = 2\left[-\log L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \mathbf{y}) + (p+1)\right] ,$$

da sowohl $\boldsymbol{\beta}$ als auch σ^2 im Modell geschätzt werden.

Zur Erinnerung ist das Minimum der Log-Likelihood Funktion

$$\begin{aligned}\log L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \mathbf{y}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\frac{\mathsf{SSE}(\hat{\boldsymbol{\beta}})}{n} - \frac{1}{2\mathsf{SSE}(\hat{\boldsymbol{\beta}})/n}\mathsf{SSE}(\hat{\boldsymbol{\beta}}) \\ &= -\frac{n}{2}\log\frac{\mathsf{SSE}(\hat{\boldsymbol{\beta}})}{n} - \frac{n}{2}\log(2\pi) - \frac{n}{2}\,.\end{aligned}$$

Wir erhalten daher

$$\mathsf{AIC} = 2\left[-\log L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \mathbf{y}) + (p+1)\right] = n\log\frac{\mathsf{SSE}(\hat{\boldsymbol{\beta}})}{n} + 2p + \mathsf{Rest}\,,$$

wobei der Rest weder von $\mathsf{SSE}(\hat{\boldsymbol{\beta}})$ noch von p abhängt, und somit derselbe für alle betrachteten Modelle ist. Deshalb berechnet `R`

$$\mathsf{AIC} = n\log\frac{\mathsf{SSE}(\hat{\boldsymbol{\beta}})}{n} + 2p\,.$$

- AIC_c, Korrigiertes Akaike's Informationskriterium:

Diese korrigierte Version ist definiert als

$$\text{AIC}_c = -2 \log L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \mathbf{y}) + 2(p+1) + 2\frac{(p+1)(p+2)}{n-p} = \text{AIC} + 2\frac{(p+1)(p+2)}{n-p}\,.$$

Dies entspricht einer bias-korrigierten Version des AIC für kleine Stichprobengrößen oder falls die Parameteranzahl im Vergleich zu n relativ groß ist.

AIC_c sollte dem AIC vorgezogen werden, falls $n/(p+1) \leq 40$. Weiters wird generell empfohlen, in der Praxis AIC_c zu verwenden, da für $n \to \infty$ folgt, dass $\text{AIC}_c \to \text{AIC}$.

Jedoch hat auch AIC die Tendenz zum Overfitting, falls n klein oder falls p groß in Relation zu n ist (der Strafterm für Modellkomplexität ist nicht stark genug). Da AIC_c einen größeren Strafterm aufweist, ist die biaskorrigierte Version fast immer dem AIC vorzuziehen.

• BIC, Bayes'sches Informationskriterium:

Schwarz (1978) schlug folgendes Kriterium vor.

$$\text{BIC} = -2 \log L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \mathbf{y}) + (p+1) \log n \,.$$

Auch BIC ist derart definiert, dass gilt: je kleiner der Wert von BIC desto besser das Modell. BIC ist sehr ähnlich zu AIC, nur ist der Faktor 2 im Strafterm jetzt durch $\log n$ ersetzt. Somit zieht BIC eher einfache Modelle vor.

• Eine sehr populäre Strategie in der Praxis ist es, Werte von R^2_{adj}, AIC, AIC_c und BIC zu berechnen und die Modelle zu vergleichen, die AIC, AIC_c und BIC minimieren, mit jenem das R^2_{adj} maximiert.

Entscheidung über Prädiktoren:

Hier gibt es 2 unterschiedliche Vorgehensweisen:

1. Betrachte alle möglichen Teilmengen.
2. Verwende schrittweise Methoden.

• Alle möglichen Teilmengen:

Stehen m Prädiktoren zur Verfügung, so betrachtet man alle 2^m möglichen Regressionsmodelle und identifiziert jenes, welches ein Anpassungsmaß maximiert oder ein Informationskriterium minimiert.

Hält man die Anzahl p der Prädiktoren in einem Modell fest, so liefern alle 4 Kriterien jenes Modell als bestes Modell, das die kleinste Fehlerquadratsumme aufweist. Stellt man jedoch Vergleiche über Modelle mit verschiedenen Anzahlen von Prädiktoren an, so können die Kriterien sehr wohl unterschiedliche Resultate generieren.

Beispiel (Brückenkonstruktion): Bevor man mit der Konstruktion beginnt, durchläuft ein derartiges Projekt viele Entwurfsphasen. Kann man die Dauer der Entwurfsphasen vorhersagen, hilft dies bei der Planung des notwendigen Budgets. Informationen von 45 Brückenprojekten stehen zur Verfügung.

y Time: Entwurfszeit in Personen-Tage
x_1 DArea: Deck area (Brückenfläche) in 1,000 sq ft
x_2 CCost: Konstruktionskosten in $ 1,000
x_3 Dwgs: Anzahl Konstruktionspläne
x_4 Length: Brückenlänge in ft
x_5 Spans: Anzahl der Brückenfelder

```
> bridge <- read.table("bridge.dat", header=TRUE)
> attach(bridge)

> m1 <- lm(log(Time) ~ log(DArea)+log(CCost)+log(Dwgs)+log(Length)+log(Spans))
> summary(m1)
```

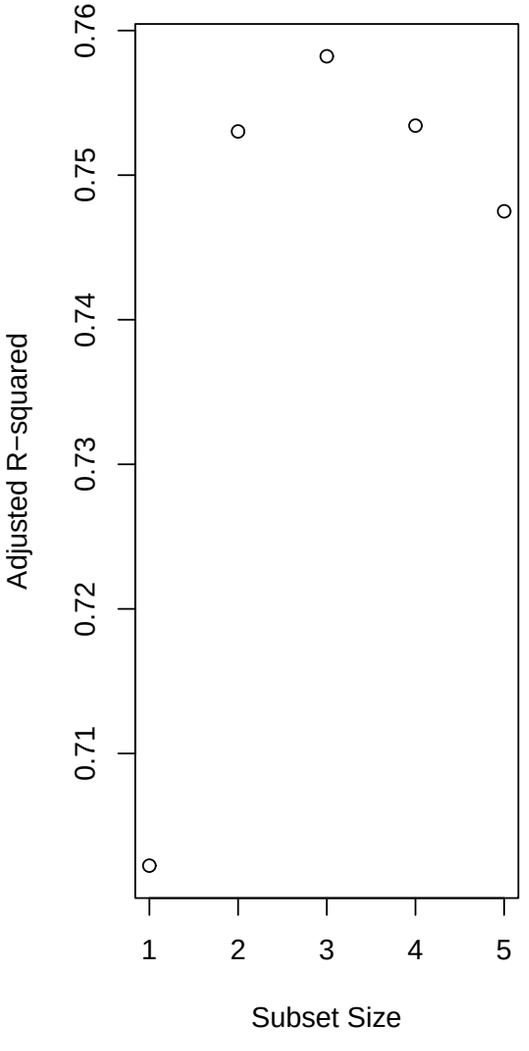
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.28590    0.61926   3.691  0.00068 ***
log(DArea)  -0.04564    0.12675  -0.360  0.72071
log(CCost)   0.19609    0.14445   1.358  0.18243
log(Dwgs)    0.85879    0.22362   3.840  0.00044 ***
log(Length) -0.03844    0.15487  -0.248  0.80530
log(Spans)   0.23119    0.14068   1.643  0.10835
---
Residual standard error: 0.3139 on 39 degrees of freedom
Multiple R-squared: 0.7762,     Adjusted R-squared: 0.7475
F-statistic: 27.05 on 5 and 39 DF,  p-value: 1.043e-11
```

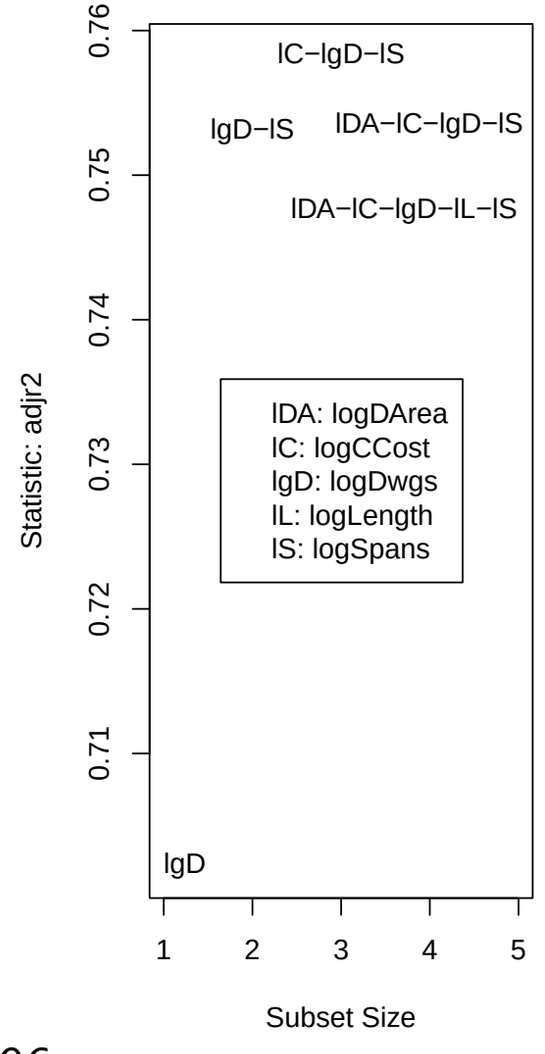
F-Statistik ist zwar hoch signifikant, aber nur 1 Slope ist signifikant (log(Dwgs) mit p-Wert < 0.001). Wir wollen daher eine Teilmenge der Prädiktoren wählen.

Identifiziere für eine feste Anzahl an Prädiktoren jenes Modell mit maximalem R^2_{adj} (minimalem SSE($\hat{\boldsymbol{\beta}}$)). Das Package `leaps` (functions for model selection) bietet hierzu einiges.

```
> logDArea <- log(DArea)
> logCCost <- log(CCost)
> logDwgs <- log(Dwgs)
> logLength <- log(Length)
> logSpans <- log(Spans)
> X <- cbind(logDArea, logCCost, logDwgs, logLength, logSpans)

> install.packages("leaps")
> library(leaps)
> b <- regsubsets(as.matrix(X),log(Time))
> (rs <- summary(b))
 Subset selection object
 5 Variables  (and intercept)
           Forced in Forced out
 logDArea      FALSE      FALSE
 logCCost      FALSE      FALSE
 logDwgs       FALSE      FALSE
 logLength     FALSE      FALSE
 logSpans      FALSE      FALSE
```

```
 1 subsets of each size up to 5
 Selection Algorithm: exhaustive
          logDArea logCCost logDwgs logLength logSpans
 1  ( 1 ) " "      " "      "*"     " "       " "
 2  ( 1 ) " "      " "      "*"     " "       "*"
 3  ( 1 ) " "      "*"      "*"     " "       "*"
 4  ( 1 ) "*"      "*"      "*"     " "       "*"
 5  ( 1 ) "*"      "*"      "*"     "*"       "*"

> par(mfrow=c(1,2))
> plot(1:5, rs$adjr2, xlab="Subset Size", ylab="Adjusted R-squared")
> rs$adjr2
 [1] 0.7022401 0.7530191 0.7582178 0.7534273 0.7475037

> library(car) # An R Companion to Applied Regression
> subsets(b, statistic=c("adjr2")) # to plot a regsubsets object
```


Adjusted R−squared
0.71
0.72
0.73
0.74
0.75
0.76
1
2
3
4
5
Subset Size


Statistic: adjr2
0.71
0.72
0.73
0.74
0.75
0.76
lC−lgD−lS
lgD−lS
lDA−lC−lgD−lS
lDA−lC−lgD−lL−lS
lDA: logDArea
lC: logCCost
lgD: logDwgs
lL: logLength
lS: logSpans
lgD
1
2
3
4
5
Subset Size

```
> om1 <- lm(log(Time)~log(Dwgs))
> om2 <- lm(log(Time)~log(Dwgs)+log(Spans))
> om3 <- lm(log(Time)~log(Dwgs)+log(Spans)+log(CCost))
> om4 <- lm(log(Time)~log(Dwgs)+log(Spans)+log(CCost)+log(DArea))
> om5 <- m1
> #Subset size=1
> n <- length(om1$residuals)
> npar <- length(om1$coefficients) + 1
> extractAIC(om1, k=2) #Calculate edf & AIC
[1]   2.00000  -94.89754
> extractAIC(om1, k=2) + 2*npar*(npar+1)/(n-npar-1) # Calculate edf & AICc
[1]   2.585366 -94.31217
> extractAIC(om1, k=log(n)) # Calculate edf & BIC
[1]   2.00000  -91.28421
```

u.s.w.

	Prädiktoren	R^2_{adj}	AIC	AIC_c	BIC
1	log(Dwgs)	0.702	-94.90	-94.31	-91.28
2	log(Dwgs)+log(Spans)	0.753	-102.37	**-101.37**	**-96.95**
3	log(Dwgs)+log(Spans)+log(CCost)	**0.758**	**-102.41**	-100.87	-95.19
4	log(Dwgs)+log(Spans)+log(CCost) +log(DArea)	0.753	-100.64	-98.43	-91.61
5	log(Dwgs)+log(Spans)+log(CCost) +log(DArea)+log(Length)	0.748	-98.71	-95.68	-87.87

Die Kriterien R^2_{adj} und AIC favorisieren ein Modell mit 3 Prädiktoren während die Kriterien AIC_c und BIC ein Modell mit 2 Prädiktoren als bestes einstufen.

Wir entscheiden uns wegen der zuvor erwähnten Gründe für das 2 Prädiktoren Modell.

```
> summary(om2)
             Estimate Std. Error t value Pr(>|t|)
 (Intercept)  2.66173    0.26871   9.905 1.49e-12 ***
 log(Dwgs)    1.04163    0.15420   6.755 3.26e-08 ***
 log(Spans)   0.28530    0.09095   3.137  0.00312 **
 ---
 Residual standard error: 0.3105 on 42 degrees of freedom
 Multiple R-squared: 0.7642,     Adjusted R-squared: 0.753
 F-statistic: 68.08 on 2 and 42 DF,  p-value: 6.632e-14
> summary(om3)
             Estimate Std. Error t value Pr(>|t|)
 (Intercept)   2.3317     0.3577   6.519  7.9e-08 ***
 log(Dwgs)     0.8356     0.2135   3.914 0.000336 ***
 log(Spans)    0.1963     0.1107   1.773 0.083710 .
 log(CCost)    0.1483     0.1075   1.380 0.175212
 ---
 Residual standard error: 0.3072 on 41 degrees of freedom
 Multiple R-squared: 0.7747,     Adjusted R-squared: 0.7582
 F-statistic: 46.99 on 3 and 41 DF,  p-value: 2.484e-13
```

Beide Prädiktoren sind signifikant im 2 Prädiktoren Modell während nur 1 Prädiktor im 3 Prädiktoren Modell signifikant ist. Das 3er Modell scheint also zu einem **overfit** der Daten zu führen und wir entscheiden uns daher für das 2er Modell.

• Schrittweises Auswählen:

Prüfe sequentielle Teilmenge aller 2^m möglichen Regressionsmodelle. Populäre Variationen dieser Methode sind die **backward elimination** und die **forward selection** Methode.

Backward Elimination: Starte mit allen möglichen Prädiktoren im Modell. Dann entferne in jedem Schritt jenen Prädiktor, so dass das resultierende Modell den kleinsten Wert des Informationskriteriums hat (dies entspricht jedesmal dem Entfernen jenes Prädiktors mit dem größten p-Wert). Fahre diese Strategie solange fort, bis alle Prädiktoren entfernt sind oder das Informationskriterium anwächst.

Forward Selection: Starte mit einem Modell, das nur den Intercept enthält. Dann nimm in jedem Schritt jenen Prädiktor ins Modell auf, so dass das resultierende Modell den kleinsten Wert des Informationskriteriums hat (dies entspricht jedesmal einem Aufnehmen jenes Prädiktors mit dem kleinsten p-Wert). Fahre diese Strategie solange fort, bis alle Prädiktoren beinhaltet sind oder das Informationskriterium anwächst.

Beide Verfahren betrachten maximal $m+(m-1)+\cdots+1 = m(m+1)/2$ mögliche Untermengen von Prädiktoren. Daher findet man nicht immer das beste Modell mit dem kleinsten Informationskriterium unter allen möglichen 2^m Modellen. Es gibt auch keine Garantie, dass beide Verfahren dasselbe Modell liefern.

Beispiel (Brückenkonstruktion):
Backward Elimination mit AIC Kriterium

```
> backAIC <- step(m1, direction="backward", data=bridge, k=2)
 Start:  AIC=-98.71
 log(Time) ~ log(DArea) + log(CCost) + log(Dwgs) + log(Length) + log(Spans)
                Df Sum of Sq    RSS      AIC
 - log(Length)  1   0.00607 3.8497 -100.640
 - log(DArea)   1   0.01278 3.8564 -100.562
 <none>                     3.8436  -98.711
 - log(CCost)   1   0.18162 4.0252  -98.634
 - log(Spans)   1   0.26616 4.1098  -97.698
 - log(Dwgs)    1   1.45358 5.2972  -86.277
```

```
Step:  AIC=-100.64
log(Time) ~ log(DArea) + log(CCost) + log(Dwgs) + log(Spans)

             Df Sum of Sq    RSS      AIC
- log(DArea)  1   0.01958 3.8693 -102.412
<none>                    3.8497 -100.640
- log(CCost)  1   0.18064 4.0303 -100.577
- log(Spans)  1   0.31501 4.1647  -99.101
- log(Dwgs)   1   1.44946 5.2991  -88.260

Step:  AIC=-102.41
log(Time) ~ log(CCost) + log(Dwgs) + log(Spans)

             Df Sum of Sq    RSS      AIC
<none>                    3.8693 -102.412
- log(CCost)  1   0.17960 4.0488 -102.370
- log(Spans)  1   0.29656 4.1658 -101.089
- log(Dwgs)   1   1.44544 5.3147  -90.128
```

Backward Elimination mit BIC Kriterium

```
> backBIC <- step(m1,direction="backward", data=bridge, k=log(n))
 Start:  AIC=-87.87
 log(Time) ~ log(DArea) + log(CCost) + log(Dwgs) + log(Length) + log(Spans)
                Df Sum of Sq    RSS     AIC
 - log(Length)  1   0.00607 3.8497 -91.607
 - log(DArea)   1   0.01278 3.8564 -91.529
 - log(CCost)   1   0.18162 4.0252 -89.600
 - log(Spans)   1   0.26616 4.1098 -88.665
 <none>                     3.8436 -87.871
 - log(Dwgs)    1   1.45358 5.2972 -77.244
```

```
Step:  AIC=-91.61
log(Time) ~ log(DArea) + log(CCost) + log(Dwgs) + log(Spans)

             Df Sum of Sq    RSS     AIC
- log(DArea)  1   0.01958 3.8693 -95.185
- log(CCost)  1   0.18064 4.0303 -93.350
- log(Spans)  1   0.31501 4.1647 -91.874
<none>                     3.8497 -91.607
- log(Dwgs)   1   1.44946 5.2991 -81.034

Step:  AIC=-95.19
log(Time) ~ log(CCost) + log(Dwgs) + log(Spans)

             Df Sum of Sq    RSS     AIC
- log(CCost)  1   0.17960 4.0488 -96.950
- log(Spans)  1   0.29656 4.1658 -95.669
<none>                     3.8693 -95.185
- log(Dwgs)   1   1.44544 5.3147 -84.708
```

```
Step:  AIC=-96.95
log(Time) ~ log(Dwgs) + log(Spans)

             Df Sum of Sq    RSS     AIC
<none>                     4.0488 -96.950
- log(Spans)  1    0.9487 4.9975 -91.284
- log(Dwgs)   1    4.3989 8.4478 -67.661
```

Backward Elimination mit BIC führt zu einem sparsameren Modell mit nur den beiden Prädiktoren `log(Dwgs)` und `log(Spans)`.

Forward Selection basierend auf AIC liefert dasselbe Modell wie Backward Elimination basierend auf AIC:

```
> mint <- lm(log(Time) ~ 1, data=bridge) # Intercept model as initial model
> forwardAIC <- step(mint,
  scope=list(lower=~1,
             upper=~log(DArea)+log(CCost)+log(Dwgs)+log(Length)+log(Spans)),
  direction="forward", data=bridge, k=2)
 Start:  AIC=-41.35
```

```
log(Time) ~ 1
              Df Sum of Sq     RSS     AIC
+ log(Dwgs)    1   12.1765  4.9975 -94.898
+ log(CCost)   1   11.6147  5.5593 -90.104
+ log(DArea)   1   10.2943  6.8797 -80.514
+ log(Length)  1   10.0120  7.1620 -78.704
+ log(Spans)   1    8.7262  8.4478 -71.274
<none>                     17.1740 -41.347

Step:  AIC=-94.9
log(Time) ~ log(Dwgs)
              Df Sum of Sq    RSS      AIC
+ log(Spans)   1   0.94866 4.0488 -102.370
+ log(CCost)   1   0.83170 4.1658 -101.089
+ log(Length)  1   0.66914 4.3284  -99.366
+ log(DArea)   1   0.47568 4.5218  -97.399
<none>                     4.9975  -94.898

Step:  AIC=-102.37
```

```
log(Time) ~ log(Dwgs) + log(Spans)
               Df Sum of Sq    RSS     AIC
+ log(CCost)    1  0.179598 3.8693 -102.41
<none>                      4.0488 -102.37
+ log(DArea)    1  0.018535 4.0303 -100.58
+ log(Length)   1  0.016924 4.0319 -100.56

Step:  AIC=-102.41
log(Time) ~ log(Dwgs) + log(Spans) + log(CCost)
               Df Sum of Sq    RSS     AIC
<none>                      3.8693 -102.41
+ log(DArea)    1  0.019578 3.8497 -100.64
+ log(Length)   1  0.012868 3.8564 -100.56
```

Interessante, hilfreiche R-Funktionen, um ein derartige Variablenselektion durchzuführen sind für die Rückwärtselimination:

```
> dropterm(om5, test="F")
 Single term deletions
 Model:
```

```
 log(Time) ~ log(DArea) + log(CCost) + log(Dwgs) + log(Length) + log(Spans)
             Df Sum of Sq    RSS      AIC F Value     Pr(F)
 <none>                    3.8436  -98.711
 log(DArea)   1   0.01278 3.8564 -100.562  0.1297 0.7207050
 log(CCost)   1   0.18162 4.0252  -98.634  1.8428 0.1824259
 log(Dwgs)    1   1.45358 5.2972  -86.277 14.7491 0.0004399 ***
 log(Length)  1   0.00607 3.8497 -100.640  0.0616 0.8052958
 log(Spans)   1   0.26616 4.1098  -97.698  2.7006 0.1083492
```

oder für die Forward Selection:

```
> addterm(om2, om5, test="F")
 Single term additions
 Model: log(Time) ~ log(Dwgs) + log(Spans)
             Df Sum of Sq    RSS     AIC F Value  Pr(F)
 <none>                    4.0488 -102.37
 log(DArea)   1  0.018535 4.0303 -100.58 0.18856 0.6664
 log(CCost)   1  0.179598 3.8693 -102.41 1.90308 0.1752
 log(Length)  1  0.016924 4.0319 -100.56 0.17209 0.6804
```

Stepwise Regression: Hierbei werden in jedem Schritt 4 Optionen betrachtet: nimm einen Prädiktor dazu, entferne einen Prädiktor, tausche einen Prädiktor im Modell gegen einen nicht im Modell, oder stoppe. Die Funktion `stepAIC` im Package `MASS` erlaubt gerade dieses Vorgehen.

```
> stepAIC(om1,
    scope=list(upper=~log(DArea)*log(CCost)*log(Dwgs)*log(Length)*log(Spans),
               lower=~1))
 Start:  AIC=-94.9
 log(Time) ~ log(Dwgs)
               Df Sum of Sq     RSS      AIC
 + log(Spans)   1    0.9487  4.0488 -102.370
 + log(CCost)   1    0.8317  4.1658 -101.089
 + log(Length)  1    0.6691  4.3284  -99.366
 + log(DArea)   1    0.4757  4.5218  -97.399
 <none>                      4.9975  -94.898
 - log(Dwgs)    1   12.1765 17.1740  -41.347
 Step:  AIC=-102.37
 log(Time) ~ log(Dwgs) + log(Spans)
```

```
                        Df Sum of Sq    RSS      AIC
+ log(CCost)             1    0.1796 3.8693 -102.412
<none>                                4.0488 -102.370
+ log(Dwgs):log(Spans)   1    0.0428 4.0060 -100.849
+ log(DArea)             1    0.0185 4.0303 -100.577
+ log(Length)            1    0.0169 4.0319 -100.559
- log(Spans)             1    0.9487 4.9975  -94.898
- log(Dwgs)              1    4.3989 8.4478  -71.274
Step:  AIC=-102.41
log(Time) ~ log(Dwgs) + log(Spans) + log(CCost)

                        Df Sum of Sq    RSS      AIC
<none>                                3.8693 -102.412
- log(CCost)             1   0.17960 4.0488 -102.370
- log(Spans)             1   0.29656 4.1658 -101.089
+ log(Dwgs):log(Spans)   1   0.02231 3.8469 -100.672
+ log(DArea)             1   0.01958 3.8497 -100.640
+ log(CCost):log(Dwgs)   1   0.01889 3.8504 -100.632
+ log(Length)            1   0.01287 3.8564 -100.562
```

```
+ log(CCost):log(Spans)  1   0.00041 3.8688 -100.417
- log(Dwgs)              1   1.44544 5.3147  -90.128

Coefficients:
(Intercept)    log(Dwgs)   log(Spans)   log(CCost)
     2.3317       0.8356       0.1963       0.1483
```

Verwendet man hingegen das BIC Kriterium, dann liefert dies jetzt auch das Modell mit den 2 Prädiktoren:

```
> stepAIC(om1,
   scope=list(upper=~log(DArea)*log(CCost)*log(Dwgs)*log(Length)*log(Spans),
              lower=~1), k=log(n))
Start:  AIC=-91.28
log(Time) ~ log(Dwgs)

               Df Sum of Sq     RSS     AIC
+ log(Spans)    1    0.9487  4.0488 -96.950
+ log(CCost)    1    0.8317  4.1658 -95.669
+ log(Length)   1    0.6691  4.3284 -93.946
```

```
+ log(DArea)    1    0.4757  4.5218 -91.979
<none>                       4.9975 -91.284
- log(Dwgs)     1   12.1765 17.1740 -39.540

Step:  AIC=-96.95
log(Time) ~ log(Dwgs) + log(Spans)
                        Df Sum of Sq    RSS     AIC
<none>                                4.0488 -96.950
+ log(CCost)             1    0.1796 3.8693 -95.185
+ log(Dwgs):log(Spans)   1    0.0428 4.0060 -93.622
+ log(DArea)             1    0.0185 4.0303 -93.350
+ log(Length)            1    0.0169 4.0319 -93.332
- log(Spans)             1    0.9487 4.9975 -91.284
- log(Dwgs)              1    4.3989 8.4478 -67.661

 Coefficients:
(Intercept)    log(Dwgs)   log(Spans)
     2.6617       1.0416       0.2853
```