SBWL Tourismusanalyse und Freizeitmarketing,

Vertiefungskurs 2: Multivariate Verfahren 1

2. Teil: Multiples Regressionsmodell

Regina Tüchler

Department für Statistik & Mathematik

Inhalt

Inhalt

Inhalt

Inhalt

Inhalt

Einleitung

Wir beginnen mit dem einfachsten Fall eines linearen Regressionsmodells, wo nur 2 Variable X und Y im Modell sind. Dieses Modell ist schon aus Statistik 1 bekannt. Da wurde einerseits die Korrelation zwischen X und Y betrachtet, als Maßzahl für die Stärke des linearen Zusammenhangs. Weiters ging es bei der linearen Regression darum, die Responsevariable Y durch die Prädiktorvariable X zu erklären. Die einfache lineare Regression wird in Abschnitt 1. wiederholt.

In Abschnitt 3. wird dieses Modell dann auf mehrere Responsevariable erweitert. Für dieses multiple Regressionsmodell werden wir wichtige statistische Annahmen kennenlernen (Abschnitt 4.),

Einleitung

die es ermöglichen die Genauigkeit der Schätzungen anzugeben (Abschnitt 5.) und mit deren Hilfe eine Auswahl aus verschiedenen Regressionsmodellen getroffen werden kann (Abschnitt 5., 7.).

Weitere Methoden des Modellvergleichs werden in Abschnitt 10. beschrieben.

1.1. Einfaches lineares Regressionsmodell

Im einfachen Regressionsmodell wird eine **Responsevariable** (**abhängige** Variable) Y durch eine **Prädiktorvariable** (**unabhängige** oder **erklärende** Variable) X erklärt.

Gegeben sei eine Stichprobe vom Umfang T:

Liste von Datenpaaren $(x_i, y_i), i = 1, \ldots, T$.

Diese Daten können in ein Streudiagramm eingetragen werden.

Beispiel: Der Bremsweg eines Autos hängt von der Geschwindigkeit ab. In Fig. 1 werden Daten aus den 20-er Jahren mit einem einfachen linearen Regressionsmodell modelliert. Wie die Abbildung zeigt, war damals ein linearer Zusammenhang noch passend.

1.1. Einfaches lineares Regressionsmodell

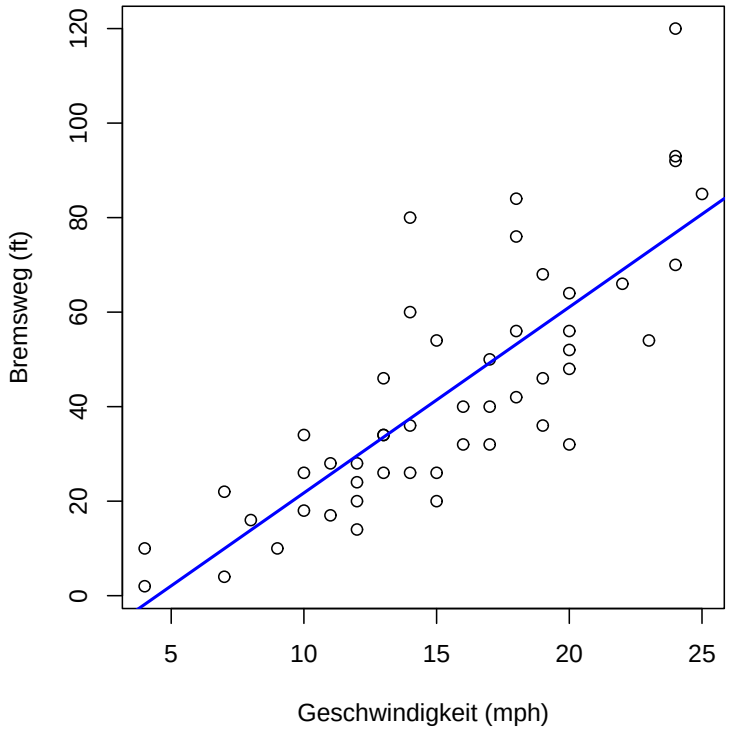


Abbildung 1: Abhängigkeit des Bremswegs von der Geschwindigkeit, (20-er Jahre)

1.1. Einfaches lineares Regressionsmodell

Ein **einfaches Regressionsmodell** ist ein Modell, in dem ein Prädiktor X die Responsevariable Y erklärt:

$$Y = f(X) + \varepsilon$$

$f(X)$ ist hier der **Strukturanteil**, der den Einfluss der Prädiktorvariablen modelliert

ε ist **zufällige Störung (Residuen)**, nicht erklärbarer Anteil mit Erwartungswert $E(\varepsilon) = 0$

$f(X)$ heißt **Regressionsfunktion**.

1.1. Einfaches lineares Regressionsmodell

Im **einfachen linearen Regressionsmodell** ist die Regressionsfunktion $f(X) = \beta_1 + \beta_2 X$ eine **Regressionsgerade**.

$$Y \quad = \quad \beta_1 + \beta_2 X + \varepsilon.$$

Daher erwartet man sich für die Responsevariable Y:

$$E(Y) \quad = \quad \beta_1 + \beta_2 x.$$

Die statistische Aufgabe ist das Schätzen der **Parameter** β_1 und β_2 aus Daten.

1.2. Prinzip der kleinsten Quadrate

Die Parameter werden mit statischen Mitteln so geschätzt, dass die Daten y_i durch die lineare Funktion $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ möglichst gut vorhergesagt werden. Als Bezeichnung für den Schätzer wird $\hat{\beta}_1$ und $\hat{\beta}_2$ verwendet.

- $\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$ heißt **Schätzwert**,

- $y_i - \hat{y}_i$ heißt **Prognosefehler**. Sie sind die **Residuen** in der Modellgleichung.

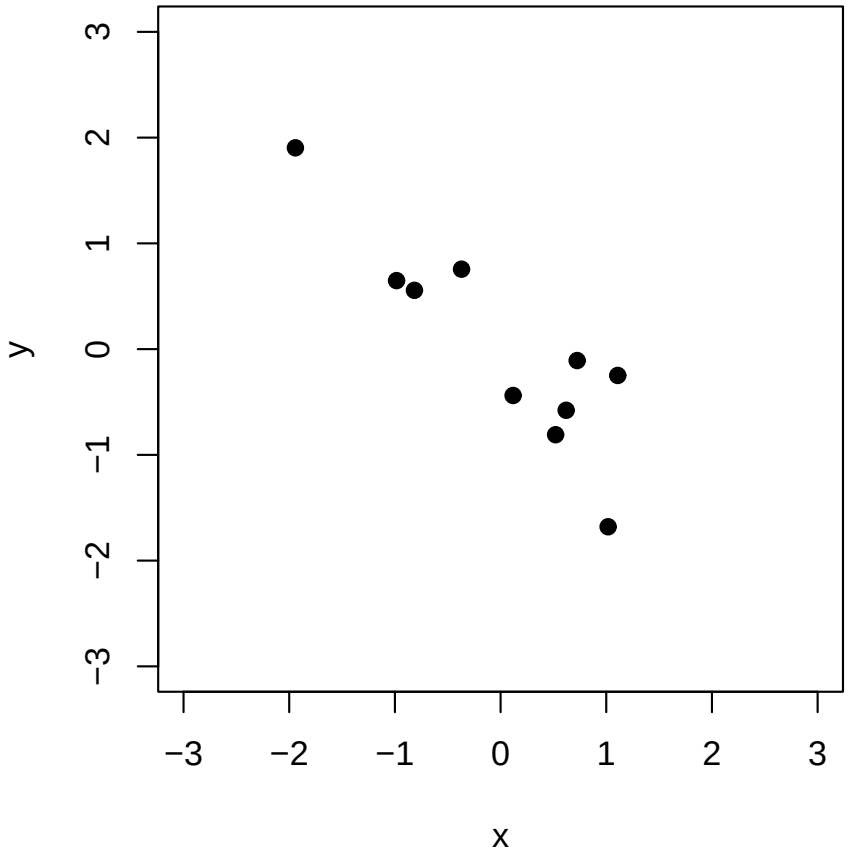
Die Methode soll so gewählt werden, dass die Residuen möglichst klein werden.

1.2. Prinzip der kleinsten Quadrate

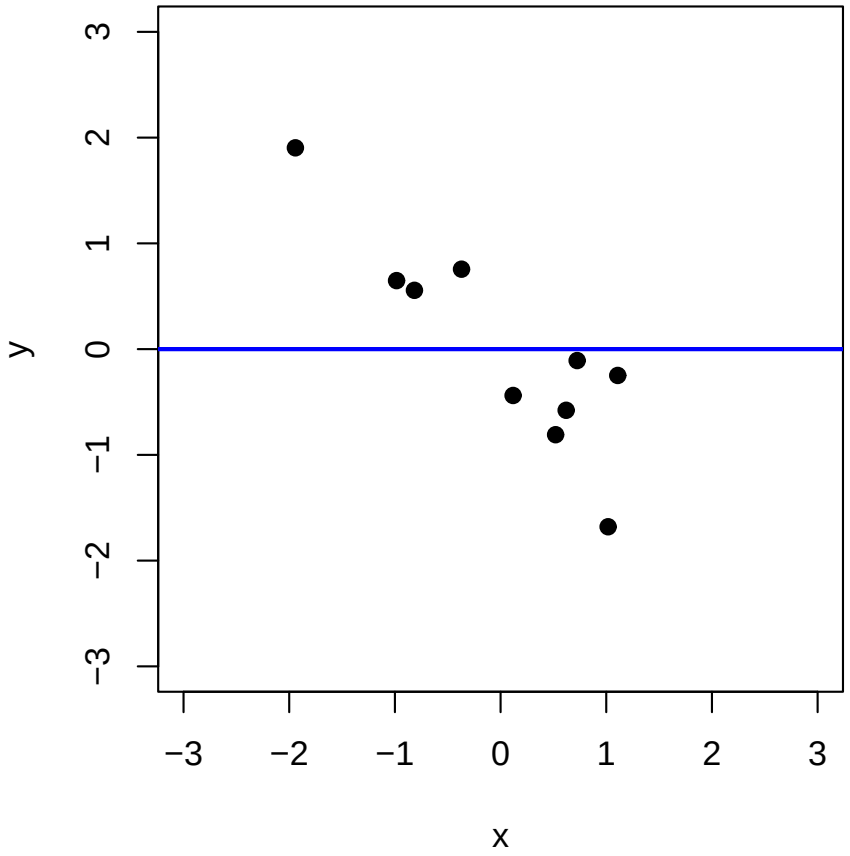
Nach dem **Prinzip der kleinsten Quadrate** (LSQ, KQ oder OLS) werden die Modellparameter β_1 und β_2 so gewählt, dass die Fehlerquadratsumme des Prognosefehlers SS_R minimal wird.

$$\begin{aligned} SS_R &= \sum_{i=1}^{T} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{T} (y_i - (\beta_1 + \beta_2 x_i))^2 \ \rightarrow \ \min \end{aligned}$$

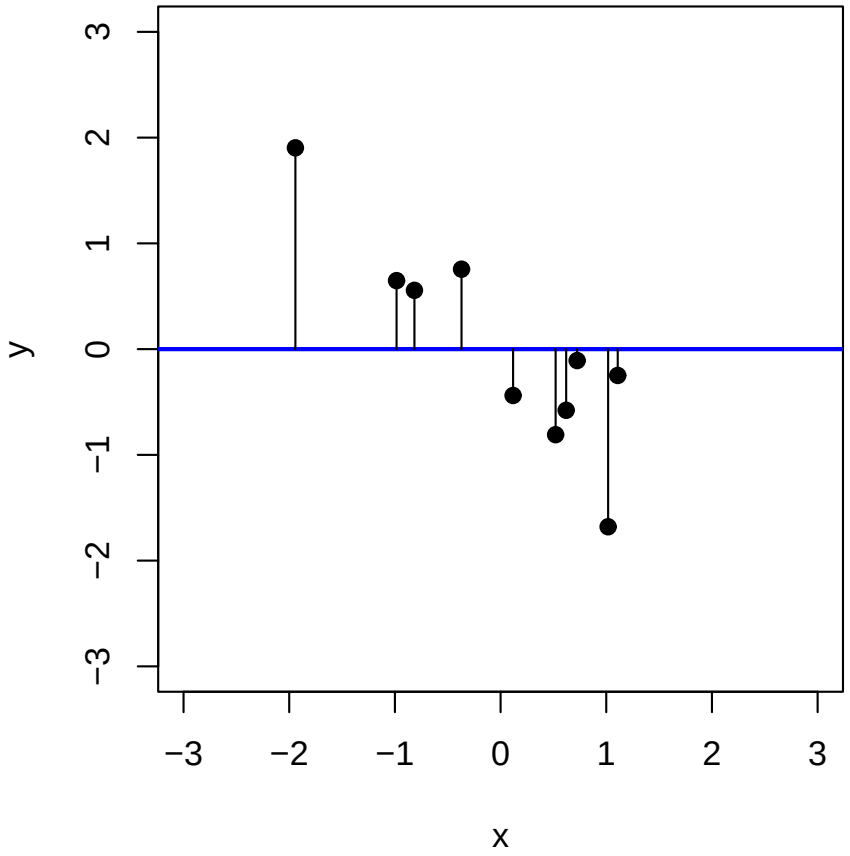
1.2. Prinzip der kleinsten Quadrate



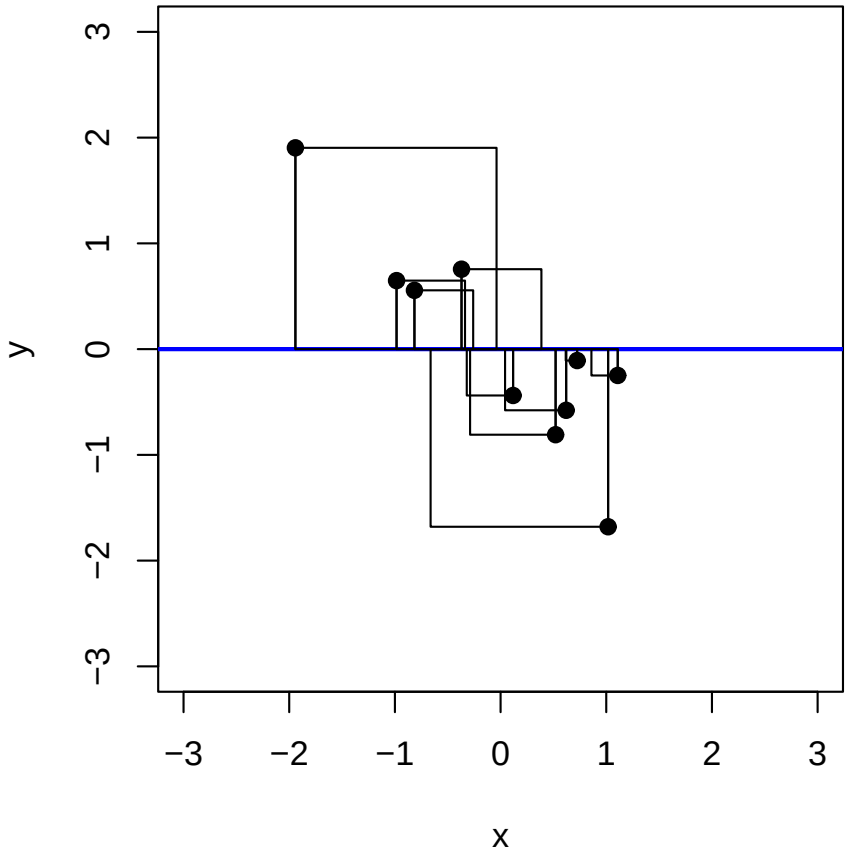
1.2. Prinzip der kleinsten Quadrate



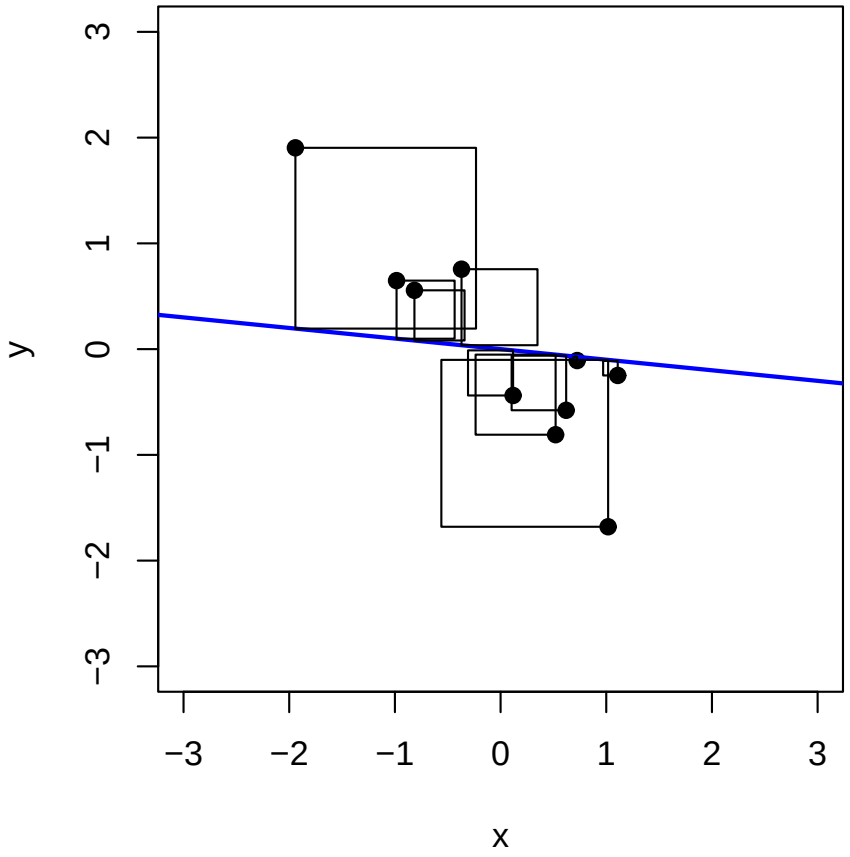
1.2. Prinzip der kleinsten Quadrate



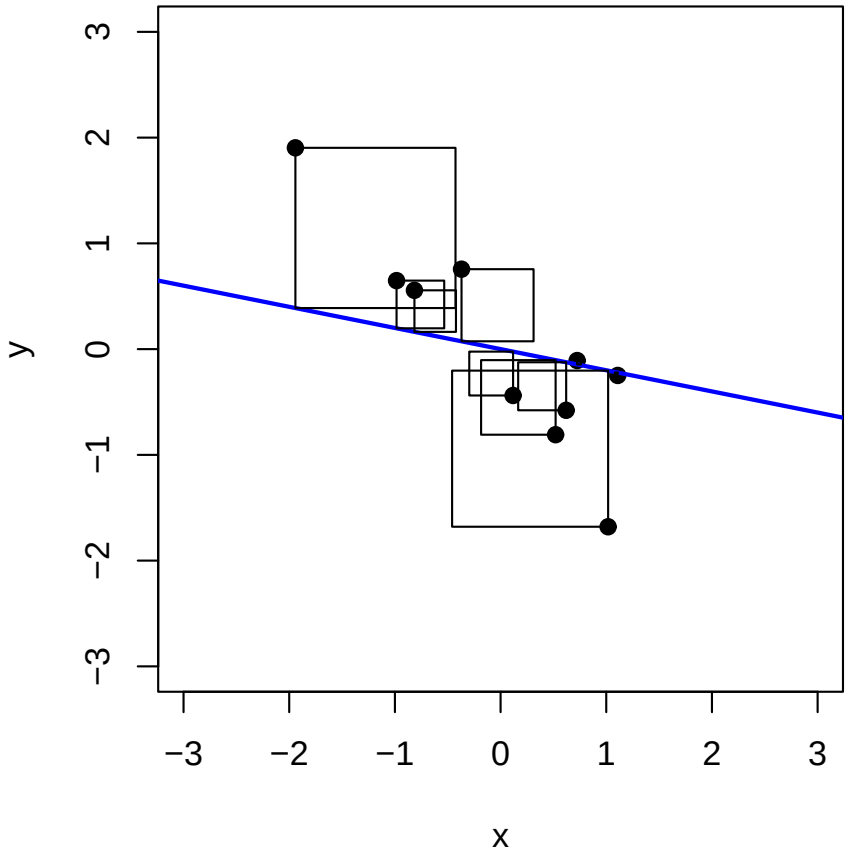
1.2. Prinzip der kleinsten Quadrate



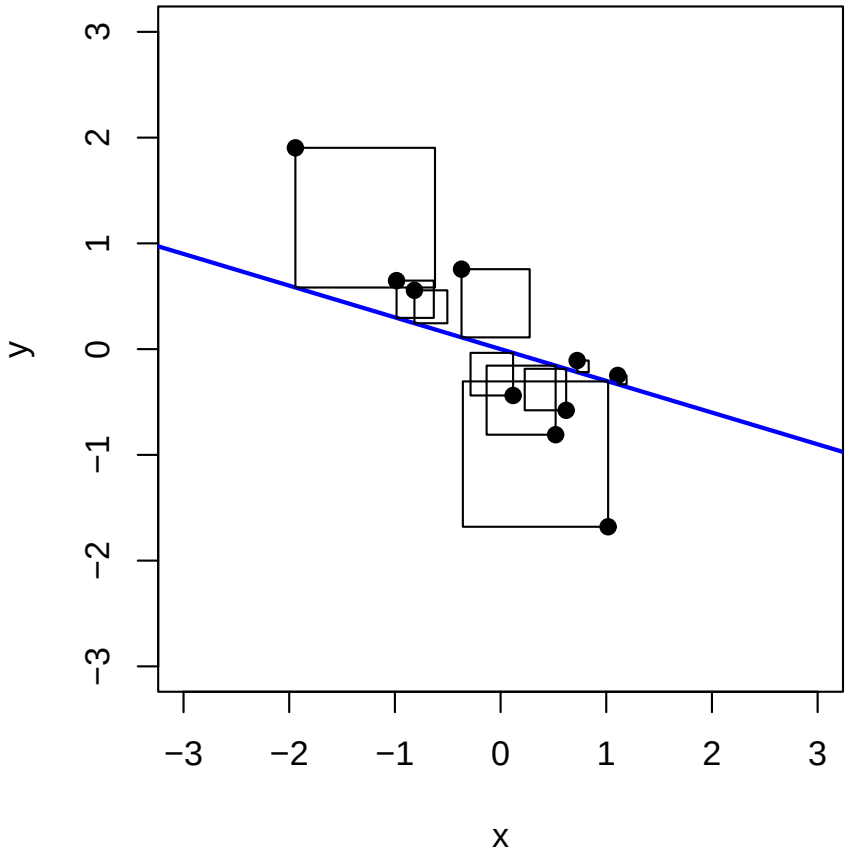
1.2. Prinzip der kleinsten Quadrate



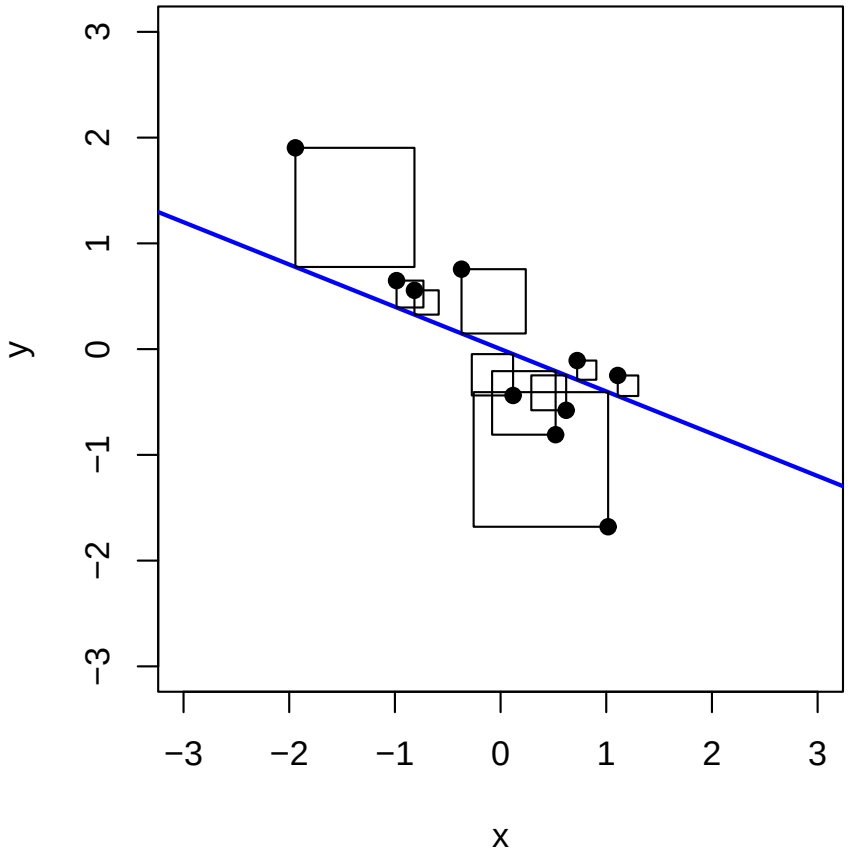
1.2. Prinzip der kleinsten Quadrate



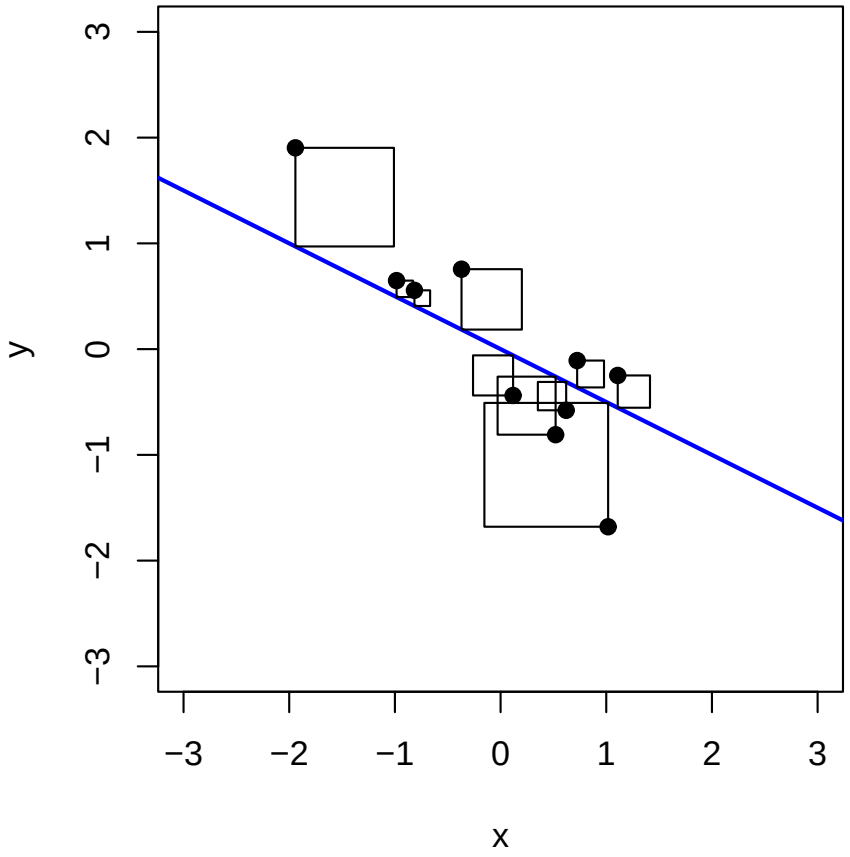
1.2. Prinzip der kleinsten Quadrate



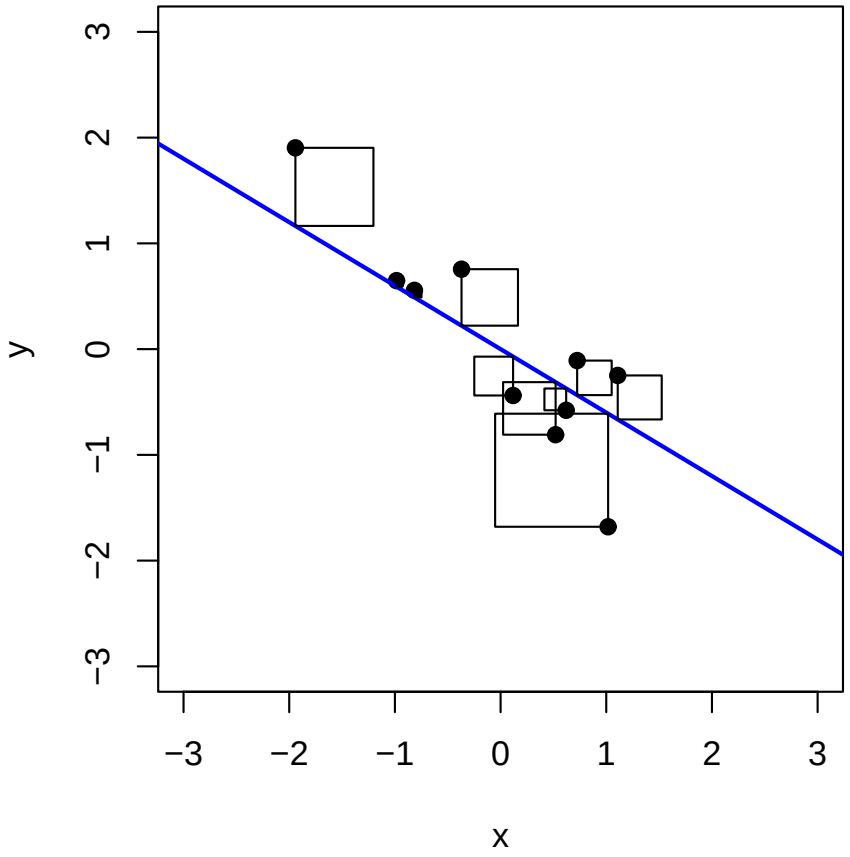
1.2. Prinzip der kleinsten Quadrate



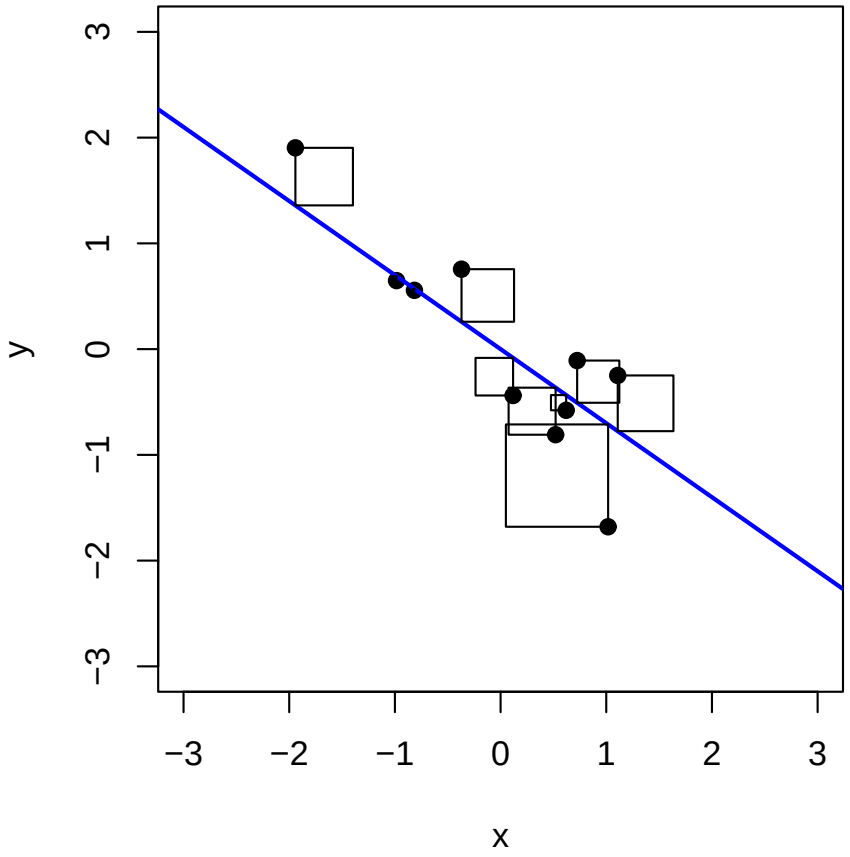
1.2. Prinzip der kleinsten Quadrate



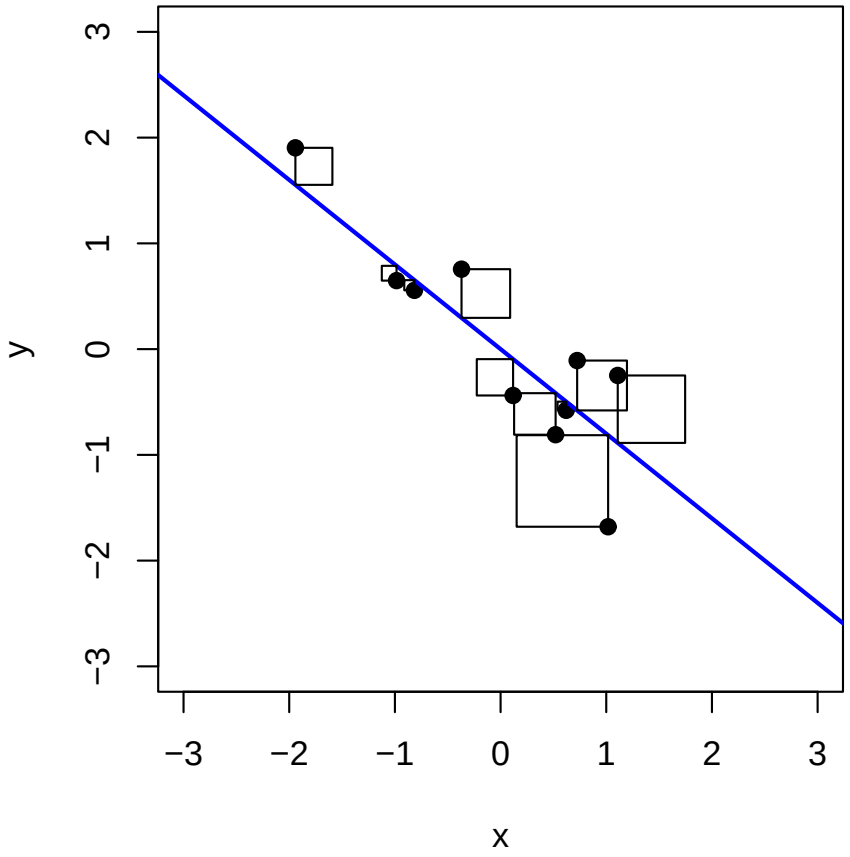
1.2. Prinzip der kleinsten Quadrate



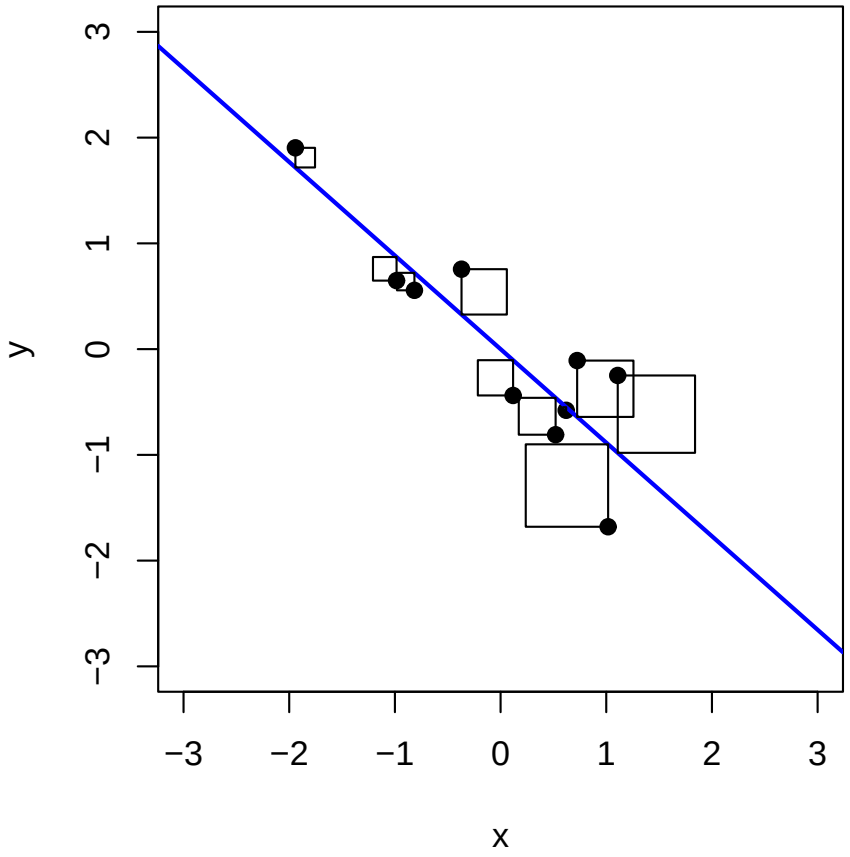
1.2. Prinzip der kleinsten Quadrate



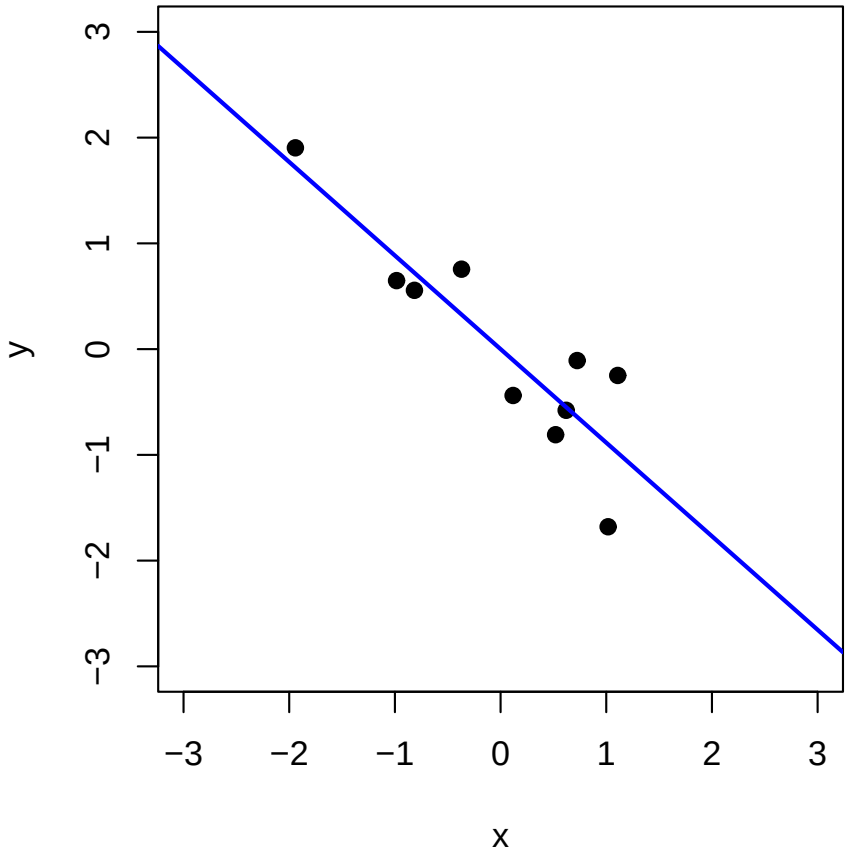
1.2. Prinzip der kleinsten Quadrate



1.2. Prinzip der kleinsten Quadrate



1.2. Prinzip der kleinsten Quadrate



1.2. Prinzip der kleinsten Quadrate

KQ-Schätzer im einfachen linearen Regressionsmodell:

Wiederholung aus Statistik 1:

SS_R minmal, wenn

$$\begin{aligned} \hat{\beta}_2 &= r\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2} \\ \hat{\beta}_1 &= \bar{y} - \beta_2 \bar{x} \end{aligned}$$

r .. empirischer Korrelationskoeffizient zwischen X und Y

s_x, s_y .. Standardabweichung von X bzw. Y

$f(x) = \hat{\beta}_1 + \hat{\beta}_2 x$ heißt empirische Regressionsgerade.

2. Einfaches lineares Regressionsmodell in R

Im '`EinfRegrinR.pdf`' von der LV-Seite gibt es eine Beschreibung der R-Befehle dieses Abschnitts.

Übung 1

1. Rechnen Sie das Beispiel zur einfachen linearen Regression in R (`EinfRegrinR.pdf`) nach.

2. Laden Sie die Daten `statlab` von der LV-Seite. (Sie finden dort die '`.rda`'-Datei mit den Daten, das sie mit `load()` in R laden können. Eine Beschreibung steht in der '`.htm`'-Datei. Um die Variablen unter ihrem Namen *** ohne Angabe von `statlab$***` ansprechen zu können, muss man `attach(statlab)` nach dem Laden der Daten eingeben.) Berechnen Sie die Regressionsgerade mit der Responsevariable `CTHGHT` und der unabhängigen Variable `MTHGHT`.

Übung 1

Wie lauten die Regressionskoeffizienten, die Schätzungen für die Responsevariable und die Residuen. Wie lautet die Prognose für die Größe eines Kindes von eine 55 bzw. 70 inch großen Mutter?

3. Zeichnen Sie das Streudiagramm für die Daten aus Beispiel 2. und zeichnen Sie die Regressionsgerade ein.

3.1. Das klassische lineare Regressionsmodell

Bisher hatten wir 2 Zufallsvariablen X und Y, die mit einer Regressionsgeraden modelliert wurden.

Jetzt wird dieses Modell erweitert. Üblicherweise steht nicht nur eine erklärende Variable zur Verfügung sondern mehrere Variable, die im multiplen Regressionsmodell als Prädiktoren verwendet werden. Solche Modelle werden jetzt betrachtet.

3.1. Das klassische lineare Regressionsmodell

Es stehen k Prädiktorvariablen $X_1, \ldots, X_k$ zur Erklärung der Responsevariablen Y zur Verfügung.

Im einfachen linearen Regressionsmodell hatten wir eine Regressionsgerade mit einem Parameter für die Konstante und einem für die Steigung. Auch im multiplen Regressionsmodell wird i. allg. eine Konstante einbezogen. Diese Konstante sei gleich X_1, daher $X_1 = 1$ in der Modellgleichung.

Gegeben sei wieder eine Stichprobe der Größe T. Das lineare Regressionsmodell lautet für $i = 1, \ldots, T$:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + \varepsilon_i.$$

3.1. Das klassische lineare Regressionsmodell

Die Parameter $\beta_1, \ldots, \beta_k$ werden wieder gemäß dem KQ-Prinzip geschätzt.

$$SS_R = \sum_{i=1}^{T} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{T} (y_i - (\beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}))^2 \rightarrow \min$$

Wir werden in Abschnitt 3.3 jene Annahmen kennenlernen, die sicher stellen, dass wir die KQ-Schätzung durchführen könnnen, und danach in Abschnitt 3.4 die KQ-Schätzung für das multiple Regressionsmodell angeben.

3.1. Das klassische lineare Regressionsmodell

Wie entsteht ε?

- ε aggregiert Variablen, die keinen Eingang ins Modell gefunden haben,
  - weil Einfluss a priori nicht bekannt ist.
  - weil keine Beobachtungen vorliegen.
  - weil die Variable schwierig zu quantifizieren ist.
- ε aggregiert Messfehler, die durch Quantifizierung von ökonomischen Variablen entstehen.
- ε steht für eine der Variablen Y immanente Zufälligkeit, die durch keine anderen Variablen erklärt werden kann.

3.2. Regressions- und lineares Modell

Anmerkungen zum linearen Regressionsmodell:

- Der für Y erwartete Wert bei fest vorgegebenen Prädiktorvariablen ist gleich dem Strukturanteil des linearen Regressionsmodells:

$$E(Y|X_2 = x_{i2}, \ldots, X_k = x_{ik}) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}$$

- Es muss eine **lineare** Funktion in den Parametern $\beta_1, \ldots, \beta_k$ vorliegen.

- Das Modell kann eine nichtlineare Funktion in den Prädiktorvariablen sein.

3.2. Regressions- und lineares Modell

- Die Responsevariable Y kann durch Transformation aus einer anderen Variablen Y^* entstanden sein $(Y = g(Y^*))$.

- Das Regressionsmodell kann durch Einführung neuer Variablennamen als lineares Modell geschrieben werden:

$$\log\, y_i^* = \beta_1 + \beta_2 \log\, x_{i2}^* + \varepsilon_i$$

wird durch $y_i = \log\, y_i^*$ und $x_{i2} = \log\, x_{i2}^*$ zu

$$y_i = \beta_1 + \beta_2 x_{i2} + \varepsilon_i$$

3.2. Regressions- und lineares Modell

Unter einem **linearen Modell** versteht man ein Modell das linear in den Prädiktorvariablen ist.

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + \varepsilon_i,\ i = 1, \ldots, T$$

Die Regressionskoeffizienten $\beta_j,\ j \geq 2$ quantifizieren die **zu erwartende absolute Veränderung** von Y, wenn sich die Prädiktorvariable X_j **absolut** um eine Einheit ändert und alle anderen Prädiktorvariablen $X_{j^*}, j^* \neq j$ gleich bleiben.

3.2. Regressions- und lineares Modell

- Für $\beta_j > 0$ führt eine Steigerung (Reduktion) von X_j im Mittel zu einer Steigerung (Reduktion) von Y.
- Für $\beta_j < 0$ führt eine Steigerung (Reduktion) von X_j im Mittel zu einer Reduktion (Steigerung) von Y.
- Für $\beta_j = 0$ hat X_j keinen Einfluss auf Y.

3.3. Standardannahmen im klass. lin. Regr.modell

In diesem Abschnitt werden die Standardannahmen des klassischen linearen Regressionsmodells vorgestellt. Diese Annahmen stellen sicher, dass das Modell sinnvoll für die Modellierung der Daten verwendet werden kann und dass die Voraussetzungen für die mathematische Durchführbarkeit der KQ-Schätzung gegeben sind.

3.3. Standardannahmen im klass. lin. Regr.modell

A1. Das Modell hat keinen systematischen Fehler.

A2. Die Fehlervarianz ist für alle Beobachtungen gleich groß.

A3. Die Komponenten des Fehler sind nicht korreliert.

A4. Die Prädiktorvariablen sind exogen und fest vorgegeben.

A5. Es bestehen keine linearen Abhängigkeiten zwischen den Variablen.

3.3. Standardannahmen im klass. lin. Regr.modell

Die Standardannahme A1.

Das Modell hat keinen systematischen Fehler. Der Erwartungswert des Fehlers ist 0:

$$E(\varepsilon_i) = 0$$

für alle i.

Wenn A1 verletzt ist, besitzt das Modell einen **Spezifikationsfehler**.

A1 häufig dann verletzt, wenn eine Prädiktorvariable nicht berücksichtigt wurde.

Gegenbeispiel:
Im wahren Modell hängt die Responsevariable Y von den Prädiktoren X und X^2 ab:

$$y_i = 1 + x_i^2 - 6x_i + \varepsilon_i$$

Wir schätzen aber ein lineares Modell, in dem nur X als Prädiktor vorkommt. Laut KQ-Schätzung ergibt sich dafür:

$$y_i = -22 + 4.9x_i + \varepsilon_i^*$$

3.3. Standardannahmen im klass. lin. Regr.modell

Hier wird die Annahme A1. verletzt, weil für den erwarteten Fehler, wenn als Schätzung $\hat{y}_i$ aus dem linearen Modell verwendet wird, gilt:

$$E(\varepsilon_i^*) = y_i - \hat{y}_i = -21 - x_i^2 + 10.9x_i \neq 0$$

In Abb. 2 sieht man die Daten (schwarz) und ihre Schätzung mit dem linearen Modell (rot). Aus dieser Abbildung und aus Abb. 3 wird deutlich, dass A1 verletzt ist:

- In den Randbereichen von X ist $E(\varepsilon_i) > 0$, sodass Y systematisch unterschätzt wird.
- Im mittleren Bereich von X ist $E(\varepsilon_i) < 0$, sodass Y systematisch überschätzt wird.

3.3. Standardannahmen im klass. lin. Regr.modell

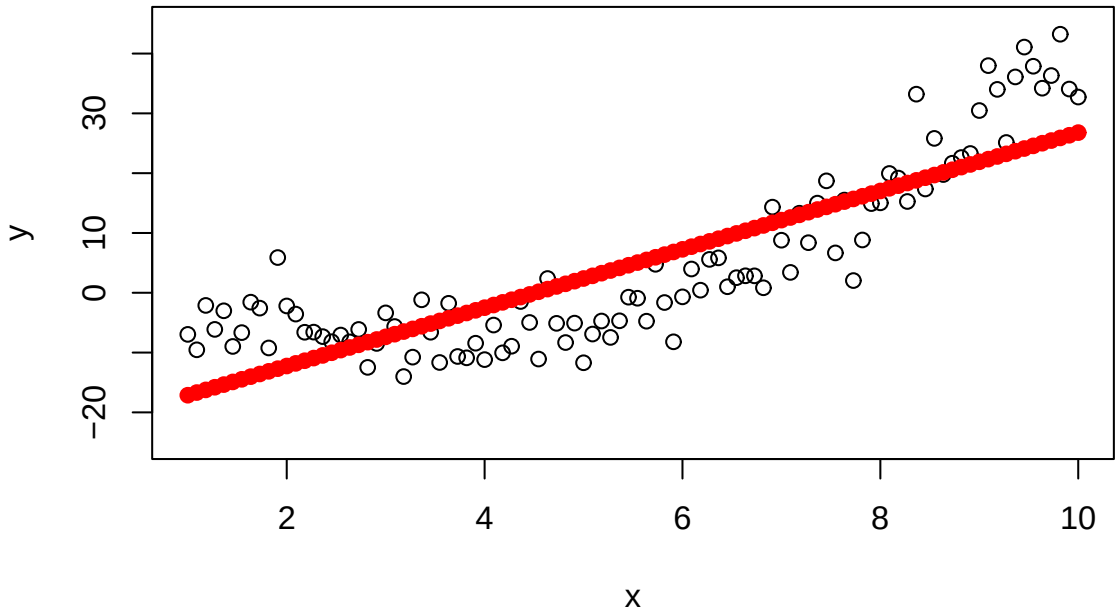


Abbildung 2: Gegenbeispiel zu A1.

3.3. Standardannahmen im klass. lin. Regr.modell

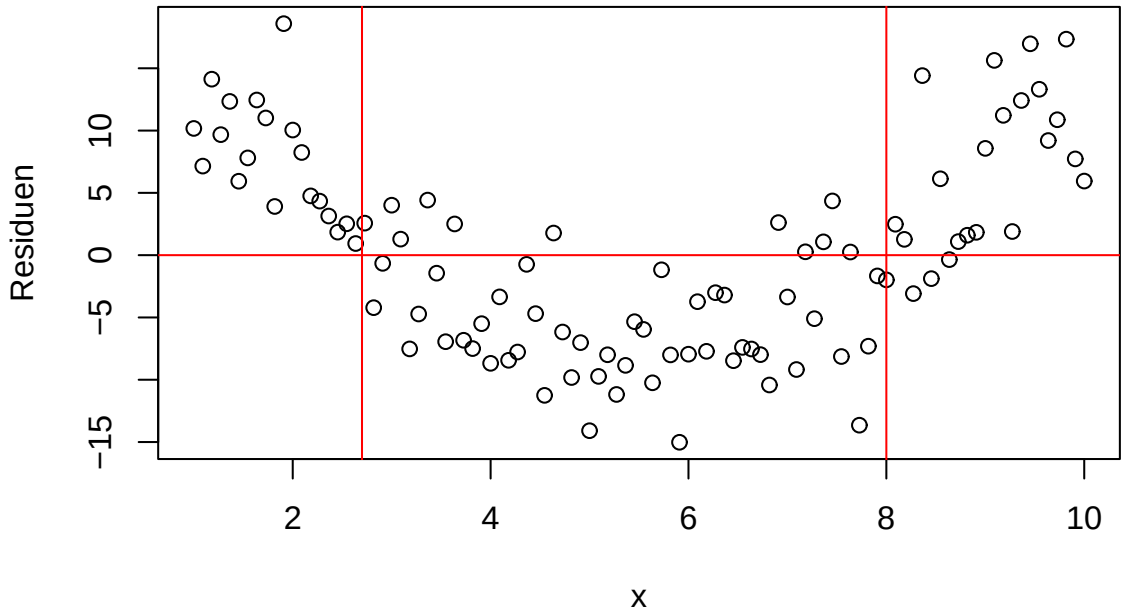


Abbildung 3: Residuen zum Gegenbeispiel zu A1.

Die Standardannahme A2.

Die Varianz des Modellfehlers ε_i ist **homoskedastisch**, d.h. gleich für alle Beobachtungen.

$$V(\varepsilon_i) = \sigma^2$$

für alle i.

Wenn A2 verletzt ist, liegen **heteroskedastische** Fehler vor.

A2 ist häufig dann verletzt, wenn die Varianz mit den Werten der Prädiktorvariablen variiert.

Gegenbeispiel:
Im wahren Modell hängt der Fehler von der Prädiktorvariable ab:

$$y_i = 0.2 + 0.8x_i + \varepsilon_i, \quad \varepsilon_i = 0.05x_i^2 u_i$$

u_i sind unabhängig identisch verteilt mit Erwartungswert 0.

In Abb. 4 sehen wir die Daten und die Regressionsgerade

$$y_i = 0.2 + 0.8x_i + \varepsilon_i^*$$

3.3. Standardannahmen im klass. lin. Regr.modell

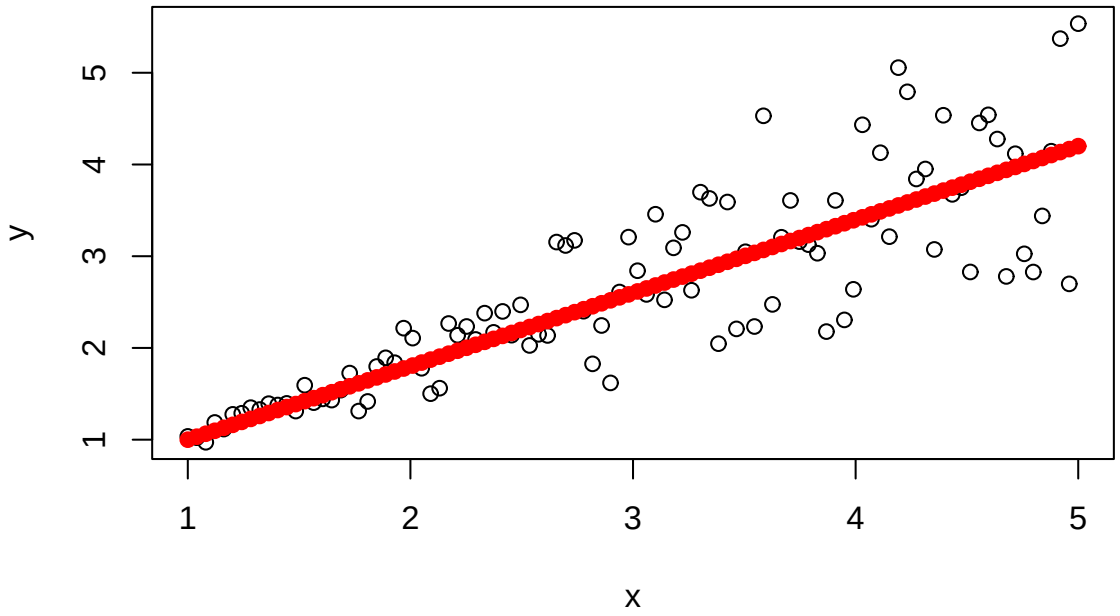


Abbildung 4: Gegenbeispiel zu A2.

3.3. Standardannahmen im klass. lin. Regr.modell

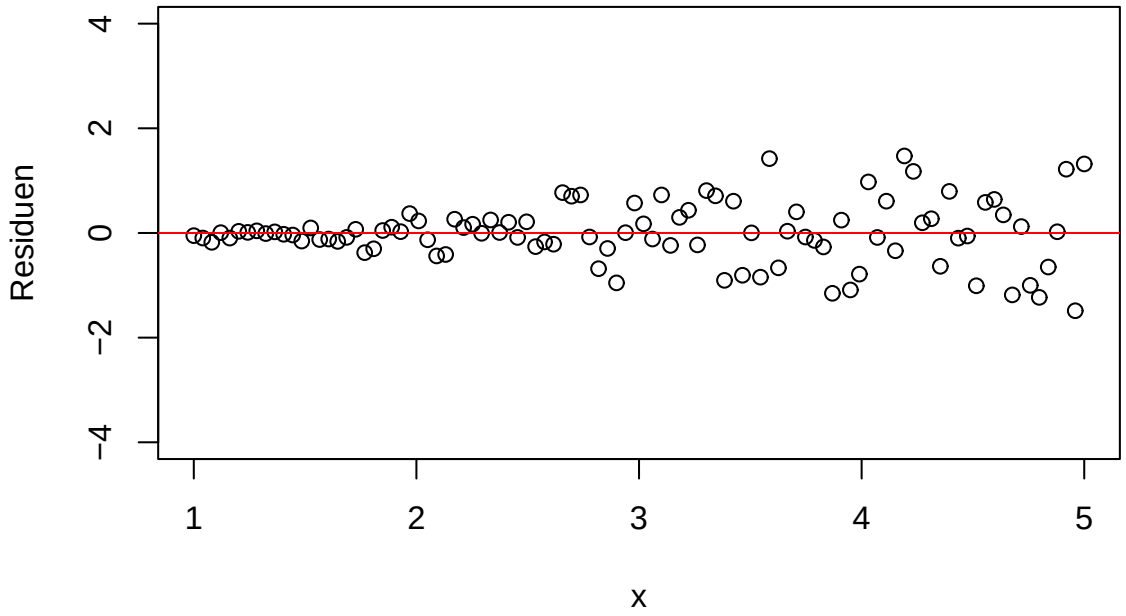


Abbildung 5: Residuen zum Gegenbeispiel zu A2.

3.3. Standardannahmen im klass. lin. Regr.modell

Die Standardannahme A3.

Die Komponenten des Fehlers sind nicht korreliert.

$$COV(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \cdot \varepsilon_j) = 0$$

für alle $i \neq j$.

D.h. Der zufällige Anteil bei der Beobachtung y_i hat keinen Einfluß auf die Größe oder das Vorzeichen des zufälligen Anteils bei der Beobachtung y_j.

Bei Verletzung von A3. liegt **Autokorrelation des Fehlers** vor.

Gegenbeispiel:

Im wahren Modell sind die Fehler von aufeinanderfolgenden Beobachtungen nicht unkorreliert, sondern es liegt eine positive Autokorrelation vor. Der Fehler ε_i im wahren Modell setzt sich als Summe von unkorrelierten Variablen u_i mit Erwartungswert 0 und Varianz σ_u^2 zusammen:

$$y_i = 0.2 + 0.8x_i + \varepsilon_i, \quad \varepsilon_i = u_i + u_{i-1}$$

für $i = 1, \ldots, T$.

Es gilt: $COV(\varepsilon_i, \varepsilon_{i-1}) = E(\varepsilon_i \cdot \varepsilon_{i-1}) = \sigma_u^2 \neq 0$

3.3. Standardannahmen im klass. lin. Regr.modell

Für Daten (x_i, y_i) aus obigem Modell wurde die Regressionsgerade

$$y_i = 0.2 + 0.8x_i + \varepsilon_i^*$$

in Abb. 6 gezeichnet (Daten-schwarz, ihre KQ-Schätzung-rot). In Abb. 7 sieht man die positive Autokorrelation Lag 1 des Fehlers.

3.3. Standardannahmen im klass. lin. Regr.modell

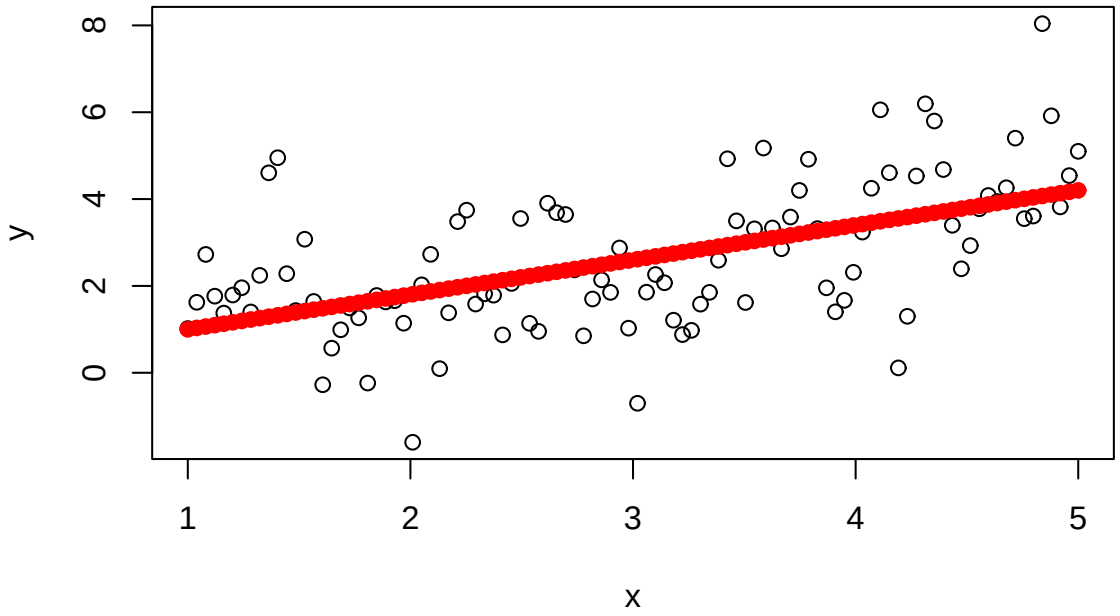


Abbildung 6: Gegenbeispiel zu A3.

3.3. Standardannahmen im klass. lin. Regr.modell

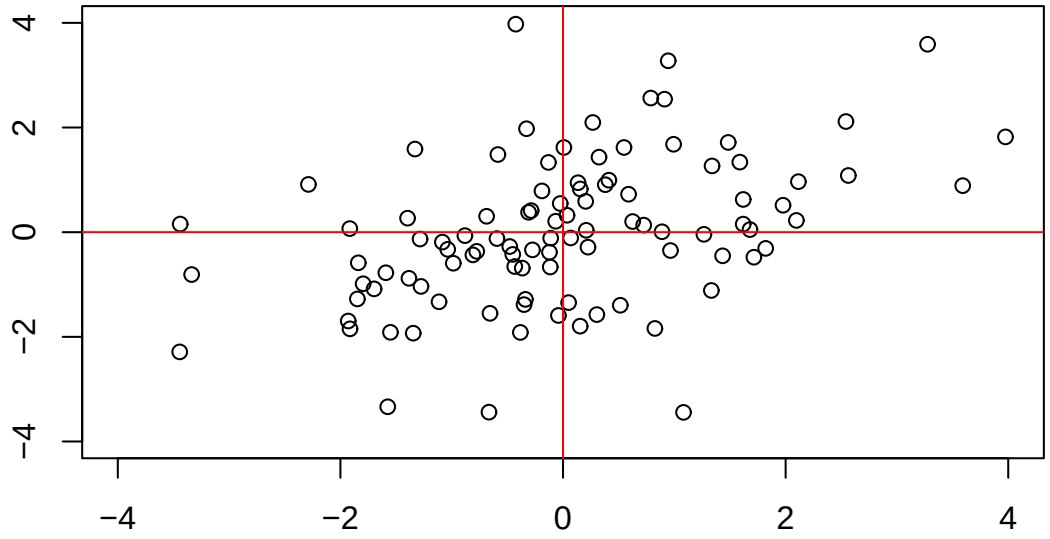


Abbildung 7: ε^*_{i-1} gegen ε^*_i.

Die Standardannahme A4.

Die Prädiktorvariablen sind exogen und fest vorgegeben.

$$COV(X_{il}, \varepsilon_i) = E(X_{il}, \varepsilon_i) = 0$$

für alle $i = 1, \ldots, T, l = 1, \ldots, k$.

Es besteht keine Abhängigkeit zwischen den Prädiktorvariablen und den zufälligen Anteilen.

Wenn A4 verletzt ist, kann der systematische Anteil und der zufällige Anteil nicht getrennt werden. Typischerweise ist das dann der Fall, wenn **zeitverzögerte Responsevariable** als Prädiktorvariable vorkommen.

Gegenbeispiel:
Im wahren Modell werden die Prädiktoren aus zeitverzögerten Responsevariablen gebildet.

$$y_i = 0.8 y_{i-1} + \varepsilon_i$$

In Abb. 8 sieht man die Regressionsgerade mit den Prädiktoren $X_i = y_{i-1}$. In Abb. 9 sieht man die positive Korrelation des Fehlers mit der Responsevariable.

3.3. Standardannahmen im klass. lin. Regr.modell

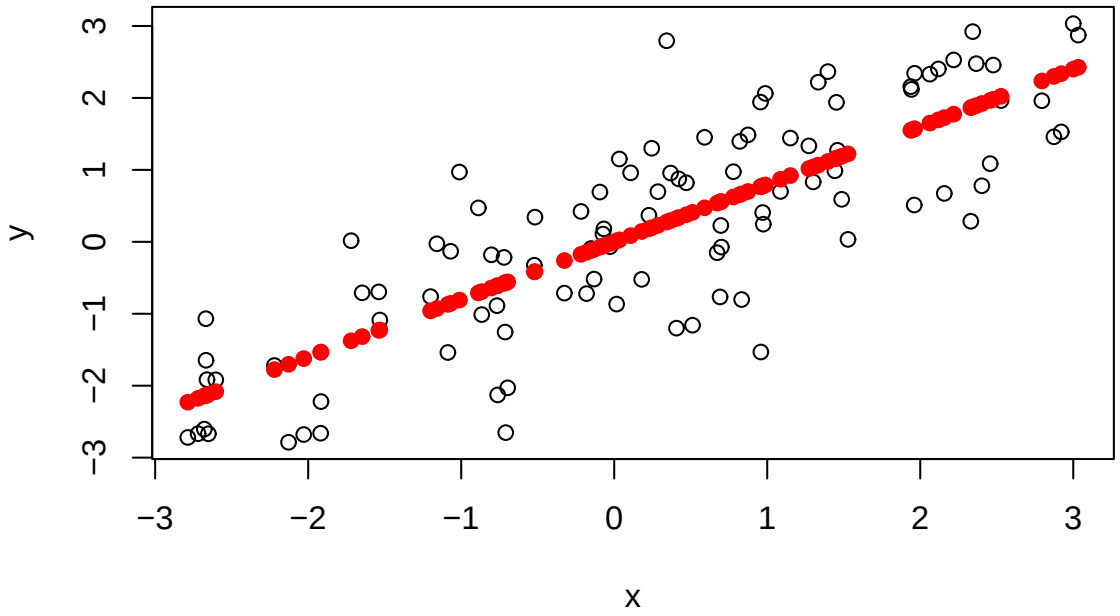


Abbildung 8: Gegenbeispiel zu A4.

3.3. Standardannahmen im klass. lin. Regr.modell

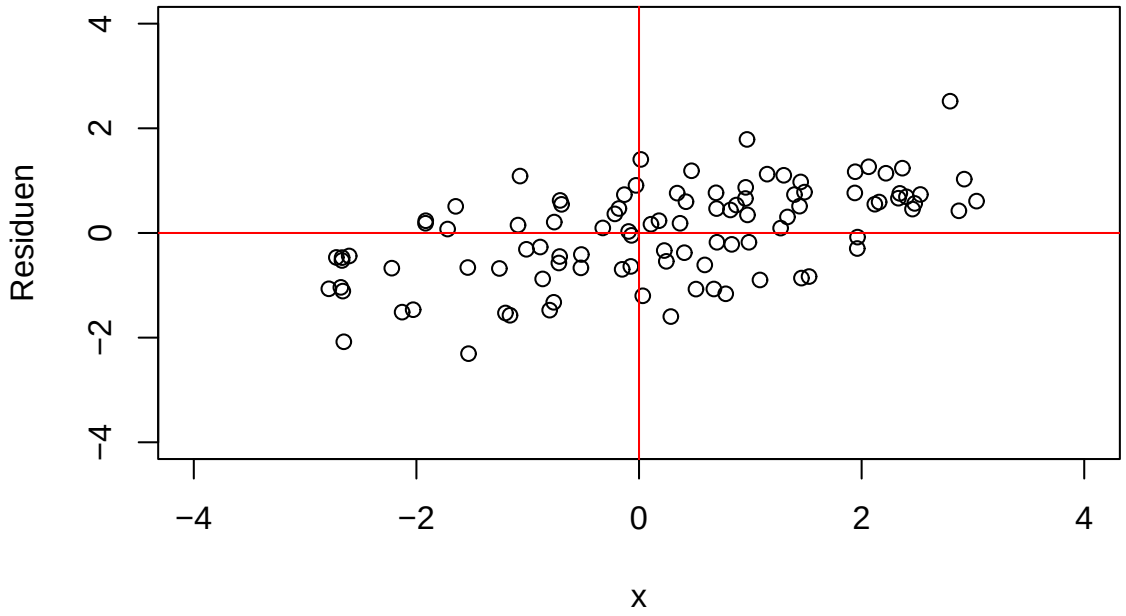


Abbildung 9: Gegenbeispiel zu A4.

3.3. Standardannahmen im klass. lin. Regr.modell

Die Standardannahme A5.

Es besteht keine lineare Abhängigkeit zwischen den Prädiktorvariablen.

$$\text{Aus} \quad \lambda_2 x_{i2} + \lambda_3 x_{i3} + \ldots + \lambda_k x_{ik} = 0$$

$$\text{folgt, dass} \quad \lambda_2 = \lambda_3 = \ldots = \lambda_k = 0 \quad \text{ist.}$$

Im Fall einer linearen Abhängigkeit der Prädiktorvariablen spricht man von **Multikollinearität**. Das Modell ist dann nicht identifizierbar.

Eine nichtlineare Abhängigkeit verletzt A5 nicht, z.B. $X_{i3} = X_{i2}^2$.

Gegenbeispiel:
Gegeben sei das Regressionsmodell:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

Zusätzlich sei $x_{i3} = 0.5 \cdot x_{i2}$. Es gibt daher Koeffizienten $\lambda_2 = -0.5$ und $\lambda_3 = 1$, sodass gilt: $\lambda_2 x_{i2} + \lambda_3 x_{i3} = 0$.

Man kann die Regressionskoeffizienten β_2 und β_3 nicht getrennt voneinander schätzen:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 0.5 x_{i2} + \varepsilon_i = \beta_1 + (\beta_2 + 0.5\beta_3) x_{i2} + \varepsilon_i$$

3.4 Kleinst-Quadrate Schätzung

Wir haben Daten mit T simultanen Beobachtungen für die Responsevariable $y_1, \ldots, y_T$ und für die Prädiktoren $x_{12}, ..., x_{1k}, \cdots, x_{T2}, ..., x_{Tk}$. Die T Gleichungen des multiplen Regressionsmodells lauten:

$$
\begin{array}{rcccccccc}
y_1 & = & 1 \cdot \beta_1 & + & x_{12} \cdot \beta_2 & \ldots & + & x_{1k} \cdot \beta_k & + & \varepsilon_1 \\
y_2 & = & 1 \cdot \beta_1 & + & x_{22} \cdot \beta_2 & \ldots & + & x_{2k} \cdot \beta_k & + & \varepsilon_2 \\
 & \vdots & & & & & & & & \\
y_T & = & 1 \cdot \beta_1 & + & x_{T2} \cdot \beta_2 & \ldots & + & x_{Tk} \cdot \beta_k & + & \varepsilon_T
\end{array}
$$

3.4 Kleinst-Quadrate Schätzung

Wir können diese T Gleichungen in Matrizenschreibweise (s. Mathematik 1) schreiben:

$$y = X \cdot \beta + \varepsilon$$

mit

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{12} & \dots & x_{1k} \\ 1 & x_{22} & \dots & x_{2k} \\ & & & \\ \vdots & & & \\ 1 & x_{T2} & \dots & x_{Tk} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{pmatrix}.$$

3.4 Kleinst-Quadrate Schätzung

Normalgleichungen:

Die Fehlerquadratsumme $\sum_{i=1}^{T} \varepsilon_i^2$ wird genau für jene β minimal, die die Normalgleichungen

$$(X^\top X)\beta = X^\top y$$

erfüllen. Da die Annahme A5 sicher stellt, dass die Matrix $(X^\top X)$ eine Inverse hat, kann man einen eindeutigen Schätzer für β bekommen als:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

(s. Mathematik 1)

Übung 2

Wir haben für das einfache lineare Regressionsmodell schon die R-Funktion `lm` kennengelernt. Sie kann auch für das multiple Regressionsmodell verwendet werden. Lösen Sie daher die folgenden Beispiele mit dieser Funktion. Laden Sie die Daten `statlab` von der LV-Seite.

1. Erklären Sie mit Hilfe eines linearen Regressionsmodells die Größe des Kindes (`CTHGHT`) durch die Größe von Mutter (`MTHGHT`) und Vater (`FTHGHT`). Wie lautet die Regressionsgleichung? Zeichnen Sie die Residuen gegen die geschätzten Werte der Größe des Kindes in ein Streudiagramm.

Übung 2

Welche Körpergröße kann für ein Kind erwartet werden, wenn die Mutter 55 und der Vater 70 inch groß sind?

2. Wählen Sie die Variable `CTWGT` als abhängige Variable und die Prädiktoren `MBAG` und `FBAG`. Wie lauten die Regressionsparameter? Zeichnen Sie die Residuen gegen die Schätzungen für `CTWGT` in ein Streudiagramm.

4. Statistik im multiplen Regressionsmodell

In diesem Kapitel wird im Abschnitt 4.1 zusätzlich zu den schon bekannten Standardannahmen noch die Annahme von normalverteilten Residuen hinzugefügt. Auf Basis dieser Annahme können wir dann im Abschnitt 4.2 und 4.3 die statistischen Eigenschaften des KQ-Schätzers und der Regressionsparameter untersuchen.

4.1. Verteilungsannahmen des Fehlers

Die Standardannahme A6

Der Modellfehler sei normalverteilt.

$$\varepsilon_i \sim N(0, \sigma^2)$$

für alle $i = 1, \ldots, T$.

Für feste Prädiktorvariablen und feste Parameter $\beta_1, \ldots, \beta_k$ schwankt y_i um etwa $\pm 2 \cdot \sigma$ um den Strukturanteil $\hat{y}_i$.[1]

Extreme Abweichungen im Vergleich zum Großteil der Daten sind unwahrscheinlich.

[1] Signifikanzniveau = 95%

4.1. Verteilungsannahmen des Fehlers

Gegenbeispiel:

Gegeben sei:

$$y_i = 0.2 + 0.8x_i + \varepsilon_i^*$$

Wir vergleichen zwei Modelle mit unterschiedlicher Fehlervarianz:

Modell 1: $\varepsilon_i^* \sim N(0, 4)$

Modell 2: $\varepsilon_i^* \sim N(0, 4)$ für 90% der Daten.

$\varepsilon_i^* \sim N(0, 400)$ für 10% der Daten.

4.1. Verteilungsannahmen des Fehlers

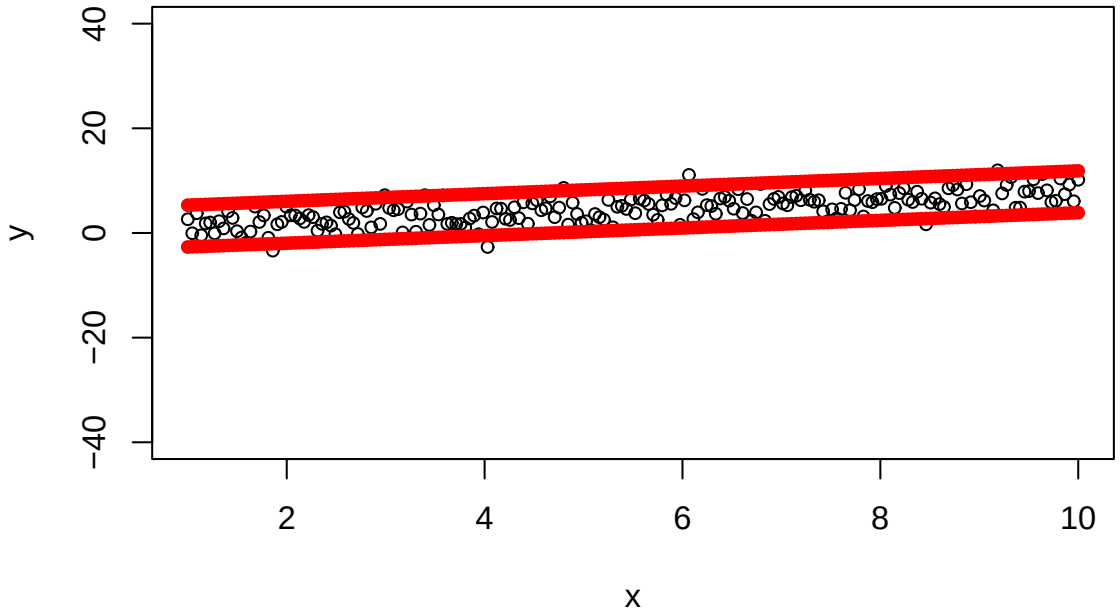


Abbildung 10: Gegenbeispiel zu A6: Modell 1.

4.1. Verteilungsannahmen des Fehlers

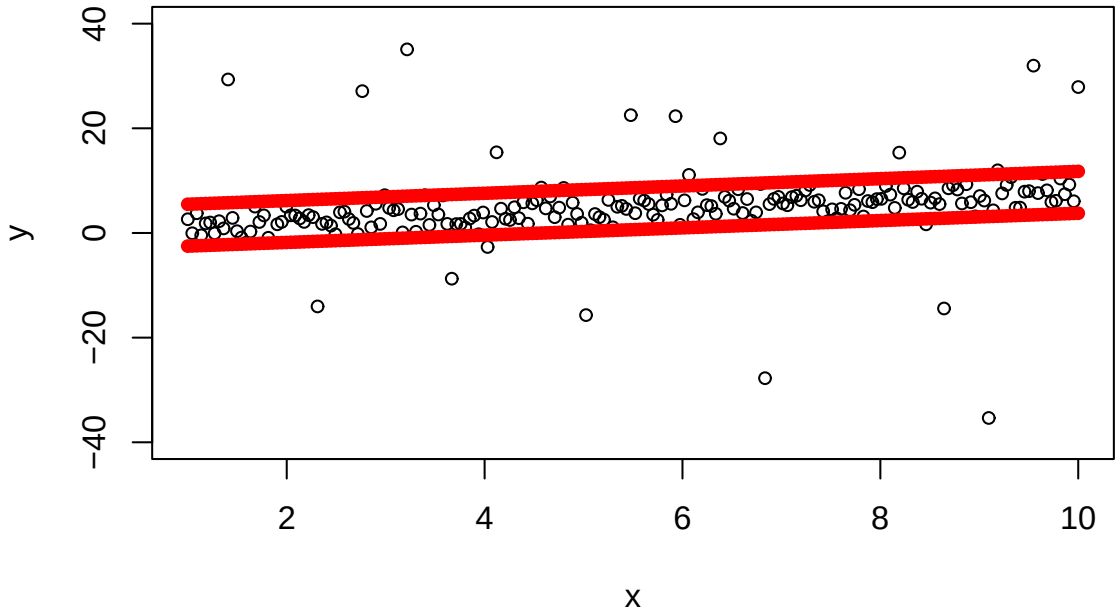


Abbildung 11: Gegenbeispiel zu A6: Modell 2.

4.2 Eigenschaften des KQ-Schätzers

Mit dem folgenden Theorem wird ausgesagt, dass die Schätzmethode mit Hilfe der kleinsten Quadrate zu einem Schätzer der Regressionsparameter führt, der wichtige statistische und für die praktische Anwendung notwendige Eigenschaften hat:

Gauss-Markov Theorem:

Unter den Annahmen A1-A5 ist der KQ-Schätzer (Abschnitt 3.4) der **BLUE** (best linear unbiased estimator). Er besitzt die folgenden Eigenschaften:

- $\hat{\beta}$ ist **linear**, d.h. eine Linearkombination von $y_1, \ldots, y_T$.

4.2 Eigenschaften des KQ-Schätzers

- $\hat{\beta}$ ist **erwartungstreu (unverzerrt, unbiased)**, d.h. im Mittel wird der wahre Parameter geschätzt.

- $\hat{\beta}$ ist **effizient** unter den linearen Schätzern, d.h. jeder andere lineare Schätzer hat eine größere Schwankungsbreite.

Wenn zusätzlich auch die Annahme A6 gilt, dann ist $\hat{\beta}$ sogar effizient unter **allen** Schätzern.

4.3. Statistik zu den Regressionsparametern

Abweichung des Schätzers $\hat{\beta}_j$ von β_j, ($j = 1, \ldots, k$):

Der KQ-Schätzer $\hat{\beta}_j$ ist erwartungstreu (Gauss-Markov Theorem). Daher ist die Abweichungen vom wahren β_j im Mittel 0:

$$E(\hat{\beta}_j) = \beta_j.$$

Unter A6 ist der KQ-Schätzer normalverteilt:

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 (X^\top X)^{-1}_{jj}).$$

($(X^\top X)^{-1}_{jj}$ ist das j-te Diagonalelement der Matrix $(X^\top X)^{-1}$.)

4.3. Statistik zu den Regressionsparametern

Aus dieser Normalverteilung kann man die statistische Schwankungsbreite angeben (s. Statistik 1):

$$|\hat{\beta}_j - \beta_j| \leq c_\alpha SD_j.$$

c_α ist der kritische Wert aus der Standardnormalverteilung zum α-Signifikanzniveau. ($c_{0.95} = 1.96 \approx 2$)

Die Standardabweichung für den j-ten Regressionsparameter ist [2]:

$$SD_j = \sqrt{\sigma^2 (X^\top X)^{-1}_{jj}}$$

[2] wird von Standard-Software automatisch berechnet.

4.3. Statistik zu den Regressionsparametern

Im Allgemeinen wird auch die Varianz von ε unbekannt sein und daher geschätzt werden (und dann bei der Berechung von SD_j eingesetzt werden):

$$\hat{\sigma}^2 = \frac{1}{T-k} \sum_{i=1}^{T} (y_i - \hat{y}_i)^2$$

Wenn σ^2 geschätzt wird, müssen die kritischen Werte c_α aus der sogenannten t-Verteilung gewählt werden. [3]

[3] mit $T - k$ Freiheitsgraden; bei großem T kann als Näherung die Standardnormalverteilung verwendet werden.

4.3. Statistik zu den Regressionsparametern

Das **Konfidenzintervall** kann nun für die einzelnen Regressionsparameter angegeben werden:

$$\hat{\beta}_j - c_\alpha \cdot SD_j \leq \beta_j \leq \hat{\beta}_j + c_\alpha \cdot SD_j$$

für $j = 1, \ldots, k$.

5. Modellwahl im einf. lin. Regressionsmodell

Zunächst wird die Modellwahl für das einfache lineare Regressionsmodell behandelt (s. Statistik 1). Es geht um die Frage, ob ein Prädiktor X überhaupt einen Beitrag zur Erklärung der Responsevariablen Y liefert und wenn ja, wie groß ein solcher Beitrag ist. Die Frage lautet also: Ist eine Regressionsgerade überhaupt sinnvoll zur Modellierung der Daten?

5.1. F-Statistik

Es werden einander 2 Modelle gegenübergestellt:
Im Modell M_1 wird die Prädiktorvariable X nicht berücksichtigt:

$$M_1 : y_i = \beta_1 + \varepsilon_i.$$

Die Schätzung für β_1 ist daher der Datenmittelwert: $\hat{\beta}_1 = \bar{y}$ und die Residuen ε_i gleichen den zufälligen Schwankungen der einzelnen Daten y_i um ihren gemeinsamen Mittelwert.

Das Modell M_2 ist das einfache lineare Regressionsmodell mit Prädiktor X:

$$M_2 : y_i = \beta_1 + \beta_2 x_i + \varepsilon_i.$$

5.1. F-Statistik

Das Ausmaß des Erklärungsbeitrages der beiden Modelle wird verglichen indem man die Fehlerquadratsummen gegenüberstellt.

Die Fehlerquadratsumme in M_1 ist: $SSE_{M_1} = \sum_{i=1}^{T}(y_i - \bar{y})^2$,

und in M_2: $SSE_{M_2} = \sum_{i=1}^{T}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{T}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2$.

Es gilt jedenfalls, dass das einfachere Modell M_1 weniger Erklärungsbeitrag (also eine größere SSE) hat als jenes Modell M_2, das eine zusätzliche erklärende Variable einbezieht. Es stellt sich aber die Frage, ob der zusätzliche Erklärungsbeitrag signifikant ist, d.h. ob M_2 die Daten im statistischem Sinn besser erklärt. Dazu wird ein statistischer Test, die ANOVA, durchgeführt.

5.1. F-Statistik

Mit Hilfe der Varianzanalyse (ANOVA) kann man testen, ob der Erklärungsbeitrag von X signifikant ist. Der statistische Test nimmt als Nullhypothese an, dass das Modell ohne Prädiktor X, also M_1 gewählt wird:

Nullhypothese H_0: $\quad M_1 : y_i = \beta_1 + \varepsilon_i$

Alternativhypothese H_1: $\quad M_2 : y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$

Falls der Test eine signifikante Testgröße liefert, kann M_1 verworfen werden und M_2 angenommen werden.[4]

[4] Vgl. die Interpretation von Signifikanz, Fehler 1. Art, Fehler 2. Art,... aus Statistik 1

5.1. F-Statistik

In der Varianzanalyse wird die Fehlerquadratsumme des Modells M_1 in 2 Teile zerlegt:

$$SSE_{M_1} = (SSE_{M_1} - SSE_{M_2}) + SSE_{M_2}.$$

SSE_{M_1} = die duch M_1 nicht erklärbare Streuung von Y, sie ist gleich der Abweichungsquadratsumme der Daten y_i von ihrem Mittelwert $\bar{y}$
und setzt sich zusammen aus:
$(SSE_{M_1} - SSE_{M_2})$ = einem Teil der durch Einbeziehen der Prädiktorvariable X zusätzlich erklärt werden kann, und
SSE_{M_2} = einem Teil der auch durch M_2 nicht erklärt werden kann.

5.1. F-Statistik

In der Notation aus Statistik 1 schreiben wir:

$$\underbrace{SSE_{M_1}}_{SS_T} = \underbrace{(SSE_{M_1} - SSE_{M_2})}_{SS^*} + \underbrace{SSE_{M_2}}_{SS_R}$$

$SS_T =$ die Gesamtstreuung
$SS^* =$ erklärbare Streuung
$SS_R =$ Reststreuung

Und leiten daraus die ANOVA-Tabelle und die Testgrösse ab.

5.1. F-Statistik

ANOVA-Tabelle:

	df	SS	MSS	F-Statistik	p-Wert
$*$	1	SS^*	$MSS^* = \frac{SS^*}{1}$		
R	$T-2$	SS_R	$MSS_R = \frac{SS_R}{T-2}$		
	$T-1$	SS_T			

F-Statistik: $F = \frac{MSS^*}{MSS_R} = (T-2)\frac{r^2_{y\hat{y}}}{1-r^2_{y\hat{y}}}$

mit $r_{y\hat{y}}$ dem Korrelationskoeffizienten zwischen den wahren y_i und den Vorhersagewerten aus M_2: $\hat{y}_i$.[5]

[5] Im einf. lin. Regress.m. gilt $r_{y\hat{y}} = r_{yx}$.

5.1. F-Statistik

In Statistik 1 wurde zur Beurteilung der F-Statistik als Faustregel verwendet: Wenn F den Wert 4 überschreitet, kann die Nullhypothese verworfen werden.

Von statistischer Software werden zu Tests p-**Werte** ausgegeben, damit beurteilt werden kann, ob eine signifikates Ergebnis vorliegt:

Kleine p-Werte sprechen gegen die Nullhypothese (z.B. beim 95%-Signifikanzniveau führt ein p-Wert, der kleiner als 0.05 ist zum Verwerfen der Nullhypothese.)

Große p-Werte sprechen für Beibehaltung der Nullhypothese.

5.2. t-Statistik

Alternativ zum Test mit Hilfe der F-Statistik aus dem vorigen Abschnitt kann man das Testproblem für das einfache lineare Regressionsmodell $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ auch so formulieren:

Nullhypothese H_0: $\beta_2 = 0$

Alternativhypothese H_1: $\beta_2 \neq 0$

Es wird also getestet, ob der Regressionskoeffizient β_2 von 0 verschieden ist.

5.2. t-Statistik

Eine Prüfgröße kann auf Basis der Überlegungen zur Schwankungsbreite von β_2 gebildet werden. Wir passen das einfache lineare Regressionsmodell an die Daten an. Der Koeffizient β_2 wird sich in der Regel von 0 unterscheiden. Es soll nun beurteilt werden, ob diese Abweichung von 0 im Rahmen der zufälligen Schwankung liegt oder ob eine signifikante Abweichung von 0 vorliegt, und daher die Regressionsgerade verwendet werden soll.

5.2. t-Statistik

Wenn der Schätzwert $\hat{\beta}_2$ um den wahren Wert 0 schwankt, ergibt sich die t-Statistik:

$$t \quad = \quad \frac{\hat{\beta}_2}{SD_2}.$$

Wenn $-c_\alpha \leq t \leq c_\alpha$ wird die Abweichung als zufällig beurteilt und die Nullhypothese beibehalten.[6]

Wenn t ausserhalb des Intervalls liegt, wird die Nullhypothese verworfen und das Regressionsmodell liefert eine signifikante Verbesserung bei der Erklärung von Y.

Es werden von statistischer Software wieder p-Werte zur Beurteilung ausgegeben.

[6] Für c_α gelten die Überlegungen vom Abschnitt über die Schwankungsbreiten.

5.3. Bestimmtheitsmaß

In den vorigen beiden Abschnitten wurde getestet, ob die Prädiktorvariable einen **signifikanten** Beitrag zur Erklärung der Responsevariable liefert. Jetzt geht es um die Frage, ob so ein signifikanter Beitrag inhaltlich **relevant** ist. Als geeignetes Maß kann das Bestimmtheitsmaß aus der ANOVA-Tabelle berechnet werden:

$$\frac{SS^*}{SS_T} \quad = \quad r^2_{y\hat{y}}$$

Das Bestimmtheitsmaß gibt also jenen Anteil an der totalen Abweichung SS_T an, der durch das einfache lineare Regressionsmodell erklärt werden kann.

5.3. Bestimmtheitsmaß

Die inhaltliche Interpretation des Bestimmtheitsmaßes muss dann je nach Anwendung erfolgen. Denn: welches Ausmaß an Erklärungsbeitrag als inhaltlich relevant anzusehen ist, kann je nach Anwendung verschieden sein.

6. Modellwahl im einf. lin. Regr.m. in R

In '`ModellWahl1inR.pdf`' von der LV-Seite werden die R-Befehle zur Modellwahl im einfachen linearen Regressionsmodell erklärt.

Übung 3

1. Führen Sie das Beispiel aus '`ModellWahl1inR.pdf`' durch.

2. Wählen Sie aus dem `statlab`-Datensatz die Variable `CTWGT` als Responsevariable, die im einfachen linearen Regressionsmodell von der Prädiktorvariable `FBAG` abhängt. Prüfen Sie die Hypothese, dass das einfache lineare Regressionsmodell keinen sigifikanten Beitrag zur Erklärung von `CTWGT` liefert, mit der F-Statistik.

3. Führen Sie den Test aus dem vorigen Beispiel mit der t-Statistik durch.

Übung 3

4. (a) Geben Sie die 95%-Konfidenzintervalle für die Regressionsparameter der Regressionsgerade aus Beispiel 2 an.
   (b) Geben Sie den Anteil an der Gesamtstreuung der Daten an, der durch die Regressionsgerade erklärt werden kann.

7. Modellwahl im multiplen Regressionsmodell

Jetzt sollen die Konzepte zur Variablenwahl vom einfachen linearen Regressionsmodell auf das multiple Regressionsmodell verallgemeinert werden. Die Variablenwahl muss nun aus den Prädiktorvariablen $X_2, \ldots, X_k$ getroffen werden. Je größer k ist, desto komplexer wird das Problem.

Wir werden damit beginnen das Bestimmtheitsmaß für multiple Regressionsmodelle zu definieren. Danach werden wir die F-Statistik für komplexere Testprobleme betrachten und die Vorgangsweise bei der Variablenwahl am einfachsten Beispiel von 2 möglichen Prädiktorvariablen zeigen. Diese kann dann auf noch mehr Prädiktoren verallgemeinert werden.

7. Modellwahl im multiplen Regressionsmodell

Zum Schluss werden wir die t-Statistiken für die Parameter im multiplen Regressionsmodell anschauen und feststellen, dass anders als im einfachen linearen Regressionsmodell, die t-Statistiken der Regressionskoeffizienten des allgemeinsten Modells nicht mehr direkt zur Vereinfachung des Modells herangezogen werden können.

7.1. Das Bestimmtheitsmaß

Der **multiple Korrelationskoeffizient** $R_{y.M}$ des Modells M mit Prädiktorvariablen $X_2, \ldots, X_k$ ist die Verallgemeinerung des Korrelationskoeffizienten $r_{y\hat{y}}$ aus dem einfachen linearen Regressionsmodell.

$$R_{y.M} \quad = \quad COR(y_i, \hat{y}_i)$$

$\hat{y}_i$ ist der Schätzwert für y_i aus dem Modell M.

$R_{y.M}$ ist wegen der positiven Koppelung der Daten y_i mit ihrer Vorhersage $\hat{y}_i$ immer positiv.

7.1. Das Bestimmtheitsmaß

Für ein Modell M mit Prädiktorvariablen $X_2, \ldots, X_k$ kann der Erklärungswert wieder durch das **Bestimmtheitsmaß** angegeben werden:

$$\frac{(SSE_{M_1} - SSE_M)}{SSE_{M_1}} \quad = \quad R^2_{y.M}$$

Das ist der Anteil an der totalen Abweichung, den das Modell M zusätzlich im Vergleich zum konstanten Modell M_1 erklärt.

Natürlich wird der Erklärungswert automatisch größer, wenn man in ein Modell zusätzliche Prädiktoren hineinnimmt. Es stellt sich aber die Frage, ob dieser zusätzliche Beitrag signifikant ist.

7.1. Das Bestimmtheitsmaß

Dazu wird auch im multiplen Regressionsmodell eine F-Statistik zum Testen verwendet.

7.2 ANOVA

Gegeben sei ein multiples Regressionsmodell M mit Prädiktorvariablen $X_2, \ldots, X_k$:

$$M : y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

Es soll die Hypothese getestet werden, ob ein Teil der Prädiktoren gleich 0 ist. (Für eine einfachere Notation nehmen wir an, dass dies die letzten m Prädiktoren sind.) D.h. als Nullhypothese wird das Modell M_0 angenommen:

$$M_0 : y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_{k-m} x_{ik-m} + \varepsilon_i$$

7.2 ANOVA

Anders ausgedrückt, lautet der Test für das Modell M:

Nullhypothese H_0: $\beta_{k-m+1} = \ldots = \beta_k = 0$

Alternativhypothese H_1: mind. ein $\beta_{k-m+1}, \ldots, \beta_k$ ist ungleich 0

Also: Es wird getestet, ob alle obigen Parameter **gemeinsam** 0 sind.
Wenn nur einer ungleich 0 ist, wird H_0 schon verworfen.

7.2 ANOVA

Als Testgröße dient die schon bekannte F-Statistik:

$$F = \frac{(SSE_{M_0} - SSE_M)/m}{SSE_M/(T-k)}$$

Auch hier werden p-Werte von statistischer Software ausgegeben.[7]

[7] Die korrekte Prüfverteilung ist die sogenannte F-Verteilung mit $(m, T-k)$ Freiheitsgraden.

7.2 ANOVA

Wir betrachten nun den einfachsten Fall eines multiplen Regressionsmodells mit nur 2 möglichen Prädiktoren. Es muss eine Auswahl aus den folgenden Modellen getroffen werden:

$$
\begin{aligned}
M_1: \quad y_i &= \beta_1 + \varepsilon_i \\
M_2: \quad y_i &= \beta_1 + \beta_2 x_{i2} + \varepsilon_i \\
M_3: \quad y_i &= \beta_1 + \beta_3 x_{i3} + \varepsilon_i \\
M_{23}: \quad y_i &= \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i
\end{aligned}
$$

7.2 ANOVA

Die Vorgangsweise ist die folgende:

1. Wir starten mit M_{23} und testen es als H_1 mit $H_0 : M_1$. Das entspricht also einem Testen der beiden Parameter β_2 und β_3 gemeinsam. Wenn die F-Statistik

   (a) nicht signifikant ist, wird M_1 angenommen und wir sind fertig.

   (b) signifikant ist, muss zumindestens eines der β_2, β_3 ungleich 0 sein $\longrightarrow$ 2.

2. Wir testen M_{23} als H_1 und M_2 als H_0. Das entspricht also einem Testen des Parameters β_3. Wenn die F-Statistik

   (a) nicht signifikant ist, wird β_3 gleich 0 gesetzt.

(b) signifikant ist, wird β_3 ungleich 0 gesetzt $\rightarrow$ 3.

3. Wir testen M_{23} als H_1 mit $H_0 : M_3$. Das entspricht also einem Testen von β_2. Wenn die F-Statistik

   (a) nicht signifikant ist, wird β_2 gleich 0 gesetzt.
   (b) signifikant ist, wird β_2 ungleich 0 gesetzt und wir wählen M_{23}.

Diese Vorgangsweise kann auf noch mehr Variablen verallgemeinert werden.

7.3. Modellwahl im multipl. Regressionsmodell in R

In R kann mit Hilfe der Funktion `anova` der Vorgang der Variablenwahl durchgeführt werden und der auf ersten Blick kompliziert anmutende Vorgang aus dem vorigen Abschnitt in eine übersichtliche Form gebracht werden.

Dies wird in '`ModellWahl2inR.pdf`' an einem Beispiel gezeigt.

Übung 4

Mit den Statlab-Daten:

1. Wählen Sie `FIT` als Responsevariable, die in einem multiplen Regressionsmodell durch `FBAG` und `MBAG` erklärt werden soll (M_{23}). Testen Sie das Modell M_{23} gegen das konstante Modell M_1. Welches dieser beiden Modelle würden Sie wählen? Welche Regressionsparameter sind signifikant von 0 verschieden?

   Es sollen nun die Details zu 1. angeschaut werden:

2. Wie schaut die ANOVA aus, wenn wir das konstante Modell M_1, das einfache lineare Regressionsmodell M_2 mit Prädiktor `FBAG`

Übung 4

und das Modell M_{23} betrachten? Welches dieser Modelle würden Sie wählen?

3. Analog zu 2. Jetzt soll aber statt M_2 das einfache lineare Regressionsmodell M_3 mit Prädiktor `MBAG` verwendet werden. Welches dieser Modelle würden Sie wählen?

4. Wenn sie jetzt die Analysen aus dem 1. bis 3. Beispiel betrachten: Welches Modell würden Sie wählen? Könnten wir auch eines der einfachen linearen Regressionsmodelle anstatt des konstanten Modells wählen?

7.4 . t-Statistik

Im einfachen linearen Regressionsmodell wurde zwischen dem konstanten Modell und einem Modell mit Regressionsgerade gewählt. Dazu standen die F-Statistik und die t-Statistik zur Verfügung.

Im multiplen Regressionsmodell wird nun aber mit Hilfe der F-Statistik die Nullhypothese getestet, ob mehrere Regressionsparameter gemeinsam 0 sind. Es handelt sich dabei im allgemeinen um mehr als einen Parameter β_j, der getestet wird. Dazu haben wir die F-Statistik verwendet.

Es ist ein häufiger Fehler auch bei einem solchen Test einfach die t-Statistiken von mehreren Parametern anzusehen, und dann all jene Parameter im multiplen Regressionsmodell gleich 0 zu setzten,

7.4 . t-Statistik

die eine nicht signifikante t-Statistik haben. Aber: eine einzelne t-Statistik

$$t_j = \frac{\hat{\beta}_j}{SD_j}, \quad j \geq 2$$

gibt Auskunft über einen einzelnen Parameter, unter der Voraussetzung, dass alle anderen im Modell bleiben!

Übung 5

Laden Sie die Daten '`tbeispiel.rda`' von der LV-Seite und rechnen Sie ein Regressionsmodell mit der Responsevariablen Y und den erklärenden Variablen X_2 und X_3.

1. Betrachten Sie die `summary` des allgemeinsten Modells, das beide Prädiktoren enthält. Interpretieren Sie die t-Statistiken der Prädiktoren.

2. Führen Sie die Modellwahl durch. Welches Modell würden Sie wählen?

8. Statistik zur Prognose

Gegeben sei das multiple Regressionsmodell unter den Standardannahmen A1 - A6:

$$y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i. \tag{1}$$

Dieses Modell wurde aus den Daten $i = 1, \ldots, T$ mit den schon bekannten Methoden geschätzt.

Wir haben daher die Regressionsparameter geschätzt: $\hat{\beta}_1, \ldots, \hat{\beta}_k$, und die Varianz von ε_i: $\hat{\sigma}^2$. Diese Schätzer werden im folgenden verwendet, um eine Prognose und deren Schwankungsbreite für neue Daten anzugeben.

8. Statistik zur Prognose

Es soll nun für ein neues Szenario, das durch neue Daten $x_{T+1,2}, \ldots, x_{T+1,k}$ bestimmt ist, die nicht für die Modellschätzung verwendet wurden, die unbekannte Responsevariable y_{T+1} vorhergesagt werden.

Die Vorhersage aus dem multiplen Regressionsmodell bezeichnen wir mit $\hat{y}_{T+1}$.

8. Statistik zur Prognose

Es sollen in diesem Abschnitt die folgenden Fragen behandelt werden:

Punktprognose: Prognosewert $\hat{y}_{T+1}$?

Schwankungsbreite: statistische Schwankungsbreite für den wahren erwarteten Wert $E(y_{T+1}|X_2 = x_{T+1,2}, \ldots, X_k = x_{T+1,k})$? v.a. bei Querschnittsdaten wichtig.

Intervallprognose: in welchem Bereich wird die zukünftige Beobachtung y_{T+1} fallen? v.a. bei Zeitreihendaten.

8. Statistik zur Prognose

Punktprognose:

Die Punktprognose ist der aus dem Modell (1) erwartete Schätzwert unter dem neuen Szenario.

In (1) werden jene Prädiktoren $x_{T+1,2}, \ldots, x_{T+1,k}$ eingesetzt, für deren Szenario eine Prognose erstellt werden soll:
Wahres Modell: $y_{T+1} = \beta_1 + \beta_2 x_{T+1,2} + \ldots + \beta_k x_{T+1,k} + \varepsilon_{T+1}$.

Die unbekannten Parameter werden durch die Schätzwerte $\hat{\beta}_1, \ldots, \hat{\beta}_k$ aus dem ursprünglichen Modell ersetzt.

Der Fehlerterm ε_{T+1} wird 0 gesetzt, da er ja Erwartungswert 0 hat.

$$\longrightarrow \quad \hat{y}_{T+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{T+1,2} + \ldots + \hat{\beta}_k x_{T+1,k}$$

8. Statistik zur Prognose

Für diese Punktprognose gilt, dass

- kein systematischer Fehler gemacht wird, unter der Voraussetzung, dass das Modell stimmt.

- der unter dem neuen Szenario zu erwartende Prognosefehler minimal wird:

$$E((y_{T+1} - \hat{y}_{T+1})^2 | X_2 = x_{T+1,2}, \ldots, X_k = x_{T+1,k}) \rightarrow \min$$

8. Statistik zur Prognose

Statistische Schwankungsbreite für die Punktprognose:

Bei der Schätzung der statistischen Schwankungsbreite der Punktprognose geht es darum, jene Unsicherheit bei der Prognose abzuschätzen, die durch die Unsicherheit der Schätzungen der Regressionskoeffizienten entsteht. Es wird ein Konfidenzintervall für die unbekannte **erwartete wahre** Prognose angegeben. Denn: Der wahre, für das neue Szenario erwartete Wert ist ja nicht bekannt. Man hat nur die Schätzungen aus (1) zur Verfügung. Daraus bekommt man die Punktprognose (s. vorige Folie) und die Schwankungsbreite dieser Punktprognose um den wahren erwarteten Wert (s. nächste Folien), aus denen die statistischen Schwankungsbreiten angegeben werden können.

8. Statistik zur Prognose

- Der wahre Erwartungswert für y_{T+1} ist für ein gewisses Szenario $x_{T+1,2}, \ldots, x_{T+1,k}$ durch das wahre Modell mit den wahren Parametern $\beta_1, \ldots, \beta_k$ gegeben[8]:

$$E(Y|X_2 = x_{T+1,2}, \ldots, X_k = x_{T+1,k}) =$$
$$= \beta_1 + \beta_2 x_{T+1,2} + \ldots + \beta_k x_{T+1,k}$$

[8] Der Erwartungswert des Fehlers ist ja 0, daher bleibt nur der Strukturanteil über.

8. Statistik zur Prognose

- Der prognostizierte Erwartungswert $\hat{y}_{T+1}$ ist durch das Modell mit den geschätzten Parametern $\hat{\beta}_1, \ldots, \hat{\beta}_k$ gegeben:

$$\hat{y}_{T+1} = \hat{\beta}_1 + \hat{\beta}_2 x_{T+1,2} + \ldots + \hat{\beta}_k x_{T+1,k}$$

Daraus entsteht die statistische Schwankung des Prognosewertes um den wahren Wert:

$$\begin{aligned} &\hat{y}_{T+1} - E(Y|X_2 = x_{T+1,2}, \ldots, X_k = x_{T+1,k}) = \\ &= (\hat{\beta}_1 - \beta_1) + (\hat{\beta}_2 - \beta_2) x_{T+1,2} + \ldots + (\hat{\beta}_k - \beta_k) x_{T+1,k} \end{aligned}$$

8. Statistik zur Prognose

Frage: Wie groß ist diese Schwankung?

Die Schwankungsbreite kann wieder durch **kritischer Wert mal Standardabweichung** abgeschätzt werden:

$$\hat{y}_{T+1} - c_\alpha \cdot SD_{\hat{y}} \leq E(Y|X_2 = x_{T+1,2}, \ldots, X_k = x_{T+1,k}) \leq \hat{y}_{T+1} + c_\alpha \cdot SD_{\hat{y}}$$

c_α ist der kritische Wert aus Abschnitt 4.3.

$SD_{\hat{y}}$ ist die Standardabweichung für dieses Prognoseproblem.[9]

[9] Sie kann mit Mitteln der Matrizenrechnung hergeleitet werden und wird von Statistiksoftware automatisch berechnet.

8. Statistik zur Prognose

Die statistische Schwankungsbreite der Punktprognose wird durch die Abweichung der Schätzungen $\hat{\beta}_j$ von ihren wahren Werten β_j bestimmt. Daher gilt, dass sie umso kleiner wird:

- je größer die Stichprobengröße T ist.
- je kleiner die Varianz des Modellfehlers σ^2 ist.
- je näher die Werte des neuen Szenarios $x_{T+1,2}, \ldots, x_{T+1,k}$ am Mittelwert der Beobachtungen, auf denen die Modellschätzung aus (1) basiert, liegen.

8. Statistik zur Prognose

Intervallprognose für zukünftige y_{T+1}:

Jetzt soll abgeschätzt werden in welchem Bereich die zukünftige, unbekannte Beobachtung y_{T+1} liegen wird. Es wird jetzt die **gesamte** Unsicherheit des Modells berücksichtigt. Das Modell besteht ja aus einem Strukturteil (erwarteter Wert für y_{T+1} von vorhin) und einem zufälligen Residuenteil (ε_{T+1}). Die statistische Schwankung für y_{T+1} um die Punktprognose $\hat{y}_{T+1}$ setzt sich daher aus den Schwankungen dieser beiden Teile zusammen:

- Statistische Schwankung der Punktprognose um den wahren Erwartungswert, d.h. $SD_{\hat{y}}$ von vorhin.
- Statistische Schwankung von ε_{T+1} um 0, d.h. $\hat{\sigma}^2$.

8. Statistik zur Prognose

Unter den Standardannahmen gilt, dass die statistische Schwankung von y_{T+1} die folgende Standardabweichung hat:

$$SD_y = \sqrt{(SD_{\hat{y}}^2 + \hat{\sigma}^2)}$$

D.h. die Varianzen der einzelnen Teile können einfach addiert werden.

Daraus ergibt sich die Schwankungsbreite für y_{T+1}[10]:

$$\hat{y}_{T+1} - c_\alpha SD_y \leq y_{T+1} \leq \hat{y}_{T+1} + c_\alpha SD_y$$

[10] Wird von Software automatisch berechnet.

8. Statistik zur Prognose

Für die statistische Schwankungsbreite für die Prognose von y_{T+1} gilt für wachsendes T:

- $SD_{\hat{y}}$ wird immer kleiner.
- **Aber**: $\hat{\sigma}^2$ bleibt erhalten.
- $SD_y \geq \hat{\sigma}^2$, D.h. der Anteil aus den Residuen bleibt erhalten und ist eine untere Schranke für die Unsicherheit der Prognose (auch für ein sehr großes T).

9. Beispiel zur Prognose in R

In 'PrognoseinR.pdf' werden die für die Punktprognose und die Schätzung der statistischen Schwankungsbreiten notwendigen R-Befehle erläutert.

Übung 6

1. In Übung 1, Bsp. 2 wurde mit den statlab-Daten das einfache lineare Regressionsmodell mit Prädiktor `MTHGHT` und Responsevariable `CTHGHT` gerechnet.
   (a) Erstellen Sie die Punktprognose für die Größe eines Kindes von einer 55 bzw. 70 inch großen Mutter mit der R-Funktion `predict`.
   (b) Geben Sie weiters die Vorhersageunsicherheit, die aufgrund der Unsicherheit der Parameterschätzungen entstehen, d.h. das Intervall der Schwankungsbreite für diese Punktprognose an (Sicherheit 95%).

Übung 6

(c) Wie lautet der Bereich für die gesamte Unsicherheit der Vorhersage? (Sicherheit 95%)

2. Analog Beispiel 1, mit `CTWGT ~ MTWGT + FTWGT`. Es sollen die Prognosen für 3 Szenarien erstellt werden:
   `(MTWGT, FTWGT) = (125, 170); (142,185); (155,175).`

10. Methoden des Modellvergleichs

Wir haben bis jetzt schon die F-Statistik zur Auswahl von Prädiktorvariablen kennengelernt. Die F-Statistik dient zum Testen von sogenannten **genesteten** Modellen. Genestete Modelle sind solche, wo das eine Modelle im anderen enthalten ist. D.h. Das eine Modell geht aus dem anderen durch Vereinfachung (Nullsetzen einiger Parameter) hervor.

Jetzt werden Kriterien vorgestellt, die sowohl zum Vergleich von genesteten als auch **nicht genesteten** Modellen geeignet sind (Abschnitt 10.2).

In Abschnitt 10.1 wird das Verhalten der Schätzungen bei Spezifikationsfehlern und bei Einbeziehung irrelevanter Prädiktoren betrachtet.

10.1. Spezifikationsfehler und irrelevante Parameter

Spezifikationsfehler:

Von einem Spezifikationsfehler spricht man, wenn die Annahme A1 nicht erfüllt ist. D.h. der Erwartungswert des Fehlers ist ungleich 0. Die häufigsten Ursachen für einen Spezifikationsfehler sind:

- relevante Prädiktorvariablen fehlen.
- falsche funktionale Form wurde gewählt.

(siehe Gegenbeispiel in Abschnitt 3.3)

Ein Spezifikationsfehler führt in der Regel zu verzerrten Schätzungen und unter Umständen auch zu ökonomisch sinnlosen Parameterschätzungen (z.B. falsches Vorzeichen).

10.1. Spezifikationsfehler und irrelevante Parameter

Einbeziehen von irrelevanten Prädiktoren:

Beispiel:

wahres Modell: $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$

gewähltes Modell: $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i^*$

Der wahre Wert von β_4 ist 0. Daher gilt: $E(\varepsilon_i^*) = E(\varepsilon_i - \beta_4 x_{i4}) = 0$.

Es gilt:

- Einbeziehen von irrelevanten Prädiktoren ist kein Spezifikationsfehler.
- Die KQ-Schätzer sind unverzerrt.
- Die Parameterschätzung der irrelevanten Prädiktoren ergibt im Mittel 0 (je größer T desto sicherer an 0).

10.1. Spezifikationsfehler und irrelevante Parameter

Aber: Die statistischen Schwankungen der KQ-Schätzungen für die relevanten Parameter sind größer als im korrekten Modell. (d.h. der Schätzer ist nicht mehr effizient.) Daher werden Schätzungen auf Basis von Modellen mit zu vielen Parametern unsicherer.

Daher: **Principle of Parsimony:**

Das Modell sollte so wenige Parameter wie möglich enthalten, ohne dabei an Erklärungswert zu verlieren.

Methoden so ein Modell zu wählen werden im folgenden Abschnitt vorgestellt.

10.2. Kriterien zum Modellvergleich

Fehlerquadratsumme:

Die Fehlerquadratsumme SSE_M des Modells M wird automatisch kleiner wenn zusätzliche Parameter ins Modell einbezogen werden (Vgl. Abschnitt 7). Man möchte aber ein möglichst sparsames Modell wählen, das trotzdem einen guten Erklärungswert hat. Daher ist SSE_M als Kriterium nicht geeignet für die Modellwahl.

Die Fehlerquadratsumme wird aber (so wie schon bei der F-Statistik zum Vergleich von genesteten Modellen) auch in Kriterien zum Vergleich nicht genesteter Modelle verwendet. Diese Kriterien berücksichtigen die Anzahl an Parametern, indem sie einen **Strafterm** für diese Anzahl beinhalten.

10.2. Kriterien zum Modellvergleich

AIC - Akaike Information Criterion:

$$AIC = T \cdot \log(SSE_M) + 2 \cdot p$$

SC - Schwarz Criterion oder**BIC - Bayesian Information Criterion:**

$$SC = T \cdot \log(SSE_M) + \log(T) \cdot p$$

p bezeichnet die Anzahl an Parametern des Modells; hier: $p = k + 1$ (k Regressionsparameter und die Modellfehlervarianz)

10.2. Kriterien zum Modellvergleich

Anwendung von AIC und SC:

- Man wählt das Modell mit dem kleinsten Wert des Kriteriums
- Da sich die beiden Kriterien beim 2. Summanden unterscheiden (d.h. bei der Art, wie die Anzahl an Parametern ins Kriterium eingeht), kann die Modellwahl abhängig von der Wahl des Kriteriums zu verschiedene Ergebnisse führen.
- AIC tendiert dazu mehr Parameter ins Modell zu nehmen als das SC.

10.2. Kriterien zum Modellvergleich

- Entscheidend für die Evidenz zugunsten eines Modells ist nicht der **absolute** Wert, sondern die **Differenz** der Kriterien, die für die zu vergleichenden Modellen berechnet wurden.

- Die Kriterien werden oft umgeformt. Die Idee bleibt aber gleich und es wird immer das Modell mit möglichst kleinem Kriterium gewählt.[11]

[11] Stat. Software berechnet AIC und SC automatisch nach einer der möglichen Formeln.

11. AIC und SC in R

In 'ModellWahl3inR.pdf' wird ein Beispiel mit Modellwahl mit Hilfe von AIC und SC in R gegeben.