1. Tutorium am 01.12.08/04.12.08

(Einfache lineare Regression)

Aufgabe (vgl. Skript Statistik I (SS 2004))

Eine Speditionsfirma will anhand von 10 zufällig ausgewählten LKW-Lieferungen untersuchen, ob ein bzw. welcher Zusammenhang zwischen der Länge des Transportweges (in km) und der Lieferzeit (in Tagen) von der Abholbereitstellung bis zum Eintreffen der Lieferung beim Empfänger besteht. Es wurden die folgenden Daten erhoben:

Nummer der Lieferung	1	2	3	4	5	6	7	8	9	10
Weglänge (in km)	825	215	1070	550	480	920	1350	325	670	1215
Lieferzeit (in Tagen)	3.5	1.0	4.0	2.0	1.0	3.0	4.5	1.5	3.0	5.0

(a) Zeichne ein Streuungsdiagramm für die Weglänge x in km (Ausgangsvariable) und die Lieferzeit y in Tagen (Zielvariable).

(b) Berechne für das Modell

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d. } \mathcal{N}(0, \sigma^2)$$

die MKQ-Schätzer $(\hat{\alpha}, \hat{\beta})$.

(c) Zeichne die Ausgleichsgerade ins Diagramm von (a) ein.

(d) Teste ob überhaupt ein signifikanter Zusammenhang zwischen der Länge des Transportweges und der Lieferzeit besteht, d. h. teste die Hypothese

$$H_0 : \ „\beta = 0“$$

zum Niveau $1 - \gamma = 0.05$. *Hinweis*: $t_{8,0.975} = 2.306$.

(e) Bestimme 95% Konfidenzintervalle für α und β.

Einfache lineare Regression

Gegeben zwei Datensätze

$$(x_1, x_2, \ldots, x_n) \text{ Ausgangsvariable und } (y_1, y_2, \ldots, y_n) \text{ Zielvariable.}$$

Vermutung: es besteht ein linearer Zusammenhang zwischen x und y:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d. } \mathcal{N}(0, \sigma^2).$$

Aufgaben zur einfachen linearen Regression:

(i) Zeichne (x_i, y_i) in ein Diagramm ein („Streuungsdiagramm“);

(ii) Berechne die MKQ-Schätzer $\hat{\alpha}$, $\hat{\beta}$ für α und β;

(iii) Zeichne die Gerade („Ausgleichsgerade“)

$$y = \hat{\alpha} + \hat{\beta}x$$

ins Diagramm ein;

(iv) Teste Hypothesen über α bzw. β, z. B.

$$H_0 : \text{ „}\alpha = 0\text{“};$$

(v) Bestimme Konfidenzintervalle für α und β.

Zur Aufgabe:

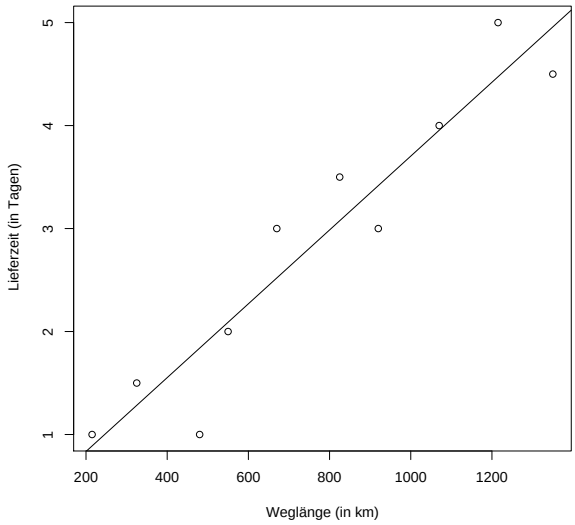


Abbildung 1: Streuungsdiagramm

(a)

(b) *Bekannt ist*: der Vektor $(\hat{\alpha}, \hat{\beta})$, mit

$$\hat{\beta} = \frac{s_{xy}^2}{s_{xx}^2}, \quad \hat{\alpha} = \bar{y}_n - \hat{\beta}\bar{x}_n$$

minimiert den mittleren quadratischen Fehler

$$e(\alpha, \beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2,$$

wobei

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{(Stichprobenmittel)}$$

und

$$\begin{aligned} s_{xx}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \quad \text{(Stichprobenvarianz von } x\text{)} \\ s_{xy}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n) \quad \text{(Stichprobenkovarianz von } (x,y)\text{)} \\ s_{yy}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2 \quad \text{(Stichprobenvarianz von } y\text{).} \end{aligned}$$

Hier:

$$\begin{aligned} \bar{x}_n &= 762 \\ \bar{y}_n &= 2.85 \\ s_{xx}^2 &= 144206.7 \\ s_{xy}^2 &= 517 \\ \hat{\beta} &= 0.003585132 \\ \hat{\alpha} &= 0.1181291 \end{aligned}$$

(c) Die Ausgleichsgerade lautet

$$y = \hat{\alpha} + \hat{\beta}x.$$

(d) *Bekannt ist*:

$$\begin{aligned} \frac{\hat{\alpha} - \alpha}{S\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)/(n(n-1)s_{xx}^2)}} &\sim t_{n-2} \\ \frac{\hat{\beta} - \beta}{S/\sqrt{(n-1)s_{xx}^2}} &\sim t_{n-2}, \end{aligned}$$

wobei

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

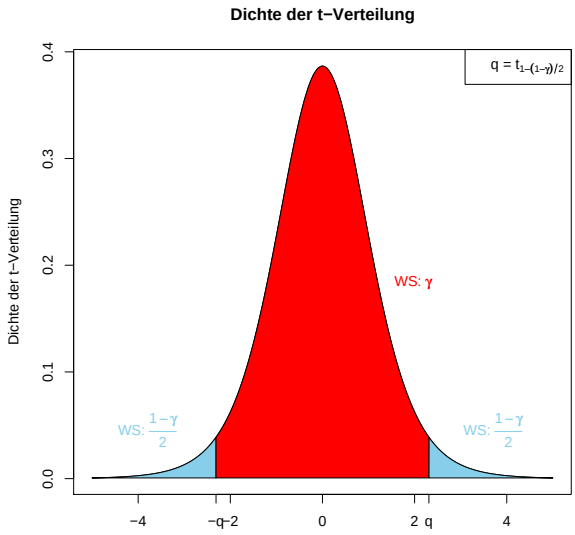
Hierbei sind $\hat{y}_i = \hat{\alpha} + \hat{\beta}x_i$ die gefitteten Werte für y_i.
Aus der Graphik erkennen wir: mit Wahrscheinlichkeit γ ist

$$-t_{n-2,1-\frac{1-\gamma}{2}} \leq \frac{\hat{\alpha} - \alpha}{S\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)/(n(n-1)s_{xx}^2)}} \leq t_{n-2,1-\frac{1-\gamma}{2}}$$

und ebenso

$$-t_{n-2,1-\frac{1-\gamma}{2}} \leq \frac{\hat{\beta} - \beta}{S/\sqrt{(n-1)s_{xx}^2}} \leq t_{n-2,1-\frac{1-\gamma}{2}}$$

Hieraus ergeben sich die t-Tests:



- Hypothese H_0 : „$\alpha = \alpha_0$“ wird zum Niveau $1-\gamma$ abgelehnt, falls

$$\frac{|\hat{\alpha} - \alpha_0|}{S\sqrt{\left(\sum_{i=1}^{n} x_i^2\right)/(n(n-1)s_{xx}^2)}} > t_{n-2,1-\frac{1-\gamma}{2}};$$

- Hypothese H_0 : „$\beta = \beta_0$“ wird zum Niveau $1-\gamma$ abgelehnt, falls

$$\frac{|\hat{\beta} - \beta_0|}{S/\sqrt{(n-1)s_{xx}^2}} > t_{n-2,1-\frac{1-\gamma}{2}}.$$

Hier:

$$\bar{x}_{10} = 762, \ \sqrt{9 \cdot s_{xx}^2} = 1139.24$$

und somit

$$\frac{|\hat{\beta}|}{S/\sqrt{9 \cdot s_{xx}^2}} = \frac{0.0036}{0.48/1139.24} = \frac{0.0036}{0.0004} = 9.00.$$

Andererseits gilt $t_{8,0.975} = 2.306$ und somit wird die Hypothese H_0 : „$\beta = 0$“ zum Niveau 5% abgelehnt, d. h. es besteht ein signifikanter Zusammenhang zwischen der Länge des Transportweges und der Lieferzeit.

Weitere Begriffe aus der Testtheorie:

- *Fehler 1. Art*: fälschliches Ablehnen der Hypothese
- *Fehler 2. Art*: fälschliches Annehmen der Hypothese
- *p-Wert*: kleinstes Signifikanzniveau für das die Nullhypothese H_0 noch abgelehnt werden kann

(e) *Aus (d)*: es gilt mit Wahrscheinlichkeit γ:

- $$\hat{\alpha} - t_{n-2,1-\frac{1-\gamma}{2}} S \sqrt{\left(\sum_{i=1}^{n} x_i^2\right) / (n(n-1)s_{xx}^2)} < \alpha$$
$$< \hat{\alpha} + t_{n-2,1-\frac{1-\gamma}{2}} S \sqrt{\left(\sum_{i=1}^{n} x_i^2\right) / (n(n-1)s_{xx}^2)}$$

- $$\hat{\beta} - t_{n-2,1-\frac{1-\gamma}{2}} S/\sqrt{(n-1)s_{xx}^2} < \beta < \hat{\beta} + t_{n-2,1-\frac{1-\gamma}{2}} S/\sqrt{(n-1)s_{xx}^2}.$$

<u>*Hier*</u>:

- $$t_{8,0.975} S \sqrt{\left(\sum_{i=1}^{10} x_i^2\right) / (10 \cdot 9 \cdot s_{xx}^2)} = 2.306 \cdot 0.48 \cdot \sqrt{7104300/(90 \cdot 144206.7)} = 0.8189$$

  und somit gilt mit Wahrscheinlichkeit 95%

  $$-0.7008 < \alpha < 0.9370.$$

- $$t_{8,0.975} S/\sqrt{(n-1)s_{xx}^2} = 2.306 \cdot 0.48/\sqrt{9 \cdot 144206.7} = 0.0009716,$$

  also mit Wahrscheinlichkeit 95%

  $$0.002613 < \beta < 0.0045567.$$

2. Tutorium am 08.12.08/11.12.08

(Multivariate lineare Regression)

Aufgabe (vgl. Ökonometrie Blatt 2 Nr. 2 (WS 2008/09))

In neun verschiedenen amerikanischen Wintersportorten wurden während einer gewissen Beobachtungszeit die Anzahl der Besucher registriert. Es wird angenommen, dass diese linear von der Gesamtlänge der zur Verfügung stehenden Pisten sowie der Liftkapazität abhängen.

Skigebiet	Pistenlänge	Liftkapazität	Besucherzahl
1	10.5	2200	19929
2	2.5	1000	5839
3	13.1	3250	23696
4	4.0	1475	9881
5	14.7	3800	30011
6	3.6	1200	7241
7	7.1	1900	11634
8	17.0	4200	36476
9	6.4	1850	12068

(a) Erstelle ein geeignetes Regressionsmodell, wobei die üblichen Normalverteilungsannahmen gelten sollen.

(b) Teste die Hypothese H_0, dass keine Abhängigkeit der Besucherzahl von den beiden Einflussgrößen besteht für $\alpha = 0.05$. Runde alle Zwischenergebnisse auf 4 Dezimalen und rechne mit den gerundeten Werten weiter.

(c) Prüfe nun einzeln die Hypothesen $H_0^{(1)}$: „Es besteht keine Abhängigkeit des Besucheraufkommens von der Pistenlänge“ und $H_0^{(2)}$: „Es besteht keine Abhängigkeit des Besucheraufkommens von der Liftkapazität“ ($\alpha = 0.05$). *Hinweis*: $F_{2,6,0.95} = 5.14$.

(d) In einem zehnten Skigebiet stehen insgesamt 15.0 km Piste zur Verfügung bei einer Liftkapazität von 3950. Wie viele Besucher können in diesem Gebiet erwartet werden? *Hinweis*: $F_{1,6,0.95} = 5.99$.

Multivariate lineare Regression

Gegeben zwei Datensätze

$$(\overrightarrow{x_1}, \overrightarrow{x_2}, \ldots, \overrightarrow{x_n}) \text{ Ausgangsvariable und } (y_1, y_2, \ldots, y_n) \text{ Zielvariable,}$$

wobei

$$\overrightarrow{x_i} = \begin{pmatrix} x_{i,2} \\ x_{i,3} \\ \vdots \\ x_{i,m} \end{pmatrix}.$$

Vermutung: es besteht ein linearer Zusammenhang zwischen x und y:

$$y_i = \beta_1 + \beta_2 x_{i,2} + \ldots + \beta_m x_{i,m} + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d. } \mathcal{N}(0, \sigma^2).$$

In Matrix-Schreibweise:

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

wobei

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{1,2} & x_{1,3} & \cdots & x_{1,m} \\ 1 & x_{2,2} & x_{2,3} & \cdots & x_{2,m} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n,2} & x_{n,3} & \cdots & x_{n,m} \end{pmatrix} \text{ „Design-Matrix“}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Aufgaben zur multivariaten Regression:

(i) Berechne den MKQ-Schätzer $\hat{\beta}$ für β;

(ii) Teste Hypothesen über β, z. B.

$$H_0 : \text{ „}\beta_1 = \beta_2 = 0\text{“}.$$

Zu (i):
Bekannt ist:

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

ist der MKQ-Schätzer für β (oft sind $(X^T X)^{-1}$ und $X^T y$ gegeben $\rightsquigarrow$ multipliziere Matrix mit Vektor).
Zu (ii):
Bekannt ist: Für den Test

$$H_0 : \text{ „}H\beta = d\text{“}$$

ist

$$T_H := \frac{(H\hat{\beta} - d)^T (H(X^T X)^{-1} H^T)^{-1} (H\hat{\beta} - d)}{sS^2} \sim F_{s,n-m},$$

wobei

$$\begin{aligned} S^2 &= \frac{1}{n-m}(Y - X\hat{\beta})^T (Y - X\hat{\beta}) = \frac{1}{n-m}(Y - \hat{Y})^T (Y - \hat{Y}) \\ n &= \text{Stichprobenumfang} \\ m &= \text{Anzahl der Spalten von } X \\ s &= \text{Rang}(H). \end{aligned}$$

Die Hypothese H_0 wird zum Niveau $1 - \gamma$ abgelehnt, falls $T_H > F_{s,n-m,\gamma}$ ist.

Zur Aufgabe:

(a) *Regressionsmodell*:

y_i: Besucherzahl (Regressand)
$x_{i,2}$: Pistenlänge (Regressor 1)
$x_{i,3}$: Liftkapazität (Regressor 2)

$$y_i = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \varepsilon_i, \quad i = 1, ..., 9 \quad \varepsilon_i \text{ i.i.d. } \mathcal{N}(0, \sigma^2)$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_9 \end{pmatrix} = \begin{pmatrix} 1 & x_{1,2} & x_{1,3} \\ 1 & x_{2,2} & x_{2,3} \\ \vdots & \vdots & \vdots \\ 1 & x_{9,2} & x_{9,3} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_9 \end{pmatrix}$$

(b) Teste

$$H_0 : \text{„}\beta_2 = \beta_3 = 0\text{“ vs. } H_1 : \text{„}\beta_2 \neq 0 \text{ oder } \beta_3 \neq 0\text{“}$$

Der Test lässt sich auch wie folgt schreiben: teste

$$H_0 : \text{„}H\beta = 0\text{“ vs. } H_1 : \text{„}H\beta \neq 0\text{“, mit } H = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ und } \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Mit

$$\hat{\beta} = (X^T X)^{-1} X^T y \stackrel{R}{=} \begin{pmatrix} -2017.5677 \\ 1098.4661 \\ 4.2282 \end{pmatrix}$$

folgt dann

$$T_H := \frac{(H\hat{\beta})^T (H(X^T X)^{-1} H^T)^{-1} (H\hat{\beta})}{2S^2} \sim F_{2,9-3},$$

mit

$$S^2 = \frac{1}{9-3}(Y - X\hat{\beta})^T (Y - X\hat{\beta}) \stackrel{R}{=} 3191332.$$

Die Hypothese H_0 wird abgelehnt, da $T_H \stackrel{R}{=} 141.773 > F_{2,6,0.95} = 5.14$. Es hat also mindestens einer der beiden Faktoren Einfluss auf die Besucherzahl.

(c) Teste

$$H_0^{(1)} : \text{„}\beta_2 = 0\text{“ vs. } H_1^{(1)} : \text{„}\beta_2 \neq 0\text{“}$$

bzw. teste

$$H_0^{(2)} : \text{„}\beta_3 = 0\text{“ vs. } H_1^{(2)} : \text{„}\beta_3 \neq 0\text{“}.$$

Die Tests lassen sich auch wie folgt schreiben: teste

$$H_0^{(1)} : \text{„}H^{(1)}\beta = 0\text{“ vs. } H_1^{(1)} : \text{„}H^{(1)}\beta \neq 0\text{“, mit } H^{(1)} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix},$$

bzw.

$$H_0^{(2)} : \text{„}H^{(2)}\beta = 0\text{“ vs. } H_1^{(2)} : \text{„}H^{(2)}\beta \neq 0\text{“, mit } H^{(2)} = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \text{ und } \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Beachte: Der Rang der Matrizen $H^{(1)}$ bzw. $H^{(2)}$ ist jeweils $s = 1$. Es gilt:

$$T_H^{(1)} \stackrel{R}{=} 2.620317 < F_{1,6,0,95} = 5.99 \quad \text{bzw.} \quad T_H^{(2)} \stackrel{R}{=} 1.874717 < F_{1,6,0,95} = 5.99.$$

Beide Hypothesen werden also nicht verworfen.

(d)

$$\hat{y}_1 = \hat{\beta}_1 + 15.0\ \hat{\beta}_2 + 3950\ \hat{\beta}_3 \overset{R}{=} 31160.9$$

Exkurs: Determinante, Rang und Inverse einer Matrix

- *Determinante von 2×2 bzw. 3×3-Matrizen*:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33}$$

- *Entwicklung nach der i-ten Zeile (zur Berechnung der Determinante von $n \times n$-Matrizen)*:

$$\begin{vmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \cdots & a_{nj} & \cdots & a_{nn} \end{vmatrix} = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \begin{vmatrix} a_{1,1} & \cdots & a_{1,j-1} & a_{1,j+1} & \cdots & a_{1,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{i-1,1} & \cdots & a_{i-1,j-1} & a_{i-1,j+1} & \cdots & a_{i-1,n} \\ a_{i+1,1} & \cdots & a_{i+1,j-1} & a_{i+1,j+1} & \cdots & a_{i+1,n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n,1} & \cdots & a_{n,j-1} & a_{n,j+1} & \cdots & a_{n,n} \end{vmatrix}$$

- *Rang einer Matrix*: der Rang einer Matrix entspricht der Anzahl linear unabhängiger Spaltenvektoren:
  - Nur $n \times n$-Matrizen können vollen Rang haben.
  - Genau dann wenn die Determinante einer $n \times n$-Matrix $\neq 0$ ist, hat die Matrix vollen Rang n.
  - Bringe Matrix in Dreiecksform um Rang abzulesen.

- *Die Inverse einer 2×2-Matrix (beachte: es können nur Matrizen mit vollem Rang invertiert werden)*:

$$A^{-1} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1} = \frac{1}{|A|} \begin{pmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{pmatrix}.$$

- *Cramersche Regel für Matrixinversion (für $n \times n$-Matrizen)*:

$$A^{-1} = \begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \cdots & a_{nj} & \cdots & a_{nn} \end{pmatrix}^{-1} = \begin{pmatrix} a'_{11} & \cdots & a'_{1j} & \cdots & a'_{1n} \\ \vdots & & \vdots & & \vdots \\ a'_{i1} & \cdots & a'_{ij} & \cdots & a'_{in} \\ \vdots & & \vdots & & \vdots \\ a'_{n1} & \cdots & a'_{nj} & \cdots & a'_{nn} \end{pmatrix},$$

  mit

$$a'_{ij} = \frac{1}{|A|} (-1)^{i+j} \left| A_{ji} \right|,$$

  wobei A_{ji} eine $(n-1) \times (n-1)$-Matrix ist, welche aus A durch Streichen der j-ten Zeile und der i-ten Spalte hervorgeht.

3. Tutorium am 15.12.08/18.12.08

(Linearisierung und RESET-Test)

Aufgabe 1 (Linearisierung von nichtlinearen Modellen)

(a) Linearisiere das nicht-lineare Modell

$$y_i = \beta_1 + \beta_2 \cos(\beta_3 x_i) + \varepsilon_i$$

mit Hilfe der Taylor-Approximation und stelle das zugehörige linearisierte Regressionsmodell auf.

(b) Bringe das nicht-lineare Modell

$$y_i = \beta_1 \cdot e^{\beta_2 x_2} \cdot x_3^{\beta_3} \cdot e^{\varepsilon_i}$$

durch eine geeignete Transformation auf ein lineares Modell.

Aufgabe 2 (Reset-Test)

Um die Linearitätsannahme von Aufgabe 2 auf Blatt 2 zu überprüfen wurde mittels R ein RESET-Test durchgeführt:

```
> resettest(Besucherzahl ~ 1 + Liftkapazitaet + Pistenlaenge, power=2:3, type="fitted")

        RESET test

data:  Besucherzahl ~ 1 + Liftkapazitaet + Pistenlaenge
RESET = 3.0807, df1 = 2, df2 = 4, p-value = 0.1550
```

Interpretiere das Ergebnis.

Linearisierung mittels Taylor

Gegeben ein nichtlineares Regressionsmodell

$$y_i = f(\vec{x}_i, \beta) + \varepsilon_i,$$

mit einer (zweimal differenzierbaren) Funktion f. Dann lässt sich f um den Punkt $\beta^0 \in \mathbb{R}^m$ wie folgt mit Hilfe der multivariaten Taylorformel entwickeln:

$$\begin{aligned} y_i &= f(\vec{x}_i, \beta) + \varepsilon_i \\ &\approx f(\vec{x}_i, \beta^0) + \sum_{k=1}^{m} \left.\frac{\partial f}{\partial \beta_k}\right|_{\beta=\beta^0, \vec{x}=\vec{x}_i} \cdot (\beta_k - \beta_k^0) + \varepsilon_i. \end{aligned}$$

Mit den Bezeichnungen

$$\begin{aligned} \tilde{x}_{i,k}^0 &:= \left.\frac{\partial f}{\partial \beta_k}\right|_{\beta=\beta^0, \vec{x}=\vec{x}_i} \\ \tilde{y}_i^0 &:= y_i - f(\vec{x}_i, \beta^0) + \sum_{k=1}^{m} \left.\frac{\partial f}{\partial \beta_k}\right|_{\beta=\beta^0, \vec{x}=\vec{x}_i} \cdot \beta_k^0 \\ &= y_i - f(\vec{x}_i, \beta^0) + \sum_{k=1}^{m} \tilde{x}_{i,k}^0 \beta_k^0, \end{aligned}$$

erhält man das *lineare Modell*

$$\tilde{y}_i^0 = \sum_{k=1}^{m} \tilde{x}_{i,k}^0 \cdot \beta_k + \varepsilon_i.$$

Zu Aufgabe 1:

(a) Hier ist $f(x_i, \beta_1, \beta_2, \beta_3) = \beta_1 + \beta_2 \cos(\beta_3 x_i)$ und

$$\begin{aligned}
\tilde{x}_{i,1}^0 &= \left.\frac{\partial f}{\partial \beta_1}\right|_{\beta=\beta^0, x=x_i} &&= 1 \\
\tilde{x}_{i,2}^0 &= \left.\frac{\partial f}{\partial \beta_2}\right|_{\beta=\beta^0, x=x_i} &&= \cos(\beta_3^0 x_i) \\
\tilde{x}_{i,3}^0 &= \left.\frac{\partial f}{\partial \beta_3}\right|_{\beta=\beta^0, x=x_i} &&= -\beta_2^0 x_i \sin(\beta_3^0 x_i),
\end{aligned}$$

und

$$\begin{aligned}
\tilde{y}_i^0 &= y_i - f(x_i, \beta^0) + \beta_1^0 \tilde{x}_{i,1}^0 + \beta_2^0 \tilde{x}_{i,2}^0 + \beta_3^0 \tilde{x}_{i,3}^0 \\
&= y_i - (\beta_1^0 + \beta_2^0 \cos(\beta_3^0 x_i)) + \beta_1^0 + \beta_2^0 \cos(\beta_3^0 x_i) - \beta_3^0 \beta_2^0 x_i \sin(\beta_3^0 x_i) \\
&= y_i - \beta_3^0 \beta_2^0 x_i \sin(\beta_3^0 x_i).
\end{aligned}$$

Das linearisierte Modell lautet:

$$\tilde{y}_i^0 = \beta_1 \tilde{x}_{i,1}^0 + \beta_2 \tilde{x}_{i,2}^0 + \beta_3 \tilde{x}_{i,3}^0 + \varepsilon_i.$$

(b) Anwenden des Logarithmus auf beiden Seiten ergibt:

$$\begin{aligned}
\log(y_i) &= \log(\beta_1 \cdot e^{\beta_2 x_2} \cdot x_3^{\beta_3} \cdot e^{\varepsilon_i}) \\
&= \log(\beta_1) + \beta_2 x_2 + \beta_3 \log(x_3) + \varepsilon_i \\
&= \tilde{\beta}_1 + \beta_2 x_2 + \beta_3 \tilde{x}_3 + \varepsilon_i,
\end{aligned}$$

wobei $\tilde{\beta}_1 = \log(\beta_1)$ und $\tilde{x}_3 = \log(x_3)$.

Der Reset-Test

Ziel Testen der Linearitätsannahme

Vorgehensweise

(1.) Berechne in dem *linearen Modell*

$$y_i = \beta_1 + \beta_2 x_{i,2} + \ldots + \beta_m x_{i,m} + \varepsilon_i$$

den MKQ-Schätzer $\hat{\beta}$ und die gefitteten Werte

$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i,2} + \ldots + \hat{\beta}_m x_{i,m}.$$

(2.) Ergänze das lineare Modell um Potenzen von $\hat{y}_i$, z. B. ergibt sich bis zur Ordnung 4 folgendes lineare Modell:

$$y_i = \beta_1 + \beta_2 x_{i,2} + \ldots + \beta_m x_{i,m} + \beta_{m+1} x_{i,m+1} + \beta_{m+2} x_{i,m+2} + \beta_{m+3} x_{i,m+3} + \varepsilon_i,$$

wobei $x_{i,m+1} = \hat{y}_i^2$, $x_{i,m+2} = \hat{y}_i^3$, $x_{i,m+3} = \hat{y}_i^4$.

Idee:

- z. B. enthält $\hat{y}_i^2$ Informationen über (alle) Terme $x_{i,k} \cdot x_{i,l}$ der Ordnung zwei;
- statt alle möglichen gemischten Terme ins erweiterte Modell aufzunehmen reicht es deshalb Potenzen von $\hat{y}_i$ aufzunehmen;
- jedes nichtlineare Modell kann mit der Taylorentwicklung (in x) durch ein polynomiales Modell approximiert werden.

(3.) Teste in dem erweiterten Modell die Hypothese

H_0 : „$\beta_{m+1} = \beta_{m+2} = \beta_{m+3} = 0$“ vs. H_1 : „$\beta_{m+1} \neq 0$ oder $\beta_{m+2} \neq 0$ oder $\beta_{m+3} \neq 0$“.

Dies führt wie im 2. Tutorium erläutert auf einen F-Test.

(4.) Wenn H_0 abgelehnt wird, so wird die Annahme der Linearität verworfen.

Zu Aufgabe 2:

Hier werden die Potenzen $\hat{y}_i^2$ und $\hat{y}_i^3$ mit ins Modell aufgenommen. Der p-Wert ist 0.1550 und die Hypothese „$\beta_4 = \beta_5 = 0$“ wird somit *nicht* verworfen. Damit wird die Annahme der Linearität *nicht* verworfen.

4. Tutorium am 22.12.08/08.01.09

(ML-Schätzer und Likelihood-Ratio-Test)

Aufgabe 1 (ML-Schätzer)
Bestimme den Maximum-Likelihood-Schätzer für (α, β) im einfachen linearen Regressionsmodell

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad \forall i = 1, ..., n \quad \text{mit } \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \varepsilon_i \text{ i.i.d. .}$$

Aufgabe 2 (Likelihood-Ratio-Test)
Vgl. Ökonometrie Blatt 2 Nr. 2 sowie Blatt 3 Nr. 3 (WS 2008/09).

(a) Berechne die Maximum-Likelihood-Schätzer für die Parameter im Box-Cox-Modell.

(b) Überprüfe, ob die Annahme eines linearen multivariaten Regressionsmodells im Rahmen des Box-Cox-Modells verworfen wird.

Der ML-Schätzer

Grundidee: Bestimme unbekannte Parameter so, dass die gemeinsame Wahrscheinlichkeit für eine gegebene Stichprobe maximiert wird.

Voraussetzung: Die Wahrscheinlichkeitsverteilung ist bekannt und die Stichprobe stochastisch unabhängig.

Vorgehensweise:

(a) Stelle Likelihood-Funktion auf: $L(\beta|x) = \prod_{i=1}^{n} f(\beta|x_i)$.

(b) Wenn vorteilhaft, benutze die Log-Likelihood-Funktion $\log(L(\beta|x))$.

(c) Suche die Parameter, die die (Log-)Likelihood-Funktion maximieren.

Der ML-Schätzer im Box-Cox-Modell

Sei $x^{(\lambda)} = \frac{x^\lambda - 1}{\lambda}$ für $\lambda \neq 0$ und $x^{(\lambda)} = \log x$ für $\lambda = 0$ (identisch für $y_i^{(\theta)}$). Das Box-Cox-Modell ist dann von folgender Form:

$$y_i^{(\theta)} = \beta_1 x_{i,1}^{(\lambda_1)} + \beta_2 x_{i,2}^{(\lambda_2)} + ... + \beta_m x_{i,m}^{(\lambda_m)} + \varepsilon_i \qquad \forall i = 1, ..., n \quad \text{mit } \varepsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2).$$

Die Likelihood-Funktion lautet folglich:

$$\begin{aligned}
L(\theta, \beta, \lambda|x) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2} e_i^2\right) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left(y_i^{(\theta)} - \sum_{j=1}^{m} \beta_j x_{ij}^{(\lambda_j)}\right)^2\right) y_i^{\theta-1} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n}\left(y_i^{(\theta)} - \sum_{j=1}^{m} \beta_j x_{ij}^{(\lambda_j)}\right)^2\right) \prod_{i=1}^{n} y_i^{\theta-1},
\end{aligned}$$

und die Log-Likelihood ist gegeben durch

$$\begin{aligned}\log\left(L(\theta,\beta,\lambda|x)\right) &= -\frac{n}{2}\log\left(2\pi\right)-\frac{n}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i^{\theta}-\sum_{j=1}^{m}\beta_j x_{ij}^{(\lambda_j)}\right)^2\\ &\quad+(\theta-1)\sum_{i=1}^{n}\left(\log y_i\right).\end{aligned}$$

Der Likelihood-Ratio-Test

Gegeben ist: $\theta=\lambda_1=\lambda_2=\ldots=\lambda_m$. Teste

$$H_0:\ „\lambda=\lambda_0“ \text{ vs. } H_1:\ „\lambda\neq\lambda_0“,$$

z. B. wird für $\lambda_0=1$ getestet, ob ein lineares Modell vorliegt. Die Teststatistik

$$T_H=-2\log\frac{L(\lambda_0,\widehat{\beta}|x)}{L(\widehat{\lambda},\widehat{\beta}|x)}$$

ist χ_1^2 -verteilt, wobei $\widehat{\lambda}$ und $\widehat{\beta}$ die ML-Schätzer sind.

Zu Aufgabe 1:

Es gilt:

$$\varepsilon_i=y_i-\alpha-\beta x_i \quad \varepsilon_i\sim\mathcal{N}(0,\sigma^2).$$

Die (Log-)Likelihood-Funktion lautet folglich:

$$\begin{aligned}L((\alpha,\beta)|x) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}e_i^2\right)\\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}e_i^2\right)\\ \log\left(L((\alpha,\beta)|x)\right) &= n\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}e_i^2.\end{aligned}$$

Nun sollen α und β so gewählt werden, dass die Log-Likelihood-Funktion maximiert wird. Dies ist gleichbedeutend mit einer Minimierung des Terms $\frac{1}{2\sigma^2}\sum_{i=1}^{n}e_i^2$ bzw. $\sum_{i=1}^{n}e_i^2$. Dies entspricht aber genau dem Ansatz für die Herleitung des schon bekannten MKQ-Schätzers für (α,β). In diesem Fall ist der ML-Schätzer also gleichzeitig der MKQ-Schätzer.

Zu Aufgabe 2:

(a) Es gilt: $\theta=\lambda_1=\lambda_2=\lambda_3$. Die Log-Likelihood-Funktion lautet dann:

$$\begin{aligned}\log\left(L(\lambda|x)\right) &= -\frac{9}{2}\log\left(2\pi\right)-\frac{9}{2}\log\left(\sigma^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{9}\left(y_i^{(\lambda)}-\sum_{j=1}^{3}\beta_j x_{ij}^{(\lambda)}\right)^2\\ &\quad+(\lambda-1)\sum_{i=1}^{9}\left(\log y_i\right)\end{aligned}$$

Der Rest der Berechnung wird R überlassen:

```
### Zunächst die Daten in Form von Vektoren
Besucherzahl <- c(19929,5829,23696,9881,30011,7241,11634,36476,12068)
Pistenlänge <- c(10.5,2.5,13.1,4.0,14.7,3.6,7.1,17.0,6.4)
Liftkapazität <- c(2200,1000,3250,1475,3800,1200,1900,4200,1850)

### Nun die lineare Regression
output <- summary(lm(formula = Besucherzahl ~ 1 + Pistenlänge + Liftkapazität))

### Die Log-Likelihood-Funktion
ll <- function(x){
    sigma <- x[1]
    beta1 <- x[2]
    beta2 <- x[3]
    beta3 <- x[4]
    lambda <- x[5]
    n <- length(Besucherzahl)
    llwert <- {-n/2*log(2*pi)-n/2*log(sigma^2)+(lambda-1)*sum(log(Besucherzahl))-
        1/(2*sigma^2)*sum(((Besucherzahl^lambda - 1)/lambda - beta1 -
        beta2*(Pistenlänge^lambda-1)/lambda - beta3*(Liftkapazität^lambda - 1)/lambda)^2)}
    return(-llwert)
}

### Optimierung mit nlm mit MKQ-Werten als Startwerte (Achtung nlm minimiert eine Funktion!!!)
nlm(ll,c(output$sigma,output$coefficient[1],output$coefficient[2],output$coefficient[3],1))$estimate
[1]  1785.609234 -2020.699701  1098.983590     5.256928     1.020303
```

In der letzten Zeile werden die ML-Schätzer $\widehat{\sigma}$, $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$ und $\widehat{\lambda}$ ausgegeben.

(b) Überprüfe, ob die Annahme eines linearen multivariaten Regressionsmodells im Rahmen des Box-Cox-Modells verworfen wird. Hierfür muss zunächst folgende Teststatistik berechnet werden:

$$T_H = -2 \log \frac{L(\lambda = 1, \widehat{\beta}, \widehat{\sigma}|x)}{L(\widehat{\lambda}, \widehat{\beta}, \widehat{\sigma}|x)}$$

Auch das wird R überlassen:

```
> {ML <- nlm(ll,c(output$sigma,output$coefficient[1],
+ output$coefficient[2],output$coefficient[3],1))$estimate}
>
> ### Die Teststatistik zum Likelihood-Ratio-Test
> T <- -2*(-ll(c(ML[1],ML[2],ML[3],ML[4],1)) + ll(ML));T
[1] 8.596016
> qchisq(0.95,df=1)
[1] 3.841459
```

Es ist $T_H > \chi^2_{1,0.95}$. Daher wird die Hypothese abgelehnt.

5. Tutorium am 12.01.09/15.01.09

(Verallgemeinerte lineare Modelle – Teil 1)

Aufgabe 1 (Exponentialfamilie und natürliche Linkfunktion) Zeige, dass

(a) die Exponentialverteilung $\mathrm{Exp}(\mu)$ mit der Dichte

$$f(y) = \begin{cases} \frac{1}{\mu} e^{-\frac{x}{\mu}}, & y > 0 \\ 0, & \text{sonst,} \end{cases}$$

sowie

(b) die negative Binomialverteilung $\mathrm{NB}(\mu, k)$ mit der Zähldichte

$$\mathbb{P}(Y = y) = \binom{k+y-1}{k-1} \mu^k (1-\mu)^y,$$

wobei $k > 0$ bekannt sei,

zur Exponentialfamilie gehören und berechne jeweils die natürliche Linkfunktion.

Aufgabe 2 (Logistisches Regressionsmodell – vgl. Ökonometrie Blatt 4 Nr. 4 (WS 2008/09))
Auf der Homepage der Vorlesung befindet sich die Datei `challenger.txt`, die folgende Werte enthält:

Temperatur	53	57	58	63	66	67	67	67	68	69	70	70
Ausfall	1	1	1	1	0	0	0	0	0	0	0	0

Temperatur	70	70	72	73	75	75	76	76	78	79	81
Ausfall	1	1	0	0	0	1	0	0	0	0	0

Der Wert *Temperatur* ist die Außentemperatur (in Fahrenheit) beim Start der 23 Space-Shuttle Flüge vor der Challenger Katastrophe und der Parameter *Ausfall* gibt an, ob mindestens einer der Dichtungsringe wegen Materialermüdung ausgefallen ist (1) oder nicht (0). Untersuche mit Hilfe eines logistischen Regressionsmodells (Logit-Modells) den *Einfluss der Temperatur auf das Auftreten solcher Materialermüdungserscheinungen*. Welche Wahrscheinlichkeit wird für das Versagen mindestens eines Dichtungsringes prognostiziert, wenn die Außentemperatur wie am Unglückstag 31° F beträgt?

Hinweis:
Für verallgemeinerte lineare Modelle gibt es in R den Befehl `glm()`; ein Bernoulli-Modell mit natürlicher Linkfunktion wird durch den Parameter `family=binomial(link=''logit'')` spezifiziert.

Die Exponentialfamilie

Die Verteilung einer Zufallsvariablen Y gehört zur *Exponentialfamilie*, falls es einen Parameter θ und Funktionen a, b gibt, so dass:

- im absolutstetigen Fall hat die Dichte von Y die Form

$$f(y) = e^{\frac{1}{\tau^2}(y\theta + a(y,\tau) - b(\theta))};$$

- im diskreten Fall hat die Zähldichte von Y die Form

$$\mathbb{P}(Y = y) = e^{\frac{1}{\tau^2}(y\theta + a(y,\tau) - b(\theta))}.$$

Verallgemeinerte lineare Modelle

Bei der Betrachtung linearer Modelle $Y = X\beta + \varepsilon$ hatten wir stets vorausgesetzt,

- dass

$$\mathbb{E}[\varepsilon] = 0, \text{ d. h. } (\mathbb{E}[Y_1], \ldots, \mathbb{E}[Y_n])^T = X\beta,$$

- und dass die Störterme verteilt sind gemäß

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

Dieses Modell kann nun mit folgenden zwei Verallgemeinerungen betrachtet werden:

(1.) die **Erwartungswerte** $\mathbb{E}[Y_1], \ldots, \mathbb{E}[Y_n]$ der Stichprobenvariablen $Y_1, \ldots, Y_n$ können über eine beliebige (monotone) Funktion g – die so genannte *Linkfunktion* – durch die Komponenten des Vektors $X\beta$ ausgedrückt werden, so dass

$$(g(\mathbb{E}[Y_1]), \ldots, g(\mathbb{E}[Y_n]))^T = X\beta.$$

(2.) Die (unabhängigen) Stichprobenvariablen $Y_1, \ldots, Y_n$ müssen nicht notwendig normalverteilt sein; wir nehmen nur an, dass die Verteilungen von $Y_1, \ldots, Y_n$ zu einer *Exponentialfamilie* gehören.

Natürliche Linkfunktion

Die Funktion g heißt *natürliche Linkfunktion*, falls $g = (b')^{-1}$, g zweimal stetig differenzierbar und $g'(x) \neq 0$ ist.

Zu Aufgabe 1:

(a) Es ist

$$\begin{aligned} f(y) &= \frac{1}{\mu} e^{-\frac{x}{\mu}} \\ &= e^{-\log(\mu) - \frac{x}{\mu}}, \end{aligned}$$

so dass

$$\tau = 1, \quad \theta = -\frac{1}{\mu}, \quad b(\theta) = \log\left(-\frac{1}{\theta}\right), \quad a(y,\tau) \equiv 0.$$

Außerdem ist $b'(\theta) = -\frac{1}{\theta}$, so dass die natürliche Linkfunktion gegeben ist durch

$$g(y) = -\frac{1}{y}.$$

(b) Es gilt

$$\begin{aligned} f(y) &= \binom{k+y-1}{k-1}\mu^k(1-\mu)^y \\ &= e^{\log\left[\binom{k+y-1}{k-1}\right]+k\log(\mu)+y\log(1-\mu)}, \end{aligned}$$

so dass

$$\tau = 1, \quad a(y,\tau) = \log\left[\binom{k+y-1}{k-1}\right], \quad \theta = \log(1-\mu) \quad \text{und} \quad b(\theta) = -k\log(\mu).$$

Da

$$\begin{aligned} & \theta = \log(1-\mu) \\ \iff \quad & e^\theta = 1-\mu \\ \iff \quad & \mu = 1-e^\theta, \end{aligned}$$

ist $b(\theta) = -k\log\left(1-e^\theta\right)$. Hieraus errechnet sich die natürliche Linkfunktion g wie folgt:

$$\begin{aligned} & b'(\theta) = \frac{ke^\theta}{1-e^\theta} \stackrel{!}{=} y \\ \iff \quad & \frac{k}{y} = \frac{1-e^\theta}{e^\theta} = e^{-\theta} - 1 \\ \iff \quad & \frac{k+y}{y} = e^{-\theta} \\ \iff \quad & \frac{y}{k+y} = e^\theta \\ \iff \quad & \theta = \log\left(\frac{y}{k+y}\right), \end{aligned}$$

so dass $g(y) = \log\left(\frac{y}{k+y}\right)$.

Zu Aufgabe 2: Wir modellieren

$$Y_i := \begin{cases} 1 & \text{falls beim } i\text{-ten Flug ein Ausfall stattfindet,} \\ 0 & \text{falls beim } i\text{-ten Flug kein Ausfall stattfindet.} \end{cases}$$

Dann ist $Y_i \sim \text{Bernoulli}(p_i)$, wobei die Ausfallwahrscheinlichkeit $p_i \in [0,1]$ unbekannt ist. Wir nehmen nun an, dass p_i von der Temperatur abhängig ist und wählen hierzu ein verallgemeinertes Regressionsmodell der Form

$$g(\mathbb{E}[Y_i]) = g(p_i) = \beta_1 + \beta_2 x_{i,2},$$

wobei $x_{i,2}$ die Temperatur beim i-ten Flug ist. Als Linkfunktion g wählen wir die natürliche Linkfunktion im Bernoulli-Modell, d. h.

$$g(y) = \log\left(\frac{y}{1-y}\right).$$

Unser verallgemeinertes lineares Modell ist damit vollständig spezifiziert, genauer gesagt lauten die Modellparameter:

- $\theta = \log\left(\frac{p}{1-p}\right)$, $\tau^2 = 1$, $a(y,\tau) \equiv 0$, $b(\theta) = \log\left(1+e^{\theta}\right)$ – da Y Bernoulli-verteilt;
- $g(y) = \log\left(\frac{y}{1-y}\right)$ – da wir die natürliche Linkfunktion wählen.

Wir schätzen nun (mit R) die Parameter β_1 und β_2 wie folgt:

```
> ##Einlesen und verallgemeinerte lineare Regression
> PFAD <- "D:/.../"
> daten <- read.table(file = paste(PFAD, "challenger.txt", collapse="", sep=""), header = TRUE)
> attach(daten)
> names(daten)
[1] "Temperatur" "Ausfall"
>
> glm.challenger <- glm(Ausfall ~ 1 + Temperatur, family=binomial(link="logit"))
> summary(glm.challenger)

Call: glm(formula = Ausfall ~ 1 + Temperatur, family = binomial(link
= "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0611  -0.7613  -0.3783   0.4524   2.2175

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
Temperatur   -0.2322     0.1082  -2.145   0.0320 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.267  on 22  degrees of freedom
Residual deviance: 20.315  on 21  degrees of freedom AIC: 24.315

Number of Fisher Scoring iterations: 5
```

Es ist also $\beta_1 = 15.0429$ und $\beta_2 = -0.2322$. Damit können wir die Ausfallwahrscheinlichkeit am Unglückstag wie folgt schätzen:

$$\begin{aligned} & \log\left(\frac{p}{1-p}\right) = 15.0429 - 0.2322 \cdot 31 = 7.8447 \\ \iff \quad & \frac{p}{1-p} = e^{7.8447} \\ \iff \quad & p = \frac{e^{7.8447}}{1+e^{7.8447}} = 0.9996083. \end{aligned}$$

Exkurs: weitere Aufgaben zur Linearisierung von nichtlinearen Modellen

Aufgabe

(a) Linearisiere mit Hilfe der Taylor-Approximation die nicht-linearen Modelle

(i) $y_i = \beta_1\beta_2 + \beta_3^2 x_i + \varepsilon_i$ um $\beta^0 = (\beta_1^0, \beta_2^0, \beta_3^0)$, sowie

(ii) $y_i = \beta_1^2 x_{i,1} + \beta_2^3 x_{i,2} + \varepsilon_i$ um $\beta_0 = (1, 2)$.

(b) Bringe das nicht-lineare Modell

$$y_i = \beta_1 \cdot x_{i,2}^{\beta_2} \cdot e^{\beta_3 x_{i,3}^2 + 2\beta_4 x_{i,4}} \cdot e^{\varepsilon_i}$$

durch eine geeignete Transformation auf ein lineares Modell.

Zur Aufgabe:

(a) (i)

$$\begin{aligned}
\tilde{x}_{i,1}^0 &:= \left.\frac{\partial f}{\partial \beta_1}\right|_{\beta=\beta^0, \vec{x}=\vec{x}_i} = \beta_2^0 \\
\tilde{x}_{i,2}^0 &:= \left.\frac{\partial f}{\partial \beta_2}\right|_{\beta=\beta^0, \vec{x}=\vec{x}_i} = \beta_1^0 \\
\tilde{x}_{i,3}^0 &:= \left.\frac{\partial f}{\partial \beta_3}\right|_{\beta=\beta^0, \vec{x}=\vec{x}_i} = 2\beta_3^0 x_i \\
\tilde{y}_i^0 &:= y_i - f(\vec{x}_i, \beta^0) + \sum_{k=1}^{m} \tilde{x}_{i,k}^0 \beta_k^0, \\
&= y_i - \beta_1^0\beta_2^0 - (\beta_3^0)^2 x_i + \beta_1^0\beta_2^0 + \beta_2^0\beta_1^0 + 2(\beta_3^0)^2 x_i \\
&= y_i + \beta_1^0\beta_2^0 + (\beta_3^0)^2 x_i,
\end{aligned}$$

und das linearisierte Modell lautet

$$\tilde{y}_i^0 = \beta_1 \tilde{x}_{i,1}^0 + \beta_2 \tilde{x}_{i,2}^0 + \beta_3 \tilde{x}_{i,3}^0 + \varepsilon_i.$$

(ii)

$$\begin{aligned}
\tilde{x}_{i,1}^0 &:= \left.\frac{\partial f}{\partial \beta_1}\right|_{\beta=\beta^0, \vec{x}=\vec{x}_i} = 2\beta_1^0 x_{i,1} = 2x_{i,1} \\
\tilde{x}_{i,2}^0 &:= \left.\frac{\partial f}{\partial \beta_2}\right|_{\beta=\beta^0, \vec{x}=\vec{x}_i} = 3(\beta_2^0)^2 x_{i,2} = 12x_{i,2} \\
\tilde{y}_i^0 &:= y_i - f(\vec{x}_i, \beta^0) + \sum_{k=1}^{m} \tilde{x}_{i,k}^0 \beta_k^0, \\
&= y_i - (\beta_1^0)^2 x_{i,1} - (\beta_2^0)^3 x_{i,2} + \beta_1^0 2x_{i,1} + \beta_2^0 12x_{i,2} \\
&= y_i + x_{i,1} + 16x_{i,2},
\end{aligned}$$

und das linearisierte Modell lautet

$$\tilde{y}_i^0 = \beta_1 \tilde{x}_{i,1}^0 + \beta_2 \tilde{x}_{i,2}^0 + \varepsilon_i.$$

(b) Anwenden des Logarithmus auf beiden Seiten ergibt:

$$\begin{aligned}\log(y_i) &= \log\left(\beta_1 x_{i,2}^{\beta_2} \cdot e^{\beta_3 x_{i,3}^2 + 2\beta_4 x_{i,4}} \cdot e^{\varepsilon_i}\right) \\ &= \log(\beta_1) + \beta_2 \log(x_{i,2}) + \beta_3 x_{i,3}^2 + 2\beta_4 x_{i,4} + \varepsilon_i \\ &= \tilde{\beta}_1 + \beta_2 \tilde{x}_{i,2} + \beta_3 \tilde{x}_{i,3} + \tilde{\beta}_4 x_{i,4} + \varepsilon_i,\end{aligned}$$

wobei $\tilde{\beta}_1 = \log(\beta_1)$, $\tilde{x}_{i,2} = \log(x_{i,2})$, $\tilde{x}_{i,3} = x_{i,3}^2$ und $\tilde{\beta}_4 = 2\beta_4$.

6. Tutorium am 19.01.09/22.01.09

(Verallgemeinerte lineare Modelle – Teil 2)

Aufgabe (verallgemeinerte lineare Modelle – vgl. Ökonometrie Blatt 5 Nr. 1 (WS 2008/09))

Auf der Homepage der Vorlesung befindet sich die Datei `swisslabor.dat`, die die Ergebnisse einer Befragung von 872 Haushalten über Gesundheit in der Schweiz enthält. Dabei wurden Daten über folgende Größen erhoben:

- Teilnahme: Ist die Person erwerbstätig?
- Alter: Alter in Jahrzehnten (Jahre geteilt durch 10)
- Bildung: Anzahl der Jahre der Berufsausbildung
- JKinder: Anzahl der Kleinkinder (unter 7 Jahre)
- AKinder: Anzahl der älteren Kinder (über 7 Jahre)
- Herkunft: Ist die Person Ausländer (also kein(e) Schweizer(in))?

Teilnahme	Alter	Bildung	JKinder	AKinder	Herkunft
no	3	8	1	1	no
yes	4.5	8	0	1	no
⋮	⋮	⋮	⋮	⋮	

(a) Betrachte das Logit-Modell und regressiere die binäre Variable „Teilnahme“ (Zielvariable) auf alle übrigen (erklärenden) Variablen. Interpretiere den Output der R-Funktion `summary()` nach Anwendung auf das Ergebnis von `glm()`.

(b) Erweitere das Logit-Modell aus Aufgabe (a), indem als weitere erklärende Variable das Quadrat des Alters mit aufgenommen wird und interpretiere wie in (a) den R-Output.

Hinweis:
Als erster Parameter in `glm()` muss hier
`Teilnahme~1+Alter+Bildung+JKinder+AKinder+Herkunft+I(Alter^2)`
angegeben werden.

(c) Welches der beiden Modelle aus (a) und (b) ist im Sinne des AIC-Kriteriums besser?

(d) Was bedeutet im R-Output `Number of Fisher Scoring iterations`?

(e) Teste mit Hilfe des Likelihood-Quotiententests aus der Vorlesung, ob die Regressionskoeffizienten β_3 (Bildung), β_4 (JKinder), β_5 (AKinder) und β_6 (Herkunft) im Logitmodell aus Aufgabe (a) gleich 0 sind (Nullhypothese: H_0 : „$\beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$“). Interpretiere das Testergebnis.

Vorgehensweise in R:
Durchführung einer Logit-Regression unter H_0:
`glm.swiss1 <- glm(Teilnahme~1+Alter,...)`
Durchführung einer Logit-Regression mit allen erklärenden Variablen:
`glm.swiss2 <- glm(Teilnahme~1+Alter+Bildung+JKinder+AKinder+Herkunft,...)`
Anwendung des Likelihood-Quotiententests:
`anova(glm.swiss1, glm.swiss2, test=''Chisq'')`
Der p-Wert steht in der Spalte `P(>|Chi|)`.

(f) Teste analog zu Teilaufgabe (e) die Hypothese H_0 : „$\beta_6 = 0$“ vs. H_1 : „$\beta_6 \neq 0$“ im Logitmodell und interpretiere das Ergebnis.

Aufgaben zu verallgemeinerten linearen Modellen:

(i) Modell aufstellen und ML-Schätzer für β mit R berechnen;

(ii) Verschiedene Modelle mit Hilfe des AIC vergleichen;

(iii) Hypothesentests der Form

$$H_0 : \text{„}H\beta = d\text{“ vs. } H_1 : \text{„}H\beta \neq d\text{“}.$$

ML-Schätzung von β

Wir betrachten ein verallgemeinertes lineares Modell

$$(g(\mathbb{E}[Y_1]), \ldots, g(\mathbb{E}[Y_n]))^T = X\beta,$$

wobei

- die Y_i unabhängig aber nicht notwendigerweise identisch verteilt sind;
- die (Zähl-) Dichte von Y_i die Gestalt

$$f_{\theta_i} = e^{\frac{1}{\tau^2}\left(y\theta_i + a(y,\tau^2) - b(\theta_i)\right)}$$

  hat (mit θ_i unbekannt);
- g die Linkfunktion ist, z. B. die natürliche Linkfunktion $g = (b')^{-1}$.

Wir wollen den *ML-Schätzer* $\hat{\beta}$ für β berechnen. Hieraus bekommt man mittels

$$\theta = (b')^{-1}\left(g^{-1}(x_i^T\beta)\right)$$

einen Schätzer für θ.

Schritt 1: *Log-Likelihood aufstellen*:

$$\begin{aligned} \log L(Y,\beta) &= \log \Pi_{i=1}^n f_{\theta_i}(Y_i) \\ &= \frac{1}{\tau^2}\sum_{i=1}^n (Y_i\theta_i + a(Y_i,\tau) - b(\theta_i)). \end{aligned}$$

Indem man $\theta = (b')^{-1}\left(g^{-1}(x_i^T\beta)\right)$ einsetzt bekommt man die Log-Likelihood als Funktion von Y und β.

Schritt 2: *numerische Maximierung der Log-Likelihood*:

1. Ableitungen: $U_i(\beta) = \frac{\partial \log L(Y,\beta)}{\partial \beta_i}, \quad i \in \{1,\ldots,m\}$.

2. Ableitungen: $W_{ij}(\beta) = \frac{\partial^2}{\partial\beta_i\partial\beta_j}\log L(Y,\beta), \quad i,j \in \{1,\ldots,m\}$.

Setzt man voraus, dass $W(\beta)$ negativ definit ist, so ergibt sich $\hat{\beta}$ als Lösung des nichtlineares Gleichungssystems $U(\beta) = 0$.

Numerische Berechnung von $\hat{\beta}$:

- Newton-Verfahren: $\hat{\beta}_{k+1} = \hat{\beta}_k - W^{-1}(\hat{\beta}_k)U(\hat{\beta}_k)$;
- Fisher-Scoring: $\hat{\beta}_{k+1} = \hat{\beta}_k + I^{-1}(\hat{\beta}_k)U(\hat{\beta}_k)$, wobei $I_{ij} = \mathbb{E}\left[U_i(\beta)U_j(\beta)\right]$ die Fisher-Informationsmatrix ist.

(Asymptotische) Tests für β

Wir testen (ähnlich wie im multivariaten linearen Modell) Hypothesen der Form

$$H_0 : \text{„}H\beta = d\text{“ vs. } H_1 : \text{„}H\beta \neq d\text{“}.$$

Dies geht mit dem Likelihood-Quotiententest wie folgt:

- sei $\tilde{\beta}$ der ML-Schätzer von β unter H_0, d. h.

$$\tilde{\beta} = \arg \max_{\beta \in \mathbb{R}^m : H\beta = d} \log L(Y, \beta)$$

- und sei $\hat{\beta}$ der ML-Schätzer unrestringiert, d. h.

$$\hat{\beta} = \arg \max_{\beta \in \mathbb{R}^m} \log L(Y, \beta),$$

dann gilt

$$\tilde{T}_n = -2 \log \left(\frac{L(Y, \tilde{\beta})}{L(Y, \hat{\beta})} \right)$$

ist asymptotisch χ^2_m-verteilt.

Akaikes Kriterium zur Modellwahl

Ziel: passendes Modell mit möglichst kleiner Anzahl an Parametern

Informationskoeffizient von Akaikes:

$$\text{AIC} = -2 \log L(Y, \hat{\beta}) + 2m$$

$\rightarrow$ je niedriger, desto besser ist Modell geeignet.

Zur Aufgabe:

(a) Es ist

$$\text{Teilnahme}_i = \begin{cases} \text{yes} & \text{falls } i\text{-ter Befragter erwerbstätig,} \\ \text{no} & \text{falls } i\text{-ter Befragter nicht erwerbstätig,} \end{cases}$$

also $Y_i \sim \text{Bernoulli}(p_i)$, wobei die Wahrscheinlichkeit p_i dass i-ter Befragter erwerbstätig ist, unbekannt ist. Wir nehmen an, dass p_i von den übrigen Variablen abhängig ist und wählen ein verallgemeinertes Regressionsmodell der Form

$$g(\mathbb{E}\left[Y_i\right]) = g(p_i) = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5} + \beta_6 x_{i,6},$$

wobei

$$\begin{aligned} x_{i,2} &= \text{Alter des } i\text{-ten Befragten;} \\ x_{i,3} &= \text{Bildung des } i\text{-ten Befragten;} \\ x_{i,4} &= \text{JKinder des } i\text{-ten Befragten;} \\ x_{i,5} &= \text{AKinder des } i\text{-ten Befragten;} \\ x_{i,6} &= \text{Herkunft des } i\text{-ten Befragten.} \end{aligned}$$

Als Linkfunktion wählen wir die natürliche Linkfunktion im Bernoulli-Modell:

$$g(y) = \log\left(\frac{y}{1-y}\right).$$

Mit R berechnen wir nun den ML-Schätzer $\hat{\beta} = \left(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5, \hat{\beta}_6\right)^T$:

```
## Einlesen des Datensatzes
> PFAD <- "D:/.../"
> daten <- read.table(file = paste(PFAD, "swisslabor.dat", collapse="", sep=""),
header = TRUE)
> attach(daten)

## (a) - Logitmodell
> glm.labor1 <- glm(Teilnahme ~ 1+Alter+Bildung+JKinder+AKinder+Herkunft,
family=binomial("logit"))
> summary(glm.labor1)

Call:
glm(formula = Teilnahme ~ 1 + Alter + Bildung + JKinder + AKinder +
    Herkunft, family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9110  -0.9783  -0.5891   1.1227   2.1989

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.085961   0.540535   3.859 0.000114 ***
Alter       -0.527066   0.089670  -5.878 4.16e-09 ***
Bildung     -0.001298   0.027502  -0.047 0.962371
JKinder     -1.326957   0.177893  -7.459 8.70e-14 ***
AKinder     -0.072517   0.071878  -1.009 0.313024
Herkunftyes  1.353614   0.198598   6.816 9.37e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1203.2  on 871  degrees of freedom
Residual deviance: 1069.9  on 866  degrees of freedom
AIC: 1081.9

Number of Fisher Scoring iterations: 4
```

Interpretation: die Variablen „Alter“, „JKinder“ und „Herkunft“ sind stark signifikant, d. h. die Hypothese H_0 : „$\beta_i = 0$“ vs. „$\beta_i \neq 0$“ wird sehr stark signifikant

verworfen. Dagegen sind die p-Werte von „Bildung“ und „AKinder“ sehr groß, was ein Indiz dafür ist, dass diese Variablen keinen Einfluss auf die Zielvariable „Teilnahme“ haben.

(b)
```
## (b) - erweitertes Logitmodell
> glm.labor2 <- glm(Teilnahme ~ 1+Alter+Bildung+JKinder+AKinder+Herkunft+I(Alter^2),
family=binomial("logit"))
> summary(glm.labor2)

Call:
glm(formula = Teilnahme ~ 1 + Alter + Bildung + JKinder + AKinder +
    Herkunft + I(Alter^2), family = binomial("logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8947  -0.9941  -0.5257   1.0759   2.1386

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.73259    1.28968  -2.894  0.00380 **
Alter        2.69498    0.66177   4.072 4.65e-05 ***
Bildung     -0.01032    0.02808  -0.368  0.71318
JKinder     -1.20891    0.17079  -7.078 1.46e-12 ***
AKinder     -0.26303    0.08284  -3.175  0.00150 **
Herkunftyes  1.25284    0.20115   6.228 4.71e-10 ***
I(Alter^2)  -0.39896    0.08202  -4.864 1.15e-06 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1203.2  on 871  degrees of freedom
Residual deviance: 1044.8  on 865  degrees of freedom
AIC: 1058.8

Number of Fisher Scoring iterations: 3
```

Unterschiede zu (a):

- „AKinder“ hat Einfluss auf die Zielvariable;
- Der Einfluss von „Alter“ wird im Logit-Modell unzureichend berücksichtigt, da „Alter^2“ ebenfalls signifikanten Einfluss auf „Teilnahme“ hat.

(c) $\mathrm{AIC}_{(a)} = 1081.9 > 1058.8 = \mathrm{AIC}_{(b)}$, damit ist Modell (b) besser.

(d) Die „Fisher-Scoring Iterations“ sind die Anzahl der Iterationsschritte zur Bestimmung von $\hat{\beta}$.

(e)
```
## (e) Test auf beta3 = beta4 = beta5 = beta6 = 0
> glm.swiss1 <- glm(Teilnahme ~ 1+Alter, family=binomial("logit"))
> glm.swiss2 <- glm(Teilnahme ~ 1+Alter+Bildung+JKinder+AKinder+Herkunft,
family=binomial("logit"))
> anova(glm.swiss1, glm.swiss2, test="Chisq")

Analysis of Deviance Table

Model 1: Teilnahme ~ 1 + Alter
Model 2: Teilnahme ~ 1 + Alter + Bildung + JKinder + AKinder + Herkunft
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       870    1195.77
2       866    1069.89   4   125.88 2.967e-26
```

Die Hypothese H_0 : „$\beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$“ wird verworfen, also hat mindestens eine der erklärenden Variablen „Bildung“, „JKinder“, „AKinder“ und „Herkunft“ Einfluss auf „Teilnahme“.

(f)
```
## (f) Test auf beta6 = 0
> glm.swiss1 <- glm(Teilnahme ~ 1+Alter+Bildung+JKinder+AKinder,
family=binomial("logit"))
> glm.swiss2 <- glm(Teilnahme ~ 1+Alter+Bildung+JKinder+AKinder+Herkunft,
family=binomial("logit"))
> anova(glm.swiss1, glm.swiss2, test="Chisq")

Analysis of Deviance Table

Model 1: Teilnahme ~ 1 + Alter + Bildung + JKinder + AKinder
Model 2: Teilnahme ~ 1 + Alter + Bildung + JKinder + AKinder + Herkunft
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       867    1120.06
2       866    1069.89   1    50.17 1.408e-12
```

Die Hypothese H_0 : „$\beta_6 = 0$“ wird verworfen, also hat „Herkunft“ Einfluss auf „Teilnahme“.

7. Tutorium am 26.01.09/29.01.09

(Verallgemeinerte lineare Modelle – Teil 3 ; Zeitreihen – Teil 1)

Aufgabe (verallgemeinerte lineare Modelle – vgl. Ökonometrie Blatt 6 Nr. 2 (WS 2008/09))

Auf der Homepage der Vorlesung befindet sich die Datei `apprentice.dat` mit Daten über die Anzahl der Lehrlinge, die für ihre Ausbildung von anderen schottischen Regionen nach Edinburgh gezogen sind. Es gibt insgesamt 33 Beobachtungen für 5 Variablen plus die Namen der Regionen in der (unbenannten) Spalte 1:

- Distance: Die Entfernung von Edinburgh (Meilen).
- Apprentices: Anzahl der Lehrlinge, die von der angegebenen Region nach Edinburgh gezogen sind.
- Population: Die Bevölkerung (in Tausend) der angegebenen Region.
- Urbanization: Der Grad der Urbanisierung, gemessen anhand des Prozentsatzes der Bevölkerung, der in städtischen Gebieten wohnt.
- Location: Die Himmelsrichtung der Region relativ zu Edinburgh (Norden, Westen, Süden).

	Distance	Apprentices	Population	Urbanization	Location
Midlothian	21	225	56	18.8	South
Westlothian	24	22	18	37.9	West
⋮	⋮	⋮	⋮	⋮	

(a) Die Zielvariable „Apprentices“ sei poissonverteilt. Stelle das verallgemeinerte lineare Regressionsmodell auf (berücksichtige alle Variablen) und bestimme die natürliche Linkfunktion. Interpretiere den Output der R-Funktion `summary()` nach der Anwendung auf das Ergebnis von `glm()`.

(b) Fasse die Gruppen „West“ und „North“ zu einer Gruppe „North“ zusammen und entscheide mit Hilfe des AIC-Ktiteriums, ob diese Abänderung im Datensatz zur Verbesserung des Modells beiträgt.

(c) Teste mit Hilfe des Likelihood-Quotiententests, ob mindestens eine der erklärenden Variablen im „neuen“ Datensatz von Teil (b) Einfluss auf die Anzahl der Lehrlinge hat und interpretiere das Testergebnis.

Aufgaben zu verallgemeinerten linearen Modellen:

(i) Modell aufstellen und ML-Schätzer für β mit R berechnen;

(ii) Verschiedene Modelle mit Hilfe des AIC vergleichen;

(iii) Hypothesentests der Form

$$H_0 : \text{„}H\beta = d\text{“ vs. } H_1 : \text{„}H\beta \neq d\text{“}.$$

Zur Aufgabe:

(a) Das verallgemeinerte lineare Regressionsmodell:

$$g\left(\mathbb{E}\left[Y_i\right]\right) = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5} + \beta_6 x_{i,6}$$

wobei $x_{i,5}$ und $x_{i,6}$ die Variablen „South“ ja/nein und „West“ ja/nein beschreiben.

Die Linkfunktion:

$$\mathbb{P}_\lambda(Y=k) = \frac{\lambda^k}{k!}e^{-\lambda} = e^{(k\log\lambda - \log(k!) - \lambda)}$$

$$\theta = \log\lambda, \qquad \tau = 1, \qquad a(k,\tau) = -\log(k!), \qquad b(\theta) = \lambda = e^\theta$$

$$g(y) = (b'(\theta))^{-1} = \log(y) \qquad \text{und} \qquad \mathbb{E}[k] = b'(\theta) = \lambda$$

Daraus folgt:

$$g(\mathbb{E}[Y_i]) = \log(\mathbb{E}[Y_i]) = \log(\lambda_i) = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5} + \beta_6 x_{i,6}$$

```
> ## Aufgabe
> ## (a)
> daten <- read.table("\\Dokumente und Einstellungen\\Michael Kochanski\\Desktop\\apprentice.dat",
header=TRUE)
> attach(daten)
> glm.move1 <- glm(Apprentices ~ 1 + Distance + Population + Urbanization + Location,
family=poisson(link="log"))
> summary(glm.move1)

Call:
glm(formula = Apprentices ~ 1 + Distance + Population + Urbanization + Location,
family = poisson(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8346  -1.4970  -0.1212   1.7419   6.9641

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.251678   0.247713  17.164  < 2e-16 ***
Distance      -0.033978   0.001931 -17.592  < 2e-16 ***
Population     0.021342   0.001523  14.014  < 2e-16 ***
Urbanization  -0.035819   0.004053  -8.837  < 2e-16 ***
LocationSouth  1.106500   0.150001   7.377 1.62e-13 ***
LocationWest   0.232365   0.183618   1.265    0.206
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
(Dispersion parameter for poisson family taken to be 1)
    Null deviance: 1350.44  on 32  degrees of freedom
Residual deviance:  256.31  on 27  degrees of freedom
AIC: 362.65

Number of Fisher Scoring iterations: 6
```

Interpretation: Die Himmelsrichtung West hat möglicherweise keinen Einfluss auf die Zielvariable.

(b)
```
> ## Aufgabe
> ## (b)
> detach(daten)
> daten[daten[,5]=="West",5] <- "North"
> attach(daten)
> glm.move2 <- glm(Apprentices ~ 1 + Distance + Population + Urbanization + Location,
family=poisson(link="log"))
> summary(glm.move2)

Call:
glm(formula = Apprentices ~ 1 + Distance + Population + Urbanization +
```

```
    Location, family = poisson(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.7607  -1.5980  -0.1649   2.3132   6.9181

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    4.251947   0.250934  16.944  < 2e-16 ***
Distance      -0.033857   0.001933 -17.519  < 2e-16 ***
Population     0.021234   0.001539  13.800  < 2e-16 ***
Urbanization  -0.033452   0.003596  -9.302  < 2e-16 ***
LocationSouth  1.052247   0.143286   7.344 2.08e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1350.44  on 32  degrees of freedom
Residual deviance:  257.89  on 28  degrees of freedom
AIC: 362.24

Number of Fisher Scoring iterations: 6
```

Interpretation: Die AIC-Werte unterscheiden sich kaum. Das Modell wurde nicht wesentlich verbessert.

(c)
```
> ## Aufgabe
> ## (c)
> glm.lqt <- glm(Apprentices ~ 1, family=poisson(link="log"))
> anova(glm.lqt, glm.move2, test="Chisq")
Analysis of Deviance Table

Model 1: Apprentices ~ 1
Model 2: Apprentices ~ 1 + Distance + Population + Urbanization + Location
  Resid. Df Resid. Dev Df Deviance  P(>|Chi|)
1        32    1350.44
2        28     257.89  4  1092.55 3.122e-235
```

Interpretation: Die Hypothese H_0 : „$\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$“ wird verworfen ($p$-Wert sehr klein), also hat mindestens eine der erklärenden Variablen Einfluss auf die Zielvariable.

Exkurs: Zeitreihen

Annahme: $(\delta_t)_{t\in\mathbb{N}}$ sei eine Folge unkorrelierter Zufallsvariablen mit konstanter Varianz und $\mathbb{E}[\delta_i] = 0$.

(a) **Autoregressiver Prozess AR(1):** Der aktuelle Zustand hängt vom letzten Zustand ab.

$$\begin{aligned} Z_t &= \rho Z_{t-1} + \delta_t \\ Z_t &= \delta_t + \sum_{i=1}^{\infty} \rho^i \delta_{t-i} \\ \mathbb{E}[Z_t] &= 0 \end{aligned}$$

(b) **Autoregressive Prozesse AR(m):** Der aktuelle Zustand hängt von früheren Zuständen ab.

$$Z_t = \delta_t + \sum_{i=1}^{m} \rho_i Z_{t-i}$$

(c) **Moving Average Prozess MA(n):** Der aktuelle Zustand hängt vom „gleitenden Mittel“ ab.

$$Z_t = \delta_t + \sum_{i=1}^{n} \rho_i \delta_{t-i}$$

(d) **Autoregressiver Moving Average Prozess ARMA(m, n):**

$$Z_t = \delta_t + \sum_{i=1}^{m} \rho_i Z_{t-i} + \sum_{i=1}^{n} \rho_i' \delta_{t-i}$$

(e) **Integrierter Autoregressiver Moving Average Prozess ARIMA(m, n):**
Sei $\Delta Z_t = Z_t - Z_{t-1}$ ein stationärer ARMA(m, n)-Prozess, dann ist $(Z_t)_{t\in\mathbb{N}}$ ein ARIMA(m, n)-Prozess.

(f) **Linearer Prozess:**
Ein linearer Prozess ist ein MA(∞)-Prozess mit $\sum_{i\in\mathbb{Z}} \rho_i^2 < \infty$.

(g) **Stationarität:** Der Zusammenhang zweier Zustände ist bei gleichbleibendem Abstand unabhängig von ihrem Zeitpunkt.

$$\begin{aligned} \mathbb{E}\left[X_t^2\right] < \infty \qquad \forall t, \qquad \mathbb{E}[X_t] = \text{const.} \\ \operatorname{Cov}(X_s, X_t) = \operatorname{Cov}(X_{s+r}, X_{t+r}) \qquad \forall s, t, r \end{aligned}$$

Folgende Prozesse sind stationär: MA(n) und lineare Prozesse. Kann ein AR(m)- oder ARMA(m, n)-Prozess als ein linearer Prozess dargestellt werden, so ist dieser ebenfalls stationär (Es gilt dann z. B. für einen AR(1)-Prozess: $|\rho_i| < 1$).

8. Tutorium am 02.02.09/05.02.09

(Zeitreihen – Teil 2)

Aufgabe (Zeitreihen – vgl. Ökonometrie Blatt 7 Nr. 3 (WS 2008/09))

Auf der Homepage der Vorlesung befindet sich die Datei `flug.dat` mit Daten über die durchschnittlichen monatlichen Passagierzahlen pro Flugzeug einer Fluggesellschaft zwischen Januar 1990 und Dezember 1993. Analysiere den Datensatz mit $S_t = \sin\left(2\pi t - \frac{\pi}{2}\right)$ und $T_t = \ln(t - 1989)$.

Aufgaben zu Zeitreihen:

(i) Interpretation der Zeitreihe bezüglich der Komponenten.

(ii) Aufstellen des Modells (Plausibilitätsprüfung).

(iii) Schätzung der Modellparameter (lineare Regression).

(iv) Schätzung der (Auto-)Korrelationsfunktion.

(v) Wahl eines geeigneten Prozesses für den stationären Anteil der Zeitreihe.

Das Zeitreihen-Modell

Das allgemeine Zeitreihen-Modell lässt sich in folgender Form darstellen:

$$Z_t = T_t + S_t + X_t.$$

Dabei bezeichnet T_t den Trend, S_t den saisonalen Anteil und X_t den stationären Anteil. Der zeitunabhängige Teil von T_t wird als Drift bezeichnet und oft als eigenständiger Teil der Zeitreihe dargestellt.

Schätzung der (Auto-)Korrelationsfunktion

Die (Auto-)Kovarianzfunktion $R(s)$ und die Autokorrelationsfunktion $B(s)$ sind für einen stationären Prozess X_t wie folgt definiert:

$$\begin{aligned} R(s) &= \mathbb{E}\left[(X_s - \mathbb{E}[X_s])(X_0 - \mathbb{E}[X_0])\right] \qquad \text{mit} \qquad s \in \mathbb{Z} \\ &= \mathbb{E}[X_s X_0] \qquad \text{da} \qquad \mathbb{E}[X_t] = 0 \qquad \forall t \end{aligned}$$

$$\text{und} \qquad B(s) = \frac{R(s)}{R(0)} \qquad \text{mit} \qquad s \in \mathbb{Z}.$$

Bei beiden Funktionen ist das Ergebnis unabhängig von t. Die Kovarianz bzw. die Korrelation zweier Zustände eines stationären Prozesses hängt also nur vom Abstand der Zustände und nicht vom Zeitpunkt ab.

Die (Auto-)Kovarianzfunktion $R(s)$ und die Autokorrelationsfunktion $B(s)$ lassen sich durch folgende (nicht erwartungstreue) Schätzer schätzen:

$$\widehat{R}(s) = \frac{1}{n}\sum_{t=1}^{n-s} X_t X_{t+s} \qquad \text{mit} \qquad s \in \{0, 1, \ldots, n-1\}$$

$$\widehat{B}(s) = \frac{\widehat{R}(s)}{\widehat{R}(0)} \qquad \text{mit} \qquad s \in \{0, 1, \ldots, n-1\}.$$

Zur Aufgabe:

(a) *Trendanteil*: zunächst stark ansteigende Werte, dann Anstieg schwächer.
saisonaler Anteil: Höchststände jeweils zur Jahresmitte, Tiefststände jeweils zum Ende des Jahres.

```
> daten <- read.table("\\Dokumente und Einstellungen\\Michael Kochanski\\Desktop\\flug.dat",
header=TRUE)
> flug <- ts(daten, start=1990, frequency=12)
> plot(flug)
```

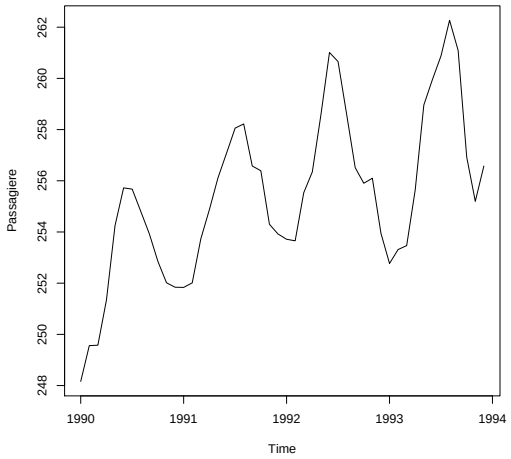


Abbildung 2: Zeitreihe – Flug

(b) Das Modell:

$$\begin{aligned} Z_t &= T_t + S_t + X_t \\ T_t &= \beta_1 + \beta_2 \ln(t - 1989) \\ S_t &= \beta_3 \sin\left(2\pi t - \frac{\pi}{2}\right) \end{aligned}$$

Der Trend wird durch die ln-Funktion modelliert, der saisonale Anteil durch eine Sinusfunktion (mit passender Periode).

(c) Schätzung der Koeffizienten β_1, β_2 und β_3 mit der bekannten R-Funktion `lm()`.

```
> t <- time(flug)
> lm.flug <- lm(flug ~ 1 + log(t-1989) + sin(2*pi*t-pi/2))
> summary(lm.flug)

Call: lm(formula = flug ~ 1 + log(t - 1989) + sin(2 * pi * t -
pi/2))

Residuals:
    Min      1Q  Median      3Q     Max
-2.2574 -0.7150  0.1162  0.7333  1.8988
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           251.1660     0.3444  729.20   <2e-16 ***
log(t - 1989)           4.2818     0.3162   13.54   <2e-16 ***
sin(2 * pi * t - pi/2)  3.1022     0.1988   15.61   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.9714 on 45 degrees of freedom Multiple
R-squared: 0.9106,     Adjusted R-squared: 0.9066 F-statistic: 229.2
on 2 and 45 DF,  p-value: < 2.2e-16
```

Interpretation: Alle p-Werte (sowohl für $H_0 : \beta_i = 0$ als auch für $H_0 : \beta_2 = \beta_3 = 0$) weisen auf einen starken Einfluss der Trend- und saisonaler-Anteil-Funktionen. Zudem bestätigt das nahe bei 1 liegende Bestimmtheitsmaß R^2 eine gute Modellwahl.

Das Modell lässt sich nun wie folgt schreiben:

$$\begin{aligned} Z_t &= \beta_1 + \beta_2 \ln(t - 1989) + \beta_3 \sin\left(2\pi t - \frac{\pi}{2}\right) + X_t \\ Z_t &= 251,166 + 4,2818 \ln(t - 1989) + 3,1022 \sin\left(2\pi t - \frac{\pi}{2}\right) + X_t \\ \mathbb{E}[Z_t] &= 251,166 + 4,2818 \ln(t - 1989) + 3,1022 \sin\left(2\pi t - \frac{\pi}{2}\right) \end{aligned}$$

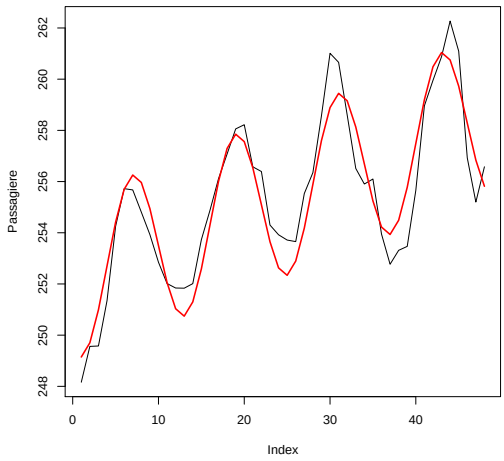


Abbildung 3: Gefittete Zeitreihe

(d) Schätzung der (Auto-)Korrelationsfunktion (mit der R-Funktion `acf`):

```
> X <- lm.flug$residuals
> acf(X)
```

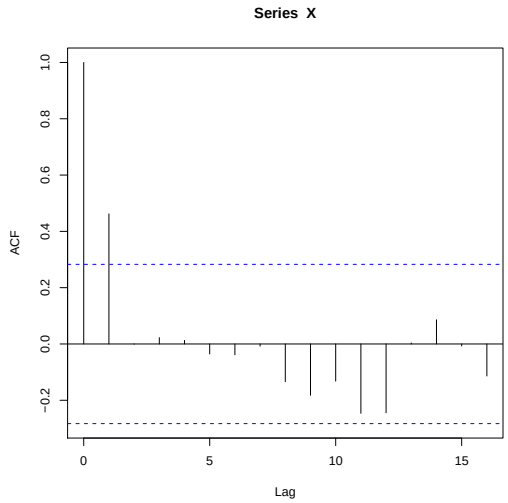


Abbildung 4: Lag

Interpretation: Der Wert s, der Wert also, der den Abstand zweier Zustände charakterisiert, wird als „Lag“ bezeichnet. Der Grafik ist zu entnehmen, dass die (Auto-) Korrelationsfunktion für die Werte $s = 0, 1$ signifikant von Null verschieden ist (da die Balken nur zu diesen Werten ausserhalb der gestrichelten Geraden liegen).

(e) Zunächst ist zu klären, warum X_t ein MA(q)-Prozess ist: Es ist bekannt, dass ein stationärer Prozess sich in Form eines linearen Prozesses darstellen lässt. Ein linearer Prozess lässt sich als MA(q)-Prozess auffassen, bei dem auch $q = \infty$ zugelassen ist. X_t ist stationär.

Die Aussage von Aufgabe 1 auf dem Übungsblatt Nr. 7 lautet: Die (Auto-) Korrelationsfunktion eines MA(q)-Prozesses ist Null für $q < |s|$ und von Null verschieden für $q \geq |s|$. Aus (d) ist bekannt, dass die (Auto-)Korrelationsfunktion für alle $s \geq 2$ Null ist, demnach sollte X_t ein MA(1)-Prozess sein.

9. Tutorium am 09.02.09/12.02.09

(Einige abschließende Bemerkungen)

Die Autokorrelationsfunktion für lineare Prozesse

Wenn X_t ein linearer Prozess der Form

$$X_t = \sum_{i=0}^{\infty} \gamma_i \delta_{t-i}$$

ist, so gilt nach Satz 3.2.2 aus der Vorlesung für die Autokovarianz- bzw. Autokorrelationsfunktion:

$$\begin{aligned} R(s) &= \mathrm{Cov}(X_s, X_0) = \sigma^2 \sum_{i=0}^{\infty} \gamma_i \gamma_{i+|s|} \\ B(s) &= \frac{R(s)}{R(0)} = \frac{\sigma^2 \sum_{i=0}^{\infty} \gamma_i \gamma_{i+|s|}}{\sigma^2 \sum_{i=0}^{\infty} \gamma_i^2} = \frac{\sum_{i=0}^{\infty} \gamma_i \gamma_{i+|s|}}{\sum_{i=0}^{\infty} \gamma_i^2}. \end{aligned}$$

Insbesondere gilt:

- für einen MA(q)-Prozess (vgl. Blatt 7/ Aufgabe 1) der Form

  $$X_t = \sum_{i=1}^{q} \alpha_i \delta_{t-i},$$

  dass

  $$R(s) = \begin{cases} \sigma^2 \sum_{i=0}^{q-|s|} \alpha_i \alpha_{i+|s|}, & \text{falls } |s| \le q, \\ 0, & \text{falls } |s| > q \end{cases}$$

  und

  $$B(s) = \begin{cases} \sum_{i=0}^{q-|s|} \alpha_i \alpha_{i+|s|} / \sum_{j=0}^{q} \alpha_j^2, & \text{falls } |s| \le q, \\ 0, & \text{falls } |s| > q. \end{cases}$$

  Die acf eines MA(q)-Prozesses ist also Null für Lags größer als q.

- für einen AR(1)-Prozess (vgl. Blatt 8/ Aufgabe 2) der Form

  $$X_t = \rho X_{t-1} + \delta_t = \sum_{i=0}^{\infty} \rho^i \delta_{t-i},$$

  der stationär ist falls $\rho < 1$, dass:

  $$R(s) = \sigma^2 \sum_{i=0}^{\infty} \rho^i \rho^{i+|s|} = \sigma^2 \rho^{|s|} \sum_{i=0}^{\infty} \rho^{2i},$$

  und

  $$B(s) = \frac{R(s)}{R(0)} = \rho^{|s|}.$$

  Die acf eines AR(1)-Prozesses klingt also exponentiell ab.

Kriterien zur Modellwahl: verallgemeinerte lineare Modelle vs. lineare Modelle

- Für verallgemeinerte lineare Modelle verwenden wir das Kriterium von Akaike. Das Modell mit dem kleineren AIC ist besser geeignet, wobei

$$\mathrm{AIC} = -2\log L(Y, \hat{\beta}) + 2m,$$

  mit $m =$ Anzahl an Parametern.

- Für lineare Modelle verwenden wir den Wert des „adjusted R^2“. Der Wert R^2 gibt an, wie viel Prozent der Varianz aus den Daten mit Hilfe des linearen Modells erklärt werden kann (er liegt zwischen $0 = 0\%$ und $1 = 100\%$). Beim adjusted R^2 wird ähnlich wie beim AIC die Anzahl an Parametern berücksichtigt (es ergibt sich wieder ein Wert zwischen 0 und 1).

Zum Hypothesentest im multivariaten Regressionsmodell

Testet man im multivariaten Regressionsmodell Hypothesen der Form

$$H_0: \text{ „}\beta_i = 0\text{“ vs. } H_1: \text{ „}\beta_i \neq 0\text{“}$$

so muss bei der Berechnung der Teststatistik lediglich eine 1×1-Matrix (also eine Zahl) invertiert werden. Vergleiche zum Beispiel Aufgabe 2 von Blatt 2: teste

$$H_0: \text{ „}\beta_3 = 0\text{“ vs. } H_1: \text{ „}\beta_3 \neq 0\text{“}.$$

Gegeben:

$$\left(X^T X\right)^{-1} = \begin{pmatrix} 0.993083 & 0.234087 & -0.001265 \\ 0.234087 & 0.144294 & -0.000646 \\ -0.001265 & -0.000646 & 0.000003 \end{pmatrix}$$

$$X^T y = \begin{pmatrix} 156775 \\ 1820952 \\ 461629875 \end{pmatrix}, \qquad S^2 = 0.32 \cdot 10^7,$$

$$\alpha = 0.05 \qquad \text{und} \qquad F_{1,6,0.95} = 5.99.$$

Benötigte Formeln:

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T y \\ T_H &= \frac{(H\hat{\beta} - d)^T (H(X^T X)^{-1} H^T)^{-1} (H\hat{\beta} - d)}{sS^2} \sim F_{s,n-m}. \end{aligned}$$

Lösung:

$$H = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}, \qquad s = 1$$

$$\begin{aligned} \hat{\beta} = \left(X^T X\right)^{-1} X^T y &= \begin{pmatrix} 0.993083 & 0.234087 & -0.001265 \\ 0.234087 & 0.144294 & -0.000646 \\ -0.001265 & -0.000646 & 0.000003 \end{pmatrix} \begin{pmatrix} 156775 \\ 1820952 \\ 461629875 \end{pmatrix} \\ &= \begin{pmatrix} -2010.0137 \\ 1238.5381 \\ 10.2343 \end{pmatrix} \end{aligned}$$

$$H\hat{\beta} - d \;=\; \begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -2010.0137 \\ 1238.5381 \\ 10.2343 \end{pmatrix} = 10.2343$$

$$\begin{aligned} (H(X^TX)^{-1}H^T)^{-1} &= \left(\begin{pmatrix} 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0.993083 & 0.234087 & -0.001265 \\ 0.234087 & 0.144294 & -0.000646 \\ -0.001265 & -0.000646 & 0.000003 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right)^{-1} \\ &= \left(\begin{pmatrix} -0.001265 & -0.000646 & 0.000003 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right)^{-1} = \frac{1}{0.000003}. \end{aligned}$$

Also ergibt sich die Teststatistik zu

$$T_H \;=\; \frac{10.2343^2}{0.000003 \cdot 0.32 \cdot 10^7} = 10.9105.$$

Achtung: Rundungsfehler sehr groß (exakter Wert: $T_H = 1.874717$)!