Ausgewählte Aufgaben aus Statistik III für Nebenfachstudierende zur Vorbereitung auf die Eingnungsprüfung für Master-Quereinsteiger
Stand: 08.08.2014

Vorbemerkungen

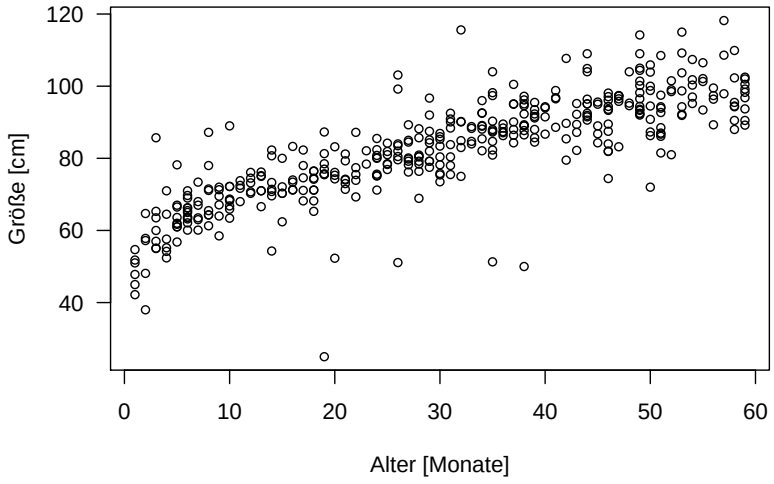
- Diese Sammlung umfasst ausgewählte Aufgaben aus der Veranstaltung *Statistik III für Nebenfachstudierende*, die besonders gut zur Vorbereitung auf den Eignungsfeststellungstest für Master-Quereinsteiger geeignet sind.
- Es wird keine Garantie für die Vollständigkeit der Aufgaben übernommen.
- Für weitergehendes Material schauen Sie bitte direkt auf der Veranstaltungshomepage von *Statistik III für Nebenfachstudierende* nach.
- Falls Sie Fehler oder Inkonsistenzen feststellen, schreiben Sie bitte eine E-Mail an die Person, die auf der Homepage der Fragestunde Mathematik/Stochastik/Statistik gelistet ist.

Inhaltsverzeichnis

Kapitel 0: Regressionsmodelle: Erinnerung und Ausblick

Aufgabe 1:

Gegeben sei eine Stichprobe von 400 Kindern aus Indien aus den Jahren 2005/06. Das folgende Streudiagramm zeigt eine erste deskriptive Analyse des Zusammenhangs zwischen Alter (`cage`) und Körpergröße (`cheight`):



Es wird eine lineare Einfachregression in R gerechnet, aus der sich folgender Output ergibt:

```
Call:
lm(formula = cheight ~ cage, data = india)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 61.05763    0.83147   73.43   <2e-16 ***
cage         0.70859    0.02418   29.30   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.139 on 398 degrees of freedom
Multiple R-squared: 0.6833, Adjusted R-squared: 0.6825
F-statistic: 858.7 on 1 and 398 DF,  p-value: < 2.2e-16
```

(a) Interpretieren Sie die geschätzten Regressionskoeffizienten.

(b) Zeichnen Sie die geschätzte Regressionsgerade in das obige Streudiagramm ein.

(c) Berechnen Sie ein 99%-Konfidenzintervall für β_1.

(d) Überprüfen Sie mit einem geeigneten Test zum Signifikanzniveau von $\alpha = 0.05$, ob es einen signifikanten Zusammenhang zwischen dem Alter und der Körpergröße gibt.

Aufgabe 2:

Gegeben sei das Standardmodell der linearen Einfachregression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{für} \quad i = 1, \ldots, n$$

mit folgenden Annahmen an die Fehlerterme:

$$\mathrm{E}(\varepsilon_i) = 0 \ , \quad \mathrm{Var}(\varepsilon_i) = \sigma^2 \quad \text{und} \quad \varepsilon_i \perp\!\!\!\perp \varepsilon_j \quad \text{für alle } i \neq j$$

(a) Zeigen Sie, dass sich die Kleinste-Quadrate-Schätzer für β_0 und β_1 wie folgt ergeben:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \text{und} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} \ .$$

(b) Zeigen Sie, dass $\hat{\beta}_0$ und $\hat{\beta}_1$ erwartungstreue Schätzer sind.

Kapitel 1: Mehrdimensionale Zufallsvariablen

1.1 Zufallsvektoren, Erwartungswertvektor und Kovarianzmatrix

Aufgabe 3:

Gegeben seien die Zufallsvariablen Y und Z mit folgender gemeinsamer Dichtefunktion:

$$f(y,z) = \begin{cases} c \cdot (y + 2z) & \text{falls } 0 \leq y \leq 2 \ , \ 0 \leq z \leq 1 \\ 0 & \text{sonst} \end{cases}$$

(a) Wie groß muss c sein, damit $f(y,z)$ tatsächlich eine Dichte ist?

(b) Bestimmen Sie die Randdichten von Y und Z.

(c) Sind Y und Z unabhängig?

(d) Bestimmen Sie die Randverteilungsfunktion von Y.

(e) Bestimmen Sie die bedingte Dichte von Y gegeben $Z = z_0$.

(f) Fassen Sie $(Y,Z)^\top$ als Zufallsvektor auf und berechnen Sie den Erwartungswertvektor.

(g) Berechnen Sie die Kovarianzmatrix von $(Y,Z)^\top$.

(h) Berechnen Sie die Korrelationsmatrix von $(Y,Z)^\top$.

Aufgabe 4:

Bei welchen der folgenden Matrizen könnte es sich um Kovarianzmatrizen handeln? Begründen Sie Ihre Antwort!

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.2 & 0.5 \\ 0.2 & 0.3 \\ 0.5 & 0.3 \end{pmatrix} \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.5 & 0.7 & 0.9 \\ 0.3 & 0.9 & 0.3 \\ 0.9 & 0.7 & 0.5 \end{pmatrix}$$

$$\boldsymbol{\Sigma}_3 = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} \qquad \boldsymbol{\Sigma}_4 = \begin{pmatrix} 0.5 & 0.7 & -0.9 \\ 0.7 & 0.9 & 0.3 \\ -0.9 & 0.3 & -0.5 \end{pmatrix}$$

1.2 Multivariate Normalverteilung

Aufgabe 5:

Sei $\boldsymbol{X} = (X_1, X_2, X_3, X_4)^\top$ multivariat normalverteilt mit

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \\ 1 \\ -1 \end{pmatrix} \quad \text{und} \quad \boldsymbol{\Sigma} = \begin{pmatrix} 4 & 2 & 1 & -1 \\ 2 & 4 & 3 & 1 \\ 1 & 3 & 4 & 2 \\ -1 & 1 & 2 & 4 \end{pmatrix} .$$

(a) Interpretieren Sie die Elemente σ_{11}, σ_{12} und σ_{41} von $\boldsymbol{\Sigma}$.

(b) Berechnen Sie die Korrelationsmatrix $\boldsymbol{R}$ und interpretieren Sie die Elemente ρ_{12} und ρ_{41}.

(c) Bestimmen Sie die Randverteilung von $\boldsymbol{Y} = (X_1, X_3)^\top$.

(d) Wie lautet die bedingte Verteilung von $\boldsymbol{Y}|\boldsymbol{Z}$ mit $\boldsymbol{Z} = (X_2, X_4)^\top$?

(e) Berechnen Sie Erwartungswert und Kovarianzmatrix der Zufallsvariablen $\boldsymbol{U} = \boldsymbol{A} \cdot \boldsymbol{X}$ mit

$$\boldsymbol{A} = \begin{pmatrix} 2 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} .$$

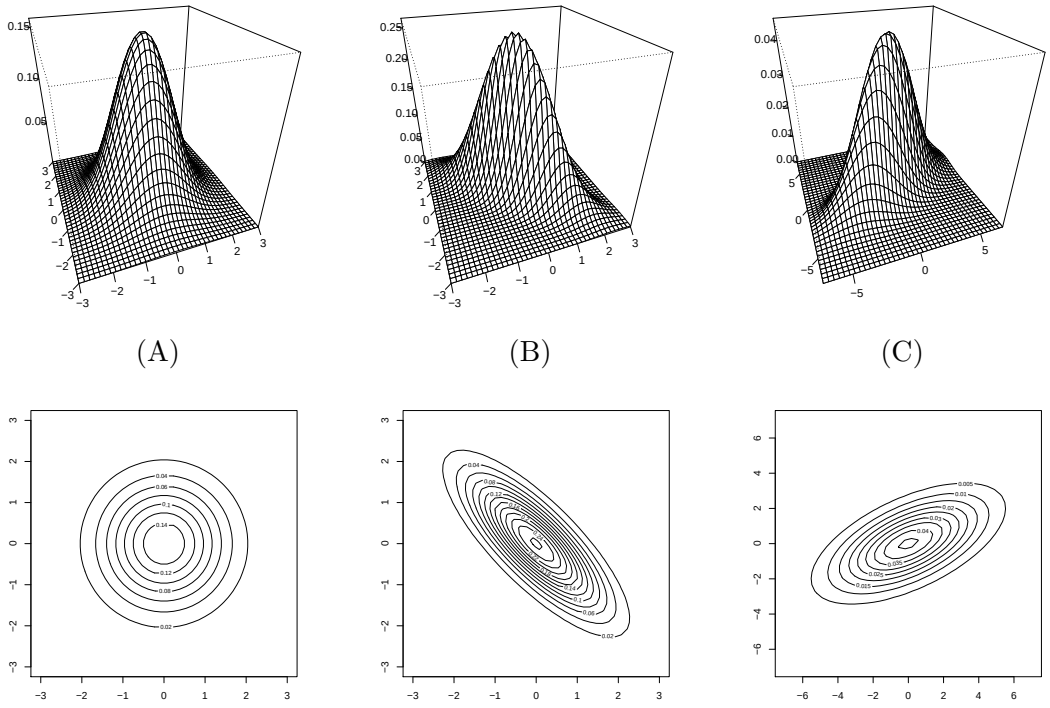
Aufgabe 6:

Gegeben seien zwei bivariat normalverteilte Zufallsvariablen

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X) \qquad \text{mit} \qquad \boldsymbol{\mu}_X = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma}_X = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$$

und $$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N(\boldsymbol{\mu}_Y, \boldsymbol{\Sigma}_Y) \qquad \text{mit} \qquad \boldsymbol{\mu}_Y = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \boldsymbol{\Sigma}_Y = \begin{pmatrix} 6.85 & 2.5 \\ 2.5 & 2.65 \end{pmatrix}.$$

Die folgenden Abbildungen zeigen Dichten und zugehörige Konturplots der Zufallsvariablen $\boldsymbol{X}$ und $\boldsymbol{Y}$ sowie von einer dritten zweidimensionalen Zufallsvariablen $\boldsymbol{Z}$:



(A) (B) (C)

Welche der Abbildungen (A), (B) oder (C) zeigt die Dichte von $\boldsymbol{X}$?
Welche der Abbildungen (A), (B) oder (C) zeigt die Dichte von $\boldsymbol{Y}$?
Begründen Sie Ihre Antworten.

Kapitel 2: Likelihood-Inferenz

Aufgabe 7:

Gegeben seien Realisationen $y_1, \ldots, y_n$ von unabhängig und identisch exponential-verteilten Zufallsvariablen $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{Exp}(\lambda)$, wobei $\lambda > 0$ unbekannt ist. Die Dichte einer exponentialverteilten Zufallsvariablen lautet:

$$f(y) = \lambda \exp(-\lambda y) \ .$$

(a) Bestimmen Sie die Likelihood $L(\lambda)$ sowie die Log-Likelihood $l(\lambda)$ für die Stichprobe.

(b) Bestimmen Sie die Scorefunktion $s(\lambda)$ und berechnen Sie den ML-Schätzer $\hat{\lambda}_{ML}$ für den unbekannten Parameter λ.

(c) Berechnen Sie Likelihood, Log-Likelihood und ML-Schätzer für die folgende konkrete Stichprobe der Größe $n = 5$:

$$y_1 = 0.21 \quad y_2 = 0.42 \quad y_3 = 0.01 \quad y_4 = 1.60 \quad y_5 = 0.20$$

(d) Bestimmen Sie die beobachtete Fisher-Information $F_{obs}(\hat{\lambda}_{ML})$ sowie die erwartete Fisher-Information $F(\hat{\lambda}_{ML})$ und vergleichen Sie diese miteinander.

Aufgabe 8:

Gegeben seien Realisationen $y_1, \ldots, y_n$ von unabhängig und identisch Bernoulli-verteilten Zufallsvariablen $Y_1, \ldots, Y_n \overset{iid}{\sim} \text{B}(\pi)$, wobei $0 < \pi < 1$ unbekannt ist. Die Wahrscheinlichkeitsfunktion einer Bernoulli-verteilten Zufallsvariablen lautet:

$$P(Y = y) = \pi^y (1 - \pi)^{1-y} \ .$$

(a) Bestimmen Sie die Likelihood $L(\pi)$ sowie die Log-Likelihood $l(\pi)$ für die Stichprobe.

(b) Bestimmen Sie die Scorefunktion $s(\pi)$ und berechnen Sie den ML-Schätzer $\hat{\pi}_{ML}$ für den unbekannten Parameter π.

(c) Bestimmen Sie die beobachtete Fisher-Information $F_{obs}(\hat{\pi}_{ML})$ sowie die erwartete Fisher-Information $F(\hat{\pi}_{ML})$ und vergleichen Sie diese miteinander.

Aufgabe 9:

Gegeben sei eine Stichprobe unabhängiger Beobachtungen $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ vom Umfang n, wobei x eine Kovariable und y eine normalverteilte Zielvariable bezeichnet. Der Zusammenhang zwischen x und y wird durch das folgende Modell der linearen Einfachregression beschrieben:

$$y_i \overset{unabh.}{\sim} N(\eta_i, \sigma^2) \qquad \text{mit} \qquad \eta_i = \beta_0 + \beta_1 x_i$$

wobei $\beta_0 \in \mathbb{R}$ und $\beta_1 \in \mathbb{R}$ unbekannte Parameter sind. Der Varianzparameter σ^2 wird als bekannt angenommen.

(a) Bestimmen Sie zunächst Likelihood-Beitrag $L_i(\beta_0, \beta_1)$ und Log–Likelihood-Beitrag $l_i(\beta_0, \beta_1)$ für die i-te Beobachtung der Stichprobe.
Berechnen Sie dann die Likelihood $L(\beta_0, \beta_1)$ und die Log–Likelihood $l(\beta_0, \beta_1)$ für die gesamte Stichprobe.

(b) Bestimmen Sie die Scorefunktion $s(\beta_0, \beta_1)$ der Stichprobe.

(c) Bestimmen Sie die ML-Schätzer für die unbekannten Parameter β_0 und β_1 und vergleichen Sie diese mit den KQ-Schätzern aus Aufgabe 2

(d) Bestimmen Sie die beobachtete Fisher-Informationsmatrix $\boldsymbol{F}_{obs}(\beta_0, \beta_1)$.

(e) Bestimmen Sie die erwartete Fisher-Informationsmatrix $\boldsymbol{F}(\beta_0, \beta_1)$.

(f) Erklären Sie anhand einer Skizze für den Fall, dass genau ein Parameter θ unbekannt ist, wie sich ein höherer Stichprobenumfang auf die Varianz des ML-Schätzers $\hat{\theta}_{ML}$ auswirkt. Inwiefern spielen in diesem Zusammenhang die Fisher-Informationsmatrizen eine Rolle?

Kapitel 3: Lineare Regressionsmodelle

3.1 Grundbegriffe

Aufgabe 10:

Zur Untersuchung des Zusammenhangs zwischen Einkommen und Attraktivität wurde der Datensatz vom Allbus 2010 (siehe `http://www.gesis.org/allbus/`) aufbereitet, sodass er die folgenden Variablen von 2290 befragten Personen enthält:

Variable	Beschreibung
`einkommen`	Nettoeinkommen in Euro
`logeinkommen`	Logarithmiertes Nettoeinkommen: log(`einkommen`)
`alter`	Alter in Jahren
`geschlecht`	Geschlecht (0 = Mann, 1 = Frau)
`attrakt`	Geschätzte Attraktivität des Befragten auf einer Skala von 1 (unattraktiv) bis 11 (attraktiv)

Zur Modellierung des Zusammenhangs zwischen dem Nettoeinkommen und den Kovariablen wird ein lineares Regressionsmodell gerechnet, es ergibt sich folgender R–Output:

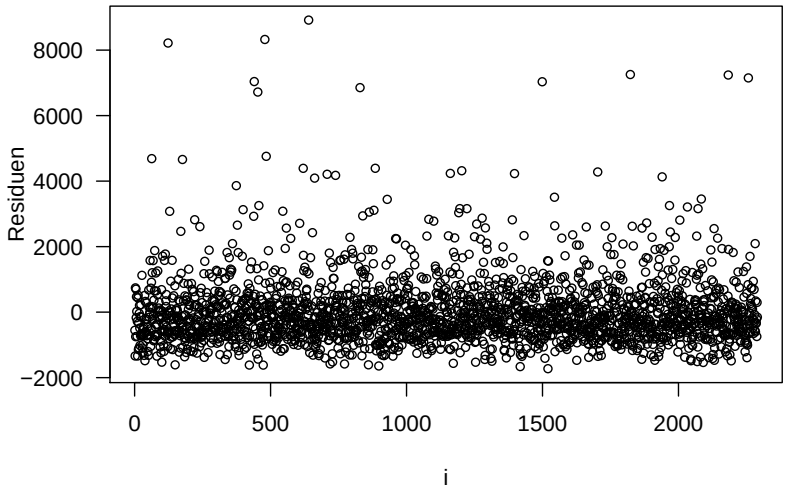
```
Call:
lm(formula = einkommen ~ alter + geschlecht + attrakt, data = allbus)

Residuals:
    Min      1Q  Median      3Q     Max
-1726.0  -614.3  -187.8   337.0  8916.7

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     956.162    122.260   7.821 7.96e-15 ***
alter             6.102      1.282   4.760 2.05e-06 ***
geschlechtFrau -701.024     42.841 -16.363  < 2e-16 ***
attrakt          68.431     11.108   6.161 8.53e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1023 on 2286 degrees of freedom
Multiple R-squared: 0.1166,Adjusted R-squared: 0.1154
F-statistic: 100.5 on 3 and 2286 DF,  p-value: < 2.2e-16
```

(a) Wie lautet die geschätzte Modellgleichung für dieses Regressionsmodell?

(b) Interpretieren Sie den geschätzten Intercept sowie die geschätzten Parameter für `alter` und `attrakt`.

(c) Prognostizieren Sie das erwartete Nettoeinkommen für eine 28-jährige Frau, die der Interviewer in Attraktivitätsstufe 7 einordnet.

(d) Die folgende Abbildung zeigt die Residuen des obigen Modells:



Welche Annahme des linearen Regressionsmodells scheint Ihnen hier verletzt zu sein? Begründen Sie Ihre Antwort.

(e) Es wird ein Modell mit logarithmiertem Nettoeinkommen als Zielvariable geschätzt, für das sich der folgende (reduzierte) Modelloutput ergibt:

```
Call:
lm(formula = logeinkommen ~ alter + geschlecht + attrakt, data = allbus)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.6534536  0.0793660  83.832  < 2e-16 ***
alter           0.0052475  0.0008322   6.306 3.43e-10 ***
geschlechtFrau -0.4943155  0.0278107 -17.774  < 2e-16 ***
attrakt         0.0471000  0.0072105   6.532 7.97e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wie ändert sich die Interpretation der Parameterschätzer in diesem Modell?

(f) Was erscheint Ihnen am verwendeten Kovariablen-Design ungünstig? Wie könnte man das Kovariablen-Design verbessern?

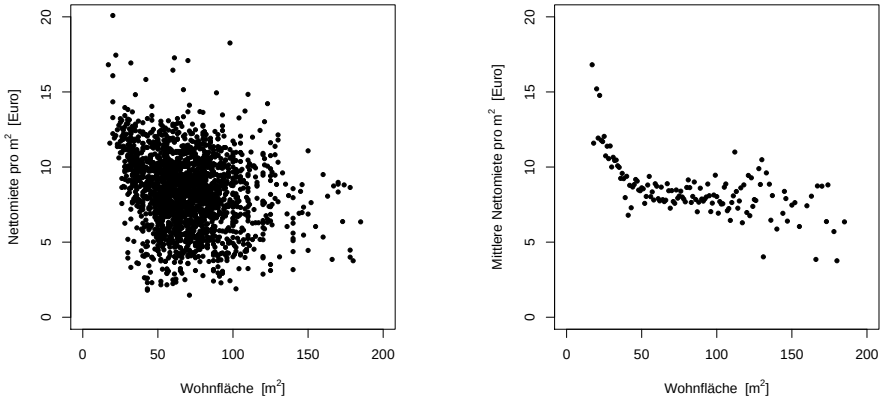
Aufgabe 11:

Zur Erstellung des Mietspiegels für München im Jahr 2003 wurden unter anderem die folgenden Merkmale erhoben:

Variablenname	Beschreibung
$nmqm$	Nettomiete pro m^2 in Euro
wfl	Wohnfläche in m^2
$lage$	Wohnlage in den Kategorien 'normal', 'gut', 'beste'

Der Zusammenhang zwischen Nettomiete pro m^2 und Wohnfläche soll mit einer linearen Einfachregression untersucht werden. Dazu wird eine im Datenarchiv des Instituts für Statistik frei verfügbare Stichprobe mit 2053 Wohnungen aus dem Originaldatensatz des Mietspiegels herangezogen. Hintergrundinformationen dazu finden Sie unter folgendem Link: `http://www.stat.uni-muenchen.de/service/datenarchiv/miete/miete03.html`.

(a) Eine erste deskriptive Analyse des Zusammenhangs zwischen Wohnfläche und Nettomiete pro m^2 ergibt die folgenden Streudiagramme:



Der Mittelwert der Variable Nettomiete pro m^2 beträgt $\overline{nmqm} = 8.39$ Euro, der Mittelwert der Variable Wohnfläche beträgt $\overline{wfl} = 69.60m^2$.

Was lässt sich aufgrund der beiden Diagramme über den Zusammenhang zwischen den beiden Merkmalen sagen?

(b) Betrachtet werden die folgenden vier Modelle:

$$(1) \quad nmqm_i = \beta_0 + \beta_1 \cdot wfl_i + \varepsilon_i$$
$$(2) \quad nmqm_i = \beta_0 + \beta_1 \cdot (wfl_i - \overline{wfl}) + \varepsilon_i$$
$$(3) \quad nmqm_i = \beta_0 + \beta_1 \cdot wfl_i^{-1} + \varepsilon_i$$
$$(4) \quad nmqm_i = \beta_0 + \beta_1 \cdot wfl_i + \beta_2 \cdot wfl_i^2 + \beta_3 \cdot wfl_i^3 + \varepsilon_i$$

Für diese Modelle ergeben sich die folgenden Parameterschätzungen:

Modell	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
(1)	9.94	-0.0222		
(2)	8.39	-0.0222		
(3)	6.61	108.2305		
(4)	15.53	-0.2373	0.0024	-0.000008

Wie lassen sich die Parameter jeweils interpretieren?

(c) Begründen Sie, weshalb der Modellierungsansatz

$$nmqm_i = \beta_0 + \beta_1 \cdot (wfl_i - \overline{wfl})^{-1}$$

zu Problemen führt und nennen Sie eine weitere Transformation der Wohnfläche, unter der sich dieselbe Problematik zeigt.

(d) Aus den Daten ergibt sich für Wohnungen mit $30m^2$ bzw. $90m^2$ Wohnfläche eine durchschnittliche Nettomiete pro m^2 von 10.00 bzw. 7.87 Euro.

Berechnen Sie, basierend auf den Parameterschätzungen in (b), jeweils die prognostizierte Nettomiete pro m^2 für die beiden betrachteten Wohnungen. Welches Modell scheint demnach am besten zu passen?

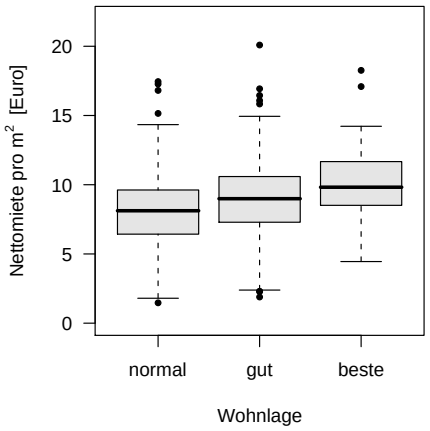
(e) Skizzieren Sie für die vier Modelle jeweils den partiellen Effekt der Wohnfläche, ausgewertet an den Stellen $wfl = 20, 40, \ldots, 180$.

(f) Die Wertebereiche der Effekte der Wohnfläche in (e) sind sehr unterschiedlich, was ihren Vergleich erschwert. Wie könnte man den jeweiligen Effekt der Wohnfläche um den Wert 0 (auf der y-Achse) zentrieren? Welche Konsequenzen ergeben sich nach der Zentrierung für die Parameterinterpretation und für die Prognose?

Aufgabe 12:

Jetzt soll mit der Mietspiegel-Stichprobe der Zusammenhang zwischen Nettomiete pro m^2 und Wohnlage mittels eines linearen Regressionsmodells untersucht werden.

Eine erste deskriptive Analyse des Zusammenhangs zwischen Wohnlage und Nettomiete pro m^2 ergibt, dass die mittlere Nettomiete pro m^2 in normaler Wohnlage 8.02 €, in guter Wohnlage 8.86 € und in bester Wohnlage 10.21 € beträgt. Außerdem ergeben sich die folgenden gruppierten Boxplots:



(a) Kodieren Sie die Kovariable *Wohnlage* gemäß Dummy- und Effekt-Kodierung mit normaler Wohnlage als Referenzkategorie. Geben Sie dabei für beide Kodierungen die entsprechenden Dummy-Variablen an.

(b) Für das Regressionsmodell ergeben sich mit den zwei angegebenen Kodierungen die folgenden Parameterschätzungen:

Modell	$\hat{\beta}_0$	$\hat{\beta}_{gut}$	$\hat{\beta}_{best}$
Dummy-Kodierung	8.02	0.84	2.20
Effekt-Kodierung	9.03	-0.17	1.18

Interpretieren Sie diese Ergebnisse. Können die Vermutungen aus der deskriptiven Analyse bestätigt werden?

(c) Wie lassen sich die Parameterschätzungen in (b) ineinander überführen?

(d) Welche Nettomiete pro m^2 prognostizieren die beiden Modelle in normaler bzw. in bester Wohnlage?

3.2 Schätzen & Testen

Aufgabe 13:

Wie in Aufgabe 10 wird der Zusammenhang zwischen Einkommen und Attraktivität im Allbus-Datensatz 2010 betrachtet. Aus didaktischen Gründen wurde der vorliegende Datensatz vorab auf 400 Beobachtungen reduziert, er enthält die folgenden Variablen:

Variable	Beschreibung
`einkommen`	Nettoeinkommen in Euro
`logeinkommen`	Logarithmiertes Nettoeinkommen: log(`einkommen`)
`alter`	Alter in Jahren
`alterc`	Zentriertes Alter: $\texttt{alter} - \overline{\texttt{alter}} = \texttt{alter} - 50.52$
`geschlecht`	Geschlecht (0 = Mann, 1 = Frau)
`attrakt`	Geschätzte Attraktivität des Befragten auf einer Skala von 1 (unattraktiv) bis 11 (attraktiv)
`attrKat`	Attraktivität in 3 Kategorien: 1-5, 6-8, 9-11

Zur Modellierung des Zusammenhangs zwischen dem logarithmiertem Nettoeinkommen und den Kovariablen wurde ein lineares Regressionsmodell gerechnet (Modell 1, R-Modelloutput siehe Seite 23 im Anhang).

(a) Berechnen Sie die Werte `A` bis `F` im Output auf Seite 3. Geben Sie dabei jeweils die zugrunde liegenden Formeln an.

(b) Bestimmen Sie ein 95%-Konfidenzintervall für den geschätzten Regressionskoeffizienten der Kovariablen `attrKat6-8`.

(c) Welche Aussagen lassen sich damit über den Zusammenhang zwischen Attraktivität und Einkommen mit Hilfe von t-Tests auf einzelne Kovariablen zum Niveau $\alpha = 0.05$ treffen?

(d) Was lässt sich für das Modell bezüglich des Tests der Hypothese
$H_0 : \beta_{\texttt{alterc}} = \beta_{\texttt{geschlecht}} = \beta_{\texttt{attrKat6-8}} = \beta_{\texttt{attrKat9-11}} = 0$
aussagen, wenn man $\alpha = 0.005$ zugrunde legt?

Aufgabe 14:

In dieser Aufgabe soll mit verschiedenen Tests überprüft werden, ob die Variable `attrKat` aus dem Modell in Aufgabe 13 insgesamt einen zum Niveau von $\alpha = 0.05$ signifikanten Zusammenhang mit dem Einkommen hat.

Dazu wurde ein weiteres lineares Regressionsmodell ohne die Kovariable `attrKat` gerechnet (Modell 2, R-Modelloutput siehe Seite 24 im Anhang).

(a) Formulieren Sie zunächst die Nullhypothese als lineare Hypothese in der Form

$$H_0 : \boldsymbol{C\beta} = \boldsymbol{d} .$$

(b) Berechnen Sie mit Hilfe der beiden Modelloutputs die Residuenquadratsummen SSE und SSE_{H_0}. Führen Sie dann zur Überprüfung der obigen Hypothese den F-Test zum Niveau von $\alpha = 0.05$ durch.

(c) Führen Sie zur Überprüfung der obigen Hypothese den Wald-Test zum Niveau von $\alpha = 0.05$ durch. Welcher Zusammenhang besteht zum F-Test?

(d) Führen Sie zur Überprüfung der obigen Hypothese den Likelihood-Quotienten-Test zum Niveau von $\alpha = 0.05$ durch.

(e) Welches der beiden Modelle würden Sie aufgrund von adjustiertem Bestimmtheitsmaß und AIC bevorzugen?

Aufgabe 15:

Eine überregionale Firma möchte die Wirksamkeit ihrer Werbemaßnahmen analysieren. Dazu untersucht sie in $n = 150$ Städten den Umsatz eines Produktes Y (in Euro) in Abhängigkeit von den Ausgaben für Werbung. Die Zielvariable wird in einem linearen Regressionsmodell in Abhängigkeit von den folgenden Kovariablen (und Intercept) untersucht:

X_1	Ausgaben für Werbung durch Inserate in lokalen Zeitungen (in Euro)
X_2	Ausgaben für Werbung durch Flyer (in Euro)
X_3	Ausgaben für Werbung durch Plakate (in Euro)

(a) Betrachtet wird die eine lineare Nullhypothese der Form $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ mit

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & -1 \end{pmatrix}.$$

Welche inhaltliche Aussage wird durch diese Hypothese getestet?

(b) Berechnen Sie für diese Nullhypothese die Wald-Teststatistik.

Aufgabe 16:

Zeigen Sie, dass sich die Fisher-Informationsmatrizen im linearen Modell mit Normalverteilungsannahme wie folgt ergeben:

$$\boldsymbol{F}_{obs}^{-1}(\hat{\boldsymbol{\beta}}) = \boldsymbol{F}^{-1}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top \boldsymbol{X})^{-1} .$$

Aufgabe 17:

Gegeben seien unabhängige und identisch verteilte Zufallsvariablen $X_1, \ldots, X_n$ mit Erwartungswert μ und Varianz σ^2. Als Schätzfunktionen für σ^2 lassen sich die empirische Varianz $\tilde{S}^2$ und die Stichprobenvarianz S^2 unterscheiden:

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad \text{und} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

(a) Zeigen Sie, dass gilt:

$$\mathrm{E}(\tilde{S}^2) = \frac{n-1}{n}\sigma^2$$

Was kann daraus für $\tilde{S}^2$ geschlossen werden?

(b) Zeigen Sie, dass gilt:

$$\mathrm{E}(S^2) = \sigma^2$$

Welche der beiden Schätzfunktionen sollte demnach bevorzugt werden?

Aufgabe 18:

Die Arbeitslosenzahlen y_t der letzten fünf Jahre seien durch folgendes Regressionsmodell beschrieben:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_t^2 + \beta_3 z_t + \beta_4 \cdot \sqrt{z_t} + \beta_5 x_t z_t + \beta_6 \cdot t + \beta_7 \cdot \sin(2\pi \cdot m_t/12), \quad t = 1, \ldots, 60$$

Dabei seien x_t und z_t Konjunkturmaßzahlen, t bezeichne die Zeit in Monaten und m_t den Kalendermonat, wobei $m_t = 1$ für Januar, $m_t = 2$ für Februar usw. steht.

(a) Drücken Sie die folgenden Aussagen durch lineare Hypothesen der Art $\boldsymbol{C\beta} = \boldsymbol{d}$ aus:

(i) "Die Arbeitslosenzahlen unterliegen keinem zeitlichen Einfluss."

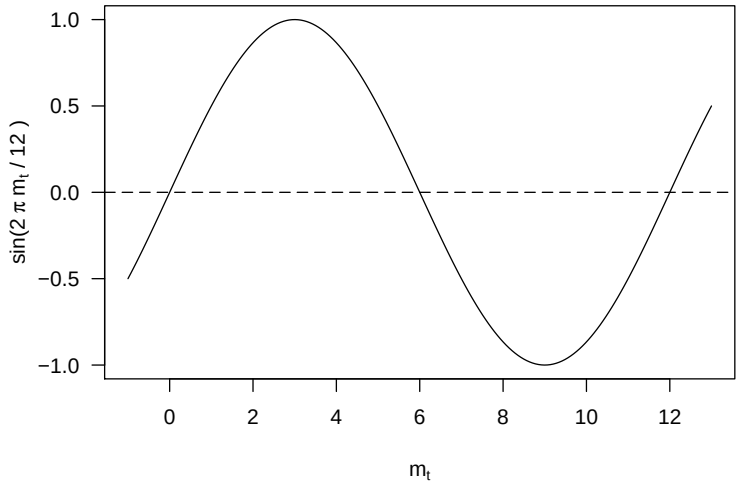
(ii) "Die Arbeitslosenzahlen hängen nicht von der Maßzahl x_t ab."

(b) Interpretieren Sie die Hypothesen

(i) $H_0 : \beta_0 = 4 \cdot 10^6$

(ii) $H_0 : \begin{pmatrix} 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \cdot \boldsymbol{\beta} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

Hinweis:
Die Funktion für die saisonale Komponente $f(m_t) = \sin(2\pi \cdot m_t/12)$ verläuft wie folgt:



Aufgabe 19:

Bei einer Untersuchung des Zusammenhangs zwischen Körpermaßen und Geweihlänge einer speziellen Fliegenart im baltischen Bernstein wurden die folgenden Merkmale bei 16 Fliegen erhoben:

Variable	Beschreibung
Geweihlänge	Geweihlänge der Fliege in mm
Breite	Körperbreite der Fliege in mm
Länge	Körperlänge der Fliege in mm
Flügel	Flügellänge der Fliege in mm
Geschlecht	Geschlecht der Fliege (0 = weiblich, 1 = männlich)

Zur Modellierung des Zusammenhangs zwischen der Geweihlänge der Fliegen und den Kovariablen wird ein lineares Regressionsmodell gerechnet, es ergibt sich folgender R–Output:

```
Call:
lm(formula = Geweih ~ Laenge + Breite + Fluegel + Geschlecht,
    data = fliegen)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.59038    0.45452  -3.499  0.00498 **
Laenge          A       0.22605   0.290  0.77697
Breite       0.19705    0.81885     C    0.81425
Fluegel      1.06291       B      3.206  0.00837 **
Geschlecht   0.55848    0.12699   4.398     D
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1706 on 11 degrees of freedom
Multiple R-squared: 0.8958,     Adjusted R-squared:   E
F-statistic: 23.65 on 4 and 11 DF,  p-value: 2.346e-05
```

(a) Berechnen Sie die Werte A bis E im R–Output. Geben Sie dabei jeweils die zugrunde liegenden Formeln an.

(b) Bestimmen Sie ein 95%-Konfidenzintervall für den geschätzten Regressionskoeffizienten der Kovariablen *Flügel*.

(c) Was lässt sich für das Modell bezüglich des Tests der Hypothese
$H_0 : \beta_{Länge} = \beta_{Breite} = \beta_{Flügel} = \beta_{Geschlecht} = 0$
aussagen, wenn man $\alpha = 0.005$ zugrunde legt?

(d) Für die Kovarianzmatrix der Parameterschätzer ergibt sich in R:

```
             (Intercept) Laenge Breite Fluegel Geschlecht
(Intercept)        0.207 -0.033  0.138  -0.072     -0.030
Laenge            -0.033    F   -0.112  -0.023      0.604
Breite             0.138 -0.112  0.671  -0.135        G
Fluegel           -0.072 -0.023 -0.135   0.110      0.009
Geschlecht        -0.030  0.604    H     0.009      0.016
```

Für die Korrelationsmatrix ergibt sich:

	(Intercept)	Laenge	Breite	Fluegel	Geschlecht
(Intercept)	I	-0.318	0.370	-0.476	-0.520
Laenge	-0.318	J	-0.606	-0.303	0.604
Breite	0.370	-0.606	K	-0.498	-0.656
Fluegel	-0.476	-0.303	-0.498	L	N
Geschlecht	-0.520	0.604	-0.656	O	M

Berechnen Sie die Werte F bis O.

(e) In einem reduzierten Modell wurden nur die beiden Kovariablen *Länge* und *Geschlecht* in die Regressionsgleichung aufgenommen, alle anderen Kovariablen wurden weggelassen. Es ergibt sich der folgende R–Output:

```
Call:
lm(formula = Geweih ~ Laenge + Geschlecht, data = fliegen)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.0448     0.5542  -1.885 0.081911 .
Laenge        0.7024     0.1219   5.762 6.58e-05 ***
Geschlecht    0.6402     0.1330   4.815 0.000338 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2401 on 13 degrees of freedom
Multiple R-squared: 0.7561,     Adjusted R-squared: 0.7186
F-statistic: 20.15 on 2 and 13 DF,  p-value: 0.0001039
```

Welcher wesentliche Unterschied fällt im Vergleich zum ersten Modell auf? Worauf könnte dieser Unterschied zurückzuführen sein?

Hinweis: Korrelationsmatrix für die Merkmale:

	Laenge	Breite	Fluegel
Laenge	1.0000	0.8498	0.8842
Breite	0.8498	1.0000	0.8702
Fluegel	0.8842	0.8702	1.0000

Kapitel 4: Generalisierte lineare Regressionsmodelle – Binäre Regression

Aufgabe 20:

In der Vorwerk-Familienstudie von 2007, die gemeinsam mit dem Institut für Demoskopie Allensbach durchgeführt wurde, ging es unter anderem um das Thema, ob das Leben mit Kindern als glücklicher empfunden wird. In diesem Zusammenhang wurde die folgende Frage gestellt:

> „Glauben Sie, dass man Kinder braucht, um wirklich glücklich zu sein – oder glauben Sie, man kann ohne Kinder genauso glücklich leben?“

788 Befragte unter 45 Jahren (379 Männer und 409 Frauen) wählten eine der folgenden Antwortmöglichkeiten: „Ja, man braucht Kinder“ oder „Nein, man kann genauso ohne Kinder glücklich leben“.

Insgesamt ergab sich die folgende Kreuztabelle der absoluten Häufigkeiten:

	Nein	Ja
Männer	212	167
Frauen	172	237

(a) Berechnen Sie mit Hilfe der Kreuztabelle die folgenden Größen:

i. Relative Häufigkeiten für Zustimmung zur Frage in Abhängigkeit vom Geschlecht

ii. Empirische Chancen, d.h. Verhältnis zwischen Zustimmung und Ablehnung zur Frage, in Abhängigkeit vom Geschlecht

iii. Empirisches Chancenverhältnis (Odds Ratio) zwischen Frauen und Männern

(b) Die Schätzung der Wahrscheinlichkeit für die Zustimmung zur Frage mittels eines Logit–Modells mit dem Geschlecht als Kovariable (Referenzkategorie “Männer”) liefert das folgende Ergebnis:

	$\hat{\beta}_0$	$\hat{\beta}_1$
Dummykodierung	-0.2386	0.5592

Interpretieren Sie die geschätzten Parameter auf der Ebene der logarithmierten Chancen, der Chancen und der Wahrscheinlichkeiten. Welche Zusammenhänge bestehen zu den berechneten Größen aus (a)?

Aufgabe 21:

Gegeben sei eine Stichprobe unabhängiger Beobachtungen $(x_1, y_1), \ldots, (x_n, y_n)$ vom Umfang n, wobei x eine Kovariable und y eine binäre Zielvariable bezeichnet. Der Zusammenhang zwischen x und y wird durch folgendes Logit-Modell beschrieben:

$$y_i \overset{ind.}{\sim} \mathrm{B}(1, \pi_i) \qquad \text{mit} \qquad \pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \qquad \text{für } i = 1, \ldots, n$$

(a) Bestimmen Sie die Likelihood $L(\beta_0, \beta_1)$ sowie die Log-Likelihood $l(\beta_0, \beta_1)$ für die unbekannten Parameter β_0 und β_1 des Regressionsmodells.

Hinweis: Bilden Sie zunächst die Likelihood $L_i(\beta_0, \beta_1)$ und die Log-Likelihood $l_i(\beta_0, \beta_1)$ bezüglich der i-ten Beobachtung.

(b) Bestimmen Sie die Scorefunktion $s(\beta_0, \beta_1)$.

(c) Mit welchem Verfahren kann der ML-Schätzer für die Parameter (β_0, β_1) bestimmt werden? Skizzieren Sie das Verfahren für einen eindimensionalen Parameter θ.

Aufgabe 22:

Für eine Analyse des Europäischen Patentamts zum Thema „Einsprüche gegen Patente“ wurden die folgenden Merkmale für 4866 erteilte Patente aus den Branchen *Biotechnologie/Pharma* und *Halbleiter/Computer* erhoben:

Zielvariable (Ausprägungen: 1 = Ja / 0 = Nein)	
`einspruch`	Einspruch gegen das Patent

Stetige Kovariablen	
`jahr`	Jahr der Patenterteilung
`azit`	Anzahl der Zitationen für dieses Patent
`aland`	Anzahl der Länder, für die Patentschutz gelten soll
`ansp`	Anzahl der Patentansprüche

Binäre Kovariablen (1 = Ja / 0 = Nein)	
`biopharm`	Patent aus Biotechnologie-/Pharma–Branche
`uszw`	US Zwillingspatent
`patus`	Patentinhaber aus den USA
`patdsg`	Patentinhaber aus Deutschland, Schweiz oder Großbritannien

Das Ziel der Untersuchung war es, die Wahrscheinlichkeit für einen Patenteinspruch in Abhängigkeit von Kovariablen zu modellieren.

Die metrischen Kovariablen wurden vor der Schätzung jeweils um ihren Mittelwert zentriert, was durch die Endung `c` an den metrischen Kovariablen erkennbar ist. Es wurden folgende zentrierte Kovariablen konstruiert:

Zentrierte Kovariablen	
jahrc	jahr - $\overline{\text{jahr}}$ = jahr - 1991.07
azitc	azit - $\overline{\text{azit}}$ = azit - 1.64
alandc	aland - $\overline{\text{aland}}$ = aland - 7.80
alandc2	aland^2 - $\overline{\text{aland}^2}$ = aland^2 - 77.76
alandc3	aland^3 - $\overline{\text{aland}^3}$ = aland^3 - 889.12
anspc	ansp - $\overline{\text{ansp}}$ = ansp - 13.13

Der folgende R–Output zeigt die Ergebnisse der Schätzung eines linearen Logit–Modells für die Wahrscheinlichkeit eines Einspruchs.

```
Call:
glm(formula = einspruch ~ jahrc + azitc + alandc + anspc + uszw +
    patus + patdsg + biopharm, family = binomial(link = "logit"),
    data = patent)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.477417   0.079461  -6.008 1.88e-09 ***
jahrc       -0.071335   0.008706  -8.194 2.53e-16 ***
azitc        0.117972   0.014218   8.297  < 2e-16 ***
alandc       0.084422   0.010666   7.915 2.47e-15 ***
anspc        0.017676   0.003387   5.219 1.80e-07 ***
uszw        -0.391677   0.067583  -5.795 6.81e-09 ***
patus       -0.151550   0.075712  -2.002   0.0453 *
patdsg       0.322948   0.082871   3.897 9.74e-05 ***
biopharm     0.680641   0.083743   8.128 4.37e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 6604.1  on 4865  degrees of freedom
Residual deviance: 5815.5  on 4857  degrees of freedom
AIC: 5833.5

Number of Fisher Scoring iterations: 4
```

(a) Interpretieren Sie die geschätzten Regressionsparameter der Variablen `alandc` und `biopharm` sowie die Konstante `(Intercept)`. Wie ändern sich jeweils die logarithmierten Chancen, die Chancen und die Wahrscheinlichkeit, falls sich die entsprechende Kovariable ändert?

(b) Testen Sie mittels eines geeigneten Tests zum Niveau $\alpha = 0.01$ die Signifikanz des Einflusses der Kovariablen `patus`. Wie klein dürfte man α wählen, damit die entsprechende Nullhypothese gerade noch abgelehnt werden könnte?

(c) Berechnen Sie die prognostizierte Wahrscheinlichkeit für einen Patenteinspruch gegen ein deutsches Patent aus der Halbleiterbranche mit US-Zwilling, das 1994 erteilt wurde, für 11 Staaten Patentschutz bietet, 2 Zitationen erhielt und 16 Patentansprüche angemeldet hat.

In einer Erweiterung des Modells wurde der Einfluss der Kovariable `alandc` mittels eines Polynoms 3. Grades modelliert. Es ergab sich der folgende Output:

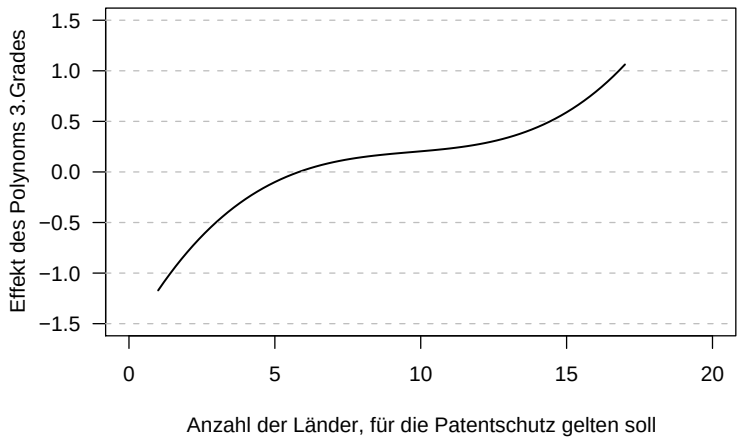
```
Call:
glm(formula = einspruch ~ jahrc + azitc + alandc + alandc2 +
    alandc3 + anspc + uszw + patus + patdsg + biopharm,
    family = binomial(link = "logit"), data = patent)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.4788516  0.0798309  -5.998 1.99e-09 ***
jahrc       -0.0678202  0.0090857  -7.464 8.36e-14 ***
azitc        0.1186959  0.0142560   8.326  < 2e-16 ***
alandc       0.5201918  0.1308003   3.977 6.98e-05 ***
alandc2     -0.0511566  0.0155445  -3.291 0.000998 ***
alandc3      0.0017597  0.0005583   3.152 0.001623 **
anspc        0.0176530  0.0033916   5.205 1.94e-07 ***
uszw        -0.3972353  0.0677318  -5.865 4.50e-09 ***
patus       -0.1670024  0.0762402  -2.190 0.028490 *
patdsg       0.2935930  0.0837841   3.504 0.000458 ***
biopharm     0.7086731  0.0850969   8.328  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

    Null deviance: 6604.1  on 4865  degrees of freedom
Residual deviance: 5804.1  on 4855  degrees of freedom
AIC: 5826.1
```

(d) Testen Sie mittels eines geeigneten Tests zum Niveau $\alpha = 0.025$, ob die komplexere Modellierung durch das Polynom für die Variable `aland` signifikant ist.

Hinweis: Die Log-Likelihood für das lineare Logit-Modell beträgt -2907.755, für das erweiterte Modell beträgt sie -2902.04.

(e) Welches Modell ist nach dem Akaike Informationskriterium `AIC` zu bevorzugen?

(f) Interpretieren Sie den in der folgenden Abbildung dargestellten, zentrierten Effekt der Variablen `aland`:



Anhang

Modelloutput zu Aufgabe 13

Modell 1:

- Modelloutput:

```
> summary(modell1)

Call:
lm(formula = logeinkommen ~ alterc + geschlecht + attrKat, data = allbusDemo)

Residuals:
Min      1Q  Median      3Q     Max
-3.3819 -0.3658  0.0155  0.3991  2.3734

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.08046    0.10782  65.669  < 2e-16 ***
alterc          0.00264    0.00200   1.320   0.1876
geschlechtFrau -0.54969    0.07037  -7.811 5.16e-14 ***
attrKat6-8      0.22644    0.10952   2.068   0.0393 *
attrKat9-11     0.24109    0.11783   2.046   0.0414 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.7004 on 395 degrees of freedom
Multiple R-squared:   0.15,Adjusted R-squared:  0.1414
F-statistic: 17.43 on 4 and 395 DF,  p-value: 3.501e-13
```

- Geschätzte Kovarianzmatrix $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})$:

```
> vcov(modell1)
                (Intercept)         alterc geschlechtFrau    attrKat6-8   attrKat9-11
(Intercept)     1.162533e-02 -2.654186e-05  -3.004223e-03 -1.012875e-02 -1.009514e-02
alterc         -2.654186e-05  4.000919e-06  -8.633630e-07  1.651667e-05  4.046524e-05
geschlechtFrau -3.004223e-03 -8.633630e-07   4.952040e-03  6.536955e-04  3.318596e-04
attrKat6-8     -1.012875e-02  1.651667e-05   6.536955e-04  1.199502e-02  9.831929e-03
attrKat9-11    -1.009514e-02  4.046524e-05   3.318596e-04  9.831929e-03  1.388520e-02
```

- Wert der Log-Likelihood an der Stelle $\hat{\boldsymbol{\beta}}$:

```
> logLik(modell1)
'log Lik.' -422.6358 (df=6)
```

Modelloutput zu Aufgabe 14

Modell 2:

- Modelloutput:

```
> summary(modell2)

Call:
lm(formula = logeinkommen ~ alterc + geschlecht, data = allbusDemo)

Residuals:
Min      1Q  Median      3Q     Max
-3.3477 -0.3912  0.0223  0.3990  2.4080

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.288749   0.050384 144.664  < 2e-16 ***
alterc          0.002072   0.001970   1.052    0.294
geschlechtFrau -0.560086   0.070313  -7.966 1.75e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.7029 on 397 degrees of freedom
Multiple R-squared:  0.1397,Adjusted R-squared:  0.1354
F-statistic: 32.23 on 2 and 397 DF,  p-value: 1.07e-13
```

- Geschätzte Kovarianzmatrix $\widehat{\mathrm{Cov}}(\tilde{\boldsymbol{\beta}})$:

```
> vcov(modell2)
(Intercept)        alterc geschlechtFrau
(Intercept)     2.538530e-03 -4.339082e-06  -2.532755e-03
alterc         -4.339082e-06  3.880939e-06  -8.266716e-07
geschlechtFrau -2.532755e-03 -8.266716e-07   4.943940e-03
```

- Wert der Log-Likelihood an der Stelle $\tilde{\boldsymbol{\beta}}$:

```
> logLik(modell2)
'log Lik.' -425.0531 (df=4)
```