Institut für Stochastik
Universität Karlsruhe
JProf. Dr. H. Holzmann
Dipl.-Math. oec. D. Engel

SS 2008

Übungen zum

Statistischen Praktikum

Aufgabe 1 (Zusatzaufgabe 1 zur logistischen Regression)
Am 28.01.1986 explodierte die amerikanische Raumfähre Challenger (OV-99) kurz nach dem Start der Mission STS-51-L mit sieben Astronauten an Bord. Als Ursache der Katastrophe wurden Materialermüdungserscheinungen an den Dichtungsringen im Bereich der Triebwerke, den so genannten O-Ringen, ausgemacht. Von diesen O-Ringen besitzen die beiden Raketentriebwerke insgesamt sechs Stück. Zum Zeitpunkt des Startes der Challenger herrschte eine Außentemperatur von nur 31°F (ca. 0°C).
In der Nacht vor der Challenger Katastrophe fand eine mehrstündige Telefonkonferenz mit den Experten des Triebwerkherstellers Morton Thiokol, des Marshall Space Flight Center der NASA und des Kennedy Space Center Raumflughafens statt. Im Wesentlichen ging es dabei um die Wettervorhersage von extrem niedrigen 31°F Außentemperatur für die Startzeit am nächsten Morgen und den Effekt einer niedrigen Außentemperatur auf die Zuverlässigkeit der O-Ringe.
Bei dieser Telefonkonferenz spielte der Datensatz `challeng`, den Sie im R-Paket `alr3` sowie unter `challeng.txt` auf der Seite zum Praktikum finden, eine wichtige Rolle. Dieser gibt in `Fail` für die vorherigen Flüge der NASA-Raumfähren an, wie viele der O-Ringe Ermüdungserscheinungen zeigten. Außerdem gibt (u.a.) `Temp` die „Außentemperatur in °F“ zur jeweiligen Startzeit an.

a) Laden Sie zunächst das Paket `alr3` mit Hilfe des Befehls `library(alr3)` und anschließend den Datensatz `challeng` mit `data(challeng)` oder lesen Sie den Datensatz `challeng.txt` von der Seite zum Praktikum ein. Bestimmen Sie die minimale und die maximale Außentemperatur bei den bisherigen Starts sowie deren Mittelwert und Median. Bei welchen Temperaturen traten in der Vergangenheit Probleme mit den O-Ringen auf (`Fail`> 0)? Beurteilen Sie, die auf dieser Grundlage getroffene Entscheidung, die Raumfähre am nächsten Morgen starten zu lassen.

b) Nach der Katastrophe nahm eine von der U.S. Regierung eingesetzte Expertenkommision (Rogers-Kommision) die Starttemperaturen der problemlos verlaufenen Flüge hinzu.
Führen Sie die Zielvariable `y` ein, die 1 ist, falls beim entsprechenden Flug ein Problem mit den Dichtungsringen aufgetreten war, das heißt `Fail`> 0, und die 0 ist, falls kein Problem vorlag. Zeichen Sie nun ein Streudiagramm von `y` gegen `Temp`. Was beobachten Sie? Wie würden Sie auf dieser Grundlage die Entscheidung für einen Start bei 31°F bewerten?

c) Passen Sie ein lineares Modell der Form $\texttt{y} = \beta_1 + \beta_2 \texttt{Temp} + \varepsilon$ an die Daten an. Ist der lineare `Temp`-Term zum 99%−Niveau signifikant? Zeichnen Sie die Regressionsgerade in das Streudiagramm aus Teilaufgabe b) ein. Warum liefert das lineare Modell keine gute Beschreibung der Daten?

d) Beschreiben Sie nun die Wahrscheinlichkeit p für den Ausfall eines Dichtungsrings in Abhängigkeit von der Außentemperatur `Temp` zum Startzeitpunkt durch ein logistisches Modell (Logit-Link). Welche Schätzwerte erhalten Sie für die Modellparameter?

e) Plotten Sie die Ausfallwahrscheinlichkeit eines Dichtungsrings, die Sie mit dem angepassten Modell aus Teilaufgabe d) erhalten, in Abhängigkeit von der Außentemperatur. Geben Sie die Wahrscheinlichkeit an, mit welcher (unter diesem Modell) bei der prognostizierten Temperatur von 31°F mit einem Ausfall zu rechnen war.

f) Wiederholen Sie die Teilaufgaben d) und e) für

   (*i*) die Probit-Linkfunktion

   (*ii*) die komplementäre Log-log-Linkfunktion

   anstatt der Logit-Linkfunktion.
   R-Hinweis: Sie erhalten die Probit-Linkfunktion bzw. die komplementäre Log-log-Linkfunktion mit Hilfe der Befehle `glm(y~Temp,family=binomial(link=‘‘probit‘‘))` bzw. `glm(y~Temp,family=binomial(link=‘‘cloglog‘‘))`.

Aufgabe 2 (Zusatzaufgabe 2 zur logistischen Regression)
Betrachten Sie nochmals den `challeng`-Datensatz aus dem Paket `alr3`, der bereits in der Zusatzaufgabe 1 untersucht wurde.

a) Fassen Sie nun jeweils die Gesamtanzahl der Flüge und diejenigen mit Ermüdungserscheinungen bei den Dichtungsringen für die Starttemperaturen 51°F bis (einschließlich) 55°F, 56°F bis 60°F,...,81°F bis 85°F zusammen und verwenden Sie den Mittelpunkt eines jeden Intervalls als erklärende Variable.

b) Bestimmen Sie die Wahrscheinlichkeit p für die einzelnen Temperaturintervalle mit denen ein Problem bezüglich der O-Ringe auftrat und passen Sie für diese Wahrscheinlichkeiten ein logistisches Modell an.

c) Erstellen Sie ein Streudiagramm der Wahrscheinlichkeit für ein Problem mit den O-Ringen in Abhängigkeit der Mittelpunkte der entsprechneden Temperaturintervallen und zeichnen Sie die Kurve der angepassten Responsefunktion aus Teilaufgabe b) ein.

d) Ersetzen Sie Wahrscheinlichkeiten p von 0 und 1 durch 0.001 beziehungsweise 0.999, so dass Sie die Logit-Linkfunktion für die Wahrscheinlichkeiten berechnen können und berechnen Sie diese, das heißt, bestimmen Sie

$$\text{Logit}(p) = \log\left(\frac{p}{1-p}\right)$$

   für die einzelnen Temperaturintervalle.

e) Passen Sie ein lineares Modell der Form

$$\log\left(\frac{p}{1-p}\right) = \beta_1 + \beta_2 \texttt{Temp} + \varepsilon =: \eta + \varepsilon$$

   an. Replizieren Sie die Werte der einzelnen Temperaturintervalle dabei entsprechend der Anzahl der Beobachtungen in diesen Intervallen. Welche Kleinste-Quadrate Schätzer $\widehat{\beta}_1$ und $\widehat{\beta}_2$ erhalten Sie für das lineare Modell?

f) Zeichnen Sie die Kurve $\dfrac{1}{1+\mathrm{e}^{-\widehat{\eta}}}$ mit $\widehat{\eta} = \widehat{\beta}_1 + \widehat{\beta}_2 \texttt{Temp}$ ebenfalls in das Schaubild aus Teilaufgabe c) ein. Stimmt diese Kurve mit derjenigen aus Teilaufgabe c) überein? Erklären Sie dies kurz.

g) Wie groß ist bei den beiden oben betrachteten Modellen die Wahrscheinlichkeit eines Problems mit den O-Ringen bei einer Starttemperatur von 31°F?

h) Betrachten Sie jetzt nur noch zwei Klassen (Gruppen) von Flügen, nämlich solche, bei denen die Außentemperatur zum Startzeitpunkt maximal 67°F betrug, und solche mit einer Außentemperatur von über 67°F beim Start. Stellen Sie die 23 Individualdaten aus dem ursprünglichen Datensatz nach dieser Gruppierung in Form einer so genannten Kontingenztabelle dar, das heißt bestimmen Sie die Anzahl der Flüge a, b, c und d in der folgenden Tabelle

	Problem	kein Problem
`Temp` ≤ 67	a	b
`Temp` > 67	c	d

und stellen Sie die Anteile graphisch in einem so genannten Mosaik-Plot dar.

i) Passen Sie an die gruppierten Daten aus Teilaufgabe h) auf die folgenden drei Arten jeweils ein logistisches Modell an:

   (*i*) „direkt“ `glm()`-Befehl auf binäre Größen (Problem: ja/nein und Temperatur maximal 67°F: ja/nein)anwenden.

   (*ii*) mit dem Befehl `glm(c(`a`/(`a`+`b`),`c`/(`c`+`d`))~c(1,0),family=binomial,weights=c(`a`+`b`,`c`+`d`))`.

   (*iii*) mit dem Befehl `glm(cbind(c(`a`,`c`),c(`b`,`d`))~c(1,0),family=binomial)`.

   Vergleichen Sie die `summary`-Outputs der drei oben angepassten Modelle.

Aufgabe 3 (Zusatzaufgabe 3 zur logistischen Regression)
Die Larve des tobacco budworms (*Heliothis virescens*) ist für große Einbußen bei der Baumwollernte in den Vereinigten Staaten sowie Mittel- und Südamerika verantwortlich. J.W. Holloway untersuchte daher 1989 in seiner Dissertation „A comparison of the toxicity of the pyrethroid trans-cypermethrin, with and without the synergist piperonyl butoxide, to adult moths from two strains of *Heliothis virescens*“ Resistenzlevels der erwachsenen Nachtfalter gegenüber dem Insektizid Cypermethrin aus der Gruppe der Pyrethroide.
In diesem Versuch wurden Pyrethroid-resistente Falter, nach Geschlecht getrennt, in Versuchseinheiten zu je 20 Tieren zwei Tage nach dem „Schlüpfen“ verschiedenen Mengen von Cypermethrin ausgesetzt. Das Ziel dieser Untersuchung war, den Effekt einer ansteigenden Dosis an Cypermethrin auf die Toxizität zu beurteilen.
Die Daten zu diesem Experiment finden Sie im Datensatz `budworm.txt` auf der Seite zum Praktikum. Die Variable `sex` gibt an, ob bei der entsprechenden Versuchseinheit männliche (`sex`=0) oder weibliche (`sex`=1) Falter betrachtet wurden. Die Variable `dose` gibt die Dosis an Cypermethrin in Milligramm an, der die Falter ausgesetzt waren und `s` ist die Anzahl der Tiere, die 72 Stunden nach der Behandlung mit dem Insektizid kein Lebenszeichen mehr von sich gaben. Weiter gibt `n` die Anzahl der Tiere in den einzelnen Versuchseinheiten wieder ($\mathtt{n}_i = 20\ \forall i$).

a) Lesen Sie den Datensatz in R ein und verschaffen Sie sich einen ersten Überblick über die Daten.

b) Passen Sie ein logistisches Modell mit Hilfe des `glm()`-Befehls an die Daten an und betrachten Sie den `summary`-Output für dieses Modell.

Das Ziel im weiteren Verlauf dieser Aufgabe ist, die Werte, welche der `summary`-Output liefert, einmal „von Hand“ zu berechnen.

c) Bestimmen Sie zunächst iterativ den Schätzer $\widehat{\beta} = (\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)^T$ für die Koeffizienten $\beta_1,\ \beta_2$ und β_3 bei der logistischen Regression, welcher Ihnen der `summary()`-Befehl unter `Estimate` liefert. Sie können hierzu, wie folgt, vorgehen:

   (*i*) Berechnen Sie mit Hilfe einer linearen Regression Startwerte $\widehat{\beta}^{(0)} = (\widehat{\beta}_1^{(0)}, \widehat{\beta}_2^{(0)}, \widehat{\beta}_3^{(0)})^T$ für die geschätzten Regressionskoeffizienten.

   (*ii*) Berechnen Sie den linearen Prädiktor $\eta = X\widehat{\beta}^{(0)}$, wobei X, wie üblich, die Designmatrix ist.

   (*iii*) Bestimmen Sie als nächstes $p_i^{(0)} = \dfrac{\exp(\eta_i)}{1+\exp(\eta_i)}$.

   (*iv*) Bestimmen Sie nun die auf Grundlage der Wahrscheinlichkeiten $p_i^{(0)}$ erwartete Anzahl $\mu_i^{(0)} = \mathtt{n}_i p_i^{(0)}$ von „Treffern“ in der $i-$ten Gruppe.

   (*v*) Berechnen Sie den verbesserten Schätzer $\widehat{\beta}^{(1)} = (\widehat{\beta}_1^{(1)}, \widehat{\beta}_2^{(1)}, \widehat{\beta}_3^{(1)})^T$ für die Regressionskoeffizienten bei der logistischen Regression mit Hilfe der Formel

$$\widehat{\beta}^{(1)} = \widehat{\beta}^{(0)} + \left(X^T \mathrm{diag}\left[\mathtt{n}_i p_i^{(0)}\left(1-p_i^{(0)}\right)\right] X\right)^{-1} X^T(\mathtt{s} - \mu^{(0)}), \tag{1}$$

   wobei $\mathrm{diag}\left[\mathtt{n}_i p_i^{(0)}\left(1-p_i^{(0)}\right)\right]$ die Diagonalmatrix mit den Einträgen $\mathtt{n}_i p_i^{(0)}\left(1-p_i^{(0)}\right)$ auf der Diagonalen bezeichnet.

   (*vi*) Wiederholen Sie die obigen Schritte (*ii*) bis (*v*) so lange, bis $\left|\widehat{\beta}_i^{(1)} - \widehat{\beta}_i^{(0)}\right| < 10^{-6}$ für alle $i = 1, 2, 3$. Setzen Sie dabei vor Beginn jedes Durchgangs $\widehat{\beta}^{(0)} = \widehat{\beta}^{(1)}$.

d) Bestimmen Sie als nächstes die Kovarianzmatrix der in Teilaufgabe c) geschätzten Regressionskoeffizienten $\widehat{\beta} = (\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3)^T$ bei der logistischen Regression gemäß der Formel

$$\mathrm{Cov}(\widehat{\beta}) = \left(X^T \mathrm{diag}\left[\mathtt{n}_i \widehat{p}_i(1-\widehat{p}_i)\right] X\right)^{-1},$$

   wobei X wieder die Designmatrix bezeichnet und

$$\widehat{p}_i = \frac{\exp(\widehat{\beta}_1 + \widehat{\beta}_2\ \mathtt{sex}_i + \widehat{\beta}_3\ \mathtt{dose}_i)}{1+\exp(\widehat{\beta}_1 + \widehat{\beta}_2\ \mathtt{sex}_i + \widehat{\beta}_3\ \mathtt{dose}_i)}$$

   ist. Berechnen Sie damit die Standardabweichungen $\mathrm{sd}(\widehat{\beta}_i)$ für die geschätzten Regressionskoeffizienten, die Sie im `summary`-Output unter `Std. Error` finden.

e) Teilen Sie nun die geschätzten Regressionskoeffizienten $\widehat{\beta}_i$ durch ihre Standardabweichungen $\mathrm{sd}(\widehat{\beta}_i)$ um die Werte z der $z-$Teststatistik (`z value`) zu erhalten.

f) Bestimmen Sie die entsprechenden $p-$Werte (`Pr (>|z|)`) zu den $z-$Werten aus Teilaufgabe e), welche durch

$$\begin{cases} 2(1-\Phi(z)) & \text{falls } z \geq 0 \\ 2\Phi(z) & \text{falls } z < 0 \end{cases}$$

   gegeben sind, wobei Φ die Verteilungsfunktion der Standardnormalverteilung bezeichnet.

g) Geben Sie die „Nulldevianz“ (`Null deviance`), das heißt die Devianz des Modells, welches nur den Intercept enthält, die gerade durch $2(\mathcal{L}_{sat} - \mathcal{L}_{null})$ gegeben ist, an.

h) Die Devianz D des angepassten logistischen Modells (`Residual Deviance`) ist durch $D = 2(\mathcal{L}_{sat} - \mathcal{L}_{logit})$ gegeben, dabei ist $\mathcal{L}_{logit}$ der Wert der Log-Likelihood-Funktion für

$$\widehat{p}_i = \frac{\exp(\widehat{\beta}_1 + \widehat{\beta}_2 \texttt{sex}_i + \widehat{\beta}_3 \texttt{dose}_i)}{1 + \exp(\widehat{\beta}_1 + \widehat{\beta}_2 \texttt{sex}_i + \widehat{\beta}_3 \texttt{dose}_i)}.$$

i) Berechnen Sie die Devianzresiduen r_i^D. Nutzen Sie dabei aus, dass r_i^D gerade die Wurzel des Beitrags der $i-$ten Beobachtung zur Devianz ist, wobei das Vorzeichen das gleiche wie das von $\frac{\texttt{s}_i}{\texttt{n}_i} - \widehat{p}_i$ ist. Sie erhalten die Zusammenfassung der Devianzresiduen, die Sie im `summary`-Output des logistischen Modells aus Teilaufgabe b) unter `Deviance Residuals` finden, am einfachsten, wenn Sie den `summary()`-Befehl auf den Vektor der gerade berechneten Devianzresiduen anwenden.

j) Berechnen Sie den Wert des AIC mit Hilfe der Formel AIC $= -2\mathcal{L}_{logit} + 2p'$, wobei p' die Anzahl der erklärenden Variablen (inklusive Intercept) angibt.