



Statistik II

Übungsblatt 6

Abgabe: Do. 4.12.2003, 10.00 Uhr, Briefkasten: Statistik II.

Bei jeder Aufgabe können maximal 5 Punkte erreicht werden.

1. "Woodpeckers"-Datensatz:

Dieser Datensatz besteht aus zwei kategoriellen Variablen 'SiteType' und 'Age', sowie zwei binären Zielgrößen 'DOWOp' und 'BBWOp', die angeben, ob gewisse Spechtarten (DO bzw. BB) an den Untersuchungsorten angetroffen wurden. Mit einem logistischen Regressionsmodell (je eines für DOWOp bzw. BBWOp) sollte untersucht werden, ob die Variablen 'SiteType' und 'Age' oder die Interaktion 'SiteType*Age' einen signifikanten Einfluß auf die Anwesenheit der Vögel haben. Mit Data Desk 6.1 sind diese logistischen Modelle jedoch nicht sinnvoll berechenbar, da keine Konvergenz erzielt wird, und die Berechnung endlos durchgeführt wird. Untersuchen Sie mit explorativen Mitteln die Struktur des Datensatzes und finden Sie mögliche problematischen Punkte heraus, die die Konvergenz verhindern.

Mit R wurde folgendes Ergebnis erzielt:

```
Call: glm(formula = DOWOp ~ SiteType * Age, family = binomial)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.5183	-0.9282	-0.8446	1.3537	1.5518

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-0.6190	0.4688	-1.320	0.1867
SiteType	-6.9462	11.9339	-0.582	0.5605
AgeB	1.3922	0.6807	2.045	0.0408 *
AgeC	1.05e-15	0.5742	1.83e-15	1.0000
SiteType.AgeB	5.7676	11.9789	0.481	0.6302
SiteType.AgeC	6.7180	11.9584	0.562	0.5743

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 132.75 on 98 degrees of freedom

Residual deviance: 120.35 on 93 degrees of freedom

AIC: 132.35

Number of Fisher Scoring iterations: 6

Ist dieses Ergebnis mit Ihren explorativen Untersuchungen verträglich?

2. "Detergent"-Datensatz:

Dieser Datensatz stammt aus einer Untersuchung zur Beurteilung von Waschmitteln: 3 Einflußgrößen *Temperatur* (Temperature), *Wasserhärte* (WaterSoftness) und *Bisherige Verwendung von Waschmittel M* (M-User) und die binäre Zielgröße *Präferenz für Waschmittel M oder X* (Preference) wurden untersucht.

- (a) Untersuchen Sie unterschiedliche logistische Modelle für diese Daten und vergleichen Sie die Modelle.
- (b) Welches Modell finden Sie am geeignetsten? Begründen Sie Ihre Antwort.
3. Interpretieren Sie die S-Form der logistischen Linkfunktion an folgendem Beispiel: Ein bestimmtes Auto wird verschiedenen Käufern j zu verschiedenen Preisen x_j angeboten. Y sei die binäre Variable, die angibt, ob der jeweilige Interessent das Auto zum gebotenen Preiskauf würde oder nicht.

In der linearen Regression $Y = X\beta + \epsilon$ schätzt man β durch $\hat{\beta}$ und erhält mit $x_j\hat{\beta}$ eine Vorhersage für Y_j . Auch in der logistischen Regression geht man so vor; es ergebe sich $x_j\hat{\beta} = -0.5$. Wie interpretieren Sie diesen Wert?

4. Folgende Tabelle zeigt die Ergebnisse eines Experiments zur Untersuchung der Wirksamkeit eines Insektizids (*Pyrethroid Trans-Cypermethrin*) gegenüber *Heliothis virescens*, der Raupe einer Motte, die sich insbesondere von den Knospen und jungen Blättern von Tabaks- und Baumwollpflanzen ernährt (im Englischen wird diese Raupe als *tobacco budworm* bezeichnet). Jeweils 20 männliche und 20 weibliche Mottenraupen wurden je einer gewissen Dosis (in μg) des Insektizids ausgesetzt und es wurde jeweils die Anzahl der diese Prozedur nicht überlebenden Raupen notiert.

Tabelle 1: Budworm-Daten

	Dosis					
Geschlecht	1	2	4	8	16	32
männlich	1	4	9	13	18	20
weiblich	0	2	6	10	12	16

Auf unserer Internetseite finden Sie eine Datei mit einigen R-Befehlen, die eine Berechnung logistischer Regressionsmodelle ermöglichen.

- (a) Betrachten Sie die Modelle `budworm.lg.i`, `budworm.lg.2` und `budworm.lg.A`.
- Wie unterscheiden sich die Modelle?
 - Für welches dieser Modelle würden Sie sich entscheiden bzw. welches Modell würden Sie aus den gegebenen Modellen als geeignet ableiten? Begründen Sie Ihre Antwort!
 - Welche weiteren Berechnungen oder Graphiken würden Sie in diesem Zusammenhang gerne untersuchen?
- (b) Die bereitgestellte Datei enthält auch Befehle zur Generierung von Graphiken der vorhergesagten Werte gegen den linearen Prädiktor. Erstellen Sie diese Graphiken und vergleichen Sie diese. Wie beurteilen Sie die Ergebnisse der Signifikanztests für die Parameterschätzer im Vergleich zu den Graphiken?
- (c) In R gibt es mehrere Möglichkeiten die Zielvariable bei logistischer Regression einzugeben. Bisher wurde die Variante einer zweispaltigen Matrix, die die Anzahl der getöteten und der überlebenden Raupen beinhaltet, verwendet. Alternativ kann auch ein Vektor der beobachteten relativen Häufigkeiten verwendet werden. Die beiden Modelle liefern fast identische Ergebnisse. Lediglich der AIC-Wert ist stark unterschiedlich. Was kann die Ursache hierzu sein? Welche Folgerung für die Verwendung von AIC ziehen Sie hieraus?
- (d) Statt eines logistischen Regressionsmodells kann auch ein Probit-Modell verwendet werden. Vergleichen Sie dieses Modell mit den bisherigen Modellen.
5. Ausreißerererkennung spielt bei der Analyse von Daten eine große Rolle. Welche Beobachtungen würden Sie in logistischen Regressionsmodellen als Ausreißer bezeichnen und mit welchen Werkzeugen und Methoden würden Sie versuchen, solche zu erkennen.