Verallg. lineare Modelle (Rg–2a) — Serie 2

1. In der untenstehenden Tabelle ist die Anzahl der an AIDS gestorbenen Personen in Australien von 1983 bis 1986 aufgeführt. Die 14 Werte umfassen jeweils Perioden von 3 Monaten.

Periode (`time`)	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Anzahl Todesfälle (`number`)	0	1	2	3	1	4	9	18	23	31	20	25	37	45

Im Folgenden soll der Zusammenhang zwischen der Anzahl Todesfälle und der Zeit studiert werden. Die Daten stehen Ihnen im Datensatz `aids.dat` zur Verfügung.

a) Formulieren Sie ein verallgemeinertes lineares Modell. Nennen Sie die Ziel- und die erklärende Variable. Wie ist die Zielvariable verteilt? Welche Linkfunktion ist geeignet?

b) Schätzen Sie das Modell mit R. Welchen Wert haben die geschätzten Koeffizienten? Was ist der Einfluss der erklärenden Variablen?

c) Untersuchen Sie die Residuen. Betrachten Sie vor allem den Termplot. Was fällt auf?

d) Verbessern Sie falls nötig das Modell.

2. Für die Kontrolle einer grossen Unternehmung wurden getätigte Transaktionen von zwei bestimmten Typen untersucht. Man ist an der aufgewendeten Zeit (resp. den Kosten) der Transaktionen interessiert. Das Data Frame `transaction.dat` enthält Daten von 261 Zweigstellen:

`Time`	Total-Zeit (in Minuten), die für die Transaktionen gebraucht wurde
`Type1`	Anzahl Transaktionen vom Typ 1
`Type2`	Anzahl Transaktionen vom Typ 2

a) Lesen Sie die Daten ein mit `read.table("http://.../transaction.dat",header=TRUE)` und schauen Sie sie zuerst an, z.B. mit `pairs`().

b) Passen Sie ein (gewöhnliches) lineares Modell mit normalverteilten Fehlern an. Überprüfen Sie die Residuen. Weshalb passt dieses Modell nicht?

R-Hinweise:

```
> r.lm <- lm(Time ~ Type1 + Type2, data=d.trans)
> plot(r.lm, panel=panel.smooth)
```

c) Passen Sie nun mit `glm` ein verallgemeinertes lineares Modell mit der Gamma-Verteilung an. Formulieren Sie das verallgemeinerte Modell. Welche Link-Funktion ist hier angebracht? Probieren Sie den kanonischen Link und den Identitätslink (`link=identity`) aus und begründen Sie, warum der zweite vorzuziehen ist.

R-Hinweise:

```
> glm(Time ~ Type1 + Type2, family=Gamma(link=...), data=d.trans)
```

3. (freiwillig) Zeigen Sie, dass die folgenden Verteilungen zur Exponentialfamilie gehören (Skript, Kap. 2.2.c). D. h. zeigen Sie, dass sich die Wahrscheinlichkeit $P\langle Y = y\rangle$ resp. die Dichte $f\langle y\rangle$ dieser Verteilungen schreiben lässt als

$$\exp\langle (y\theta - b\langle\theta\rangle)/\phi + c\langle y;\phi\rangle\rangle\,.$$

Bestimmen Sie zu diesem Zweck θ, ϕ und die Funktionen $b\langle\theta\rangle$ und $c\langle y;\phi\rangle$ der

a) Poisson-Verteilung: $P_\lambda\langle Y = y\rangle = \frac{\lambda^y}{y!}\, e^{-\lambda}$.

b) Gamma-Verteilung:

$$f_{\eta,\sigma}\langle y\rangle = \frac{1}{\Gamma\langle\eta\rangle\sigma}\,(y/\sigma)^{\eta-1} e^{\,-y/\sigma}\;,$$

für $y \geq 0$ und $\sigma > 0$, $\eta > 0$.
Hinweis: $\theta = -1/\eta\sigma$