Angewandte Regression — Serie 9

1. Für die Bewirtschaftung von Hartholz-Wäldern in Australien besteht der Anspruch, sowohl den Umweltschutz wie auch die Holzproduktion zu berücksichtigen. Diese Forderungen legen nahe, dass unterschiedliche Formen der Bewirtschaftung in den verschiedenen Regionen zum Zuge kommen sollen. Um Gebiete identifizieren zu können, die sich für eine bestimmte Bewirtschaftungsart eignen, müssen zuerst die Faktoren, welche die Vielfalt von Pflanzen und Tieren beeinflussen, studiert werden. Einige dieser Variablen, welche in 151 verschiedenen Gebieten von 3 ha Grösse beobachtet wurden, sind im Datensatz `species.dat` enthalten.

`Diversity`	Anzahl verschiedene Arten Beutelratten (*possums*)
`Shrubs`	Anzahl Sträucher
`Stumps`	Anzahl Baumstrunke
`Stags`	Anzahl hohle Baumstämme
`Bark`	Index für die Menge abgenagter Baumrinden
`Habitat`	Index für die Eignung des Lebensraumes
`BAcacia`	Grundfläche von Akazien
`Eucalyptus`	Eukalyptus-Art (r=*regnans*, d=*delegatensis*, n=*nitens*)
`Aspect`	Ausrichtung des Gebietes (SW-NW, SE-SW, NW-NE, NW-SE)

Wir wollen den Zusammenhang zwischen der Anzahl Arten von Beutelratten Y_i und den anderen Variablen modellieren.

a) Verschaffen Sie sich zuerst einen Überblick über die Daten.

R-Hinweise: `par(mfrow=c(2,4)); plot(Diversity ~ ., data=d.species)`

b) Welches Modell scheint Ihnen hier sinnvoll? Wie lautet die Verteilung der Y_i und welche Link-Funktion(en) ist/sind geeignet?

c) Passen Sie nun das Modell mit `glm()` oder `regr()` an und kommentieren Sie das Resultat. Sind einige/alle/keine erklärenden Variablen relevant? Führen Sie die entsprechenden Tests durch.

R-Hinweise:

```
r.p <- glm(Diversity ~ ., family=poisson, data=d.species)          # oder
r.p <- regr(Diversity ~ ., family="poisson", data=d.species)
summary(r.p)
```

Die Quantile der χ^2-Verteilung lassen sich mit `qchisq()` berechnen. Für die Überprüfung, ob die einzelnen erklärenden Variablen das Modell wesentlich beeinflussen, gibt es zum Beispiel die R-Funktion `drop1()`.

d) Überprüfen Sie die Residuen.

R-Hinweise: `TA.plot()`, `termplot(..., partial=T)`, `plot(`*`regr-Objekt`*`)`

e) Vereinfachen Sie das Modell, indem Sie nicht-signifikante Variablen schrittweise weglassen.

R-Hinweise: `step()`

2. In der Nicht-Lebensversicherung, zb Haftpflicht, ist die Berechnungen des Gesamtaufwandes der Schäden eines Schadenjahres (Accident Year) ein zentrales Thema, da die Zahlungen der Schäden sich auf mehrere Jahre (Developement Year) verteilen können. Diese Zahlungen werden in einem Dreieck dargestellt.

```
         Development Year (DevYear)
AccYear  0       1       2      3      4       5      6    7     8     9
      0  5946975 3721237 895717 207760 206704  62124 65813 14850 11130 15813
      1  6346756 3246406 723222 151797  67824  36603 52752 11186 11646
      2  6269090 2976223 847053 262768 152703  65444 53545  8924
      3  5863015 2683224 722532 190653 132976  88340 43329
      4  5778885 2745229 653894 273395 230288 105224
      5  6184793 2828338 572765 244899 104957
      6  5600184 2893207 563114 225517
      7  5288066 2440103 528043
      8  5290793 2357936
      9  5675568
```

Lese-Beispiel: im Schadenjahr= 4 sind
im Jahr = 4 (Developement Year= 0) 5778885 Geldeinheiten bezahlt worden,
im Jahr = 5 (Developement Year= 1) 2745229 Geldeinheiten bezahlt worden,
im Jahr = 6 (Developement Year= 2) 653894 Geldeinheiten bezahlt worden,
im Jahr = 7 (Developement Year= 3) 273395 Geldeinheiten bezahlt worden,
im Jahr = 8 (Developement Year= 4) 230288 Geldeinheiten bezahlt worden,
im Jahr = 9 (Developement Year= 5) 105224 Geldeinheiten bezahlt worden.
Bemerkung: die Summe der Diagonale im Dreieck entspricht der totalen Zahlungen in demselben Jahr.

Wir möchten nun das Dreieck zu einem Viereck ergänzen. Dafür benützen wir verschiedene in der Praxis gebräuchlichen GLM-Modelle. Die Daten sind in `RunOff.dat` zu finden.

Bemerkung: Was wir hier *kurz* diskutieren, ist im Aktuariat in der Tat ein resp mehrere full-time job. Es ist auch nicht der Sinn dieser Übung alles vollständig zu machen, sondern einen Einblick in eine andere Branche zu erhalten.

Quelle: V. Wüthrich and Michael Merz, *Stochastic Claims Reserving Methods in Insurance*, Wiley, New York, p. 201-232 .

a) Schauen Sie sich die Daten im Scatterplot an. Was fällt auf?

b) Machen Sie die folgenden Aufgaben für die untenstehenden Modelle:

- Modelle für `AccYear` und `DevYear` als Faktoren und als numerische Werte;
- Modelle für untransformierte und transformierte Daten (sofern möglich und sinnvoll);
- Residuenanalyse für die Modelle;
- Schreiben Sie Ihr favorisiertes Modell auf;
- Welche Faktoren sind auf dem 5%-Niveau signifikant? Beobachtung?
- GLM-Modelle für verschiedene Linkfunktionen (weshalb ist die log-Linkfunktionen so beliebt?) - siehe dazu Tabelle 12.3 im R-Skript;
- Berechnen Sie die Zahlungen der restlichen Developement Years für das Accident Year= 9.

Modell 1 (lineare Regression): Machen Sie die gewöhnliche lineare Regression mit der first-Aid-Funktion log für die Zielvariable.

Bemerkung: da wir die Zahlungen logarithmiert haben, nennt man dieses Modell auch das Lognormal-Modell.

Modell 2 (GLM-Gamma): Eine wichtige Verteilung die zur Modellierung von Schadenaufwand benutzt wird, ist die Gamma Verteilung. Die Gamma Verteilung gehört auch zur Exponentialfamilie:

Verteilung	kanonischer Link	Varianz-Funktion
Gamma	$1/\mu$	μ^2

R-Hinweise:

```
r.p <- glm(... , family=Gamma, data=...)                    # oder
r.p <- regr(... , family=Gamma, data=...)
                                          # mit verschieden Linkfunktionen
r.p <- glm(... , family=Gamma(link=log), data=...)          # oder
r.p <- regr(... , family=Gamma(link=log), data=...)
```

Modell 3 (GLM-Poisson): Eine weit verbreitete Modellierung ist mit der Poisson-Verteilung - sie gehört auch zur Exponential Familie:

Verteilung	kanonischer Link	Varianz-Funktion
Poisson	$\log(\mu)$	μ

Modell 4 (GLM-Tweedie): Eine ganze Familie von Verteilungen erhält man mit einem zusätzlichen Parameter p:

Verteilung	kanonischer Link	Varianz-Funktion
Tweedie	μ^{1-p}	μ^p

Um diese Familie anwenden zu können, brauchen Sie `tweedief.R`, das Sie mit `source("ftp://stat.ethz.ch/WBL/Source-WBL-2/R/tweedief.R"))` laden können. Diese R-Funktion finden Sie im Package `statmod`, das Sie zu Hause installieren können - siehe `r-project.org`. Machen Sie Ihre Modelle mit verschiedenen p's (zb $p = 1.2$, $p = 1.5$).

R-Hinweise:

```
r.p <- glm(... , family=tweedie(var.power=p ,link.power=1-p)), data=...)
```

Bemerkung: Diese Verteilungsfamilie enthält die Normalverteilung ($p = 0$), Poisson($p = 1$) und Gamma($p = 2$) als Spezialfälle.

3. (Fakultativ) 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 das Modell, falls nötig.