Angewandte Regression — Serie 8

1. Das Data Frame `heart.dat` enthält die nach Alter sortierten Daten von 99 Personen. Für jede Altersgruppe age_k ist die totale Anzahl Personen (m_k) und die Anzahl Personen $\tilde{y}_k$ mit Symptomen einer Herzkrankheit gegeben.

a) Schauen Sie sich die Daten $\tilde{y}/m \sim age$ an.

b) Schätzen Sie die Parameter einer einfachen logistischen Regression, welche die Wahrscheinlichkeit, Symptome zu zeigen, mit dem Alter in Beziehung setzt. Testen Sie die Hypothese, dass das Alter (`age`) die Wahrscheinlichkeit, Symptome zu zeigen, beeinflusst.

R-Hinweise:
Falls ($\tilde{Y}_k \sim \mathcal{B}(m_k, \pi_k)$) ist mit $m_k > 1$, wird die Zielvariable als Matrix (mit 2 Spalten) eingegeben, wobei in der ersten Spalte die Anzahl "Erfolge" ($\tilde{Y}_k$) und in der zweiten Spalte die Anzahl "Misserfolge" ($m_k - \tilde{Y}_k$) stehen. Sie können die logistische Regression entweder mit der Funktion `glm()` rechnen, oder mit der Funktion `regr()`

```
> r.glm <- glm(cbind(y, m-y)~age, family=binomial, data=d.heart)
> r.glm.regr <- regr(cbind(y,m-y)~age, data=d.heart, method="glm",
                     family="binomial")
```

c) Überprüfen Sie die Residuen. Für den Tukey-Anscombe Plot können Sie die Funktion `TA.plot()` benutzen. Diese befindet sich im Package `sfsmisc`, das zuerst mit `library(sfsmisc)` geladen werden muss.

R-Hinweise:

```
> TA.plot(r.glm.regr, res=..., labels="*", show.call=F)
> termplot(r.glm.regr, partial=TRUE, rug=TRUE)
```

Bemerkung: Wenn Sie mit `regr()` arbeiten, erhalten Sie mit `plot(r.glm.regr)` nicht nur die beiden oben beschriebenen Plots frei Haus, sondern zusätzlich noch den Leverage-Plot.

d) Zeichnen Sie die logistische Regressions-Kurve und schätzen Sie das Alter, bei welchem Sie erwarten würden, dass 10%, 20%, ..., 90% der Personen Symptome zeigen. Diskutieren Sie diese Resultate.

R-Hinweise:
Verwenden Sie die Funktion `predict()` zur Bestimmung der erwarteten Wahrscheinlichkeiten:

```
> r.pred <- predict(r.glm.regr, newdata=data.frame(age=0:100),
  type="response")
```

Die tatsächlich beobachteten Werte kann man direkt aus dem Dataframe holen und zeichnen:

```
> plot(d.heart$age, d.heart$y/d.heart$m, xlim=c(0,100), ylim=c(0,1))
```

In diesen Plot können die erwarteten Werte mit `lines(..., r.pred)` dann eingezeichnet werden.

Quelle: D.W. Hosmer and S. Lemeshow (1989), *Applied Logistic Regression*, Wiley, New York, p. 3.

2. 38 Käfer der Arten *Haltica Oleracea* und *Haltica Carduorum* wurden gefangen und ausgemessen. Die Variablen im Data Frame `floh.dat` sind:

`x1`	Abstand der Transversalrille vom hinteren Rand des Prothorax
`x2`	Länge der Flügeldecke
`unt`	Code für den Untersucher (0 oder 1)
`art`	Art des Käfers (0 = *Oleracea*, 1 = *Carduorum*)

Wir betrachten `art` als die abhängige Variable und versuchen, die Wahrscheinlichkeit eines Käfers, zur Art *Carduorum* (`art` = 1) zu gehören, als Funktion der erklärenden Variablen `x1` und `x2` zu schätzen.
Das gefundene logistische Modell kann anschliessend dazu benützt werden, neue ausgemessene Käfer zu klassieren. Falls die Wahrscheinlichkeit, zur Art *Carduorum* zu gehören, grösser als 0.5 ist, wird der Käfer dieser Art zugeordnet.

Bemerkung: Diese Fragestellung wird üblicherweise mit einer Diskriminanz- oder Identifikationsanalyse gelöst, wie wir es im Block Multivariate Statistik I gemacht haben. Die logistische Regressionsrechnung kann aber ebenfalls benützt werden.

a) Betrachten Sie das Streudiagramm von `x2` gegen `x1`. Unterscheiden sich die beiden Arten?

R-Hinweise:
```
> d.floh <- read.table("http://.../floh.dat", header=T)
> plot(d.floh$x1, d.floh$x2, type="n")
> text(d.floh$x1, d.floh$x2, labels=d.floh$art, col=d.floh$art+1)
```

b) Formulieren Sie ein logistisches Regressionsmodell für die Wahrscheinlichkeit eines Käfers, zur Art *Carduorum* zu gehören, unter Einbezug der Variablen `unt`, um mögliche Untersuchereffekte zu berücksichtigen.

c) Schätzen Sie die Parameter des Modells, und prüfen Sie (auf dem 5%-Niveau), ob die Untersucher-Variable `unt` im Modell benötigt wird. Reduzieren Sie falls nötig das Modell.

R-Hinweise:
```
> glm(art ~ ..., family=binomial, data=d.floh)
```
Für einen allfälligen Modellvergleich:
```
> 1 - pchisq(..., ...)
```
oder direkt mit:
```
> anova(r.glm, r.glm2, test="Chi")
```

d) Untersuchen Sie, ob die Abhängigkeit der Zielgrösse von den erklärenden Grössen linear ist.

R-Hinweise:
```
> termplot(..., partial=TRUE)
```

e) Welcher Art würden Sie einen Käfer mit `x1` = 197 und `x2` = 303 zuordnen? Schätzen Sie die Wahrscheinlichkeit, dass dieser Käfer zur Art *Carduorum* gehört.