Angewandte Regression — Serie 5

1. In einer Untersuchung an 78 Schülern und Schülerinnen wurde der Zusammenhang zwischen schulischen Leistungen, Intelligenzquotienten und der Selbsteinschätzung (erfasst mit einem psychologischen Test) studiert. Die Datei
`http://stat.ethz.ch/Teaching/Datasets/WBL/concept.dat`
enthält die folgenden Angaben:

gpa	Punktezahl in einem Schultest
iq	IQ–Test
alter	Alter in Jahren
sex	1=weiblich, 2=männlich
total	Gesamtscore im psycholog. Test
c1	Teilscore "Verhalten"
c2	Teilscore "Status"
c3	Teilscore "Aussehen"
c4	Teilscore "Ängstlichkeit"
c5	Teilscore "Beliebtheit"
c6	Teilscore "Zufriedenheit"

a) Es ist anzunehmen, dass der Intelligenzquotient IQ positiv mit dem Schulerfolg korreliert. Welche weiteren Variablen haben einen signifikanten Zusammmenhang mit dem GPA? (Berechnen Sie für die Variable `gpa` ein Modell, das alle erklärenden Variablen enthält)

b) Führen Sie für das Modell in a) eine Rückwärts-Elimination durch.
R-Hinweis: Benützen Sie die Funktion `step()`.

c) Untersuchen Sie die Residuen. Gibt es Ausreisser? Sind Transformationen nötig?

d) Stellen Sie die Residuen gemeinsam mit zwei Eingangsvariablen dar mit Hilfe der Funtion `plres2x`. Mit einiger Geduld können Sie das mit allen Paaren von (bedeutenden) Eingangsvariablen machen.
Gibt es Hinweise auf Wechselwirkungen?

e) Prüfen Sie numerisch, ob es sich lohnt, im reduzierten Modell quadratische Terme oder Wechselwirkungen einzufügen. **R-Hinweis:** Benützen Sie die Funktion `add1()`. Sie prüft die gestellte Frage für `regr`-Objekte direkt – allerdings nur, wenn Sie die neue Version der Funktion `drop1.regr` zur Verfügung haben, was Sie mit
`source("ftp://stat.ethz.ch/WBL/Source-WBL-2/R/drop1.regr.R")` erreichen.

f) Wiederholen Sie gegebenenfalls die Schritte **a)**, **b)** und **c)**.

2. Wir fahren mit dem Datensatz `asphalt` von der Serie 4 fort.

`RUT`	Abnutzung des Belags in inches pro 1 Mio. Räder
`VISC`	Viskosität des Asphalts
`ASPH`	Anteil des Asphalts im Oberflächenbelag (in %)
`BASE`	Anteil des Asphalts im Unterbelag (in %)
`FINES`	Anteil der Feinteile im Oberflächenbelag (in %)
`VOIDS`	Anteil der Hohlräume im Oberflächenbelag (in %)
`RUN`	Indikatorvariable, welche die zwei Versuchsreihen unterscheidet

Quelle: R.V. Hogg and J. Ledolter, Applied Statistics for Engeneers and Physical Scientists, Maxwell Macmillan International Editions, 1992, p.393

a) Vereinfachen Sie das Modell mit dem Backward-Verfahren. Welches Modell resultiert? Führen Sie das Forward-Verfahren durch. Welches Modell resultiert hier?

b) Welches Siegermodell (kleinster Cp-Wert) liefert das "All Subsets"-Verfahren?

R-Hinweis zum Forward-Verfahren:

```
r.start <- regr(Zielvariable ~ 1, data=d.asphalt)
step(r.start, scope=Formel), direction="forward")
```

R-Hinweis zum "All Subsets"-Verfahren:

```
library(leaps)
r.allsub <- regsubsets(formula(regr-Objekt), data=d.asphalt, nbest=2)
                                  # von jeder Modellgrösse werden nur die zwei besten aufgelistet
summary(r.allsub)                 # zeigt mit "*" die Variablen im Modell
summary(r.allsub)$cp              # gibt die zugehörigen Cp-Werte an
```

3. Der Datensatz `mort` enthält Daten über die jährliche Sterbewahrscheinlichkeit der schwedischen Bevölkerung in den Jahren 1951-2005.

`Year`	Beobachtungsjahr
`Age`	Alter
`DeathProz`	empirische Sterbewahrscheinlichkeit in einem Jahr
`Gender`	Geschlecht (0 =Mann,1 =Frau)

In den Lebensversicherungen werden diese Tafeln zur Bestimmung von Prämien benutzt.

Quelle: P. De Jong, G. Z. Heller, Generlized Linear Models for Insurance Data

a) Betrachten Sie die Sterbewahrscheinlichkeit, log-transformiert und nicht transformiert, in Abhängigkeit des Alters für die Jahre 1951 und 2005. Benützen Sie verschiedene Farben für das Geschlecht. Beobachtung? Wo wird man Schwierigkeiten im Modellieren erhalten?

b) Wir bezeichnen mit q_x die jährliche Sterbewahrscheinlichkeit im Alter x. In der aktuariellen Welt werden häufig lineare Modelle für q_x oder $log(q_x)$ (und auch Kombinationen davon, die wir hier aber in der linearen Regression nicht behandeln) untersucht. Diese Modelle sind unter dem Namen Gompertz-Makeham-Class bekannt. Machen Sie zu den beiden Modellen lineare Regressionen, mit Transformationen, mit Interaktionen und quadratischen (und allenfalls höheren) Termen. Machen Sie dazu Residuenanalysen. Wählen Sie Ihr bestes Modell und notieren Sie Ihre Beobachtungen.