Regression I — Serie 5

1. Der Datensatz `asphalt` enthält Daten zur Abnutzung von Strassenbelägen. In einem Experiment wurden 31 verschiedene Strassenbeläge gemischt. Als Zielvariable wurde die Tiefe der Radspur in inches gemessen, nachdem eine Million Räder über den Belag gefahren waren. Die Variablen sind:

`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) Betrachten Sie die Streudiagramme von jeder Variablen gegen jede (Scatterplot-Matrix). Was fällt Ihnen auf? Welche Variablen sollten transformiert werden?
Benützen Sie für die weiteren Aufgaben die transformierten Variablen.

R-Hinweis:
Benutzen Sie den R-Befehl `pairs` um die Scatterplot-Matrix zu erzeugen. Um festzustellen ob die zwei Versuchsreihen sich unterscheiden, kann man die Punkte die zum ersten `RUN` gehören rot einfärben und die anderen blau. Dies kann mit dem folgenden R-Code erreicht werden:

```
t.index <- d.asphalt$RUN == 0
t.color <- rep("blue", dim(d.asphalt)[1])
t.color[t.index] <- "red"
pairs(d.asphalt,col = t.color)
```

b) Suchen Sie mit verschiedenen schrittweisen Verfahren (vorwärts, rückwärts, all subsets) ein Modell. Vergleichen Sie die Resultate.

R Anleitung für das Backward-Verfahren:

```
> r.asphalt <- lm("Volles Modell", data = d.asphalt)
> summary(step(r.asphalt, direction = "backward"))
```

R Anleitung für das Vorwärts-Verfahren:

```
> r.start <- lm("Zielvariable" ~ 1, data = d.asphalt) # Das Start-Modell ist Y = α + E
step(r.start, scope = list(upper="Volles Modell", direction="forward"))
```

R Anleitung für das All Subsets-Verfahren:

```
library(leaps)                    # damit wir Funktion regsubsets brauchen können
r.sub.as <- regsubsets("Modell", data = d.asphalt, nbest = 2)
> r.cp <- summary(r.sub.as)$cp                                # Cp-Werte der Modelle
```

2. Im Skript (Abschnitt 1.1.g) ist das Beispiel der basischen Böden beschrieben. Die Daten, die gegenüber dem Skript leicht geändert wurden, stehen im Dataframe `Basisch` zur Verfügung. Es zeigt sich, dass es sinnvoll ist, die quadrierte Höhe (Variable `h.quad`) als Zielvariable zu benützen. Die beiden erklärenden Variablen sind `ph` und `l.sar` = log(SAR), (SAR=sodium absorption ratio).

Es soll untersucht werden, ob die Messung der zusätzlichen Grösse SAR "etwas bringt".

a) Testen Sie, ob der Koeffizient von `l.sar` sich signifikant von 0 unterscheidet!

b) Ist die Varianz der Fehler konstant? Sind die Fehler normalverteilt?

c) Gibt es einflussreiche Beobachtungen?

R-Hinweis:

```
> r.infl <- lm.influence(name des fits)        # r.infl$hat enthält die Leverage H_ii
> plot(r.infl$hat, resid(name des fits))       # Residuen gegen Leverage H_ii
```

Oder:

```
> source("ftp://stat.ethz.ch/NDK/Source-NDK-9/R/regr.R")     # R-Funktionen von Werner Stahel
> regr...
> plot(name des regr-Objektes)
```

d) Ändern Sie, wenn nötig, das Modell oder die Daten. Untersuchen Sie wieder die Residuen.