Serie 4

1. Der Datensatz `asphalt.dat` 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-Anleitung:

```
> asphalt<-read.table.url(
          ``http://stat.ethz.ch/Teaching/Datasets/asphalt.dat'',header=T)

> pairs(...,pch=20)                    # 'pch=' zeichnet den 'character' 20 in den Plot
```

(andere Grafikparameter siehe ?par)

b) Suchen Sie mit verschiedenen schrittweisen Verfahren (vorwärts, rückwärts) ein Modell. Vergleichen Sie die Resultate. Welches Modell würden Sie wählen? Kommentieren Sie das Ergebnis.
Bemerkung: Beide Verfahren benutzen den AIC zur Elimination bzw. zur Aufnahme einer Variabeln aus bzw. in das Modell.

R-Anleitung für das Rückwärts-Verfahren:

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

R-Anleitung für das Vorwärts-Verfahren:

```
> start.lm <- lm(``Zielvariable'' ~ 1,data=asphalt)    # Das Start-Modell ist
```

$Y = \alpha + \epsilon$

```
> summary(step(start.lm, scope=list(upper= "Volles Modell"),
                  direction=``forward''))
```

2. Betrachten Sie das in Aufgabe 1 gefundene Modell und prüfen Sie mit einer Residuenanalyse ob

a) die Varianzen der Fehler gleich sind und

b) die Fehler normalverteilt sind.

c) Tragen Sie die Residuen gegen alle erklärenden Variablen auf. Fällt Ihnen etwas auf?

R-Anleitung:

```
> asphalt.mod <- lm(``Endmodell aus Aufg.1'', data=asphalt)
> plot(fitted(...), resid(...))                       # Tukey-Anscombe-Plot
> qqnorm(resid(...))                                  # Normalplot
> par(mfrow=c(2,2))                                   # 4 Bilder auf eine Seite
> plot(``erklärende Variable'', ``Residuen'')
```

3. Wir betrachten das Modell zum Vergleich zweier Gruppen:

$$y_i \;=\; \alpha + \beta \cdot 1_{[x_i=2]} + \epsilon_i$$

Dabei ist $x_i \in \{1,2\}$ ein Gruppenindikator, d.h. β gibt den Unterschied im Erwartungswert der beiden Gruppen. Wir nehmen an, dass die Fehler korreliert sind und zwar mit $\mathrm{Cov}(\epsilon_i, \epsilon_j) \;=\; 0.8^{|i-j|}$. Wir wollen die Standardabweichung von $\widehat{\beta}$ für 3 Methoden und 3 verschiedene Anordnungen der Gruppen vergleichen.

Methoden:

a) β geschätzt mit der gewöhnlicher kleinste Quadrate Methode, Korrelation (fälschlicherweise) ignoriert.

b) β geschätzt mit der gewöhnlicher kleinste Quadrate Methode, Korrelation berücksichtigt.

c) β geschätzt mit der verallgemeinerter kleinste Quadrate Methode.

Anordnungen:

i) $x_i = 1$ für $i \leq \frac{n}{2}$ und $x_i = 2$ für $i > \frac{n}{2}$.

ii) x_i in zufälliger Reihenfolge, d.h. eine der $\binom{n}{\frac{n}{2}}$ Reihenfolge wird zufällig ausgewählt.

iii) Zufällige Reihenfolge in jedem aufeinanderfolgenden Paar, d.h. $x_{2i-1} = 1$ und $x_{2i} = 2$ mit Wahrscheinlichkeit $\frac{1}{2}$ und $x_{2i-1} = 2$ und $x_{2i} = 2$ mit Wahrscheinlichkeit $\frac{1}{2}$ (unabhängig für die verschiedenen i's).

Die 3 Anordnungen mit n=30 sind im Datensatz `korrelation.dat` in den drei Kolonnen $X1$, $X2$ und $X3$ vorhanden.

Stellen Sie die Matrix des linearen Modells in den drei Fällen und die Kovarianzmatrix Σ auf und berechnen Sie die Standardabweichung von $\widehat{\beta}$ mit R.

R-Anleitung:

```
> korrelation<-read.table.url(''../korrelation.dat'',header=T)
> eins <- rep(1,30)                  # erzeugt Vektor (1,...,1) der Länge 30
> a-1          # zieht von jeder Komponente des Vektors a 1 ab: z.B. (1,1,2)-1 = (0,0,1)
> cbind(a,b)   # Bildet aus Vektoren a & b eine Matrix durch kolonnenweises zusammenbinden
> X%*%Y                              # Multiplikation der Matrizen X und Y
> t(X)                               # Transponierte Matrix von X
> solve(X)                           # Inverse Matrix von X
> outer(1:30,1:30,FUN=''-'')         # Berechnet 30×30 Matrix deren (i,j)ter Eintrag i-j ist
> sigma <- 0.8^(abs(outer(1:30,1:30,FUN=''-'')))           # Berechnung von Σ
```

Vorbesprechung : Freitag 26.5. 13.15 im HG D 1.1.

Abgabe: Mittwoch, 7. Juni 2000 vor der Vorlesung.

Präsenz: Jeweils Donnerstag, 12.00 bis 13.00 Uhr im LEO C12.1, Leonhardstr. 27, oder nach Vereinbarung: Marcel Wolbers (`wolbers@stat.math.ethz.ch`), LEO C14, Tel. 632 22 52 und Isabelle Flückiger (`isabelle@stat.math.ethz.ch`), LEO C13, Tel. 632 42 76.