Serie 6

1. In Serie 4, Aufgabe 2,wurde festgestellt, dass die Streuung der Residuen und damit ungefähr auch der zufällige Fehler E im Modell $W = \alpha v + \beta v^2 + E$ von v abhängig ist. In erster Näherung gilt $\mathrm{Var}(E_i) = c \cdot v_i^2$.

a) Der obige Mangel (Streuung proportional zu v) lässt sich durch gewichtete Regression beheben. Wie lautet das Modell (unter $\mathrm{Var}(E_i) = c \cdot v_i^2$), wenn wir $Z_i := W_i/v_i$ setzen?
Bem: Die allgemeine Situation sei $\mathrm{Cov}(E) = \sigma^2 \Sigma$ (mit bekannter Matrix Σ, hier $\Sigma = \mathrm{diag}(v_1^2, \ldots, v_n^2)$), und das transformierte (gewichtete) Modell gilt für $Z := \Sigma^{-\frac{1}{2}} W$.
Führe die Rechnung für dieses lineare Modell für Z mit S-plus durch.
Vergleiche $\hat\beta$ und die Schätzung für $\sigma(\hat\beta)$ mit denen, die man durch eine ungewichtete Regression für das Modell $W = \alpha v + \beta v^2 + E$ erhält.

b) Rechne zum Vergleich eine Regression mit Gewichten $\frac{1}{v^2}$.
S-plus-**Anleitung:** Benütze die Option `weights =` $\frac{1}{v^2}$ im Befehl `lm`.

c) Rechne auch eine Regression mit den Gewichten $\frac{1}{v}$.
Vergleiche $\hat\alpha$, $\hat\beta$ und die Schätzungen für $\sigma(\hat\alpha)$, $\sigma(\hat\beta)$ mit den in b) erhaltenen.

2. Eine Asphalt produzierende Firma möchte den Abnützungsgrad von Asphaltbelägen studieren. Die erhobenen Daten sind in der Reihenfolge $x1$, $x2$, $x3$, $x4$, $x5$, $x6$, y als Dataframe `d.asphalt` gegeben. Dabei ist y, die abhängige Variable, eine "mittlere Spurrinne" (in 'Zoll pro Million Raddurchgänge'). Die Variablen $x1$ und y sollen logarithmiert werden. Die Variable $x4$ ist eine sogenannte "Dummy"-Variable zur Unterscheidung von zwei verschiedenen Versuchsreihen.

a) Was fällt Dir bei diesen Daten auf, und was sollte besonders nachgeprüft werden bei der Auswertung? **S$^+$-Anleitung:** Funktion `pairs(..)` verwenden.

b) Vergleiche die zwei Versuchsreihen ($x4 = 0$ und $x4 = 1$), indem Du die Steigungen auf Gleichheit testest. Interpretation?
S$^+$-Anleitung: Die Funktion `summary.aov(`*Modell*`)` angewandt auf die Modelle

```
lm(log(y) ~ log(x1) + x2 + x3 + x4 + x5 + x6)
lm(log(y) ~ log(x1) + x2 + x3 + x5 + x6, subset= x4==0)
lm(log(y) ~ log(x1) + x2 + x3 + x5 + x6, subset= x4==1)
```

liefert die Grössen, aus denen man die Teststatistik berechnen kann.

c) Verwende zur Variablen-Auswahl einen $C(p)$-Search. Untersuche den $C(p)$-Plot und gib die besten Regressionsgleichungen zusammen mit R^2 an. Kommentar?
S$^+$-Anleitung: Erstelle zuerst die transformierten Variablen:

```
cx <- cbind(log(x1),x2,x3,x4,x5,x6); cy <- log(y)
dimnames(cx) <- list(NULL,paste(1:6)).
```

Benütze die Funktion `leaps`: `allRg <- leaps(cx,cy,method='Cp')`. Dann:
`n.plot(allRg$size, allRg$Cp, nam=allRg$label, ylim=c(0,25)); abline(0,1)`.

Abgabe: Mo, 16. Dez. 96.
Präsenz, Hilfe: Im SOL F5 (Sonneggstr.33), gemäss Vereinbarung: Tel. (63) 2-3408 oder via e-mail To: `maechler@stat.math.ethz.ch`.

Neu: Übungsblatt via WWW in `ftp://ftp.stat.math.ethz.ch/Teaching/maechler/Stat-Meth/`. `96-6.ps` ist Serie 6 (WS 96/97) als "PostScript" Datei. Im "Dateimanager" öffnen, oder zum Drucker schicken!