Serie 5

1. An den Stellen Libby und Newgate des Kootenay Rivers wurde der Wasserdurchfluss im Monat Januar in den Jahren 1931–43 gemessen (`Newgate` sei die Zielvariable und `Libby` die Erklärende). Mit diesen Daten, die als Datensatz `fluss.dat` zur Verfügung stehen, vergleichen wir mehrere robuste Regressionmethoden, wobei wir deren Robustheit in Bezug auf die Auswirkungen der Veränderung des 4. Punktes (77.6,44.9) zu (40.0,44.9) einerseits und zu (77.6,32.0) andererseits untersuchen wollen.

a) Berechne in den drei Fällen die Schätzungen für den Achsenabschnitt α und die Steigung β mit KQ-, L_1- , Huber, Mallows-Typ- und "least median of squares" (LMS)-Regression.

b) Zeichne für die drei Positionen des Ausreissers die Geraden für alle sechs Regressions-Schätzer und diskutiere kurz die Resultate.

R-Anleitung:
Die verschiedenen Verfahren sind in verschiedenen R-"libraries":

- L_1-Regression: mittels dem Befehl `rq` aus der `library("quantreg")`
- Huber-Regression: mittels dem Befehl `rlm` aus der `library("MASS")`
- Mallows-Typ-Regression: nicht verfügbar; wir basteln uns den Schätzer wie unten angegeben selbst
- LMS-Regression: mittels dem Befehl `lmsreg` aus der `library("lqs")`

Die Syntax der Befehle ist jeweils analog wie der bekannte `lm`-Befehl (s. das Beispiel zur Mallows-Typ-Regression unten (das ebenfalls den `rlm`-Befehl verwendet) und die entsprechenden help-Seiten).
Die Werte der 4. Beobachtung lassen sich z.B. ändern mittels: `fluss[4,] <- c(40.0,44.9)`.

```
# Daten einlesen:
fluss <-read.table.url("http://stat.ethz.ch/Teaching/Datasets/fluss.dat",header=T)
# Libraries holen:
library("MASS"); library("lqs"); library("quantreg")
# Mallows-Typ-Regression:
x.loc.scale <- hubers(fluss$Libby)   #robust location and scale of x-Variable
weights.mal <- wt.huber((fluss$Libby-x.loc.scale$mu)/x.loc.scale$s)
fluss.mal <- rlm(fluss$Newgate ~ fluss$Libby, weights=weights.mal)
```

2. **a)** Bestimme den Grenzwert der asymptotischen Kovarianzmatrix des Huber-Schätzers für $c \to 0$ unter der Annahme, dass ε eine stetige positive Dichte hat.
Bemerkung: Man kann zeigen, dass dies tatsächlich die Kovarianzmatrix des L_1-Schätzers ist.

b) Vergleiche die Genauigkeit (asymptotische Kovarianz) des L_1- und des KQ-Schätzers in den folgenden zwei Fällen:

1) ε normalverteilt

2) ε t_3-verteilt

Hinweis: Die Dichte der t_3-Verteilung ist $f(x) = \frac{2}{\sqrt{3}\pi}(1+\frac{x^2}{3})^{-2}$; eine t_3-verteilte Z.V. X hat $\mathrm{Var}[X] = 3$.

3. Der Dampfverbrauch pro Monat `Steam` einer Fabrik soll als Funktion der Variablen „Betriebstage pro Monat“ (`Operating.Days`) und „Mittlere Aussentemperatur pro Monat“ (`Temperature`) beschrieben werden.

a) Führe eine Regressionsrechnung mit dem Kleinste Quadrate Schätzer und dem MM-Schätzer durch und vergleiche die Resultate.

R-Anleitung:

```
> D.steam <- read.table.url("http://stat.ethz.ch/Teaching/Datasets/
            dsteam.dat",header=T)
> library("lqs"); library("MASS")
> fit.MM <- rlm(Steam ~ Operating.Days+Temperature, data=D.steam, method="MM")
                                                          # MM-Schätzer
```

b) Führe eine Residuenanalyse durch: Trage die Residuen in beiden Fällen gegen die “fitted values” und die erklärenden Variablen auf. Was fällt auf?

c) Welche Beobachtungen beeinflussen die KQ-Parameter besonders stark? Betrachte die Cook's-Distanzen der einzelnen Beobachtungen.

R-Anleitung:

```
> plot(lm-Objekt)                          # Cook's distance ist der 4. Plot
```

d) Es gibt zwei Beobachtungen, deren Anzahl Betriebstage pro Monat extrem klein sind (Betriebsferien?). (Diese entsprechen den zwei Beobachtungen mit der grössten Cook's-Distanz.) Nehmen wir an, es haben Betriebsferien stattgefunden. Wir berücksichtigen diese besonderen Umstände mit einer Dummy-Variablen (`Working.Holidays`).

Schätzen Sie das Modell nochmals mit der robusten MM- und der Kleinste-Quadrate-Methode. Bestehen immer noch Unterschiede zwischen den beiden Lösungen?

e) Zeichnen Sie die verschiedenen Lösungen (“fitted values”) im Plot `Steam` vs Beobachtungsreihenfolge ein. Weil die Daten Monat für Monat erhoben wurden, entspricht die Reihenfolge dem Zeitverlauf.

R-Anleitung:

```
> matplot(cbind(D.steam$Steam, fitted(lm-Objekt), fitted(rlm-Objekt)), type="l",
            ylim=c(0,13), lty=c(1,2,4), ylab="")
> legend(10,2.7, c("Beobachtungen", "LS", "MM"), lty=c(1,2,4))
                                        # Legende, lty='line type'
```

Vorbesprechung : Freitag 9.6. 13.15 im HG D 1.1.

Abgabe: Freitag 16. 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.