Serie 3

1. In dieser Aufgabe wird der Datensatz `catheter.dat` benützt. Es handelt sich um Daten aus der Medizin. Die Variable `x1` bezeichnet die Grösse eines Patienten (in cm), `x2` das Gewicht (in kg) und `y` die optimale Länge eines Katheters (in cm), der für eine Herzoperation verwendet wird. Man möchte gerne die Katheter-Länge aus den Patienten-Daten schätzen.
Mache eine Regression von `y` auf `x1` und `x2`. Kommentiere den Output, insbesondere die verschiedenen Tests, die man ablesen kann.
Finde eine Erklärung für die beobachteten Phänomene und prüfe diese mit geeigneten Plots nach.

Datensatz einlesen:
```
catheter <- read.table.url(
                "http://stat.ethz.ch/Teaching/Datasets/catheter.dat", header=T)
```

2. **a)** $(Y, X_1, X2)$ sei ein dreidimensionaler Zufallsvektor. $\hat{Y} = \alpha + \beta X_1$ und $\hat{X}_2 = \gamma + \delta X_1$ seien die besten linearen Prognosen von Y bzw. $X2$ gestützt auf X_1 (vgl. Serie 1, Aufgabe 1). Zeige, dass die Korrelation der Fehler $Y - \hat{Y}$ und $X2 - \hat{X}_2$ gerade die partielle Korrelation von Y und X_2 bei festem X_1 ist.

b) Zeige, dass die beste lineare Prognose von Y gestützt auf X_1 und X_2 die Form $\bar{Y} = \hat{Y} + \theta(X_2 - \hat{X}_2)$ hat.

c) Analog kann man mit Daten $(y_i, x_{i,1}, x_{x,2})$ die Regression von y auf x_1 und x_2 mit Hilfe von einfachen Regressionen durchführen. Prüfe das mit dem Datensatz von Aufgabe 1 nach und erstelle insbesondere den Plot von $y_i - \hat{\alpha} - \hat{\beta} x_{i,1}$ gegen $x_{i,2} - \hat{\gamma} - \hat{\delta} x_{i,1}$. Dieser Plot wird auch "partial residual plot" genannt.

3. Der Datensatz von Forbes zeigt Messungen von Siedepunkt (in °F) und Luftdruck (in inches of mercury) an verschiedenen Orten in den Alpen. Die Daten stehen als Datensatz `forbes.dat` mit den Variablen `Temp` und `Press` zur Verfügung.

a) Zeichne für das Modell $\texttt{Press} = \alpha + \beta \texttt{Temp} + \epsilon$ die Residuen gegen die angepassten Werte des Luftdruckes und zeichnen Sie den Normalplot der Residuen. Kommentar?

b) Betrachte das exponentielle Modell $\texttt{Press} = \alpha \mathrm{e}^{\beta \texttt{Temp}} \mathrm{e}^{\epsilon}$. Linearisiere dieses Modell und führe eine lineare Regression durch. Erstelle anschliessend den Tukey-Anscombe-Plot sowie den Normalplot der Residuen. Kommentar?

c) Identifiziere und entferne den Ausreisser. Führe nun nochmals dasselbe durch wie in b).

R-Anleitung:
> `r.forbes <- lm(...,data=forbes)`
Die Residuen erhält man mittels `resid`(*r.forbes*), die angepassten Werte mittels `fitted`(*r.forbes*).
Identifizieren der Punkte im Tukey-Anscombe-Plot mittels `identify`(*Residuen,Fitted*), und dann mit der linken Maustaste in der Nähe des zu identifizierenden Punktes klicken. Beenden des Identifikationsmodus mittels Klicken der mittleren Maustaste im Grafikfenster. (Achtung: Solange der Identifikationsmodus aktiv ist, kann man im R nicht weiterfahren!)
Für den Normalplot: `qqnorm`(*Residuen*).
Beide Plots lassen sich auch mittels `plot`(*r.forbes*) kreiieren. Siehe die help zu `plot.lm`(.)

4. Während einer Woche im September 1993 wurde in der Südröhre des Gubrist-Tunnels nördlich von Zürich Messungen durchgeführt. Das Ziel ist es, Emissionsfaktor von NOx für die verschiedenen Fahrzeugtypen (Lastwagen und Nicht-Lastwagen) zu bestimmen.

Der Datensatz `gubrist-ueb.dat` enthält die folgenden Variablen

`PLKW`	Lastwagen-Anteil am Gesamtverkehr
`VFZ`	Fahrzeuggeschwindigkeit (in km/h)
`VLUFT`	Luftgeschwindigkeit im Tunnel (in m/s)
`NOX`	mittlerer Emissionsfaktor pro Fahrzeug von NOx (in mg/km)

Wir betrachten das Modell $\texttt{NOX}_i = \alpha + \beta \cdot \texttt{PLKW}_i + \epsilon_i$. (Der Index i bezeichnet eine Zeitperiode.)

a) Überlege, wie man aus den Koeffizienten α und β die Emmissionsfaktoren erhält. (Anders gefragt: Wie gross ist der Erwartungswert von `NOX` wenn der Lastwagen-Anteil 0 bzw. 1 wäre?)

b) Betrachte das Streudiagramm `NOX` gegen `PLKW`. Was fällt auf?
```
gubrist <- read.table.url("http://stat.ethz.ch/Teaching/
   gubrist-ueb.dat",na.strings = ".",header=T)  # Datensatz einlesen
```

c) Untersuchungen haben gezeigt, dass bei kleinem Verkehrsaufkommen die Schadstoffe nicht gleichmässig durch den Tunnel bewegt werden. Wir lassen deshalb Beobachtungen mit einer Luftgeschwindigkeit von weniger als 5 m/s in der Auswertung weg, indem wir einen neuen Dataframe ohne diese Daten erzeugen. Betrachte für die verbleibenden Daten nochmals das Streudiagramm `NOX` gegen `PLKW`.
```
vluft <- !(is.na(gubrist[,'VLUFT']))&(gubrist[,'VLUFT'] >= 5)
gubrist2 <- gubrist[vluft,]
```

d) Schätze die Koeffizienten im Modell $\texttt{NOX}_i = \alpha + \beta \cdot \texttt{PLKW}_i + E_i$ und zeichne die Regressionsgerade ins Streudiagramm ein. Wie gross sind die geschätzten Emissionsfaktoren für Lastwagen und Nicht-Lastwagen?
```
gubrist.lm ~ lm(...,data=gubrist2)
abline(gubrist.lm)    # Gerade in bestehenden Plot einzeichnen
```

e) Sind die Fehler normalverteilt und unabhängig (zeitlich)?
```
n <- length(resid(gubrist.lm))            # Anzahl Residuen
plot(1:n, resid(gubrist.lm), type=''l'')  # Residuen gegen Index (Zeit)
```

f) Würdest Du die Daten transformieren? Wenn ja, wie?

Vorbesprechung : Freitag 12. Mai 13.15 im HG D 1.1.

Abgabe: Mittwoch, 24. Mai 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.