Serie 12

1. In dieser Aufgabe betrachten wir 4 Datensätze die von Anscombe konstruiert wurden. In jedem der Datensätze gibt es eine Zielvariable Y und eine erklärende Variable X.

a) Stelle jeden der 4 Datensätze als Streudiagramm dar, zeichne die Regressionsgerade ein und kommentiere die Ergebnisse.

b) Vergleiche die Schätzungen von β_0, β_1 und σ^2, sowie das sogenannte "Gütemass" R^2, das später genauer besprochen wird.

R-Hinweise:

```
data(anscombe) ## Einlesen des Datensatzes
```

Die Regression kann man mit

```
reg <- lm(y1~x1, data = anscombe) oder
reg <- lm(anscombe$y1 ~ anscombe$x1)
summary(reg)
```

berechnen und numerisch auswerten. Mit `par(mfrow=c(2,2))` wird das Grafikfenster so eingeteilt, dass alle 4 Bilder nebeneinander passen. Den Scatterplot und die Regressionsgerade erhält man mit

```
plot(anscombe$x1, anscombe$y1)
abline(reg)
```

Die Schätzungen für die Koeffizienten β_0, β_1 und σ, sowie das Gütemass R^2 erhält man mit

```
summary(reg)
```

2. Der Datensatz von Forbes zeigt Messungen von Siedepunkt (in ^{o}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) Trage in einem Streudiagramm den Druck gegen die Temperatur auf. Macht es Sinn, diese Daten mit einer Regressionsgeraden zu modellieren?
R-Anleitung:

```
> forbes <- read.table("http://stat.ethz.ch/Teaching/Datasets/forbes.dat",
                       header=TRUE)
> par(mfrow = c(3,1)) # Ermoeglicht 3 Grafiken untereinander zu platzieren.
> plot(forbes[,"Temp"], forbes[,"Press"])
```

b) Berechne die Koeffizienten der Regressionsgeraden und trage die Regressionsgerade ins Streudiagramm ein.

```
> forbes.fit <- lm(Press ~ Temp, data = forbes) #Regression berechnen
> summary(forbes.fit) # Regressionsoutput zeigen
> abline(forbes.fit)  # Regressionsgerade einzeichnen
```

c) Zeichne den Tukey-Anscombe-Plot (Residuen gegen angepasste Werte) und den Normalplot der Residuen. Gibt es Hinweise, dass die Modellannahmen verletzt sind?

```
> plot(fitted(forbes.fit), resid(forbes.fit), main="Tukey-Anscombe Plot")
> abline(h=0)
> qqnorm(resid(forbes.fit))
```

d) Logarithmiere nun den Druck. Trage in einem Streudiagramm den logarithmierten Druck gegen die Temperatur auf, berechne die Regressionsgerade und trage sie ins Diagramm ein.

```
> forbes[,"Logpress"] <- log(forbes[,"Press"])
```

e) Zeichne wiederum den Tukey-Anscombe und den Normalplot. Wie steht es nun mit den Modellannahmen? Gibt es Ausreisser?

f) Identifiziere und entferne den Ausreisser. Berechne die Regressionsgerade neu und zeichne nochmals alle Plots. Sind jetzt die Modellvoraussetzungen erfüllt?
Ein Ausreisser ist eine Beobachtung, die nicht in das Modell passt (z.B. wegen Tippfehler). Ausreisser identifizieren mit Hilfe des Befehls identify: Dazu schliesse man zuerst alle Grafikfenster. Nach Ausführung des identify Befehls (wie unten beschrieben) mit der linken Maustaste auf den Ausreisser klicken, dann erscheint die Nummer des Ausreissers. R fährt nach dem identify Befehl erst weiter, wenn dieser mittels klicken der mittleren Maustaste in der Grafik beendet worden ist.

```
> plot(fitted(forbes.fit), resid(forbes.fit))
> identify(fitted(forbes.fit), resid(forbes.fit))
> forbes <- forbes[-..,] # Ausreisser entfernen: .. mit Beobachtungsnummer ersetzen
```

3. In der folgenden Tabelle stehen die Weltrekorde der Männer über 13 verschiedene Laufdistanzen, Stand 1974.

Distanz (m)	100	200	400	800	1000	1500	2000
Zeit (s)	9.9	19.8	43.8	103.7	136.0	213.1	296.2

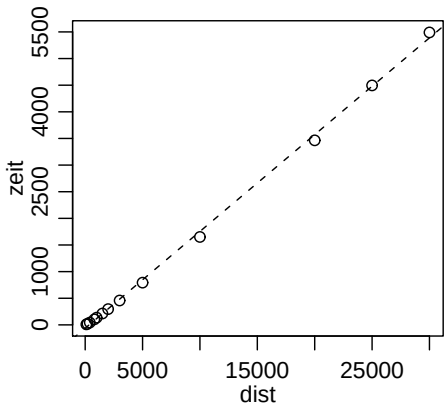
Distanz (m)	3000	5000	10000	20000	25000	30000
Zeit (s)	457.6	793.0	1650.8	3464.4	4495.6	5490.4

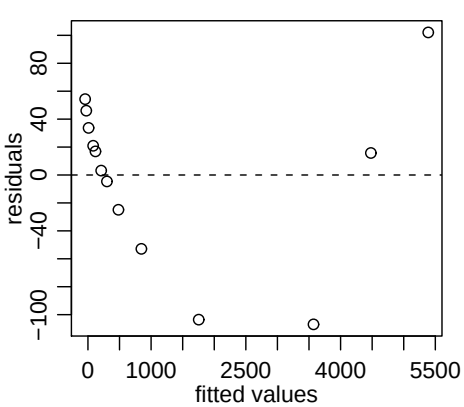
An diese Daten wurde folgendes Regressionsmodell angepasst:

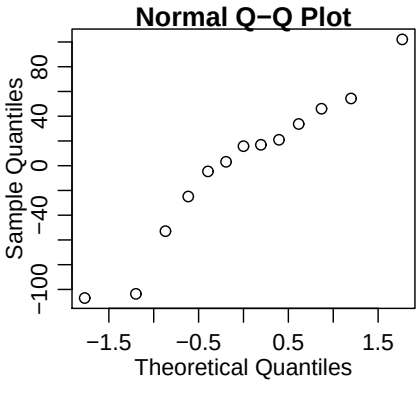
$$\text{Zeit}_i = \beta_0 + \beta_1 \cdot \text{Distanz}_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

Der Regressionsoutput und die Diagnoseplots sehen folgendermassen aus:

```
Call:
lm(formula = zeit ~ dist)

Residuals:
    Min      1Q  Median      3Q     Max
-106.95  -24.90   15.77   33.71  102.08

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -62.59296   21.81098   -2.87   0.0152 *
dist          0.18170    0.00173  105.05   <2e-16 ***

Residual standard error: 62.68 on 11 degrees of freedom
Multiple R-squared: 0.999,Adjusted R-squared: 0.9989
F-statistic: 1.103e+04 on 1 and 11 DF,  p-value: < 2.2e-16

Residuals:
    Min      1Q  Median      3Q     Max
-106.95  -24.90   15.77   33.71  102.08
```







a) Gibt es einen signifikanten Zusammenhang zwischen Distanz und Zeit, d.h. ist β_1 signifikant von 0 verschieden?

b) Eines der folgenden 4 Intervalle ist das 95%-Vertrauensintervall für β_1. Welches?
i) [0.1800, 0.1834] ii) [0.1779, 0.1855] iii) [0.1765, 0.1869] iv) [0.1800, 0.1852]

c) Wie gross ist das Residuum der 5. Beobachtung (1000m)?

d) Dürfen wir die berechnete Regressionsgerade benutzen, um zu schliessen, dass 1974 der Weltrekord über 100km (100000m) ungefähr bei 18000s gelegen wäre?

e) Wie gross ist die geschätzte Standardabweichung der Fehler E_i? Was heisst das für die Brauchbarkeit des Modells?

f) Was folgerst Du aus der Darstellung der Residuen gegen angepasste Werte?

g) Formuliere ein Modell, das vermutlich besser zu diesen Daten passen würde.

Besprechung: Donnerstag, 05. Dezember.

Abgabe: Die Übung kann auf freiwilliger Basis abgegeben werden - Bitte markieren Sie die Aufgaben, die korrigiert werden sollen.