Exercise Sheet 11

1. A hole is drilled into permafrost, and the temperature measured at the following depths: 0, 0.2, 0.5, 0.6, 0.8, 0.9, 1.2 and 6 m. In summer, these temperatures could be e.g. 6, 4.2, 0.6, -2.1, -5.2, -7.3, -8.9 and 15°C.
To help you, each part of this problem is supplemented by the appropriate R code.

R hints: replace each .. by the correct values!

a) Make a scatterplot of temperature against depth. What do you see? Give one (two) possible interpretation(s) of these data.

```
> depth <- c(..); temp <- c(..)
> plot(depth, temp, main="Scatterplot")
```

b) Compute the empirical correlation of the data without the outlier, and compare it to the empirical correlation of the full data set ($\rho = 0.6$).

```
> depthWO <- depth[-..];  tempWO <- temp[-..] # Omit one observation
> cor(..,..)
```

c) Fit a straight line to the data. (Estimate the intercept β_0 and slope β_1 of the line using the method of least squares, and plot the line in the scatter plot.) Then omit the outlier and fit the line using the remainder of the data.

```
> fit1 <- lm(temp ~ depth) # Fit a linear model
> fit2 <- lm(tempWO ~ depthWO) # Fit a linear model without the outlier
> ## Plot a dashed regression line for fit1
> abline(fit1,lty=2)
> abline(fit2) # Plot the regression line for fit2
```

d) Could the slope be 0? Test $H_0 : \beta_1 = 0$ against $H_A : \beta_1 \neq 0$, with and without the outlier.

```
> summary(fit1) # Plot the regression output for fit1
> summary(fit2)
```

An investigation of the relationship between traffic flow x (1000's of cars per 24 hours) and lead content y of bark on trees near the highway ($\mu g/g$ dry weight) yielded the data in the x_i and y_i columns below:

x_i	8.3	8.3	12.1	12.1	17.0	17.0	17.0	24.3	24.3	24.3	33.6
y_i	227	312	362	521	640	539	728	945	738	759	1263

The data have been made available in the file `strasse.dat`, containing the explanatory variable `verkehr` (traffic) and the target variable `blei` (lead).

2. **a)** Draw a scatter plot of these two variables.
Instructions for R:

```
> highway <- read.table("http://stat.ethz.ch/Teaching/Datasets/strasse.dat",
+                        header=TRUE)
> plot(highway[,"verkehr"], highway[,"blei"], main="Scatter plot")
```

b) Estimate the coefficients β_0 and β_1, as well as the error variance σ.

```
> highway.fit <- lm(blei ~ verkehr, data=highway) # Fit a linear model
> summary(highway.fit) # Show the linear model
> abline(highway.fit) # Add the regression line to the existing scatter plot
```

c) Does the slope β_1 differ significantly from zero?

d) Predict the lead concentration for the traffic intensity $x = 40$.

3. The following data give the income, number of cows and area for a number of American farms.

Income (Dollar)	960	830	1260	610	590	900	820	880	860	760
Number of cows (cows)	18	0	14	6	1	9	6	12	7	2
Size of farm (acres)	60	220	180	80	120	100	170	110	160	230

Income (Dollar)	1020	1080	960	700	800	1130	760	740	980	800
Number of cows (cows)	17	15	7	0	12	16	2	6	12	15
Size of farm (acres)	70	120	240	160	90	110	220	110	160	80

To these data, the linear regression model

$$Dollar_i = \beta_0 + \beta_1 cows_i + \beta_2 acres_i + E_i$$

with i.i.d. $E_i \sim \mathcal{N}(0, \sigma^2)$ was fitted
This is part of the output from R:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  285.457     81.379   3.508   0.0027 **
cows          32.569      3.728     ??? 1.08e-07 ***
acres          2.138      0.394   5.434 4.47e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 76.45 on ??? degrees of freedom
Multiple R-squared: 0.8179,        Adjusted R-squared: 0.7965
F-statistic: 38.17 on ??? and ??? DF,  p-value: 5.165e-07
```

1) The size of a farm has a statistically significant influence on its income.

a) True   b) False

2) The number of cows on a farm has a statistically significant influence on its income.

a) True   b) False

3) What is the outcome of the test of the null hypothesis $H_0 : \beta_2 = 0$ against the alternative $H_A : \beta_2 \neq 0$?

a) Keep H_0 b) Reject H_0

4) How many degrees of freedom are there in this model fit?

a) ∞ b) 20   c) 18   d) 17   e) 3

5) Which of the following is an exact 95% confidence interval for β_1?

    a) $32.569 \pm 2.11 \cdot 3.7276$ b) $32.569 \pm 1.96 \cdot 3.7276$

    c) $32.569 \pm \frac{2.11}{\sqrt{17}} \cdot 5.45$ d) None of these

6) How high an income would you predict for a 100-acre farm without cows?

    a) 213   b) 285

    c) 499   d) 325

7) In a simple linear regression model using the area of a farm as the only explanatory variable, would it (the area) have a significant influence on the income?

    a) Definitely   b) Definitely not

    c) It isn't clear

Question Hour:

1) Wednesday, 27.07.2011, 14:00 - 15:00, HG G 26.1.

2) Wednesday, 03.08.2011, 14:00 - 15:00, HG G 26.1.