Series 3

1. The file `gas.dat` contains the gas consumption (in kWh) and the differences of temperature (in ℃) inside and outside of 15 houses which are heated with gas. The measures were collected over a long time span and then averaged.

a) Read in the data from the internet using
`read.table("http://stat.ethz.ch/Teaching/Datasets/gas.dat", header = TRUE)`.
Hint: Alternatively the data can be downloaded from the web using a browser and read in from the local drive using also `read.table()`. This could be necessary if you get an error reading it directly (e.g. caused by a stringent firewall).
Illustrate the data graphically. What does the relation look like?

b) Fit a simple linear regression in order to explain gas consumption by temperature, using the R-command `mod1 <- lm(...)`. Compare the output when calling `mod1` and `summary(mod1)`.

c) Write down the fitted model equation.

d) According to the fitted model, what kind of consumption do you expect when the difference in temperature is 14℃ ?

e) What are the model assumptions? Check the model assumptions by performing residual analysis. Does it look satisfactory?
Hints:
`plot(fitted(mod1),resid(mod1)), abline(h=0)`,
`plot(gas$temp,resid(mod1)), abline(h=0)` and
`qqnorm(resid(mod1))`, `qqline(resid(mod1))`.
Or `plot(mod1)`, which generates directly the above plots and an additional one.

f) Which of the following statements are **not** correct and why?

(i) If the residuals vs. fitted values plot is fan-shaped, then the constant variance assumption is violated.

(ii) If the QQ-plot deviates from a straight line, then the normality assumption is violated.

(iii) If the normality assumption is not fulfilled, then the p-values and confidence intervals cannot be trusted.

(iv) Least square estimators of regression coefficients are not unbiased if the error distribution is not Gaussian.

(v) The R^2-value, produced by R, is wrong if at least one of the model assumptions is not correct.

2. The article "Characterization of Highway Runoff in Austin, Texas, Area" gave a scatter plot of x=rainfall volume and y=runoff volume for a particular location. The values are:

x	5	12	14	17	23	30	40	47	55	67	72	81	96	112	127
y	4	10	13	15	15	25	27	46	38	46	53	70	82	99	100

a) Produce a scatterplot of runoff volume vs. rainfall volume. Do you think a simple linear regression is plausible here.

b) Now fit a simple linear regression model. Use it for predicting the runoff volume when the rainfall volume takes the value 50. Also compute the 95% prediction interval for this case.

c) How much of the observed variation in runoff volume can be attributed to the simple linear association between runoff and rainfall volume?

d) Is there a significant linear association between runoff and rainfall volume? Moreover, use a statistical test to determine whether there is a 1:1 relation between runoff and rainfall. If no, why do you think it is not a 1:1 relation?

e) Produce a plot of the residuals vs. fitted values and a normal plot. If you inspect it very carefully, you can notice that some of the assumptions for simple linear regression are violated. Explicitly mention these.

f) Runoff and rainfall volume are both variables which can only take positive values. Both are skewed to the right, though only slightly so here. Taking logs on both variables could thus be beneficial. Fit a simple linear regression model for the transformed variables and compare the results with the ones from the initial model, i.e. repeat all the steps a)-e). Does the log-transformed model fit better?

3. (Continuation of exercise 2 on sheet 2) Various data on cars and their fuel consumption is stored in the data set `my.mtcars.rda`. In this exercise, we will continue to look at the connection between the engine power (variable `hp`) and fuel consumption (variable `l.100km`).

a) Perform a log transformation of both variables and do a regression analysis (fit and plot the model, perform residual diagnostics). How does it compare with the original model without the transformation?

b) The linear regression model for the transformed variables is

$$\log(\texttt{l.100km}) = \beta_0 + \beta_1 \log(\texttt{hp}) + \epsilon$$

Express this as a relation between the original variables `l.100km` and `hp`.

c) Plot the model curve of the log-transformed model in a scatter plot of the original model.

Preliminary discussion: Monday, October 13.

Deadline: Monday, October 20.