Series 3

1. 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. Also try to give a guesstimate of R^2.

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 assoociation 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) Do a plot of 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 here.
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). Which of the two fits better?

g) There is another important advantage of the model in f). It's easy to see what it is if you plot the 95% prediction interval as a prediction band for the two models, both on the original scale.

2. The file `farm.dat` contains the size A (in acres), the number of cows C and the income I (in $) of 20 farms in the US. Read in the data from the web using
`read.table("http://stat.ethz.ch/Teaching/Datasets/farm.dat", header = TRUE)`.

a) Compute an ordinary linear regression of I versus C. Does the income depend on the number of cows?

b) Give the confidence intervals for the expected income without any cows, with 20 cows and with $\bar{C} = 8.85$ cows. Give also a prediction interval for the income of a farm having no cows or with $\bar{C} = 8.85$ cows.

c) Compute an ordinary linear regression of I versus A and of C versus A and also a multiple linear regression of I versus A and C.
Explain the differences between the three results.

3. The data in teengamb.rda come from a study on teenage gambling in Great Britain. The goal is to fit a multiple regression model, where the gambling expenses (in pounds per year) is the response variable, and sex (0=male, 1=female), status (socio-economic status score, based on the parents employment), income (in pounds per week) and verbal score (number of correctly answered questions from 12 on use of language) are the predictors.

a) Use some visualization methods to gain a first overview on the data. Also decide which transformations are necessary.

b) Make sure that all predictors are from the correct data type in R.

c) Perform a multiple linear regression with all predictors (some of which may be transformed).

d) What portion of the variation in the response is explained by the predictors?

e) Which of the observations has the largest positive residual? What are the properties of that person?

f) Compute median and mean of the residuals. Any comments?

g) Now also compute the correlation of the residuals with the fitted values, as well as with the predictor income.

h) If all the other predictors remain the same, what is the difference in the predicted gambling expenses between a male and a female? Also give a confidence interval for that difference.

i) Start with an empty model, only containing the intercept. Then add the predictors step by step, one at each time. Use the following sequence: income, sex, verbal, status. After every step, write down the estimated error variance, R-squared and adjusted R-squared. Finally, display these graphically.

Preliminary discussion: Monday, October 17.

Deadline: Monday, October 24.