Series 2

1. **Robust Regression and Leverage Points:** We want to describe the monthly consumption of steam (`Steam`) of a certain firm as a function of the variables "monthly operating days" (`Operating.Day`) and "average outside temperature" (`Temperature`). The dataset is `dsteam.dat` and can be found at `http://stat.ethz.ch/Teaching/Datasets/cas-das/dsteam.dat`.

   a) Make a pairs-plot. Do you see any outliers? Any leverage points?

   b) Run a regression analysis with the least squares method (`lm(...)`) and with the MM-estimator (`lmrob(...)` from the package `robustbase`) on the dataset `dstream`. Compare the results.
   (optional) Use the regression M-method (`rlm(..., method = "M")` from the package `MASS`) and compare the results with those of the other estimators.

   c) Compare the residual analyses of the two fits of exercise **b)**.

   d) There are two observations with small numbers of monthly operating days. Let's assume the firm had holidays during these months. The dummy variable `Working.Holidays` has been introduced for this assumption.
   Fit again the corresponding regression models by the LS- and the MM-method. Are there still differences in the results?

   e) Draw the fitted values into the plot `Steam` vs sequence of the observations. Since the data contains monthly data, the plot corresponds to a time series plot.
   R-Hints:
   ```
   t.ylim <- range(d.steam$Steam, fitted(r.ls), fitted(r.MM))
   plot(d.steam$Steam, type = "l", ylim = t.ylim, ylab = "")
   lines(fitted(r.ls), lty = 2, col = 2)
   lines(fitted(r.MM), lty = 3, col = 3)
   ```

2. **Model selection:** The data is the result of an experiment with the purpose of finding the factors for the specific gravity of pine wood. For the experiment a very thin cross-section of tree trunks has been analysed. The number of fibers per mm^2 of spring wood (`nFrFas`) and summer wood (`nSoFas`), the portion of spring wood (`AntFr`), the portion of absorbed light in the spring wood (`FrLicht`) and summer wood (`SoLicht`) and the specific gravity of the wood has been measured. The original data (from Draper and Smith 1966) can be found in `wood.dat`, the contaminated data (from Rousseeuw and Leroy, 1987) in `woodRous.dat` (`http://stat.ethz.ch/Teaching/Datasets/cas-das`).

   a) Read the data into R and transform the explanatory variables according to Tukey's "first aid transformations".
   R-Hints: Transformations
   - Squareroot for numbers: `sqrt(...)`
   - Arcus-Sinus-Funktion for portions: `asin(sqrt(...))`

   b) Make a pairs-plot. Can you see any outliers?

   c) Fit a suitable regression model by the least-square method. Make a residual analysis. What do you see? Run a model selection and report your final model. Make residual analysis for this model: What attracts your attention?
   R-Hints: Use `drop1(... , test = "F")` or `step(...)`.

   d) Redo the regression analysis and the model selection by the MM method. What are your findings?
   R-Hints: Use `lmrob(...)` from the package `robustbase` and use `plot(...)` for the residual analysis. To compare the two models `M1` and `M2` use `anova(M1, M2, test = ....)` for `test = "Wald"` and `test = "Deviance"`.

   e) Compare the results of the two regression analyses.

   f) Compare the outliers of the two regression analyses with the original data `wood.dat`. Which regression analysis can correctly identify the modified observation? Note, that the original data has to be transformed in the same way as the modified data. Which regression analysis finds the outliers?

g) *(optional)* i) Remove the outliers you found and fit classical (non-robust) regression on the rest of the data set.Run a stepwise model selection with AIC.
ii) Run a step wise model selection with the robustified final prediction err using the orignial data.
iii) Compare your final models to the one found in **d)**.

3. In the file `http://stat.ethz.ch/Teaching/Datasets/cas-das/Synthetisch.dat` a synthetic data set is provided with the response variable `y` and the explanatory variables `x1` and `x2`.

a) Fit the model

$$Y_i = \beta_o + \beta_1 \, x1_i + \beta_2 \, x2_i + E_i$$

with the least squares method to the data. Run a residual analysis and identify deviations from the assumptions. Note the estimated coefficients and the estimated standard deviation.

b) Fit the same model as in part **a)** to the data, but use the regression SMDM-estimator. Run again a residual analysis and identify deviations from the assumptions. Write down also the estimated coefficients and the estimated standard deviation.

c) Compare the two analyses.

4. In this exercise we are looking at the data `aptitude`. For this data set there are the following three variables on 27 subjects:

`PASS`:	binary response, 1 if the subject passed the exam at the end of the course and 0 otherwise
`SCORE`:	numeric, represents scores on an aptitude test for a course
`EXP`:	numeric, represents months of relevant previous experience

We are interested in the question whether passing or failing the test on the end of the course can be explained by the scores of the aptitude test and the months of relevant previous experience.

Read in the data via

```
> aptitude <- read.table("http://stat.ethz.ch/Teaching/Datasets/cas-das/aptitude.dat",
                         header = TRUE)
```

a) Which regression model should be used for this kind of data?

b) Fit this regression model with classical (non-robust) estimates. What are the estimates of the parameters and which of the variables seem to have a significant effect on passing the exam?

c) Now, plot `PASS` vs. `EXP`. Which observations could be seen as "outliers", meaning that they have an unexpected value of the response variable `PASS` for the value of their explanatory variable `EXP`?
R-Hint: `identify()`

d) Now fit the regression model with robust *Mqle* estimates (and $w(x_i) = 1$ for all observations i - default in R). What are your estimates now? Did they change compared to **b)**?
R-Hint:

```
> library(robustbase)
> fit <- glmrob(..., family = ..., method = "Mqle")
```

e) Plot the weights of the residuals the algorithm used. Which residuals got the smallest weights? Compare with your result from **c)**.

f) In which situation is it helpful to use weights on the design points, i.e., use different $w(x_i)$? Use a plot to see whether this is the case or not.

5. In this exercise we are looking at the data set `epilepsy` from the R-package `robustbase`. It can be loaded via `data(epilepsy)`. It is data from a clinical trial of 59 patients with epilepsy. We are interested in the following variables:

`Ysum`:	Total number of epilepsy attacks patients have during the 4 follow-up periods.
`Age10`:	Age of the patients divided by 10
`Trt`:	A factor with levels `placebo` and `progabide` indication whether the anti-epilepsy drug Progabide has been applied or not
`Base4`:	Number of epileptic attacks recorded during 8 week period prior to randomization divided by 4

We are interested in the question whether the total number of epilepsy attacks is dependent on the treatment and the other two explanatory variables.

a) Which regression model can be used for analysing this kind of data?

b) Fit this regression model with classical (non-robust) estimates. Which of the two-way-interactions with the factor `Trt` could be present and should thus be included in the model? Include these, and look at the estimates of the fitted model.

c) Now fit the regression model with robust *Mqle* estimates and $w(x_i) = \sqrt{1 - h_{ii}}$ for all observations i. Do your estimates change?

d) Check whether the interactions should be included in the model. If not, leave them away and check again your model estimates.
R-Hint: `anova(..., test = "QD")`

e) Plot the weights $w(x_i)$. Which observation gets the smallest weight? Can you find out why?

Exercise hour: Monday, June 10, afternoon.