Test Robust & Nonlinear Regression

Name:

Please write **all required results**, R-codes and sketches **on paper**. The R script as well as plots drawn in R **cannot** be printed or handed in (they will not be graded).

If there is something you don't understand or if there are problems with R, **ask us** before losing too much time.

Good luck!

Robust Statistics

1. *(6 points)* The data in the dataframe `RP` are from a study, correlating three measurements made on tyre rubber samples: `Hard` = Hardness (in Shore units), `TS` = Tensile Strength (in kg/sq m), and `AL` = Abrasion Loss (in gm/hr). Because both `Hard` and `TS` are easier and less costly to measure than `AL` a model predicting `AL` as a function of `Hard` and `TS` has been developed:

$$\texttt{AL}_i = b_0 + b_1 \cdot \texttt{Hard}_i + b_2 \cdot \texttt{tTS}_i + E_i \quad with \quad E_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

where the variable `tTS` is transformed Version of `TS`.

The data can be read in as follows:

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

a) *(3 points)* Fit this regression model using the least-squares and the regression-MM-estimator (`lmrob()` using default settings). Calculate the robust 95% confidence interval. (Write down the **R-Code** in your solution!) Compare the robust 95% confidence interval with the estimated least-squares coefficients. What do you conclude?

b) *(3 points)* Perform residual analyses based on both estimation procedure. Discuss the results. Are the conclusions different?

2. *(14 points)* We will work on a fish catch dataset which contains measurements on 115 fish caught in the lake Laengelmavesi, Finland. For these fishes of 3 species the weight, length, height, and width were measured. Three different length measurements are recorded: from the nose of the fish to the beginning of its tail, from the nose to the notch of its tail and from the nose to the end of its tail. The dataset `Fish2` contains the log-transformed measurements `lWeight`, `lLength1`, `lLength2`, `lLength3`, `lHeight`, and `lWidth`. The last variable, `Species`, represents the grouping structure of the three species called `A`, `B` and `C`. But there might be some doubt whether all fishes are assigned correctly. The data can be read in as follows:

```
> Fish2 <- read.table("http://stat.ethz.ch/Teaching/Datasets/WBL/Fish2.dat",
                      header = TRUE)
```

a) *(4 points)* Based on the first 6 variables in `Fish2`, estimate the covariance matrix classically and robustly using the R function `CovRobust`.

(i) Which method is used in the R function `CovRobust` when applied to the dataset `Fish2`?

(ii) Note the entries for the covariance between the variable `lWidth` and the variables `lWeight`, `lHeight` and `lWidth` for both estimation methods and compare them.

b) *(4 points)* Calculate the robust and the classical Mahalanobis distances for the observations and identify the outliers.

(i) Write down the **R-Code** for doing the calculation.

(ii) How many outliers are found with the classical and how many are found with the robust approach based on the 0.95 quantile?

c) *(6 points)* Run a linear discriminant analysis with the classical method and with the R function `rlda(...)`.

(i) Display the data with respect to the first and second discriminant variables and discuss the results (draw a sketch of the plots in your solution!).

(ii) What does the R function `rlda()` do differently compared with the classical method?

Nonlinear Regression

3. *(22 points)* In fisheries biology, there are several theoretical models describing the relationship between the size of the spawning stock (called the spawning biomass) and the resulting number of fish (called the recruitment). The dataframe `d.hake` contains stock and recruitment data for hake in the period 1982-1996. The three variables are the spawning biomass `S` (in 1000 tonnes), the number of fish `Y` (in million fish), and `year` (which we will not use).

In this exercise, we will use the Beverton-Holt model

$$h\langle S, a, k\rangle = \frac{a \cdot S}{(1 + \frac{S}{k})} \ ,$$

where a and k are unknown parameters, to analyse the hake data. The data can be read in as follows:

```
> d.hake <- read.table("http://stat.ethz.ch/Teaching/Datasets/WBL/dhake.dat",
                       header = TRUE)
```

a) *(2 points)* Fit the Beverton-Holt model to the data with the starting value $a = 5$ and $k = 40$. (Write down the **R-Code** in your solution!). Provide the estimated parameters a and k.

b) *(4 points)* Determine the approximate 95%-confidence-interval for k based on

- the Wald approach
- profile likelihood
- bootstrap method.

How do the three solutions differ? (Write down the **R-Code** in your solutions!)

c) *(2 points)* Visualize the correlation between the two estimated coefficients based on the bootstrap simulation. What do you observe?

d) *(3 points)* Use the profile t-plot and the profile-traces to assess the linear approximation in exercise **b)** (write down the **R-Code** in your solution!). What are your conclusions?

e) *(3 points)* Perform a residual analysis and assess the fit based on the Tukey-Anscombe plot and the QQ-normal-plot (draw a sketch of the two plots in your solution!). Can we rely on the inferential result of the previous sub-exercises?

f) *(4 points)* Calculate a prediction and the 95% prediction interval for $h\langle S_0, a, k\rangle$ at $S_0 = 35$ (i.e., at 35'000 tonnes). (Write down the **R-Code** in your solution!)

g) *(4 points)* Suppose we hadn't given you starting values for fitting the Beverton-Holt model in **b)**. How would you proceed to determine the starting values for a and k yourself? Explain your approach and report your obtained starting values (write down the **R-Code** in your solution!). **Hint**: a is the slope of the Beverton-Holt model at $S = 0$.