MATH1712 Probability and Statistics II
Practical

http://www1.maths.leeds.ac.uk/~voss/2017/MATH1712/
Jochen Voss, J.Voss@leeds.ac.uk
2017/18, semester 2

- For this practical you have to examine a data set using R, and to present your results in a short report.
- The practical counts towards the coursework mark for the module. There will be 20 marks in total (16 for contents and 4 for presentation).
- You should work on your report during the week from 5th to 9th March. (The practical replaces the homework which would normally be due on Wednesday 15th March.) The deadline for handing in your report is **Friday, 16th March, 5pm**.
- There are timetabled drop-in sessions for the practical where you can get help with the use of R.

In this practical we will consider a data set about fishing vessels which the UK government provides at

`https://www.gov.uk/government/statistical-data-sets/vessel-lists-over-10-metres`

The web page provides several versions of the data set, for the practical we will use the data from January 2018.

Task 1. Import the data into R and give an overview over the data, using summary statistics as appropriate. To load the data into R, it may be easiest to first load the data into Microsoft Excel and then to save the relevant section of the spreadsheet as a csv file. Your report should at least address the following issues:

- Give some general information about the data set.
- Convince yourself and the reader that you have imported the data correctly.
- Classify all variables as either numerical or categorical.
- Produce histograms of the "overall length" and the "vessel capacity units".

Task 2. Use least squares regression to fit a linear model which can be used to predict "vessel capacity units" from "overall length". Use this model to predict the capacity units of a (previously unseen) vessel which has an overall lengths of 28 metres. Your report should at least address the following issues:

- Describe the procedure used to fit the model.
- Describe the fitted model.
- Use the model to produce a prediction.
- Discuss the accuracy of your prediction.

Task 3. Discuss how appropriate a linear model is for the data, and how model fit can be improved. Your report should at least address the following issues:

- Include a scatter plot which shows the data together with the regression line fitted in task 2.
- Based on the plot, identify possible problems with the model fit.
- Produce an improved estimate for the capacity units of a 28 metres long vessel.

p.t.o.

Task 4. To assess how good the predictions of a fitted model will be for new data, the following procedure can be used: The model is fitted using only half of the data, and then the other half of the data is used to assess the quality of predictions made by the model. This procedure is called cross validation. Perform cross validation for the linear regression model considered in task 2, and assess how accurate predictions made using the model will be.

- Split the data into two halves, the "training set" and the "test set".
- Fit a linear model, as in task 2, using only the training set.
- For each vessel length in the test set, use your model to predict vessel capacity units.
- Using any method you see fit, compare the predicted capacity units to the actual capacity units of the test set.

Some guidance for writing your report:

a) Clearly mark your report with your name and your student ID.

b) Staple the pages of your report together, do not use plastic sleeves *etc.*

c) Your report must be typeset (not handwritten) and must not exceed five pages, including all figures, R code *etc.* Since space is limited, focus your discussions on the essentials and think about what is most important to include in your report.

d) Clearly explain what you do and discuss your results. Write complete sentences, including correct punctuation.

e) Use plots to illustrate your results, and discuss each plot you include in the report. Take care to use good axis labels, captions, *etc.* for your plots and make sure that your plots are meaningful and easy to interpret.

f) Include and explain all R code you use to derive your results, do not include unused or unnecessary R code. Include the code in the main text rather than into an appendix.