RESEARCH SCHOOL OF
FINANCE, ACTUARIAL STUDIES AND APPLIED STATISTICS
College of Business & Economics, The Australian National University

REGRESSION MODELLING

(STAT2008/STAT6038)

Assignment 1 for 2014

Instructions

- This assignment is worth 20% of your overall marks for your course (for all students, enrolled in either STAT2008 or STAT6038). If you wish, you may work together with another student in doing the analyses and present a single (joint) report. If you choose to do this then both of you will be awarded the same total mark. A STAT2008 student may work with a STAT6038 student. You may NOT work in groups of more than two students and the usual ANU examination rules on plagiarism still apply with respect to people not in your group.

- Research School of Finance, Actuarial Studies and Applied Statistics Cover Sheets for individual and group assignments are also available on Wattle. Please complete and attach a copy of the appropriate cover sheet to the front of your assignment.

- Assignments should be written or typed on sheets of A4 paper stapled together at the top left-hand corner (do not submit the assignment in plastic covers or envelopes). Your assignment may include some carefully edited computer output (e.g. graphs) showing the results of your data analysis and a discussion of those results. Please be selective about what you present – only include as many pages and as much computer output as necessary to justify your solution and be concise in your discussion of the results. Clearly label each part of your report with the question number and the part of the question that it refers to.

- Unless otherwise advised, use a significance level of 5%.

- Marks may be deducted if these instructions are not strictly adhered to, and marks will certainly be deducted if the total report is of an unreasonable length, i.e. more than 10 pages including graphs. You may include as an appendix any *R* commands you used to produce your computer output (or the details from whatever statistical software you choose to use). This appendix and the cover sheet are in addition to the above page limits; but the appendix will generally not be marked, only checked if there is some question mark over what you have actually done.

- Assignments will be marked by your allocated tutor. Assignments should be submitted in the relevant assignment box located next to the Research School of Finance, Actuarial Studies and Applied Statistics office by **5pm on Thursday 3 April 2014**. The tutors will NOT answer any further questions about this assignment after this deadline.

- Late assignments will be accepted without deduction of marks until **12 noon on Monday 7 April 2014**. Assignments will NOT be accepted after 12 noon on Monday 7 April without an extension. Extensions will usually be granted on medical or compassionate grounds on production of appropriate evidence and must have the written permission of the lecturer and your tutor by no later than Thursday 3 April.

Data

The data to be used in this year’s assignments come from the recommended text by Julian J. Faraway (Linear Models with *R*, Chapman & Hall/CRC, 2005) and are all stored in the Faraway library, which is available from CRAN (the *Comprehensive R Archive Network*, the original Australian mirror site for which is located here in Canberra at the CSIRO). You can access Faraway’s stored library of data and functions by starting *R* and typing the following commands:

```
install.packages()
# select the Australian CSIRO CRAN mirror and choose the faraway package

library(faraway) # this attaches the faraway library to your search path
search()

ls(pos="package:faraway") # lists the contents of the faraway package

help(prostate)
help(teengamb)
# Faraway has provided brief help files on all of the datasets, which
# include a description of the variables and the original source

prostate
teengamb
# shows the contents of the data to be used in this assignment

attach(prostate)
attach(teengamb)
# attaches the data to your search path, so you can reference the variables
```

Further details (such as the other packages you will need to load if you wish to use all of the stored functions described in the Faraway text) are available in Appendix A on page 217 of the Faraway text.

Question 1 **(20 marks)**

The dataset `prostate` comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy (a surgical procedure). Of the variables included in this dataset, `lcavol` (log of the cancer volume) is a measure of the size of the cancer tumour and `lpsa` (log of the prostate specific antigen measure) is the result of a diagnostic blood test for prostate cancer.

(a) Plot `lpsa` against `lcavol`. Is there a significant correlation between `lpsa` and `lcavol`? Use *R* to conduct a suitable hypothesis test and present and interpret the results. **(4 marks)**

(b) Fit a simple linear regression model with `lcavol` as the response variable and `lpsa` as the predictor. Construct a plot of the residuals against the fitted values, a Q-Q plot of the residuals and a bar plot of the leverages for each observation. Comment on the model assumptions and on any unusual data points. **(4 marks)**

(c) Produce the ANOVA (Analysis of Variance) table for the SLR model in part (b) and interpret the results of the F test. Are these results consistent with the hypothesis test you conducted in part (a)? **(4 marks)**

(d) What are the estimated coefficients of the SLR model in part (b) and the standard errors associated with these coefficients? Interpret the values of these estimated coefficients and perform t-tests to test whether or not these coefficients differ significantly from zero. What do you conclude as a result of these t-tests? **(4 marks)**

(e) Plot `lcavol` against `lpsa`. Include the fitted SLR model from part (b) as a line on the plot and also show 95% confidence intervals for the mean or expected value of `lcavol` (do NOT plot the 95% prediction intervals). Do the results of a PSA test appear to be a reliable predictor of the size of the prostate cancer tumour? **(4 marks)**

Question 2 **(20 marks)**

The dataset `teengamb` concerns a study of teenage gambling in Britain. For this assignment, we are interested in whether a teenager's `income` (measured in UK £ per week) can be used to predict the amount they will `gamble` (gambling expenditure measured in UK £ per year), at least for the teenagers who do regularly gamble.

(a) Plot `gamble` against `income`. Describe the correlation shown in the plot. Would you expect a simple linear regression model to be a reasonable model for the relationship shown in the plot? **(4 marks)**

(b) Fit a simple linear regression model with `gamble` as the response variable and `income` as the predictor. Construct a plot of the residuals against the fitted values, a Q-Q plot of the residuals and a bar plot of Cook's Distances for each observation. Comment on the model assumptions and on any unusual data points. **(4 marks)**

(c) In question 1, a natural log (to the base *e*) transformation had already been applied to both the response and predictor variables and appeared to produce reasonable results. In this example, there are a number of teenagers who are not regular gamblers (their annual expenditure on gambling is very small or even zero). What is the problem with applying a log transformation in this situation? Exclude any teenager who spends less than £1 per year on gambling and fit another simple linear regression model with log(`gamble`) as the response variable and log(`income`) as the predictor. Check the same plots you produced for the earlier model in part (b). Are the same problems still apparent? **(4 marks)**

(d) Produce the ANOVA table and the table of the estimated coefficients for the revised SLR model in part (c). Interpret the values of the estimated coefficients for this SLR model and the results of the overall F test and the t-tests on the estimated coefficients. **(4 marks)**

(e) Use the revised SLR model from part (c) to predict the annual expenditure on gambling for three British teenagers, who were not included in the original study, but who have weekly incomes of £1, £5 and £20, respectively. Find 95% prediction intervals for these predictions. Do you think this revised SLR model is a good model for making all three of these predictions? **(4 marks)**