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

REGRESSION MODELLING

(STAT2008/STAT4038/STAT6038)

Sample Assignment 2

Instructions

- This is NOT your current assignment. It is a sample assignment using questions taken from assignments from earlier years. Assignment instructions tend to change slightly from year to year – read the full instructions on this year's assignment sheet and follow those instructions carefully.

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("faraway") # installs the faraway package and dataset library.
# This can take some time, especially on an ANU InfoCommons computer, so show some patience.

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 265 of the Faraway text.

Copies of the datasets (as .csv files) and help files are also available on Wattle, in case you have trouble loading the Faraway library.

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). In this assignment we are going to fit an appropriate multiple linear regression model to examine factors affecting lcavol (log of the cancer volume), which is a measure of the size of the cancer tumour (measured in ml).

(a) All of the other variables in the prostate dataset could potentially be included as predictors (explanatory variables) in a multiple regression model with lcavol as the response variable. Produce suitable plots and/or summary R output to investigate the contents of the variables svi, gleason and pgg45. How are these variables distributed? Discuss any potential problems with including these variables in a multiple regression model. **(3 marks)**

(b) Find an appropriate multiple linear regression model with lcavol as the response variable and lweight, age, lbph, lcp and lpsa as possible predictors. To simplify this exercise, exclude the variables mentioned in part (a) from consideration, assume that all the other variables are already measured on an appropriate scale (i.e. no further transformations are necessary), that an additive model is appropriate (i.e. no interaction terms or quadratic/higher order terms are needed), but do NOT exclude any potential outliers. Do NOT present output for multiple models, choose just ONE model!
Produce the ANOVA (Analysis of Variance) table for your chosen model and summary output showing the estimated coefficients and use these to justify your choice of model. Why have you included the explanatory variables that are included in your model and why have you chosen to exclude other possible predictors? **(4 marks)**

(c) For your chosen multiple regression model, construct a plot of the externally Studentised residuals against the fitted values and a normal Q-Q plot of the internally standardised residuals and use these plots to comment on the model assumptions.
Also produce selected statistics and/or a plot to investigate and discuss possible outliers and influential observations. Do NOT try to present a table of various statistics showing all 97 observations (though you could select just one statistic and present a relevant plot which shows all 97 observations). **(5 marks)**

(d) Perform a "nested model" F test to see whether or not any of the subset of possible predictors you have excluded from your chosen model would be a significant addition to your chosen model. If your chosen model includes 4 or 5 of the possible predictors (lweight, age, lbph, lcp and lpsa) then perform a test of the last two or three predictors as an addition to a model that already contains the other variables.
Review the above test results and the output in parts (b), (c) and (d) and also compare your chosen model with the simple linear regression model shown in the models solutions to Question 1 of Sample Assignment 1. Is your chosen a model an improvement in terms of reliably predicting the size of a prostate cancer tumour? **(5 marks)**

(e) Add an interaction term between lpsa and lcp to your chosen model (if your chosen model does not already include linear terms in lpsa and lcp, also add those terms to the model). Is this term a significant addition to the model? Interpret the coefficients of the terms involving both lpsa and lcp (and their interaction) in this expanded model and in your chosen model. **(3 marks)**

Question 2 **(20 marks)**

The dataset teengamb concerns a study of teenage gambling in Britain. In this assignment we are going to fit an appropriate multiple linear regression model to examine factors affecting the amount that teenagers will gamble (gambling expenditure measured in UK £ per year), including both teenagers who do and who do not regularly gamble.

(a) Transform gamble by creating a new variable trans.gamble <- log(gamble + 1). Compare histograms of gamble and trans.gamble and comment on which is more likely to be suitable for inclusion in a multiple regression model.
Assume that the researchers who collected the data believe that gambling expenditure differs by sex and is also strongly affected by factors such as education and socio-economic status. This is why they collected the variables verbal and status (as measures of education and socio-economic status respectively) and any multiple regression model will include status, verbal and sex as predictors so we can test these assertions (and control for the effects of these factors).
This leaves income as the only remaining observed variable (covariate). Construct an added variable plot to assess income as a possible addition to a multiple regression model for trans.gamble that already includes sex, verbal and status as predictors. Does this added variable plot suggest a transformation is required for income? The transformation we used in Question 2 of Sample Assignment 1 was log(income). Construct a different added variable plot for log(income). Is this an improvement?
(5 marks)

(b) Fit the multiple linear regression model with trans.gamble as the response variable and sex, verbal, status and log(income) as predictors. Construct a plot of the externally Studentised residuals against the fitted values, a normal Q-Q plot of the internally Studentised residuals and a bar plot of Cook's Distances for each observation. Comment on the model assumptions and on any unusual data points. Calculate appropriate influence statistics for the most unusual data point and comment on these statistics, but do NOT refine the model by removing this observation as a possible outlier. **(5 marks)**

(c) Produce the ANOVA table and the summary table of estimated coefficients for the multiple linear regression model in part (b). Interpret the overall and sequential F tests and the t-tests and the values of the estimated coefficients of the model. Are the earlier assertions in part (a) about sex, education and socio-economic status supported in the context of this model? **(5 marks)**

(d) To help the researchers interpret the model, plot gamble against income, with different plotting symbols for the two values of sex. Include your model on this plot by calculating predicted values for trans.gamble, for the full range of income values and for both values of sex, holding verbal and status at their mean values. Suitably back-transform the predictions and include them on the plot separately for both males and females. Also include point-wise 95% confidence intervals (but not 95% prediction intervals) on the plot. **(5 marks)**
