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

REGRESSION MODELLING

(STAT2008/STAT4038/STAT6038)

Assignment 2 for 2017

Instructions (please read carefully as some instructions have changed since Assignment 1)

- This assignment is worth either 20% of your overall marks for this course (for all students, enrolled in STAT2008, STAT4038 or STAT6038).
- If you wish, you may work together with another student in doing the analyses and submit a single (joint) report. If you choose to do this, then both of you will be awarded the same total mark. Students enrolled under different course codes may work together. **Do NOT submit multiple copies of the same report to different markers**, if you have different tutors, just pick one. 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 Statistics (RSFAS) assignment cover sheets are available on Wattle. Please complete and attach a copy of the cover sheet to the front of your report. **Remember to keep a copy of your assignment for your own records.**
- Assignments should be written, typed or printed 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 12 pages including graphs. You may include as an appendix, any `R` commands you used to produce your computer output. 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 about what you have actually done.
- Assignments will be marked by your tutor (or one of your two tutors, for joint assignments). One copy of your assignment should be submitted to the box labelled with the name of your tutor, located next to the RSFAS office, by **3 pm on Friday 19 May 2017**. You may ask any of the tutors or me (Ian McDermid) questions about this assignment, in person, up to the deadline (3 pm on Friday 19 May 2017), after which we will NOT answer any further questions about this assignment, until after the marked assignments have been returned to students. Answers to questions in writing sent to me via e-mail or posted on Wattle, will be posted on Wattle, but must be received no later than 12 noon on Thursday 18 May 2017.
- Late assignments will NOT be accepted after the deadline without an extension. Extensions will usually be granted on medical or compassionate grounds on production of appropriate evidence, but must have my permission by no later than 12 noon on Thursday 18 May 2017. Even with an extension, all assignments must be submitted (and receipt acknowledged by the RSFAS Office), before 2pm on Monday 22 May 2017, which is the start of the first tutorial in week 12, which is when the assignment solutions will be released on Wattle, so that the solutions can be discussed in the tutorials in week 12.

Data (this section is the same as for Assignment 1)

The data for the first question (available in the file moorhen.csv on Wattle) is presumably from an old consulting project (not one of mine, which might explain the poor documentation). The original consultant hopefully got the permission of the owners to use it in teaching, as it has been used for this purpose before. I am using it here as an example of a situation I have often found myself in as a consultant, which is having to work with poorly described data and where I have to speculate on aspects of the interpretation and the research question. Working as a consultant is considerably easier when you are able to directly collaborate with the clients on these issues.

Many of the projects I have worked on as a statistician have involved data that was considered private (such as health data) or data to which access was restricted (for example, data designated "commercial-in-confidence"). For these reasons, it is not always easy to source realistic data for use in teaching statistics and so groups of statisticians maintain repositories of examples of real data that are in the "public domain". In many countries, there are Internet repositories of data available for use in the teaching of introductory statistics.

The data for the second question comes from such a repository: the data archive associated with the Journal of Statistics Education (JSE), a publication of the American Statistical Association (www.amstat.org/publications/jse/jse_data_archive.htm).

Datasets in the JSE data archive are typically accompanied by a file which give a description of the variables included in the data (the "meta-data") and are also often accompanied by an associated article in the journal (and occasionally even by references to other sources). The fat data, which we will be using in question 2 of this year's assignments, includes both of the above accompanying documents.

You can download a text file containing the fat data and the associated documents from the JSE website (www.amstat.org/publications/jse/jse_data_archive.htm) or the data is also available on Wattle in the file fat.csv, which includes a header row with the variable names. I have also downloaded a copy of the meta-data text file (fat.txt), and made this file available on Wattle.

Alternatively, the fat data are also stored in the UsingR package from the recommended text by John Verzani (Using *R* for Introductory Statistics, 2nd Edn, Chapman & Hall/CRC, 2014). The UsingR package 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 install the UsingR package by typing the following commands in R:

```
install.packages("UsingR") # installs the UsingR package
library(UsingR) # attaches both the UsingR and the HistData libraries to your search path
search()

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

help(fat) # there are brief help files on all of the datasets in the above libraries
fat # typing the name shows the contents of the dataset
attributes(fat) # check that the variable names match the description
summary(fat) # always a sensible bit of exploratory data analysis
attach(fat) # attaches the data sets to your search path, so you can reference the variables
```

Further details are available in the sections titled "External packages" and "Data sets" on pages 15-18, towards the end of "Chapter 1. Getting started" in the Verzani text.

Question 1 **(20 marks)**

Following on from the analysis of the moorhen data in Question 1 of Assignment 1, we would like to use all available variables to try and build a multiple regression model with Shield area as the response variable. The e-mail from the scientists that came with the data doesn't really describe the variables Stern, Hb and TandT, except to say that they are "three lineal measurements" taken on each bird. Adult is an indicator of whether the bird is a juvenile (0) or adult (1) bird.

Use R to further analyse the moorhen data and answer these questions:

(a) Produce both a scatterplot matrix and a correlation matrix for the variables in the moorhen data and comment on any important relationships between the variables (assuming you are planning to do a multiple regression analysis). **(3 marks)**

(b) Fit a multiple linear regression model with Shield as the response variable and with all the other variables in the data as explanatory variables. Present the main residual plot of the residuals against the fitted values for this model. Are there are any obvious problems with underlying assumptions? **(3 marks)**

(c) Now fit a multiple linear regression model with ln(Shield) as the response variable, still using all the other variables (not log transformed) as explanatory variables. Again present the main residual plot of the residuals against the fitted values for this new model. Does the transformation applied to the response variable appear to have corrected any problems you identified in part (b)? **(3 marks)**

(d) Examine (but do not present) the ANOVA (Analysis of Variance) table and summary output for the model in part (c). Now adjust the order of the explanatory variables in the model in part (c), so that you can test the following "nested" hypotheses:

$H_0 : \beta_{\text{Weight}} = \beta_{\text{Hb}} = \beta_{\text{TandT}} = 0$

$H_0 : \beta_{\text{Hb}} = \beta_{\text{TandT}} = 0$

$H_0 : \beta_{\text{TandT}} = 0$

Present the ANOVA table for the re-ordered model and discuss the result of the partial (nested) F-tests for the above hypotheses. Do your results suggest some possible modification(s) you could make to the model? **(3 marks)**

(e) Now fit the multiple linear regression model with ln(Shield) as the response variable and with Adult and Stern as explanatory variables. For this model, construct a plot of the internally Studentised residuals against the fitted values, a normal Q-Q plot of the residuals, and a bar plot of Cook's distances for each observation. Are there any obvious problems with the underlying assumptions? **(3 marks)**

(f) Plot Shield against Stern, using different plotting symbols for juvenile and adult birds. Use the model from part (e) to predict the expected Shield area for both juvenile and adult birds over the full range of possible Stern measurements and include these on your plot as two different curves (using different line types). Include appropriate titles, axis labels, a legend and a brief discussion of your plot. **(3 marks)**

(g) Consider two birds, one an adult bird and the other a juvenile, but who have the same Stern measurement. Present the table of coefficients for the model in part (e). Use this table to estimate the ratio of the expected Shield area of the adult bird to the expected Shield area of the juvenile bird with the same Stern measurement. Find a 95% confidence interval for this estimate. **(2 marks)**

Question 2 **(20 marks)**

The dataset fat contains estimates of the percentage of adipose tissue (body.fat) and other related measurements taken on a sample of 252 adult men. The measurements are described in the help file for this dataset, help(fat), which also contains a link to another file, fat.txt, which contains yet more information about the data. Read both of these documents carefully to gain a better understanding of the contents of this dataset. You may also like to search the internet for a description of some of the terms used (e.g. "adipose tissue").

For this assignment, we are interested in using all available information in the data to build a multiple linear regression model which can be used to estimate the percentage of body.fat, which is not easy to measure directly, as it has to be estimated using an underwater weighing technique.

(a) The file, fat.txt, suggests that there is an error in the height measurement for case 42, which should be 69.5 inches, rather than 29.5 inches. Apply this correction to the data. In fitting a multiple regression model with body.fat as the response variable, why should you not include case, body.fat.siri or density as possible explanatory variables? Is there a potential problem with including all three of weight, height and BMI as explanatory variables (hint: try looking up the definition of BMI)? What about including ffweight as a predictor in a model that already includes weight? **(4 marks)**

(b) Using body.fat (or some transformed version of body.fat) as your response variable and using just age, weight, height, neck, chest, abdomen, hip, thigh, knee, ankle, bicep, forearm and wrist as possible predictors, try to find a multiple regression model that is a "good" fit to these data. Note that weight is considered key and must be included, but BMI and any other variables not listed above should NOT be included. You are welcome to apply suitable scale transformations (e.g. log) to any or all of the explanatory variables, but do not use any interaction terms or higher order terms in any of the variables (except possibly for weight and height, as a way of including some proxy for BMI in the model). Do NOT use any "automatic" variable selection method in choosing your final model and do NOT present output from such a method as a justification for your choice of final model. Also, at this stage, do NOT delete any potential outliers.
Choose a promising candidate model and present a plot of the internally Studentised residuals against the fitted values, a Q-Q plot of the residuals and a bar plot of Cook's distances for each observation. Discuss these plots and comment on the model assumptions and on any unusual data points. **(4 marks)**

(c) You may now delete no more than 3 potential outliers, but only if you can suggest a sensible justification for each exclusion. Refit the model to the reduced data set and again check residual plots. At this stage, you might decide to vary any transformations used and revisit the issue of which potential outliers to exclude. Choose just ONE final model and to justify your choice, present and discuss residual plots for your chosen model (with outliers removed). **(4 marks)**

(d) Present the ANOVA table and the table of the estimated coefficients for your chosen model from part (c). Interpret the values of the estimated coefficients for this model and the results of the overall F test and the t-tests on the estimated coefficients. **(4 marks)**

(e) Assume you have four groups of new individuals categorised as "underweight", "normal", "overweight" and "obese". Assuming that the four new groups have the same average for weight and the other measurements as the corresponding groups in the original sample, use your chosen model to repeat the predictions and confidence intervals made using the simple linear model in part (e) of Question 2 of Assignment 1. Is your chosen model a good model for making all of these predictions? **(4 marks)**