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

REGRESSION MODELLING

(STAT2008/STAT4038/STAT6038)

Assignment 1 for 2017

Instructions

- This assignment is worth either 15% or 20% of your overall marks for this course (for all students, enrolled in STAT2008, STAT4038 or STAT6038). It will be worth only 15% rather than 20%, if you attempt the optional mid-semester Wattle quiz in week 6, which is worth 5%, and your mark on the quiz is better than your mark on this assignment.
- 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. Students enrolled under different course codes may work together. 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 31 March 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 31 March 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 30 March 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 30 March 2017. Even with an extension, all assignments must be submitted reasonably close to the original deadline to allow time for the marking to be completed prior to week 7 (which starts on Tuesday 18 April 2017), when the assignment solutions will be released and discussed.

Data

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)**

Moorhens are those blue-purple-red water birds often seen down near Lake Burley Griffin in Commonwealth Park. They are characterised by large, fleshy red shields that protrude from their heads. Some scientists have collected various measurements on a group of 43 moorhens in Commonwealth Park in the file moorhen.csv, which is available on Wattle. The scientists have sent the data to you for analysis.

The e-mail accompanying the data is a little light on the details, but there is a suggestion that moorhens form a fairly hierarchical society and that shield size is a relevant indicator of a bird's status within their group, so the variable of most interest (the response variable) is the area of each bird's Shield (units not specified, but presumably in mm^2). An alternative explanation might be that a bird's status is more strongly related to their overall size (which could be measured by the bird's Weight, presumably in mg) and that bigger birds simply have larger shields.

In this assignment, we will concentrate on the relationship between Weight and Shield (we will investigate the other available variables in Assignment 2). Read the data into R and conduct the following analyses:

(a) Plot Shield against Weight (which means that Shield should be the response variable on the Y or vertical axis and Weight should be the explanatory variable on the X or horizontal axis). Use the identify() function in R to identify any unusual data points on the plot. Discuss why you chose these observation(s) as being unusual. **(2 marks)**

(b) Is there a significant correlation between Weight and Shield? Use the cor.test() function to conduct a suitable hypothesis test. Clearly specify the hypotheses you are testing and present and interpret the results. **(2 marks)**

(c) Experiment with applying natural log transformations (to the base e, which is the default for the log() function in R) and square root transformations to one or both of Weight and Shield, and repeat the analysis in parts (a) and (b). Do NOT show all of your results, just pick whichever one you think is the best choice of scale for the two variables and show and discuss the results for your chosen combination. **(4 marks)**

(d) Fit a simple linear regression (SLR) model with your chosen transformation of Shield as the response variable and your chosen transformation of Weight as the predictor. Construct a plot of the residuals against the fitted values, a normal Q-Q plot of the residuals, a bar plot of the leverages for each observation and a bar plot of Cook's distances for each observation. Use these plots to comment on the model assumptions and on any unusual data points. **(3 marks)**

(e) Produce the ANOVA (Analysis of Variance) table for the SLR model in part (d) and interpret the results of the F test. What is the coefficient of determination for this model and how should you interpret this summary measure? **(3 marks)**

(f) What are the estimated coefficients of the SLR model in part (d) 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? **(3 marks)**

(g) Repeat part (a) and again plot Shield against Weight, but this time extend both X and Y axes to include the origin. Now include the transformed SLR model from part (d) as a curve on your plot and also include the untransformed SLR of Shield against Weight as a line on the plot. Use different line types for the two curves and also include an appropriate legend on the plot. What are you overall conclusions about the relationship between Weight and Shield, and the broader research questions discussed in the second paragraph of this question? **(3 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 include a derived variable, BMI or body mass index, which is frequently used as a measure of obesity and is based on simple weight and height measurements.

For this assignment, we are interested in whether or not BMI, which is relatively easy to measure, can be used to predict the percentage of body.fat, which has to be estimated using an underwater weighing technique?

(a) Plot body.fat against BMI. 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 (SLR) model with body.fat as the response variable and BMI 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) A natural log (to the base *e*) transformation (to one or both of the response and predictor variables) is often used to adjust the scale of the variables prior to fitting an SLR model. Now fit another SLR model with body.fat as the response variable and log(BMI) as the predictor. What would be the problem with also applying a log transformation to the response variable? 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 transformed 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) Body mass index values less than 18.5 are typically categorised as "underweight"; from 18.5 to 25 as "normal", 25 to 30 as "overweight" and over 30 as "obese". Use the transformed SLR model from part (c) to predict the body.fat percentage for groups of males with typical BMI values 17.25 ("moderately underweight"), 21.75 ("normal"), 27.5 ("overweight") and 32.5 ("moderately obese"), respectively. Find 95% confidence intervals for these predictions. Do you think this SLR model is a good model for making these predictions? **(4 marks)**