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

GENERALISED LINEAR MODELS

(STAT3015/STAT4030/STAT7030)

Assignment 1 for 2016

Instructions

- This assignment is worth 15% of your overall marks for your course (for all students, enrolled in STAT3015, STAT4030 or STAT7030). 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 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.
- 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 10 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 the course tutor, Yang Yang. Assignments should be submitted in the assignment box labelled with name of this course and your tutor's name located next to the Research School of Finance, Actuarial Studies and Statistics office by **3 pm on Friday 2 September 2016**. You may ask the tutor or me (Ian McDermid) questions about this assignment, in person, up to the deadline (3 pm on Friday 2 September 2016), 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 1 September 2016.
- 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 1 September 2016. Even with an extension, all assignments must be submitted reasonably close to the original deadline of 3 pm on Friday 2 September 2016 to allow time for the marking to be completed prior to week 8 (Monday 19 to Friday 23 September 2016), when the assignment solutions will be released and discussed.

Data

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 that was 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 to be used in this year's assignments come from one such 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 fruitfly data, which we will be using in question 1 of this year's assignments, includes both of the above accompanying documents.

You can download a text file containing the fruitfly 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 fruitfly.csv, which includes a header row with the variable names. I have also downloaded a copy of the meta-data text file (fruitfly.txt), and made this file available on Wattle.

Question 1 **(22 marks)**

Read the description of the fruitfly data in the text file fruitfly.txt, which is available on Wattle (you may also choose to read the other articles referenced in this file, which are all available on-line or through the ANU library e-resources). The title of the original study conducted by Linda Partridge and Marion Farquhar: "Sexual activity reduces lifespan of male fruitflies" (*Nature*, Vol. 294, 10 December 1981, pp. 580-582), provides a brief description of their key research question (locate and read this article for further details).

The original data for this study is available in the file fruitfly.csv, which is also available on Wattle. Read these data into R and create a new factor variable (Activity) to summarise the levels of sexual activity, as follows:

$$\text{Activity} = \begin{cases} \text{A}, & \text{Partners} = 0 \,\&\, \text{Type} = 9 \\ \text{B}, & \text{Partners} = 1 \,\&\, \text{Type} = 0 \\ \text{C}, & \text{Partners} = 1 \,\&\, \text{Type} = 1 \\ \text{D}, & \text{Partners} = 8 \,\&\, \text{Type} = 0 \\ \text{E}, & \text{Partners} = 8 \,\&\, \text{Type} = 1 \end{cases}$$

Note that there is a copy of the data in the faraway library, where the above levels are listed as "isolated", "one", "low", "many" and "high", respectively. However, the data in the faraway library is missing the first observation, so do NOT use that data for this assignment.

(a) Fit an ordinary (normally distributed) additive linear model with Longevity as the response variable, Activity as an exploratory factor and Thorax as a continuous covariate. Produce a plot of the residuals against the fitted values for this model. Are there any obvious problems with this plot? **(1 mark)**

(b) Refit the model in part (a), applying a log() transformation to the response variable (using the default in R, which are logarithms to base *e*). For this modified model, use the rstandard() function to produce a plot of the internally Studentised residuals against the fitted values; using different plotting characters for the five different levels of Activity. Does the log transformation appear to have to have corrected any problems identified in part (a)? Identify unusual observations on your plot and discuss any other interesting features of the plot. **(3 marks)**

(c) Produce a normal quantile plot of the residuals from the model in part (b). Also produce a bar plot of Cook's Distances for each of the observations. Use the rstudent() and hatvalues() functions to calculate the externally Studentised residuals and leverage values for any observations that stand out on these additional residual plots and compare with appropriate cut-offs. Comment on the plots and statistics you have just produced and discuss whether or not there are any outliers. Do NOT refit the model to exclude any outliers. **(3 marks)**

(d) Give the algebraic equation for the underlying population model fitted in part (b), including any assumptions about the error distribution, full details of the variables included in the model and the constraints applied to any factor variables. Is this an example of an ANOVA model or an ANCOVA model? **(2 marks)**

(e) Compare the model in part (b) with a multiplicative model that includes an interaction term between the factor variable (Activity) and the covariate (Thorax). Describe how this additional term modifies the relationship between the response variable and the covariate for the different levels of the factor variable. Is this additional term a significant improvement to the model? Give full details of an appropriate hypothesis test. **(2 marks)**

Question 1 continued

(f) Produce a plot of the data on the original scale (not the log scale) with different plotting characters for the five levels of Activity. Include five curves on this plot, to represent the fitted model from part (b) for the five levels of Activity. Also highlight on the plot any potential outliers you identified in part (c). **(2 marks)**

(g) Present the ANOVA table and the summary table of the coefficients for the model in part (b). Use these tables and the plot in part (f) to discuss the results of the analysis you have conducted so far. **(3 marks)**

(h) Now modify the model in part (b) to include the ID variable as a random effect in an additive mixed effects model. Describe the changes to the underlying population model described in part (d). Discuss whether or not this is an appropriate treatment of the ID variable. (Hint – you may need to investigate possible relationships between ID and the other variables). **(3 marks)**

(i) Present and examine the summary output (analysis of variance table and table of coefficients) for the new mixed effects model in part (h). How has this changed from the summary output presented in part (g)? Calculate the intraclass correlation coefficient for the mixed effects model and comment on the results. **(3 marks)**