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 2017

Instructions

- This assignment is worth either 15% or 20% of your overall marks for your course (for all students, enrolled in STAT3015, STAT4030 or STAT7030). 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 in groups of up to three students (i.e. 1, 2 or 3) in doing the analyses and present a single (joint) report. If you choose to do this, then all of your group 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 three 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 on sheets of A4 paper stapled together at the top left-hand corner (do NOT submit the assignment in plastic covers or envelopes) or scanned as a .pdf file. 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. Each group should submit just one copy of the assignment either online via Wattle or in the assignment box for this course located next to the RSFAS office by **3 pm on Friday 1 September 2017**. You may ask the tutor or me (Ian McDermid) questions about this assignment, in person, up to the deadline (3 pm on Friday 1 September 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 August 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 the permission of both your tutor and me by no later than 12 noon on Thursday 30 August 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, when the assignment solutions will be released and discussed.

Question 1 **(20 marks)**

Neter et al in the text *Applied Linear Statistical Models* (4th edn, Irwin, 1996, p.1159) describe the results of a marketing experiment to investigate the effect of colour of paper (blue, green or orange) on the response rates for questionnaires distributed by the "windshield method" in supermarket parking lots. A representative sample of 15 supermarket car parks were chosen in a metropolitan area and 5 car parks were assigned at random to each of the three colours. The entire experiment was repeated in a different week, with the same colours assigned to the same car parks.

The data are available in the file qcolour.csv, which is available on Wattle. The variables are:

- rrate – the observed response rates (the percentage of questionnaires returned);
- colour – of the paper (blue, green or orange);
- size – of the car park (measured by counting the number of parking spaces); and
- week – week A or week B.

(a) Using R, fit an ordinary (normally distributed) additive linear model with rrate as the response variable, colour and week as exploratory factors and size as a continuous covariate. Do not vary the default contrasts used by R. For this model produce: a plot of the residuals against the fitted values for this model; a normal quantile plot of the residuals; a bar plot of Cook's Distances for each of the observations; and a plot of the standardised residuals against the leverage values. Are there any obvious problems with these plots? **(2 marks)**

(b) Is week an important factor in the model in part (a)? Present appropriate R output to support your conclusion. Re-fit the model in part (a), without week as a factor. Produce a plot of the data with rrate on the vertical axis and size on the horizontal axis, using plotting symbols as follows: B for blue questionnaires in week A; b for blue questionnaires in week B; G for green questionnaires in week A; g for green questionnaires in week B, O for orange questionnaires in week A and o for orange questionnaires in week B. What are the fitted regression lines from the reduced model (excluding week) for each of the three colours? Include these lines on the plot and also include an appropriate legend. **(2 marks)**

(c) For the reduced model in part (b), give the algebraic equation for the underlying population model, including any assumptions about the error distribution, details of transformations (if any) applied to the variables and the constraints applied to any factor variables. Present appropriate R output that gives a summary of the fitted coefficients of this model and interpret the significance of these coefficients. Are the contrasts that have been used in the model a good choice to address the research question for this experiment? **(4 marks)**

(d) Use the reduced model in part (b) to estimate the response rate for questionnaires of all three different colours which have been distributed in a car park with 250 spaces and find 95% confidence intervals for these estimates. **(2 marks)**

(e) Compare the reduced model in part (b) with a multiplicative model that includes an interaction term between the factor variable (colour) and the covariate (size). Describe how this changes the algebraic equation in part (c). Is this additional term a significant improvement to the model? Present some R output and give full details of an appropriate hypothesis test. What do your results suggest about the relationship between the response rates and the explanatory variables and factors? **(2 marks)**

Question 1 continued

(f) Fit the reduced model in part (b) to the data for week A only and repeat the analysis in part (c). What would you conclude about the effect of questionnaire colour on response rates, if there had only been one round of this experiment? **(3 marks)**

(g) Now modify the reduced model in part (b) to include week as a random effect in an additive mixed effects model for the full data (not just week A). Describe the changes to the underlying population model described in part (c). Present and examine the summary output (analysis of variance table and table of coefficients) for the new mixed effects model. How has this changed from the summary output presented in part (c)? Calculate the intra-class correlation coefficient for the mixed effects model and comment on the results. **(3 marks)**

(h) Finally, discuss the results of all the above analysis. You might decide to discuss the fit of the various models, whether or not there has been an appropriate treatment of each of the variables and/or aspects of the experimental design. (Hint – we are after a good concise discussion of the important issues. This part of the question is worth as many marks as most of the other parts, so very short discussions will not get full marks; though long and winding discussions that miss the important points and even worse, cause you to exceed the overall page limit, will also not get full marks). **(2 marks)**