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

GENERALISED LINEAR MODELS

(STAT3015/STAT4030/STAT7030)

Assignment 2 for 2016

Instructions

- This assignment is worth 20% 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 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 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 21 October 2016**. You may ask the tutor or me (Ian McDermid) questions about this assignment, in person, up to the deadline (3 pm on Friday 21 October 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 20 October 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 20 October 2016. Even with an extension, all assignments must be submitted reasonably close to the original deadline of 3 pm on Friday 21 October 2016 as the assignment solutions will be released and discussed in week 13 (which starts Monday 24 October 2016).

Question 1 **(16 marks)**

Question 1 of Assignment 1 for this year (2016) involved fitting an ordinary (normally distributed) linear model to the fruitfly data in the text file fruitfly.csv, which is available on Wattle. Refer to the model solutions for Assignment 1, which are also available on Wattle. The model solutions argue that the best ordinary (normally distributed) linear model for these data is the additive fixed effects model from part (b) of question 1 of the previous assignment.

However, there were arguably still some problems with the fit of this model. Modify this model, which is described algebraically in part (d) of question 1 of the previous assignment, so that the model becomes a (non-normally distributed) generalised linear model (GLM) – you will need to decide the best family to use to model the error distribution and the appropriate link function with which to transform the response variable.

(a) For your chosen GLM, present an algebraic description of the underlying population model. Also, present the model object (i.e. present some output in which you have simply typed the name of the model, so that we can see the details of the sample/fitted model that you have chosen). Briefly outline how you decided on your chosen model. **(2 marks)**

(b) For your chosen GLM, present a plot of (suitably) standardised residuals against the linear predictor values and identify on this plot any observations that you consider to be outstanding. Discuss any observations you decide to identify. Also produce a normal quantile plot of the standardised residuals and if you consider that there is some issue with possible outliers, also produce some outlier plot. Use these plots to assess the overall fit of your chosen GLM. **(4 marks)**

(c) Present the analysis of deviance table and present a hypothesis test on the residual deviance from your chosen GLM to decide if there is any evidence of significant over or under-dispersion. Do the results of this test confirm your assessment of the overall fit of the model? **(3 marks)**

(d) In the analysis of deviance table, examine the drop-in-deviance associated with each of the terms in your model and give details of hypotheses tests to determine the significance of including each term (and the associated variables) in the model. Also present the table of coefficients and briefly comment on what your model suggests about the relationship between the response variable and the explanatory variables. Are the results from the two tables consistent? **(3 marks)**

(e) Categorise male fruit flies with Thorax lengths in the range [0.64, 0.73] as "small"; in the range [0.74, 0.84] as "medium"; and in the range [0.85, 0.94] as "large". The data includes 21 "small" sized fruit flies (with a median Thorax length of 0.68 mm); 59 "medium" sized flies (median 0.82 mm); and 45 "large" flies (median 0.88). Using these medians to represent a typical fruit fly of each size category, use your chosen GLM to estimate the expected Longevity for each combination of size and the five different levels of Activity. Find both 95% confidence intervals and 95% prediction intervals for these estimates and compare the estimates with the mean and the range of observed Longevity values for each combination of size and Activity. Present your results in a table and comment on how well your model estimates the observed data. **(4 marks)**

Question 2 **(17 marks)**

Probably the most famous maritime disaster of the twentieth century was the sinking of the RMS Titanic after it hit an iceberg at 11:40pm on 14 April 1912. Details of the disaster are in a series of related articles on *Wikipedia* (https://en.wikipedia.org/wiki/RMS_Titanic), which are both extensive and (unusually) well referenced.

One of the main internet references used in the Wikipedia article is the *Encyclopedia Titanica* (www.encyclopedia-titanica.org). In both the *Encyclopedia Titanica* and the other Titanic related articles on *Wikipedia* (which are linked to the main article) there are extensive lists of the passengers and crew (both the survivors and the victims), but neither source appears to have complete lists. There are also numerous inconsistencies between the sources; typical of internet data compiled by different people from a variety of sources.

The data in the Excel spreadsheet file RMStitanic2016.xlsx have been compiled by collating data from both the above internet sources. I first started collating these data a few years ago to present a talk to commemorate the 100th anniversary of the sinking and since then I have been constantly revising the data.

The questions and model solutions for Assignment 2 of 2015 are available on Wattle. In question 2 of this old Assignment 2, I asked students to analyse an earlier version of the Titanic data and fit an appropriate generalised linear model (GLM) to examine how the survival of the passengers (crew survival was definitely different) related to their age, sex and passenger class. My preferred GLM for modelling passenger survival is included in the file of R Code: Assignment2_2015_Q2.R.

My most recent version of the Titanic data are also available on Wattle – make sure you have the current (2016) files, do not use the older versions of my data which are included with the older (2015) assignment. In the Excel spreadsheet RMStitanic2016.xlsx, data on the survival of the passengers is summarised using an Excel Pivot Table and this summary has been saved in the file titanic2016.csv. To understand these data, you should examine both of the above internet sources, the Excel spreadsheet and the stored R code.

The new version of the aggregate or summary data in the file titanic2016.csv includes an indicator variable English: equal to 1 for a group of passengers, if Home_Country = "England"; and is 0 otherwise. For this Assignment, you need to modify my preferred GLM for passenger survival to include this indicator variable. The aim is to form a new GLM which can be used to test some of the key assertions in the internet article *"English manners cost Titanic lives"* (ABC Science, Wednesday 28 January 2009). There is a link to this article on Wattle. I chose this particular article as it includes a link to the original paper, which resulted in this article and a number of other media articles around the same time.

For various reasons, if being English (and having English "manners") did lead to lower survival rates, then the effect on survival may have been different for different groups of passengers, based on their age, sex and passenger class. There is also a suggestion that speaking English (presuming that all of the English spoke English) may have resulted in better survival in "steerage" (passengers in third class cabin accommodation). This suggestion is in paragraph 12 of the section on "Departure of the lifeboats (00:45-02:05)" in the Wikipedia article on "Sinking of the RMS Titanic" and a footnote attributes this reference to: Howells, RP (1999). *The Myth of the Titanic*. New York: Palgrave Macmillan.

Your task is to review and modify the stored R code for question 2 of Assignment 2 for 2015 to model the revised data and examine the effects on survival of being English, controlling for the effects of age, sex and passenger class. In your answer, address the detailed questions shown on the following page:

Question 2 continued

(a) Experiment with possible models that control for effects of age, sex and passenger class (i.e. models that include those variables as explanatory variable) and which include the English indicator variable. If being English had different effects on survival for different ages, sexes and passenger classes, then you made need to include interaction terms between these variables and the English indicator variable. Choose just one final GLM to address the research question, and present a couple of analysis of deviance tables (no more than two or three) for candidate models to justify your choice of final model.

Remember that if you wish to address a research question about the effects of being English on survival, then your final model must include at least a main effects term involving the English indicator variable. Present the model object for your chosen model (i.e. present some output in which you have simply typed the name of the model, so that we can see the details of the sample/fitted model that you have chosen). **(2 marks)**

(b) For your chosen GLM, present a plot of (suitably) standardised residuals against the linear predictor values and identify on this plot any observations that you consider to be outstanding. Also produce a normal quantile plot of the standardised residuals and if you consider that there is some issue with possible outliers, also produce some outlier plot. Use these plots to assess the overall fit of your chosen GLM. Discuss any outliers that you decide to identify and discuss what was different about the survival of passengers in the group represented by this observation. If a potential outlying group consists of just one or two passengers, locate the name of these passengers in the data and look up their biographies on *Encyclopedia Titanica*. Do these biographies suggest what might have been unusual about the survival of these passengers? **(4 marks)**

(c) Present the analysis of deviance table and present a hypothesis test on the residual deviance from your chosen GLM to decide if there is any evidence of significant over or under-dispersion. Do the results of this test confirm your assessment of the overall fit of the model? **(3 marks)**

(d) In the analysis of deviance table, examine the drop-in-deviance associated with each of the terms in your model and give details of hypotheses tests to determine the significance of including each term (and the associated variables) in the model. Also present the table of coefficients and briefly comment on what your model suggests about the relationship between the response variable and the explanatory variables. Are the results from the two tables consistent? **(3 marks)**

(e) Finally, the file passengers2016.csv, available on Wattle, contains individual level data on the passengers. Note that the main explanatory variables all have slightly different names to avoid confusion with the aggregate data in titanic2016.csv and I have added two additional explanatory variables: ESC (for "English-speaking country"), to indicate if a person from that Home_Country could be expected to speak English; and Nat_Group (for "Nationality group"), a categorical variable which groups the countries listed in Home_Country into broad 1912 geographical groups.

Use this individual level passenger data to fit a binary response GLM to explore differences in survival for different groupings of nationalities. Choose just one model and present the model object and an analysis of deviance table for your chosen model. Discuss the fit of your model and your conclusions, but do not present a lot of other output unless it is directly relevant to the discussion. **(4 marks)**