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 2017

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 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 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. 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 20 October 2017**. You may ask the tutor or me (Ian McDermid) questions about this assignment, in person, up to the deadline (3 pm on Friday 20 October 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 19 October 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 19 October 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 12, when the assignment solutions will be released and discussed.

Question 1 **(15 marks)**

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 Question 1 of this Assignment come from such a repository: OzDASL, the Australasian Data and Story Library, which is part of the Statistical Science Web (www.statsci.org/data/index.html), maintained by Gordon Smyth, a member of the Statistical Society of Australia.

Data on "Ear Infections in Swimmers" are available on Wattle in the file earinf.txt, or can be downloaded from OzDASL (http://www.statsci.org/data/oz/earinf.html). The data were collected in Sydney, New South Wales (NSW), which is a large city on the east coast of Australia with a number of suburban surf beaches, that are used by the residents of Sydney and visitors for recreational swimming. Wastewater (stormwater run-off and treated sewerage) is disposed of via outlets offshore from the beaches. Some wastewater also ends up in the rivers and bays that surround Sydney, many of which are also used for swimming.

In 1990, the NSW Water Board conducted a pilot Surf/Health survey of 287 swimmers, which collected the following variables:

Swimmer	whether the survey respondent reported themselves to be a frequent ("Freq") or an occasional ("Occas") swimmer
Location	where the person usually swims ("Beach" or "NonBeach")
Age	the swimmer's age group ("15-19", "20-24" or "25-29")
Sex	the swimmer's sex ("Male" or "Female")
Infections	the number of self-diagnosed ear infections. Over half of the respondents reported zero infections; however, 13 of the 287 reported more than 5 infections, with 2 people reporting 16 and 17 infections, respectively.

(a) Use R to fit the first of the two GLMs described on OzDASL. Produce a series of residual plots for this model [hint: do not do anything fancy, the default plots will be fine in this instance] and comment on these plots. **(2 marks)**

(b) An Analysis of Deviance table is presented for this model on OzDASL. Based on this table, is there any evidence of significant under or over-dispersion for this model? Conduct an appropriate hypothesis test and comment on your results. **(2 marks)**

(c) If there is evidence of significant under or over-dispersion, present a suitably corrected Analysis of Deviance table. Does this corrected table suggest any possible refinements that you might make to this model? **(2 marks)**

(d) Refine the model as suggested in part (c) and change the link function to a square root transformation. Examine the fitted variance weights from this model. Can you explain what is going on? In particular, why has the weights= option not been used? **(2 marks)**

(e) Present residual plots, an Analysis of Deviance table and other summary output for the new refined model in part (d) [hint: now I am expecting residuals and tables to be suitably standardised and corrected for under or over-dispersion, wherever possible]. Use these pieces of R output to discuss the overall fit of this model. **(5 marks)**

(f) What are the safest situations to go swimming in Sydney? Where should NSW Water concentrate on improving the water quality? Produce 95% confidence and prediction intervals for these situations and comment on your results. **(2 marks)**

Question 2 **(15 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.

There are also other sources readily available on the internet, including the Titanic Inquiry Project (www.titanicinquiry.org), which includes both the original American and British inquiries into the disaster and the *Encyclopedia Titanica* (www.encyclopedia-titanica.org), which contains extensive biographies of everyone involved. In other Titanic related articles on *Wikipedia* (which are linked to the main article) and in the other internet sources, there are extensive lists of the passengers and crew (both the survivors and the victims), but there are numerous inconsistencies between the sources; which is typical of internet data compiled by different people using a variety of sources.

The data in the Excel spreadsheet file RMStitanic2017.xlsx (available on Wattle) have been compiled by collating data from all 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 for both 2015 and 2016 are also available on Wattle. In Question 2 of both these old assignments, I asked students to analyse earlier versions of the Titanic data and to fit a series of generalised linear models (GLMs) to examine how the survival of the passengers (crew survival was definitely different) related to their age, sex and passenger class. My preferred GLMs for modelling passenger survival are included in the files of R code that accompany these old assignments.

My most recent versions of the Titanic data are also available on Wattle – for this assignment, make sure you use the files marked with the current year (2017), do not use the older versions of my data which are included with the older assignments. To understand the data, you should examine the above internet sources, and the various materials from the old assignments.

Other statisticians and data analysts have also shown an interest in the data. In particular, Kaggle (www.kaggle.com), which conducts competitions involving "real-world machine learning problems", uses a version of the Titanic data as their "entry-level competition" (https://www.kaggle.com/c/titanic/data). You may have to join Kaggle to access this last link, which contains a description of the data. It doesn't cost any money to join Kaggle (you just have to agree to receive the occasional e-mail), but in case you don't want to do this, I have made a copy of this web-page available on Wattle.

Kaggle divides the Titanic data into a training set; which you should use to build a model; and a test set; which has the key survival data omitted – you have to use your model to predict whether each passenger in the test data survived or not. I have combined the Kaggle training and test data and matched it with my version of the passenger data. I have added in some of the variables from my data used in the old assignments and I have also added in the missing survival indicator for the test set (which is definitely against the purpose of the Kaggle competition). The combined data are available on Wattle in the file titanic_combined2017.csv.

Note that two groups of crew members had passenger cabins – the 9 members of the "Guarantee" group from shipbuilders Harland & Wolff, who had cabins in 1st or 2nd class (and who all died in the disaster) and the 8 members of the "Orchestra", who all had 2nd class cabins (and who also all died in the disaster). The Kaggle data includes the "Guarantee" group amongst the passengers, but excludes the "Orchestra". I have renamed the ID variable and the Age variable in the Kaggle data to distinguish them from my versions. There are numerous inconsistencies between the two Age variables – I possibly still have some work to do on my version, but unlike the Kaggle version, there are no missing values in my version.

Question 2 continued

(a) Read the combined data into R. One passenger is described as having a 2nd class cabin in the Kaggle data, but has a 1st class cabin in my version of the data. Who is this passenger? Look up the relevant biography in the *Encyclopedia Titanica*. Does this explain the discrepancy? **(1 mark)**

(b) Separate the data into the Kaggle training and test sets. With the training portion of the data, experiment with fitting a binary response GLM that relates passenger survival to age, sex and passenger class and also makes use of one or more of the additional explanatory variables available in the Kaggle data [hint: start with one of the models described in part (e) of Question 2 in Assignment 2 for 2016]. Present an appropriate Analysis of Deviance table and discuss whether or not your chosen additional Kaggle variables are a significant addition to the model. **(2 marks)**

(c) Chose one of the binary response GLMs you experimented with in part (b). Present summary output for your chosen model and discuss why you chose that particular model. Present a series of residual plots for your chosen model [hint: consider using the binnedplot() function from library(arm) when presenting the main residual plot] and use these plots to discuss the overall fit of your model. **(5 marks)**

(d) Present the analysis of deviance table for your chosen model in part (c). Does it make sense to test for over or under-dispersion in the context of this model? **(2 marks)**

(e) Assuming you chose a binary response GLM with the default logit link function in part (c), a linear predictor value of greater than 0 will indicate a passenger who the model predicts is likely to have survived and a linear predictor value of less than 0 will indicate a passenger who is likely to have not survived the disaster. How should you interpret the fitted values from your model? Classify the passengers in the training data into predicted survivors (those that your model predicts are more likely to survive) and predicted non-survivors and compare these numbers with the observed data on survival [hint: round the fitted values from your model to the nearest whole number and use the table() function to count the numbers in the different categories].

Calculate the true positive rate (*sensitivity*) of your chosen model on the training data, i.e. the proportion of passengers your model correctly predicted as a survivor out of the number of passengers in the training set who actually survived. Also calculate the true negative rate (*specificity*), i.e. the correctly predicted non-survivors as a proportion of the actual non-survivors. Also calculate the overall *accuracy* of your chosen model on the training data, i.e. the total number of both survivors and non-survivors correctly predicted by your model, as a proportion of the total number of passengers in the training set. **(2 marks)**

(f) Finally, use the model you fitted to the training data to predict the likely survival of passengers in the test data [note: use your chosen model from part (c), which you fitted to the training data, do NOT re-fit the same model to the test data or to the combined data]. Use the actual survival in the test data to calculate the *sensitivity*, *specificity* and *accuracy* of your chosen model on the test data and compare with the results of part (e). Discuss these results. Do you think you are likely to win the Kaggle competition with your chosen GLM? **(3 marks)**
