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

GENERALISED LINEAR MODELS

(STAT3015/STAT7030)

Assignment 2 for 2015

Instructions

- This assignment is worth 20% of your overall marks for this course (for all students, enrolled in either STAT3015 or STAT7030). If you wish, you may work together with another student in doing the analyses and present joint reports for the two questions in this assignment. If you choose to do this, then both of you will be awarded the same total mark. A STAT3015 student may work with a STAT7030 student. 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.
- You should submit separate reports for Question 1 (to be marked by Xu Shi) and Question 2 (to be marked by Ian McDermid). 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 both reports. Please submit your report for Question 1 to the submission box for this course marked with Xu Shi's name and your report for Question 2 in the box with my name (Ian McDermid). There will be NO online submission for Assignment 2. Remember to keep a copy of both your assignment reports.
- Reports 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). The reports 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.
- Reports should be submitted using the relevant submission assignment box located next to the RSFAS office by **4 pm on Friday 23 October 2015**. You may ask your tutor or me (Ian McDermid) questions about this assignment, in person, up to the deadline (4 pm Friday 23 October 2015), 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 posted on Wattle or sent to me via e-mail will be posted on Wattle, but must be received no later than 4 pm Wednesday 21 October 2015.
- 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 4 pm Wednesday 21 October 2015. Even with an extension, all assignments must be submitted reasonably close to the above deadline to allow time for the marking to be completed prior to week 13, when the assignment solutions will be released and discussed.

Question 1 **(20 marks)**

The data in the file **Geriatric.txt** were collected in a study conducted on 100 geriatric patients (subjects) all of whom were at least 65 years of age and in reasonably good health.

Half of the subjects were randomly assigned to one of the two interventions:
education only (**training** = 0) or education plus aerobic exercise training (**training** = 1).

Three control variables were also collected for each subject: **gender** (0=female, 1=male), a **balance** index (higher values indicate greater stability), and a **strength** index (higher values indicate greater strength). Each subject kept a diary recording the number of **falls** during the six months of the study.

With the number of **falls** as the response variable, use *R* to fit an appropriate generalised linear model to these data. Refine this model until you have a model that you feel is a reasonable fit to the data. Present and discuss the following *R* output for your chosen model:

(a) 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 model you have chosen). Briefly outline how you decided on your chosen model.

(b) Present an appropriate plot to illustrate your model or the residuals from your model (i.e. just one plot – you choose which plot is the important one to discuss) and briefly discuss any interesting features of this plot. For this example, you should probably try to use some of the methods covered in the *R* examples discussed in lectures, and apply these methods to identify which observations or residuals on the plot correspond to which values of the response and/or explanatory variables.

(c) Examine the residual deviance from your model: is there any evidence of significant over or under-dispersion?

(d) Present the analysis of deviance table and the table of coefficients and briefly comment on what your models suggests about the relationship between the response variable and the explanatory factors. In particular, did the aerobic **training** intervention have a significant effect on the number of **falls**, controlling for **gender**, **balance** and/or **strength**? Do all of the variables in your model have a significant effect on the response variable?

(e) Use your chosen model to estimate the expected number of **falls** for each of the different combinations of **training** and **gender** and for the mean values of the other control variables (**balance** and **strength**) for each of these combinations, and find 95% confidence intervals for these estimates. Given the results of your analyses in parts (b) and (c), is it appropriate to use your chosen model to make these predictions?

Question 2 **(20 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. There are more details on the disaster in the relevant Wikipedia article (http://en.wikipedia.org/wiki/Titanic), which is 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 in other Titanic related articles on Wikipedia (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 **titanicnew.xls** 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. My most recent data are available on Wattle – make sure you have the right files, as older versions of my data (and data collated by other people) have been used for previous assignments. Data on just the passengers are summarised using an *Excel* Pivot Table and this summary has been saved in the text file **titanicnew.txt**. To understand these data, you should examine both of the above internet sources and the *Excel* spreadsheet.

Parts (a) to (e) below present a more general version of the issues addressed in parts (a) to (e) of Question 1. Fit an appropriate generalised linear model (GLM) to the summary data on passengers in **titanicnew.txt**; to examine how the survival of the passengers relates to their age, sex and passenger class. Address the issues in parts (a) to (d) below, in relation to your chosen model, as parts (a) to (d) of this question.

The file **passengers.txt** contains individual level rather than aggregate or summary data (the original source is still the same data as was summarised in **titanicnew.txt**). In part (e) of this question, discuss the issues in part (e) below and also fit your chosen GLM from parts (a) to (d) to the individual level data. Compare confidence intervals for suitable predictions using both fitted versions of the model (the model fitted to the aggregate data in **titanicnew.txt** and the same model fitted to the individual level data in **passengers.txt**) and discuss any differences between the two versions.

(a) 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 model you have chosen). Then give a brief description of your chosen model and also briefly discuss any decisions you made in refining your chosen model. You are not required to find a model that is a perfect fit to the data (which may not be possible using the techniques you have covered so far in this course), so do not spend a lot of time on what may be a futile attempt to find the "perfect" model, simply choose ONE reasonable model to present and discuss.

(b) Present appropriate plots of the residuals from your model and use these plots to briefly discuss the fit of your chosen model. Identify and discuss any (problem) observations that stand out on these residual plots. You may also examine some related statistics to further investigate any problems you identify with the fit of the model (for example, a test for significant over or under-dispersion). Do not present any plot or statistic that you do not discuss, however, do briefly mention in your discussion, any diagnostics that you examined where there were no obvious problems.

(c) Present appropriate summary output for your model. Include the analysis of deviance table and a table of coefficients and any other output you consider necessary to illustrate the results of your analysis. Discuss what this summary output indicates about possible answers to the underlying research questions (you may have to decide for yourself what these are, in the absence of clear instructions from the "client").

(d) Choose some meaningful way to present the model and the results of your analysis for a general (non-statistical) audience. For example, you might choose to present and discuss a plot of the data on the original scale with your model superimposed on the plot (preferably with suitable confidence intervals).

(e) Given your assessment of the fit of the model in response to issue (b), discuss how persuasive (conclusive) you consider the results presented in parts (c) and (d). Would your chosen model be appropriate to use for making predictions? Illustrate this discussion by using the model to make some predictions (e.g. for "typical" groups in your data).