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

GENERALISED LINEAR MODELS

(STAT3015/STAT4030/STAT7030)

Solutions to Assignment 2 for 2017

Question 1 **(15 marks)**

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.

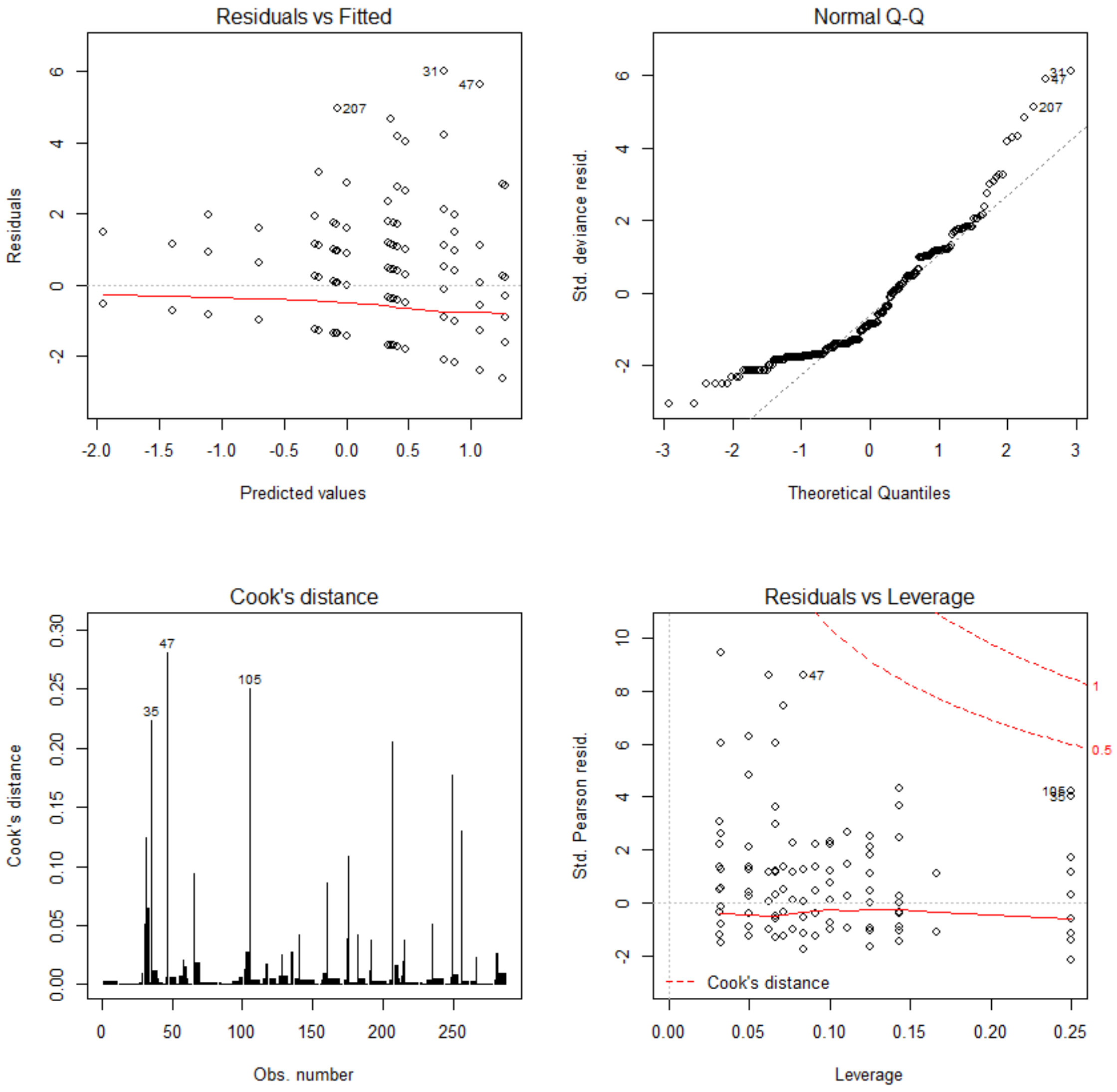
(a) Use R to fit the first of the two GLMs described on OzDASL. Produce a series of residual plots for this model [hint: do not do anything fancy, the default plots will be fine in this instance] and comment on these plots. **(2 marks)**

The plots are shown and discussed on the next page. To aid in the discussion, here is a list of the 13 observations mentioned above, each of which reported more than 5 ear infections:

```
> ear.infections[Infections>5,]
    Swimmer Location   Age    Sex Infections
29    Occas NonBeach 15-19   Male          6
30    Occas NonBeach 15-19   Male         11
31    Occas NonBeach 15-19   Male         16
35    Occas NonBeach 15-19 Female         10
47    Occas NonBeach 20-24   Male         17
65    Occas NonBeach 25-29   Male         10
105    Freq NonBeach 15-19 Female          6
160   Occas    Beach 15-19   Male          9
174   Occas    Beach 15-19 Female          6
175   Occas    Beach 15-19 Female          9
207   Occas    Beach 25-29   Male          9
215   Occas    Beach 25-29 Female          6
249    Freq    Beach 15-19 Female         10
>
> table(Infections)
Infections
  0   1   2   3   4   5   6   9  10  11  16  17
151  40  39  26  13   5   4   3   3   1   1   1
```

Not surprisingly, a number of the observations in this list are highlighted on the plots.

Question 1, part (a) continued



The apparent curved line structure in the main Residuals vs Fitted plot is a result of the fact that the response variable, a count of the number of reported ear infections, only takes on a limited number of values (see the table at the bottom of the previous page) – there are curves for 0 infections, 1 infections, 2 infections and so on as you go from the bottom to the top of the graph.

With an assumed dispersion of 1, the residuals are already approximately standardised (or you could look at the Normal Q-Q plot or the Residuals vs Leverage plot and see the actual standardised deviance or Pearson residual values). A number of the observations have very large positive (standardised) residuals – the highlighted ones appear in the table on the previous page and the others are probably also in this list.

The Normal Q-Q plot is based on a relatively large sample size and shows some departures for the assumption of asymptotic normality, suggesting potential problems with how the model is modelling the dispersion (error variance). The Cook's distance and Residuals vs Leverage plots suggest this is a problem with a relatively large group of observations, rather than just one or two potential "outliers", which reinforces the idea that the model is not correctly modelling the error variance.

Question 1 continued

(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)**

```
> round(anova(glm.inf,test="F"),2)
Analysis of Deviance Table

Model: poisson, link: log

Response: Infections

Terms added sequentially (first to last)

                         Df Deviance Resid. Df Resid. Dev     F Pr(>F)    
NULL                                       286     824.51                 
Swimmer                   1    34.70       285     789.81 34.70 <2e-16 ***
Location                  1    25.16       284     764.65 25.16 <2e-16 ***
Age                       2     8.58       282     756.07  4.29   0.01 ** 
Sex                       1     0.63       281     755.43  0.63   0.43    
Swimmer:Location          1     1.69       280     753.74  1.69   0.19    
Swimmer:Age               2     6.38       278     747.36  3.19   0.04 *  
Location:Age              2     3.92       276     743.44  1.96   0.14    
Swimmer:Sex               1     0.23       275     743.21  0.23   0.63    
Location:Sex              1    11.12       274     732.09 11.12 <2e-16 ***
Age:Sex                   2     1.78       272     730.31  0.89   0.41    
Swimmer:Location:Age      2     3.67       270     726.63  1.84   0.16    
Swimmer:Location:Sex      1     0.24       269     726.39  0.24   0.62    
Swimmer:Age:Sex           2     0.19       267     726.20  0.10   0.91    
Location:Age:Sex          2    13.94       265     712.26  6.97 <2e-16 ***
Swimmer:Location:Age:Sex  2     8.54       263     703.72  4.27   0.01 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Warning message:
In anova.glm(glm.inf, test = "F") :
  using F test with a 'poisson' family is inappropriate
```

The current version of R does not produce the output shown on OzDASL. The above “F” statistics and p-values are actually mislabelled χ^2 statistics, assuming that the dispersion equals 1. However, the residual deviance and degrees of freedom are the same as shown on OzDASL and can still be used in the usual "goodness of fit" test:

```
> glm.inf$deviance/glm.inf$df.residual
[1] 2.67574
> glm.inf$deviance
[1] 703.7197
> glm.inf$df.residual
[1] 263
> c(qchisq(0.025, glm.inf$df.residual), qchisq(0.975, glm.inf$df.residual))
[1] 219.9720 309.8145
```

An alternative estimate of the dispersion is found by dividing the residual deviance by the residual degrees of freedom:

$$\hat{\phi}_{alt} = \frac{703.72}{263} = 2.68$$

This is considerably larger than 1, suggesting over-dispersion, which we can confirm with a formal test of the following hypotheses:

$$H_0 : \phi = 1 \quad H_A : \phi \neq 1$$

The observed residual deviance (scaled using the assumed dispersion of 1) is 703.72 and lies well above a two-sided 95% interval for the χ^2 distribution with 263 degrees of freedom (219.97, 309.82), so we reject the null hypothesis and conclude there is evidence of significant over-dispersion. The tail of observations with a very large number of reported ear infections may not have affected estimation of the “Poisson” mean, but have inflated the estimate of the variance, so that it is larger than the mean.

Question 1 continued

(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)**

For the data to follow a Poisson distribution, then the mean should be the same as the variance. The variance problems appear to the result of a reasonable sized group of observations, rather than just one or two outliers, so the data appear to be genuinely over-dispersed. Instead of trying to modify the model, we could adjust for this “extra Poisson variance” by adjusting the model output to account for the over-dispersion. We could get the current version of R to produce the approximate “F” table shown on OzDASL (see the appendix of R commands for details), but my preferred approach is χ^2 inference using the alternative estimate of the dispersion (which is also an approximation, see the appendix for more discussion):

```
> alt.est.disp <- glm.inf$deviance/glm.inf$df.residual
> anova(glm.inf, dispersion=alt.est.disp, test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: Infections

Terms added sequentially (first to last)

                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                                        286     824.51              
Swimmer                    1   34.699       285     789.81 0.0003169 ***
Location                   1   25.160       284     764.65 0.0021664 ** 
Age                        2    8.582       282     756.07 0.2011651    
Sex                        1    0.635       281     755.43 0.6262242    
Swimmer:Location           1    1.693       280     753.74 0.4264137    
Swimmer:Age                2    6.383       278     747.36 0.3033909    
Location:Age               2    3.920       276     743.44 0.4807044    
Swimmer:Sex                1    0.227       275     743.21 0.7706846    
Location:Sex               1   11.120       274     732.09 0.0414909 *  
Age:Sex                    2    1.783       272     730.31 0.7166244    
Swimmer:Location:Age       2    3.674       270     726.63 0.5033329    
Swimmer:Location:Sex       1    0.239       269     726.39 0.7650498    
Swimmer:Age:Sex            2    0.194       267     726.20 0.9644218    
Location:Age:Sex           2   13.943       265     712.26 0.0738675 .  
Swimmer:Location:Age:Sex   2    8.538       263     703.72 0.2028249    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The above scaled drop-in-deviance tests are significant for the mains effects of both Swimmer (p-value 0.00032 < α = 0.05) and Location (p-value 0.0022 < α = 0.05):

$$\mathrm{H}_0 : \tau_{reference\ level} = \tau_{other\ level} = 0 \quad \mathrm{H}_\mathrm{A} : \text{not both } \tau_j = 0$$

$$\text{Test Statistic}: \frac{\Delta_{Deviance}}{\hat{\phi}_{CV}} \sim \chi_1^2\left(0.95\right) = 3.84$$

In both these cases we would reject the null hypotheses and conclude that these terms in the model are leading to significantly different numbers of infections. Conversely, the similar scaled drop-in-deviance test associated with most of the other terms in the model suggest that these terms could be refined out of the model. The only exception is the marginally significant (p-value 0.041 < α = 0.05) two-way Swimmer: Location interaction term. This differs from the “F” table shown on OzDASL, where this term has a p-value of 0.06, however, once we have made the change to the link function suggested in part (d) and carefully refined the model by deleting one non-significant term at a time, we can refine the model to one which just has main effects for Swimmer and Location.

Question 1 continued

(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)**

```
> infections.glm

Call:  glm(formula = Infections ~ Location + Swimmer, family = poisson(sqrt))

Coefficients:
     (Intercept)  LocationNonBeach      SwimmerOccas  
          0.8479            0.2820            0.3418  

Degrees of Freedom: 286 Total (i.e. Null);  284 Residual
Null Deviance:          824.5 
Residual Deviance: 767.1     AIC: 1145
> unique(infections.glm$weights)
[1] 4
```

If the Poisson distribution is a good model for the observed counts (which arguably it may not be, as the model is still over-dispersed), then the square root transformation should have the effect of stabilising the variance. It is not difficult to show that:

If $Y \sim \text{Poisson}\left(\text{mean } \mu \text{ and variance } \sigma^2 = \lambda\right)$ and

$U = g(Y) = \sqrt{Y} = Y^{\frac{1}{2}}$, and $g'(Y) = \tfrac{1}{2}Y^{-\frac{1}{2}} = \dfrac{1}{2\sqrt{Y}}$, then by the delta method

$E(U) = E\left[g(Y)\right] \approx g(\mu) = \sqrt{\lambda}$ and

$Var(U) = Var\left[g(Y)\right] \approx \left[g'(\mu)\right]^2 \sigma^2 = \left[\dfrac{1}{2\sqrt{\lambda}}\right]^2 \lambda = \tfrac{1}{4}$, which is a constant.

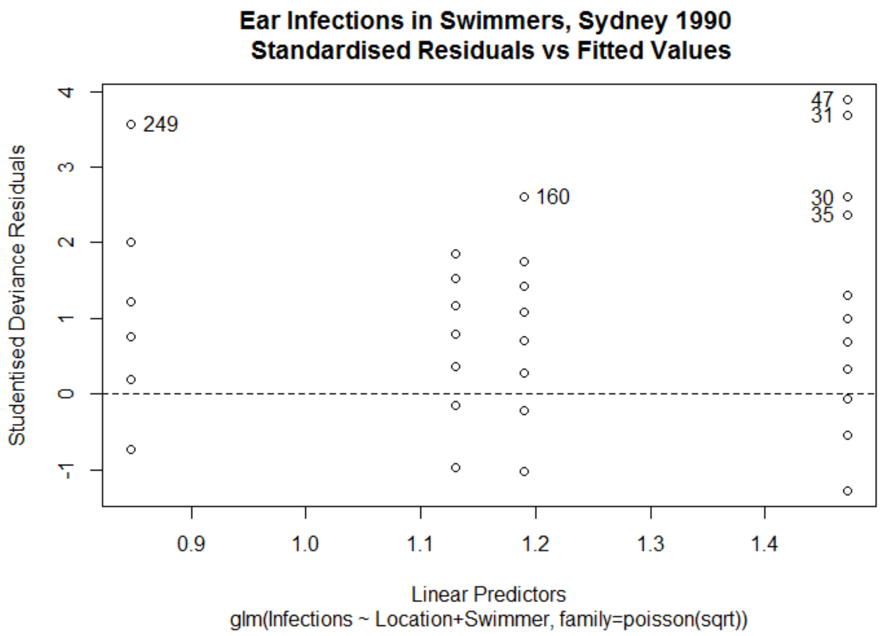
If you examine the variance weights for any Poisson count model with a square root link function, you will find they are all set at 4, the reciprocal of the above variance. All GLMs incorporate variance weights, but here we are dealing with Poisson counts, not with Poisson rates measured on different sized groups, so there is no need to adjust the weights by the size of the groups in order to make the dispersion a constant equal to 1 (the additional weights in this instance would all be 1 anyway, as the observations all represent individuals or groups of size 1).

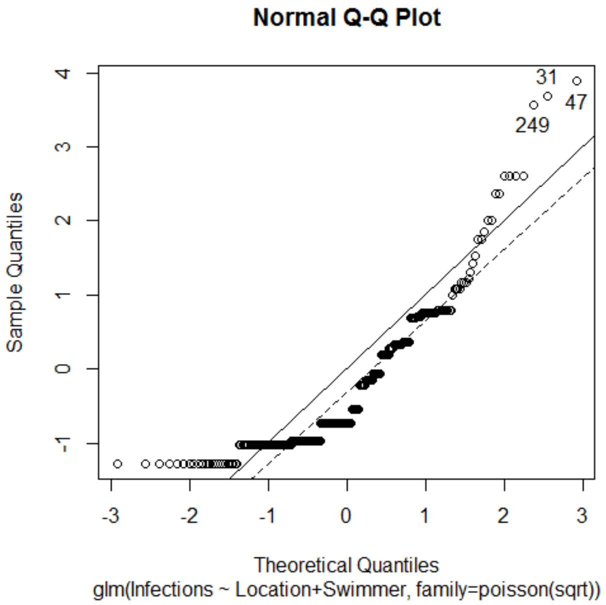
(e) Present residual plots, an Analysis of Deviance table and other summary output for the new refined model in part (d) [hint: now I am expecting residuals and tables to be suitably standardised and corrected for under or over-dispersion, wherever possible]. Use these pieces of `R` output to discuss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

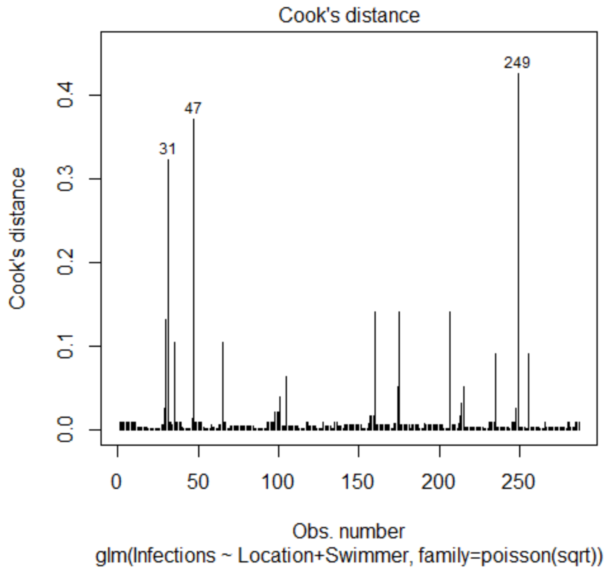
The residual plots are shown and discussed on the next page. Again the points identified on those plots are from the list of observations with more than 5 infections:

```
> dataplusmodel <- data.frame(ear.infections, Fitted=fitted(infections.glm),
"Std Residuals"=std.residuals, "Cooks D"=cooks.distance(infections.glm))
> dataplusmodel[Infections>5, ]
    Swimmer Location   Age    Sex Infections    Fitted Std.Residuals    Cooks.D
29    Occas NonBeach 15-19   Male          6 2.1658150      1.306261 0.02480357
30    Occas NonBeach 15-19   Male         11 2.1658150      2.601545 0.13167428
31    Occas NonBeach 15-19   Male         16 2.1658150      3.687136 0.32290528
35    Occas NonBeach 15-19 Female         10 2.1658150      2.363625 0.10355132
47    Occas NonBeach 20-24   Male         17 2.1658150      3.887772 0.37127471
65    Occas NonBeach 25-29   Male         10 2.1658150      2.363625 0.10355132
105    Freq NonBeach 15-19 Female          6 1.2766404      1.847677 0.06264838
160   Occas    Beach 15-19   Male          9 1.4153727      2.603937 0.14030939
174   Occas    Beach 15-19 Female          6 1.4153727      1.747381 0.05126563
175   Occas    Beach 15-19 Female          9 1.4153727      2.603937 0.14030939
207   Occas    Beach 25-29   Male          9 1.4153727      2.603937 0.14030939
215   Occas    Beach 25-29 Female          6 1.4153727      1.747381 0.05126563
249    Freq    Beach 15-19 Female         10 0.7189492      3.571312 0.42544614
```

Question 1, part (e) continued







In the list on the previous page, a number of the observations (with the same number of infections) have the same fitted values and standardised residuals as others in the list (65 is the same as 35 and both 175 and 207 are the same as 160), so are hidden “behind” the identified points on the main residual plot. All observations that reported a large number of infections (9, 10, 11, 16 or 17) have standardised residual values (supposedly corrected for the over-dispersion) in excess of 2.36 standard deviations and the fitted values appear to be consistently under-estimating the observed infections (218 of the 287 or 76% of the fitted values are less than 2), so it appears that we have still have some problems with the “mean” part of the model, as well as with the “variance” part of the model.

The model is still over-dispersed and the problems with the “variance” model and the over-dispersion are still evident in the Normal Q-Q plot and the stand-outs on the Cook’s distance plot are the two observations that reported 16 and 17 infections (observations 31 and 47) and the one frequent female beach swimmer that reported a large number of infections (observation 249).

Question 1, part (e) continued

The test for over-dispersion and a suitably adjusted Analysis of Deviance table are:

```
> infections.glm$deviance
[1] 767.0611
> infections.glm$df.residual
[1] 284
> c(qchisq(0.025, infections.glm$df.residual),
 qchisq(0.975, infections.glm$df.residual))
[1] 239.2108 332.5759
>
> alt.est.disp <- infections.glm$deviance/infections.glm$df.residual
> alt.est.disp
[1] 2.700919
>
> anova(infections.glm, dispersion=alt.est.disp, test="Chisq")
Analysis of Deviance Table

Model: poisson, link: sqrt

Response: Infections

Terms added sequentially (first to last)

         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       286     824.51
Location  1   24.149       285     800.36 0.0027881 **
Swimmer   1   33.298       284     767.06 0.0004461 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The remaining terms in the model are highly significant. There is no easy way to adjust the summary output for the over-dispersion, but the results are at least consistent, and the important coefficients are both positive, suggesting that there are more infections amongst “Non-Beach” swimmers rather than “Beach” swimmers and also more infections amongst “Occas” rather than “Freq” swimmers:

```
> summary(infections.glm)

Call:
glm(formula = Infections ~ Location + Swimmer, family = poisson(sqrt))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0813  -1.5979  -1.1991   0.5353   6.3550

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.84791    0.05107  16.603  < 2e-16 ***
LocationNonBeach  0.28198    0.05905   4.775 1.80e-06 ***
SwimmerOccas      0.34179    0.05904   5.789 7.07e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 824.51  on 286  degrees of freedom
Residual deviance: 767.06  on 284  degrees of freedom
AIC: 1145.5

Number of Fisher Scoring iterations: 5

>
> round(tapply(fitted(infections.glm), list(Location, Swimmer), mean), 2)
         Freq Occas
Beach    0.72  1.42
NonBeach 1.28  2.17
```

Finally, the fitted values in each of the four Location by Swimmer categories are similar to those shown on OzDASL for their reduced model. However, given the underlying problems, both models are probably better used as exploratory models for examining relationships rather than as reliable predictive models.

Question 1 continued

(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)**

Despite my reservations about using this model for prediction, here are predictions for the four Location by Swimmer categories:

```
> data.frame(newd,Sample=as.vector(ssize),Obs.Mean=as.vector(obs.means),Min=as.vector(mins),
Max=as.vector(maxs),Estimate=temp2$fit^2,CI=temp2$ci.fit^2,PI=temp2$pi.fit^2)
  Location Swimmer Sample  Obs.Mean Min Max  Estimate  CI.lower  CI.upper PI.lower  PI.upper
1    Beach    Freq     72 0.8194444   0  10 0.7189492 0.5585896 0.8995174        0  7.945779
2 NonBeach    Freq     71 1.1408451   0   6 1.2766404 1.0586270 1.5150514        0  9.615124
3    Beach   Occas     75 1.2800000   0   9 1.4153727 1.1893091 1.6610939        0  9.989024
4 NonBeach   Occas     69 2.3478261   0  17 2.1658150 1.8761632 2.4762502        0 11.851914
```

The model appears to be doing a reasonable job of predicting the observed means, which are all within the confidence intervals, and the prediction intervals are very wide (as might be expected), but still don't cover the most extreme values.

These data are not particularly useful for answering the above research questions. If I interpret the first question as being "where should I personally go swimming", then the prediction intervals (for an individual prediction) are wide, but it appears I could still expect to get some ear infections, but slightly less infections if I go swimming at a "Beach" location, rather than a "Non-Beach" location, regardless of which type of swimmer I am. This still doesn't answer the question of which "Beach"? [Personally I tend to do most of my swimming in Sydney at Coogee, rather than Bondi, even now they have extended the wastewater outlet, so that it comes out a lot further off-shore from Bondi].

There is also a suggestion in the above results that frequent swimmers get less infections than occasional swimmers at the same locations, but is that because I am more likely to become a frequent swimmer (rather than an occasional swimmer), only if I am someone who is less susceptible to getting ear infections?

Similarly, NSW Health presumably wants to concentrate on cleaning up the areas that generate the highest average number of infections as measured by the confidence intervals and the confidence intervals for "Non-Beach" are higher than (and not overlapping with) the confidence intervals for the "Beach" areas, for both types of swimmers. But again this doesn't address the issue of which "Non-Beach" area (the Nepean-Hawkesbury, the Harbour, Botany Bay, the Georges River, Cook's River etc)? [Personally, I would prohibit swimming in the Cook's River, which looks distinctly unpleasant, and concentrate on cleaning up some of the more popular swimming spots, but you would need to collect different, more detailed data to start addressing these issues.]

Question 2 **(16 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)**

```
> table(passengers$Pclass, passengers$Class)

    1st Class 2nd Class 3rd Class
  1       323         0         0
  2         1       276         0
  3         0         0       709
>
> passengers[passengers$Pclass==2&passengers$Class=="1st Class",]
     Kaggle_Set Kaggle_Id Survived Pclass                                         Name
1297       test      1297        1      2 Nourney, Mr. Alfred (Baron von Drachstedt")"
      Sex Kaggle_Age SibSp ParCh        Ticket    Fare Cabin Embarked ID_No Age
1297 male         20     0     0 SC/PARIS 2166 13.8625   D38        C  1481  20
       Age_Group      Group     Class Home_Country English ESC Nat_Group   Boarded
1297 20to24years Passengers 1st Class      Germany       0   0  European Cherbourg
```

The biography of Mr Alfred Nourney (alias "Baron von Drachstedt") on the *Encyclopedia Titanica* says that he boarded as a second class passenger, but requested and paid for an upgrade to first class, so he was a first class passenger (as per my version of the data) at the time of the disaster. Fitting Class as a factor variable (my version) rather than Pclass, (which by default would be treated a continuous covariate) also allows for different effects between 1st and 2nd class than between 2nd and 3rd class. I also prefer to use Age over Kaggle_Age, because of all the missing values in the latter.

(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)**

Here is an Analysis of Deviance table that I suspect a lot of students would have considered:

```
> train <- passengers[passengers$Kaggle_Set=="train",]
> attach(train)

> train.glm3 <- glm(Survived ~ Age + Sex + Class + Age:Sex + Sex:Class +
 Age:Class + SibSp + ParCh, family = binomial)
>
> anova(train.glm3, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Survived

Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        890    1186.66
Age        1    1.946       889    1184.71   0.16304
Sex        1  266.941       888     917.77 < 2.2e-16 ***
Class      2  114.069       886     803.70 < 2.2e-16 ***
SibSp      1   18.492       885     785.21 1.706e-05 ***
ParCh      1    0.459       884     784.75   0.49799
Age:Sex    1   17.432       883     767.32 2.978e-05 ***
Sex:Class  2   19.622       881     747.70 5.486e-05 ***
Age:Class  2    5.723       879     741.97   0.05718 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Question 2, part (b) continued

The above model was the third in a series of models that I fit as part of a process of model building that I have outlined with extensive comments in the `R` appendix. Reporting on the model building process is like presenting an argument for your preferred model and it usually a good idea to choose and present one model part-way through the process to present and discuss the issues encountered in the process.

It is a bit of a surprise that one of the extra variables available in the Kaggle data, `SibSp`, turns out to be highly significant, based on the drop-in-deviance test shown in the Analysis of Deviance table (p-value = 0.000017 $< \alpha = 0.05$). Before I am happy to include this variable in any model, I need to think of some reason why it might be important in explaining the relative survival of different passengers. Note that `ParCh`, which has a definition closely related to that of `SibSp`, is not significant in the Analysis of Deviance table (p-value = 0.5 $> \alpha = 0.05$).

For various reasons, which I discuss in more detail in the appendix, it would require considerable work to extract useful information out of most of the other extra variables available on Kaggle, particularly `Name`, `Ticket` and `Embarked`, and I am happy for students to rule them out. The same goes for `Cabin`, which looks a lot more promising, but is predominantly missing for 77% of the training data.

Finally, `Fare`, a possible continuous covariate, turns out on further investigation, to have some strange values – many of the large fares cover an entire group of passengers, but the same fare is recorded against all passengers in the group and some passengers with very good 1st class cabins did not have to pay for their passage (such as Mr Bruce Ismay, President of the White Star Line, “owner” of the ship). Much of the information available in the current version of `Fare` is related to `Class`, and `Fare` turns out not to be significant as an addition to a model that already contains terms involving `Class`.

(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)**

At this stage, students will hopefully present a model that includes `SibSp`, or some related derived variable, but which excludes the other variables from the Kaggle data, based on the above discussion and on a model building process such as the one I outline in the appendix.

My chosen model involves a new derived variable. I first added the variables `SibSp` and `ParCh` and then added an extra 1 (to account for the current passenger themselves) for a total estimate of the size of the family group (`Family_Size`) that each passenger was travelling with. This turned out to be a better fit treated as a categorical variable, `Family`, with categories: “Alone” (`Family_Size` = 1), “Couple” (`Family_Size` = 2), “Small” (`Family_Size` = 3 or 4) and “Large” (`Family_Size` > 4). Note that 69 out of 82 (84%) of the “Large” families were in 3rd Class.

```
> table(passengers$Family, passengers$Class)

         1st Class 2nd Class 3rd Class
  Alone        161       157       472
  Couple       104        52        79
  Large         11         2        69
  Small         48        65        89
```

Question 2, part (c) continued

Here is the required summary output for my chosen model:

```
> summary(train.glm8)

Call:
glm(formula = Survived ~ Age + Sex + Class + Family + Age:Sex + 
    Sex:Class, family = binomial)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.1114  -0.6100  -0.4136   0.3736   2.8075  

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)             4.36759    0.80522   5.424 5.83e-08 ***
Age                    -0.02196    0.01380  -1.591  0.11172    
Sexmale                -2.56332    0.91245  -2.809  0.00497 ** 
Class2nd Class         -1.19544    0.74182  -1.612  0.10707    
Class3rd Class         -3.62937    0.65247  -5.562 2.66e-08 ***
FamilyCouple           -0.04787    0.25807  -0.185  0.85285    
FamilyLarge            -2.12470    0.45967  -4.622 3.80e-06 ***
FamilySmall             0.50888    0.27190   1.872  0.06126 .  
Age:Sexmale            -0.03723    0.01745  -2.134  0.03286 *  
Sexmale:Class2nd Class -0.86051    0.83517  -1.030  0.30285    
Sexmale:Class3rd Class  1.48078    0.72121   2.053  0.04005 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1186.66  on 890  degrees of freedom
Residual deviance:  727.33  on 880  degrees of freedom
AIC: 749.33

Number of Fisher Scoring iterations: 6
```

Note that the only category of Family with a significant coefficient is the one for “Large” families (the significant p-value indicates that it differs from the reference category “Alone”), which appear to have significantly worse survival (the coefficient is negative). Apart from the fact that most of the “Large” families were in 3rd Class, where survival was worse (based on the analysis in the earlier assignments), a plausible explanation might also be that larger families were harder to accommodate in the life boats, especially if they insisted on remaining together as a group.

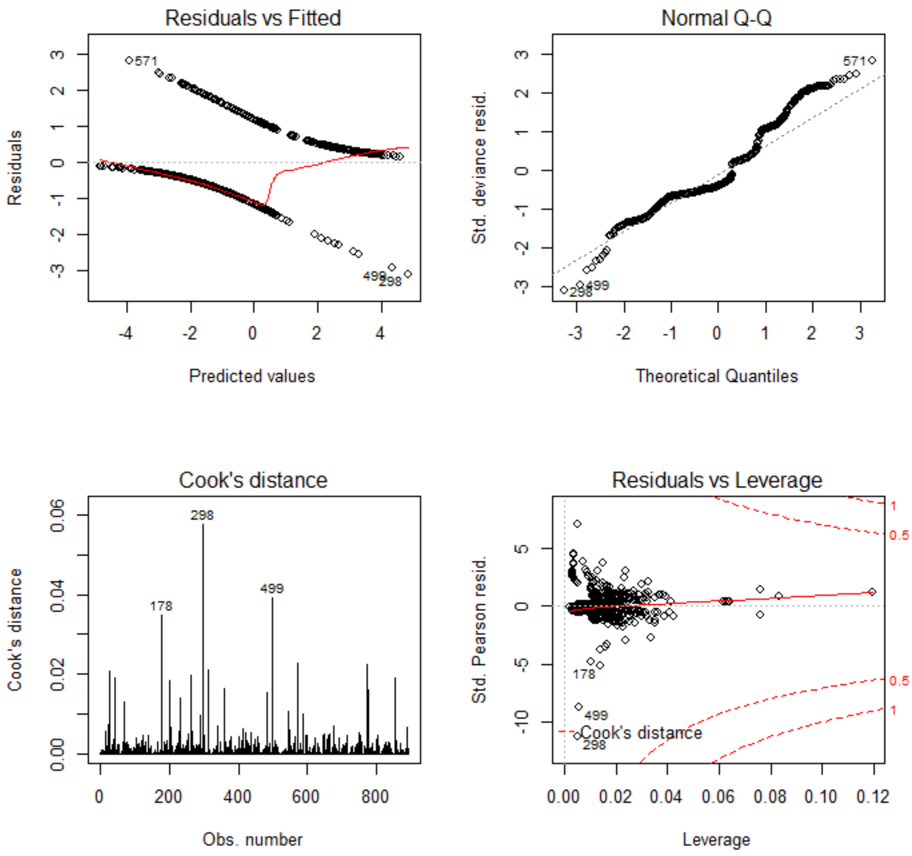
A set of residual plots produced using the plot() function are shown at the top of the next page. As I argue in a lot more detail in the comments in the appendix, when dealing with binary response models, these “standard” plots are not much help with assessing the underlying assumptions. The main residual plot consists of two curves. Passengers who survived are in the upper curve of positive residuals and those who didn't survive are in the lower curve of negative residuals:

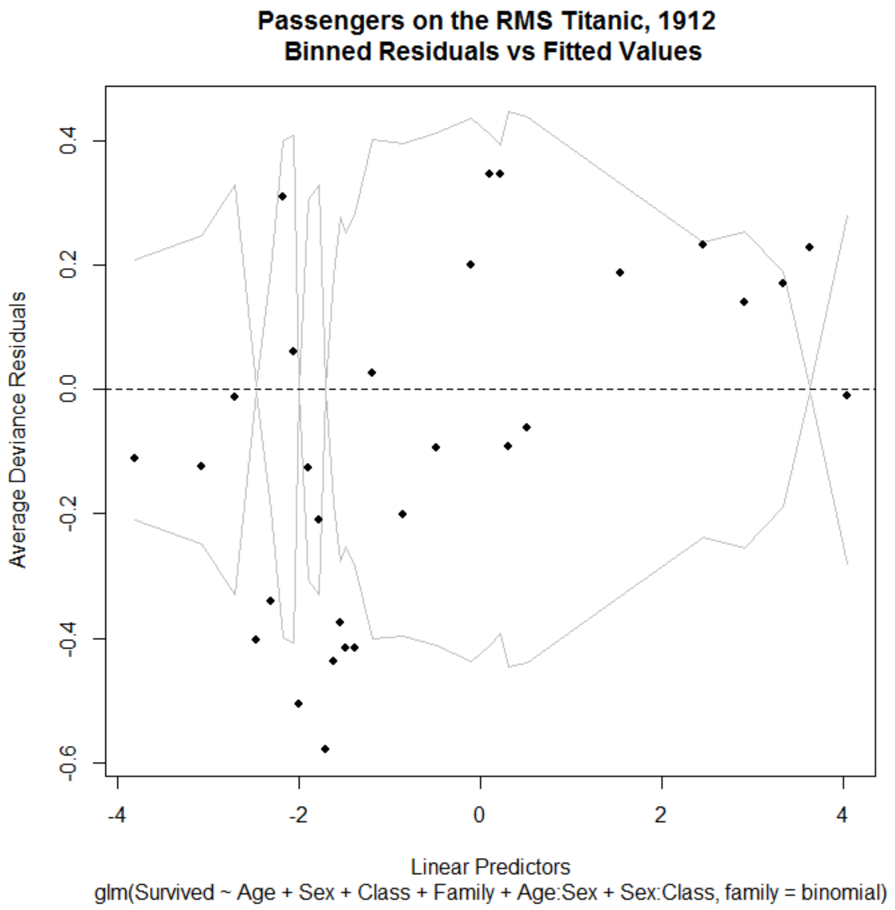
```
> range(residuals(train.glm8)[(train$Survived == 1)])
[1] 0.1420648 2.8075304
> range(residuals(train.glm8)[(train$Survived == 0)])
[1] -3.1114289 -0.1257168
```

However, if you do identify the three observations highlighted in the Cook’s distance plot, you will find that they are the only three female passengers in 1st Class (out of the passengers included in the training data) who did not survive.

I have also included a “binned” version of the main residual plot on the following page, which I argue (in a lot more detail in the appendix) does not indicate any serious departures from the assumptions of independence and constant dispersion (variance). There are better reasons, which I will discuss in part (d), why I consider the overall fit of this model to be reasonable.

Question 2, part (c) continued





Question 2 continued

(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)**

```
> anova(train.glm8, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Survived

Terms added sequentially (first to last)

          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                        890    1186.66
Age        1    1.946       889    1184.71     0.163
Sex        1  266.941       888     917.77 < 2.2e-16 ***
Class      2  114.069       886     803.70 < 2.2e-16 ***
Family     3   40.904       883     762.80 6.854e-09 ***
Age:Sex    1   15.391       882     747.40 8.740e-05 ***
Sex:Class  2   20.070       880     727.33 4.383e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The drop-in-deviance tests are all highly significant (except for the main effect term for Age, but Age is also involved in a significant two-way interaction term). The results are also consistent with the table of coefficients presented in part (c), where there were significant differences in at least one of the levels associated with each of the significant terms.

Whilst we can trust the drop-in-deviance tests, where the associated degrees of freedom are the degrees of freedom required to fit each additional term, we cannot trust the goodness of fit test which uses the residual deviance and the residual degrees of freedom in the context of a binary response model – the usual asymptotic arguments for binomial models do not apply and we do not have a good handle on the true (residual) degrees of freedom available to estimate the actual dispersion.

I think the best argument for the fit of this model is the fact that very similar models fitted to the aggregated data (where we could trust the residual plots and the goodness of fit test) were a reasonable fit. I do not believe the GLMs for these data have gone from being over-dispersed when fit to aggregate data, to being genuinely under-dispersed when fit to equivalent individual level data.

Overall, the above argument, the highly significant drop-in-deviance tests and the reasonable binned residual plot in part (c) suggest that the overall fit of this model is reasonable, making the model okay to use as an exploratory model for investigating factors that lead to passenger survival on the Titanic.

Question 2, continued

(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)**

The fitted values range from 0 to 1 and can be interpreted for each observation as the probability (estimated from the model) that the passenger is in the `Survived` = 1 category. If we round the fitted values, either down to 0 or up to 1, then we have the values of `Survived`, as predicted by the model. We can cross-tabulate these against the actual values of `Survived`:

```
> table(round(fitted(train.glm8),0), Survived)
   Survived
      0   1
  0 485  93
  1  64 249
>
> train.results <- as.vector(table(round(fitted(train.glm8),0), Survived))
> train.results
[1] 485  64  93 249
> sum(train.results)
[1] 891
>
> # So, the sensitivity (proportion of passengers correctly predicted as
> # having Survived) is:
>
> train.sensitivity <- train.results[4]/(train.results[3] + train.results[4])
> train.sensitivity
[1] 0.7280702
>
> # The specificity (proportion of passengers correctly predicted as having
> # NOT Survived) is:
>
> train.specificity <- train.results[1]/(train.results[1] + train.results[2])
> train.specificity
[1] 0.8834244
>
> # And overall accuracy is:
>
> train.accuracy <- (train.results[1] + train.results[4])/sum(train.results)
> train.accuracy
[1] 0.8237935
```

Question 2 continued

(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)**

```
> test.predicted <- predict(train.glm8,
      newdata=passengers[passengers$Kaggle_Set=="test",], type="response")
> test.Survived <- passengers[passengers$Kaggle_Set=="test",]$Survived
>
> table(round(test.predicted,0), test.Survived)
   test.Survived
      0   1
  0 213  51
  1  47 107
>
> test.results <- as.vector(table(round(test.predicted,0), test.Survived))
> test.results
[1] 213  47  51 107
> sum(test.results)
[1] 418
>
> test.sensitivity <- test.results[4]/(test.results[3] + test.results[4])
> test.sensitivity
[1] 0.6772152
> train.sensitivity
[1] 0.7280702
>
> test.specificity <- test.results[1]/(test.results[1] + test.results[2])
> test.specificity
[1] 0.8192308
> train.specificity
[1] 0.8834244
>
> test.accuracy <- (test.results[1] + test.results[4])/sum(test.results)
> test.accuracy
[1] 0.7655502
> train.accuracy
[1] 0.8237935
```

As expected, the results on the test data are all lower than the results on the training data. It is not really sensible to use any model based on the Titanic data as a predictive model. The inquiries following the disaster started a century of improvements in ship building and maritime safety and society has certainly changed since the early twentieth century. There is no way that a model built on the Titanic data should be used to try to predict what is likely to happen in a disaster of similar size over a hundred years later.

The only point in building a predictive model is to try and win the Kaggle competition, but I suspect the best predictive accuracy you could achieve with just a GLM would be somewhere in the 75% to 80% range. You will not do any better without also employing other related techniques from a course in Big Data Analytics, Statistical Learning or Data Mining (such courses are also offered by my school, RSFAS). Many of these techniques involve over-fitting the model, but like a fully saturated GLM, over-fit models could still be used in an exploratory fashion to investigate factors that influenced survival on the Titanc.

I sincerely doubt that you could achieve accuracy close to 100% without ignoring the Kaggle request “not to cheat” by matching to known survival outcomes. With this in mind, I offer any student who has managed to beat my cross-validated accuracy of 76.5% and who sends me a copy of their code, a bonus 0.5 mark to make up for marks lost elsewhere in question 2, but anyone who claims to have close to 100% accuracy will get 0 marks for this part of question 2.