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 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)**

$$E\left[\,\text{log(Longevity)}_{ij}\,\right] = \beta_0 + \tau_j + \beta_1 \text{Thorax}_{ij}$$

Where j represents the 5 different levels of Activity : j = {"A", "B" "C", "D", "E"}, and i = 1, 2, …, 25 (equivalent to the different values of the ID variable) represents the observations within each of the five different Activity groups. In the R code, I have not specified any special contrasts, so the default treatment contrasts will have been used with Activity group "A" as the reference level, so the constraint applied is $\tau_A = 0$.

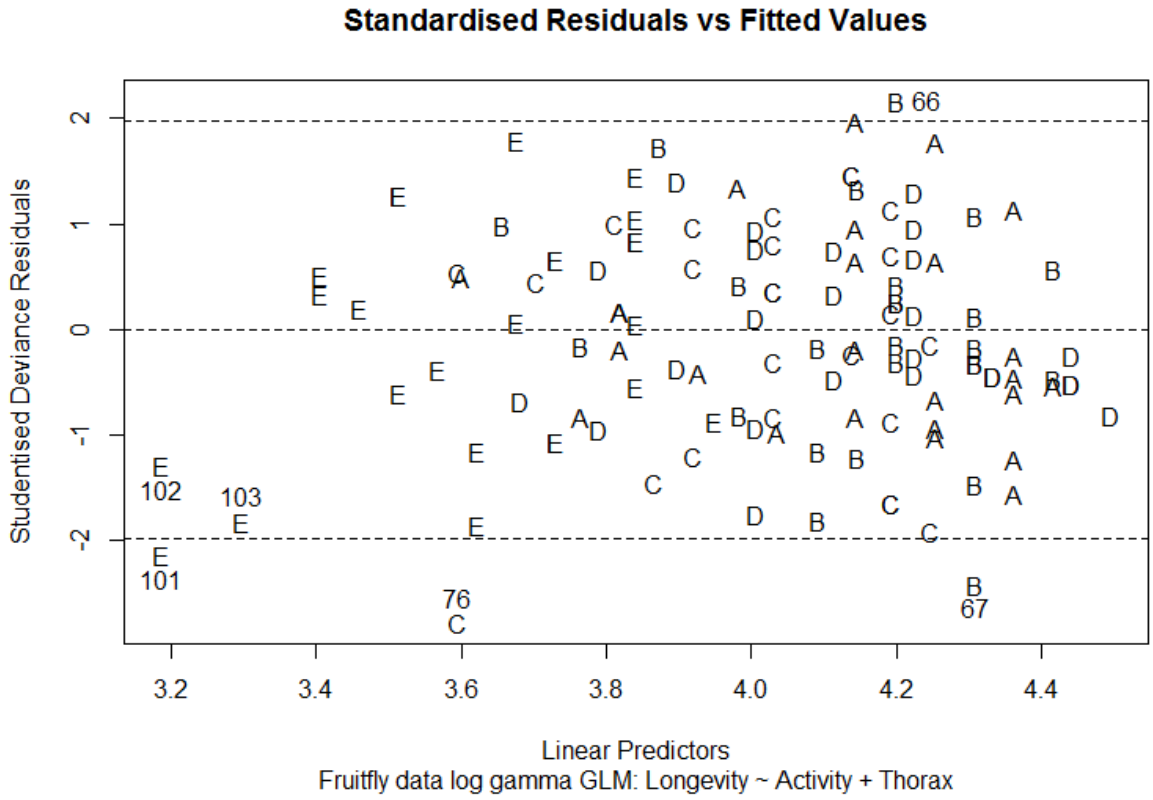
This gamma GLM with a log link function has the same "mean" model as the ordinary linear model from the previous assignment, the main difference is that the errors are now assumed to be independent and from a gamma distribution, which should allow more flexibility in modelling the variance/deviance. The model object is:

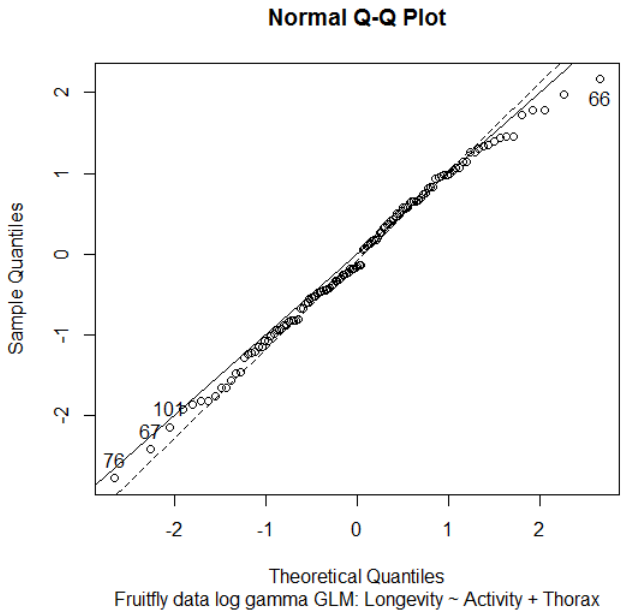
```
> fruitfly.glm

Call:  glm(formula = Longevity ~ Activity + Thorax, family = Gamma(link = log))

Coefficients:
(Intercept)    ActivityB    ActivityC    ActivityD    ActivityE       Thorax
    1.86201      0.05561     -0.11659      0.07834     -0.41380      2.71802

Degrees of Freedom: 124 Total (i.e. Null);  119 Residual
Null Deviance:          13.49
Residual Deviance: 4.331    AIC: 948.3
```

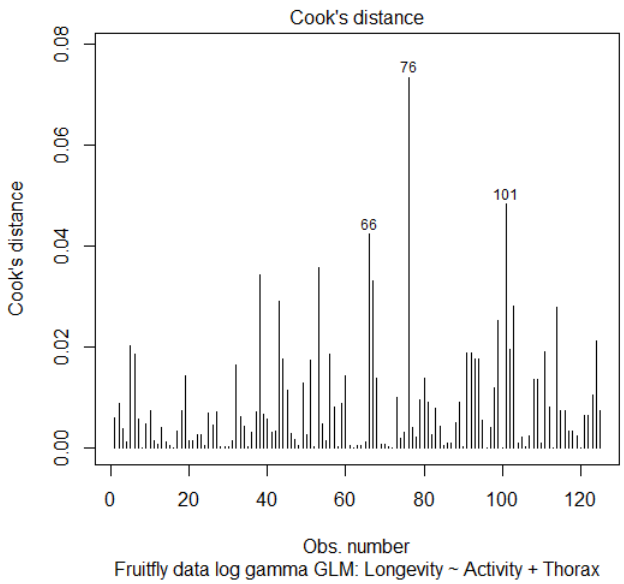
A quick check of the default residuals plots (not shown), suggests that this switch to a GLM has not introduced any new problems and may even have helped with some of the problems with the model in the previous assignment, however, we will take a closer look at these issues in the remaining parts of this question.

Question 1 continued

(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)**







The above plots look very similar to the plots for the ordinary linear model in Assignment 1, with the same observations identified (and for the same reasons), so maybe switching to a gamma GLM has not really provided any improvement.

Question 1 continued

(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? **(3marks)**

```
> anova(fruitfly.glm)
Analysis of Deviance Table

Model: Gamma, link: log

Response: Longevity

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev
NULL                       124    13.4910
Activity  4   4.3197       120     9.1713
Thorax    1   4.8401       119     4.3312
> summary(fruitfly.glm)$dispersion
[1] 0.03541252
> df <- fruitfly.glm$df.residual
> fruitfly.glm$deviance/df
[1] 0.03639623
```

The assumed dispersion used to calculate the model is given in the summary output and is shown above, this is the CV estimate: $\hat{\phi}_{CV} = 0.03541252$. An alternative estimate can be found by dividing the residual deviance by the residual degrees of freedom:

$$\hat{\phi}_{alt} = \frac{4.3312}{119} = 0.0363963$$

These two estimates are fairly close, suggesting no over or under-dispersion, but we could confirm this by comparing the residual deviance (scaled using the CV estimate) with a χ^2 distribution with the residual degrees of freedom. This comparison is a test of whether the dispersion is what it was assumed to be (the CV estimate):

$$H_0 : \phi = \hat{\phi}_{CV} \quad H_A : \phi \neq \hat{\phi}_{CV}$$

```
> fruitfly.glm$deviance/summary(fruitfly.glm)$dispersion
[1] 122.3057
> c(qchisq(0.025,df), qchisq(0.975,df))
[1]  90.69959 151.08442
```

The residual deviance lies within the central 95% interval, on the χ^2 distribution with 119 degrees of freedom, so we do not reject the null hypothesis and conclude there is no evidence of significant over or under-dispersion.

This suggests no additional problems with this GLM, but arguably the ordinary linear model in Assignment 1 was just a good as a fit to the data as this GLM, and was also arguably a simpler model.

Question 1 continued

(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)**

```
> scaled.dev <- anova(fruitfly.glm)$Deviance/summary(fruitfly.glm)$dispersion
> chisq.pvalues <- 1 - pchisq(scaled.dev, anova(fruitfly.glm)$Df)
> cbind(anova(fruitfly.glm), "Scaled Dev"=scaled.dev, "Pr(>Chi)"=chisq.pvalues)
         Df Deviance Resid. Df Resid. Dev Scaled Dev Pr(>Chi)
NULL     NA       NA       124  13.490983         NA       NA
Activity  4 4.319691       120   9.171292   121.9820        0
Thorax    1 4.840140       119   4.331151   136.6788        0
```

In terms of the underlying model in part (a), the scaled drop-in-deviance associated with Activity can be used to test:

$$\mathrm{H}_0 : \tau_A = \tau_B = \tau_C = \tau_D = \tau_E = 0 \quad \mathrm{H}_\mathrm{A} : \text{not all } \tau_j = 0$$

$$\frac{\Delta_{Deviance}}{\hat{\phi}_{CV}} = \frac{4.319691}{0.03541252} = 121.982 \quad \sim \quad \chi^2_4\left(0.95\right) = 9.488$$

And the scaled drop-in-deviance associated with Thorax can be used to test:

$$\mathrm{H}_0 : \beta_1 = 0 \quad \mathrm{H}_\mathrm{A} : \beta_1 \neq 0$$

$$\frac{\Delta_{Deviance}}{\hat{\phi}_{CV}} = \frac{4.840140}{0.03541252} = 136.679 \quad \sim \quad \chi^2_1\left(0.95\right) = 3.841$$

In both cases, as the observed test statistics are greater than the critical values from the relevant χ^2 distributions (or because the corresponding *p*-values are less than $\alpha = 0.05$), we reject the null hypotheses and conclude that the terms involving these explanatory variables are significant additions to the model.

```
> round(summary(fruitfly.glm)$coefficients,8)
              Estimate Std. Error   t value   Pr(>|t|)
(Intercept)  1.86200508 0.18996141  9.802017 0.00000000
ActivityB    0.05561073 0.05327632  1.043817 0.29868588
ActivityC   -0.11658571 0.05322714 -2.190343 0.03044901
ActivityD    0.07833938 0.05365486  1.460061 0.14690828
ActivityE   -0.41379937 0.05382648 -7.687655 0.00000000
Thorax       2.71801734 0.22272213 12.203625 0.00000000
```

The table of coefficients tells a consistent story to the analysis of deviance table, with the *t* value associated with Thorax testing the same hypothesis as above and the parameters associated with the levels of Activity testing whether that level differs from Activity A, the reference level:

$$\mathrm{H}_0 : \tau_j = 0 \quad \mathrm{H}_\mathrm{A} : \tau_j \neq 0, \quad j = \{\mathrm{B}, \mathrm{C}, \mathrm{D}, \mathrm{E}\}$$

The relevant *t* values can be all be compared to a two-tailed critical value from a Student's *t* distribution with the residual degrees of freedom: $\pm t_{119}\left(0.975\right) = \pm 1.9801$

Not all of the Activity levels differ from the reference level, but levels C and E, the two groups involving sexual activity are both significantly below the reference level, suggesting reduced Longevity in these groups, suggesting that sexual activity does lead to a reduced lifespan for male fruit flies.

Question 1 continued

(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)**

The following table required a lot of R coding (see the Appendix for details). Personally I think I prefer the graphical presentation used in Assignment 1, however this table includes confidence and prediction intervals (as well as the estimates) for the five different levels of Activity and 15 curves would have made any graph far too "busy":

```
> data.frame(DataSummary,Estimate=exp(temp2$fit),CI=exp(temp2$ci.fit),PI=exp(temp2$pi.fit))
     Size Activity Count     Mean Min Max Thorax Estimate CI.lower CI.upper PI.lower  PI.upper
1   large        A    12 71.83333  58  96   0.88 70.37657 65.16038 76.01032 48.10405 102.96140
2   large        B     9 72.00000  46  92   0.88 74.40113 68.79842 80.46010 50.84164 108.87783
3   large        C    11 64.90909  48  81   0.88 62.63189 57.99986 67.63385 42.81190  91.62766
4   large        D     6 75.33333  70  81   0.88 76.11153 70.15959 82.56839 51.97626 111.45405
5   large        E     7 51.00000  42  60   0.88 46.52831 42.84580 50.52733 31.76700  68.14883
6  medium        A     8 64.00000  47  89   0.82 59.78646 55.47442 64.43368 40.88303  87.43044
7  medium        B    13 63.00000  42  97   0.82 63.20542 58.66394 68.09848 43.22345  92.42494
8  medium        C    11 54.54545  36  68   0.82 53.20719 49.36623 57.34700 36.38351  77.81013
9  medium        D    14 64.42857  39  86   0.82 64.65844 59.99881 69.67995 44.21512  94.55394
10 medium        E    13 37.84615  26  54   0.82 39.52684 36.66915 42.60723 27.02815  57.80533
11  small        A     5 43.00000  37  47   0.68 40.86430 36.92291 45.22642 27.77345  60.12545
12  small        B     3 51.00000  42  65   0.68 43.20117 39.15375 47.66698 29.38493  63.51355
13  small        C     3 35.00000  21  44   0.68 36.36734 32.84392 40.26875 24.71396  53.51564
14  small        D     5 46.00000  35  63   0.68 44.19432 40.27543 48.49452 30.10170  64.88464
15  small        E     5 23.80000  16  33   0.68 27.01676 24.65685 29.60255 18.40811  39.65130
```

Even though some of the estimates are a little different to the observed means, the observed means mainly lie within 95% confidence intervals, and the full range of observed data values typically lie within the 95% prediction intervals.

Notable exceptions are for the high sexual activity group (E) in both the "large" and "small" categories. Residuals from this group stood out on the residual plots (and in the previous assignment), so there are some caveats on the usefulness of this model for making predictions at the extreme ends of the data.

However, the general trends are obvious and comparisons between the two sexual activity groups (E = high activity, C = low activity) and the three control groups (note that typically the 95% confidence intervals do not over-lap between groups E, C and the other categories), lead to the conclusion that sexual activity does indeed lead to lower longevity in male fruit flies.

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.

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)**

The aggregate data has changed from the data used in the 2015 Assignment, but re-fitting the final preferred model from that assignment to the new data, results in a model in which all the key terms are still significant (not shown, but details included in the Appendix). Adding the English indicator variable to this model gives the following analysis of deviance table:

```
> anova(titanic2.glm, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Prop.Surv

Terms added sequentially (first to last)


                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                88     722.76
Mean.Age          1     1.81        87     720.95   0.17869
Sex               1   367.52        86     353.43 < 2.2e-16 ***
P.Class           2   151.44        84     201.99 < 2.2e-16 ***
English           1     2.27        83     199.72   0.13185
Mean.Age:Sex      1    25.45        82     174.27 4.531e-07 ***
Sex:P.Class       2    30.51        80     143.75 2.369e-07 ***
Mean.Age:P.Class  2     8.76        78     135.00   0.01255 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

I also experimented with all possible interactions between English and the other variables (again details not shown, but available in the Appendix), but after refining this model, we arrive back at the above model.

Note that this model includes two non-significant terms: a main effects term for Mean.Age, which is included as age is also involved in two significant interactions terms; and the English indicator variable, which is required if we want to use this model to address the research question.

Note that if we assume that the model used in the 2015 Assignment correctly controls for the effects of age, sex and passenger class on survival, then the conclusion from the above appears to be that being English did not have a significant effect on survival, once we have controlled for these other important factors.

Question 2, part (a) continued

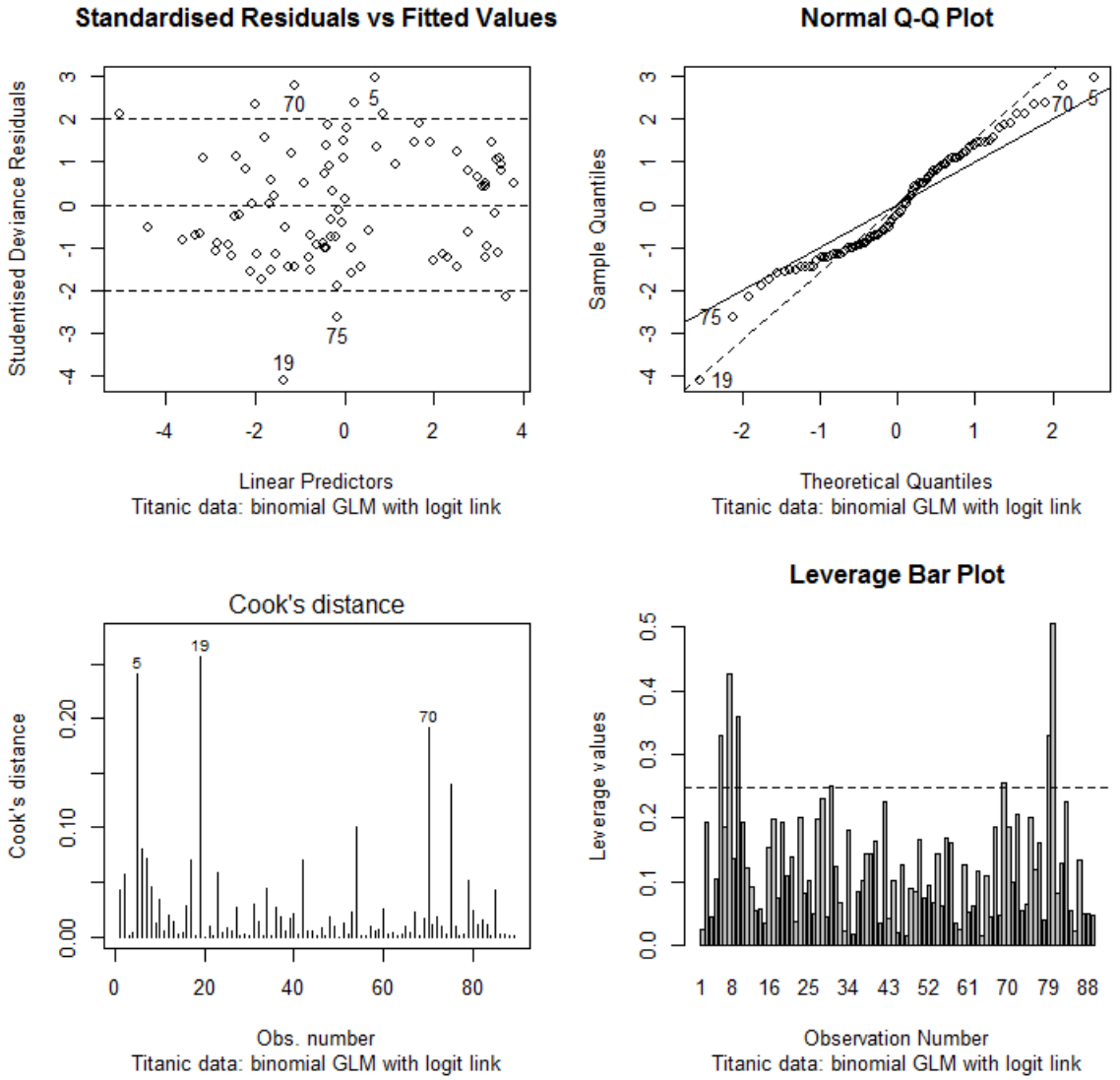
Now some more output – not really necessary to the above argument, but they do complete the requirements of the question. Firstly, an extra analysis of deviance table, which clearly shows that the English indicator variable has no effects, even in combination with some of the other factors (i.e. there are no apparent additional effects to being English, even within particular age, sex and passenger class groups):

```
> anova(titanic3.glm, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Prop.Surv

Terms added sequentially (first to last)


                         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                        88     722.76
Mean.Age                  1     1.81        87     720.95   0.17869
Sex                       1   367.52        86     353.43 < 2.2e-16 ***
P.Class                   2   151.44        84     201.99 < 2.2e-16 ***
English                   1     2.27        83     199.72   0.13185
Mean.Age:Sex              1    25.45        82     174.27 4.531e-07 ***
Sex:P.Class               2    30.51        80     143.75 2.369e-07 ***
Mean.Age:P.Class          2     8.76        78     135.00   0.01255 *
Mean.Age:English          1     0.58        77     134.42   0.44744
Sex:English               1     1.45        76     132.97   0.22774
P.Class:English           2     3.61        74     129.35   0.16444
Mean.Age:Sex:English      1     2.19        73     127.16   0.13857
Sex:P.Class:English       2     1.40        71     125.76   0.49591
Mean.Age:P.Class:English  2     1.72        69     124.04   0.42416
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

And the model object for my chosen model:

```
> titanic2.glm

Call:  glm(formula = Prop.Surv ~ Mean.Age + Sex + P.Class + Mean.Age:Sex +
    Sex:P.Class + Mean.Age:P.Class + English, family = binomial,
    weights = Passengers)

Coefficients:
           (Intercept)                Mean.Age                   Sexmale
             3.658e+00              -8.183e-03              -2.525e+00
         P.Classsecond            P.Classthird                 English
             4.936e-01              -3.469e+00              -3.111e-01
      Mean.Age:Sexmale   Sexmale:P.Classsecond    Sexmale:P.Classthird
             -3.442e-02              -7.645e-01               1.702e+00
Mean.Age:P.Classsecond   Mean.Age:P.Classthird
             -5.150e-02              -5.262e-05

Degrees of Freedom: 88 Total (i.e. Null);  78 Residual
Null Deviance:          722.8
Residual Deviance: 135      AIC: 295.6
```

(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)**



There are numerous potential outliers, with 9 of the 89 residuals lying more than 2 standard deviations away from the model, which represents more than 10% of the data, suggesting possible problems with over-dispersion, which we will further investigate in part (c), below.

On the above plots, I have only identified 4 of the most extreme observations, which all have absolute standardised residual values in excess of 2.5.

Question 2, part (b) continued

The 4 identified observations are:

```
> titanic[c(5,19,70,75),]
     Age.Group Mean.Age P.Class    Sex English Survived Passengers
5  00to13years      2.1  second   male       0        7          7
19 14to19years     17.4   third   male       0        2         56
70 40to49years     45.8   first   male       1        9         17
75 40to49years     42.5   third female       0        2         13
```

Survival appears to have relatively good for observations 5 and 70 and relatively poor for observations 19 and 75. The Cook's distances for these observations are large, but not extreme. Observations 5 and 70 (just) have more than twice the average leverage on the leverage bar plot, but they are by no means the most influential observations in the data.

All four of the above groups involve more than just a couple of passengers, but it wouldn't hurt to go sample a few of the relevant biographies on *Encyclopedia Titanica*., so that you appreciate that there were real people behind some of these "statistics".

(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)**

```
> anova(titanic2.glm)
Analysis of Deviance Table

Model: binomial, link: logit

Response: Prop.Surv

Terms added sequentially (first to last)


                Df Deviance Resid. Df Resid. Dev
NULL                                88     722.76
Mean.Age         1     1.81        87     720.95
Sex              1   367.52        86     353.43
P.Class          2   151.44        84     201.99
English          1     2.27        83     199.72
Mean.Age:Sex     1    25.45        82     174.27
Sex:P.Class      2    30.51        80     143.75
Mean.Age:P.Class 2     8.76        78     135.00
> summary(titanic2.glm)$dispersion
[1] 1
> df <- titanic2.glm$df.residual
> titanic2.glm$deviance/df
[1] 1.730733
```

For a binomial GLM, the assumed dispersion is 1. An alternative estimate can be found by dividing the residual deviance by the residual degrees of freedom:

$$\hat{\phi}_{alt} = \frac{135.0}{78} = 1.730733$$

Which is larger than the assumed dispersion, confirming the suggestion of over-dispersion in part (b), but again, we could confirm this by comparing the residual deviance (scaled using the CV estimate) with a χ^2 distribution with the residual degrees of freedom:

Question 2, part (c) continued

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

```
> titanic2.glm$deviance/summary(titanic2.glm)$dispersion
[1] 134.9972
> c(qchisq(0.025,df), qchisq(0.975,df))
[1]  55.46562 104.31594
```

The residual deviance lies above the central 95% interval, on the χ^2 distribution with 78 degrees of freedom, so we do reject the null hypothesis and conclude that there is evidence of significant over-dispersion with this model. We could attempt to compensate for this over-dispersion by using the alternative estimate of the dispersion as the scaling factor in the drop-in-deviance tests, rather than the assumed dispersion of 1, when we examine the analysis of deviance table in part (d).

(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)**

```
> scaled.dev <- anova(titanic2.glm)$Deviance/(titanic2.glm$deviance/df)
> chisq.pvalues <- 1 - pchisq(scaled.dev, anova(titanic2.glm)$Df)
> cbind(anova(titanic2.glm), "Scaled Dev"=scaled.dev, "Pr(>Chi)"=chisq.pvalues)
                 Df   Deviance Resid. Df Resid. Dev Scaled Dev     Pr(>Chi)
NULL             NA         NA        88   722.7580         NA           NA
Mean.Age          1   1.808497        87   720.9495   1.044931 0.3066774404
Sex               1 367.517131        86   353.4324 212.347661 0.0000000000
P.Class           2 151.442969        84   201.9894  87.502207 0.0000000000
English           1   2.270555        83   199.7189   1.311903 0.2520503342
Mean.Age:Sex      1  25.453605        82   174.2653  14.706834 0.0001255903
Sex:P.Class       2  30.511517        80   143.7538  17.629245 0.0001485451
Mean.Age:P.Class  2   8.756566        78   134.9972   5.059455 0.0796807295
```

The above table is adjusted to use the alternative estimate of the dispersion as the scaling factor. The original unadjusted analysis of deviance table was presented and discussed in part (a). The only thing that has changed is that one of the second order interaction terms is now not significant, but I do not feel that justifies making further refinements to the model.

The key test on the term involving the English indicator variable is still not significant:

$$\mathrm{H}_0 : \beta_{English} = 0 \quad \mathrm{H}_\mathrm{A} : \beta_{English} \neq 0$$

$$\frac{\Delta_{Deviance}}{\hat{\phi}_{alt}} = \frac{2.270555}{1.730733} = 1.311903 \quad \sim \quad \chi^2_1(0.95) = 3.841$$

As the observed test statistic is less than the critical value from the χ^2 distribution (or because the corresponding *p*-value is greater than $\alpha = 0.05$), we do not reject the null hypotheses and conclude that the term involving this explanatory variables is not a significant addition to the model, so there is no evidence that being English had any bearing on passenger survival, once we have controlled for the other factors already included in the model (a fairly complicated combination of age, sex and passenger class). If the researchers behind the referenced news article found such evidence, I suspect it was because they had not properly controlled for the other important factors affecting passenger survival.

Question 2, part (d) continued

```
> round(summary(titanic2.glm)$coefficients,8)
                           Estimate Std. Error     z value   Pr(>|z|)
(Intercept)             3.65839698 0.80519249  4.54350607 0.00000553
Mean.Age               -0.00818347 0.01747401 -0.46832267 0.63955386
Sexmale                -2.52539036 0.73013608 -3.45879408 0.00054260
P.Classsecond           0.49356003 0.96215708  0.51297240 0.60797063
P.Classthird           -3.46870785 0.77032116 -4.50293725 0.00000670
English                -0.31110386 0.19262926 -1.61503947 0.10630215
Mean.Age:Sexmale       -0.03442153 0.01429890 -2.40728620 0.01607157
Sexmale:P.Classsecond  -0.76447346 0.67210461 -1.13743226 0.25535763
Sexmale:P.Classthird    1.70214665 0.56347077  3.02082513 0.00252087
Mean.Age:P.Classsecond -0.05149793 0.02119013 -2.43027953 0.01508718
Mean.Age:P.Classthird  -0.00005262 0.01621125 -0.00324602 0.99741006
> c(qt(0.025,df),qt(0.975,df))
[1] -1.990847  1.990847
```

The table of coefficients tells a consistent story to the analysis of deviance table, with some (but not all) of the differences between the levels involved in the various interaction terms being significant. The key test on the term involving the English indicator variable is still not significant, with the associated *t* value testing the same hypothesis as the above scaled drop-in-deviance test (as it is a simple binary indicator variable). This *t*-test has not been adjusted for the over-dispersion, but as with the adjusted drop-in-deviance test, the adjustment would simply increase the associated *p*-value, which is already considerably greater than $\alpha = 0.05$.

(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)**

If you fit binary response models to the individual data, you get models that have very similar coefficients to equivalent binomial GLMs, fitted to appropriately aggregated data (the differences are mainly due to rounding that occurs when you summarise some of the covariates, for example Mean_Age is a categorised version of Age). To fit binomial GLMs you do need to re-aggregate the data in an appropriate fashion. The big disadvantage of binary response models (which are arguably easier to fit), is that the residual plots and the goodness-of-fit test on the residual deviance that we applied in parts (b) and (c) respectively, to check the models, are no longer useful.

ESC and Nat_Group are like the English indicator variable used in the earlier parts of this question, in that all three are different ways of re-categorising the information available in the Home_Country variable, which has 45 different levels. It only makes sense to use one of these variables in the same model. Including ESC in place of the English indicator variable, produces very similar results (not shown, but details included in the Appendix) and there are no significant differences in survival between passengers from ESC and from non-ESC countries. Including Nat_Group gives more interesting results:

Question 2, part (e) continued

```
> anova(passengers3.glm, 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                         1301     1734.3
Age           1     2.18      1300     1732.1   0.13942
Gender        1   367.32      1299     1364.8 < 2.2e-16 ***
Cabin         2   152.45      1297     1212.3 < 2.2e-16 ***
Nat_Group     6    16.42      1291     1195.9   0.01165 *
Age:Gender    1    26.14      1290     1169.8 3.183e-07 ***
Gender:Cabin  2    31.10      1288     1138.7 1.762e-07 ***
Age:Cabin     2     9.14      1286     1129.5   0.01036 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> summary(passengers3.glm)

Call:
glm(formula = Survived ~ Age + Gender + Cabin + Age:Gender +
    Gender:Cabin + Age:Cabin + Nat_Group, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6663  -0.6568  -0.4421   0.3784   3.2592

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)            3.945115   0.812676   4.854 1.21e-06 ***
Age                   -0.011650   0.017345  -0.672 0.501774
Gendermale            -2.598688   0.727402  -3.573 0.000354 ***
Cabinsecond            0.397345   0.965434   0.412 0.680654
Cabinthird            -3.787881   0.795118  -4.764 1.90e-06 ***
Nat_GroupBritish      -0.401031   0.265237  -1.512 0.130542
Nat_GroupEuropean     -0.211906   0.266060  -0.796 0.425765
Nat_GroupIrish         0.076695   0.324971   0.236 0.813429
Nat_GroupOES          -0.907137   0.453703  -1.999 0.045564 *
Nat_GroupOther         0.920069   0.512466   1.795 0.072594 .
Nat_GroupOttoman       0.498488   0.342654   1.455 0.145728
Age:Gendermale        -0.034448   0.014207  -2.425 0.015323 *
Gendermale:Cabinsecond -0.778085   0.674190  -1.154 0.248458
Gendermale:Cabinthird  1.758706   0.567489   3.099 0.001941 **
Age:Cabinsecond       -0.048594   0.021119  -2.301 0.021397 *
Age:Cabinthird         0.005472   0.016197   0.338 0.735460
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

 (Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1734.3  on 1301  degrees of freedom
Residual deviance: 1129.5  on 1286  degrees of freedom
AIC: 1161.5

Number of Fisher Scoring iterations: 6

> levels(Nat_Group)
[1] "American" "British"  "European" "Irish"    "OES"     "Other"    "Ottoman"
```

The drop-in-deviance test in the analysis of deviance table suggests there are significant differences between different nationality groups ($\chi^2_6 = 16.42$, $p = 0.01165$) and one of the "*t*-tests" ($z = -1.999$, $p = 0.045564$) suggests that the other English-speaking group, which included 2 Australians, had worse survival than the American reference group.