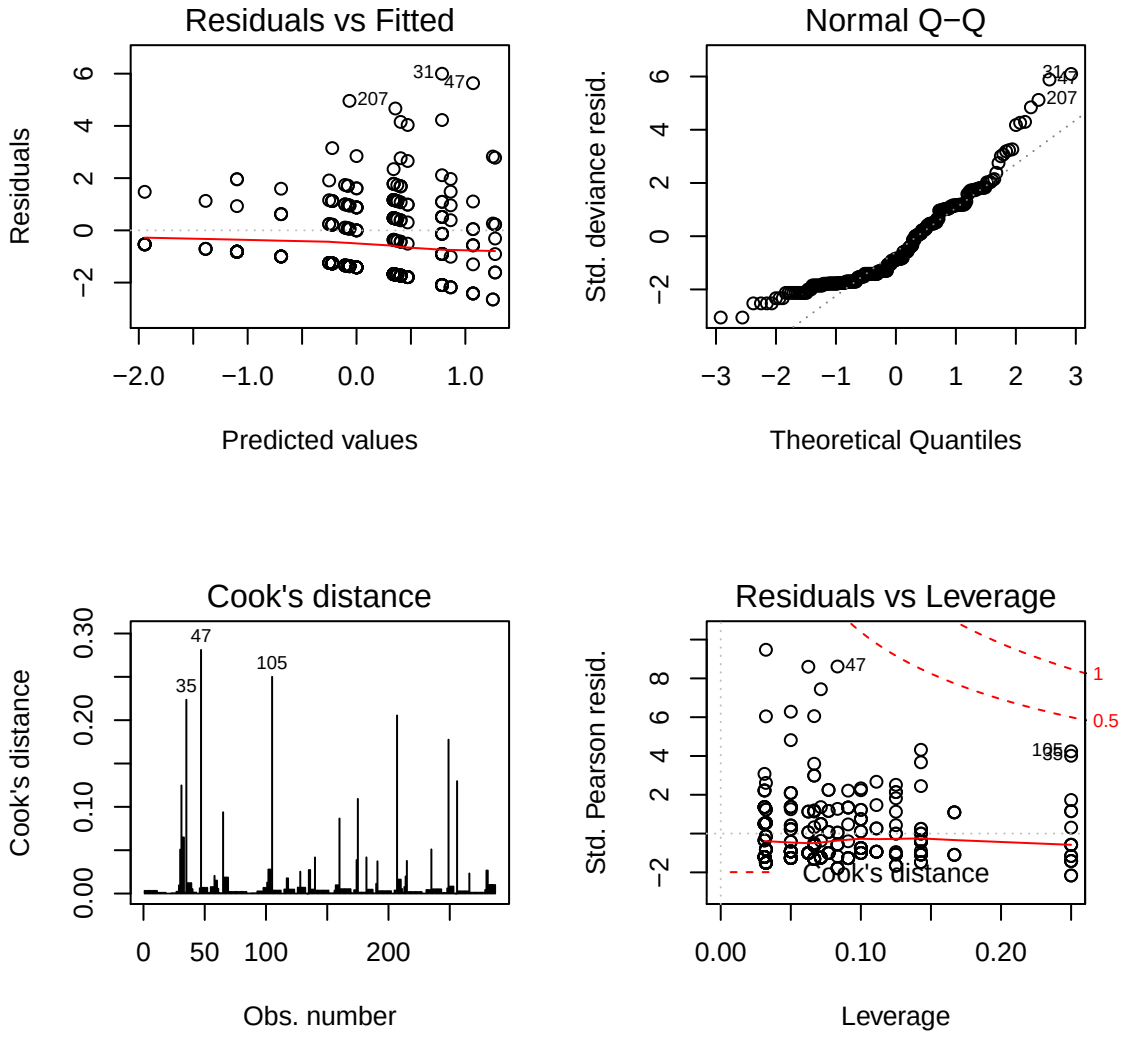
STAT7030 Assignment 2

Question 1

(a)

Firstly we fit the model from the given webpage and generate some basic diagnostic plots as requested:



- The residuals vs fitted plot shows that the data has a clear overdispersion since the spread of data points are not constant.
- The normal q-q plot demonstrates the sample is not a normal distribution, but this should not be a serious concern since we are fitting a generalized linear model here.
- In Cook's distance plot, all of the values of Cook's distances are not large (not exceeding 0.30), and no relatively large value detected as well. So, there is no potential influential points shown in Cook's distance plot.

- Also, no potential high-leverage points observed in Residuals vs Leverage plot since no points go beyond the Cook's distance line.

(b)

```
## 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 3.848e-09 ***
## Location                   1   25.160       284     764.65 5.277e-07 ***
## Age                        2    8.582       282     756.07 0.0136927 *
## Sex                        1    0.635       281     755.43 0.4256263
## Swimmer:Location           1    1.693       280     753.74 0.1932591
## Swimmer:Age                2    6.383       278     747.36 0.0411125 *
## Location:Age               2    3.920       276     743.44 0.1408603
## Swimmer:Sex                1    0.227       275     743.21 0.6335094
## Location:Sex               1   11.120       274     732.09 0.0008540 ***
## Age:Sex                    2    1.783       272     730.31 0.4100133
## Swimmer:Location:Age       2    3.674       270     726.63 0.1593097
## Swimmer:Location:Sex       1    0.239       269     726.39 0.6249411
## Swimmer:Age:Sex            2    0.194       267     726.20 0.9076170
## Location:Age:Sex           2   13.943       265     712.26 0.0009382 ***
## Swimmer:Location:Age:Sex   2    8.538       263     703.72 0.0139971 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the ANOVA table output, we can calculate the dispersion $\phi = 703.72/263 = 2.67574$, while the rule of thumb for over/underdispersion is $(0.7383876, 1.2616124)$. Obviously, we have overdispersion.

```
# rule of thumb for overdispersion
inf.glm$deviance/inf.glm$df.residual
```

```
## [1] 2.67574
```

```
c(1-3*sqrt(2/inf.glm$df.residual),1+3*sqrt(2/inf.glm$df.residual))
```

```
## [1] 0.7383876 1.2616124
```

If we want to conduct a formal test for over/underdispersion,

$$H_0 : \phi = 1 \text{ vs } H_a : \phi \neq 1.$$

```
# formal test for over/underdispersion
inf.glm$deviance/summary(inf.glm)$dispersion # where summary(inf.glm)$dispersion==1
```

```
## [1] 703.7197
```

```r
c(qchisq(0.025, inf.glm$df.residual),qchisq(0.975, inf.glm$df.residual))
```

```
## [1] 219.9720 309.8145
```

Using formal test for over/under-dispersion, as the observed deviance 703.7191 lies outside the interval $(219.9720, 309.8145)$, we would reject the null hypothesis, and conclude that there is evidence of significant over-dispersion.

(c)

```r
est.dispersion <- inf.glm$deviance/inf.glm$df.residual
```

By part (b), we believe there is evidence of significant overdispersion. Therefore, some modifications can be applied so that the dispersion is allowed to be larger than 1. The detailed modifcation is to replace `dispersion` parameter in `anova()` command with the devaince of our model divided by the degrees of freedom of the model residuals.

```
## 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
```

We can have the following finds from the Analysis of Deviance table above:

- On one hand, the table shows very small p-values for `Swimmer` term and `Location` term and both are smaller than 0.05, so we consider that the refined model can contain these two terms. For the interaction terms, the p-value for `Location*Sex` is smaller than 0.05, but the variable `Sex` is not significant in the model (p-value $= 0.6262242 < 0.05$), we do not contain this interaction term in the refined mode. For other variables and interactions, the p-values suggest their effect might be trivial as well. By the principal of parsimony, we could start from a simple model, i.e. with variables `Swimmer` and `Location` only.

To be specific, our refined model is:

$$\log(\text{Infections}) = \beta_0 + \beta_1 \cdot \text{Swimmer} + \beta_2 \cdot \text{Location}$$

```
## [1] 2.675741 1.261612
```

- On the other hand, even after our modification, the problem of overdispersion still exists. The real problem (also for other Poisson regression) is excess zeros. So we can do the Zero-Inflation test using `testZeroInflation()` function in `DHARMa` package.

```
##
##  DHARMa zero-inflation test via comparison to expected zeros with
##  simulation under H0 = fitted model
##
## data:  simulateResiduals(inf.glm, refit = T)
## ratioObsExp = 1.6557, p-value < 2.2e-16
## alternative hypothesis: more
```

The test result indicates that zero inflation does exist for our case, and this could be problematic as we are going to insist on fitting a GLM with poisson model. One possible solution is to fit other models such as negative binomial model or quasi-poisson. But nevertheless, this definitely needs further investigation.

(d)

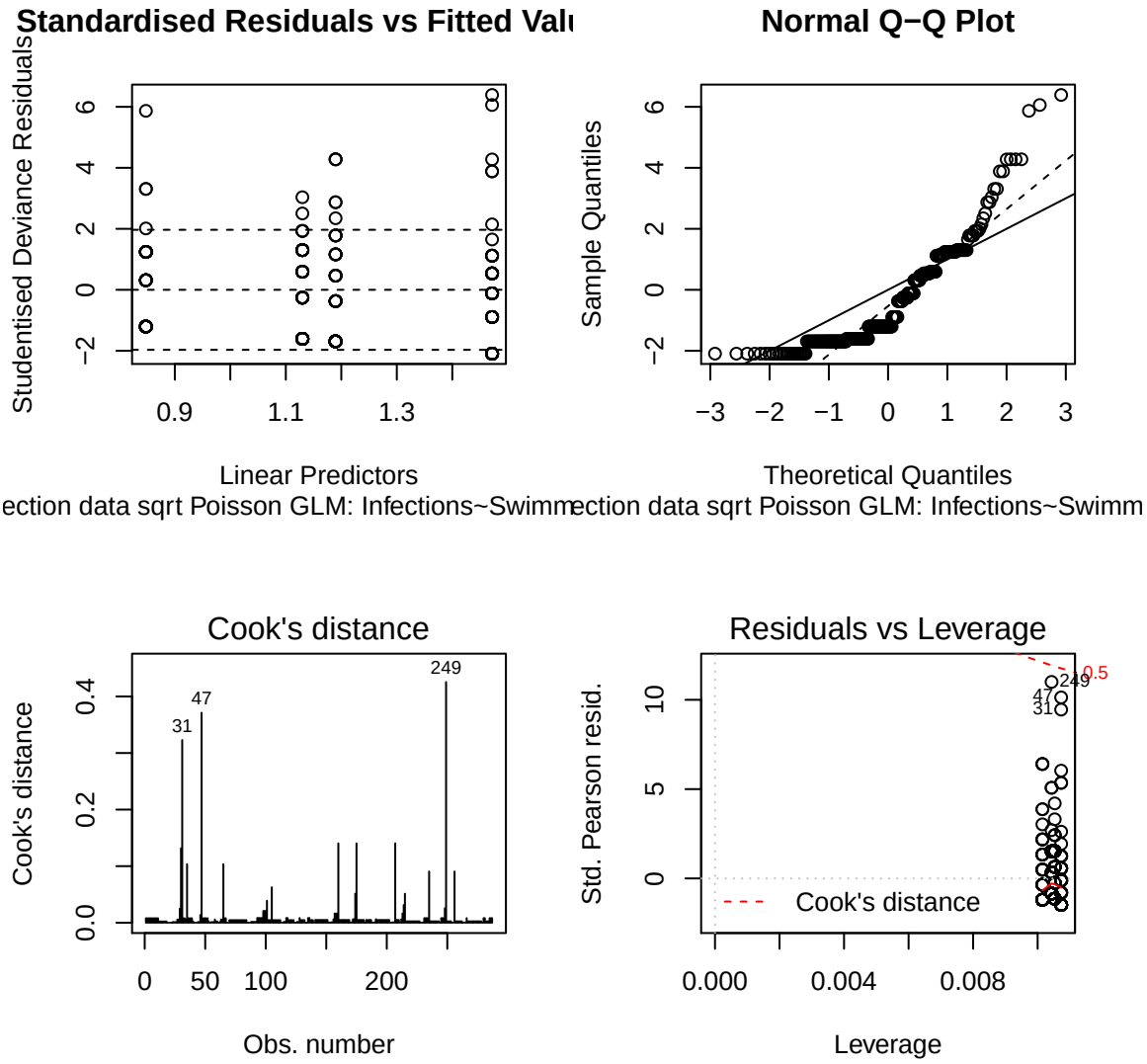
```
## [1] 4
```

After fitting the model with refined model and new link function (`sqrt()`), the fitted variance weights are the same, which are all 4. This makes sense, as mentioned in the *BRICK*,

$$w_i^2 = \frac{1}{V(\mu_i)g'(\mu_i)^2} = \frac{1}{\mu \cdot \left(\frac{1}{\mu \cdot \ln e}\right)^2} = \mu$$

which is constant, so that `weights=` option is not required here.

(e)



- The standardised residuals vs linear predictors actually tells nothing about the situation of overdispersion. But it indeed highlights the suspicion of overdispersion as we are familiar that residuals associated with fitted values less than 2 might cause distortion in plot. In this plot, all of our fitted values are quite small. Recall the Zero-Inflation test we conducted previously, it confirms our concern. We can do some analytic test to check it really holds an overdispersion. In fact, the refined model still has the problem of overdispersion as we can see the dispersion is beyond the critical range $767.0611 > 332.5759$.

```
## [1] 767.0611
```

```
## [1] 239.2108 332.5759
```

- The modified normal q-q plot indicates some deviation from the assumption of normality, but this is totally okay.
- The Cook's distance plot shows that the 31st, 47th, 249th data points have relative large values, and we may consider them as potential influential points.
- The standardised residuals vs leverage plot raises no questions as well.

Recall that we agree upon the existence of overdispersion, so naturally we do the transformation for `dispersion=` in `anova()` to display interpretable p-values. A summary table is also generated.

```
## [1] 2.700919
```

```
## 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
## Swimmer   1   34.699       285     789.81 0.000338 ***
## Location  1   22.748       284     767.06 0.003706 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## glm(formula = Infections ~ Swimmer + Location, family = poisson(link = sqrt),
##     data = dat)
##
## 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    0.08393  10.103  < 2e-16 ***
## SwimmerOccas      0.34179    0.09702   3.523 0.000427 ***
## LocationNonBeach  0.28198    0.09705   2.905 0.003668 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 2.700919)
##
##     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
```

By the R output above, we claim that both variables `Swimmer` and `Location` are significant. So this model seems adequate as a fitting model, but still we cannot rule out the fact that this Poisson GLM model is full of problems like zero-inflation, etc.

(f)

Since we have two variables `Swimmer` and `Location`, and each of them has two unique values, so in all we have 4 combinations. We are going to print out the 95% confidence interval and 95% prediction interval in the following order:

- `Freq` and `Beach`;
- `Freq` and `NonBeach`;
- `Occas` and `Beach`;
- `Occas` and `NonBeach`.

The 95% confidence interval is

```
##       lower     upper
## 1 0.5585896 0.8995174
## 2 1.0586270 1.5150514
## 3 1.1893091 1.6610939
## 4 1.8761632 2.4762502
```

The 95% prediction interval is

```
##   lower     upper
## 1     0  7.945779
## 2     0  9.615124
## 3     0  9.989024
## 4     0 11.851914
```

Note that, for PI, we set the lower bound of original PI to 0 if they are below 0. Since a negative value after taking power of 2 would result in a positive value, thus those small positive values near 0 would be left out of PI, e.g. if $(-0.5, 0.2)$ takes square it becomes $(0.25, 0.04)$ which does not make sense. So we have to compromise a little bit.

According to the CI and PI above:

- The safest situtaion would be a **frequent swimmer swimming near beach**.
- The water quality at **non-beach** (rivers, etc.) places needs to be improved is , basically rivers.
- The salt in sea water kills microorganisms and bacteria, so swimming in sea is generally safer than in river.
- For frequent swimmers, they could gradually gain some immunity against those infections through more exposure under such circumstances. Additionally, frequent swimmers are more professional in the sense of protection, so they probably use earplugs. However, an occasional swimmer could consider that as an extra cost, so they are more likely to catch an ear infection.

Question 2

(a)

```
## integer(0)
```

```
## [1] 1297
```

```
## integer(0)
```

The passenger with mismatched class information is Mr Alfred Nourney. He boarded the Titanic at Cherbourg as a second class passenger but he was dissatisfied with the second class cabin. Therefore, he went to purser and asked to be transferred to first class. He was then assigned cabin D-38 (for about £38 surcharge). Thus it explains the discrepancy.

(b)

Since we are going to fit the data into a binomial response GLM, we pick the binomial `family` and `logit` link function.

For the variable selection, we start from 3 basic variables which are `Age`, `Sex` and `Class`. We beleive three of these quite accurately describes the feature of an individual on Titanic. Furthermore, we select `SibSp`, `ParCh` and `Fare` out of the remaining variables. `SibSp` and `ParCh` reflect an individual's family status on the boat, and `Fare` is a good complementary variable for `Class` although we highly suspects it might be overshadowed by `Class`.

By far, we have a starting model with `Survived` as response (actually as link function) and `Age`, `Sex`, `Class` and no interactions as predictors. Also, we have full model with the same response but `Age`, `Sex`, `Class`, `SibSp`, `ParCh`, `Fare` and their mutual interactions as predictors. Note that the "full" model looks "too full" because we know that higher order interactions probably will not provide any useful information thus being redundant. But we are going to treat it as a "draft" and just leave the model to automatic model selection functions anyway.

We use `step()` function, `AIC` as criterion, to do backward elimination from full model to reduced model. To save time, we also included some "not-so-full" full model to exclude those meaningless interactions.

The following is the model we manually picked out from (automatic) model selection step and we are going to further investigate its Analysis of Deviance table.

Now the candidate model contains `Age`, `Sex`, `Class`, `SibSp`, `Age:Sex` and `Sex:Class`.

```
## 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
## factor(Sex)                1  266.941       888     917.77 < 2.2e-16 ***
## factor(Class)              2  114.069       886     803.70 < 2.2e-16 ***
## SibSp                      1   18.492       885     785.21 1.706e-05 ***
## Age:factor(Sex)            1   16.703       884     768.50 4.371e-05 ***
## factor(Sex):factor(Class)  2   20.332       882     748.17 3.845e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We also notice that the variable `ParCh` and `Fare` have already been erased and p-values for all terms except `Age` are pretty satisfying. Then we decide to reorder the terms so that `Age` can be significant. After several trials, we have finally reached a all-around candidate model:

```
## 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
## factor(Class)              2  103.547       888    1083.11 < 2.2e-16 ***
```

```
## factor(Sex)                1  256.220       887     826.89 < 2.2e-16 ***
## SibSp                      1    7.740       886     819.15   0.00540 **
## Age                        1   33.940       885     785.21 5.684e-09 ***
## factor(Class):factor(Sex)  2   31.779       883     753.43 1.257e-07 ***
## factor(Sex):Age            1    5.256       882     748.17   0.02187 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So now the model is

$$\text{logit}(\text{Survived}) = \beta_0 + \beta_1 \text{Class} + \beta_2 \text{Sex} + \beta_3 \text{SibSp} + \beta_4 \text{Age} + \beta_5 \text{Class} \cdot \text{Sex} + \beta_6 \text{Sex} \cdot \text{Age}.$$

and all variables `SibSp` (from Kaggle), `Class:Sex` (from original variables), `Sex:Age` (from original variables) are significant addition to the model.

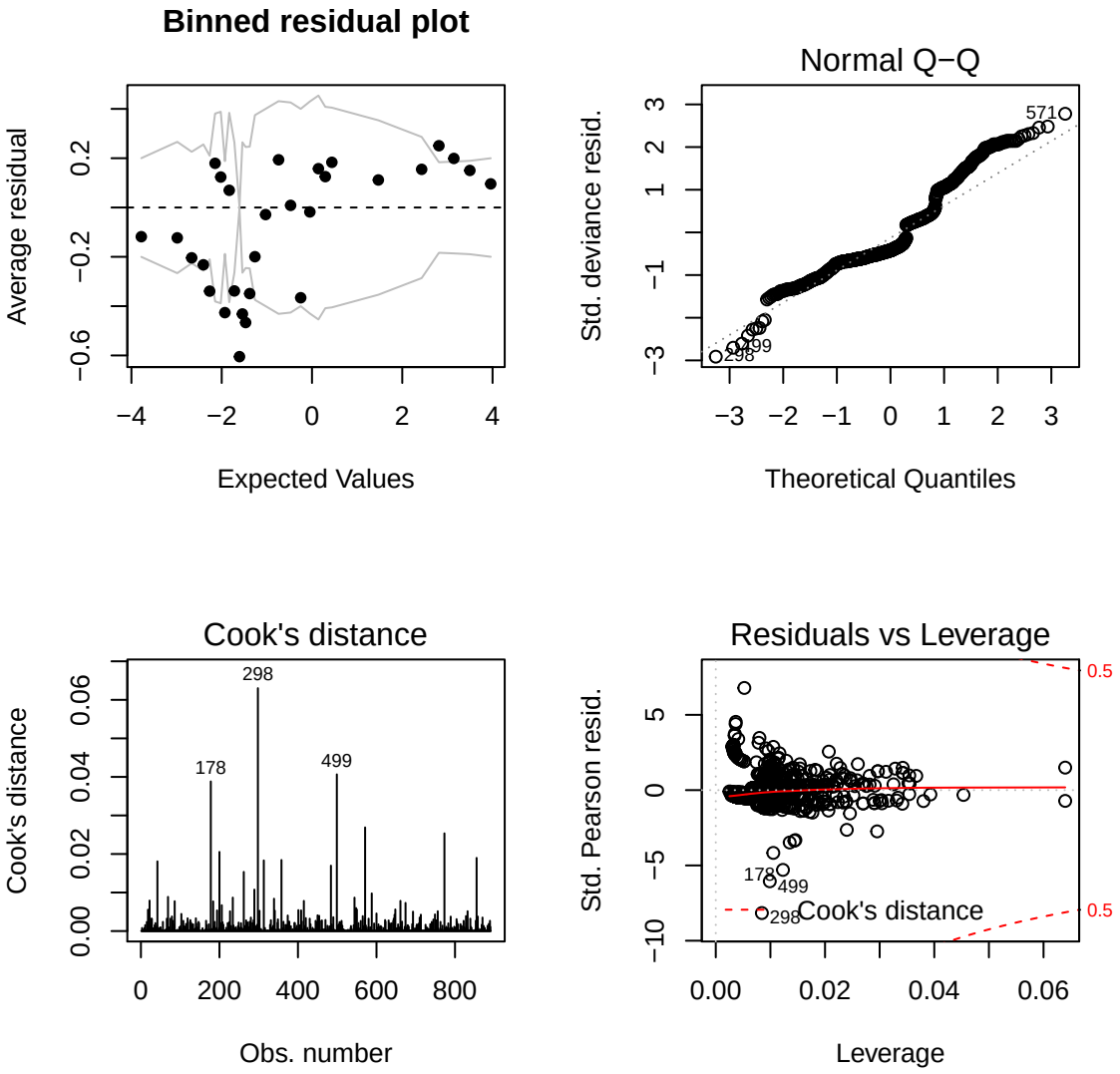
(c)

```
##
## Call:
## glm(formula = Survived ~ factor(Class) + factor(Sex) + SibSp +
##     Age + factor(Sex):factor(Class) + Age:factor(Sex), family = binomial(link = logit),
##     data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9002  -0.6400  -0.4210   0.3775   2.7733
##
## Coefficients:
##                                      Estimate Std. Error z value
## (Intercept)                           4.62482    0.79422   5.823
## factor(Class)2nd Class               -1.15511    0.73317  -1.575
## factor(Class)3rd Class               -3.72210    0.65017  -5.725
## factor(Sex)male                      -2.49830    0.89845  -2.781
## SibSp                                -0.38210    0.10415  -3.669
## Age                                  -0.02605    0.01300  -2.004
## factor(Class)2nd Class:factor(Sex)male -0.82278    0.82495  -0.997
## factor(Class)3rd Class:factor(Sex)male  1.45951    0.71845   2.031
## factor(Sex)male:Age                  -0.03802    0.01652  -2.302
##                                      Pr(>|z|)
## (Intercept)                          5.78e-09 ***
## factor(Class)2nd Class               0.115141
## factor(Class)3rd Class               1.04e-08 ***
## factor(Sex)male                      0.005425 **
## SibSp                                0.000244 ***
## Age                                  0.045047 *
## factor(Class)2nd Class:factor(Sex)male 0.318581
## factor(Class)3rd Class:factor(Sex)male 0.042208 *
## factor(Sex)male:Age                  0.021343 *
## ---
## 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:  748.17  on 882  degrees of freedom
## AIC: 766.17
##
## Number of Fisher Scoring iterations: 6
```

Now we take a look at the summary table of our candidate model, and p-values of `2nd Class` and `2nd Class:male` are greater than 0.05. Just keep in mind that there might be not such a huge difference between 1st and 2nd Class if holding other variables constant. Similarly, the interaction of 2nd Class and Male is not significant either. But generally the model here is legit, as we do emphasize the significant difference in survival status of the 3rd Class.

Then a series of diagnostic plots are produced below.



- In binned residual plot, we notice that some data points lie outside of the 95% confidence interval, but generally they are not far away from the shaded boundaries which can be considered as 2.5% and 97.5% quantiles. A further investigation is needed to determine if they are outliers. Also, a slight trend as a whole can be concluded from the plot that negative average residuals at the mid-lower end, and positive average residuals at the higher end.
- The normal q-q plot suggests no problem.
- The Cook's distance plot indicates the 298th data point seems to have a relatively large value in Cook's

distance, it is a potential influential data point.
- In residuals vs leverage plot, seems that 298th data does not seem to be high leverage point again. And nothing particular worth paying attention to.

In conclusion, our candidate model has no big problems and we should stick with it.

(d)

The Analysis of Deviance table was generated in part (b) already.

There is no overdispersion for ungrouped data. The overdispersion is not possible if total number of groups is 1. If the response only takes valeu 0 and 1 (which is, in this case), then it must be distributed as Bernoulli(p_i) and its variance is $p_i(1-p_i)$. Therefore, we should assume the `scale=1`, and not discuss overdispersion here at all.

(e)

The fitted value of our model should range from 0 to 1 and means the probability of having survived which indicates the larger the fitted value is the more likely the passenger will have survived.

By comparing the numbers of predicted survivors and non-survivors with the observed true numbers of survival status, we shall have the *sensitivity, specificity* and *accuracy* of our candidate model with training data.

```
##                 train
## sensitivity 0.6842105
## specificity 0.9016393
## accuracy    0.8181818
```

(f)

The three prediction measurements are included in the table below. Also, the values of those from training data are also included for a direct comparison.

```
##                 train      test
## sensitivity 0.6842105 0.6518987
## specificity 0.9016393 0.8576923
## accuracy    0.8181818 0.7799043
```

All the *sensitivity*, *specificity* and *accuracy* in test data decrease a little bit in training data which is somehow inevitable. This is probably by chance, if we are really "unlucky" or "lucky" enough, the testing accuracy should fluctuate over the accuracy of training data. Meanwhile the decrease is not large and we believe the model is generally consistent.

Our model seems to be good at true negative rate, bad at true positive rate. This can be explained that it probably found the pattern (or deterministic cause) that a passenger did not survive, but failed on the other hand. Maybe the key features to determine the survivor is way too complicated.

Addtionally, if we really want to see the "prediction" power of our GLM in training data, we can do a handy 10-fold cross validation.

```
library(boot)
cost_classification <- function(r, pi) mean(abs(r-pi) > 0.5)
cv.res <- cv.glm(data=train,
                 glmfit=glm(Survived~factor(Class)+factor(Sex)+SibSp+Age+
                                factor(Sex):factor(Class)+Age:factor(Sex),
```

```
                            family=binomial(link=logit),data=titanic),
                  K=10,cost=cost_classification)
cv.res$delta[1]
```

```
## [1] 0.1907969
```

So the overall accuracy should be around 81%.

About the competition: we are not confident enough to win, because the turned-in accuracy of other competitors on Kaggle are extremely high. Although we could not rule out the possibility that those algorithms (mostly machine learning approach) in fact overfit the data. Frankly speaking, a well-tuned machine learning algorithm still can beat our naïve GLM. This is probably due to the limitation of GLM, as a statistical approach, whose main concern is to infer parameters. But for machine learning approach, it focuses on prediction as an ultimate goal.

Finally, we would like to elaborate on our candidate model selection. Without doubt, model selection is important, but the difference bewteen a "very good model" and a "pretty good model" could be trivial. Our model did well as a GLM, but some other models could have very similar performace. But again, the reason why we select ours, it's because we can interpret it with no trouble. The variables we keep in the end, low-order interactions included, are not hard to understand. Admittedly, one model with twenty or more variables might improve the accuracy by one or two percentages, but it **does not make sense** when talking about what these variables mean.

And we would like to end this discussion with a quote from Megan Risdal, whose Titanic prediction project has the highest voting score on Kaggle:

> When I submit the predicted survival data from various models that built in the course to Kaggle competition, i have got approximately the same score. Now I realize that why data scientist used to spend most of their time into feature engineering and exploratory analysis compare to actual model building. Model that we are using is definitely important, however more than that understanding our data and feature engineering is crucial.

By the way, her model prediction accuracy for testing data is also around 0.7727273.

References

- Wikipedia page of Zero-Inflation, https://en.wikipedia.org/wiki/Poisson_regression#Overdispersion_and_zero_inflation.
- Mr Alfred Nourney, Encyclopedia Titanic, https://www.encyclopedia-titanica.org/titanic-survivor/alfred-nourney.html.
- Exploring Survival on the Titanic by Megan Risdal, https://www.kaggle.com/mrisdal/exploring-survival-on-the-titanic/notebook.