RMS Titanic Data Analysis

Model Fit and Prediction Accuracy in the Regression Family

Group Member:

Jennifer (Xin Bei) She (30618136)
Wendy Wang (30938138)
Chihiro Hanawa (53417135)
Aaron Yen (65332124)

Linear Regression: Statistics 306
April 7th, 2016

Section 1: Abstract

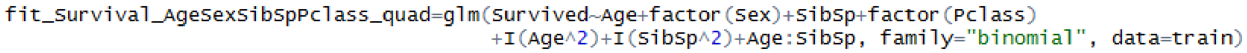
The sinking of the RMS Titanic was one of the largest passenger liners in service. When it sank on April 14, 1912, it resulted in the deaths of more than 1500 out of 2224 passengers and crew[1]. One of the reasons that the shipwreck led to such a great loss of many people is because of a lack of enough lifeboats for passengers and the low temperature of the weather. Although survival involved some luck, some groups are more likely to survive than others. However, there are not a lot of online sources, except a nearly complete dataset found on Kaggle.com, that could provide us with more information of this tragedy. We thought it would be interesting to create suitable regression models to further analyze this historical event.

The goal of our project is to form a prediction equation for passenger survival on the titanic.

We decided to transform numeric explanatory variables that have heavily skewed tails over large ranges. We also performed a preliminary logistic regression with all the explanatory variables to decide on the baseline categories for the categorical variables. The result is that the Fare variable is transformed to log(1+Fare), and all the other numerical variables stayed the same. All the baseline categories for the categorical data stayed the same.

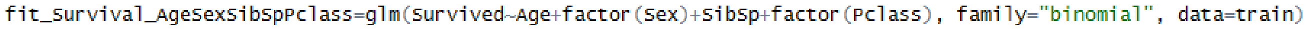
We looked at AIC, in-sample & out-of-sample misclassification rates and calibration of fit and made comparison with models that have quadratic terms. Using AIC and out-of-sample misclassification rates as criteria, the three models that we considered the best are:

AgeSexSibSpPclass_quad model:



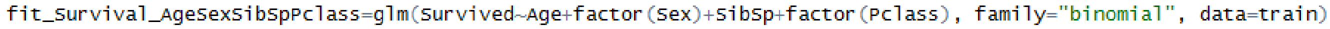
```
fit_Survival_AgeSexSibSpPclass_quad=glm(Survived~Age+factor(Sex)+SibSp+factor(Pclass)
                                        +I(Age^2)+I(SibSp^2)+Age:SibSp, family="binomial", data=train)
```

AgeSexSibSpPclass model:



```
fit_Survival_AgeSexSibSpPclass=glm(Survived~Age+factor(Sex)+SibSp+factor(Pclass), family="binomial", data=train)
```

AgeSexParchSibSpPclass model:



```
fit_Survival_AgeSexSibSpPclass=glm(Survived~Age+factor(Sex)+SibSp+factor(Pclass), family="binomial", data=train)
```

[1] We got this information from this source: https://www.kaggle.com/c/titanic

Section 2: Analysis

2.1 Summary and Key Features of Data

The response variable used is Survival (0 = no, 1 = yes.)

The explanatory variables used are as follows[2]:

Explanatory Variables	Explanation and unit (if applicable)
Pclass	Passenger Class. A proxy for social-economic status (1 = 1st Upper, 2 = 2nd Middle, 3 = 3rd Lower)
Sex	Gender of passenger (Male = 0 , Female = 1)
Age	Age of passengers in years (fractional if age < 1)
SibSp	Number of siblings/spouses aboard; siblings includes brothers, sisters, stepbrothers and stepsisters of passenger; spouse includes husband or wife of passenger (mistresses and fiancees are ignored)
Parch	Number of parents/children abroad; parents includes mother and father of passenger; children includes sons, daughters, stepsons, and stepdaughters of passenger (other relationships such as friends, neighbours, nannies ignored)
Fare	Passenger fare (in US dollars)

In particular for categorical data, the frequency tables are listed below.

Survival Categories	Frequency (unit: the number of passengers)
Survived	290
Death	424

Table 1: Frequency table for the response variable "Survival". For "Survived", it is expressed as 1 in the dataset; for "Death", it is expressed as 0 in the dataset.

Pclass Categories	Frequency (unit: the number of passengers)
First Class	186
Second Class	173

[2] We got this information from this source: https://www.kaggle.com/c/titanic

Third Class	355

Table 2: Frequency table for the explanatory variable "Pclass". For "First Class", it is expressed as 1 in the dataset; for "second class", it is expressed as 2 in the dataset; for "third class", it is expressed as 3 in the dataset.

Gender Categories	Frequency (unit: the number of passengers)
Female	261
Male	453

Table 3: Frequency table for the explanatory variable "Sex". For "female", it is expressed as 1 in the dataset,; for "male", it is expressed as 0 in the dataset.

From the basic summary of the data above, it appears that a lot more passengers did not survive than who did survive the titanic accident. The number of third class passengers is about twice as many as that of first and second class passengers. There are significantly more male passengers than female passengers who survived.

For numerical data, the frequency tables are listed below.

Quantiles (Unit: the number of passengers)	Age	# of Sibling/Spouses	Parch	Fare
Min	0.42	0	0	0
1st Qu.	20.12	0	0	8.05
Median	28	0	0	15.74
Mean	29.7	0.5126	0.4314	34.69
3rd Qu.	38	1	1	33.38
Max.	80	5	6	512.3

Table 4: Frequency table for all other continuous explanatory variables: Age, SibSp, Parch, and Fare

We can see that the distributions of the number of siblings/spouses on board, number of parents/children on board and fare price are right skewed, and the distribution of age is approximately normal. In addition, the range of Fare is very large.

2.2 Data Selection

2.2.1 Continuous/Non-Categorical Data

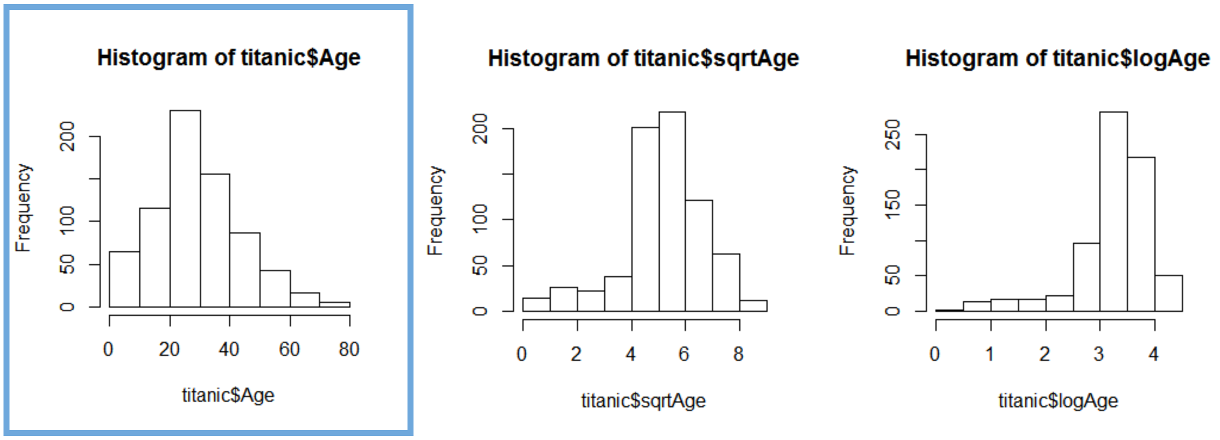


Fig. 1: Histograms for the explanatory variable for Age, square root of "Age", and log of "Age". Blue Highlighted histogram is the selected data.

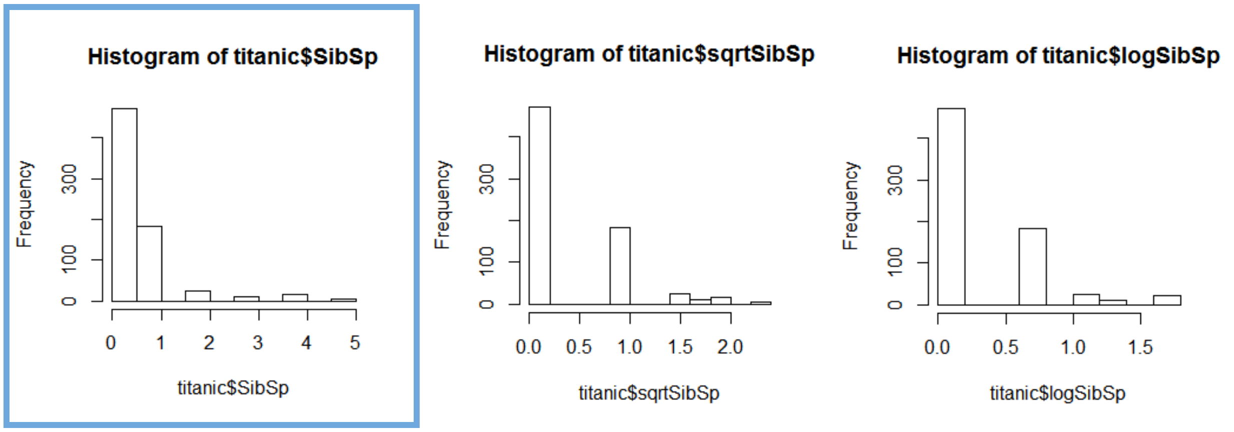


Fig. 2: Histograms for the explanatory variable for "Sibsp", square root of "SibSp", and log of "SibSp". Blue Highlighted histogram is the selected data.

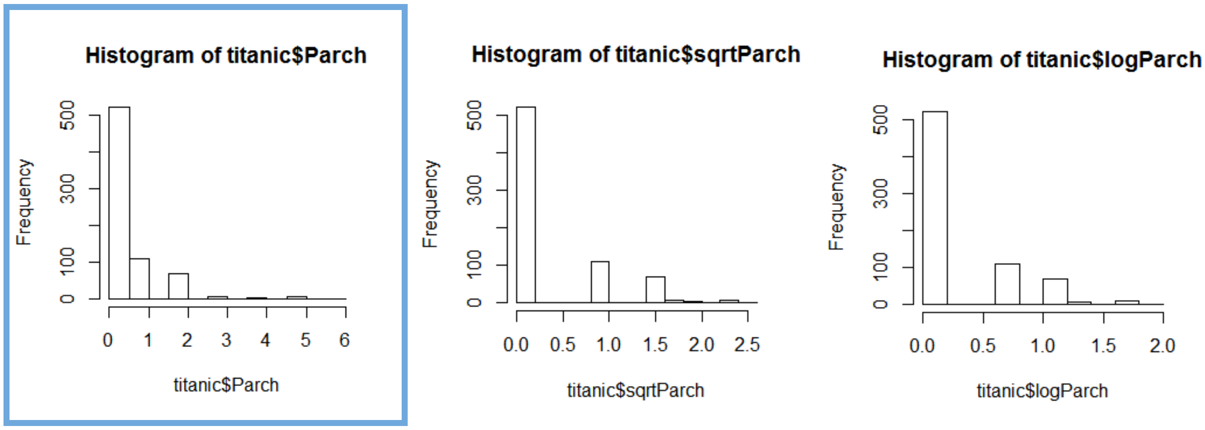


Fig. 3: Histograms for the explanatory variable for "Parch", square root of "Parch", and log of "Parch". Blue Highlighted histogram is the selected data.

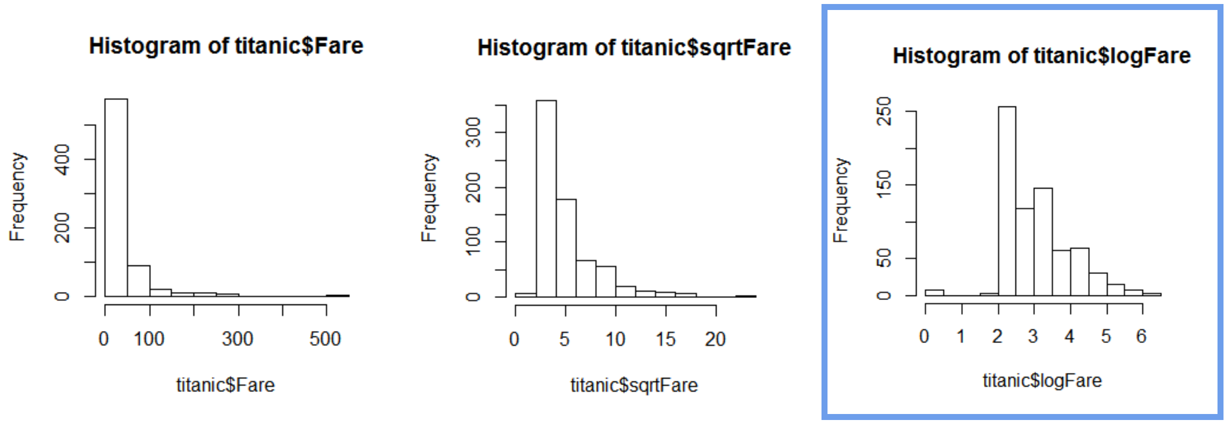


Fig. 4: Histograms for the explanatory variable for “Fare”, square root of “Fare”, and log of “Fare”. Blue Highlighted histogram is the selected data.

For “SibSp” and “Parch”(*See Fig. 2, Fig. 3*), all histograms show right skewness, but the range is small enough, and transformations did not alter the data enough for a significant result. We decided to use the non-transformed data for the model for “SibSp” and “Parch” explanatory variables. “Fare” *(See. Fig. 4)* also shows right skewness in all the plots, but “logFare” looks the most symmetric, so the log transformed data for “Fare”, “logFare” is used. However, for “Age” (*See Fig. 1*), the non-transformed data is the most symmetric and was used for the model.

2.2.2 Categorical Variables

For “Pclass” variable, category 1 (out of 1, 2, 3) is used as baseline since it has the largest beta. For “Sex” variable, we choose category 0 (out of 0, 1) for baseline.

2.3 Relationships Among Variables

2.3.1 Boxplots for Continuous Explanatory Variables vs. Survival

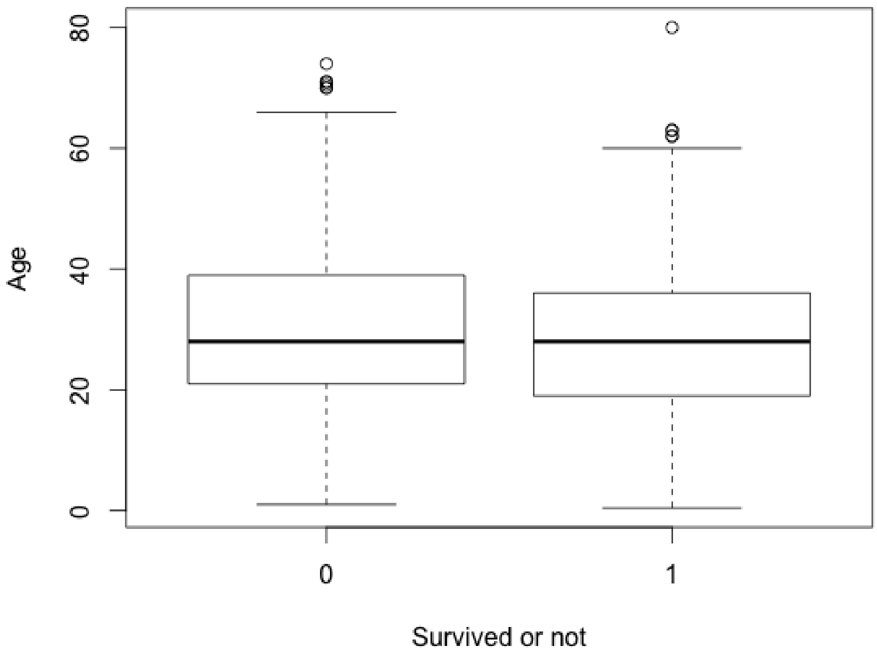


Fig 5: Boxplot between Survival and Age

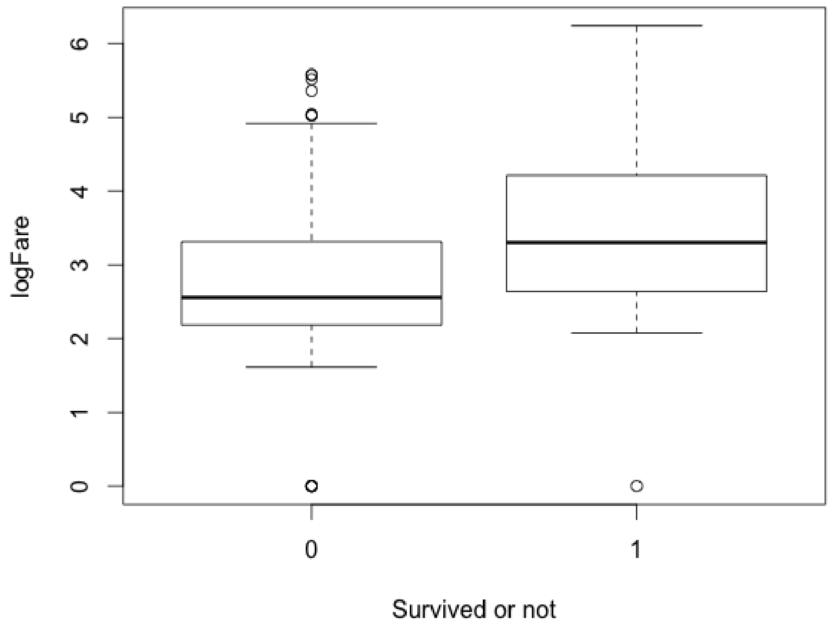


Fig 6: Boxplot between Survival and Log Fare

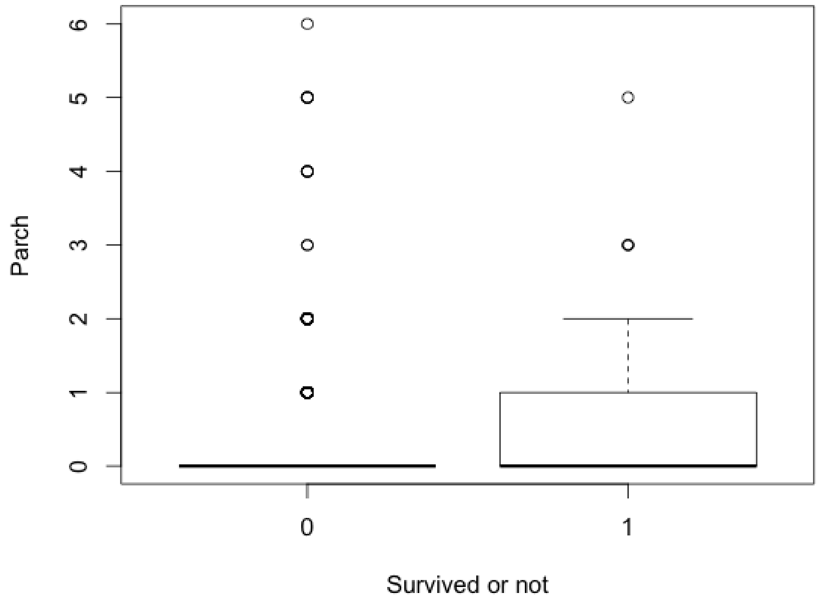


Fig 7: Boxplot between Survival and Parch

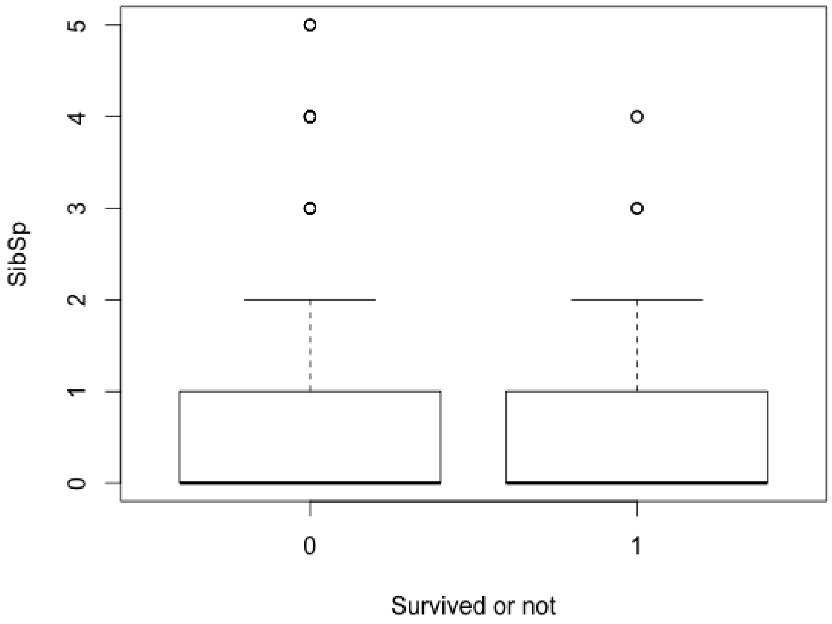


Fig 8: Boxplot between Survival and SibSp

We used the selected explanatory data to make boxplots with “Survival” response variable. In Fig 5, we can see that the average age and age range from the passengers who survived and dead are very similar. In Fig 6, it appears that the average fare is higher in survived than in death. It tells us that the more expensive the fare is, the more likely to survive. In Fig 7 and Fig 8, the mean of “Parch” and “SibSp” are around the same for both survived and dead.

2.3.2 Contingency Tables for Discrete Explanatory Variables vs. Survival

Frequency (unit: the number of passengers)	First Class	Second Class	Third Class
Dead	64	90	270
Survived	122	83	85

Table 5: Pclass vs Survival. In dataset, First Class, Second Class and Third Class are labeled as 1, 2 and 3; Dead and Survive are labeled as 0 and 1

Frequency (unit: the number of passengers)	Male	female
Dead	360	64
Survived	93	197

Table 6: Gender vs Survival. In dataset, Male and Female are labeled as 0 and 1; Dead and Survive are labeled as 0 and 1

We cannot see a huge correlation between age, the number of siblings/spouses aboard and the number of parents/children aboard, in relation to survival. Those who survived tend to have paid higher fare. Passengers in lower Pclass had higher survival rate. Females also had higher survival rate.

2.4 Model Fitting

We separated the data into two halfs (n=357 each): training data and holdout data using randomization.

We can get the correlation of explanatory variables based on the training data.

	Survived	Pclass	Sex	Age	SibSp	Parch	logFare
Survived	1.0000	-0.3462	0.5658	-0.0653	0.01189	0.07715	0.3855
Pclass	-0.3462	1.0000	-0.2003	-0.3840	0.0625	0.0681	-0.6738
Sex	0.5658	-0.2003	1.0000	-0.0685	0.1470	0.2716	0.3790
Age	-0.0653	-0.3840	-0.0685	1.0000	-0.2910	-0.0959	0.1656
SibSp	0.0119	0.0625	0.1470	-0.2910	1.0000	0.2742	0.2845
Parch	0.0771	0.0681	0.2716	-0.0959	0.2742	1.0000	0.2679
logFare	0.3855	-0.6738	0.3790	0.1656	0.2845	0.2679	1.0000

Table 7: Correlation of response variable and explanatory variables

“Sex” is the most correlated explanatory variable for survival. “logFare” is also correlated with “Survival”, because the passengers with expensive fares are prioritized to be rescued. “Pclass” is negatively correlated with “Survival” because first class people are more likely to survive than people in third class. “Parch”, “Age” and “SibSp” have small correlation with survival. “Parch” and “SibSp” are positively correlated with survival, but the correlation is very low. “Age” is negatively correlated, meaning younger passengers are more likely to survive than older passengers.

2.4.1 Exhaustive Comparison of Models (Deviance and AIC)

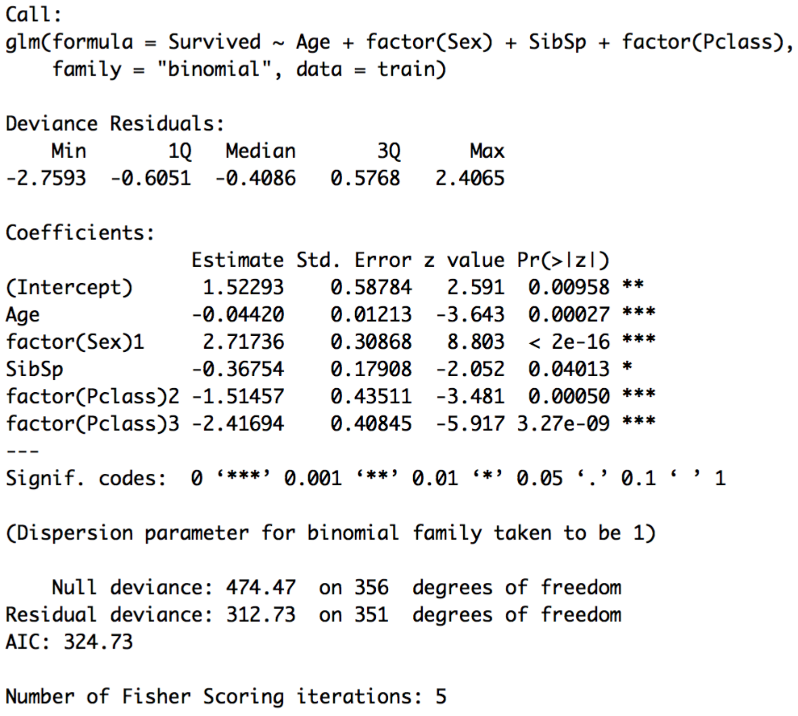
```
      subsetvec                        deviancevec       aicvec
 [1,] "Age"                           "472.9420"        "476.9420"
 [2,] "Sex"                           "357.47822"       "361.4782"
 [3,] "SibSp"                         "474.4230"        "478.423"
 [4,] "logFare"                       "417.9568"        "421.9568"
 [5,] "Pclass"                        "431.0220"        "437.0220"
 [6,] "Parch"                         "472.3989"        "476.3989"
 [7,] "Age_Sex"                        "357.1065"        "363.1065"
 [8,] "Age_SibSp"                      "472.9205"        "478.9205"
 [9,] "Age_Pclass"                     "410.0132"        "418.0132"
[10,] "Age_Parch"                      "471.1743"        "477.1743"
[11,] "Age_logFare"                    "410.0931"        "416.0931"
[12,] "Sex_logFare"                    "354.3691"        "360.3691"
[13,] "Sex_Parch"                      "354.3691"        "360.3691"
[14,] "Sex_Pclass"                     "328.6201"         "336.6201"
[15,] "Sex_SibSp"                      "354.6964"         "360.6964"
[16,] "Pclass_logFare"                  "414.7041"         "422.7041"
[17,] "Pclass_SibSp"                    "430.5980"        "438.5980"
[18,] "Pclass_Parch"                    "426.9023"         "434.9023"
[19,] "SibSp_logFare"                    "413.5539"         "419.5539"
[20,] "SibSp_Parch"                      "472.3684"         "478.3684"
[21,] "logFare_Parch"                    "417.6389"         "423.6389"
[22,] "Age_Sex_Pclass"                   "317.2775"          "327.2775"
[23,] "Age_Sex_logFare"                   "336.5957"         "344.5957"
[24,] "Age_Sex_Parch"                    "353.7798"          "361.7798"
[25,] "Age_Sex_SibSp"                     "353.4454"         "361.4454"
[26,] "Sex_SibSp_logFare"                   "330.9288"       "338.9288"
[27,] "Sex_SibSp_Parch"                    "352.6430"         "360.6430"
[28,] "Sex_SibSp_Pclass"                    "327.3128"        "337.3128"
[29,] "Parch_logFare_SibSp"                "413.5306"          "421.5306"
[30,] "Parch_logFare_Pclass"                "414.6420"          "424.6420"
[31,] "logFare_SibSp_Pclass"                  "412.6413"          "422.6413"
[32,] "Age_SibSp_Pclass"                       "409.3893"         "419.3893"
[33,] "Age_SibSp_logFare"                      "397.4695"         "405.4695"
[34,] "Age_SibSp_Parch"                        "470.9155"         "478.9155"
[35,] "Age_Pclass_logFare"                     "398.8925"         "408.8925"
[36,] "Age_Pclass_Parch"                       "406.9044"         "416.9044"
[37,] "Age_logFare_Parch"                       "408.9645"        "416.9645"
[38,] "Sex_Pclass_logFare"                      "327.7993"        "337.7993"
[39,] "Sex_Pclass_Parch"                        "327.4899"        "337.4900"
[40,] "Sex_logFare_Parch"                      "331.9108"          "339.9108"
[41,] "Parch_Pclass_SibSp"                     "426.8835"          "436.8835"
[42,] "Age_Sex_SibSp_Parch"                    "351.2989"          "361.2989"
[43,] "Age_Sex_SibSp_logFare"                   "322.2067"        "332.2067"
[44,] "Age_Sex_SibSp_Pclass"                   "312.7303"         "324.7303"
[45,] "Sex_logFare_Pclass_SibSp"                "324.9926"        "336.9926"
[46,] "Sex_logFare_Pclass_Parch"                "325.3908"        "337.3908"
[47,] "logFare_Pclass_SibSp_Parch"               "412.5735"        "424.5735"
[48,] "Age_Sex_Pclass_logFare"                  "316.9044"          "328.9044"
[49,] "Age_Sex_Pclass_Parch"                    "316.0878"          "328.08776"
```

```
[50,] "Age_Sex_logFare_Parch"         "327.9291"       "337.9291"
[51,] "Age_SibSp_Pclass_logFare"      "392.0925"       "404.0925"
[52,] "Age_SibSp_Pclass_Parch"        "405.1565"       "417.1565"
[53,] "Age_SibSp_logFare_Parch"       "397.1500"       "407.1500"
[54,] "Age_Pclass_logFare_Parch"      "398.7697"       "410.7697"
[55,] "Sex_logFare_Sibsp_Parch"       "325.5524"       "335.5524"
[56,] "Sex_Pclass_Sibsp_Parch"        "326.5720"       "338.5720"
[57,] "Age_Sex_Parch_SibSp_Pclass"    "312.2887"       "326.2887"
[58,] "Age_Sex_Parch_SibSp_logFare"   "315.4805"       "327.4805"
[59,] "Age_Sex_SibSp_Pclass_logFare" "309.9711"      "323.9711"
[60,] "Age_Sex_Parch_Pclass_logFare"  "314.8208"       "328.8208"
[61,] "Age_Parch_SibSp_Pclass_logFare""391.8545"     "405.8545"
[62,] "Sex_Parch_SibSp_Pclass_logFare""322.6237"     "336.6237"
[63,] "Age_Sex_Parch_SibSp_Pclass_logFare" "308.1252"  "324.1252"
```

The highlighted models are the best models based on AIC are AgeSexSibSpPclasslogFare (323.97), AgeSexParchSibSpPclasslogFare (324.13), AgeSexSibSpPclass (324.73) and AgeSexParchSibSpPclass (326.29). Deviance says that AgeSexParchSibSpPclasslogFare (309.97) fits better than AgeSexParchSibSpPclass (312.29).

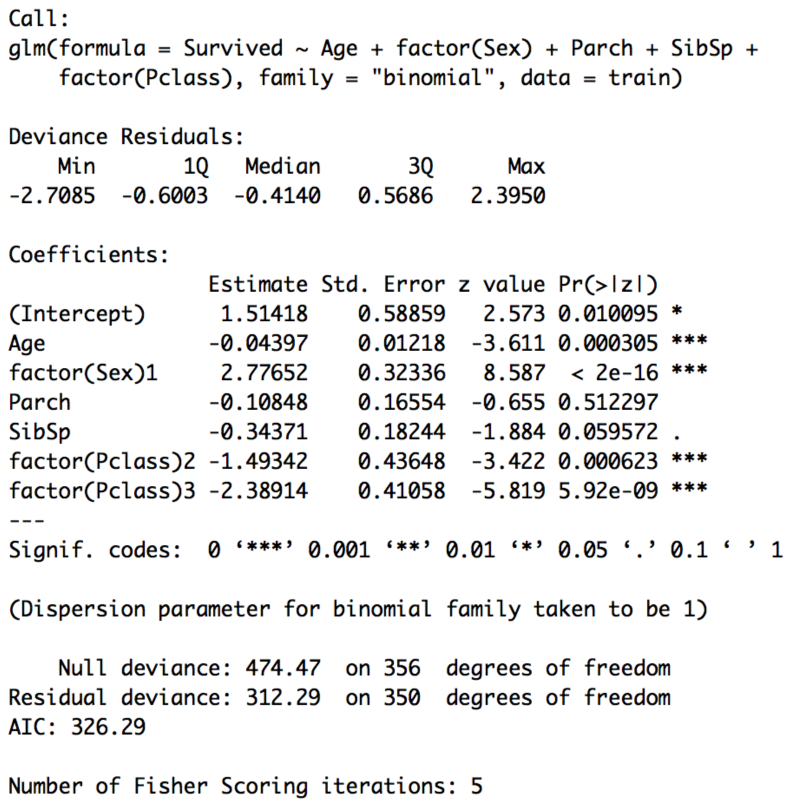
2.4.2 Summary Statistics for Best Models

[44,] "Age_Sex_SibSp_Pclass" model



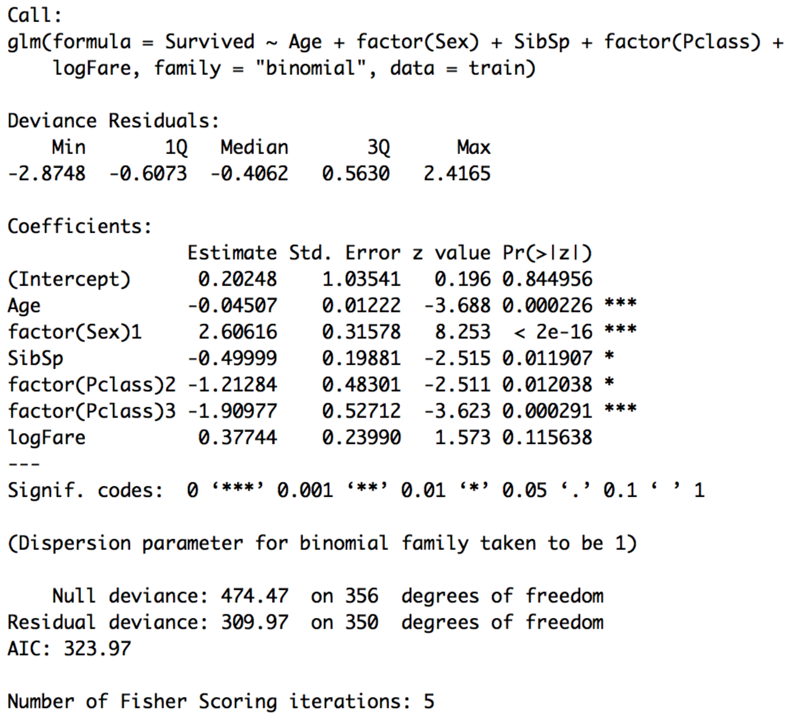
```
Call:
glm(formula = Survived ~ Age + factor(Sex) + SibSp + factor(Pclass), 
    family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7593  -0.6051  -0.4086   0.5768   2.4065  

Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)      1.52293    0.58784   2.591  0.00958 ** 
Age             -0.04420    0.01213  -3.643  0.00027 ***
factor(Sex)1     2.71736    0.30868   8.803  < 2e-16 ***
SibSp           -0.36754    0.17908  -2.052  0.04013 *  
factor(Pclass)2 -1.51457    0.43511  -3.481  0.00050 ***
factor(Pclass)3 -2.41694    0.40845  -5.917 3.27e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 474.47  on 356  degrees of freedom
Residual deviance: 312.73  on 351  degrees of freedom
AIC: 324.73

Number of Fisher Scoring iterations: 5
```

[57,] "Age_Sex_Parch_SibSp_Pclass" model



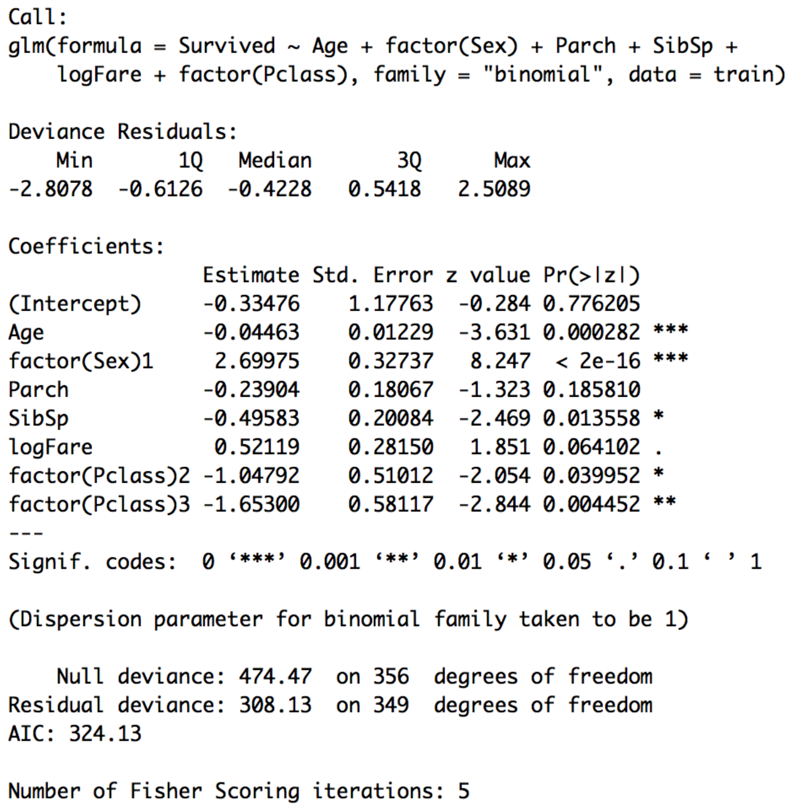
```
Call:
glm(formula = Survived ~ Age + factor(Sex) + Parch + SibSp + 
    factor(Pclass), family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.7085  -0.6003  -0.4140   0.5686   2.3950  

Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)      1.51418    0.58859   2.573 0.010095 *  
Age             -0.04397    0.01218  -3.611 0.000305 ***
factor(Sex)1     2.77652    0.32336   8.587  < 2e-16 ***
Parch           -0.10848    0.16554  -0.655 0.512297    
SibSp           -0.34371    0.18244  -1.884 0.059572 .  
factor(Pclass)2 -1.49342    0.43648  -3.422 0.000623 ***
factor(Pclass)3 -2.38914    0.41058  -5.819 5.92e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 474.47  on 356  degrees of freedom
Residual deviance: 312.29  on 350  degrees of freedom
AIC: 326.29

Number of Fisher Scoring iterations: 5
```

[59] "Age_Sex_SibSp_Pclass_logFare" model



```
Call:
glm(formula = Survived ~ Age + factor(Sex) + SibSp + factor(Pclass) + 
    logFare, family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.8748  -0.6073  -0.4062   0.5630   2.4165  

Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)      0.20248    1.03541   0.196 0.844956    
Age             -0.04507    0.01222  -3.688 0.000226 ***
factor(Sex)1     2.60616    0.31578   8.253  < 2e-16 ***
SibSp           -0.49999    0.19881  -2.515 0.011907 *  
factor(Pclass)2 -1.21284    0.48301  -2.511 0.012038 *  
factor(Pclass)3 -1.90977    0.52712  -3.623 0.000291 ***
logFare          0.37744    0.23990   1.573 0.115638    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 474.47  on 356  degrees of freedom
Residual deviance: 309.97  on 350  degrees of freedom
AIC: 323.97

Number of Fisher Scoring iterations: 5
```

[63] "Age_Sex_Parch_SibSp_Pclass_logFare" model



```
Call:
glm(formula = Survived ~ Age + factor(Sex) + Parch + SibSp + 
    logFare + factor(Pclass), family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.8078  -0.6126  -0.4228   0.5418   2.5089  

Coefficients:
                Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -0.33476    1.17763  -0.284 0.776205    
Age             -0.04463    0.01229  -3.631 0.000282 ***
factor(Sex)1     2.69975    0.32737   8.247  < 2e-16 ***
Parch           -0.23904    0.18067  -1.323 0.185810    
SibSp           -0.49583    0.20084  -2.469 0.013558 *  
logFare          0.52119    0.28150   1.851 0.064102 .  
factor(Pclass)2 -1.04792    0.51012  -2.054 0.039952 *  
factor(Pclass)3 -1.65300    0.58117  -2.844 0.004452 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 474.47  on 356  degrees of freedom
Residual deviance: 308.13  on 349  degrees of freedom
AIC: 324.13

Number of Fisher Scoring iterations: 5
```

2.5 Misclassification

We chose cut-offs 0.3, 0.4, and 0.5 because they were close to the rate of survival 0.3809. We divided the total misclassification tables with in-sample misclassification and out-of-sample misclassification. **In table 7 and 8, the numbers mean the misclassification for both misclassified to be survived and dead for each best models.**

In-Sample mis-class rate:

	Total Misclass Rate with Cut-Off 0.5	Total Misclass Rate with Cut-Off 0.4	Total Misclass Rate with Cut-Off 0.3
AgeSexSibSpPclass	0.2045	0.2045	0.2213
AgeSexParchSibSpPclass	0.1989	0.2073	0.2185
AgeSexSibSpPclasslogFare	0.1933	0.1989	0.2129
AgeSexParchSibSpP	0.1877	0.2073	0.2045

classlogFare			

Table 8: In-Sample misclassification for the best 4 models with cut off 0.3, 0.4, and 0.5.
Note:
AgeSexSibSpPclass model means the model using explanatory variables of age, sex, sibsp and pclass
AgeSexParchSibSpPclass model means the model using explanatory variables of age, sex, parch, sibsp and pclass
AgeSexSibSpPclasslogFare model means the model using explanatory variables of age, sex, sibsp and log fare
AgeSexParchSibSpPclasslogFare model means model using explanatory variables of age, sex, parch, sibsp, pclass and log fare
The blue highlighted number shows the smallest total-misclassification rate among the table

AgeSexParchSibSpPclasslogFare Model with 0.5 cutoff has the smallest in-sample misclassification rates, which means it fits the best, but overall the fits of the models look fine.

Out-of-sample mis-class rate:

	Total Misclass Rate with Cut-Off 0.5	Total Misclass Rate with Cut-Off 0.4	Total Misclass Rate with Cut-Off 0.3
AgeSexSibSpPclass	0.2185	0.2101	0.2269
AgeSexParchSibSpPclass	0.2157	0.2129	0.2213
AgeSexSibSpPclasslogFare	0.2241	0.2213	0.2353
AgeSexParchSibSpPclasslogFare	0.2185	0.2213	0.2353

Table 9: Out-of-Sample misclassification for the best 4 models with cut off 0.3, 0.4, and 0.5.
Note:
AgeSexSibSpPclass model means the model using explanatory variables of age, sex, sibsp and pclass
AgeSexParchSibSpPclass model means the model using explanatory variables of age, sex, parch, sibsp and pclass
AgeSexSibSpPclasslogFare model means the model using explanatory variables of age, sex, sibsp and log fare
AgeSexParchSibSpPclasslogFare model means model using explanatory variables of age, sex, parch, sibsp, pclass and log fare
The blue highlighted number shows the smallest total-misclassification rate among the table

AgeSexSibSpPclass with 0.4 cutoff has the smallest out-of-sample misclassification rate. In addition, cutoff of 0.4 looks the best for most of the models (AgeSexSibSpPclass, AgeSexParchSibSpPclass, AgeSexSibSpPclasslogFare).

2.6 Calibration of Fit

We use another method to validate the model by using the method of Calibration of Fit. We divided the probability of survival with 10 bins of different ranges and the observed proportion is summarized as below:

AgeSexSibSpPclass model:

Bins	(0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1]
Observed prop	0.0972	0.1556	0.1765	0.3636	0.5000	0.2222	0.7000	0.7273	0.9259	0.9706

Table 10: Calibration of fit for AgeSexSibSpPclass model

Note: AgeSexSibSpPclass model means the model using explanatory variables of age, sex, sibsp and pclass
The purple highlighted number shows the observed proportions that are out of range

AgeSexParchSibSpPclass model:

Bins	(0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1]
Observed prop	0.0896	0.1579	0.1714	0.4167	0.4118	0.2667	0.6111	0.7600	0.9231	0.9714

Table 11: Calibration of fit for AgeSexParchSibSpPclass model

Note: AgeSexParchSibSpPclass model means the model using explanatory variables of age, sex, parch, sibsp and pclass
The purple highlighted number shows the observed proportions that are out of range

AgeSexSibSpPclasslogFare model:

Bins	(0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1]
Observed prop	0.0833	0.1505	0.2059	0.3684	0.4500	0.3500	0.6667	0.8333	0.8750	0.9722

Table 12: Calibration of fit for AgeSexSibSpPclasslogFare model

Note: AgeSexSibSpPclasslogFare model means the model using explanatory variables of age, sex, sibsp, pclass and log fare
The purple highlighted number shows the observed proportions that are out of range

AgeSexParchSibSpPclasslogFare model:

Bins	(0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1]
Observed prop	0.0896	0.1250	0.2308	0.5263	0.2941	0.4706	0.5625	0.8148	0.8636	0.9730

Table 13: Calibration of fit for AgeSexParchSibSpPclassLogFare model

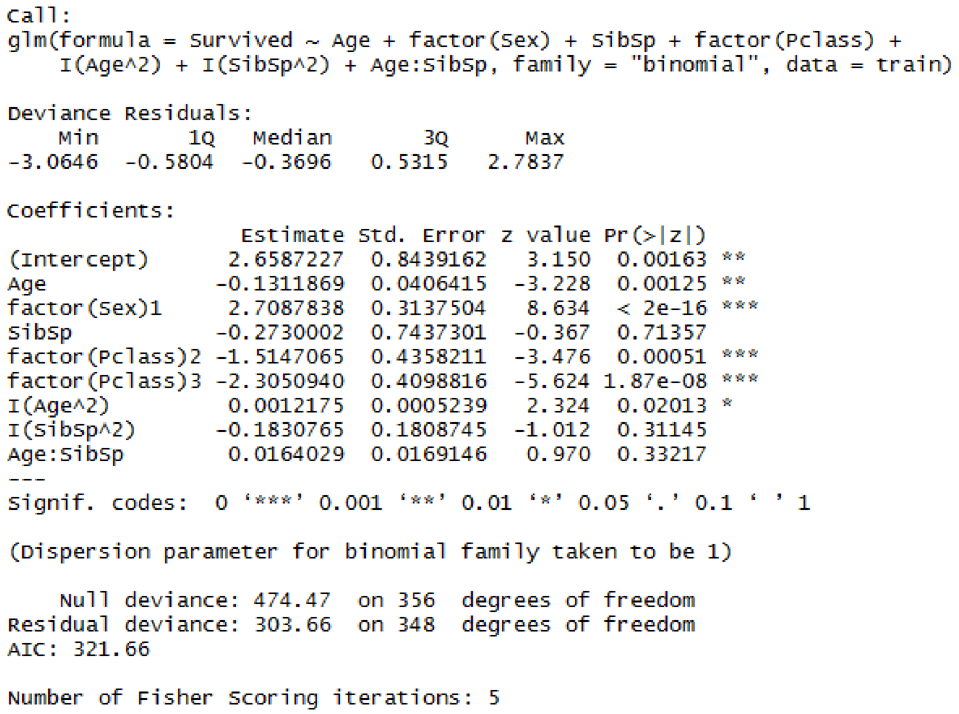
Note: AgeSexParchSibSpPclassLogFare model means the model using explanatory variables of age, sex, parch, sibsp, pclass and log fare
The purple highlighted number shows the observed proportions that are out of range

AgeSexSibSpPclasslogFare model has the best calibration (with only 2 bin out of range). Overall fit for the models look okay.

2.7 Quadratic Model

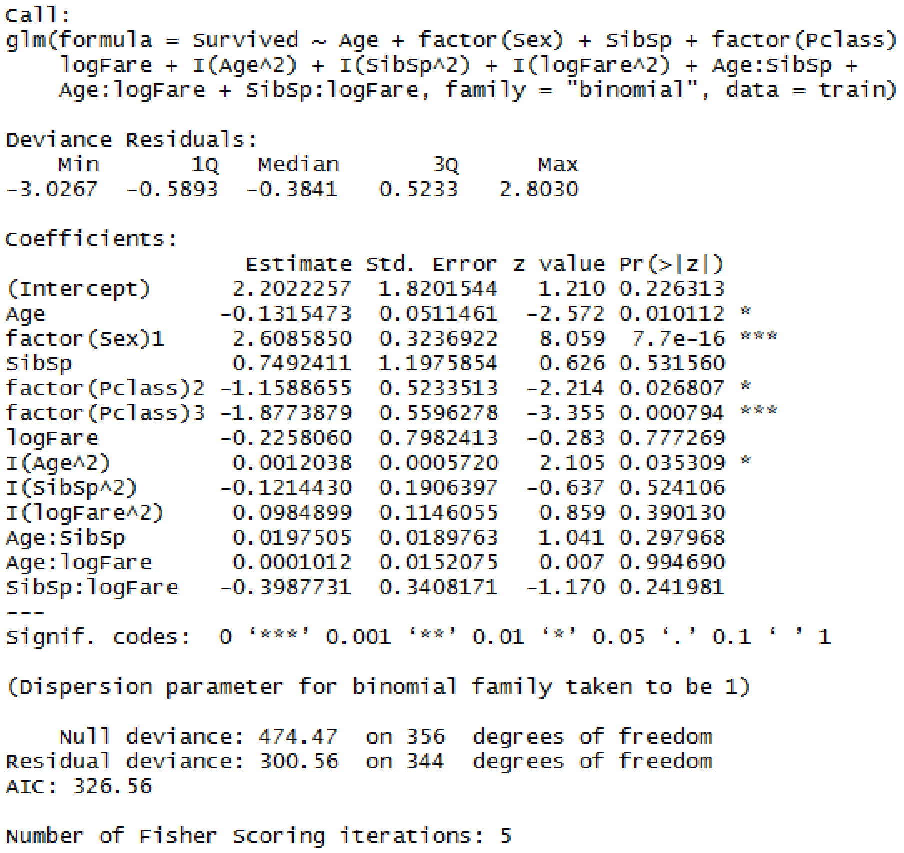
We chose the two linear models: AgeSexSibSpPclass (model with the smallest out-of-sample misclassification rate) and AgeSexSibSpPclasslogFare (model with the smallest AIC), and added quadratic terms using the numerical explanatory variables.

"Age_Sex_SibSp_Pclass_quad"



```
Call:
glm(formula = Survived ~ Age + factor(Sex) + SibSp + factor(Pclass) + 
    I(Age^2) + I(SibSp^2) + Age:SibSp, family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.0646  -0.5804  -0.3696   0.5315   2.7837  

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      2.6587227  0.8439162   3.150  0.00163 ** 
Age             -0.1311869  0.0406415  -3.228  0.00125 ** 
factor(Sex)1     2.7087838  0.3137504   8.634  < 2e-16 ***
SibSp           -0.2730002  0.7437301  -0.367  0.71357    
factor(Pclass)2 -1.5147065  0.4358211  -3.476  0.00051 ***
factor(Pclass)3 -2.3050940  0.4098816  -5.624 1.87e-08 ***
I(Age^2)         0.0012175  0.0005239   2.324  0.02013 *  
I(SibSp^2)      -0.1830765  0.1808745  -1.012  0.31145    
Age:SibSp        0.0164029  0.0169146   0.970  0.33217    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 474.47  on 356  degrees of freedom
Residual deviance: 303.66  on 348  degrees of freedom
AIC: 321.66

Number of Fisher Scoring iterations: 5
```

"Age_Sex_SibSp_Pclass_logFare_quad"



```
Call:
glm(formula = Survived ~ Age + factor(Sex) + SibSp + factor(Pclass) + 
    logFare + I(Age^2) + I(SibSp^2) + I(logFare^2) + Age:SibSp + 
    Age:logFare + SibSp:logFare, family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.0267  -0.5893  -0.3841   0.5233   2.8030  

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)      2.2022257  1.8201544   1.210 0.226313    
Age             -0.1315473  0.0511461  -2.572 0.010112 *  
factor(Sex)1     2.6085850  0.3236922   8.059  7.7e-16 ***
SibSp            0.7492411  1.1975854   0.626 0.531560    
factor(Pclass)2 -1.1588655  0.5233513  -2.214 0.026807 *  
factor(Pclass)3 -1.8773879  0.5596278  -3.355 0.000794 ***
logFare         -0.2258060  0.7982413  -0.283 0.777269    
I(Age^2)         0.0012038  0.0005720   2.105 0.035309 *  
I(SibSp^2)      -0.1214430  0.1906397  -0.637 0.524106    
I(logFare^2)     0.0984899  0.1146055   0.859 0.390130    
Age:SibSp        0.0197505  0.0189763   1.041 0.297968    
Age:logFare      0.0001012  0.0152075   0.007 0.994690    
SibSp:logFare   -0.3987731  0.3408171  -1.170 0.241981    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 474.47  on 356  degrees of freedom
Residual deviance: 300.56  on 344  degrees of freedom
AIC: 326.56

Number of Fisher Scoring iterations: 5
```

AgeSexSibSpPclass_quad has lower AIC (321.66) than any of the other models. However, AgeSexSibSpPclasslogFare_quad has worse AIC compared to all the linear models (326.56).

2.7.1 Misclassification

In-Sample: *refer to table 7 for comparison to previous models

	Total Misclass Rate with Cut-Off 0.5	Total Misclass Rate with Cut-Off 0.4	Total Misclass Rate with Cut-Off 0.3
AgeSexSibSpPclass_quad	0.1597	0.1793	0.2017
AgeSexSibSpPclasslogFare_quad	0.1765	0.1765	0.1961

Table 14: In-Sample misclassification for the quadratic models based on the best linear models with cut off 0.3, 0.4, and 0.5.
Note:
AgeSexSibSpPclass_quad model means the model using explanatory variables of age, sex, sibsp and pclass, age^2, sibsp^2, age:sibsp
AgeSexSibSpPclasslogFare model means the model using explanatory variables of age, sex, sibsp and log fare, age^2, sibsp^2, lg fare^2, age:sibsp, age:log fare, sibsp:log fare

Adding quadratic variables lowered the in-sample misclassification rate for both models. AgeSexSibSpPclass_quad with 0.5 cutoff has the smallest in-sample misclassification rates out of all the models. Overall fit look okay.

Out-of-Sample: *refer to table 8 for comparison to previous models

	Total Misclass Rate with Cut-Off 0.5	Total Misclass Rate with Cut-Off 0.4	Total Misclass Rate with Cut-Off 0.3
AgeSexSibSpPclass_quad	0.2129	0.2157	0.2465
AgeSexSibSpPclasslogFare_quad	0.2269	0.2185	0.2269

Table 15: Out-of-sample misclassification for the quadratic models based on the best linear models with cut off 0.3, 0.4, and 0.5.
Note:
AgeSexSibSpPclass_quad model means the model using explanatory variables of age, sex, sibsp and pclass, age^2, sibsp^2, age:sibsp
AgeSexSibSpPclasslogFare model means the model using explanatory variables of age, sex, sibsp and log fare, age^2, sibsp^2, lg fare^2, age:sibsp, age:log fare, sibsp:log fare

Adding quadratic variables lowered the out-of-sample misclassification rates for some of the categories compared to the original models, however the best out-of-sample misclassification rate remains linear AgeSexSibSpPclass with 0.4 cut-off.

2.7.2 Calibration of Fit

AgeSexSibSpPclass_qaud model

Bins	(0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1]
Observed prop	0.1325	0.0813	0.0556	0.3571	0.3333	0.5909	0.8000	0.8095	0.8636	0.9512

Table 16: Calibration of fit for AgeSexSibSpPclass_quad model

Note: AgeSexSibSpPclass_quad model means the model using explanatory variables of age, sex, sibsp, and quadratic of pclass
The purple highlighted number shows the observed proportions that are out of range

AgeSexParchSibSpPclasslogFare_quad model

Bins	(0,0.1]	(0.1,0.2]	(0.2,0.3]	(0.3,0.4]	(0.4,0.5]	(0.5,0.6]	(0.6,0.7]	(0.7,0.8]	(0.8,0.9]	(0.9,1]
Observed prop	0.1098	0.1023	0.1600	0.3478	0.5000	0.4800	0.7143	0.8000	0.9032	0.9429

Table 17: Calibration of fit for AgeSexParchSibSpPclassLogFare_quad model

Note: AgeSexParchSibSpPclassLogFare_quad model means the model using explanatory variables of age, sex, parch, sibsp, pclass and quadratic of log fare
The purple highlighted number shows the observed proportions that are out of range

The tested quadratic models have more bins that are out of range, but the magnitude of "deviation" from the intervals seem to be still small. Overall fit look okay.

c variables. Based on the fitted models, we got several best models as follows:

	AIC	In-sample misclass rate			Out-of-sample misclass rate			# of calibration bins out of range
Cut-offs		0.5	0.4	0.3	0.5	0.4	0.3	
AgeSexSibSpPclass	324.73	0.2045	0.2045	0.2213	0.2185	0.2101	0.2269	3
AgeSexParchSibSpPclass	326.29	0.1989	0.2073	0.2185	0.2157	0.2129	0.2213	4
AgeSexSibSpPclasslogFare	323.97	0.1933	0.1989	0.2129	0.2241	0.2213	0.2353	2
AgeSexParchSibSpPclasslogFare	324.13	0.1877	0.2073	0.2045	0.2185	0.2213	0.2353	5
AgeSexSibSpPclass_quad	321.66	0.1597	0.1793	0.2017	0.2129	0.2157	0.2465	6
AgeSexSibSpPclasslogFare_quad	326.56	0.1765	0.1765	0.1961	0.2269	0.2185	0.2269	5
Best Model	AgeSexSibSpPclass_quad	AgeSexSibSpPclass_quad	AgeSexSibSpPclasslogFare_quad	AgeSexSibSpPclasslogFare_quad	AgeSexSibSpPclass_quad	AgeSexSibSpPclass	AgeSexParchSibSpPclass	AgeSexSibSpPclasslogFare

2.8. Conclusion and Discussion

In this study we use two statistical methods to produce multiple models. We produced 4 models using logistic regression with partly transformed log variables, and based on the 4 models, we selected the best two models by the lowest AIC and smallest out-of-sample misclass rate, and modified those best two models by using logistic regression but with partly transformed quadrati

						🟦		
							🟨	
								🟪
	🟥	🟥			🟥			
			🟩	🟩				
	🟥	🟥	🟩	🟩	🟥	🟦	🟨	🟪

Table 18: Summary for best 6 models

Note:
AgeSexSibSpPclass model means the model using explanatory variables of age, sex, sibsp, pclass
AgeSexParchSibSpPclass model means the model using explanatory variables of age, sex, parch, sibsp, pclass
AgeSexSibSpPclassLogFare model means the model using explanatory variables of age, sex, sibsp, pclass and log fare
AgeSexParchSibSpPclassLogFare model means the model using explanatory variables of age, sex, parch, sibsp, pclass and log fare
AgeSexParchSibSpPclassLogFare_quad model means the model using explanatory variables of age, sex, parch, sibsp, pclass and quadratic of log fare
AgeSexSibSpPclasslogFare_quad model means the model using explanatory variables of age, sex,, sibsp, pclass and quadratic of log fare
The cells with the same color indicate the same model.

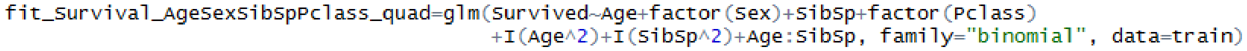
Based on the above table, we concluded that there is not a single model that performed significantly better than all the others. For example, even though the out-of-sample misclass rate is lower for one model compared to another, it could be due to random chance or the test data size being too small, since the differences are quite small.

In addition, we notice that some models are overfitted and we can eliminate those. For instance, we notice that the AgeSexParchSibSpPclass_quad model fitted the training data very well, because the in-sample misclass rate performs the best among the rest models, but for the out-of-sample misclass rate, it is not the best. Therefore we eliminated the AgeSexParchSibSpPclass_quad model.

We also prefer to decide on the best model based on out-of-sample misclass rate because this validation is for prediction, rather than fitting of the training set. However, we would also take considerations in the other models based on AIC, in-sample misclass rate and # of calibration bins out of range.

Therefore, we suggested a few best models and listed below:

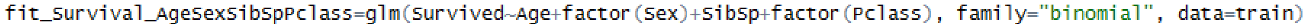
AgeSexSibSpPclass_quad model:



```
fit_survival_AgeSexSibSpPclass_quad=glm(Survived~Age+factor(Sex)+SibSp+factor(Pclass)
                                       +I(Age^2)+I(SibSp^2)+Age:SibSp, family="binomial", data=train)
```

This model has the smallest AIC and its in-sample and out-of-sample misclassification rates are quite small, meaning it fits very well to the training data, yet it is also quite good for prediction.

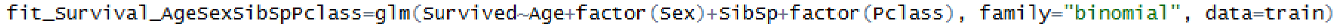
AgeSexSibSpPclass model:



```
fit_survival_AgeSexSibSpPclass=glm(Survived~Age+factor(Sex)+SibSp+factor(Pclass), family="binomial", data=train)
```

This model has one of the smallest AIC and it has the smallest out-of-sample misclassification rate at a cutoff of 0.4 over all cutoffs, meaning it fits well to the training data, and it is very good for prediction (although the small difference in misclassification rates could be due to chance).

AgeSexParchSibSpPclass model:



```
fit_survival_AgeSexSibSpPclass=glm(Survived~Age+factor(Sex)+SibSp+factor(Pclass), family="binomial", data=train)
```

This model has a larger AIC compared to the other models tested but it has some of the smallest out-of-sample misclassification rates, meaning it is very good for prediction and does not overfit.

Section 3: Appendix

3.1 Explanation for R codes

The below R code is for summary of variables in section 2.1.

```
titanic_na <- read.delim("~/cleaned_data.txt")
# remove rows where Age is NA
titanic=na.omit(titanic_na)
View(titanic)

ntot=nrow(titanic)
names(train)
sum(titanic$Survived)
mean(titanic$Survived)

#categorical data
table(titanic$Survived) #response variable
table(titanic$Pclass)
table(titanic$Sex)

#continuous/non-categorical data
summary(titanic$Age)
summary(titanic$SibSp)
summary(titanic$Parch)
summary(titanic$Fare)
```

The below R code is for data transformation and plotting graphs in section 2.2.

```
#trasforming continuous/non-categorical data
titanic$sqrtAge=sqrt(titanic$Age)
titanic$sqrtSibSp=sqrt(titanic$SibSp)
titanic$sqrtParch=sqrt(titanic$Parch)
titanic$sqrtFare=sqrt(titanic$Fare)

titanic$logAge=log(1+titanic$Age)
titanic$logSibSp=log(1+titanic$SibSp)
titanic$logParch=log(1+titanic$Parch)
titanic$logFare=log(1+titanic$Fare)

##histograms/deciding on transformations for categorical/non-categorical data
hist(titanic$Age)
hist(titanic$sqrtAge)
hist(titanic$logAge)
#non-transformed data is the most symmetric, and will be used

hist(titanic$SibSp)
hist(titanic$sqrtSibSp)
hist(titanic$logSibSp)
#non-transformed data is right skewed, but range is small enough and the
transformations
#did not really help much, so the non-transformed data will be used

hist(titanic$Parch)
hist(titanic$sqrtParch)
hist(titanic$logParch)
#non-transformed data is right skewed, but range is small enough and the
transformations
#did not really help much, so the non-transformed data will be used

hist(titanic$Fare)
hist(titanic$sqrtFare)
hist(titanic$logFare)
#non-transformed data is right skewed, and the log-transformed data seems to be the
most
#symmetric, so the log-transformed data will be used
```

The below R code is for box plots between continuous explanatory variables and response variable, and for creating contingency tables for categorical variables. We tried ggplot methods as well, but did not end up using it.

```
## plotting
#install.packages("ggplot2", dependencies = TRUE)
library(ggplot2)

## boxplots for continuous/non-categorical variables
plot(factor(titanic$Survived),titanic$Age, xlab = "Survived or not", ylab = "Age")
plot(factor(titanic$Survived),titanic$Parch, xlab = "Survived or not", ylab =
"Parch")
plot(factor(titanic$Survived),titanic$SibSp,  xlab = "Survived or not", ylab =
"SibSp")
plot(factor(titanic$Survived),titanic$logFare,  xlab = "Survived or not", ylab =
"logFare")

## tables for categorical variables
table(factor(titanic$Survived),factor(titanic$Pclass))
table(factor(titanic$Survived),factor(titanic$Sex))

## bargraphs for categorical variables
titanic$notSurvived=1-titanic$Survived
ggplot(titanic, aes(x=titanic$Pclass, y=titanic$Survived)) +
geom_bar(stat="identity")
ggplot(titanic, aes(x=titanic$Sex, y=titanic$Survived)) + geom_bar(stat="identity")
ggplot(titanic, aes(x=titanic$Pclass, y=titanic$notSurvived)) +
geom_bar(stat="identity")
ggplot(titanic, aes(x=titanic$Sex, y=titanic$notSurvived)) +
geom_bar(stat="identity")
```

The below is the R code for creating the training and hold-out set. Our training and hold-out sets have same size, and summary for the training set is explained in section 2.4.

```
set.seed(12345)
# 357 for training set, 357 for holdout
iperm=sample(ntot,ntot) # random permutation of 1...ntot
n=357
train=titanic[iperm[1:n],]
hold=titanic[iperm[(n+1):ntot],]
View(train)
View(hold)

names(train)
sum(train$Survived)
mean(train$Survived)

#correlation
attach(train)
summcor=cor(train[,c(2:7,16)],train[,c(2:7,16)])
print(summcor)
detach(train)
```

For selecting best models in 2.4, we first tried fitting all the combinations of explanatory variables manually (in section 2.4.1.), and then tried using stepAIC method (not included in the main part of the report). We ended up selecting similar models. Below is the R code.

```
## We tried variable selection manually first
##fitting 1 explanatory variable
fit_Survival_Age=glm(Survived~Age, family="binomial", data=train)
summSurvival_Age=summary(fit_Survival_Age)
fit_Survival_Sex=glm(Survived~factor(Sex), family="binomial", data=train)
```

```
summSurvival_Sex=summary(fit_Survival_Sex)
fit_Survival_SibSp=glm(Survived~SibSp, family="binomial", data=train)
summSurvival_SibSp=summary(fit_Survival_SibSp)
fit_Survival_Parch=glm(Survived~Parch, family="binomial", data=train)
summSurvival_Parch=summary(fit_Survival_Parch)
fit_Survival_logFare=glm(Survived~logFare, family="binomial", data=train)
summSurvival_logFare=summary(fit_Survival_logFare)
fit_Survival_Pclass=glm(Survived~factor(Pclass), family="binomial", data=train)
summSurvival_Pclass=summary(fit_Survival_Pclass)

##fitting 2 explanatory variables
fit_Survival_AgeSex=glm(Survived~Age+factor(Sex), family="binomial", data=train)
summSurvival_AgeSex=summary(fit_Survival_AgeSex)
fit_Survival_AgeSibSp=glm(Survived~Age+SibSp, family="binomial", data=train)
summSurvival_AgeSipSp=summary(fit_Survival_AgeSibSp)
fit_Survival_AgePclass=glm(Survived~Age+factor(Pclass), family="binomial",
data=train)
summSurvival_AgePclass=summary(fit_Survival_AgePclass)
fit_Survival_AgeParch=glm(Survived~Age+Parch, family="binomial", data=train)
summSurvival_AgeParch=summary(fit_Survival_AgeParch)
fit_Survival_AgelogFare=glm(Survived~Age+logFare, family="binomial", data=train)
summSurvival_AgelogFare=summary(fit_Survival_AgelogFare)

fit_Survival_SexPclass=glm(Survived~factor(Pclass)+factor(Sex), family="binomial",
data=train)
summSurvival_SexPclass=summary(fit_Survival_SexPclass)
fit_Survival_SexSibSp=glm(Survived~factor(Sex)+SibSp, family="binomial",
data=train)
summSurvival_SexSipSp=summary(fit_Survival_SexSibSp)
fit_Survival_SexParch=glm(Survived~factor(Sex)+Parch, family="binomial",
data=train)
summSurvival_SexParch=summary(fit_Survival_SexParch)
fit_Survival_SexlogFare=glm(Survived~factor(Sex)+logFare, family="binomial",
data=train)
summSurvival_SexlogFare=summary(fit_Survival_SexParch)

fit_Survival_PclassSibSp=glm(Survived~factor(Pclass)+SibSp, family="binomial",
data=train)
summSurvival_PclassSibSp=summary(fit_Survival_PclassSibSp)
fit_Survival_PclassParch=glm(Survived~factor(Pclass)+Parch, family="binomial",
data=train)
summSurvival_PclassParch=summary(fit_Survival_PclassParch)
fit_Survival_PclasslogFare=glm(Survived~factor(Pclass)+logFare, family="binomial",
data=train)
summSurvival_PclasslogFare=summary(fit_Survival_PclasslogFare)

fit_Survival_SibSplogFare=glm(Survived~logFare+SibSp, family="binomial",
data=train)
summSurvival_SibSplogFare=summary(fit_Survival_SibSplogFare)
fit_Survival_SibSpParch=glm(Survived~SibSp+Parch, family="binomial", data=train)
summSurvival_SibSpParch=summary(fit_Survival_SibSpParch)

fit_Survival_logFareParch=glm(Survived~logFare+Parch, family="binomial",
data=train)
summSurvival_logFareParch=summary(fit_Survival_logFareParch)

##fitting 3 explanatory variables
fit_Survival_AgeSexSibSp=glm(Survived~Age+factor(Sex)+SibSp, family="binomial",
data=train)
summSurvival_AgeSexSibSp=summary(fit_Survival_AgeSexSibSp)
fit_Survival_AgeSexlogFare=glm(Survived~Age+factor(Sex)+logFare, family="binomial",
data=train)
summSurvival_AgeSexlogFare=summary(fit_Survival_AgeSexlogFare)
fit_Survival_AgeSexParch=glm(Survived~Age+factor(Sex)+Parch, family="binomial",
data=train)
summSurvival_AgeSexParch=summary(fit_Survival_AgeSexParch)
fit_Survival_AgeSexPclass=glm(Survived~Age+factor(Sex)+factor(Pclass),
family="binomial", data=train)
summSurvival_AgeSexPclass=summary(fit_Survival_AgeSexPclass)

fit_Survival_AgeSibSpPclass=glm(Survived~Age+SibSp+factor(Pclass),
family="binomial", data=train)
summSurvival_AgeSibSpPclass=summary(fit_Survival_AgeSibSpPclass)
fit_Survival_AgeSibSplogFare=glm(Survived~Age+SibSp+logFare, family="binomial",
data=train)
summSurvival_AgeSibSplogFare=summary(fit_Survival_AgeSibSplogFare)
fit_Survival_AgeSibSpParch=glm(Survived~Age+SibSp+Parch, family="binomial",
data=train)
summSurvival_AgeSibSpParch=summary(fit_Survival_AgeSibSpParch)

fit_Survival_AgePclasslogFare=glm(Survived~Age+factor(Pclass)+logFare,
family="binomial", data=train)
summSurvival_AgePclasslogFare=summary(fit_Survival_AgePclasslogFare)
fit_Survival_AgePclassParch=glm(Survived~Age+factor(Pclass)+Parch,
family="binomial", data=train)
summSurvival_AgePclassParch=summary(fit_Survival_AgePclassParch)

fit_Survival_AgelogFareParch=glm(Survived~Age+logFare+Parch, family="binomial",
data=train)
summSurvival_AgelogFareParch=summary(fit_Survival_AgelogFareParch)

fit_Survival_SexSibSplogFare=glm(Survived~logFare+factor(Sex)+SibSp,
family="binomial", data=train)
summSurvival_SexSibSplogFare=summary(fit_Survival_SexSibSplogFare)
fit_Survival_SexSibSpParch=glm(Survived~SibSp+factor(Sex)+Parch, family="binomial",
data=train)
summSurvival_SexSibSpParch=summary(fit_Survival_SexSibSpParch)
fit_Survival_SexSibSpPclass=glm(Survived~SibSp+factor(Sex)+factor(Pclass),
family="binomial", data=train)
summSurvival_SexSibSpPclass=summary(fit_Survival_SexSibSpPclass)

fit_Survival_SexPclasslogFare=glm(Survived~factor(Sex)+factor(Pclass)+logFare,
family="binomial", data=train)
summSurvival_SexPclasslogFare=summary(fit_Survival_SexPclasslogFare)
fit_Survival_SexPclassParch=glm(Survived~factor(Sex)+factor(Pclass)+Parch,
family="binomial", data=train)
```

```
summSurvival_SexPclassParch=summary(fit_Survival_SexPclassParch)

fit_Survival_SexlogFareParch=glm(Survived~factor(Sex)+logFare+Parch,
family="binomial", data=train)
summSurvival_SexlogFareParch=summary(fit_Survival_SexlogFareParch)

fit_Survival_ParchlogFareSibSp=glm(Survived~Parch+SibSp+logFare, family="binomial",
data=train)
summSurvival_ParchlogFareSibSp=summary(fit_Survival_ParchlogFareSibSp)
fit_Survival_ParchlogFarePclass=glm(Survived~logFare+factor(Pclass)+Parch,
family="binomial", data=train)
summSurvival_ParchlogFarePclass=summary(fit_Survival_ParchlogFarePclass)

fit_Survival_ParchPclassSibSp=glm(Survived~Parch+factor(Pclass)+SibSp,
family="binomial", data=train)
summSurvival_ParchPclassSibSp=summary(fit_Survival_ParchPclassSibSp)

fit_Survival_logFareSibSpPclass=glm(Survived~SibSp+logFare+factor(Pclass),
family="binomial", data=train)
summSurvival_logFareSibSpPclass=summary(fit_Survival_logFareSibSpPclass)

##fitting 4 explanatory variables
fit_Survival_AgeSexSibSpParch=glm(Survived~Age+factor(Sex)+SibSp+Parch,
family="binomial", data=train)
summSurvival_AgeSexSibSpParch=summary(fit_Survival_AgeSexSibSpParch)
fit_Survival_AgeSexSibSplogFare=glm(Survived~Age+factor(Sex)+SibSp+logFare,
family="binomial", data=train)
summSurvival_AgeSexSibSplogFare=summary(fit_Survival_AgeSexSibSplogFare)
fit_Survival_AgeSexSibSpPclass=glm(Survived~Age+factor(Sex)+SibSp+factor(Pclass),
family="binomial", data=train)
summSurvival_AgeSexSibSpPclass=summary(fit_Survival_AgeSexSibSpPclass)

fit_Survival_AgeSexPclasslogFare=glm(Survived~Age+factor(Sex)+factor(Pclass)+logFar
e, family="binomial", data=train)
summSurvival_AgeSexPclasslogFare=summary(fit_Survival_AgeSexPclasslogFare)
fit_Survival_AgeSexPclassParch=glm(Survived~Age+factor(Sex)+factor(Pclass)+Parch,
family="binomial", data=train)
summSurvival_AgeSexPclassParch=summary(fit_Survival_AgeSexPclassParch)

fit_Survival_AgeSexlogFareParch=glm(Survived~Age+factor(Sex)+logFare+Parch,
family="binomial", data=train)
summSurvival_AgeSexlogFareParch=summary(fit_Survival_AgeSexlogFareParch)

fit_Survival_AgeSibSpPclasslogFare=glm(Survived~Age+SibSp+factor(Pclass)+logFare,
family="binomial", data=train)
summSurvival_AgeSibSpPclasslogFare=summary(fit_Survival_AgeSibSpPclasslogFare)
fit_Survival_AgeSibSpPclassParch=glm(Survived~Age+SibSp+factor(Pclass)+Parch,
family="binomial", data=train)
summSurvival_AgeSibSpPclassParch=summary(fit_Survival_AgeSibSpPclassParch)

fit_Survival_AgeSibSplogFareParch=glm(Survived~Age+SibSp+logFare+Parch,
family="binomial", data=train)
summSurvival_AgeSibSplogFareParch=summary(fit_Survival_AgeSibSplogFareParch)
```

```
fit_Survival_AgePclasslogFareParch=glm(Survived~Age+factor(Pclass)+logFare+Parch,
family="binomial", data=train)
summSurvival_AgePclasslogFareParch=summary(fit_Survival_AgePclasslogFareParch)

fit_Survival_SexlogFarePclassSibSp=glm(Survived~SibSp+factor(Sex)+logFare+factor(Pc
lass), family="binomial", data=train)
summSurvival_SexlogFarePclassSibSp=summary(fit_Survival_SexlogFarePclassSibSp)
fit_Survival_SexlogFarePclassParch=glm(Survived~Parch+factor(Sex)+logFare+factor(Pc
lass), family="binomial", data=train)
summSurvival_SexlogFarePclassParch=summary(fit_Survival_SexlogFarePclassParch)

fit_Survival_SexlogFareSibspParch=glm(Survived~factor(Sex)+logFare+SibSp+Parch,
family="binomial", data=train)
summSurvival_SexlogFareSibspParch=summary(fit_Survival_SexlogFareSibspParch)

fit_Survival_SexPclassSibspParch=glm(Survived~factor(Sex)+factor(Pclass)+SibSp+Parc
h, family="binomial", data=train)
summSurvival_SexPclassSibspParch=summary(fit_Survival_SexPclassSibspParch)

fit_Survival_logFarePclassSibSpParch=glm(Survived~Parch+SibSp+logFare+factor(Pclass
), family="binomial", data=train)
summSurvival_logFarePclassSibSpParch=summary(fit_Survival_logFarePclassSibSpParch)

##fitting 5 explanatory variables
fit_Survival_AgeSexParchSibSpPclass=glm(Survived~Age+factor(Sex)+Parch+SibSp+factor
(Pclass), family="binomial", data=train)
summSurvival_AgeSexParchSibSpPclass=summary(fit_Survival_AgeSexParchSibSpPclass)

fit_Survival_AgeSexParchSibSplogFare=glm(Survived~Age+factor(Sex)+Parch+SibSp+logFa
re, family="binomial", data=train)
summSurvival_AgeSexParchSibSplogFare=summary(fit_Survival_AgeSexParchSibSplogFare)

fit_Survival_AgeSexSibSpPclasslogFare=glm(Survived~Age+factor(Sex)+SibSp+factor(Pcl
ass)+logFare, family="binomial", data=train)
summSurvival_AgeSexSibSpPclasslogFare=summary(fit_Survival_AgeSexSibSpPclasslogFare
)

fit_Survival_AgeSexParchPclasslogFare=glm(Survived~Age+factor(Sex)+Parch+factor(Pcl
ass)+logFare, family="binomial", data=train)
summSurvival_AgeSexParchPclasslogFare=summary(fit_Survival_AgeSexParchPclasslogFare
)

fit_Survival_AgeParchSibSpPclasslogFare=glm(Survived~Age+Parch+SibSp+factor(Pclass)
+logFare, family="binomial", data=train)
summSurvival_AgeParchSibSpPclasslogFare=summary(fit_Survival_AgeParchSibSpPclasslog
Fare)

fit_Survival_SexParchSibSpPclasslogFare=glm(Survived~factor(Sex)+Parch+SibSp+factor
(Pclass)+logFare, family="binomial", data=train)
summSurvival_SexParchSibSpPclasslogFare=summary(fit_Survival_SexParchSibSpPclasslog
Fare)
```

```
#fitting 6 explanatory variables
fit_Survival_AgeSexParchSibSpPclasslogFare=glm(Survived~Age+factor(Sex)+Parch+SibSp
+logFare+factor(Pclass), family="binomial", data=train)
summSurvival_AgeSexParchSibSpPclasslogFare=summary(fit_Survival_AgeSexParchSibSpPcl
asslogFare)
#model evaluation
subsetvec=c("Age", "Sex", "SibSp", "logFare", "Pclass", "Parch",

            "Age_Sex","Age_SibSp", "Age_Pclass", "Age_Parch", "Age_logFare",
            "Sex_logFare", "Sex_Parch", "Sex_Pclass", "Sex_SibSp",
            "Pclass_logFare", "Pclass_SibSp", "Pclass_Parch",
            "SibSp_logFare", "SibSp_Parch",
            "logFare_Parch",

            "Age_Sex_Pclass", "Age_Sex_logFare", "Age_Sex_Parch", "Age_Sex_SibSp",
            "Sex_SibSp_logFare", "Sex_SibSp_Parch",
"Sex_SibSp_Pclass","Parch_logFare_SibSp",
            "Parch_logFare_Pclass","logFare_SibSp_Pclass","Age_SibSp_Pclass",
"Age_SibSp_logFare",
            "Age_SibSp_Parch","Age_Pclass_logFare", "Age_Pclass_Parch",
"Age_logFare_Parch",
            "Sex_Pclass_logFare","Sex_Pclass_Parch", "Sex_logFare_Parch",
"Parch_Pclass_SibSp",

            "Age_Sex_SibSp_Parch", "Age_Sex_SibSp_logFare", "Age_Sex_SibSp_Pclass",
            "Sex_logFare_Pclass_SibSp",
"Sex_logFare_Pclass_Parch","logFare_Pclass_SibSp_Parch",
            "Age_Sex_Pclass_logFare",
"Age_Sex_Pclass_Parch","Age_Sex_logFare_Parch",
            "Age_SibSp_Pclass_logFare",
"Age_SibSp_Pclass_Parch","Age_SibSp_logFare_Parch",
            "Age_Pclass_logFare_Parch", "Sex_logFare_Sibsp_Parch",
"Sex_Pclass_Sibsp_Parch",

            "Age_Sex_Parch_SibSp_Pclass", "Age_Sex_Parch_SibSp_logFare",
"Age_Sex_SibSp_Pclass_logFare",
            "Age_Sex_Parch_Pclass_logFare", "Age_Parch_SibSp_Pclass_logFare",
"Sex_Parch_SibSp_Pclass_logFare",

            "Age_Sex_Parch_SibSp_Pclass_logFare")

## create a vector of deviances for each model
deviancevec=c(summSurvival_Age$deviance,summSurvival_Sex$deviance,summSurvival_SibS
p$deviance,summSurvival_logFare$deviance, summSurvival_Pclass$deviance,
summSurvival_Parch$deviance,

              summSurvival_AgeSex$deviance, summSurvival_AgeSipSp$deviance,
summSurvival_AgePclass$deviance, summSurvival_AgeParch$deviance,
summSurvival_AgelogFare$deviance,
              summSurvival_SexlogFare$deviance, summSurvival_SexParch$deviance,
summSurvival_SexPclass$deviance, summSurvival_SexSipSp$deviance,
              summSurvival_PclasslogFare$deviance,
summSurvival_PclassSibSp$deviance, summSurvival_PclassParch$deviance,
              summSurvival_SibSplogFare$deviance, summSurvival_SibSpParch$deviance,
summSurvival_logFareParch$deviance,

              summSurvival_AgeSexPclass$deviance,
summSurvival_AgeSexlogFare$deviance, summSurvival_AgeSexParch$deviance,
summSurvival_AgeSexSibSp$deviance,
              summSurvival_SexSibSplogFare$deviance,
summSurvival_SexSibSpParch$deviance, summSurvival_SexSibSpPclass$deviance,
summSurvival_ParchlogFareSibSp$deviance,
              summSurvival_ParchlogFarePclass$deviance,
summSurvival_logFareSibSpPclass$deviance, summSurvival_AgeSibSpPclass$deviance,
summSurvival_AgeSibSplogFare$deviance,
              summSurvival_AgeSibSpParch$deviance,
summSurvival_AgePclasslogFare$deviance, summSurvival_AgePclassParch$deviance,
summSurvival_AgelogFareParch$deviance,
              summSurvival_SexPclasslogFare$deviance,
summSurvival_SexPclassParch$deviance, summSurvival_SexlogFareParch$deviance,
summSurvival_ParchPclassSibSp$deviance,

              summSurvival_AgeSexSibSpParch$deviance,
summSurvival_AgeSexSibSplogFare$deviance, summSurvival_AgeSexSibSpPclass$deviance,
              summSurvival_SexlogFarePclassSibSp$deviance,
summSurvival_SexlogFarePclassParch$deviance,
summSurvival_logFarePclassSibSpParch$deviance,
              summSurvival_AgeSexPclasslogFare$deviance,
summSurvival_AgeSexPclassParch$deviance, summSurvival_AgeSexlogFareParch$deviance,
              summSurvival_AgeSibSpPclasslogFare$deviance,
summSurvival_AgeSibSpPclassParch$deviance,
summSurvival_AgeSibSplogFareParch$deviance,
              summSurvival_AgePclasslogFareParch$deviance,
summSurvival_SexlogFareSibspParch$deviance,
summSurvival_SexPclassSibspParch$deviance,

              summSurvival_AgeSexParchSibSpPclass$deviance,
summSurvival_AgeSexParchSibSplogFare$deviance,
summSurvival_AgeSexSibSpPclasslogFare$deviance,
              summSurvival_AgeSexParchPclasslogFare$deviance,
summSurvival_AgeParchSibSpPclasslogFare$deviance,
summSurvival_SexParchSibSpPclasslogFare$deviance,

              summSurvival_AgeSexParchSibSpPclasslogFare$deviance)

## create a vector of AICs for each model
aicvec=c(summSurvival_Age$aic,summSurvival_Sex$aic,summSurvival_SibSp$aic,summSurvi
val_logFare$aic, summSurvival_Pclass$aic, summSurvival_Parch$aic,

         summSurvival_AgeSex$aic, summSurvival_AgeSipSp$aic,
summSurvival_AgePclass$aic, summSurvival_AgeParch$aic,summSurvival_AgelogFare$aic,
         summSurvival_SexlogFare$aic, summSurvival_SexParch$aic,
summSurvival_SexPclass$aic, summSurvival_SexSipSp$aic,
         summSurvival_PclasslogFare$aic, summSurvival_PclassSibSp$aic,
summSurvival_PclassParch$aic,
         summSurvival_SibSplogFare$aic, summSurvival_SibSpParch$aic,
summSurvival_logFareParch$aic,

         summSurvival_AgeSexPclass$aic, summSurvival_AgeSexlogFare$aic,
summSurvival_AgeSexParch$aic, summSurvival_AgeSexSibSp$aic,
         summSurvival_SexSibSplogFare$aic, summSurvival_SexSibSpParch$aic,
summSurvival_SexSibSpPclass$aic, summSurvival_ParchlogFareSibSp$aic,
         summSurvival_ParchlogFarePclass$aic, summSurvival_logFareSibSpPclass$aic,
summSurvival_AgeSibSpPclass$aic, summSurvival_AgeSibSplogFare$aic,
         summSurvival_AgeSibSpParch$aic, summSurvival_AgePclasslogFare$aic,
summSurvival_AgePclassParch$aic, summSurvival_AgelogFareParch$aic,
         summSurvival_SexPclasslogFare$aic, summSurvival_SexPclassParch$aic,
summSurvival_SexlogFareParch$aic, summSurvival_ParchPclassSibSp$aic,

         summSurvival_AgeSexSibSpParch$aic, summSurvival_AgeSexSibSplogFare$aic,
summSurvival_AgeSexSibSpPclass$aic,
         summSurvival_SexlogFarePclassSibSp$aic,
summSurvival_SexlogFarePclassParch$aic, summSurvival_logFarePclassSibSpParch$aic,
         summSurvival_AgeSexPclasslogFare$aic, summSurvival_AgeSexPclassParch$aic,
summSurvival_AgeSexlogFareParch$aic,
         summSurvival_AgeSibSpPclasslogFare$aic,
summSurvival_AgeSibSpPclassParch$aic, summSurvival_AgeSibSplogFareParch$aic,
         summSurvival_AgePclasslogFareParch$aic,
summSurvival_SexlogFareSibspParch$aic, summSurvival_SexPclassSibspParch$aic,

         summSurvival_AgeSexParchSibSpPclass$aic,
summSurvival_AgeSexParchSibSplogFare$aic,
summSurvival_AgeSexSibSpPclasslogFare$aic,
         summSurvival_AgeSexParchPclasslogFare$aic,
summSurvival_AgeParchSibSpPclasslogFare$aic,
summSurvival_SexParchSibSpPclasslogFare$aic,

         summSurvival_AgeSexParchSibSpPclasslogFare$aic)

DevAicData=cbind(subsetvec, deviancevec, aicvec)

## we did Variable selection using stepAIC method too, and ended up with the same
best models
library(MASS)
stepAIC(fit_Survival_AgeSexParchSibSpPclasslogFare, direction = "both",trace = 1)
```

The below R code is the summary of the best 3 models based on 3 different methods.

```
# summary of best 3 models based on manual deviance (not so useful - decreases with
the # of variables)
summSurvival_AgeSexParchSibSpPclasslogFare #308.13
summSurvival_AgeSexSibSpPclasslogFare #309.97
summSurvival_AgeSexParchSibSpPclass #312.29

# summary of best 3 models based on manual AIC
summSurvival_AgeSexSibSpPclasslogFare #323.97
summSurvival_AgeSexParchSibSpPclasslogFare #324.13
summSurvival_AgeSexSibSpPclass #324.73
```

```
# summary of best 3 models based on stepAIC method
summSurvival_AgeSexSibSpPclasslogFare #323.97
summSurvival_AgeSexParchSibSpPclasslogFare #324.13
summSurvival_AgeSexSibSpPclass #324.73
```

The below R code is for in-sample misclassification for the best models (`fit_Survival_AgeSexSibSpPclass, fit_Survival_AgeSexParchSibSpPclass,`
`# fit_Survival_AgeSexSibSpPclasslogFare, fit_Survival_AgeSexParchSibSpPclasslogFare`) in section 2.5. We set the boundary to 0.3, 0.4 and 0.5, and calculate the misclassification rate for each of them.

```
#in-sample misclassification
pred4=predict(fit_Survival_AgeSexSibSpPclass, type="response")
pred5_1=predict(fit_Survival_AgeSexParchSibSpPclass, type="response")
pred5_2=predict(fit_Survival_AgeSexSibSpPclasslogFare, type="response")
pred6=predict(fit_Survival_AgeSexParchSibSpPclasslogFare,type="response")

#compare mean of predictions vs mean of actual training data
print(summary(pred4))
print(summary(pred5_1))
print(summary(pred5_2))
print(summary(pred6))

#boundary of 0.5 0.4 0.3 for misclassification
tab4a=table(train$Survived,as.numeric(pred4>0.5))
tab4b=table(train$Survived,as.numeric(pred4>0.4))
tab4c=table(train$Survived,as.numeric(pred4>0.3))

tab5_1a=table(train$Survived,as.numeric(pred5_1>0.5))
tab5_1b=table(train$Survived,as.numeric(pred5_1>0.4))
tab5_1c=table(train$Survived,as.numeric(pred5_1>0.3))

tab5_2a=table(train$Survived,as.numeric(pred5_2>0.5))
tab5_2b=table(train$Survived,as.numeric(pred5_2>0.4))
tab5_2c=table(train$Survived,as.numeric(pred5_2>0.3))

tab6a=table(train$Survived,as.numeric(pred6>0.5))
tab6b=table(train$Survived,as.numeric(pred6>0.4))
tab6c=table(train$Survived,as.numeric(pred6>0.3))

#convert to rates
tab4a/apply(tab4a,1,sum)
tab5_1a/apply(tab5_1a,1,sum)
tab5_2a/apply(tab5_2a,1,sum)
tab6a/apply(tab6a,1,sum)

tab4b/apply(tab4b,1,sum)
tab5_1b/apply(tab5_1b,1,sum)
tab5_2b/apply(tab5_2b,1,sum)
tab6b/apply(tab6b,1,sum)
```

```
tab4c/apply(tab4c,1,sum)
tab5_1c/apply(tab5_1c,1,sum)
tab5_2c/apply(tab5_2c,1,sum)
tab6c/apply(tab6c,1,sum)

#table of misclassification rates
numMis4 = c(tab4a[1,2]+tab4a[2,1],tab4b[1,2]+tab4b[2,1],tab4c[1,2]+tab4c[2,1])/n
numMis5_1 =
c(tab5_1a[1,2]+tab5_1a[2,1],tab5_1b[1,2]+tab5_1b[2,1],tab5_1c[1,2]+tab5_1c[2,1])/n
numMis5_2 =
c(tab5_2a[1,2]+tab5_2a[2,1],tab5_2b[1,2]+tab5_2b[2,1],tab5_2c[1,2]+tab5_2c[2,1])/n
numMis6 = c(tab6a[1,2]+tab6a[2,1],tab6b[1,2]+tab6b[2,1],tab6c[1,2]+tab6c[2,1])/n
```

The below R code is for out-of-sample misclassification for the best models (`fit_Survival_AgeSexSibSpPclass, fit_Survival_AgeSexParchSibSpPclass,`
`# fit_Survival_AgeSexSibSpPclasslogFare, fit_Survival_AgeSexParchSibSpPclasslogFare`) in section 2.5. We set the boundary to 0.3, 0.4 and 0.5, and calculate the misclassification rate for each of them.

```
#out-of-sample misclassification
pred4.hold=predict(fit_Survival_AgeSexSibSpPclass,type="response",newdata=hold)
pred5_1.hold=predict(fit_Survival_AgeSexParchSibSpPclass, type="response", newdata
= hold)
pred5_2.hold=predict(fit_Survival_AgeSexSibSpPclasslogFare, type="response",
newdata = hold)
pred6.hold=predict(fit_Survival_AgeSexParchSibSpPclasslogFare,type="response",newda
ta=hold)

htab4a=table(hold$Survived,as.numeric(pred4.hold>0.5))
htab4b=table(hold$Survived,as.numeric(pred4.hold>0.4))
htab4c=table(hold$Survived,as.numeric(pred4.hold>0.3))

htab5_1a=table(hold$Survived,as.numeric(pred5_1.hold>0.5))
htab5_1b=table(hold$Survived,as.numeric(pred5_1.hold>0.4))
htab5_1c=table(hold$Survived,as.numeric(pred5_1.hold>0.3))

htab5_2a=table(hold$Survived,as.numeric(pred5_2.hold>0.5))
htab5_2b=table(hold$Survived,as.numeric(pred5_2.hold>0.4))
htab5_2c=table(hold$Survived,as.numeric(pred5_2.hold>0.3))

htab6a=table(hold$Survived,as.numeric(pred6.hold>0.5))
htab6b=table(hold$Survived,as.numeric(pred6.hold>0.4))
htab6c=table(hold$Survived,as.numeric(pred6.hold>0.3))

#convert to rates
htab4a/apply(htab4a,1,sum)
htab5_1a/apply(htab5_1a,1,sum)
htab5_2a/apply(htab5_2a,1,sum)
htab6a/apply(htab6a,1,sum)
```

```
htab4b/apply(htab4b,1,sum)
htab5_1b/apply(htab5_1b,1,sum)
htab5_2b/apply(htab5_2b,1,sum)
htab6b/apply(htab6b,1,sum)

htab4c/apply(htab4c,1,sum)
htab5_1c/apply(htab5_1c,1,sum)
htab5_2c/apply(htab5_2c,1,sum)
htab6c/apply(htab6c,1,sum)

#table of misclassification rates
hNumMis4 =
c(htab4a[1,2]+htab4a[2,1],htab4b[1,2]+htab4b[2,1],htab4c[1,2]+htab4c[2,1])/(ntot-n)
hNumMis5_1 =
c(htab5_1a[1,2]+htab5_1a[2,1],htab5_1b[1,2]+htab5_1b[2,1],htab5_1c[1,2]+htab5_1c[2,
1])/(ntot-n)
hNumMis5_2 =
c(htab5_2a[1,2]+htab5_2a[2,1],htab5_2b[1,2]+htab5_2b[2,1],htab5_2c[1,2]+htab5_2c[2,
1])/(ntot-n)
hNumMis6 =
c(htab6a[1,2]+htab6a[2,1],htab6b[1,2]+htab6b[2,1],htab6c[1,2]+htab6c[2,1])/(ntot-n)
```

The below R code is for calibration of fit in section 2.6.

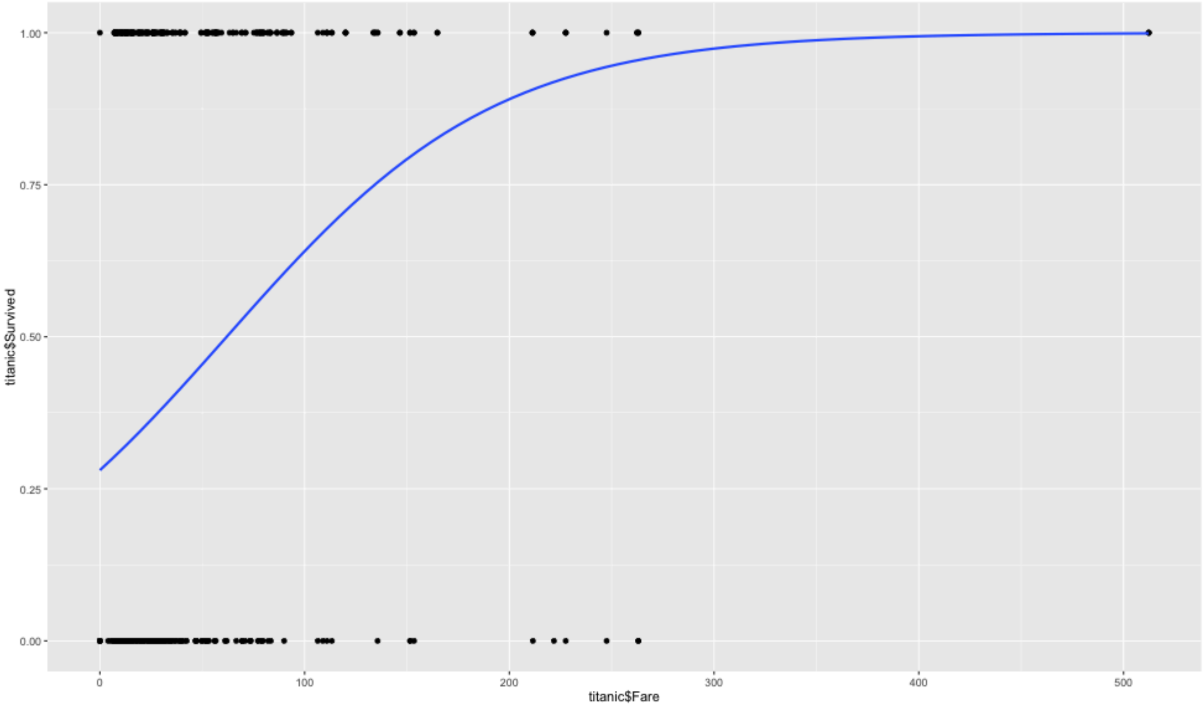
```
# Hosmer-Lemeshow calibration check
prcateg4=cut(pred4,breaks=c(0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1))
print(table(prcateg4))

HLsumm4=tapply(train$Survived,prcateg4,mean)
print(HLsumm4)

prcateg5_1=cut(pred5_1,breaks=c(0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1))
print(table(prcateg5_1))

HLsumm5_1=tapply(train$Survived,prcateg5_1,mean)
print(HLsumm5_1)

prcateg5_2=cut(pred5_2,breaks=c(0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1))
print(table(prcateg5_2))

HLsumm5_2=tapply(train$Survived,prcateg5_2,mean)
print(HLsumm5_2)

prcateg6=cut(pred6,breaks=c(0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1))
print(table(prcateg6))

HLsumm6=tapply(train$Survived,prcateg6,mean)
print(HLsumm6)
```

The below R code is for quadratic models(`fit_Survival_AgeSexSibSpPclass_quad`, `fit_Survival_AgeSexSibSpPclasslogFare_quad`) in section 2.7. Again, we are getting the summary, and in-sample and out-of-sample misclassification rates when cut-off is 0.3, 0.4 and 0.5 for both models, as explained in section 2.7.1. Then, as explained in section 2.7.2, we are checking the calibration for both models.

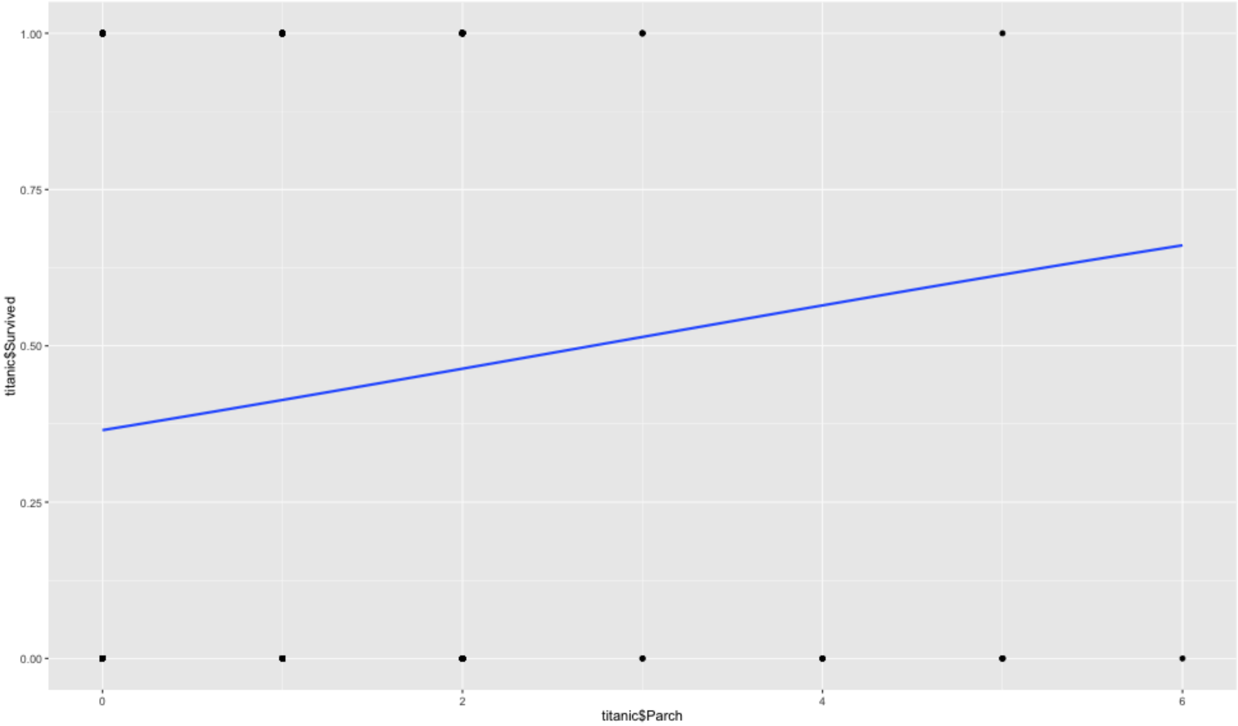
```
#quadratic model of best model based on best out-of-sample # of misclassifications
fit_Survival_AgeSexSibSpPclass_quad=glm(Survived~Age+factor(Sex)+SibSp+factor(Pclas
s)+I(Age^2)+I(SibSp^2)+Age:SibSp, family="binomial", data=train)
summSurvival_AgeSexSibSpPclass_quad=summary(fit_Survival_AgeSexSibSpPclass_quad)

#quadratic model of best model based on AIC
fit_Survival_AgeSexSibSpPclasslogFare_quad=glm(Survived~Age+factor(Sex)+SibSp+facto
r(Pclass)+logFare+I(Age^2)+I(SibSp^2)+I(logFare^2)+Age:SibSp+Age:logFare+SibSp:logF
are, family="binomial", data=train)
summSurvival_AgeSexSibSpPclasslogFare_quad=summary(fit_Survival_AgeSexSibSpPclasslo
gFare_quad)

#do summary for quadratic models (deviance, AIC)
pred4_quad=predict(fit_Survival_AgeSexSibSpPclass_quad, type="response")
pred5_2_quad=predict(fit_Survival_AgeSexSibSpPclasslogFare_quad, type="response")

#compare mean of predictions vs mean of actual training data
print(summary(pred4_quad))
print(summary(pred5_2_quad))

#boundary of 0.5 0.4 0.3 for misclassification
tab4a_quad=table(train$Survived,as.numeric(pred4_quad>0.5))
tab4b_quad=table(train$Survived,as.numeric(pred4_quad>0.4))
tab4c_quad=table(train$Survived,as.numeric(pred4_quad>0.3))

tab5_2a_quad=table(train$Survived,as.numeric(pred5_2_quad>0.5))
tab5_2b_quad=table(train$Survived,as.numeric(pred5_2_quad>0.4))
tab5_2c_quad=table(train$Survived,as.numeric(pred5_2_quad>0.3))

#convert to rates
tab4a_quad/apply(tab4a_quad,1,sum)
tab5_2a_quad/apply(tab5_2a_quad,1,sum)

tab4b_quad/apply(tab4b_quad,1,sum)
tab5_2b_quad/apply(tab5_2b_quad,1,sum)

tab4c_quad/apply(tab4c_quad,1,sum)
tab5_2c_quad/apply(tab5_2c_quad,1,sum)

#table of misclassification rates
numMis4_quad =
c(tab4a_quad[1,2]+tab4a_quad[2,1],tab4b_quad[1,2]+tab4b_quad[2,1],tab4c_quad[1,2]+t
ab4c_quad[2,1])/n
numMis5_2_quad =
c(tab5_2a_quad[1,2]+tab5_2a_quad[2,1],tab5_2b_quad[1,2]+tab5_2b_quad[2,1],tab5_2c_q
uad[1,2]+tab5_2c_quad[2,1])/n

# out-of-sample misclassification
pred4_quad.hold=predict(fit_Survival_AgeSexSibSpPclass_quad,type="response",newdata
=hold)
pred5_2_quad.hold=predict(fit_Survival_AgeSexSibSpPclasslogFare_quad,
type="response", newdata = hold)

htab4a_quad=table(hold$Survived,as.numeric(pred4_quad.hold>0.5))
htab4b_quad=table(hold$Survived,as.numeric(pred4_quad.hold>0.4))
htab4c_quad=table(hold$Survived,as.numeric(pred4_quad.hold>0.3))

htab5_2a_quad=table(hold$Survived,as.numeric(pred5_2_quad.hold>0.5))
htab5_2b_quad=table(hold$Survived,as.numeric(pred5_2_quad.hold>0.4))
htab5_2c_quad=table(hold$Survived,as.numeric(pred5_2_quad.hold>0.3))

#convert to rates
htab4a_quad/apply(htab4a_quad,1,sum)
htab5_2a_quad/apply(htab5_2a_quad,1,sum)

htab4b_quad/apply(htab4b_quad,1,sum)
htab5_2b_quad/apply(htab5_2b_quad,1,sum)

htab4c_quad/apply(htab4c_quad,1,sum)
htab5_2c_quad/apply(htab5_2c_quad,1,sum)

#table of misclassification rates
hNumMis4_quad =
c(htab4a_quad[1,2]+htab4a_quad[2,1],htab4b_quad[1,2]+htab4b_quad[2,1],htab4c_quad[1
,2]+htab4c_quad[2,1])/(ntot-n)
hNumMis5_2_quad =
c(htab5_2a_quad[1,2]+htab5_2a_quad[2,1],htab5_2b_quad[1,2]+htab5_2b_quad[2,1],htab5
_2c_quad[1,2]+htab5_2c_quad[2,1])/(ntot-n)

# Hosmer-Lemeshow calibration check
prcateg4_quad=cut(pred4_quad,breaks=c(0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1))
print(table(prcateg4))

HLsumm4_quad=tapply(train$Survived,prcateg4_quad,mean)
print(HLsumm4_quad)

prcateg5_2_quad=cut(pred5_2_quad,breaks=c(0,.1,.2,.3,.4,.5,.6,.7,.8,.9,1))
print(table(prcateg5_2))

HLsumm5_2_quad=tapply(train$Survived,prcateg5_2_quad,mean)
print(HLsumm5_2_quad)
```

3.2 Additional Scatter Plots

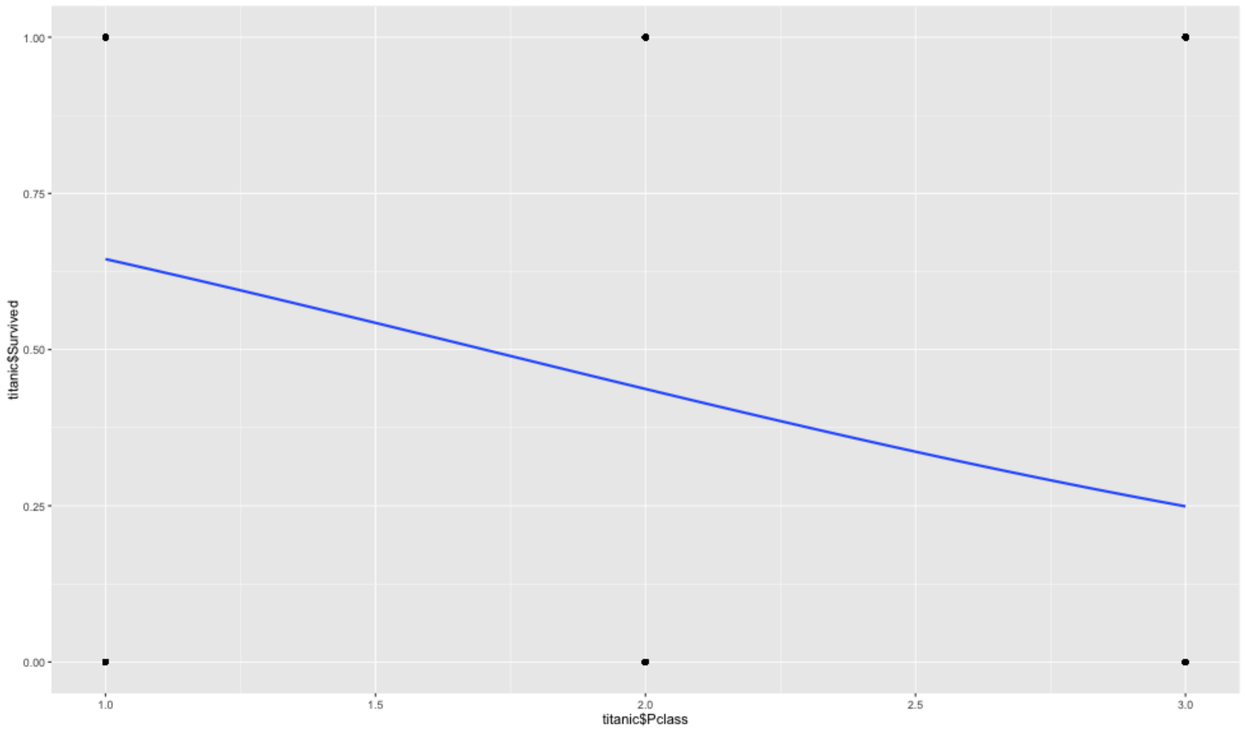
Scatter plot between Survival vs. Fare



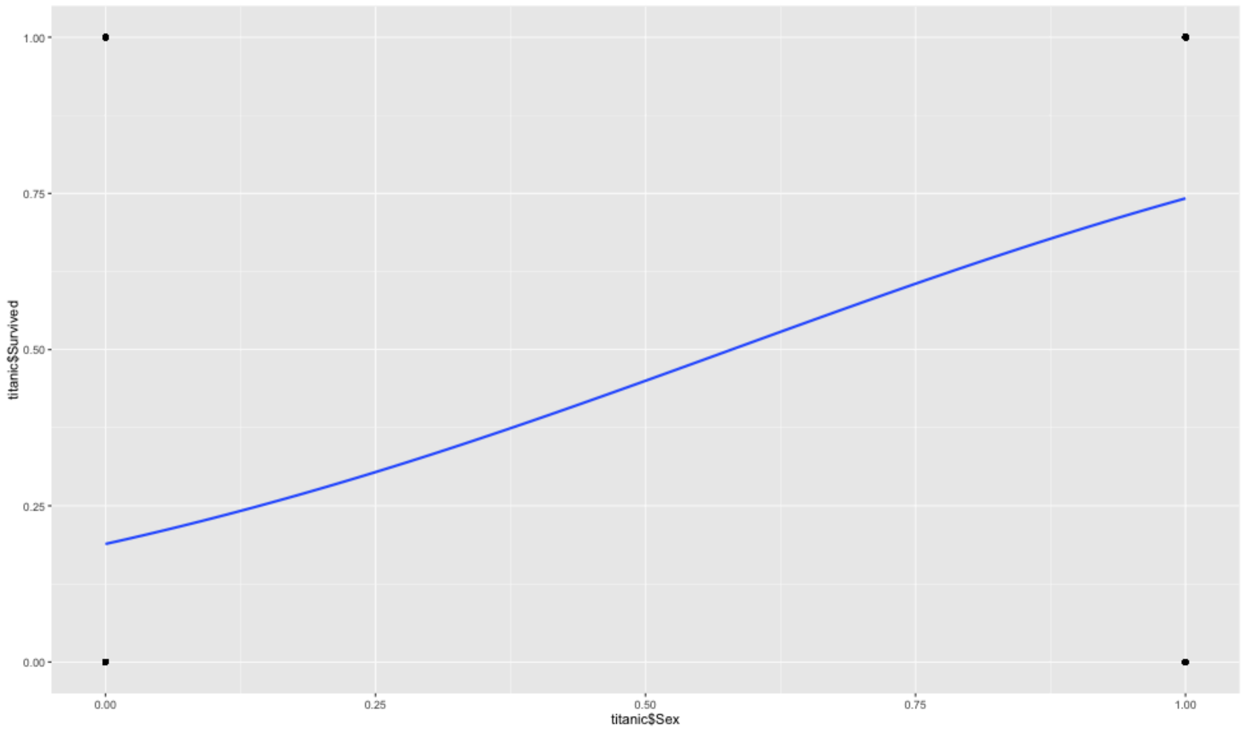
Scatter plot of Survival vs. Parch



Scatter plot of Survival vs. Pclass



Scatter plot of Survival vs Sex


titanic$Survived
titanic$Sex