**Predicting Titanic Survival**

Frannie K

# Question

Kaggle posted a competition to predict passenger survival on Titanic (<https://www.kaggle.com/c/titanic>). Dependent variable is the survival itself and a set of independent variables is a mix ore scale, ordinal and nominal variables, which are:

* Pclass – passenger class {1,2,3}
* Name – Name of passenger
* Sex – Sex of passenger
* Age – Age of passenger
* Sibsp - Number of siblings/spouses aboard
* Parch – number of parent/children aboard
* Ticket – Ticket number
* Fare – Passenger Fare
* Cabin – Cabin Number
* Embarked – Port of Embarkation {C,Q,S}

62% (549 out of total 891) have survived the ship wreckage. To determine what factors played a role in the survival, I have used exploratory analysis to reduce down the number of input variables then used discriminant analysis to predict membership of each observation.

# Data Treatment

Please refer to the Rmarkdown (Figure 1) for data cleansing steps undertaken. These include:

* Treatment of missing values
* Mending object class (e.g. factor, numeric, character.. etc)
* In SPSS, the data underwent another manual adjustment on type of data (scale, ordinal, nominal)



Figure 1

# Exploratory treatment

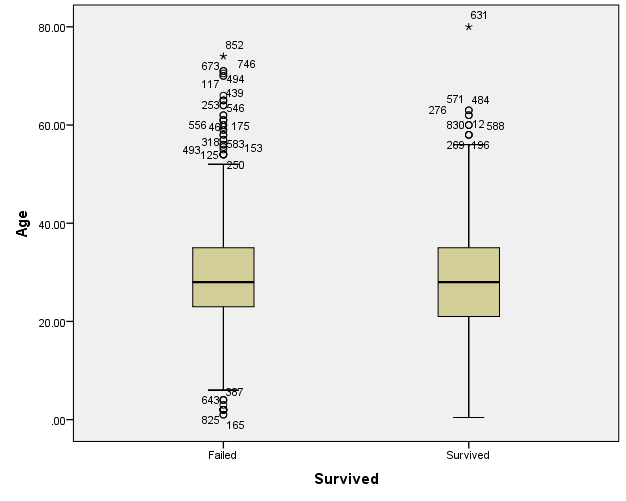
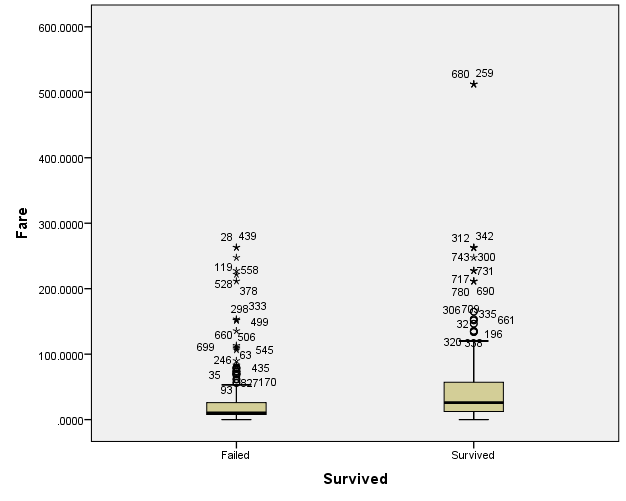
The independent variables are distinguished between1) parametric and 2) non-parametric (categorical) variables.

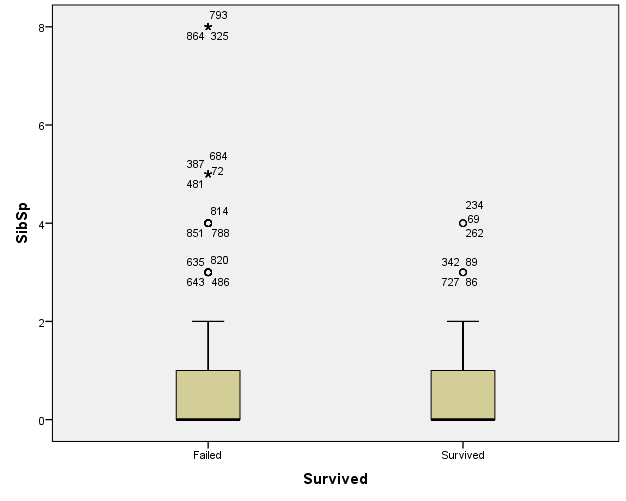
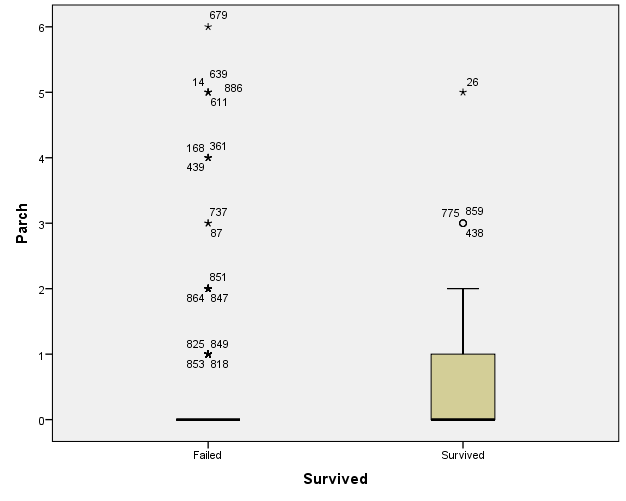
1. parametric variables consist of Age and Fare but as later pointed out, Fare is highly correlated with Pclass and therefore is substituted by categorical Pclass variable. For these, we analyse descriptive stats and ANOVA to test the significance of correlation with Survival rate. Table 1 displays the descriptive stats. It is noted the mean age of all passengers in training group is 29.4, the mean fare they paid across all passenger class is $32 and on average each passenger had less than 1 sibling/spouse and less than 1 parent/children associated.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | | | | | | | | |
|  | N | Range | Minimum | Maximum | Mean | Std. Deviation | Variance | Skewness | | Kurtosis | |
| Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Statistic | Std. Error | Statistic | Std. Error |
| Age | 891 | 79.58 | .42 | 80.00 | 29.3616 | 13.01970 | 169.512 | .510 | .082 | .994 | .164 |
| Fare | 891 | 512.3292 | .0000 | 512.3292 | 32.204208 | 49.6934286 | 2469.437 | 4.787 | .082 | 33.398 | .164 |
| SibSp | 891 | 8 | 0 | 8 | .52 | 1.103 | 1.216 | 3.695 | .082 | 17.880 | .164 |
| Parch | 891 | 6 | 0 | 6 | .38 | .806 | .650 | 2.749 | .082 | 9.778 | .164 |
| Valid N (listwise) | 891 |  |  |  |  |  |  |  |  |  |  |

Table 1

Before going into statistical significance test, we also look at whether the five-number summary and deviation metrics differ by Survival. This is shown through the boxplots below. From initial review, Age is more dispersed in Failed but the mean value and interquartile range (IQR) seem largely similar. Fare seems slightly higher in Survived category on average and by the data points in the above upper 1.5IQR. SibSp seems largely the same across the Survival status and Parch mean was the same across the category at zero but those with 1-2 parents or children survived more and anything more 2 tend to die more. Based on this, I decide to remove SibSp from the input variable set as it seems the variable has little impact on the Survival rate.

One-way ANOVA is used to measure the significance of correlation between the scale metrics and the Survival rate. Table 2 shows the difference in the variance of the variable when Survival rate is a factor. At 5% significance level, only Age variance is statistically equal across the category.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test of Homogeneity of Variances** | | | | |
|  | Levene Statistic | df1 | df2 | Sig. |
| Age | 2.616 | 1 | 889 | .106 |
| Fare | 82.918 | 1 | 889 | .000 |
| SibSp | 14.987 | 1 | 889 | .000 |
| Parch | 6.550 | 1 | 889 | .011 |

Table 2

Table 3 shows the ANOVA (or two sample t-test to be exact) result and in line with the graphic analysis, SibSp means of all groups are equal. So this confirms that using SibSp in predicting Survival will **not** result in accurate results.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVA** | | | | | | |
|  | | Sum of Squares | df | Mean Square | F | Sig. |
| Age | Between Groups | 635.654 | 1 | 635.654 | 3.762 | .053 |
| Within Groups | 150230.470 | 889 | 168.988 |  |  |
| Total | 150866.123 | 890 |  |  |  |
| Fare | Between Groups | 145508.888 | 1 | 145508.888 | 63.031 | .000 |
| Within Groups | 2052289.905 | 889 | 2308.538 |  |  |
| Total | 2197798.793 | 890 |  |  |  |
| SibSp | Between Groups | 1.350 | 1 | 1.350 | 1.111 | .292 |
| Within Groups | 1080.928 | 889 | 1.216 |  |  |
| Total | 1082.278 | 890 |  |  |  |
| Parch | Between Groups | 3.853 | 1 | 3.853 | 5.963 | .015 |
| Within Groups | 574.405 | 889 | .646 |  |  |
| Total | 578.258 | 890 |  |  |  |

Table 3

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Independent Samples Test** | | | | | | | | | | | |
|  | | Levene's Test for Equality of Variances | | t-test for Equality of Means | | | | | | | |
| F | Sig. | T | df | Sig. (2-tailed) | Mean Difference | Std. Error Difference | 95% Confidence Interval of the Difference | |
| Lower | Upper |
| Age | Equal variances assumed | 2.616 | .106 | 1.939 | 889 | .053 | 1.73680 | .89550 | -.02075 | 3.49435 |
| Equal variances not assumed |  |  | 1.897 | 671.154 | .058 | 1.73680 | .91574 | -.06126 | 3.53486 |
| Fare | Equal variances assumed | 82.918 | .000 | -7.939 | 889 | .000 | -26.2775207 | 3.3098484 | -32.7735484 | -19.7814930 |
| Equal variances not assumed |  |  | -6.839 | 436.702 | .000 | -26.2775207 | 3.8422488 | -33.8291189 | -18.7259225 |
| SibSp | Equal variances assumed | 14.987 | .000 | 1.054 | 889 | .292 | .080 | .076 | -.069 | .229 |
| Equal variances not assumed |  |  | 1.194 | 877.088 | .233 | .080 | .067 | -.051 | .212 |
| Parch | Equal variances assumed | 6.550 | .011 | -2.442 | 889 | .015 | -.135 | .055 | -.244 | -.027 |
| Equal variances not assumed |  |  | -2.479 | 758.568 | .013 | -.135 | .055 | -.242 | -.028 |

1. Nonparametric variables such as Pclass, Sex and Embarked are analysed through median and chi-squared test to measure the significance of correlation with Survival rate.

Median values for all observations are respectively, {3rd class, Male, “S”} but when we look at a subset of just ‘survived’, the median values are {2nd class, Female, “S”} which shows a deviation. Interestingly, this proves a hypothesis made in the preface that “*… some groups of people were more likely to survive than others, such as women, children, and the upper-class”,* I thought the fact that median survived class is 2nd and not 1st is due to the pure quantum (number) of 2nd class passengers outweighing the 1st class’. As I find out through bar charts, the number of survivors across the passenger class is largely the same with relatively small variation, hence explains why 2nd class became the median.

Chi-squared tests of each input variables against survival is as listed below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | Survived | | Total |
| Failed | Survived |
| Pclass | 1st class | Count | 80 | 136 | 216 |
| Expected Count | 133.1 | 82.9 | 216.0 |
| Std. Residual | -4.6 | 5.8 |  |
| 2nd class | Count | 97 | 87 | 184 |
| Expected Count | 113.4 | 70.6 | 184.0 |
| Std. Residual | -1.5 | 1.9 |  |
| 3rd class | Count | 372 | 119 | 491 |
| Expected Count | 302.5 | 188.5 | 491.0 |
| Std. Residual | 4.0 | -5.1 |  |
| Total | | Count | 549 | 342 | 891 |
| Expected Count | 549.0 | 342.0 | 891.0 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square Tests** | | | |
|  | Value | df | Asymp. Sig. (2-sided) |
| Pearson Chi-Square | 102.889a | 2 | .000 |
| Likelihood Ratio | 103.547 | 2 | .000 |
| Linear-by-Linear Association | 101.967 | 1 | .000 |
| N of Valid Cases | 891 |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 70.63. | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | Survived | | Total |
| Failed | Survived |
| Sex | male | Count | 468 | 109 | 577 |
| Expected Count | 355.5 | 221.5 | 577.0 |
| Std. Residual | 6.0 | -7.6 |  |
| female | Count | 81 | 233 | 314 |
| Expected Count | 193.5 | 120.5 | 314.0 |
| Std. Residual | -8.1 | 10.2 |  |
| Total | | Count | 549 | 342 | 891 |
| Expected Count | 549.0 | 342.0 | 891.0 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square Tests** | | | | | |
|  | Value | df | Asymp. Sig. (2-sided) | Exact Sig. (2-sided) | Exact Sig. (1-sided) |
| Pearson Chi-Square | 263.051a | 1 | .000 |  |  |
| Continuity Correctionb | 260.717 | 1 | .000 |  |  |
| Likelihood Ratio | 268.851 | 1 | .000 |  |  |
| Fisher's Exact Test |  |  |  | .000 | .000 |
| Linear-by-Linear Association | 262.755 | 1 | .000 |  |  |
| N of Valid Cases | 891 |  |  |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 120.53. | | | | | |
| b. Computed only for a 2x2 table | | | | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Crosstab** | | | | | |
|  | | | Survived | | Total |
| Failed | Survived |
| Embarked | C | Count | 75 | 93 | 168 |
| Expected Count | 103.5 | 64.5 | 168.0 |
| Std. Residual | -2.8 | 3.6 |  |
| Q | Count | 47 | 30 | 77 |
| Expected Count | 47.4 | 29.6 | 77.0 |
| Std. Residual | -.1 | .1 |  |
| S | Count | 427 | 219 | 646 |
| Expected Count | 398.0 | 248.0 | 646.0 |
| Std. Residual | 1.5 | -1.8 |  |
| Total | | Count | 549 | 342 | 891 |
| Expected Count | 549.0 | 342.0 | 891.0 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square Tests** | | | |
|  | Value | df | Asymp. Sig. (2-sided) |
| Pearson Chi-Square | 25.964a | 2 | .000 |
| Likelihood Ratio | 25.364 | 2 | .000 |
| Linear-by-Linear Association | 25.022 | 1 | .000 |
| N of Valid Cases | 891 |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 29.56. | | | |

|  |  |  |  |
| --- | --- | --- | --- |
| **Chi-Square Tests** | | | |
|  | Value | df | Asymp. Sig. (2-sided) |
| Pearson Chi-Square | 25.964a | 2 | .000 |
| Likelihood Ratio | 25.364 | 2 | .000 |
| Linear-by-Linear Association | 25.022 | 1 | .000 |
| N of Valid Cases | 891 |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 29.56. | | | |

On all cases, the Pearson Chi-Square test came within 5% significance level which cause rejection of null hypothesis that the variables are independent. Therefore, I will use all three variables in discriminant analysis.

Lastly, I confirm the high correlation between Fare and Pclass through ANOVA again.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **ANOVA** | | | | | |
| Fare | | | | | |
|  | Sum of Squares | df | Mean Square | F | Sig. |
| Between Groups | 776030.057 | 2 | 388015.028 | 242.344 | .000 |
| Within Groups | 1421768.736 | 888 | 1601.091 |  |  |
| Total | 2197798.793 | 890 |  |  |  |

# Prediction

Adding more variables are likely to result in better model fit. However, we know from various studies, over-fitting a model does not always result in better forecasting. Moreover, in real-world application, data can be scarce and input variables may be limited. Therefore, I used the significance of correlations between independent variables and dependent variable to filter only the key variables for modelling as follows:

* Pclass – passenger class {1,2,3}
* Sex – Sex of passenger
* Age – Age of passenger
* Parch – number of parent/children aboard
* Embarked – Port of Embarkation {C,Q,S}

To determine the test group’s survival, I’ve used discriminant analysis. The model can intake all types of input variables and the dependence variable is categorical as is the case here.

Here are the test specifics:  
Classification 🡪 Discriminant🡪 Grouping Variable = Survived (0 1) 🡪 Independents = as per above 🡪 Statistics: Function Coefficients: Fisher’s, Unstandardised 🡪 Leave-one out classification

During the process, standardization of coefficients allow ranking of importance of variables in explaining the dependent variable grouping.

The result is as follows:

|  |  |  |
| --- | --- | --- |
| **Test Results** | | |
| Box's M | | 129.472 |
| F | Approx. | 8.576 |
| df1 | 15 |
| df2 | 2107980.360 |
| Sig. | .000 |
| Tests null hypothesis of equal population covariance matrices. | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Wilks' Lambda** | | | | |
| Test of Function(s) | Wilks' Lambda | Chi-square | df | Sig. |
| 1 | .609 | 438.966 | 5 | .000 |

|  |  |
| --- | --- |
| **Standardized Canonical Discriminant Function Coefficients** | |
|  | Function |
| 1 |
| Pclass | .596 |
| Sex | -.857 |
| Age | .287 |
| Parch | .116 |
| Embarked | .129 |

|  |  |  |
| --- | --- | --- |
| **Classification Function Coefficients** | | |
|  | Survived | |
| Failed | Survived |
| Pclass | 5.470 | 4.224 |
| Sex | 7.115 | 10.626 |
| Age | .325 | .289 |
| Parch | .068 | -.168 |
| Embarked | 3.942 | 3.669 |
| (Constant) | -21.800 | -22.137 |
| Fisher's linear discriminant functions | | |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classification Resultsa,c** | | | | | |
|  |  | Survived | Predicted Group Membership | | Total |
|  |  | Failed | Survived |
| Original | Count | Failed | 443 | 106 | 549 |
| Survived | 88 | 254 | 342 |
| % | Failed | 80.7 | 19.3 | 100.0 |
| Survived | 25.7 | 74.3 | 100.0 |
| Cross-validatedb | Count | Failed | 442 | 107 | 549 |
| Survived | 88 | 254 | 342 |
| % | Failed | 80.5 | 19.5 | 100.0 |
| Survived | 25.7 | 74.3 | 100.0 |
| a. 78.2% of original grouped cases correctly classified. | | | | | |
| b. Cross validation is done only for those cases in the analysis. In cross validation, each case is classified by the functions derived from all cases other than that case. | | | | | |
| c. 78.1% of cross-validated grouped cases correctly classified. | | | | | |

For the full result, please refer to the attached SPSS output.



Finally, I apply the Fisher’s linear discriminant function to the test set. Of the two functions, whichever group function produces the larger value is where the observation or data-point sits in. The results are in the attached spreadsheet. According to Kaggle, this is 75.598% correct.

