Investigating the Titanic Passenger Data Set

Paul Simpson
30/10/2016

Exploring the Data

I looked at a data set comprising a sample of passengers who boarded the Titanic for its fateful maiden voyage (67.7% of all passengers that boarded the Titanic). Among the variables included, for most passengers: age; cabin class (a proxy for social class; *Pclass*); sex; variables indicating family groupings (number of siblings and spouses - *Sibsp*, number of parents and children – *Parch*, surname, and fare paid for the journey); and point of embarkation (Queensland; Southampton; or Cherbourg).

I will focus on the question: *What factors influenced the survival of passengers aboard the Titanic?* There many interesting questions that could be addressed, I motivate these as I explore the data (Table 1).

Starting with the age distributions of survivors and fatalities from the data set (Figure 1); while the distributions appear similar, there are differences in both their shape and means. The mean age of fatalities was 30.63±14.17 years (mean ± standard deviation) and the mean age of survivors was 28.34± 14.95 years (compare the hashed lines, Figure 1). Interestingly, while the ages of many of the passengers who perished were in the range of 17 to 34 years, there were more passengers older than 38 years who perished than survived (Figure 1).

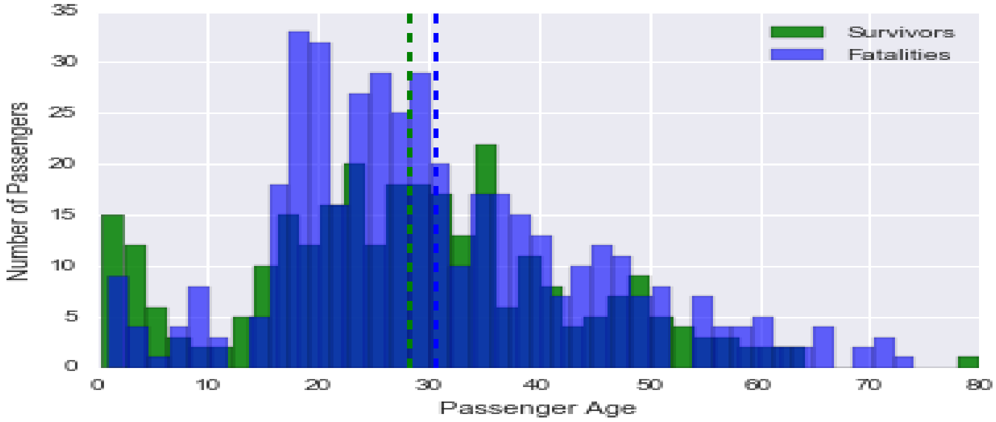

Figure 1, Age distribution of survivors (green) and fatalities (blue) in the Titanic passenger data set. Hashed lines indicate means.

Fortunately, the dataset contains more interesting variables which may be used to understand what factors influenced passenger survival – although keeping with the aphorism of *"women and children first"* I will continue to consider age as an important factor.

Adding a passenger's sex into the mix yields some further intrigue; while it is immediately apparent that more men perished than did women (Figure 2, compare the magenta bars in the upper and lower panels). Interestingly, the mean age of surviving men and women was more or less equal ($\mu_{men,surv}$=27.28±16.50 years vs. $\mu_{women,surv}$=28.84±14.18 years) whereas there was a bigger difference in age between men and women who perished ($\mu_{men,per}$=31.62±14.05 years vs. $\mu_{women,per}$=25.04±13.62 years).

It is worth pointing out that male passengers comprised 63% of the data set, given that most of the survivors were women (68% of survivors were female), this supports the idea that women (and probably children) were preferentially loaded into lifeboats. Also of note: seniors (those older than 60 years) were more likely to survive if they were women (100% chance) compared men (9.5% chance), however this is probably because there were not many old women included in the data set.

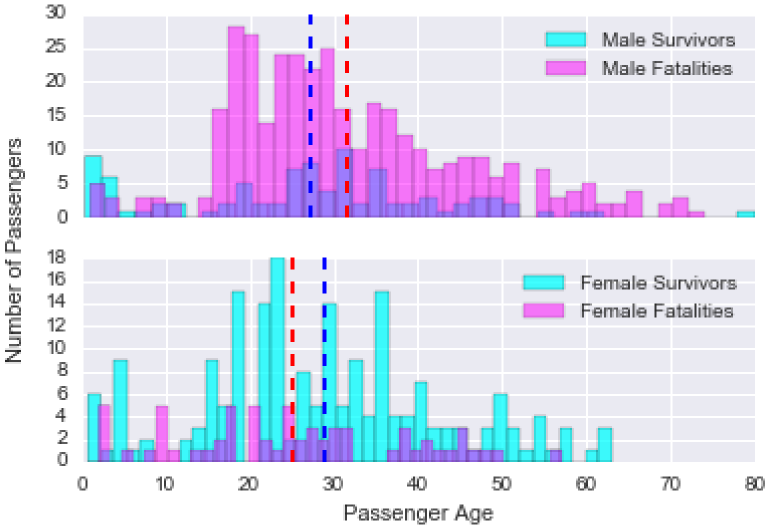

Figure 2, Age distribution of survivors (cyan) and fatalities (magenta) for male (upper) and female (lower) passengers in the Titanic passenger data set. Hashed lines denote the mean age of survivors and fatalities for each gender group.

Thankfully, for those who require more than an aphorism to understand why some survived a tragedy and others were not even sent to a pauper's grave, the data set also has information on a passenger's socioeconomic class. Here the picture builds on what we have seen; women were far more likely to survive than men. Tellingly however, a passenger's chances of surviving the disaster increases with higher level of socioeconomic class (black dots, Figure 3).

Breaking passengers down by sex and class, things get more interesting: middle and upper class women (in second or first class cabins, respectively) were twice as likely to survive as lower class women (in third class cabins); whereas upper class men were twice as likely to survive as lower and middle class men, but still less likely to survive than lower class women. The middle class is where gender differences in survival really stand out; while women were always more likely to survive, in the middle class they are almost 6 times more likely to survive than men. Overall, there appears to be a linear relationship between socioeconomic class and the probability of surviving (black dots, Figure 3).

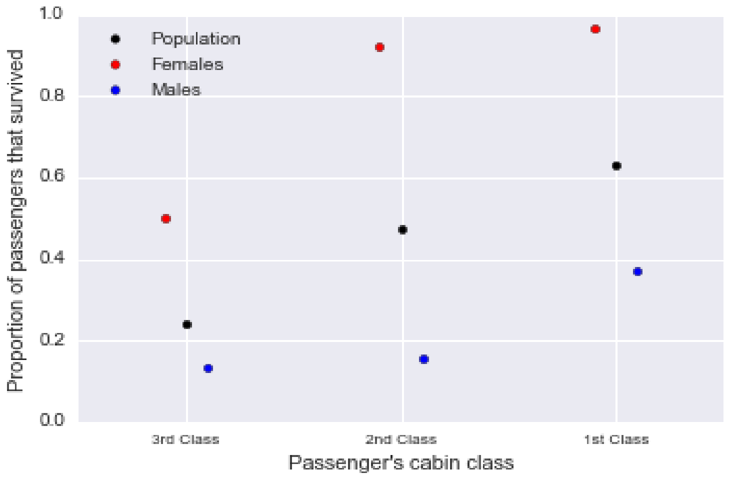

Figure 3, Proportion of all passengers (black dots), female passengers (red dots) and male passengers (blue dots) included in the Titanic data set that survived the disaster from each passenger class.

Given that there appears to be a better predictor (socioeconomic class) to explain a passenger's probability of survival, I want to see how this correlates with age (runner-up for best predictor; Figure 4).

Unsurprisingly, the mean age of passengers increased with socioeconomic class – passengers in lower classes were younger than those in higher classes (red dots, Figure 4). What stands out is that male passengers who survived, tended to be much younger than those who perished (μ_{surv} 16.02±19.55 years vs. μ_{surv}=33.37±12.15 years). Of note, the standard errors for the age of upper class women who perished is very large because there are so few of them, the same can be said for middle class men who survived.

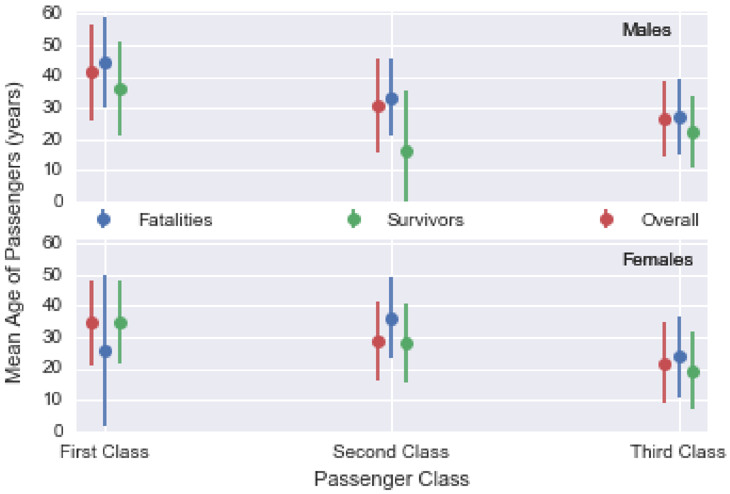

Figure 4, The mean age of male (upper) and female (lower) passengers aboard the Titanic. Blue corresponds with survivors, green corresponds with fatalities, and red corresponds with the age of survivors and fatalities; bars denote standard errors.

Table 1, Hypotheses and corresponding null hypotheses regarding the factors influencing the probability that a passenger would survive the disaster aboard the Titanic.

	Hypothesis	Null-hypothesis
1	The very old and very young were less likely to survive than working-age, adult passengers.	Passenger age had no effect on the probability of survival.
2	Adult females and children are more likely to survive than working age male, and elderly passengers.	The interaction between age and sex no effect on the probability of survival.
3	Passengers from higher socio-economic classes were more likely to survive than those from lower socio-economic classes.	Socio-economic class had no effect on the probability of survival.
4	Adult females and children of higher socio-economic classes are more likely to survive than those from lower socio-economic classes.	The interaction between socio-economic class, sex, and age had no effect on the probability of survival.

Cleaning the Data

To investigate this data set, I will start by cleaning the data. It is important to note that while the data are fairly complete, there are missing entries for three fields: passenger age; port of embarkation: and cabin (see Table 2). While one could infer the port of embarkation from a passenger's surname, this could easily incorrectly assign a port of embarkation to the passenger. The lack of entries for cabin number is

I will clean the data by creating three data sets: one where passengers with missing ages are removed; one where passengers with missing ports of embarkation are removed; and one where passengers with any missing data are removed. These data sets will each be used to answer different questions that arise from the data (see Table 3).

For analyses with categorical variables, non-integer variables, I created dummy variables for each of the levels of the categorical variable minus one using Pandas' "get_dummies" function. This was done for each data set containing any of the Sex, Embarked or Pclass variables. Additionally, an intercept column was created for each data set so that an intercept could be estimated in any regression analyses.

Table 2, A summary of the missing entries for various fields for passengers in the Titanic data set.

Field	Description	Number of Missing Values
Age	Age of passenger	177
Embarked	Port of Embarkation	2
Cabin Number	Location of the cabin on the ship	687

Table 3, Cleaned data sets used to test the hypotheses outlined in Table 1 pertaining to the factors that influence the survivorship of passengers aboard the Titanic.

Name	Fields that have been cleaned	Number of Passengers	Used to test hypothesis(/es)
titanic_df	None	891	3
clean_age_df	Age only	714	1
clean_embark_df	Embarkation only	889	5
clean_age_embark_df	Age and embarkation	712	2,4

Were the very old, and very young less likely to survive than working-age, adults?

I used a logistic regression of the binary variable, Survived on the continuous variable, Age to determine how age influences the probability of surviving the disaster. Logistic regression is used to determine the probability of an event, like a passenger surviving an event, based on one or more predictors, such as a passenger's age. Several assumptions must be met to use this approach (see Appendix); namely: that the data are independent; the response follows a binomial distribution; and the mean of the distribution is a function of the predictors (in this case, passenger age).

Table 4, Results of the generalized linear model regression of survival on passenger age using a binomial distribution.

Coefficient (β)	Estimate	Std Err	Z Score	Pr > |Z|	95% Confidence Interval		Odds Ratio
					Lower	Upper	
Age	-0.0110	0.005	-2.057	0.040	-0.021	-0.001	0.989096
Intercept	-0.0567	0.174	-0.327	0.744	-0.397	0.283	0.944855

Age had a marginally significant effect on the probability of surviving the disaster (β_{Age}=-2.057, P=0.04); in particular there is a 1.1% reduction in the probability of surviving the disaster for every year that a passenger has (based on the odds ratio, Table 4). The model could not estimate an intercept (β_{Int}=-0.327, P=0.744), so there is no estimate of mean age for survivors. Accordingly, a dummy coding scheme was used to make contrasts comparing the probability of survival among the groups of very old passengers (over 45 years), very young passengers (under 15 years), and working age passengers (between 15 and 45 years). Only one comparison was statistically significant: very young passengers were significantly more likely to survive than working age passengers ($C_{Adult:Child}$=0.2059, P<0.05).

In conclusion, younger passengers were more likely to survive than older passengers; however passenger age alone was not a good predictor of survival for this data set (based on a pseudo R^2=0.0045).

Were adult females, and children more likely to survive than adult men, and the elderly?

Factoring in sex; a passenger's sex had a significant effect on the probability of surviving the disaster – in particular, for every female passenger that survived, roughly 0.5 male passengers were saved (based on product of the odds ratio of the intercept and Sex_male coefficients; Table 5).

There was a significant negative interaction between a passenger's age, and gender on the probability of survival ($\beta_{Age:Sex}$=-0.0411, P<0.005). To wit, older male passengers were less likely to survive than younger male passengers. While female passengers were twice as likely to survive, on average, than male passengers, a female passenger's age did not affect her probability of surviving (Table 5).

Investigating the interaction between age and sex, I grouped passengers according to age and sex classes: females between 15 and 45 years (adult_fem); males between 15 and 45 years (adult_male); all passengers under 15 years (child); and all passengers above 45 years (senior). I then performed contrasts using a dummy coding scheme, where the estimated probability of survival for each group was compared to that for adult females (Table 6). Accordingly, adult females had a significantly higher probability of surviving than all other groups (see Pr>|t| column in Table 6), including children (C_{Child}=-0.1703, P<0.005).

Overall, this model fared better than the previous model; it explained 51 times more of the variation in the data than the previous model which considered only age (pseudo R^2=0.2324).

Table 5, Results of the generalized linear model regression of survival on passenger age and sex using a binomial distribution.

Coefficient (β)	Estimate	Std Err	Z Score	Pr > |Z|	95% Confidence Interval		Odds Ratio
					Lower	Upper	
Intercept	0.5938	0.310	1.913	0.056	-0.014	1.202	1.810858
Age	0.0197	0.011	1.863	0.062	-0.001	0.040	1.019897
Sex_male	-1.3178	0.408	-3.226	0.001	-2.118	-0.517	0.267737
Age : Sex_Male	-0.0411	0.014	-3.034	0.002	-0.068	-0.015	0.959715

Table 6, Dummy-coded contrasts of the Age by Sex interaction from the full, generalized linear binomial regression model of survival on passenger age and sex.

Contrast	Estimate	Std Err	t	Pr > |t|	95% Confidence Interval	
					Lower	Upper
Adult Females vs. Males	-0.5842	0.039	-15.156	3.67 e-45	-0.660	-0.508
Adult Females vs. Children	-0.1703	0.056	-3.047	0.002	-0.280	-0.061
Adult Females vs. Seniors	-0.3917	0.052	-7.535	1.50 e-13	-0.494	-0.290

Were passengers from higher socio-economic classes more likely to survive than those from lower socio-economic classes?

A logistic regression of survival probability on passenger class indicates that passenger class had a significant effect on the probability of surviving (β_{Class}=-0.8501, P<0.001; Table 7). If one considers the odds ratios, each level below first class corresponds with a reduction in survival probability by 57%. In short, those who paid for the best cabins, had the best chances of surviving.

Table 7, Results of the generalized linear binomial regression model of survival on passenger class.

Coefficient (β)	Estimate	Std Err	Z Score	Pr > |Z|	95% Confidence Interval		Odds Ratio
					Lower	Upper	
Intercept	1.4468	0.087	6.975	3.06 e-12	1.040	1.853	4.249450
Pclass	-0.8501	0.207	-9.755	1.75 e-22	-1.021	-0.679	0.427369

When one includes a passenger's sex in this analysis, sex, passenger class and the interaction of these two variables have a significant effect on the probability of survival (β_{Sex}=-6.0503, P<0.001; β_{Class}=-2.0674, P<0.001; $\beta_{Sex:Class}$=1.4306, P<0.001; Table 8). For all passenger classes, females realized a significantly higher survival probability than did males.

Table 8, Results of the generalized linear binomial regression model of survival on passenger class and sex.

Coefficient (β)	Estimate	Std Err	Z Score	Pr > |Z|	95% Confidence Interval		Odds Ratio
					Lower	Upper	
Intercept	6.1052	0.852	7.163	7.87 e-13	4.435	7.776	448.19943
Pclass	-2.0674	0.311	-6.655	2.83 e-11	-2.676	-1.459	0.126510
Sex_male	-6.0503	0.908	-6.664	2.66 e-11	-7.830	-4.271	0.002357
Pclass:Sex_male	1.4306	0.340	4.207	2.59 e-05	0.764	2.097	4.181031

Were adult females and children of higher socio-economic classes more likely to survive than those from lower socio-economic classes?

Putting everything together, I ran a logistic regression model of survival on passenger class, age and sex (Table 9). Passenger class was the only significant effect in the model (β_{Class}=-1.6801, P<0.05) where each decrease in class resulted in a 91% decrease in the probability of survival (based on Odds ratio; Table 9).

Table 9, Results of the full generalized linear binomial regression model of survival on passenger age, sex and socioeconomic class.

Coefficient (β)	Estimate	Std Err	Z Score	Pr > |Z|	95% Confidence Interval		Odds Ratio
					Lower	Upper	
Intercept	5.5897	2.211	2.529	0.011	1.257	9.922	267.647165
Age	0.0274	0.067	0.410	0.682	-0.103	0.158	1.027728
Sex_male	-3.9937	2.372	-1.683	0.092	-8.463	0.656	0.018431
Pclass	-1.6801	0.782	-2.149	0.032	-3.212	-0.148	0.186355
Age:Pclass	-0.0190	0.024	-0.775	0.438	-0.067	0.029	0.981212
Age:Sex_male	-0.0549	0.070	-0.780	0.435	-0.193	0.083	0.946535
Sex_male: Pclass	0.9712	0.860	1.129	0.259	-0.714	2.657	2.641056
Sex_male:	0.0093	0.027	0.346	0.730	-0.043	0.062	1.009331

Pclass:Age

A subsequent step-wise model simplification was done, removing statistically insignificant, higher order interaction terms and re-running successively simplified models, in an attempt to identify the best-fitting model (Table 10). To select a best model, a Bayesian information criteria score (BIC), an estimate of model fit based on log-likelihood, model parameters and number of data points, was computed for each. The difference in BIC between each model and the full model (ΔBIC) was calculated, and the model with the greatest ΔBIC (smallest BIC) was selected as the best model.

Here, the model with the fewest parameters (Reduced 3; Table 10) is best, it explained 35% of the variation in the data. Males were far less likely to survive than females regardless of their class ($\beta_{Class:Sex}$=1.4796, P<<0.001; Figure 5). Interestingly, when considering a passenger's class, sex, and age; age was the only factor that did not significantly interact with any of the other factors in the full model ($\beta_{Age:Class:Sex}$=0.0093, P>0.5; $\beta_{Age:Sex}$=-0.05493, P>0.4; $\beta_{Age:Class}$=-0.0190, P>0.4; Table 9). This is surprising as age significantly interacted with sex (Figure 3), resulting in adult women and children realizing a higher probability of survival than adult men or seniors. Moreover, age appeared to interact with passenger class (Figure 4), only not in a way that affected survival.

Table 10, The Generalized linear regression models of survival on passenger age, sex and socio economic class using a binomial distribution and subsequent model refinements.

Model	Model Df	Residual Df	Log-likelihood	BIC Score	ΔBIC	Pseudo R^2
Full	7	706	-310.43	673.4200	0.0000	0.3563
Reduced 1	6	707	-310.49	666.9702	6.4498	0.3562
Reduced 2	5	708	-311.12	661.6596	11.7604	0.3549
Reduced 3	4	709	-311.99	656.8407	16.5793	0.3531

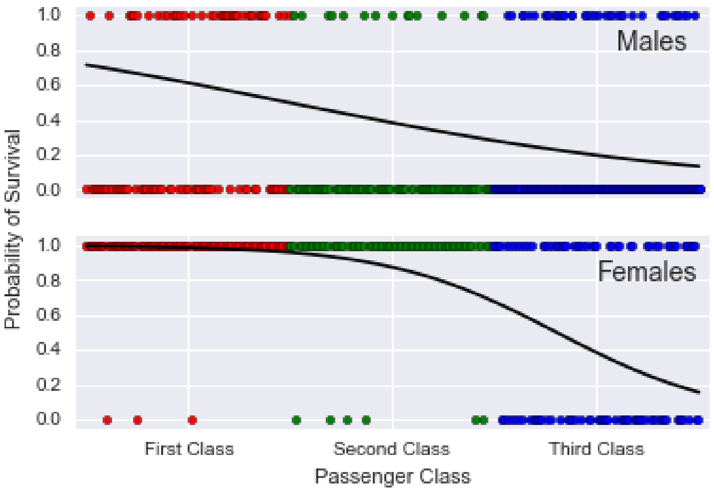

Figure 5, The probability of survival of male (upper) and female (lower) passengers in first (red), second (green) and third (blue) class. Dots denote individual passengers and whether they survived (1.0) or perished (0.0); the black line denotes the logit regression line.

Conclusions

I was able to draw some tidy conclusions from this data set; there were problems with the data based on how the sample was collected. Importantly, the data reflected a sample (67.7%) of the total number of passengers aboard the ship; however when conducting analyses of age many passengers did not have entries for this variable (19.9%) meaning that any analysis pertaining to questions of age, was only able to use roughly 80% of the data (or 56% of the data for the entire population).

Other factors that could have influenced survival, and would have been worth looking at were: a passenger's cabin location (relative to sea-level or the nearest life vessels); and whether a passenger was traveling with his or her family, and what the composition of this family was (perhaps larger families had greater mortality because lone mothers would only be able to gather so many of her children amidst the chaos of the capsizing of the vessel).

While the old adage suggests that women and children should be loaded into lifeboats first; on the Titanic, it was more likely for adult women to be saved than any other group of passengers, including children. If the Titanic's captain wanted to know what would most influence the demographic composition of the survivors, he might have proposed assign ranks in the adage – *adult women first, then perhaps children*.

Literature Cited

Resource on logistic regression in Python, using the statsmodels package:

- http://blog.yhat.com/posts/logistic-regression-and-python.html

Assumptions of generalized linear models come from:

- https://onlinecourses.science.psu.edu/stat504/node/216
- https://www.r-bloggers.com/checking-glm-model-assumptions-in-r/

How to compute contrasts and backwards differences from generalized linear models in Python:

- http://statsmodels.sourceforge.net/devel/examples/notebooks/generated/contrasts.html

Bayesian information criteria:

Kadane, J.B. and N.A. Lazar (2004). Methods and criteria for model selection. *Journal of the American Statistical Association* **99:279-290**

Appendices

The data set

https://www.kaggle.com/c/titanic/data

Assumptions of Generalized Linear Models

1. Data are independent
2. Residuals follow the correct distribution from the exponential family
3. The variance structure of the data is correctly specified
4. There is a linear relationship between the response and linear predictor

Assumptions of Logistic Regression Model

1. The dependent variable (survivorship) is assumed to come from a binomial distribution.
2. The mean, μ of this distribution is a function of the independent variables $\mathbf{x}$ and some combination of unknown parameters $\boldsymbol{\beta}$ via a logit link function.