Categorical Data Analysis
of
Survival in the Titanic Voyage

Statistics 138

Project I

Dr. Rahman Azari

Soloman Wong

SID: 999042564

Audrey Chu

SID: 999138148

Meixin Deng

SID: 913359547

Hiu Man Lam

SID: 913358632

Abstract

To analyze our chosen categorical dataset through examining different models including log-linear. The chosen ‘Titanic’ dataset is from R and provides information on the fate of passengers on the fatal maiden voyage of the ocean liner, Titanic, summarized according to gender, age, and survival. The purpose of the project is to determine the relationship of these factors. Specifically, we expect to find higher survival rates for females and children. Our findings may be applied to other voyages as well.

I. Introduction

This study analyzes the dataset of passengers from the April 15 1912 Titanic voyage containing three categorical variables: sex, age, and survival. According to Alan Agresti’s *Categorical Data Analysis*, a categorical variable has a measurement scale consisting of a set of categories. In our study, sex, age, and survival are categorical variables that create a three-dimensional array. The variables each have two levels as follows:

Sex:	Male, Female
Age:	Child, Adult
Survival:	No, Yes

It is important to analyze this data to find a possible patterns in the likelihood of survival among 2201 passengers and crewmembers of different genders and ages. An initial summary of the data shows that there is a significant amount of more males at 1731 than females at 470. Similarly, there is also a many more adults (2092) than children (109). 1490 passenger and crew members did not survive while 711 did, meaning 67.70% of passengers died. This is important to consider when looking the odds ratio and examining different models to fit our data.

II. Materials and Methodology

Conditional and Marginal Odds Ratio

The data used in this report describes the outcome of the Titanic sinking in 1912 and contains 2201 observations on three variables. The outcomes are shown in the 2 X 2 X 2 contingency table below:

		Survived	
Sex	Age	NO	Yes
Female	Adult	109	316
	Child	17	28

Male	Adult	1329	338
	Child	35	29

A logistic regression allows us to establish a relationship among binary outcome variables through the usage of multiple odds ratios. Since this is a joint distribution, we calculated the conditional odds ratio, which is defined to be

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{P(Y=1|X=1)/P(Y=2|X=1)}{P(Y=1|X=2)/P(Y=2|X=2)}$$
$$= \frac{P(X=1|Y=1)/P(X=2|Y=1)}{P(X=1|Y=2)/P(X=2|Y=2)}.$$

In our case, we calculated the two conditional odds ratio for adult female survival and adult male survival. Thus θ(Adult female no survival) = 109*28/(17*316) = 0.568131 = 56.81%. This means that the odds for an adult female not surviving is 56.8% times that for a child female. In contrast, θ(Adult male no survival) = 1329*29/(338*35) =3.257904, meaning the odds for an adult male not surviving is 325.79% times that for a child male.

We then looked at the marginal odds ratio between adults and children to find that θ(adult no survival) is 2.41% = (1438*57)/(654*52) as shown in the two by two contingency table in the appendix. Here, the odds for an adult not surviving is 3.41 times that for a child not surviving. And the likelihood ratio of a female vs a male surviving is ((316/109) / (1 + (316/109))) / ((338/1329) / (1 + (338/1329))) = 3.67. This means that females were about 3.67 times as likely to survive. Since the chance of survival was apparently related to Sex and Age, we will be using logistic regression to look for the relationship among the variables and try to decide which of the potential effects are useful to build a model predicting a person's likelihood of surviving.

Log-linear Model

In our consideration of a 2 X 2 X 2 contingency table that cross-classifies a multinomial sample of 2201 subjects on three categorical variables.
With the utilization of the loglm() function in R, we fit a log-linear model for homogenous association and survival.

III. Results

For our log-linear models, the degrees of freedom is the number of cell country subtracted by the number of model parameters. The total number of nonredundant parameters is thus I*J*K = 2*2*2 = 8, which is also the total number of cell counts. G^2 and X^2 values allow us to test whether a model holds by comparing the fitted cell values to observed counts. The Akaike's Information Criterion (AIC) judges a model by how closely it's fitted values tend to be in relation to the expected values. For the Titanic data, the table below shows our results of testing the fit for various models.

Log-linear Model	df	Likelihood Ratio (G^2)	p-value	Pearson	p-value	AIC	Quadratic (R^2)	R^2 (adj)
Survived, Sex, Age (Mutual Independence Model)	4	479.96	0	505.085	0	539.2	0.5513	0.2147
Survived & Sex, Age (2-way Interaction)	3	45.50	<0.0001	50.23	<0.0001	106.8	0.684	0.2628
Survived & Age, Sex (2-way Interaction)	3	460.40	0	475.74	0	521.7	0.6073	0.0836
Sex & Age, Survived (2-way Interaction)	3	456.68	0	477.28	0	517.9	0.6858	0.2669
Survived & Sex, Survived & Age (Conditional Independence Model)	2	25.935	<0.0001	40.896	<0.0001	89.22	0.74	0.901
Survived & Sex, Sex & Age (Conditional Independence Model)	2	22.212	<0.0001	26.176	<0.0001	85.49	0.8186	0.365
Survived & Age, Sex & Age (Conditional Independence Model)	2	437.12	0	463.89	0	500.4	0.7418	0.0962
Survived & Sex, Survived & Age, Sex & Age (Homogeneous Association Model)	1	16.319	<0.0001	18.806	<0.0001	81.60	0.8746	0.1219
Survived & Sex & Age (Saturated Model)	0	0	1	0	1	67.28	1	NA

Our final results showed that all tested models have a very small p-value, all below α = .05. We can conclude that none of our models so far are a good fit of the data. According to AIC, the homogenous association model that includes all two-way effects provides the best fit, when compared to the other current models. Additionally, the R^2 from the quadratic model supports that the homogenous association model is the best fit. R^2 - adjusted suggests that the 'Survived & Sex' and 'Survived & Age' conditional associations provide a better fit. The statistics comparing the models also provide us with strong evidence of ‘Survived & Sex’ and ‘Sex & Age’ conditional associations. However, it is important to know that all models tests lack the ‘Survived & Sex’ term and thus are a poor fit due to their large values of likelihood ratio, Pearson chi-squared, and AIC.

95% Confidence interval:

Wald test:	log Odds Ratio CI	Odds Ratio CI	p-value
Survived & Sex	[-2.529, -2.059]	[0.080,0.127]	<0.0001
Survived & Age	[0.110, 1.002]	[1.116, 2.724]	0.01449
Sex & Age	[-1.184, -0.272]	[0.306, 0.762]	0.00174

The 95% confidence intervals for the log odds ratio and odds ratio show that the value of 1 is not within any of the intervals. Since the value of 1 is not within the interval, we reject the null hypothesis to conclude that all three factors have associations between them. Thus, the categorical variables age and gender are statistically significant in predicting the variable of survival.

IV. Conclusion

Overall, we find the both sex and age contribute to the survival of the passengers and crewmembers through multiple likelihood tests and log-linear models. We conclude that a female’s survival rate is a lot higher than male, which corroborates our initial hypothesis. This final conclusion can be supported through many reasons, such as the notion that adult males believe that it is their responsibility to protect their wives or children. Also like supported earlier, the proportion of males is more than half of the passengers and so we can infer that many of these men were crewmembers and thus may have had job duties and other authority that would affect their ability to survive.

Next Steps & Possible Improvements

As our p-values for tested models did not pass the $\alpha = .05$ threshold, it would be reasonable to test a quadratic and possible cubic model. Our chosen models did not happen to fit the dataset very well; however, by adding other variables we may be able to increase the accuracy of our model estimation (as shown in the code of our appendix). Next steps of this report can be to consider other categorical variables. For the purpose of this project and due to our current knowledge of categorical data analysis, we kept to a three-way analysis. A four-way categorical analysis would be interesting in that it may provide further insight to the relationship of passengers and survival rate on the Titanic ship. Economic class, for example, can suggest location of the passenger--which in turn can affect the ability to survive.