Comparison of Logistic Regression Model and Log Linear Model applied to Titanic Survival data

S. Park

Department of Applied Statistics

1. Introduction

Titanic Survival data provides information about passengers who survived from Titanic. The data set provides information on the fate of passengers on the fatal voyage of the ocean liner "Titanic". It contains variables including economic status, sex, age and survival status of passengers. Using the information about characteristics of the passengers, I would like to create statistical models that can predict the survival status. I would like to preprocess the data and apply logistic regression model and then log linear model. I would compare the performance of logistic regression models so that I could select the most appropriate model. Also, I would like to illustrate the mosaic plots of loglinear models to check the degrees of association between variables.

2. Exploratory Data Analysis

2.1 Variable Explanation

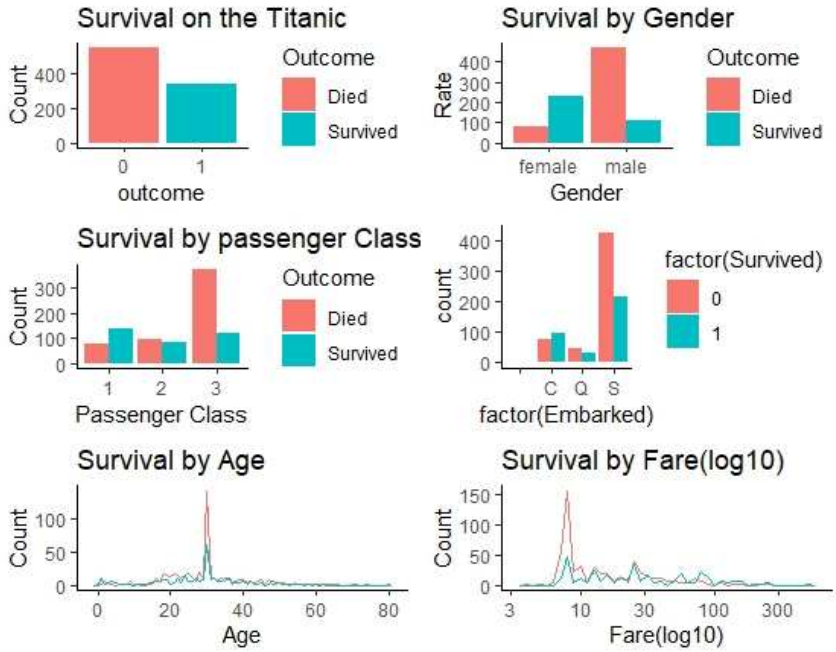
Explanation for variables is in Table 1. There are 12 variables in the "titanic::titanic_train data". Also, *Sex* and *Embarked* are saved in character type of values. So I converted those variables into factor type. Dummy variables used in conversion are explained in Table 1.

Table 1
Variable explanation

Variable name	Variable information
PassengerId	Passenger Id
survival	Survival (0 = No; 1 = Yes)
pclass	Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
name	Name
sex	Sex (0 = female; 1 = male)
age	Age
sibsp	Number of Siblings/Spouses Aboard
parch	Number of Parents/Children Aboard
ticket	Ticket Number
fare	Passenger Fare
cabin	Cabin
embarked	Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton) (C: Q=0, S=0; Q: Q=1, S=0; S: Q=0, S=1)

2.2 Survival rate by characteristics of passengers

Figure 1
Survival by variables



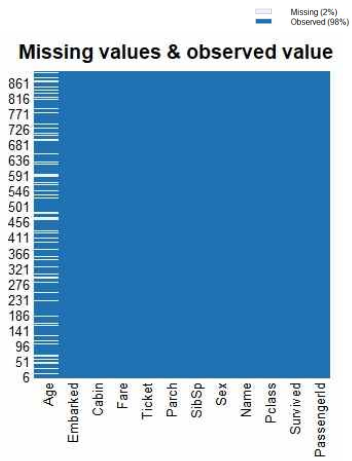
From Figure 1, I could see that most of the passengers were dead. Female passengers were more likely to survive than male passengers. Also, there were more opportunities for higher-class passengers to survive. The passengers in District C had a slightly higher survival rate than those on board in other areas. The passengers under age 16 tend to have a high survival rate, while other age groups had a high risk of death. Passengers who paid high fares generally had a higher survival rate than those who paid relatively less.

2.3 Variable significance and imputation

Table 2
Variable configuration

Variable name	Number of missing values	Number of unique values
PassengerId	0	891
survival	0	2
pclass	0	3
name	0	891
sex	0	2
age	177	89
sibsp	0	7
parch	0	7
ticket	0	681
fare	0	248
cabin	0	148
embarked	0	4

Table 3
Visualization of missing values



From Table 2, there are 177 missing values only in *Age*. I checked the virtual map of missing *Age* values in Table 3. To proceed to modeling steps, I imputed the missing values. I decided to impute the missing values of *Age* with the mean of observed *Age* values.

3. Models Interpretation and Comparison

3.1 Logistic regression model

The first model that I applied on the data is Logistic regression model. In Table 4, I checked the P-value of variables and select the significant ones in the model. I could see that the significant variables under significance level 0.05 are *Pclass, Sexmale, Age,* and *SibSp.*

Table 4
Logisitic regression full model

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	17.572941	610.227586	0.029	0.9770	
Pclass	-1.100058	0.143529	-7.664	1.80e-14	***
Sexmale	-2.718695	0.200783	-13.540	< 2e-16	***
Age	-0.039901	0.007854	-5.080	3.77e-07	***
SibSp	-0.325777	0.109384	-2.978	0.0029	**
Parch	-0.092602	0.118708	-0.780	0.4353	
Fare	0.001918	0.002376	0.807	0.4194	
EmbarkedC	-12.287753	610.227400	-0.020	0.9839	
EmbarkedQ	-12.321829	610.227451	-0.020	0.9839	
EmbarkedS	-12.706570	610.227384	-0.021	0.9834	

Table 5
ANOVA table of full logistic regression model

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)	
NULL			890	1186.66		
Pclass	1	102.254	889	1084.40	< 2.2e-16	***
Sex	1	257.206	888	827.20	< 2.2e-16	***
Age	1	21.869	887	805.33	2.918e-06	***
SibSp	1	14.481	886	790.84	0.0001415	***
Parch	1	0.513	885	790.33	0.4738915	
Fare	1	1.606	884	788.73	0.2051114	
Embarked	3	4.538	881	784.19	0.2089609	

In Table 5, adding *Pclass, Sex, SibSp* and *Age* in the model significantly reduced the residual deviance. Also, I tried Criterion-based method that forwardly extract variables using the standard of AIC and BIC. From Table 6, the model which obtained the smallest AIC and BIC was same. It was the model that includes *Pclass, Sex, SibSp* and *Age*. So, I decided to select *Pclass, Sex*, *Sibsp* and *Age* in the final logistic regression model.

Table 6
Criterion-based method

Model	AIC	BIC
Survived ~ 1 (only intercept)	1188.66	1193.45
Survived ~ Sex	921.8	931.39
Survived ~ Sex + Pclass	833.2	847.57
Survived ~ Sex + Pclass + Age	813.33	832.5
Survived ~ Sex + Pclass + Age + SibSp	**800.84**	**824.81**

The final model chosen is equation (1) below. I applied the estimates of coefficients so that I can interpret the model. In the odds, p_x means the probability of *Survived.*

$$\ln\frac{p_x}{1-p_x} = 5.1920 - 1.1724\,Pclass - 2.7398\,Sexmale - 0.0398\,Age - 0.3578\,SibSp \quad (1)$$

To interpret the model (1), A unit increase in *Pclass* reduced the log odds by 1.1724. As *Sexmale* is a dummy variable, being *male* reduced the log odds of *Survived* by 2.7398. Also, a unit increase in *Age* reduced the log odds by 0.0398. A unit increase in the number of *Siblings/Spouses* aboard reduced the log odds by 0.3578.

3.2 Loglinear model

The second model is loglinear model. I used four variables *Survived, Pclass, Sex,* and *Age. Pclass i*s justified as *Class* in this section. I fitted three models as follows.

$$\log m_{ijkl} = \mu + \lambda_{ijk}^{ClassAgeSex} + \lambda_{l}^{Survived} \tag{2}$$

$$\log m_{ijkl} = \mu + \lambda_{ijk}^{ClassAgeSex} + \lambda_{il}^{ClassSurvived} + \lambda_{jl}^{AgeSurvived} + \lambda_{kl}^{SexSurvived} \tag{3}$$

$$\log m_{ijkl} = \lambda_{ijk}^{ClassAgeSex} + \lambda_{il}^{ClassSurvived} + \lambda_{jl}^{AgeSurvived} + \lambda_{kl}^{SexSurvived} + \lambda_{jkl}^{AgeSexSurvived} \tag{4}$$

Interpretation for model (2) is that there is strong association between non-survived crew and third class passengers. From Figure2, high-class passengers were more likely to be saved. Also, those under age 16 were more likely to survive in every class. In the first class, female passengers seemed to survive more than male passengers.
Interpretation for model (3) is that it considers possible association between *gender, age,* and *class with survival.* Figure 3 illustrates that the remaining association decreased greatly.

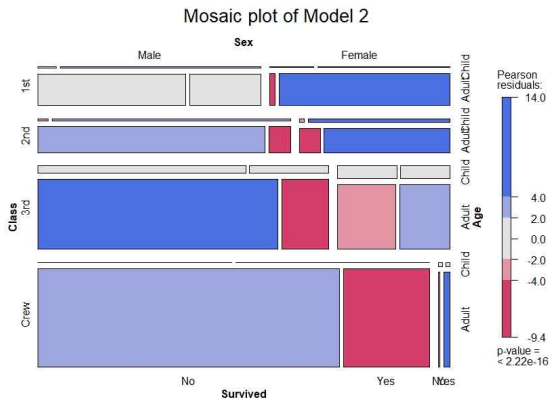


Figure 2

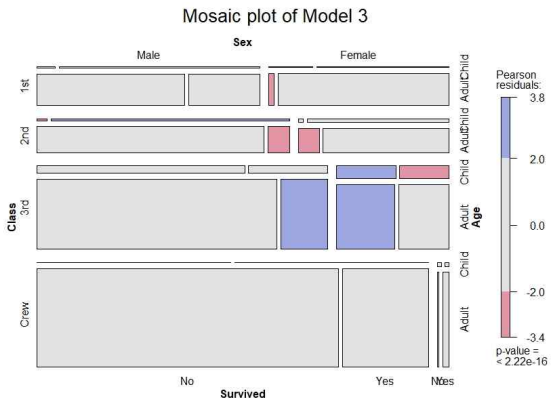


Figure 3

The model (4) considered possible association between *Age, Sex* with *Survived.* From Figure 4, I could see that the remaining association decreased more than model (3).

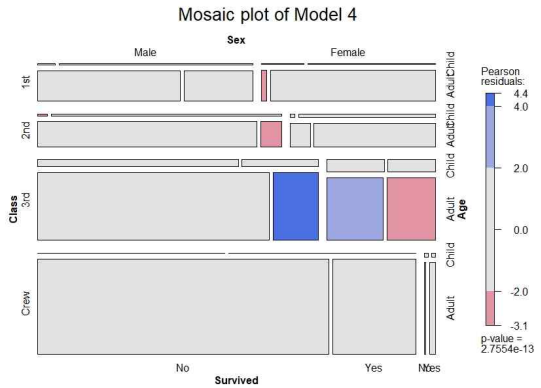


Figure 4

From Table 7 below, I concluded that the overall association decreased greatly after adding different association combinations to the model.

Table 7
ANOVA test on three Loglinear models

LR tests for hierarchical log-linear models

Model 1:
~(Class * Age * Sex) + Survived
Model 2:
~(Class * Age * Sex) + Survived * (Class + Age + Sex)
Model 3:
~(Class * Age * Sex) + Survived * (Class + Age * Sex)

	Deviance	df	Delta(Dev)	Delta(df)	P(> Delta(Dev)
Model 1	659.31623	15			
Model 2	105.55308	10	553.76314	5	0e+00
Model 3	85.51508	9	20.03800	1	1e-05
Saturated	0.00000	0	85.51508	9	0e+00

4. Summary and Conclusion

After I interpreted the results from logistic regression model and loglinear model, I could obtain information about the association between variables and some factors which affected the survival status of the passengers. There were strong association between Class, Age, and Sex with Survival status. Also, logistic regression model and loglinear model both showed similar tendency about the passengers who survived. First, the passengers in higher class were more likely to survive than the lower class passengers. Second, when age is low, that is, children were more likely to be saved than adults. Lastly, female passengers got more chances to survive than male passengers. I could infer that the cultural spirit of protecting women and children first affected the number of survivors.

5. References

Jeffrey S. (2020). The sinking of the Titanic.
David Julian. (2016), Building Machine Learning Systems with Python.
RMichy A. (2015), How to perform a Logistic Regression in R. *DataScience+*