A mathematical essay on Logistic Regression

Assignment 2 : EE4708 Data Analytics Laboratory

Madhur Jindal
Inter Disciplinary Dual Degree Program in Data Science
Indian Institute of Technology (IIT) Madras
Chennai, India
me18b059@smail.iitm.ac.in

***Abstract*—This document is a mathematical essay on Logistic Regression wherein it is applied on a dataset containing different attributes of the passengers of the famous 'unsinkable' RMS Titanic, which sank on its maiden voyage after colliding with an iceberg. The aim for the assignment is to analyze the data for the total of 2224 passengers, out of which 1502 died due to unsufficient presence of lifeboats, and hypothesize whether some groups of people (divided by name, age, gender, socio-economic class, etc.) were more likely to survive and discuss any insights and observations.**

I. Introduction

Classification is the process of learning a mapping from related input features to discrete data classes or categories. Logistic regression, despite its name, is a technique that can be used to handle binary classification problems. Logistic regression is a process of modelling the probability of a discrete outcome given an input variable. Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. The probability is computed by taking the logistic of a linear regression function. The binary logistic model has a dependent variable with two possible values, wherein the corresponding probability of the value labeled '1' can vary between 0 and 1, hence the function that converts the log-odds to probability is the logistic function. The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification, though it can be used to make a classifier, for instance by choosing a cutoff value and classifying the outputs with probability greater than the cutoff as one class, below the cutoff as other.

II. Logistic Regression

A. Different Approaches

1) **Generative model approach**
$(I) Modelp(x, C_k) = p(x|C_k)p(C_k)$
(I) Use Bayes' theorem $p(C_k|x) = \frac{p(x|C_k)p(C_k)}{p(x)}$
(D) Apply optimal decision criteria

2) **Discriminative model approach**
$(I) Modelp(C_k|x) directly$
(D) Apply optimal decision criteria

3) **Direct regression approach**
(D) Learn a function that maps each x to a class label directly from training.

Identify applicable funding agency here. If none, delete this.

B. Model structure

The most general form of Logistic Regression model is:

$$f(x, \theta) = \hat{y} = \sigma(\theta_0 + \sum_{j=1}^{m-1} \theta_j \phi_j(x))$$
or
$$log(\frac{p}{1-p}) = \theta_0 + \sum_{j=1}^{m-1} \theta_j \phi_j(x)$$

where $\hat{y}$ is the predicted output for a given input, x = $(x_1, ..., x_d)^T$, is a d-dimensional input feature vector, $\phi_j(.)$ represents a basis function, $\theta = (\theta_1, ..., \theta_m)^T$ are the model parameters and m is the order of the linear model, and σ is the sigmoid funtion:

$$\sigma(x) = \frac{1}{1+e^{-}x}$$

C. Cost Function

Cost function is a measure of how wrong the model is in terms of its ability to estimate the relation between inputs and outputs. We use the logarithmic loss function for Logistic Regression which serves many advantages.

The formula logarithmic loss for a dataset with N samples is given by:

$$J(\theta) = \sum_{i=1}^{N} (-y * log(\hat{y}) - (1 - y) * log(1 - \hat{y}))$$

where $\hat{y}_i$ is the predicted output and y_i is the observed or given output. If y = 1, (1-y) term will become zero, therefore $-log(\hat{y})$ alone will be present. If y = 0, (y) term will become zero, therefore $-log(1 - \hat{y})$ alone will be present.

The objective is to find θ that minimizes the cost function J.

Linear regression uses mean squared error as its cost function. If this is used for logistic regression, then it will be a non-convex function of parameters (theta). Gradient descent will converge into global minimum only if the function is convex.

D. Solution Approaches

Following are the different approaches for finding the parameters of the desired regression model:

1) **Maximum Likelihood approach:** An probabilistic approach to estimating the model parameters.
2) **Gradient Descent approach:** An optimization approach to minimizing the cost function and finding the parameters.
3) **Grid Search:** A brute force search approach to finding the optimal parameters.

E. Residual Analysis

Randomness and unpredictability are the two main components of a regression model.

$$\text{Prediction} = \text{Deterministic} + \text{Statistic}$$

Deterministic part is covered by the predictor variable in the model. Stochastic part reveals the fact that the expected and observed value is unpredictable. There will always be some information that are missed to cover. This information can be obtained from the residual information.

Let's explain the concept of residue through an example. Consider, we have a dataset which predicts sales of juice when given a temperature of place. Value predicted from regression equation will always have some difference with the actual value. Sales will not match exactly with the true output value. This difference is called as residue.

Residual plot helps in analyzing the model using the values of residues. It is plotted between predicted values and residue. Their values are standardized. The distance of the point from 0 specifies how bad the prediction was for that value. If the value is positive, then the prediction is low. If the value is negative, then the prediction is high. 0 value indicates prefect prediction. Detecting residual pattern can improve the model.

Non-random pattern of the residual plot indicates that the model is,

- Missing a variable which has significant contribution to the model target
- Missing to capture non-linearity (using polynomial term)
- No interaction between terms in model

Characteristics of a residue

- Residuals do not exhibit any pattern
- Adjacent residuals should not be same as they indicate that there is some information missed by system.

F. Metrics for model evaluation - Confusion Matrix

A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in the dataset. It gives you insight not only into the errors being made by your classifier but more importantly the types of errors that are being made.

		Prediction outcome		
		p	**n**	**total**
actual value	**p′**	True Positive	False Negative	P′
	n′	False Positive	True Negative	N′
	total	P	N	

we can assign the event row as "positive" and the no-event row as "negative". We can then assign the event column of predictions as "true" and the no-event as "false".

This gives us:

"true positive" for correctly predicted event values. "false positive" for incorrectly predicted event values. "true negative" for correctly predicted no-event values. "false negative" for incorrectly predicted no-event values.

1) Null-Hypothesis and P-value: The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (¡ 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.

Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

III. Problem - Understanding and Modelling

We are presented with a problem involving shipwreck data of the Titanic. On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there weren't enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

We are given attributes of the passengers of the Titanic including Name, Sex, Age, Passenger class, Number of Siblings/ Spouse onboard, Parents and children onboard, Ticket number, Fare, Cabin, and the port where they embarked and the final dependent column being whether the particular passenger survived or not. We need to build a predictive model that answers the question: "what sorts of people are more likely to survive?" using the given passenger data, and discuss any insights and observations.

A. Data Pre-processing & Feature Generation

We start with reading the data to a pandas dataframe. We observe a total of 891 data points, with 12 columns in total, including the different features related to the person onboard. On visualizing the distributions of the people survived we see that around 38% of the total passengers survived the wreck.

(Figure 1) We then move on to checking for the null values in the data and find that we have three columns Age, Cabin and Embarked with missing data.

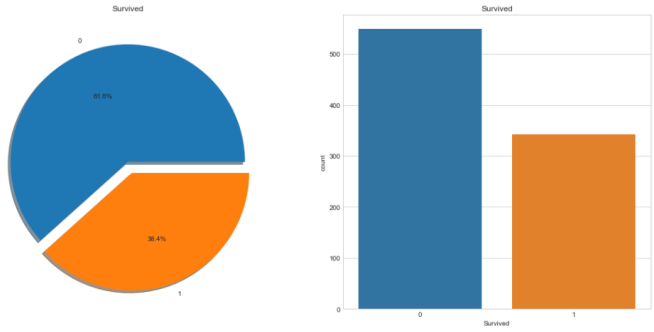

Fig. 1. Survived Distribution

Handling Age, we see that there are a total of 177 missing values that we need to take care of. We tend to check the correlation of Age with the other columns and find out that Pclass has the highest absolute correlation of about 0.37, thus we group age by Pclass and Sex and impute the missing values by the median computed for each group.

We then check for the missing values in Embarked and see only two values are missing. Checking Cabin we see that most of the values (77%) are missing. Thus we need to create a new class M (missing), and also extract the cabin class from the cabin column which includes cabin class followed by number. On visualizing the final Cabin column we see that most of the people in the missing class belong to the Pclass 3 and have little chance of survival. We see a higher chance of survival in most of the other cabin classes, with the least being in cabin A and the highest chance being in the cabin class B. We also observe that most of the people in the classes C, E, G, D, A, B belong to the Pclass 1, and have high chances of survival, thus indicating that Pclass might be a factor resulting in higher chances of survival.

On exploring the feature Embarked, we see that it contains three different classes S, C and Q. (Figure 2) while most of the people embarked from the S class and majority of them being from Pclass 3. The Passengers from C look to be lucky as a good proportion of them survived. The reason for this maybe the rescue of all the Pclass1 and Pclass2 Passengers. The Embark S looks to the port from where majority of the rich people boarded. Still the chances for survival is low here, that is because many passengers from Pclass3 around 81% didn't survive. Port Q had almost 95% of the passengers were from Pclass3.

We move on to feature generation, we start with checking that many people had the same Ticket numbers, thus we create a new feature representing the number of passengers having the same ticket number, which might be a representation of what was the size of the group in which they were travelling (Figure 3). On careful observation we see that different ticket frequency has different rates of survival with the highest being

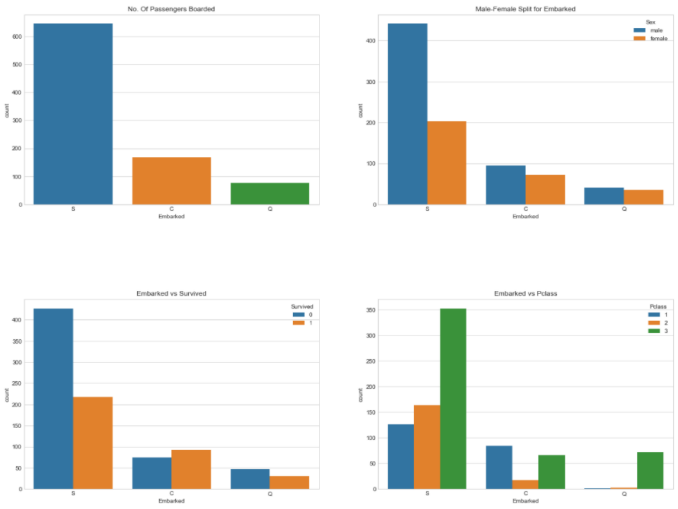

Fig. 2. Embarked

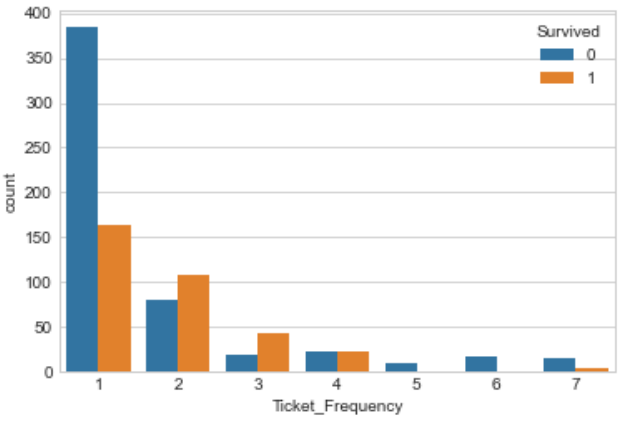

Fig. 3. Count-plot for Ticket Frequency

for the group of two and the chance for groups of more than 4 being very less.

Following that we extract the title for each of the passengers from the Name column and find out that people with different titles, had different chances of getting away. We create three categories with the largest being Mr, then clubbing Miss, Mrs and Miss into one, followed by Master and the others grouped into one (Figure 4). We can see that most of the people are form Mr group where the probability is low while the probability of Miss, Mrs, Ms and Master is higher for survival. We have very less people belonging to other classes. We also create a new feature followed from the title classes as Married wherein every passenger with Mrs class is termed as married and we see that the probability of survival for the married class is high (Figure 5).

We created a Child feature for people below 18 years of age, now we can see that children have a higher chance of getting out (Figure 6). We also create a new feature Family size, which is essentially No. of Parents + Children + Siblings + Spouse + 1, post which we create three different categories

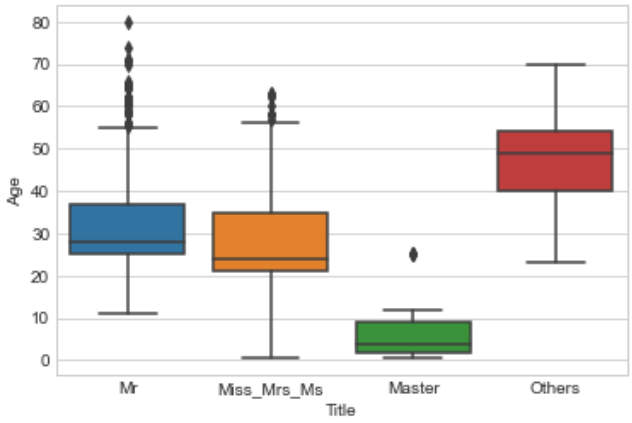

Fig. 4. Boxplot of various titles

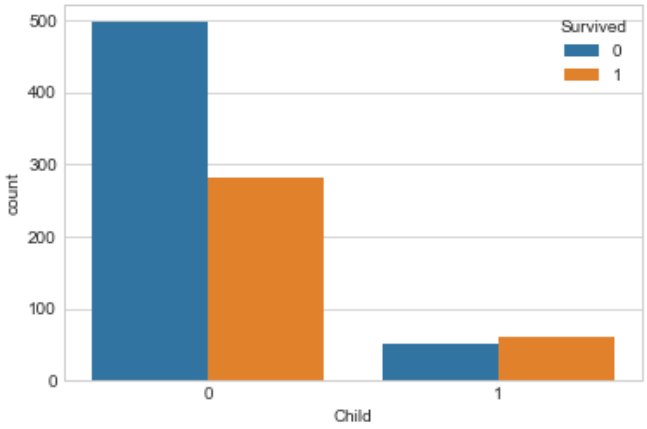

Fig. 6. Countplot - Child or Not

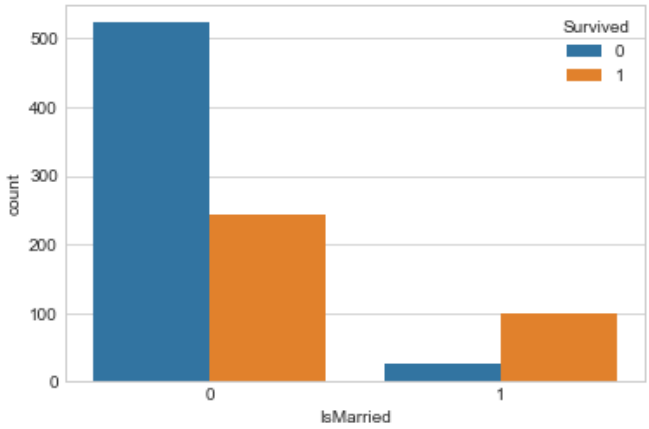

Fig. 5. Countplot - Married or Not

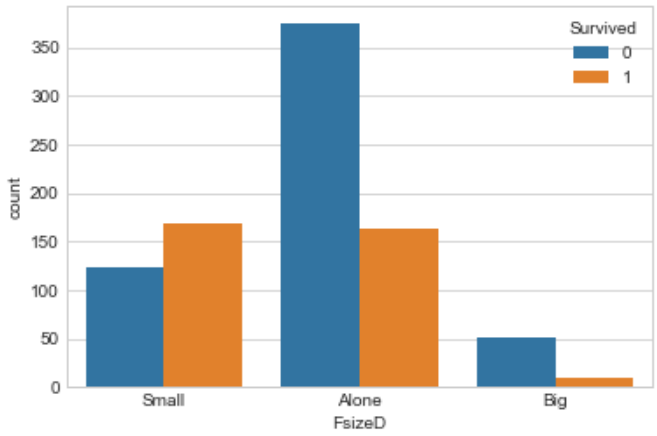

Fig. 7. Countplot for Family Size

for the family size class as Alone, Small and Big. We see that people with small family size have the highest chance of Surviving, and the ones with a big family, which the lease chance (Figure 7).

B. Visualization

We start by creating countplots for the categorical variables with hue as the target columns, to check for the distribution of each of the categories in the classes, plus commenting on whether the different categories have different survival probabilities. Starting with Pclass (Figure 8), We see that the people belonging to the Pclass 1 have a higher chance of survival, which keeps on decreasing as we go down the classes. Following which we explore Sex, wherein we observe Proportion of male and female: 2/3 vs 1/3. Male is much less likely to survive, with only 20% chance of survival. For females, ¿70% chance of survival. Obviously, Sex is an important feature to predict survival (Figure 9).

We then explore Age (Figure 10), which is not a categorical variable, thus we create histogram and boxplots for the same, resulting in insights such as passengers are mainly aged 20-40 and younger passengers tends to survive. on exploring SibSp (FIgure 11) we see that most of the passengers are travelling alone and the ones with 1 sibling/spouse are more likely to survive. Visualizing Parch we see that ¿70% passengers travel without parents/children, wherein passengers travelling with parents/ children are more likely to survive than those not (Figure 12). Creating distplots and boxplots for Fare, we see that there are people paying too much for their tickets (outliers) and we can see that money gets you a better chance of survival as evident from the graphs (Figure 13).

Finally as part of postprocessing, we convert the age variable into bins of 5 and also convert the Fare variable into bins of 4. Finally we create dummies for the categorical variables so that they are meaningfull to the machine.

C. Statistical Logistic Regression Modelling

We use the Statsmodels library in python to model the Logistic regression as it gives us the added benefits of getting the significance values of the regression coefficients. Once we have processed all the independent variables we feed

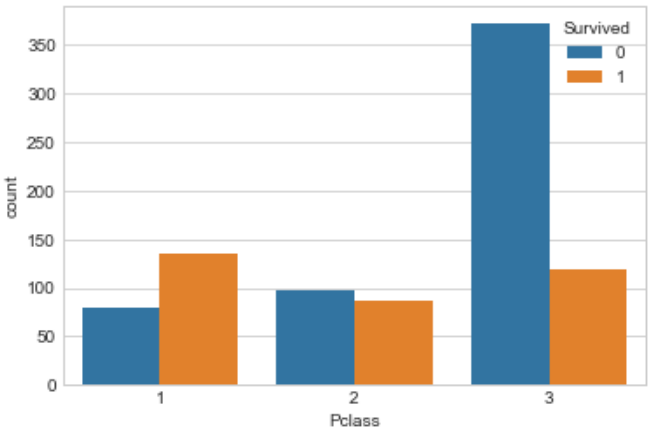

Fig. 8. Countplot for Passenger Class

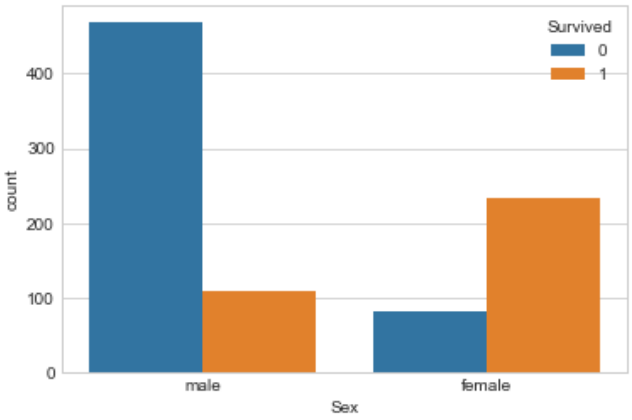

Fig. 9. Countplot - Male vs Female

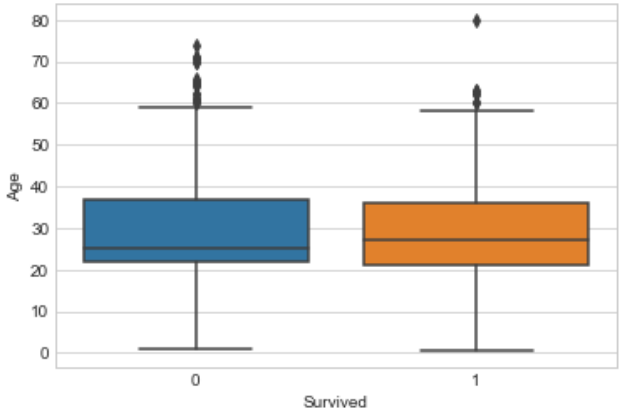

Fig. 10. Boxplot - Age

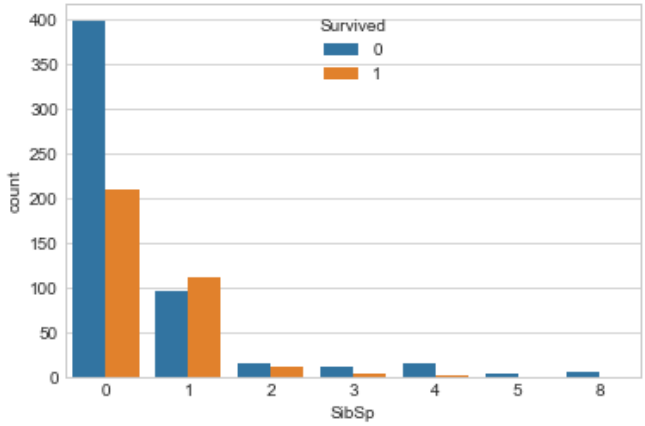

Fig. 11. Countplot - No. of Siblings/Spouse

them into the statsmodels logit formula, and then proceed to analyze the summary. We get a Log-Likelihood value of -350, with the method being maximum likelihood estimation. We see many coefficients having p values of greater than 0.05 (corresponding to 95% significance), suggesting that the coefficients are not statistically significant. Finally, we check for the best threshold and find out 0.56 to be the one with 79% error, and the confusion matrix.

IV. Conclusion

Having modelled a statistical model using different features, let's get to them one by one. Starting with fare range, we see that it has a p value greater than 0.05 and thus it does not contribute to the prediction by our model. We see that Pclass does contribute and has a negative coefficient thus verifying out initial observation that people with Pclass 1 are more likely to survive which goes down with Pclass. Neither of Parch, SibSp, Ticket frequency, IsMarried, Child are significant. following this we see that Age band has a negative coefficient which signifies that younger people are more likely to survive, thus backing our initial hypothesis. We see neither that Cabin, nor the Embarked port are statistically significant, while Mr has a high negative coefficient, thus stating that male passengers were less likely to survive than females and Others too has a negative coefficient thus backing our claim even stronger. We see Big has a negative coefficient thus saying that passengers with bigger families onboard were less likely to survive. Further avenues can be checking for more features that could explain the target variable better and look for minimizing the multicollinearity.

References

[1] Logistic regression, Data Analytics Laboratory EE4708, July - November 2021
[2] Logistic regression by Prof. Manikandan, Pattern Recognition and Machine Learning, July - November 2021
[3] https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
[4] https://en.wikipedia.org/wiki/Logistic_regression
[5] https://machinelearningmastery.com/confusion-matrix-machine-learning/

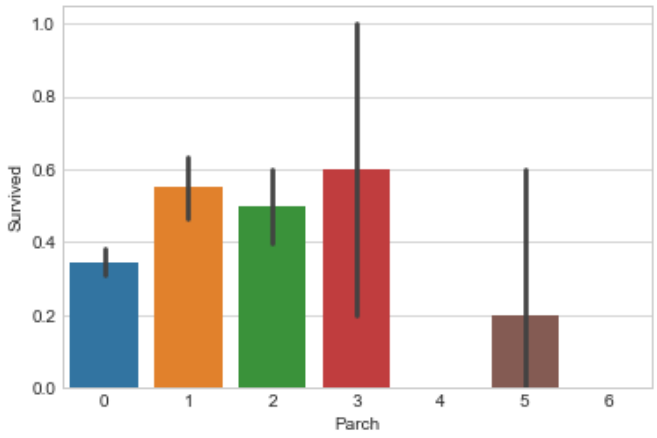

Fig. 12. Probability of Survival vs No. of Parents/Siblings

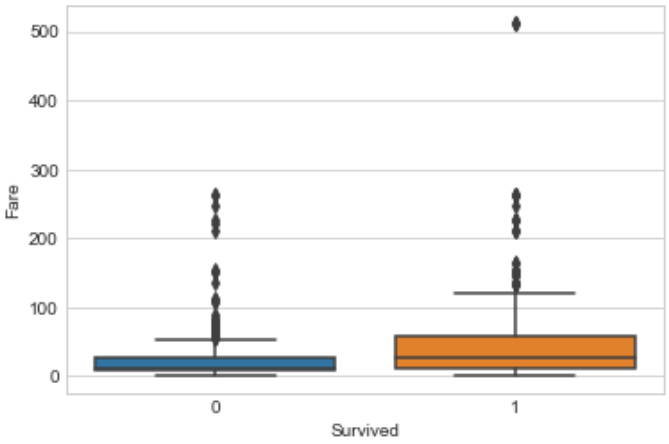

Fig. 13. Boxplot - Fare