Application of Logistic Regression on Passenger Survival Data of the Titanic Liner

Sajjida Reza, MS Scholar, College of Management Sciences, PAF KIET, Karachi, Pakistan

***Bilal Sarwar**, Faculty of Management Sciences, BUITEMS, Quetta, Pakistan

Raja Rub Nawaz, Karachi University Business School, Karachi, Pakistan

S. M. Nabeel Ul Haq, Faculty of Management Sciences, BUITEMS, Quetta, Pakistan

*Corresponding author's email: bilal.sarwar@buitms.edu.pk

ARTICLE DETAILS

History

Revised format: May 2021

Available Online: Jun 2021

Keywords

Binary, Dichotomous, Generalized Linear Model (GLM), Logistic Regression, Titanic liner

JEL Classification

C4, C49

ABSTRACT

Purpose: This empirical research aims to predict the distinguishing variables of passengers who did or did not survive while traveling in the famous Titanic liner, which sunk in 1912.

Design/Methodology/Approach: The binary logistic regression analysis empirically analyzes the secondary dataset available for 1046 passengers. Variables such as passenger's gender, age, family composition, ticket class, number of parents with/without children, and number of siblings and/or spouses were opted to examine the differences between the binary dependent variable (Passenger Survived/ Not Survived).

Findings: The study results indicate that all the variables are statistically significant in the model, with passenger's gender being the most significant predictor followed by passenger's ticket class. The survival chances of passengers decreased for male passengers compared to their counterparts (female passengers) for the sample data [$\text{Exp}(\beta)=0.080$], for the passengers of age more than 21 years compared to passengers of age less than and equal to 21 years [$\text{Exp}(\beta)=0.576$], and for passengers with ticket class second and third compared to first-class ticket holders [$\text{Exp}(\beta)=0.412$]. In contrast, there was a greater chance of survival for families traveling together with parents, siblings, spouses compared to single travelers [$\text{Exp}(\beta)=1.823$].

Implications/Originality/Value: The study is a classic example of the application of binary logistic regression analysis using EVIEWS software.

© 2021, The authors, under a Creative Commons Attribution-NonCommercial 4.0

Recommended citation: Reza, S., Sarwar, B., Nawaz, R. R., and Haq, S. M. N. U. (2021). Application of Logistic Regression on Passenger Survival Data of the Titanic Liner. *Journal of Accounting and Finance in Emerging Economies*, 7 (4), 861-867.

Introduction and Literature Review

On April 14th, 1912, the Titanic liner sank, killing 1523 people and leaving only 705 survivors. The goal of this study is to determine passenger survival using a data set of 1046 while considering the passenger's gender, age, family composition, ticket class, number of parents with/without children, and number of siblings and/or spouses. The predictors are analyzed by applying the binary logistic regression analysis to determine their predictive value in the passengers' survival (binary response variable, survived=1, not survived=0). The membership group in the outcome variable and the change to the analytic strategy are two primary research concerns that emerge in binary logistic regression studies (Warner, 2012).

Brief technical analytics was performed considering logistic regression (LR) and its diverse application in the business domain using the Web of Science (WoS) database (Table 1). The period selected was from 1980 to 2021. The filter criteria applied restricted the search to research articles in the English language, only which yielded 289 research studies for analytics. Initially, the focus was on the Web of Science category in which the LR has been applied. The result reflects that 35.64% of research papers having logistic regression as the primary statistical inference technique are falling under 'multidisciplinary sciences'. The next WoS category is 'economics' highlighting 21.10% of research articles. 'Management' is third in number having 16.27% of research papers bearing LR as the choice of statistical technique. The analytics of the WoS database yielded 25 primary categories but only the top three have been reported here.

Additionally, the 'research area' was also explored. 'Business economics' is the top discipline where 40.5% of research work has made use of the LR technique. This is followed by the domains of 'science technology' and 'social sciences' with respective 35.64% and 14.18% research studies.

Table 1. Analytic Results

#	Result Categories	Records	% of 289
Web of Science			
1.	Multidisciplinary Sciences	103	35.64
2.	Economics	61	21.10
3.	Management	47	16.26
Research Area			
1.	Business Economics	117	40.48
2.	Science Technology	103	35.64
3.	Social Sciences	41	14.18

Researchers frequently use more than one response variable in their studies. There are situations where a standard set of predictor factors can predict several response variables (Breiman & Friedman, 1997). Like, the study of anxiety and depression illnesses. In modern western countries, mental diseases are quite common, and there is typically a significant degree of comorbidity among them. In the Netherlands, to be precise. Thus, data on a substantial number of respondents' personalities and mental problems were obtained in the Study for Depression and Anxiety (Penninx et al., 2008; Spinhoven et al., 2010).

In the study of drug consumption patterns, Fehrman, Muhammad, Mirkes, Egan, and Gorban (2017) are curious about the drug consumption profiles of their individuals and how they connect to personality traits like sensation-seeking and impulsivity. Therefore, they gathered information on the use of 18 different medicines.

The treatment effect is determined through clinical trials, where the outcome can be binary, i.e., cured or not cured. On the other hand, treatments can have side effects, categorized as present or absent. Therefore, it is critical to look at therapy and side effects simultaneously (Molenberghs & Verbeke, 2006).

Young adults are prone to psychosocial difficulties. The Strengths and Difficulties Questionnaire (SDQ), a relatively brief questionnaire, can screen for these difficulties in community settings, such as during large-scale general health check-ups. The SDQ is divided into two sections: self-report and parent-report. It should have good validity qualities; that is, it should anticipate specific psychosocial issues to be effective as a screening tool. Vugteveen, De Bildt, Hartman, and Timmerman (2018) looked at the SDQ's validity concerning four diagnoses.

Because the dependent variable is categorical and the independent variables (predictors) are continuous or discrete, the study demonstrates logistic regression analysis for research purposes. Furthermore, for special contextual reasons, the categorical dependant variable in logistic regression could be polytomous or multinomial (Meyers, Gamst, & Guarino, 2016).

Methodology

The data of the Titanic disaster of 1309 passengers for this study were accessed. The study utilized 1046 responses, and the sample with missing information was dropped from the analysis.

The criterion variable in this study, passenger's survival is a binary response variable, survived=1, not survived=0, and the secondary data is analyzed by applying the binary logistic regression analysis to determine their predictive value of the set of predictors in the passengers' survival using EVIEWS software.

Validation is required for statistical models designed for prediction. For example, cross-validation involves randomly dividing the data into two halves, one for model building and the other for model validation. Validation approaches are also required when investigators want to see if a model built in another location (for example, a hospital or medical center) is suitable for predicting their environment; this is referred to as external validation (Miller, Hui, & Tierney, 1991).

One of the most often used tools for binary classification is logistic regression (Agresti, 2003; Berkson, 1953; Cox, 1958). The logistic regression model was established in the second part of the twentieth century, even though the logistic function has been known since the early nineteenth century (Cramer, 2002). Adaptations of logistic regression models have been devised to make them more flexible, for example, through base expansion, or less flexible, for instance, through regularisation (Friedman, Hastie, & Tibshirani, 2001).

An outcome or dependent variable Y_i is modeled as a linear function of a series of predictors $X_1, X_2, \dots, X_k$ using the equation $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_k X_k$ in ordinary least squares regression analysis (OLS). When the dependent variable is dichotomous and coded (0,1), the model is logit $(Y) = \ln[\Pr(y = 1)/\Pr(Y = 0)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_k X_k$ in logistic regression (Hosmer Jr, Lemeshow, & Sturdivant, 2013; Studenmund, 2014).

The independent variables chosen, among the other variables, in the study are presented in Table 2.

Empirical Results and Discussion

Table 3 presents the sample characteristics of 1046 passengers out of 1309 passengers from the Titanic disaster dataset.

Table 2. Variables of the study

#	Description
---	-------------

1. Gender		
0	=	Female
1	=	Male
2. Age		
1	≤	21 years
2	>	21 years
3. Passenger Family Composition		
1	=	Travelling Alone
2	=	Travelling with spouses or siblings
3	=	Travelling with parents or children
4	=	Travelling with complete family (Siblings or spouses and parents or children)
4. Ticket Class		
1	=	1 st class ticket
2	=	2 nd class ticket
3	=	3 rd class ticket
5. Number of parents without/with kids		
6. Number of spouses and or siblings		

The survival rate of passengers was only 40.8%, while 51.2% of the passengers perished in the famed shipping wreck of history. In the usable dataset of 1046 passengers for the study, 62.9% of passengers were male, while 37.1% were female. The dataset was divided into two age groups, i.e., age less than and equal to 21 years and more than 21 years for the present study, and the results suggest that 27.7% of the passengers belonged to the first category (aged ≤ 21 years) while 72.3% passengers from the second category, i.e., aged more than 21 years. Most passengers (48%) of the famous sinking incident were third-class ticket holders, 25% were second-class passengers, and 27% of passengers were first-class ticket holders (see Table 3).

Table 3. Sample Characteristics

Characteristics of Passengers ($n = 1046$)	Frequency	Percent
<i>Disaster Survival Status</i>		
Survived	427	40.8
Not Survived	619	51.2
Total	1046	100.0
<i>Gender</i>		
Male	658	62.9
Female	388	37.1
Total	1046	100.0
<i>Age</i>		
≤ 21 Years	290	27.7
> 21 Years	756	72.3
Total	1046	100.0
<i>Ticket Class</i>		
First	284	27.0
Second	261	25.0
Third	501	48.0
Total	1046	100.0

The results regarding the survival of passengers in Table 4 indicate that a greater number of female passengers survived the tragic event (62.9%) in comparison to male passengers (37.1%). Furthermore, Table 4 also reveals that most passengers from the age group of more than 21 years (72.3%) survived the unfortunate sinking of the Titanic in comparison to passengers aging less than and equal to 21 years (29.3%). Finally, passengers with a first-class ticket were at the top of surviving the catastrophe of the sinking of the Titanic with a survival rate of 42.4%, followed by 30.7% passengers holding a Third-class ticket and 26.9% passengers with a second-class ticket.

Table 4. Survival Status

Characteristics of Passengers (<i>n</i> = 427)	Frequency	Percent
<i>Gender</i>		
Male	135	31.6
Female	292	68.4
Total	427	100.0
<i>Age</i>		
≤ 21 Years	125	29.3
> 21 Years	302	70.3
Total	427	100.0
<i>Ticket Class</i>		
First	181	42.4
Second	115	26.9
Third	131	30.7
Total	427	100.0

To empirically test the objective of the present research, binary logistic regression was applied to the dataset. Table 5 describes the binary logistic regression analysis results in the current piece of research work. The passenger survival in the sinking of Titanic was the outcome variable (0-Not survived; 1-survived); passenger's gender, their age, family composition, ticket class, parents' children traveling along, and spouse/ siblings traveling were taken to be the predictor variables.

The results indicate the statistically significant relationship of all the predictors and model significance at a 1% [χ^2 (6) = 436.606, $p < .001$]. Furthermore, the regression slope for gender, age, passenger ticket class, passengers traveling with or without their kids, and passengers traveling with spouses or siblings are negative, implying that probability of the survival of a passenger was higher for female passengers, passengers aged more than 21 years, and passengers with the first-class ticket.

Table 5. Binary Logistic Regression Analysis

Predictor Variables	β	Wald Chi-Square Test	Exp(β)	% Δ
Gender	-2.523***	216.100	.080	92%
Age	-.551***	7.959	.576	42.4%
Family composition	.601***	17.680	1.823	82.3%
tclass	-.887***	73.260	.412	58.8%
parch	-.430***	7.236	.650	35%
sibsp	-.646***	20.779	.524	47.6%

Note: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

The logistic regression could use two indicators, such as Cox and Snell R^2 and Nagelkerke R^2 , to check the variability in the outcome variable due to the predictors. It has been unanimously recognized that Cox and Snell R^2 indicator underestimates the real value (De La Viña & Ford, 2001; De Noble, Galbraith, Singh, & Stiles, 2007; Pituch & Stevens, 2015; Zewude & Ashine, 2016). The Nagelkerke pseudo R^2 indicates that 46% variation in the model is collectively explained by the set of predictors employed in the binary logistic regression model classifying the passengers who survived with those who did not.

The first predictor in the model, gender, had a negative sign for its partial slope coefficient is statistically significant at a 1% level of significance ($\beta = -2.523$; p -value = 0.000), indicating that a male passenger had a lower chance of survival. Gender is contributing the most in the model (Wald $\chi^2 = 216.10$). Age predicts the membership group (survival) significantly at a significance level of 1% ($\beta = -0.551$; p -value = 0.005), having a negative partial slope coefficient implying that passengers falling in the age group of more than 21 years had a lower chance of survival. The third predictor, family composition of passengers is statistically significant at a 1% level of significance ($\beta = 0.601$; p -value = 0.000). Ticket class of passengers is also statistically

significant in the model at 1% level of significance ($\beta = -0.887$; p -value = 0.000). Finally, while family composition ($\beta = 0.601$; p -value = 0.000) is positive, implying probability of survival was higher for a passenger traveling alone.

The null hypothesis for statistical significance testing of independent variables used in the study is that each independent variable is unrelated to the log odds ratio outcome variable. The null hypothesis test in logistic regression is the Wald χ^2 statistic explains the projected change in log-odds for each 1-unit change in the predictor variable (Warner, 2012).

Pampel (2000) proposed that the formula for the distance of $\text{Exp}(\beta)$ from 1 shows the effect size. The score of the independent variable increases by one unit, the percentage change in the odds ratio is easily determined in percentage terms. Furthermore, Thompson (1999) and Kirk (2003) emphasized that researchers should carefully evaluate each independent variable's effect size and strength used in the model to ensure that it is substantial enough to be regarded as meaningful and practically significant.

Table 5 provides the odds of passenger survival [see $\text{Exp}(\beta)$]. For example, the odds of survival of passengers decrease by 92% for male passengers compared to their counterparts (female passengers) for the sample data [$\text{Exp}(\beta)=0.080$], the probability of survival of passengers aged more than 21 years compared to passengers of age 21 years decreased by 42.4% [$\text{Exp}(\beta)=0.576$]. Furthermore, the probability of the survival of passengers with a greater ticket class decreases by 58.8% [$\text{Exp}(\beta)=0.412$]. Moreover, the likelihood of survival of passengers traveling together with parents, siblings, spouses increases by 82.3% compared to single travellers [$\text{Exp}(\beta)=1.823$].

Conclusions

In the present study, the authors developed a binary logistic regression function to distinguish from a set of factors of passengers traveling on the Titanic liner who survived or did not survive the catastrophe. This research concludes that all the six predictors, passenger's gender, age, family composition, ticket class, passengers traveling with or without their kids, and passengers traveling with spouses or siblings were significantly contributing in determining the inter-group differences implying that the passengers who survived the sinking of the Titanic were linked with their gender, their age group, their family composition, ticket class, number of children accompanying their parents and number of sibling and spouses accompanying passengers. The binary logistic regression model with survival as a binary dependent variable has 46% variance with the set of predictors employed.

References

- Agresti, A. (2003). *Categorical data analysis* (Vol. 482): John Wiley & Sons.
- Berkson, J. (1953). A Statistically Precise and Relatively Simple Method of Estimating the Bio-Assay with Quantal Response, Based on the Logistic Function. *Journal of the American Statistical Association*, 48(263), 565-599. doi:10.1080/01621459.1953.10483494
- Breiman, L., & Friedman, J. H. (1997). Predicting Multivariate Responses in Multiple Linear Regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1), 3-54. doi:<https://doi.org/10.1111/1467-9868.00054>
- Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2), 215-232. doi:10.1111/j.2517-6161.1958.tb00292.x
- Cramer, J. S. (2002). The origins of logistic regression. doi:10.2139/ssrn.360300
- De La Viña, L., & Ford, J. (2001). Logistic Regression Analysis of Cruise Vacation Market Potential: Demographic and Trip Attribute Perception Factors. *Journal of Travel Research*, 39(4), 406-410. doi:10.1177/004728750103900407
- De Noble, A., Galbraith, C. S., Singh, G., & Stiles, C. H. (2007). Market justice, religious

- orientation, and entrepreneurial attitudes. *Journal of Enterprising Communities: People and Places in the Global Economy*, 1(2), 121-134. doi:10.1108/17506200710752548
- Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., & Gorban, A. N. (2017). The five factor model of personality and evaluation of drug consumption risk *Data science* (pp. 231-242): Springer.
- Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1): Springer series in statistics New York.
- Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398): John Wiley & Sons.
- Kirk, R. E. (2003). The importance of effect magnitude. *Handbook of research methods in experimental psychology*, 83-105.
- Meyers, L. S., Gamst, G., & Guarino, A. J. (2016). *Applied multivariate research: Design and interpretation*: Sage publications.
- Miller, M. E., Hui, S. L., & Tierney, W. M. (1991). Validation techniques for logistic regression models. *Statistics in Medicine*, 10(8), 1213-1226. doi:10.1002/sim.4780100805
- Molenberghs, G., & Verbeke, G. (2006). *Models for discrete longitudinal data*: Springer Science & Business Media.
- Pampel, F. C. (2000). Sage Publications i. *Logistic Regression: A Primer: SAGE Publications*.
- Penninx, B. W. J. H., Beekman, A. T. F., Smit, J. H., Zitman, F. G., Nolen, W. A., Spinhoven, P., . . . Van Dyck, R. (2008). The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *International Journal of Methods in Psychiatric Research*, 17(3), 121-140. doi:10.1002/mpr.256
- Pituch, K. A., & Stevens, J. P. (2015). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS*: Routledge.
- Spinhoven, P., Elzinga, B. M., Hovens, J. G. F. M., Roelofs, K., Zitman, F. G., van Oppen, P., & Penninx, B. W. J. H. (2010). The specificity of childhood adversities and negative life events across the life span to anxiety and depressive disorders. *Journal of Affective Disorders*, 126(1), 103-112. doi:<https://doi.org/10.1016/j.jad.2010.02.132>
- Studenmund, A. H. (2014). *Using econometrics a practical guide*: Pearson.
- Thompson, B. (1999). Statistical Significance Tests, Effect Size Reporting and the Vain Pursuit of Pseudo-Objectivity. *Theory & Psychology*, 9(2), 191-196. doi:10.1177/095935439992007
- Vugteveen, J., De Bildt, A., Hartman, C. A., & Timmerman, M. E. (2018). Using the Dutch multi-informant Strengths and Difficulties Questionnaire (SDQ) to predict adolescent psychiatric diagnoses. *European Child & Adolescent Psychiatry*, 27(10), 1347-1359. doi:10.1007/s00787-018-1127-y
- Warner, R. M. (2012). *Applied statistics: From bivariate through multivariate techniques*: Sage Publications.
- Zewude, B. T., & Ashine, K. M. (2016). Binary Logistic Regression Analysis in Assessment and Identifying Factors That Influence Students' Academic Achievement: The Case of College of Natural and Computational Science, Wolaita Sodo University, Ethiopia. *Journal of Education and Practice*, 7(25), 3-7.