RESEARCH ARTICLE

THE TRAGEDY OF TITANIC: A LOGISTIC REGRESSION ANALYSIS.

Dina Ahmed Mohamed Ghandour¹ and May Alawi Mohamed Abdalla².
1. Lecturer at University of Medical Sciences & Technology, Faculty of Business Administration, Khartoum, Sudan, DBA Candidate.
2. Pharmacist at Samasu Medical & Educational Services, Khartoum Sudan, and currently is a DBA Candidate.

Abstract

The sinking of Titanic is one of the most infamous shipwrecks in history. On April 15, 1912, during her maiden voyage, the Titanic sank after colliding with an iceberg, killing 1,502 of the 2,228 passengers and crew. This sensational tragedy shocked the international community and motivated the adoption of better maritime safety regulations. However there are many reasons that the shipwreck led to such loss of life and there was some elements of luck involved in surviving the sinking as some groups of people were more likely to survive than others. The main aim of this research is to identify the Impact of gender, passenger class, Accompany, age on a person's likelihood of surviving the shipwreck. Secondary data was used as the main data collection tool and it was analyzed by fitting a logistic regression model using a statistical package, SPSS. Findings of this study showed that some passenger groups were more likely to survive than others, with respect to certain demographic characteristic and whether the passenger was traveling in the first, second or third class.

Background:

RMS Titanic was a British passenger liner that sank in the North Atlantic Ocean in the early morning of 15 April 1912, after colliding with an iceberg during her maiden voyage from Southampton to New York City. Of the estimated 2,228 passengers and crew on board, more than 1,500 died, making it one of the deadliest commercial peacetime maritime disasters in modern history. The largest ship afloat at the time it entered service, the RMS Titanic was the second of three Olympic class ocean liners operated by the White Star Line, and was built by the Harland and Wolff shipyard in Belfast. Thomas Andrews, her architect, died in the disaster. Under the command of Edward Smith, who went down with the ship, Titanic carried some of the wealthiest people in the world, as well as hundreds of emigrants from Great Britain and Ireland, Scandinavia and elsewhere throughout Europe seeking a new life in North America. The first-class accommodation was designed to be the pinnacle of comfort and luxury, with an on-board gymnasium, swimming pool, libraries, high-class restaurants and opulent cabins. A high-power radiotelegraph transmitter was available for sending passenger “Marconi grams” and for the ship's operational use. (1)

Although Titanic had advanced safety features such as watertight compartments and remotely activated watertight doors, there were not enough lifeboats to accommodate all of those aboard, due to outdated maritime safety...
regulations. *Titanic* only carried enough lifeboats for 1,178 people—slightly more than half of the number on board, and one third of her total capacity.(3) After leaving Southampton on 10 April 1912, *Titanic* called at Cherbourg in France and Queenstown (now Cobh) in Ireland before heading west to New York. On 14 April, four days into the crossing and about 375 miles (600 km) south of Newfoundland, the ship hit an iceberg at 11:40 p.m. ship's time. The collision caused the ship's hull plates to buckle inwards along its starboard (right) side and opened five of its, sixteen watertight compartments to the sea; it could only survive four flooding. Meanwhile, passengers and some crew members were evacuated in lifeboats, many of which were launched only partially loaded. A disproportionate number of men were left aboard because of a "women and children first" protocol for loading lifeboats. At 2:20 a.m., it broke apart and foundered—with well over one thousand people still on board. (4) The disaster was greeted with worldwide shock and outrage at the huge loss of life and the regulatory and operational failures that had led to it. Public inquiries in Britain and the United States led to major improvements in maritime safety. (5)

One of their most important legacies was the establishment in 1914 of the International Convention for the Safety of Life at Sea (SOLAS), which still governs maritime safety today. Additionally, several new wireless regulations were passed around the world in an effort to learn from the many missteps in wireless communications—which could have saved many more passengers. The below table explores Titanic profile:

<table>
<thead>
<tr>
<th>Name</th>
<th>RMS Titanic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Owner</td>
<td>White Star Line</td>
</tr>
<tr>
<td>Port of Registry</td>
<td>Liverpool, UK</td>
</tr>
<tr>
<td>Route</td>
<td>Southampton to New York City</td>
</tr>
<tr>
<td>Ordered</td>
<td>17 September 1908</td>
</tr>
<tr>
<td>Builder</td>
<td>Harland and Wolff, Belfast</td>
</tr>
<tr>
<td>Cost</td>
<td>$7.5 million ($300 million in 2017)</td>
</tr>
<tr>
<td>Yard Number</td>
<td>401</td>
</tr>
<tr>
<td>Laid Down</td>
<td>31 March 1909</td>
</tr>
<tr>
<td>Launched</td>
<td>31 May 1911</td>
</tr>
<tr>
<td>Completed</td>
<td>2 April 1912</td>
</tr>
<tr>
<td>Maiden Voyage</td>
<td>10 April 1912</td>
</tr>
<tr>
<td>In Service</td>
<td>10–15 April 1912</td>
</tr>
<tr>
<td>Identification</td>
<td>Radio call sign &quot;MGY&quot;</td>
</tr>
<tr>
<td>Fate</td>
<td>Hit an iceberg 11:40 p.m. (ship's time) 14 April 1912 on her maiden voyage and sank 2 h 40 min later</td>
</tr>
<tr>
<td>Status</td>
<td>Wreck</td>
</tr>
</tbody>
</table>

**Research Question:**
This research was conducted to answer the following question regarding Titanic Shipwreck:
Were Some Passenger Groups More Likely to Survive than Others?

**Objective of the study:**
The General Objective of this research is to:
• Explain the Impact of gender, passenger class, Accompany, age on a person’s likelihood of surviving the shipwreck.

**Methodology:**
This study is based on analytical and quantitative methods.

**Target Population:**
Titanic Passengers

(1, 2, 3, 4, 5) https://en.wikipedia.org/wiki/RMS_Titanic
Population Size:-
Of the 2,228 passengers on board, a data on 1,309 was found and a full coverage of this data was analyzed to determine the survival status of those passengers on board (1,309 passengers)
Hosmer and Lemeshow recommended a Population size / sample size greater than 400

Data Collection:-
Secondary data was obtained from the internet regarding Titanic passengers; the data can be downloaded from the following link:


Key Variables of the study:-
Dependent variable:- Survival Status (survived=1, not survived=0).

Independent / explanatory variables:-
- Age (Code: Adult = 1, Child = 0)
- Gender (Code: Female=1, Male = 0)
- Passenger class (Code: 1st class = 1, 2nd class = 0, 3rd class= 1, 3rd class is the reference class so if 1st class = 0 and 2nd class = 0, the person must have been in 3rd class)
- Travelling Alone (Code: 0 Travelling alone, 1=Travelling with family)

Data Analysis:- The collected data was analyzed by fitting a logistic regression model using SPSS

Testing Logistic Regression Assumptions on Titanic Data Set:-

Assumption # 1
Logistic regression does not assume a linear relationship between the dependent and independent variables, this assumption can be represented in the following diagrams;
**Survive vs Accompany**

- Survived
- Not Survived

- Travelling Alone
- Traveling with family

**Survive vs Pclass**

- Survived
- Not Survived

- 1st class
- 2nd class
- 3rd class
Assumption#2:-
The dependent variable should be measured on a dichotomous scale (i.e. Survival status: Survived vs. Not Survive)

Assumption#3:-
There must be one or more independent variables, which can be either continuous (i.e., an interval or ratio variable) or categorical (i.e., an ordinal or nominal variable).

Independent Variables in Titanic Data Set: Categorical Variables: Passenger Class, Age, Gender, Travelling alone.

Assumption#4:-
Absence of Multicollinearity (it refers to predictors that are correlated with other predictors in the model)
Can be tested by:-
1. Correlation Matrix &
2. VIF (Variance Inflation Factor)

1. Correlation Matrix:-
The below table show that since Correlation Coefficient < 0.9, multicollinearity doesn’t exist
2. **VIF (Variance Inflation Factor):** Collinearity exist if VIF > 5

**Coefficients a**

<table>
<thead>
<tr>
<th>Model</th>
<th>Tolerance</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 pclass</td>
<td>.912</td>
<td>1.096</td>
</tr>
<tr>
<td>Age</td>
<td>.927</td>
<td>1.079</td>
</tr>
<tr>
<td>Accompany</td>
<td>.954</td>
<td>1.048</td>
</tr>
<tr>
<td>Gender Numeric</td>
<td>.949</td>
<td>1.054</td>
</tr>
</tbody>
</table>

**a. Dependent Variable: Survived**

**Assumption#5:**
The categories (groups) must be mutually exclusive and exhaustive; a case can only be in, one group and every case must be a member of one of the group.

**Assumption#6:**
Large sample /Population sizes are required for logistic regression to provide sufficient numbers in both categories of the response variable.
With small sample sizes, Hosmer-Lemeshow test has low power. Hosmer recommended a sample sizes greater than 400.
(Titanic Population size is 1309 Passengers on Board)
Analysis: SPSS output and Interpretation:-

Table 1 & 2 respectively: - Output: Initial Model.

### Case Processing Summary

<table>
<thead>
<tr>
<th>Unweighted Cases</th>
<th>N</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selected Cases</td>
<td>1309</td>
<td>100.0</td>
</tr>
<tr>
<td>Included in Analysis</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing Cases</td>
<td>0</td>
<td>.0</td>
</tr>
<tr>
<td>Total</td>
<td>1309</td>
<td>100.0</td>
</tr>
<tr>
<td>Unselected Cases</td>
<td>0</td>
<td>.0</td>
</tr>
<tr>
<td>Total</td>
<td>1309</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Illustration to the above Table:-
The table shows that there is 1309 passengers in the sample and No missing data

### Dependent Variable Encoding

<table>
<thead>
<tr>
<th>Original Value</th>
<th>Internal Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Survived</td>
<td>0</td>
</tr>
<tr>
<td>Survived</td>
<td>1</td>
</tr>
</tbody>
</table>

Illustration: This table tells how SPSS has coded our outcome variable: *Survival 1, Not Survived 0*

Block 0: Beginning Block (Constant Only Model):-
Block zero means that there are no predicted variables included in the model, it’s the intercept model (Null model).

### Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Survived</td>
<td>Not Survived</td>
</tr>
<tr>
<td>Survived</td>
<td>899</td>
<td>0</td>
</tr>
<tr>
<td>Not Survived</td>
<td>500</td>
<td>0</td>
</tr>
</tbody>
</table>

* Overall Percentage: 61.8%

Illustration to the above Table:-
- **Observed:** Indicates the number of 0’s & 1’s that are observed in the dependent variable
- **Predicted:** SPSS has predicted that all cases are 0 on the dependent variable

The classification table helps in assessing the performance of the model by cross tabulating the observed response categories with the predicted response categories.
The table suggests that if we knew nothing about our variables and guessed that no one will survive we would be correct 61.8%.

Block 0: Beginning Block cont.

### Variables in the Equation

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>-0.481</td>
<td>0.057</td>
<td>71.551</td>
<td>1</td>
<td>0.000</td>
<td>0.618</td>
</tr>
</tbody>
</table>

Illustration to the above Table:-

1. The variable in the equation table only includes the constant so each person has the same chance of survival.
2. The Wald $X^2$ statistics is used to test the significance of B coefficient.

To test the significance of the coefficient (intercept) we set the following:-

- **Ho**: the intercept = 0
- **Ha**: the intercept ≠ 0
- **Sig**\(=0.000\) < \(\alpha = 0.05\)

So reject H0 and accept the Ha, which means that the intercept doesn't pass through the origin.

\[
\ln\left( \frac{P}{1-P} \right) = \beta_0
\]

1. The null model is: \(\text{Logit} (P) = -0.481\)

Block 0: Beginning Block cont.

### Variables not in the Equation

<table>
<thead>
<tr>
<th>Variables</th>
<th>Score</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gender(1)</td>
<td>365.887</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>Accompany(1)</td>
<td>17.956</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>AgeGroup(1)</td>
<td>13.715</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>pclass</td>
<td>127.859</td>
<td>2</td>
<td>0.000</td>
</tr>
<tr>
<td>pclass(1)</td>
<td>102.222</td>
<td>1</td>
<td>0.000</td>
</tr>
<tr>
<td>pclass(2)</td>
<td>3.377</td>
<td>1</td>
<td>0.066</td>
</tr>
<tr>
<td>Overall Statistics</td>
<td>457.600</td>
<td>5</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Illustration to the above table:

- The variables not in the equation table tells us whether each independent variable improves the model.
- This table present the information for the variables that were not included in step zero model.

Block 1: Method = Enter:-
Omnibus Tests of Model Coefficients

<table>
<thead>
<tr>
<th></th>
<th>Chi-square</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td>Step</td>
<td>499.369</td>
<td>5</td>
</tr>
<tr>
<td>Block</td>
<td></td>
<td>499.369</td>
<td>5</td>
</tr>
<tr>
<td>Model</td>
<td></td>
<td>499.369</td>
<td>5</td>
</tr>
</tbody>
</table>

Illustration to the above table:
The Chi Square test compares the fit of the model

In our case model chi square has 5 degree of freedom a value of 499.369 and sig < 0.05 which tested by the following hypothesis:

Ho: The model is not a good fitting model
Ha: The model is a good fitting model

Sig=.000 < α = .05

So:-
Since p value (sig) of less than 0.05 for block means that block 1 model is a significant improvement to the block 0 model.

Block 1: Method = Enter

Model Summary

<table>
<thead>
<tr>
<th></th>
<th>-2 Log likelihood</th>
<th>Cox &amp; Snell R Square</th>
<th>Nagelkerke R Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td>1241.659</td>
<td>.317</td>
<td>.431</td>
</tr>
</tbody>
</table>

Illustration to the above Table:-
✓ There is no close analogs statistic in logistic regression to coefficient of determination to measure the usefulness of the model
✓ The model summary table provides some approximations such ;

Cox & Snell & Nagelkerke R square:-

Decision Rule:-
Cox & Snell R square <1
Nagelkerke R square from 0 -1

Interpretation:-
It is indicating that 32% and 43% of the variation in survival can be explained by the model i.e. the value of 0.32 & 0.43 indicates that the model is useful in predicting survival

Block 1: Method = Enter
Hosmer and Lemeshow Test

<table>
<thead>
<tr>
<th>Step</th>
<th>Chi-square</th>
<th>df</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31.752</td>
<td>6</td>
<td>.000</td>
</tr>
</tbody>
</table>

Illustration to the above Table:
The Hosmer & Lemeshow Test is a commonly used test for assessing the goodness of fit of a logistic regression model but has a low power in assessing the significance of the model.

Main problems of Hosmer & Lemeshow:
- The non-significant chi-square is indicative of good fit of the model in case of small sample size.
- Even with good fit the test may be significant if sample size is large.
- Even with poor fit the test may not be significant if sample size is small.

Block 1: Method = Enter

Classification Table

<table>
<thead>
<tr>
<th>Observed</th>
<th>Predicted</th>
<th>Survived</th>
<th>Not Survived</th>
<th>Percentage Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td></td>
<td>Survived</td>
<td>680</td>
<td>129</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not Survived</td>
<td>155</td>
<td>345</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overall Percentage</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

a. The cut value is .500

Illustration to the above Table:
The overall predictive capacity increased from 61.8% to 78%

Important terms in the Table:
- Sensitivity: percentage of occurrences correctly predicted: $345/500 = 69\%$
- Specificity: percentage of nonoccurrence's correctly predicted: $680/809 = 84\%$

Variables in the Equation

<table>
<thead>
<tr>
<th>Variables</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>Sig.</th>
<th>Exp(B)</th>
<th>95% C.I. for Exp(B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pclass</td>
<td>1.824</td>
<td>.175</td>
<td>109.559</td>
<td>2</td>
<td>.000</td>
<td>6.194</td>
<td>(4.394, 8.733)</td>
</tr>
<tr>
<td>pclass(1)</td>
<td>.886</td>
<td>.180</td>
<td>108.286</td>
<td>1</td>
<td>.000</td>
<td>2.425</td>
<td>(1.704, 3.481)</td>
</tr>
<tr>
<td>pclass(2)</td>
<td>.256</td>
<td>.149</td>
<td>281.504</td>
<td>1</td>
<td>.000</td>
<td>12.252</td>
<td>(9.143, 16.418)</td>
</tr>
<tr>
<td>Gender(1)</td>
<td>.267</td>
<td>.117</td>
<td>148.435</td>
<td>1</td>
<td>.000</td>
<td>1.791</td>
<td>(1.441, 2.223)</td>
</tr>
<tr>
<td>Accompany(1)</td>
<td>.087</td>
<td>.157</td>
<td>.304</td>
<td>1</td>
<td>.551</td>
<td>1.090</td>
<td>(0.802, 1.484)</td>
</tr>
<tr>
<td>AgeGroups(1)</td>
<td>1.021</td>
<td>.265</td>
<td>14.853</td>
<td>1</td>
<td>.000</td>
<td>2.775</td>
<td>(1.651, 4.662)</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.308</td>
<td>.186</td>
<td>154.636</td>
<td>1</td>
<td>.000</td>
<td>.099</td>
<td></td>
</tr>
</tbody>
</table>

a. Variable(s) entered on step 1: pclass, Gender, Accompany, AgeGroups.
Illustration to the above Table:-
Wald (Sig):-
- In the sig column, the p-values are all below 0.05 apart from the test for the variable Accompany, \( p = 0.581 \).
- This means that there is no relationship between that variable and survival.
- Class is tested as a whole (P class) and then 1st and 2nd class compared to the reference category 3rd class.

Exp (B): Interpretation of Odd Ratio:-
When interpreting the differences, look at the exp (B) column which represents the odds ratio for the individual variable.

Pclass: Those in 1st class were 6.194 times more likely to survive than those in second class.
Gender: Female are 12.25 times more likely to survive than the men
Age Group: Children are 2.775 times more likely to survive than the Adult

The full Model Being Tested is

\[
\text{Logit} (P) = \ln \left( \frac{P}{1-P} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k
\]

Logit (P) = -2.038 + 1.824 X Pclass1 + .886 X Pclass2 + 2.506 X Gender +1.021 X Age group

Findings:-
The chance of survival was apparently related to the below factors:
- Passenger Class (First, Second, Third class)
- Age Group (Child or Adult)
- Gender (Male, Female)

The observed survival percentage is directly related to:-
- Economic status with higher status (first class) associated with higher survival probability
- Women had a higher survival rate than men
- Children age group had a higher survival rate

Recommendations:-
This analysis could be extended as another possible explanatory variable could be included whether or not a passenger got on a lifeboat or not. This seems to be a significant determinant of survival.

Acknowledgment:-
We would like to take this opportunity to appreciate and thank Dr. Arbab Faris for his guidance and support.

References:-