

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/318391323>

THE TRAGEDY OF TITANIC: A LOGISTIC REGRESSION ANALYSIS.

Article in International Journal of Advanced Research · June 2017

DOI: 10.21474/IJAR01/4558

CITATIONS

0

READS

1,502

2 authors, including:



Dina Ghandour

University of Medical Sciences and Technology

9 PUBLICATIONS 2 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Audit Expectation Gap (AEG) [View project](#)



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.

Manuscript Info

Manuscript History

Received: 11/06/2017

Final Accepted: 17/06/2017

Published: June 2017

Key words:-

Ship Wreck, Survival Status, Logistic Regression, Titanic Passengers

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.

Copy Right, IJAR, 2017. All rights reserved.

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.(1) 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. (2).



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

Corresponding Author: - Dina Ahmed Mohamed Ghandour.

Address: - Lecturer, Faculty of Business Administration, University of Medical Sciences & Technology, Khartoum, Sudan.

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:

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

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:

Source: - Hind, Philip. Encyclopedia Titanic. Online-only resource. Retrieved 01Feb2012 from <http://www.encyclopedia-titanica.org/>

Key Variables of the study:-

Dependent variable:- Survival Status (survived=1, not survived=0).

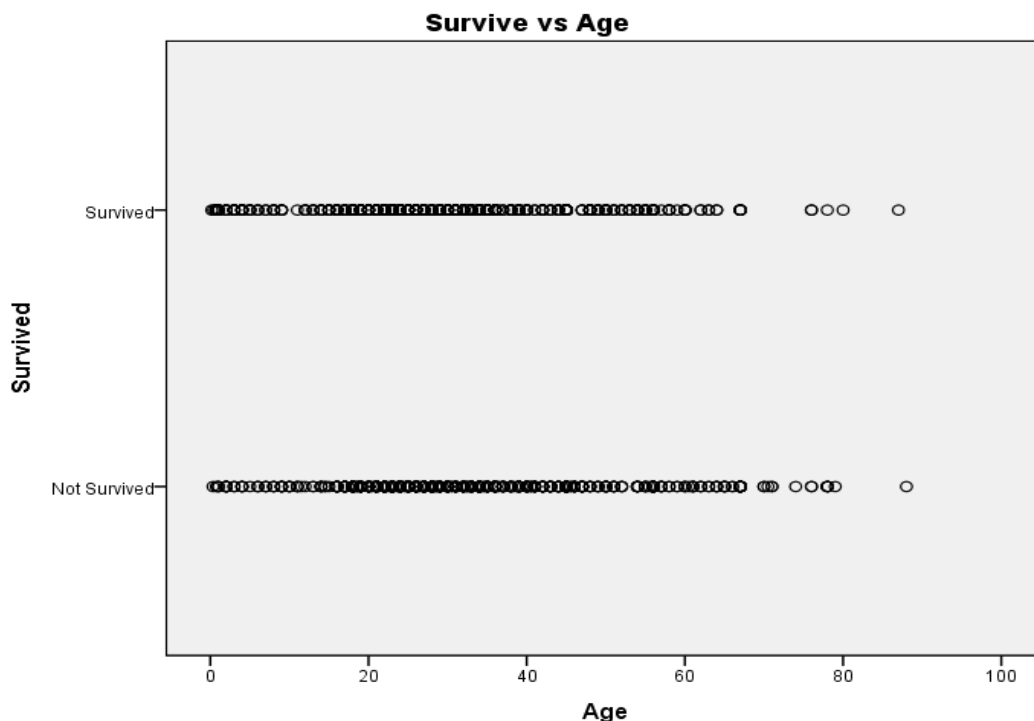
Independent / explanatory variables:-

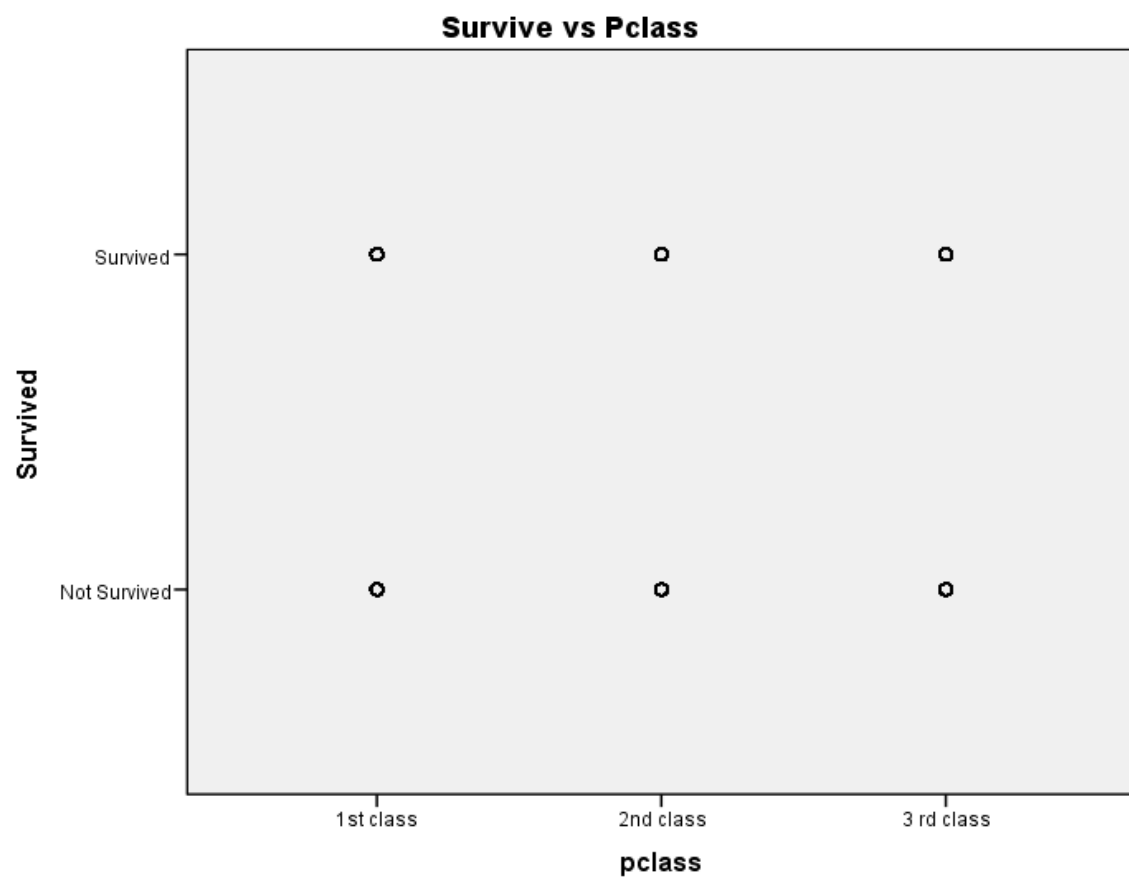
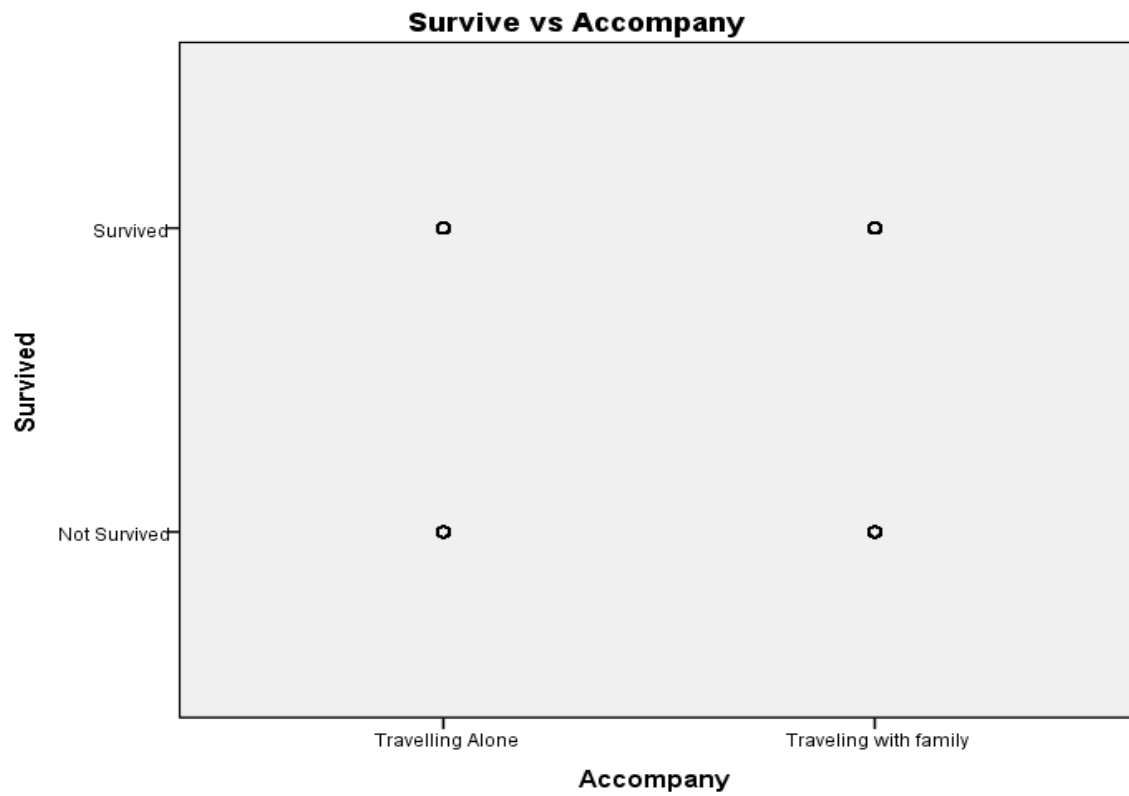
- ❖ Age (Code: Adult = 1 , Child = 0)
- ❖ Gender (Code : Female=1 , Male = 0)
- ❖ Passenger class (Code : 1st class = 1 , 2nd class = 0 , 3rd class= 1 , 3rd class is the reference class so if 1st class = 0 and 2nd class = 0, the person must have been in 3rd class)
- ❖ Travelling Alone (Code: 0 Travelling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

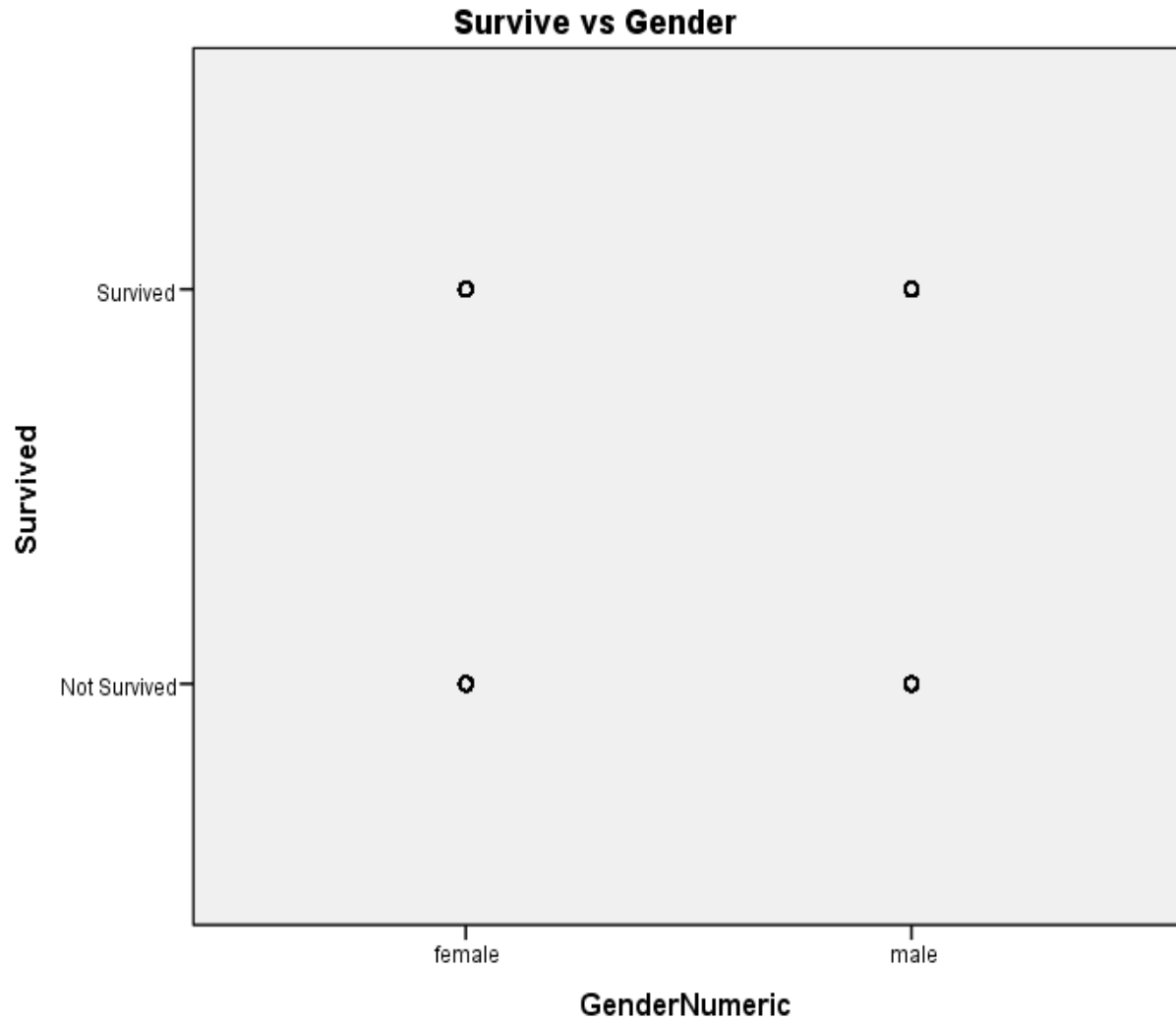
Data Analysis:- The collected data was analyzed by fitting a logistic regression model using SPSS

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;





**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

Correlation Matrix							
		Constant	pclass(1)	pclass(2)	Gender(1)	Accompany (1)	AgeGroups(1)
Step 1	Constant	1.000	-.527-	-.415-	-.524-	-.702-	-.318-
	pclass(1)	-.527-	1.000	.372	.251	.120	.177
	pclass(2)	-.415-	.372	1.000	.121	.067	.084
	Gender(1)	-.524-	.251	.121	1.000	.152	.060
	Accompany(1)	-.702-	.120	.067	.152	1.000	.195
	AgeGroups(1)	-.318-	.177	.084	.060	.195	1.000

2. **VIF (Variance Inflation Factor):** Collinerity exist if $VIF > 5$
Coefficients^a

Model		Collinearity Statistics	
		Tolerance	VIF
1	pclass	.912	1.096
	Age	.927	1.079
	Accompany	.954	1.048
	GenderNumeric	.949	1.054

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

Unweighted Cases ^a		N	Percent
Selected Cases	Included in Analysis	1309	100.0
	Missing Cases	0	.0
	Total	1309	100.0
Unselected Cases		0	.0
Total		1309	100.0

Illustration to the above Table:-

The table shows that there is 1309 passengers in the sample and No missing data

Dependent Variable Encoding

Original Value	Internal Value
Not Survived	0
Survived	1

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^{a,b}

Observed			Predicted		
			Survived		Percentage Correct
			Not Survived	Survived	
Step 0	Survived	Not Survived	809	0	100.0
		Survived	500	0	.0
	Overall Percentage				61.8

a. Constant is included in the model.

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

	B	S.E.	Wald	df	Sig.	Exp(B)
Step 0 Constant	-.481-	.057	71.551	1	.000	.618

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=.000 < α = .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

	Score	df	Sig.
Step 0 Variables Gender(1)	365.887	1	.000
Accompany(1)	17.956	1	.000
AgeGroup(1)	13.715	1	.000
pclass	127.859	2	.000
pclass(1)	102.222	1	.000
pclass(2)	3.377	1	.066
Overall Statistics	457.600	5	.000

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

		Chi-square	df	Sig.
Step 1	Step	499.369	5	.000
	Block	499.369	5	.000
	Model	499.369	5	.000

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

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1241.655 ^a	.317	.431

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

Step	Chi-square	df	Sig.
1	31.752	6	.000

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 none 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^a

Observed			Predicted		
			Survived		Percentage Correct
			Not Survived	Survived	
Step 1	Survived	Not Survived	680	129	84.1
		Survived	155	345	69.0
	Overall Percentage				78.3

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

	B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. for EXP(B)	
							Lower	Upper
Step 1 ^a								
pclass			109.559	2	.000			
pclass(1)	1.824	.175	108.289	1	.000	6.194	4.394	8.733
pclass(2)	.886	.180	24.201	1	.000	2.425	1.704	3.451
Gender(1)	2.506	.149	281.504	1	.000	12.252	9.143	16.418
Accompany(1)	.087	.157	.304	1	.581	1.090	.802	1.484
AgeGroups(1)	1.021	.265	14.853	1	.000	2.775	1.651	4.662
Constant	-2.308	.186	154.636	1	.000	.099		

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

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

$$\text{Logit}(P) = -2.038 + 1.824 \times \text{Pclass1} + .886 \times \text{Pclass2} + 2.506 \times \text{Gender} + 1.021 \times \text{Age group}$$

Findings:-**The chance of survival was apparently related to the below factors:**

- a. Passenger Class (First, Second, Third class)
- b. Age Group (Child or Adult)
- c. Gender (Male, Female)

The observed survival percentage is directly related to:-

- a. Economic status with higher status (first class) associated with higher survival probability
- b. Women had a higher survival rate than men
- c. 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:-

1. Austin, J. T., Yaffee, R. A., & Hinkle, D. E. (1992). Logistic regression for research in higher education. Higher Education: Handbook of Theory and Research, 8, 379–410.
2. Bagley, S. C., White, H., & Golomb, B. A. (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. Journal of Clinical Epidemiology, 54(10), 979-985
3. Beavis, Debbie, 2002. Who on Titanic? The Definitive Passenger Lists. Ian Allen.
4. Beesley, Lawrence, 1912. The Loss of the SS Titanic. In Jack Winocour, ed., The Story of the Titanic as Told by Its Survivors, (1960) Dover.
5. Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 13: Receiver operating characteristic curves. Critical Care (London, England), 8(6), 508-512. <http://dx.doi.org/10.1186/cc3000>
6. Bewick, V., Cheek, L., & Ball, J. (2005). Statistics review 14: Logistic regression. Critical Care (London, England), 9(1), 112-118. <http://dx.doi.org/10.1186/cc3045>
7. Eberhardt, L. L., & Breiwick, J. M. (2012). Models for population growth curves. ISRN Ecology, 2012, 1-7. <http://dx.doi.org/doi:10.5402/2012/815016>

8. Garson, G. D. (2009). "Logistic Regression" from Statnotes: Topics in Multivariate Analysis. Retrieved 6/5/2009 from <http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm>.
9. George, D., & Mallery, P. (2006). SPSS for windows step by step: A simple guide and reference (6th ed.). Boston: Allyn and Bacon.
10. Giancristofaro, R. A., & Salmaso, L. (2003). Model performance analysis and model validation in logistic regression. *Statistica*, 63(2), 375-396.
11. Gleicher D, Stevans LK (2004) Who survived Titanic? A logistic regression analysis. *International Journal of Maritime History* 16:61–94.
12. Hall W (1986) Social class and survival on the S.S. Titanic. *Soc Sci Med* 22:687–690.
13. Harrell, F. E., LEE, K. L., & MARK, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, *Statistics in Medicine*, 15, 361-387.
14. George, D., & Mallery, P. (2006). SPSS for windows step by step: A simple guide and reference (6th ed.). Boston: Allyn and Bacon.
15. Patetta, M. (2002) Categorical Data Analysis Using Logistic Regression Course Notes, Copyright © 2002 by SAS Institute Inc., Cary, NC 27513, USA.