Econ 140 Group Project- The Sinking of the Titanic

Jin Huang, Albert Tsen, Wen-Hsuan Wu

Sinking of the Royal Mail Ship Titanic was one of the greatest fears of all times throughout the history. It was on April 15th, 1912, when the RMS Titanic ocean liner, made in Great Britain, severely clashed into an enormous iceberg, and sank deeply into the ocean. The voyage started from Southampton in England, and the destination was New York City, USA. There was a large number of people onboard, but unfortunately more than one thousand of them did not survive during this tragedy. Regardless of how far back the history is, there are still ongoing research on whether or not certain personal attributes may have contributed to passenger's survivability.

All of our team members have the same curiosity over the likelihood of survivability in the tragedy, thus this paper serves the purpose of analyzing the determinants of survivors. The subject invokes our interest because survival rates vary considerably between individuals. The analysis helps us analyze human behavior in a life-and-death situation. For example, there has been a widespread belief that the social norm of "women and children first" provides women and children a survival advantage over men under emergency circumstances, while others believe that when facing a life-threatening situation, human behavior is captured by the expression of "every man for himself." Some economists argue that "the tendency to act selfishly arises when survival is at stake" (Frey, Savage, and Torgler). In that case, we would expect physically people, especially male adults, to have a higher chance of being saved.

In order to find out the answers, our team dives deeper into the relationships between the likelihood of survivability and all the possible factors influencing the control variable (Y or the dependent variable). In order to obtain unbiased results, we use the dataset created by Thomas E. Cason from the University of Virginia, which is consisted of 1310 samples. However, due to the fact that the deaths of crew and their personal information were not available in the dataset, our team can only predict the survivability of the passengers in our analysis. We brainstormed from different scenarios and thought about the story in various angles, and the factors (X or the independent variables) that we have come up with are passengers' ages, genders, and whether or not they had boarded onto lifeboats. Knowing lifeboats were located close to the first-class cabin, we assume that cabin allocation is a significant explanatory variable that determines survivability. Nonetheless, limited information with regards to the cabin allocation is available from the data, thus we decide not to take it into account in our regression analysis. Another intriguing factor that our team has thought of was whether the passengers wore jeans at the time of the accident. Denim absorbs water and becomes relatively heavy, which might increase a person's risk of drowning. Yet, the lack of reliable data terminates our willingness to invest further, therefore we can not take the factor into account of our data analysis.

To run the regression, we make survivability as a binary dependent variable, with survived being 1 and dead being 0. For independent variables, we break down age variables into 4 different categories: child, teenage, midage, and old, with 20 years in each bracket, and set each of them as a binary variable. Similarly, gender will be our binary independent variable, with female being 1 and male being 0; lifeboat will be another binary independent variable, with on boat being 1 and not on boat being 0.

In the first section, we proceed by providing the background and story of Titanic, different assumptions made by different economists, and how our team will conduct the entire process of finding the relevant answer. In section II, we will explore the relationships by making

a personal assumption and run varias regression analysis with different variables. In section III, we will conclude our overall analysis and result. In the first set of hypothesis, our estimate suggests that teenagers (20-40 years old) have a higher possibility of surviving than people in other age brackets due to higher tolerance to hunger and lower temperature. In the second set of hypothesis, we predict that female has a higher chance to be survived than male, assuming male were yielding female to leave first since it has been rooted in social behavior for long. Next, in order to ensure our prediction is correct, we run regression of survivability on both gender and age, and we exclude the intercept of male and old people as our default variable. Finally, we will run the regression with every single variable to testify whether there is any of them that dominates the other regressors.

2-1 Regression of Survivability on Age Categories

First, we intend to understand how different age brackets affect a person's probability of surviving. Under normal circumstances, we assume people in the 20-40 years old bracket have the highest chance to survive. With a better functional physical mechanism, young people's body can store more excess energy and water than those who are older; this naturally offer them a longer time to confront hunger in any extreme cold weather. Because of this, we can confidently predict that the coefficient of teenage variable will be a lot bigger than the coefficients of others.

Note that it is important to be aware of the existence of dummy variable trap in the analysis. Thus, in order to avoid multicollinearity issue that might provide us a wrong result, we exclude old variable and take that as our default variable for comparison purpose. We will keep it the same in further process of our regression analysis.

```
. regress survived child teenage midage
```

Source	SS	df	MS	Number of obs	=	1,046
				F(3, 1042)	=	2.13
Model	1.53720752	3	.512402505	Prob > F	=	0.0953
Residual	251.152085	1,042	.241028872	R-squared	=	0.0061
				Adj R-squared	=	0.0032
Total	252.689293	1,045	.241807935	Root MSE	=	.49095

survived	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
child	.1711111	.0842435	2.03	0.042	.0058049	.3364173
teenage	.090625	.0802756	1.13	0.259	-.0668953	.2481453
midage	.1097561	.0848615	1.29	0.196	-.0567628	.276275
_cons	.3	.0776255	3.86	0.000	.1476798	.4523202

Based on this regression analysis, we can claim that, on average, people who are 0~59 years old have a slightly higher probability, ranging from 9% ~ 17%, to be survived than those who are above 60 years old. However, we can barely tell if people in their early 20s have a higher chance to be survived than people in their late 40s because the coefficients of three age dummy variables are close to each others. As a result, a better body condition is probably not a crucial factor of life saving. To find out the real reason that has the most impact on survivability,

our team decides to run a regression analysis of survivability on gender and test if, in general, females have a higher chance of survivability relative to their opposite sex.

2-2 Regression of Survivability on Gender

Our team would like to explore the association between survivability and gender. To dive into this topic, we have developed different predictions using various subject theories. As stated before, some economists argue that "the tendency to act selfishly will arise when stakes are high." This theory applies to our case because when survival is at stake, people become even more selfish. On the other hand, sociobiologists think that genetic influences become more powerful under such circumstances, resulting in more women with childbearing age being saved than men. Our team agrees more on sociobiologists' perspective, and we also believe that the social norm: " ladies first" has been rooted in Western society for long.

```
. regress survived female
```

Source	SS	df	MS	Number of obs	=	1,309
				F(1, 1307)	=	507.06
Model	86.3746201	1	86.3746201	Prob > F	=	0.0000
Residual	222.639895	1,307	.170344219	R-squared	=	0.2795
				Adj R-squared	=	0.2790
Total	309.014515	1,308	.236249629	Root MSE	=	.41273

survived	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	.5364832	.0238247	22.52	0.000	.4897445	.583222
_cons	.1909846	.0142151	13.44	0.000	.1630977	.2188715

As from the regression results, the probability of surviving is 53.6% higher for females than for males. Being a female has a highly significant effect on being saved. Even though we still get a low number on the adjusted R square, the result can temporarily demonstrate the social phenomenon of the order "ladies first" and prove that our prediction based on sociobiologists' theory is correct.

2-3 Regression of Survivability on Gender and Age

After running the regression respectively on age and gender, we know younger people have a little bit more chance to be survived and females have a 53.6% higher survival rate than male. Now, we would like to dive deeper into the relationship between survivability, gender and age by testing with all the combinations. Our baseline prediction is that old male will have the least chance to be survived based on what we have assumed above.

```
. regress survived female child teenage midage
```

Source	SS	df	MS	Number of obs	=	1,046
				F(4, 1041)	=	106.45
Model	73.354822	4	18.3387055	Prob > F	=	0.0000
Residual	179.334471	1,041	.172271345	R-squared	=	0.2903
				Adj R-squared	=	0.2876
Total	252.689293	1,045	.241807935	Root MSE	=	.41506

survived	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	.5451146	.026698	20.42	0.000	.4927266	.5975026
child	.0714763	.071388	1.00	0.317	-.0686045	.2115571
teenage	.0503092	.0678952	0.74	0.459	-.0829179	.1835364
midage	.0655486	.0717762	0.91	0.361	-.0752938	.2063911
_cons	.1500935	.0660355	2.27	0.023	.0205156	.2796714

Since all the coefficients are greater than zero, we know that old male population has the least chance of being survived. However, since the adjusted R square is still pretty small (0.2876), we cannot claim that this regression fully reflects the actual picture. Thus, we further develop our analysis by taking another regression by adding onboat variable into our model. We want to find out if successfully boarding a lifeboat is the most influential factor.

[Note: If we were to compare the analysis of the last two datasets above, we would notice that the coefficients of the female have only changed slightly. This strongly explains that genders have nothing to do with the age brackets. In other words, there is no correlation between genders and age brackets.]

```
. regress survived female child teenage midage onboat
```

Source	SS	df	MS	Number of obs	=	1,046
				F(5, 1040)	=	2111.11
Model	230.025737	5	46.0051475	Prob > F	=	0.0000
Residual	22.663555	1,040	.02179188	R-squared	=	0.9103
				Adj R-squared	=	0.9099
Total	252.689293	1,045	.241807935	Root MSE	=	.14762

survived	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	.0933809	.010888	8.58	0.000	.0720159	.1147459
child	.0187599	.0253978	0.74	0.460	-.0310769	.0685967
teenage	.00453	.024154	0.19	0.851	-.0428661	.0519261
midage	.0201858	.0255339	0.79	0.429	-.0299179	.0702896
onboat	.9086189	.0107161	84.79	0.000	.8875914	.9296465
_cons	.0017346	.0235516	0.07	0.941	-.0444795	.0479486

After including the variable onboat in our regression, we notice that the coefficient of onboat is 0.9 and the adjusted R-squared is 0.9 as well. Since the coefficient of onboat is much

larger than any other variable and adjusted R-squared is large enough, onboat seems to be our dominant variable of surviving. The model is nearly perfect because 90% of variability of the dependent data can be explained by the change in independent variables.

```
. regress survived female onboat
```

Source	SS	df	MS	Number of obs	=	1,309
				F(2, 1306)	=	6051.33
Model	278.916589	2	139.458295	Prob > F	=	0.0000
Residual	30.0979257	1,306	.023045885	R-squared	=	0.9026
				Adj R-squared	=	0.9025
Total	309.014515	1,308	.236249629	Root MSE	=	.15181

survived	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	.0940952	.0100109	9.40	0.000	.0744561	.1137343
onboat	.9072141	.0099253	91.40	0.000	.8877428	.9266854
_cons	.0123398	.0055819	2.21	0.027	.0013893	.0232903

To make sure, again, that gender doesn't play the biggest role in our model, we only select gender and onboat as our independent variables this time. We notice that the coefficient of female decreased significantly from 0.53 to 0.09. After realizing the difference in the coefficient of genders, we conclude that without the variable onboat, gender captures too much of the effect, indicating that conditionally being on boat, gender does not affect survivability.

```
. regress survived onboat
```

Source	SS	df	MS	Number of obs	=	1,309
				F(1, 1307)	=	11261.70
Model	276.880554	1	276.880554	Prob > F	=	0.0000
Residual	32.1339606	1,307	.024586045	R-squared	=	0.8960
				Adj R-squared	=	0.8959
Total	309.014515	1,308	.236249629	Root MSE	=	.1568

survived	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
onboat	.9523171	.0089739	106.12	0.000	.9347123	.9699218
_cons	.0291262	.0054624	5.33	0.000	.0184103	.0398422

Knowing gender captures too much of the effect before we include the variable onboat, we have come to a conclude that onboat is the most reasonable and the most influential factor in explaining survivability. The probability of surviving is 95% higher for those boarded onto a lifeboat than for those who did not. "Failure to secure a seat virtually guaranteed death since the average ocean temperature was about 2 degrees Celsius (35 degrees Fahrenheit), just above the freezing point; any survivors of the sinking vessel left in the water would have quickly frozen to death" (Savage).

In order to measure the overall significance, we performed a hypothesis test.

survived = $\beta_0 + \beta_1$female + $\beta_2$child + $\beta_3$teenage + $\beta_4$midage + $\beta_5$onboat

H0: β1 = β2 = β3 = β4 = β5 = 0

H1: one of the restrictions does not hold

F-stat = (0.91^2 / 5) / [(1-0.91^2) / (1046-5-1)] = 1002

F0.05, (5, 1000) = 2.22

We reject the null hypothesis that the variables do not affect survivability. The variables are important factors in our regression.

From the regression results, it is clear that the survival rate of women is more than that of men, with accounts documenting many instances where "men stood back to provide women and children safety priority when boarding lifeboats." In fact, there were "reports of officers using guns to uphold Captain Smith's order to save women and children first" (Connor). Therefore, we are able to conclude that the sinking of the Titanic supports the social norm of "women and children first." In addition to that, we know that boarding onto a lifeboat or not significantly impact the surviving rate. With that said, the extent of the tragedy is mainly due to too few lifeboats on the Titanic. The only 20 available lifeboats were definitely not enough to accommodate two thousand passengers and crew members. Not only does the sinking of the Titanic reveal the aspects of human nature in a life-threatening situation, it also points out the extreme arrogance of the creators who believed the ship was "unsinkable" thus there was no need to plan ahead for vital disasters. In the wake of the accident, improvements to passengers' safety has been made for modern ships, including enough number of life jackets and lifeboats for the entire number of passengers and crew. The history of the Titanic should be remembered, and the lesson must be learned.

References

Cason, Thomas E. "Titanic Dataset, v3.5." 16 Apr. 1999, biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic3.xls. University of Virginia

Connor, Steve. "Women and Children First? It's Every Man for Himself on a Sinking Ship." *The Independent*, Independent Digital News and Media, 30 July 2012, www.independent.co.uk/news/world/world-history/women-and-children-first-its-every-man-for-himself-on-a-sinking-ship-7987975.html.

Savage, David. *ECONOMICS OF MARITIME DISASTERS: ESSAYS ON THE TITANIC AND LUSITANIA*. eprints.qut.edu.au/35730/1/David_Savage_Thesis.pdf.

Frey, Bruno S., et al. "Surviving the Titanic Disaster: Economic, Natural and Social Determinants." *Center for Research in Economics, Management and the Arts*, www.crema-research.ch/papers/2009-03.pdf.