Surviving the Titanic: A Statistical Analysis of Age, Sex, and Class of Passengers

By: Hannah Brown, Abby Fruzyna, Jude Piltingsrud, and Greyson Willhite

Introduction

RMS Titanic was most commonly known for its grand scale, elegant features, and its disastrous sinking in 1912. During the beginning of this fateful trip, the Titanic held over 2,200 passengers and crew; however, only roughly 700 survived the catastrophe, leaving over 1,500 people to perish in icy waters.[1] When disasters like these strike, many people hold the assumption that women and children must be saved first, and have observed this notion when learning about the Titanic or from various movies and TV shows. Contrary to popular belief, the sinking of the Titanic was not the origin of this iconic phrase and routine protocol. Roughly 60 years before the Titanic, the HMS Birkenhead sailed the waters while carrying 638 people. The iron-hulled ship struck rocks and ultimately perished, but not before the colonel gave out an order to his soldiers to not jump and remain on the ship to avoid endangering the few lifeboats beneath them.[2] By the time rescue boats arrived only 193 people remained alive, but all 26 women and children aboard had been saved.[2] This event gave rise to the "Birkenhead Drill," meaning that during any sinking, women and children are to go first.[2]

With this procedure popularized before the Titanic, the passengers and crew would have been familiar with it, but was it followed? Often when facing a traumatic event, people hold to the belief of 'every man for himself' or 'every person for oneself.' So, how likely is it that the passengers and crew of the Titanic followed this procedure? Therefore, we are trying to

investigate whether or not this protocol of “save the women and children first” was truly followed during the sinking. There are two possible hypotheses that we will be testing:

1. The protocol, or the Birkenhead Drill, was followed. The survival rates were influenced by a passenger's age, gender, and socioeconomic status, with women and children statistically significantly more likely to survive.
2. The protocol, or the Birkenhead Drill, was not followed. Age and gender did not play a role in the survival rates of the passengers.

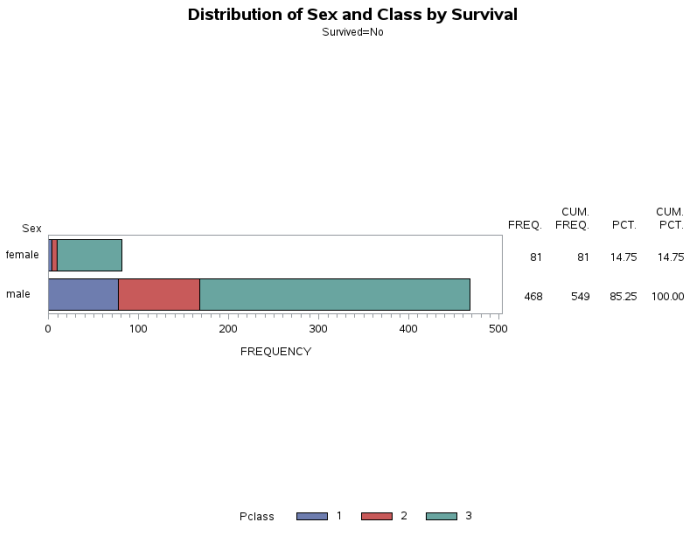
The questions posed above and the hypotheses to be tested will be addressed by drawing from the ‘Titanic Survive Model’ data set. This specific data set consists of 12 variables and 891 passengers who were on the Titanic. For our purposes, we decided to limit the variables to socioeconomic class, sex, age, and fare. The socioeconomic class variable details whether the passenger was 1st, 2nd, or 3rd class, with gender being either male or female, and age ranging from adolescence to late adulthood. We emitted variables such as passenger ID and name, where they boarded, how many children they had, and ticket number.

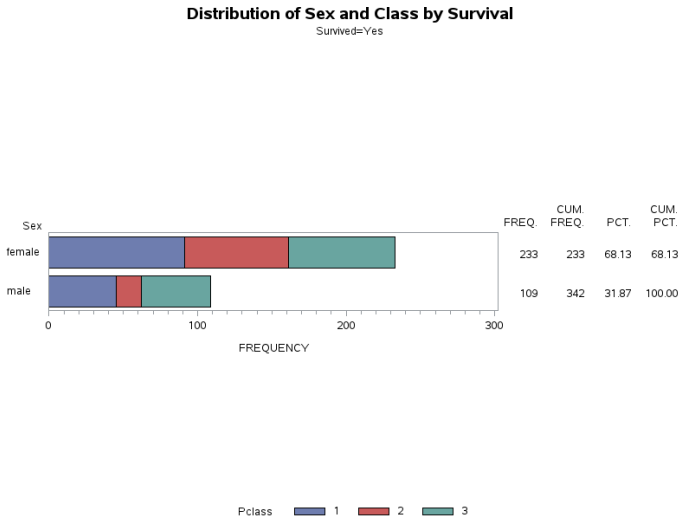
Analysis

1.1 Sex and Socioeconomic Class

Firstly, we decided to use a stacked bar chart to visualize the difference between males and females, whether their survival rates were relatively equal or higher in one group over the other. In the results titled “Distribution of Sex and Class by Survival,” a disparity between the two sexes becomes clear, showcasing a much higher rate of female survival and male death (68.13% of survivors were female and 82.25% of casualties were male). This is supported by the

data tables, which show a 74.20% survival rate for females and an 18.89% survival rate for males. Thus, supporting the hypothesis that the Titanic followed the protocol.





Secondly, the breakdown of males that survived by class (total male survival = 18.89%) showed that the percentage of surviving men that were 2nd class (2.95%) was less than half the amount of the 3rd class (8.15%) and 1st class (7.80%). These discrepancies are caused by the much smaller sample size for the 2nd and 1st classes, n = 122 and 108, respectively, in comparison to the 3rd class, n = 347. When looking at the percentage of men in each class that survived, a more accurate distribution is found, with 36.89% of 1st class men, 15.74% of 2nd class men, and 13.54% of 3rd class men surviving. This class bias is seen in female survival rates as well, with 96.81% of 1st class women, 92.11% of 2nd class women, and 50.00% of 3rd class women surviving. There is a large deviation in the difference of survival rate between sexes for 3rd class passengers, showing a less extraneous difference of 36.46% in survival between men and women. Very small in comparison to the 1st and 2nd classes, having a 59.92% and 74.37%

difference in survival, respectively. See graphs titled “Distribution of Survival and Class by Sex” below.

Distribution of Survival and Class by Sex

The FREQ Procedure

Sex=female

Frequency Percent Row Pct Col Pct	Table of Survived by Pclass			
	Pclass			
Survived	1	2	3	Total
No	3 0.96 3.70 3.19	6 1.91 7.41 7.89	72 22.93 88.89 50.00	81 25.80
Yes	91 28.98 39.06 96.81	70 22.29 30.04 92.11	72 22.93 30.90 50.00	233 74.20
Total	94 29.94	76 24.20	144 45.86	314 100.00

Distribution of Survival and Class by Sex

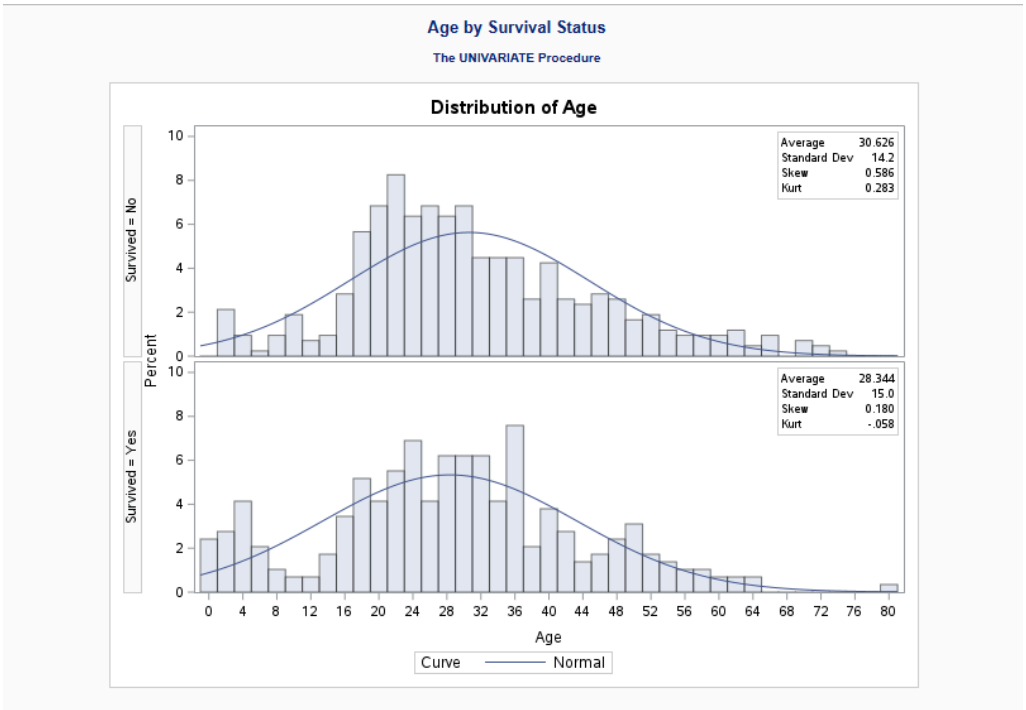
The FREQ Procedure

Sex=male

Frequency Percent Row Pct Col Pct	Table of Survived by Pclass			
	Pclass			
Survived	1	2	3	Total
No	77 13.34 16.45 63.11	91 15.77 19.44 84.26	300 51.99 64.10 86.46	468 81.11
Yes	45 7.80 41.28 36.89	17 2.95 15.60 15.74	47 8.15 43.12 13.54	109 18.89
Total	122 21.14	108 18.72	347 60.14	577 100.00

1.2 Age

Next, age and survival status were looked at to determine if age played a role in the chances of survival. First, in the graph titled “Age by Survival Status,” we decided to look at the age averages of those who survived and those who did not. It appears that the average age of those who did not survive was only a few years greater than the age of those who did (avg = 30.626 and avg = 28.344, respectively). However, there appears to be a wide variability in age seen with the standard errors, with kurtosis and skew suggesting that the age distributions of non-survivors and survivors do not appear to be statistically different.

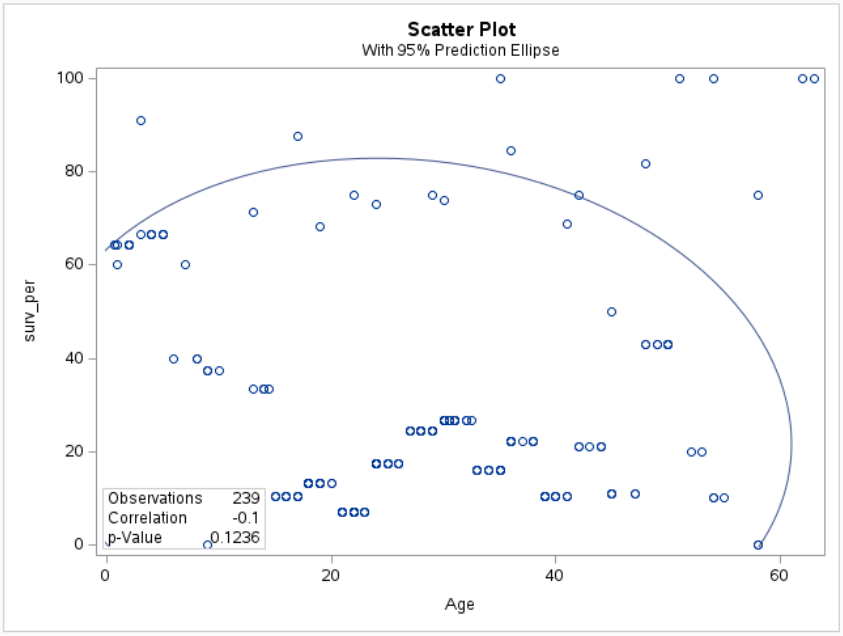


To confirm this, scatter plots of age and percent survival, separated by sex, were made to visualize and quantify the correlation between these variables. See graphs titled “Age vs Percent Survival” below. The female plot produced a correlation coefficient of -0.1 (p value = 0.1236), indicating a negligible or very weak correlation between age and percent survival. The male plot produced a correlation coefficient of -0.358 (p value = <.0001), indicating a weak negative correlation between age and percent survival. This weak correlation for males is likely caused by the higher percentage of survival seen in males 15 and younger, following the societal obligation to save women and children first.

Age vs Percent Survival

The CORR Procedure

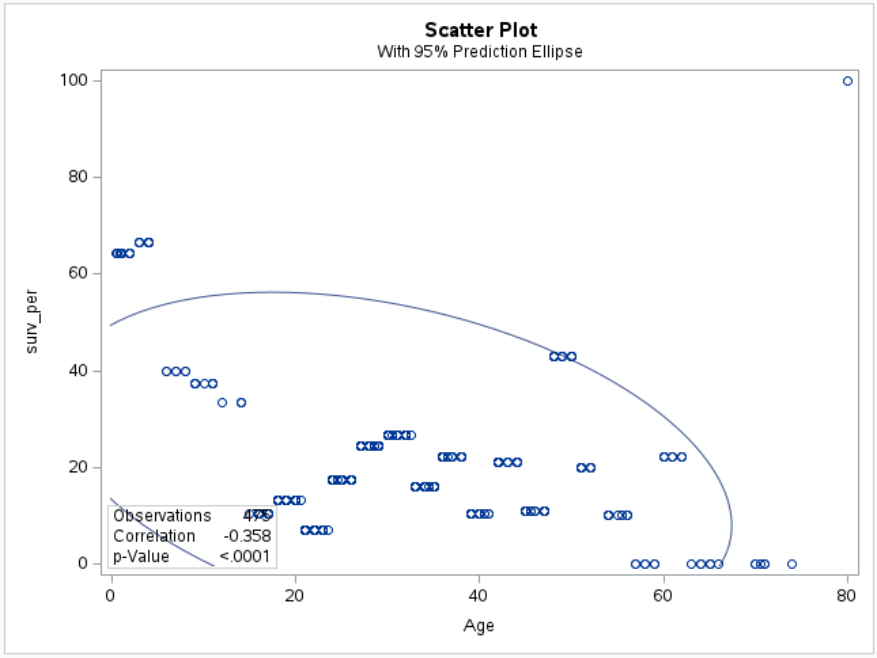
Sex=female



Age vs Percent Survival

The CORR Procedure

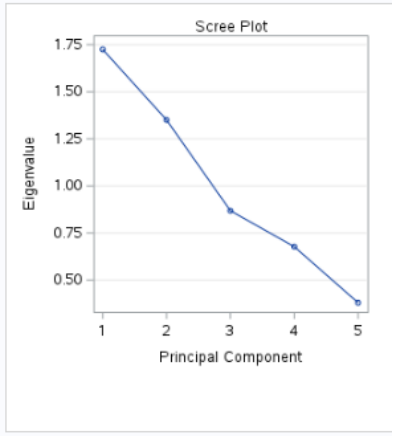
Sex=male



1.4 Logistic Regression and PCA

Here we will dive deeper into our project by building a logistic regression model to predict the passenger survival rate based on the attributes within our dataset. This started with some basic pre-processing to get the data into shape for our model to be able to work efficiently. This included some one-hot encoding for our binary variables sex and survived, replacing sex values of ‘male’ with zero and ‘female’ with one; for survival, we defined ‘yes’ as one and ‘no’ as 0. The next step was to deal with missing values. Initially, we would look to replace missing values with either their mode or mean. Trying mode we found that this had a heavy influence on the distributions of our attributes with missing values, while replacing with the mean values left us with a dataset more accurate to our initial distributions.

The first statistical output we decided to use was PCA (Principal Component Analysis). This is a method often used in dimensionality reduction to increase robustness and accuracy in the model without sacrificing the global structure of the relationships between the attributes. As this is an unsupervised learning technique, we will not consider the survived attribute in the test, as it is our target variable. PCA develops a correlation matrix representing the linear combinations of the variables that capture the most variance within the data. We looked at the individual amounts of variance explained by each component by looking at the cumulative eigenvalues of the correlation matrix along with the scree plots, and were able to determine that four principal components were able to capture 92.41% of the variance. The first component was sex, which captures 34.5% of variance. The second is Fare, capturing about 27.01% of variance. The third is Pclass, capturing 17.36% of variance. The fourth is Age, capturing 13.54% of variance. The fifth component, which is SibSp, only captured about 7.59% of variance. For the sake of robustness and accuracy of the model, we chose to exclude SibSp from our models.



The PRINCOMP Procedure

Observations	891
Variables	5

Simple Statistics

	sex_encoded	Fare	Pclass	Age	SibSp
Mean	0.6475869809	32.20420797	2.308641975	29.69911765	0.523007856
StD	0.4779900709	49.69342860	0.836071241	13.00201523	1.102743432

Correlation Matrix

	sex_encoded	Fare	Pclass	Age	SibSp
sex_encoded	1.0000	-.1823	0.1319	0.0842	-.1146
Fare	-.1823	1.0000	-.5495	0.0916	0.1597
Pclass	0.1319	-.5495	1.0000	-.3313	0.0831
Age	0.0842	0.0916	-.3313	1.0000	-.2326
SibSp	-.1146	0.1597	0.0831	-.2326	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.72491457	0.37452059	0.3450	0.3450
2	1.35039398	0.48234335	0.2701	0.6151
3	0.86805062	0.19085350	0.1736	0.7887
4	0.67719713	0.29775342	0.1354	0.9241
5	0.37944371		0.0759	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
sex_encoded	-.228121	0.452971	0.826547	-.242503	-.028114
Fare	0.605583	-.267608	0.245142	-.307932	0.637760
Pclass	-.669572	-.077766	-.065896	0.177219	0.714056
Age	0.363765	0.543531	0.022091	0.722512	0.223019
SibSp	-.024028	-.649406	0.501892	0.541240	-.181268

Now that our data was properly formatted and we had done some further analysis of our variables with PCA, it was time to start building our models. The initial idea was to include all four of our principal components and take a look at our results from there. The first model built using all four variables displayed the relationship between sex and survival rate, as it was our first principal component. The equation for the first model is Y = 4.6641 - 2.6071X1 + 0.000577X2 - 1.1501X3 - 0.0332X4 (B1=sex_encoded, B2=Fare, B3=Pclass, B4=Age). Shown below by the graph of predicted probabilities for survival based on sex, there is a clear relationship between sex and survival rate (P1). Our model predicted that women had a much higher survival rate compared to men, which is backed by our initial exploratory data analysis.

This will also be considered later in our conclusion when talking about our initial hypothesis. Another takeaway from this model was our ROC curve, which looks great; the area under the curve is .8478, which is considered to be quite high (graph R1). This means that our model is doing a good job at classifying a passenger's survival based on our coefficients within the model. However, there was one issue, and this was found when looking at the chi-sq values in our ANOVA table for our model, which will be included below (Anova 1). Fare, which was our second principal component, had a very low Wald-Chi Square value (0.0805), followed by our pr>chi-sq value being 0.7767, which is very high. This leads us to believe that Fare is not a statistically significant factor when predicting survival, which is likely due to variable Pclass already capturing a lot of the effect that fare may have. Another indicator is that the 95% wald confidence limits contain 1 (0.997, 1.005). Before dropping fare as a variable, there was another necessary step. We decided to look at a model without Pclass, but with Fare; this model ended up having a worse ROC curve, so we decided to move forward with using Pclass, Age, and Sex_encoded as our predictors.

The next model, just using those three predictors, had an equation of Y = 4.7319 - .0334X1 - 2.6120X2 - 1.1685X3 (B1=age, B2=sex_encoded, B3=Pclass)(Anova 2). Our ROC curve stayed the same as our first model, further backing our decision that fare was not statistically significant (Graph R1). Our predicted probability survival graph for this model considered the survival rate by age (Graph P2). Here we saw another clear relationship between age and probability of survival; when age increases, the survival rate decreases significantly. This relates back to our initial hypothesis, showing that it is likely true that Children were among the first on the lifeboats. However we must look at other factors before concluding this blanket statement to be true, in a survival scenario like this we must consider that younger more able

bodied bodied people may be more likely to survive due to them being in better shape and better health, not necessarily just because they were more likely to get a spot in the lifeboat. Using the same model, the last plot we looked at showed the predicted probability of survival based on Pclass. Here we see another relationship: the lower the class (referring to Pclass =3), the lower the survival rate. Considering some situational factors, we think this makes perfect sense. The lower classes were positioned lower in the boat, likely leaving them farther away from the lifeboats, and first at risk of the filling vessel (Graph P3).[3] Likely, by the time the people in the lower classes residing in the lower levels of the boat were able to reach the higher levels where the lifeboats were positioned, they were already filled, thus decreasing their survival rate. Through building these models, we find ourselves with a solid understanding of the dataset and some evidence to further back our hypotheses.

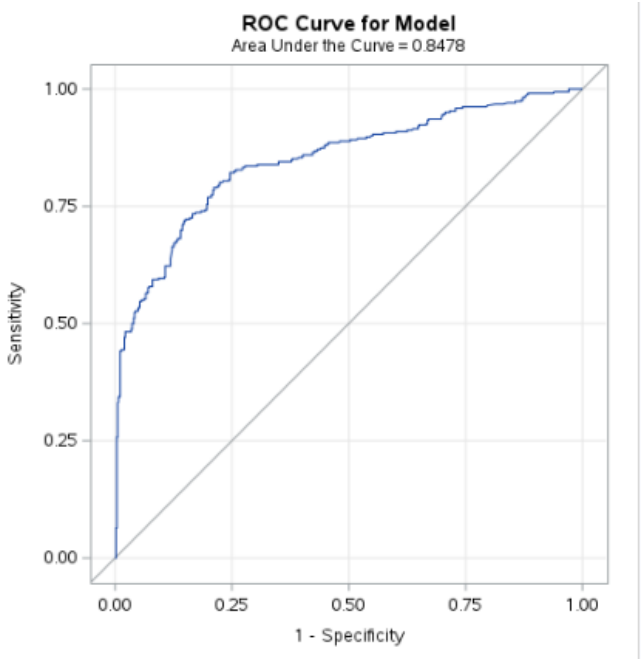
APPENDIX FOR LINEAR REGRESSION SECTION BELOW

APPENDIX FOR 1.4

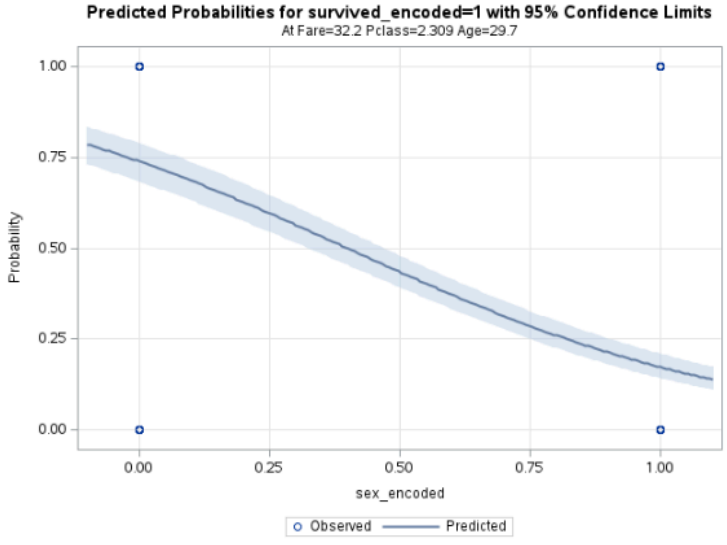
Anova 1

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	4.6641	0.5089	84.0119	<.0001
sex_encoded	1	-2.6071	0.1874	193.6181	<.0001
Fare	1	0.000577	0.00204	0.0805	0.7767
Pclass	1	-1.1501	0.1352	72.3248	<.0001
Age	1	-0.0332	0.00738	20.3085	<.0001

Graph R 1



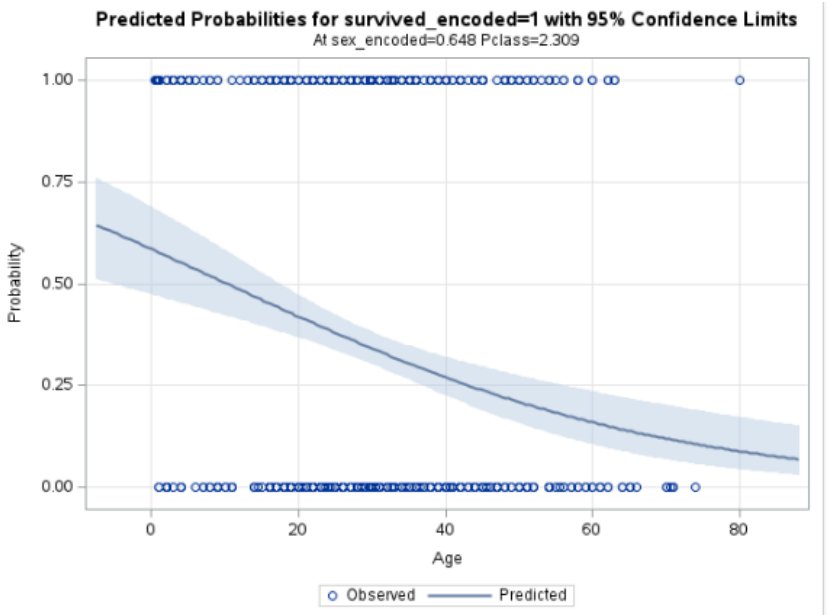
Graph P 1



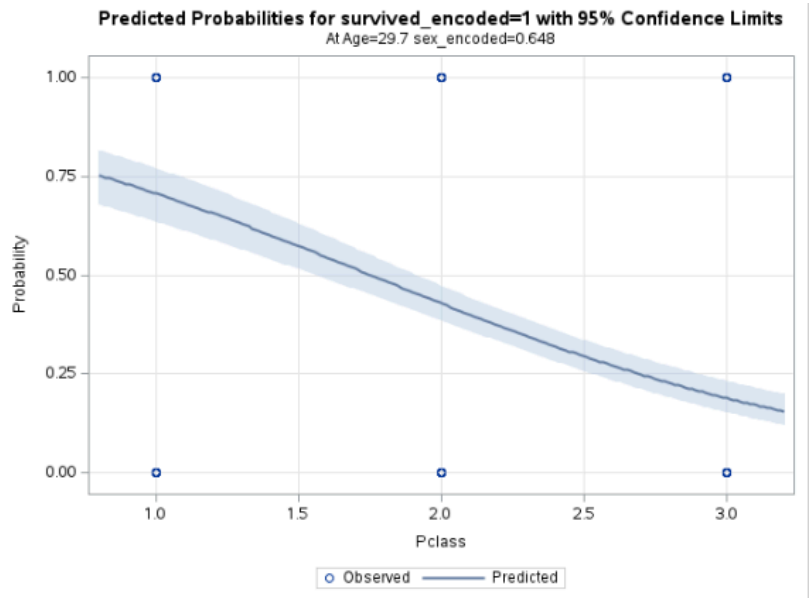
Anova 2

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	4.7319	0.4498	110.6634	<.0001
Age	1	-0.0334	0.00735	20.6968	<.0001
sex_encoded	1	-2.6120	0.1866	195.9157	<.0001
Pclass	1	-1.1685	0.1189	96.5091	<.0001

Graph P2



Graph P3



Conclusion

Our analysis revealed statistically significant insights about survival of different groups in the Titanic crisis. Female passengers had a significantly higher survival rate than males, aligning with the notion of the Birkenhead Drill by prioritizing the lives of women over men. Age also played a significant role, with female survivors not having correlation and male survivors having slight negative correlation. With the use of a logistic model, a clear relationship between age and survival is seen, once again aligning with the Birkenhead Drill protocol. The slight correlation seen with age and male survival can be hypothesized to be due to a confounding variable such as physical health, but the lack of any correlation in the female survivors doesn't provide support for this hypothesis. Class played a large role in survival as well, heavily correlated in the female population with higher classes having higher survival probability, and decently weighted in favor of 1st and 2nd class survival in the male population. This analysis supports the hypothesis that age, sex, and class were statistically significant predictors of survival. Limitations to the study include potential confounding variables like physical health, group dynamics, and lifeboat proximity. These limitations emphasize the importance of attributing survival predictions with caution. Understanding the factors that influence survival during a tragedy like the Titanic provides important information on human behavior during a crisis and may assist progress in developing effective safety protocols to increase overall survival probability. Future research could investigate possible confounding variables in this study to determine statistical significance and further develop understanding of survival probability.

Citations

(1) Munson, O. How many people died on the Titanic? Facts about the death toll and the survivors. USA Today, (2022). https://www.usatoday.com/story/news/2022/11/19/how-many-people-died-titanic-how-many-survived/10605754002/

(2) National Army Museum.Women and children first. https://www.nam.ac.uk/explore/birkenhead-sinking

(3) Malatin, T. 107 #72: Third class passengers were kept below as Titanic sank and were prevented from entering the lifeboats. (2019). https://timmaltin.com/2019/04/23/third-class-passengers-titanic/

Code

```
data titanic;
    infile "/home/u64127570/sasuser.v94/titanic-passengers.csv" dlm=";" dsd missover firstobs=2;
*most finicky dataset ever;
    length Name $ 60;
    input PassengerId : Survived $ : Pclass : Name $ : Sex $ : Age : SibSp : Parch : Ticket $ :
Fare; *keeps everything seperated so missing values can be dealt with;
run;
data titanic2; *dealing with missing values, replacing with -99 to easily sort out when using;
    set titanic;
    if Age = "" then Age = -99;
    if SibSp = "" then SibSp = -99;
    if Parch = "" then Parch = -99;
```

```
    if Fare = "" then Fare = -99;
run;
*sorted data set by survival;
proc sort data = titanic2 out= tit_surv;
            by Survived;
run;
data tit_surv01;
    set tit_surv;
    if Survived = 'Yes' then SurvivedNum = 1;
    else if Survived = 'No' then SurvivedNum = 0;
run;
*sorted data set by sex;
proc sort data = titanic2 out= tit_sex;
            by Sex;
run;
*sorted data set by pclass;
proc sort data = titanic2 out= tit_class;
            by Pclass;
run;
data tit_class;
    set tit_class;
    if Survived = 'Yes' then surv_num = 1;
    else if Survived = 'No' then surv_num = 0;
```

```
run;
*sorted data set by fare;
proc sort data = titanic2 out= tit_fare;
        by Fare;
run;
data fare_groups;
        set tit_fare;
        length faregroup $10;
    do i = 0 to 520 by 5; *sas for loop;
       if fare >= i and fare < i + 5 then faregroup = cats(i, '-', i + 5);
    end;
run;
*changing from qual to quant data;
data fare_groups;
    set fare_groups;
    if Survived = 'Yes' then surv_num = 1;
    else if Survived = 'No' then surv_num = 0;
run;
*repeating the above for age;
data age_groups;
        set titanic2;
        where Age ne -99;
        length agegroup $10;
```

```
    do i = 0 to 90 by 3; *sas for loop;
        if age >= i and age < i + 3 then agegroup = cats(i, '-', i + 3);
    end;
run;
data age_groups;
    set age_groups;
    if Survived = 'Yes' then surv_num = 1;
    else if Survived = 'No' then surv_num = 0;
run;
/* HBAR and VBAR charts */
proc gchart data=tit_surv;
            title "Distribution of Age and Class by Survival";
    by Survived;
    vbar Age/ group= Pclass; *vbar version- I prefer this one, easier to see the differences in
distrubution;
    where Age ne -99;
run;
/* Stats */
proc freq data = tit_sex ;
            title "Distribution of Survival and Class by Sex";
            table Survived*Pclass;
            by Sex;
run;
```

```
proc univariate data = tit_surv noprint; * this makes 2 graphs, survived(yes) and survived(no);
        title "Age by Survival Status";
        class Survived;
        where Age ne -99;
        histogram Age/normal midpoints= (0 to 81 by 2);
        inset mean= "Average"(6.3) std="Standard Dev"(4.3) skewness="Skew"(5.3) kurtosis =
"Kurt"(5.3)  /pos = NE;
run;
proc sort data = age_groups;
        by sex;
run;
proc means data=age_groups noprint;
    class agegroup;
    var surv_num;
    by sex;
    output out=summary sum=surv_c n=tot_c;
run;
data per;
    set summary;
    if _TYPE_ = 1;*total per group instead of entire dataset;
    surv_per= (surv_c/tot_c) * 100;
run;
*got sorting errors even tho it should be sorted, resorting;
```

```
proc sort data=age_groups;
    by ageGroup;
run;
proc sort data=per;
    by ageGroup;
run;
data age_groups_per;
    merge age_groups(in=a) per(in=b); *merge;
    by agegroup;
    if a; *only age_groups vars kept;
run;
proc sort data = age_groups_per;
        by sex;
run;
proc corr data = age_groups_per plots=scatter;
        title "Age vs Percent Survival";
        where Age ne -99;
        var age surv_per;
        by sex;
run;
data titanic_dummy;
set titanic;
run;
```

```
/* Data pre processing, replaced missing age values for the mean, used one-hot encoding to
transform categorical variables such as survival and sex to binary attributes so the model can
properly evaluate them*/
PROC MEANS DATA=titanic_dummy NOPRINT;
 VAR age;
 OUTPUT OUT=mean_age MEAN=mean_age;
RUN;
DATA titanic_imputed;
 IF _N_ = 1 THEN SET mean_age;
 SET titanic_dummy;
 IF age = . THEN age = mean_age;
RUN;
DATA titanic_encoded;
 SET titanic_imputed;
 IF sex = 'male' THEN sex_encoded = 1;
 ELSE sex_encoded = 0;
RUN;
Data full_encoded;
	set titanic_encoded;
	if Survived = 'Yes' then survived_encoded = 1;
	else survived_encoded = 0;
run;
/* ran PCA here to determine how many components to use for logistic regression model.
```

age was the first principle component, accounting for about 35% of covariance based on passenger surviving

You can see in the covariance matrix that age, fare, pclass, and sex account for 90% of variance (these four variables

account for roughly 91% of the variability) which is a good threshold to evaluate the model at so I only used

those four in the model. */

```
proc princomp data=full_encoded out=pca_out ;
   var  sex_encoded fare Pclass Age sibsp;
run;
/* logistic model displaying sex, here we notice that fare is statistically insignificant in the model*/
PROC LOGISTIC DATA=full_encoded plots = (effect roc);
 MODEL survived_encoded (EVENT='1') = sex_encoded age Pclass fare ;
 output out=pred_data pred=predicted;
RUN;

/*  logistic regression model displaying age*/
proc logistic data=full_encoded plots = (effect);
 model survived_encoded (event='1') = age sex_encoded Pclass ;
 output out=pred_data pred=predicted;
run;
/* logistic model displaying Pclass*/
```

```
proc logistic data=full_encoded plots = effect;
 model survived_encoded (EVENT='1') =  Pclass age sex_encoded ;
 output out=pred_data2 pred=predicted2;
run;
```