# MATH 4044 – Statistics for Data Science

## Practical Week 13 Solutions

### Question 1

The sinking of the Titanic is a famous event. Many well-known facts – from the proportions of first-class passengers to the 'women and children first' policy, and the fact that that policy was not entirely successful in saving the women and children in the third class – are reflected in the survival rates for various classes of passengers. These data were originally collected by the British Board of Trade in their investigation of the sinking. There is no complete agreement among primary sources as to the exact numbers on board, the number rescued, and the number lost.

The data for this question is stored in a SAS data file called `titanic.sas7bdat`. For each person on board the fatal maiden voyage of the ocean liner Titanic, this dataset records gender, age [adult/child], economic status [first/second/third class, or crew] and whether or not that person survived. Specifically, variables in the data file are as follows:

| Variable | Description |
|---|---|
| *Class* | 0 = 'crew', 1 = 'first', 2 = 'second', 3= 'third' |
| *Age* | 1 = 'adult', 0 = 'child' |
| *Gender* | 1 = 'male', 0 = 'female' |
| Survived | 1 = 'Yes', 0 = 'No' |

**[Source:** *Journal of Statistics Education* data archive.]

(a) Obtain mosaic plots of Survived versus Gender, Age and Class. Also consider tests of independence. Comment on chances of survival by gender, age and class. How well did the 'women and children first' policy work?

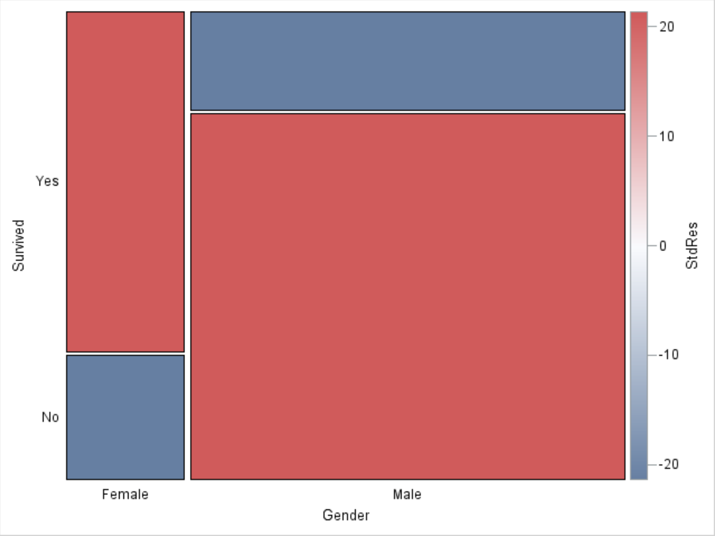


**Figure 1.** Mosaic plot of Survived versus Gender

The mosaic plot in Figure 1 indicates a higher survival rate for females than for males. More specifically, there was a much higher than expected number of female passengers who survived the disaster. In contrast, there was a much lower than expected number of male survivors. This is confirmed by counts and percentages reported in Table 1. From Table 2, the chi-squared statistic for the test of independence is 456.87 (d.f. = 1) with the corresponding *P*-value < 0.001, confirming a highly statistically significant relationship between survival and gender.

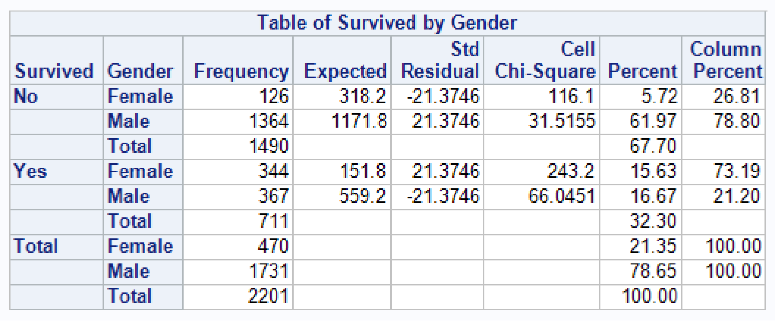
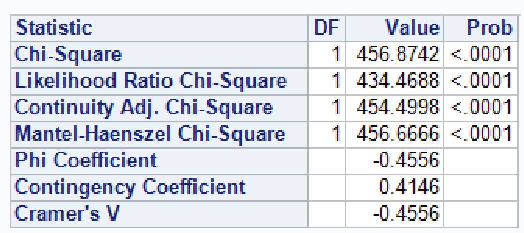


Table of Survived by Gender

| Survived | Gender | Frequency | Expected | Std Residual | Cell Chi-Square | Percent | Column Percent |
|---|---|---|---|---|---|---|---|
| No | Female | 126 | 318.2 | -21.3746 | 116.1 | 5.72 | 26.81 |
| | Male | 1364 | 1171.8 | 21.3746 | 31.5155 | 61.97 | 78.80 |
| | Total | 1490 | | | | 67.70 | |
| Yes | Female | 344 | 151.8 | 21.3746 | 243.2 | 15.63 | 73.19 |
| | Male | 367 | 559.2 | -21.3746 | 66.0451 | 16.67 | 21.20 |
| | Total | 711 | | | | 32.30 | |
| Total | Female | 470 | | | | 21.35 | 100.00 |
| | Male | 1731 | | | | 78.65 | 100.00 |
| | Total | 2201 | | | | 100.00 | |

**Table 1.** Frequency table for Survived by Gender



| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 456.8742 | <.0001 |
| Likelihood Ratio Chi-Square | 1 | 434.4688 | <.0001 |
| Continuity Adj. Chi-Square | 1 | 454.4998 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 456.6666 | <.0001 |
| Phi Coefficient | | -0.4556 | |
| Contingency Coefficient | | 0.4146 | |
| Cramer's V | | -0.4556 | |

**Table 2.** Chi-square test results for Survived versus Gender

The mosaic plot in Figure 2 representing the relationship between survival and age indicates that there were relatively few children on board, and their survival rate was higher than that of adults. More specifically, there was a higher than expected number of children who survived, and a much lower than expected number of adults who survived, confirmed by counts and percentages reported in Table 3. From Table 4, the chi-squared statistic for the test of independence is 20.96 (d.f. = 1) with the corresponding *P*-value < 0.001, confirming a highly statistically significant relationship between survival and age.

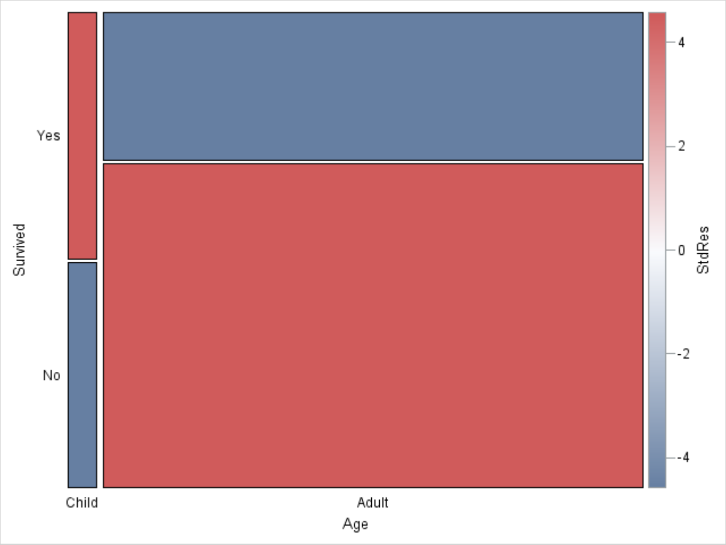


**Figure 2.** Mosaic plot of Survived versus Age

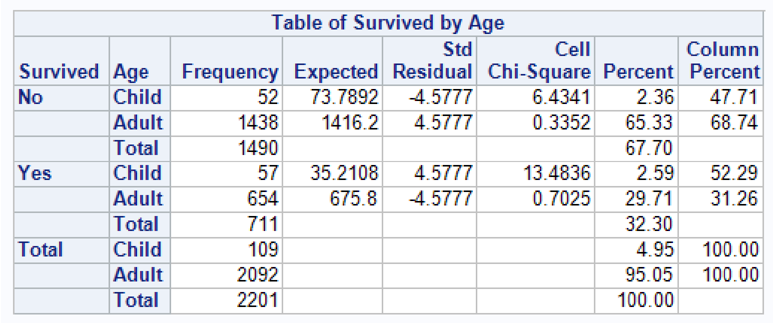
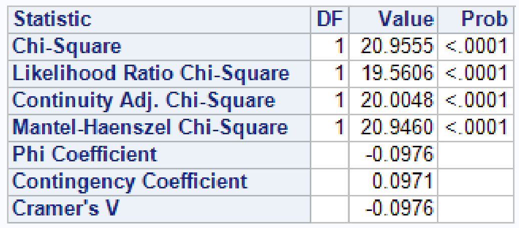


Table of Survived by Age

| Survived | Age | Frequency | Expected | Std Residual | Cell Chi-Square | Percent | Column Percent |
|---|---|---|---|---|---|---|---|
| No | Child | 52 | 73.7892 | -4.5777 | 6.4341 | 2.36 | 47.71 |
| | Adult | 1438 | 1416.2 | 4.5777 | 0.3352 | 65.33 | 68.74 |
| | Total | 1490 | | | | 67.70 | |
| Yes | Child | 57 | 35.2108 | 4.5777 | 13.4836 | 2.59 | 52.29 |
| | Adult | 654 | 675.8 | -4.5777 | 0.7025 | 29.71 | 31.26 |
| | Total | 711 | | | | 32.30 | |
| Total | Child | 109 | | | | 4.95 | 100.00 |
| | Adult | 2092 | | | | 95.05 | 100.00 |
| | Total | 2201 | | | | 100.00 | |

**Table 3.** Frequency table for Survived by Age



| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 1 | 20.9555 | <.0001 |
| Likelihood Ratio Chi-Square | 1 | 19.5606 | <.0001 |
| Continuity Adj. Chi-Square | 1 | 20.0048 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 20.9460 | <.0001 |
| Phi Coefficient | | -0.0976 | |
| Contingency Coefficient | | 0.0971 | |
| Cramer's V | | -0.0976 | |

**Table 4.** Chi-square test results for Survived versus Age

The relationship between survival and class is the most interesting one. The mosaic plot in Figure 3 indicates the highest survival rate for passengers in first class. There was significantly higher than expected number of first class passengers who survived the disaster and a significantly lower than expected number of first class passengers who perished. In contrast, there was a much lower than expected number of male survivors. Third class passengers seem to have had much lower chance of survival, similar to the ship's crew. Counts of those who survived and perished in second class were close to expected counts assuming independence. The survival rate was higher than in the third class and lower than in the first class. This is confirmed by counts and percentages reported in Table 5. From Table 6, the chi-squared statistic for the test of independence is 190.40 (d.f. = 3) with the corresponding *P*-value < 0.001, confirming a highly statistically significant relationship between survival and class.

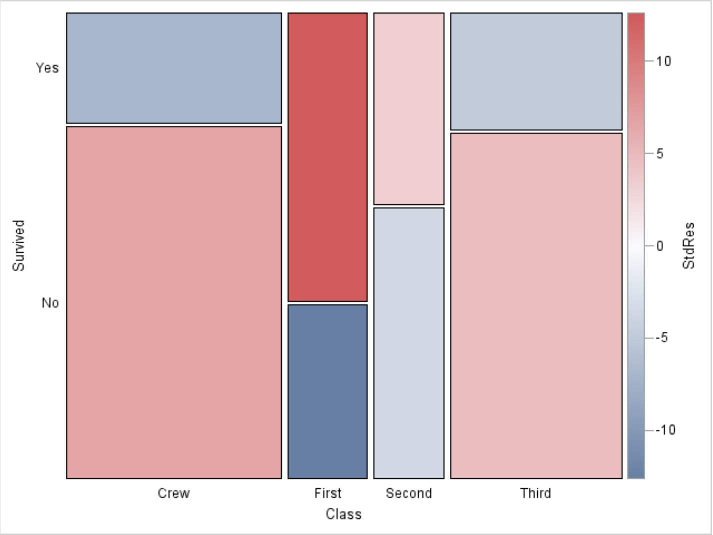
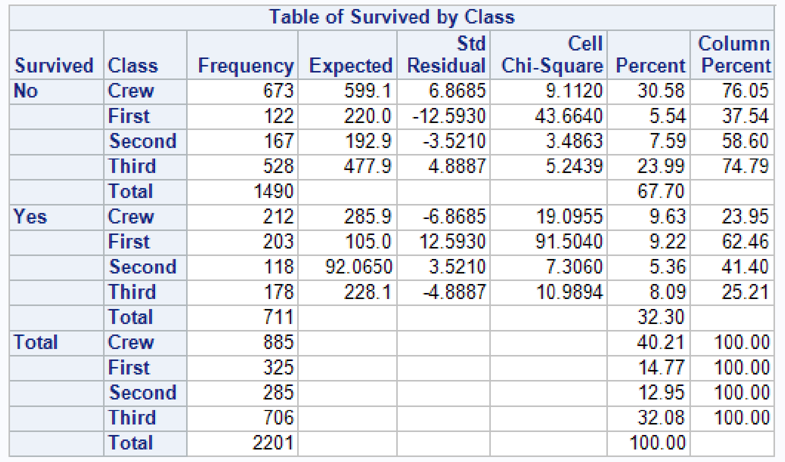
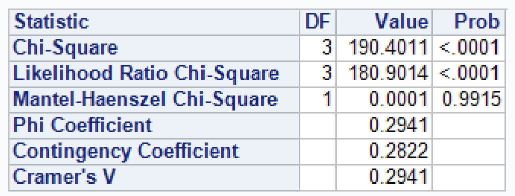


**Figure 3.** Mosaic plot of Survived versus Class



| Table of Survived by Class | | | | | | | |
|---|---|---|---|---|---|---|---|
| Survived | Class | Frequency | Expected | Std Residual | Cell Chi-Square | Percent | Column Percent |
| No | Crew | 673 | 599.1 | 6.8685 | 9.1120 | 30.58 | 76.05 |
| | First | 122 | 220.0 | -12.5930 | 43.6640 | 5.54 | 37.54 |
| | Second | 167 | 192.9 | -3.5210 | 3.4863 | 7.59 | 58.60 |
| | Third | 528 | 477.9 | 4.8887 | 5.2439 | 23.99 | 74.79 |
| | Total | 1490 | | | | 67.70 | |
| Yes | Crew | 212 | 285.9 | -6.8685 | 19.0955 | 9.63 | 23.95 |
| | First | 203 | 105.0 | 12.5930 | 91.5040 | 9.22 | 62.46 |
| | Second | 118 | 92.0650 | 3.5210 | 7.3060 | 5.36 | 41.40 |
| | Third | 178 | 228.1 | -4.8887 | 10.9894 | 8.09 | 25.21 |
| | Total | 711 | | | | 32.30 | |
| Total | Crew | 885 | | | | 40.21 | 100.00 |
| | First | 325 | | | | 14.77 | 100.00 |
| | Second | 285 | | | | 12.95 | 100.00 |
| | Third | 706 | | | | 32.08 | 100.00 |
| | Total | 2201 | | | | 100.00 | |

**Table 5.** Frequency table for Survived by Class



| Statistic | DF | Value | Prob |
|---|---|---|---|
| Chi-Square | 3 | 190.4011 | <.0001 |
| Likelihood Ratio Chi-Square | 3 | 180.9014 | <.0001 |
| Mantel-Haenszel Chi-Square | 1 | 0.0001 | 0.9915 |
| Phi Coefficient | | 0.2941 | |
| Contingency Coefficient | | 0.2822 | |
| Cramer's V | | 0.2941 | |

**Table 6.** Chi-square test results for Survived versus Class

Therefore, based on all of the above, the policy of 'women and children first' appears to have worked best for passengers in the first class, and not so well in the third class.

(b) Fit and interpret a logistic model for the probability of surviving the Titanic disaster with three main effects of Gender, Age and Class.

Logistic model was fit to predict the probability of survival, i.e. $p = P$ (Survived = 1). Dummy variables for categorical predictors Gender, Age and Class were defined using reference coding. Odds ratios for survival will therefore be estimated relative to female passengers, children and crew.

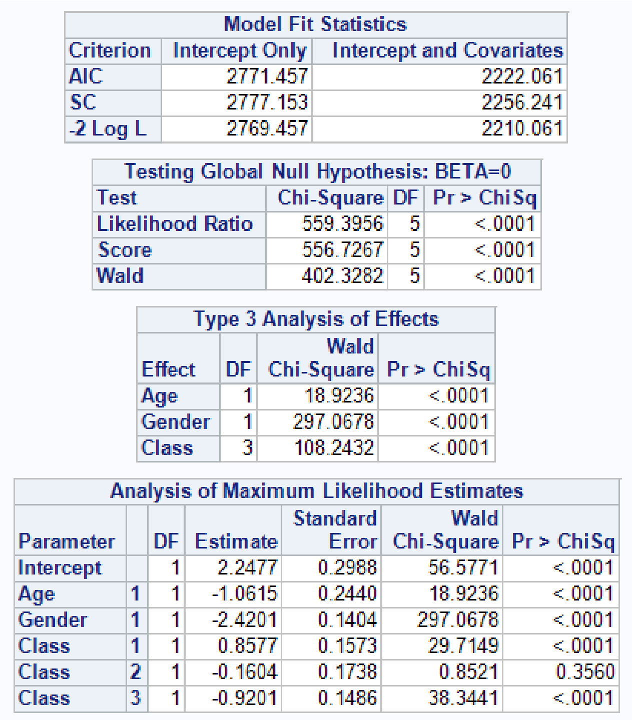
Model fit statistics in Table 7 indicate the model with intercept only to be inferior to the model that includes categorical predictors Gender, Age and Class. All three tests for the global hypothesis of zero beta indicate a highly statistically significant model, P-value < 0.001. Type 3 analysis of effects shows that all three predictors are statistically significant (*P*-value < 0.001).

From the parameters estimates section in Table 7, the estimated model for log odds of survival is

log (p/(1-p)) = 2.25 – 1.06xAge – 2.42xGender + 0.86xFirst – 0.16xSecond – 0.92xThird

This equation confirms that women and children in first class had a significantly higher chance of survival compared to all other passengers and crew.

Model performance statistics in Table 8 are based on analysis of all possible pairs of passengers in which one survived and the other did not. These statistics show 67.7% concordant pairs (passenger with the higher predicted probability of survival is the one who actually survived), 15.7% discordant pairs (the model predicted higher probability of survival for the passenger in the pair who perished) and 16.6% ties (the model predicts the same probability of survival for both passengers in the pair). Based on the *c* statistic, the probability is 76% that a passenger who survived has higher predicted probability than does a passenger who did not survive. Therefore, the model works quite well overall.



Model Fit Statistics

| Criterion | Intercept Only | Intercept and Covariates |
|---|---|---|
| AIC | 2771.457 | 2222.061 |
| SC | 2777.153 | 2256.241 |
| -2 Log L | 2769.457 | 2210.061 |

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 559.3956 | 5 | <.0001 |
| Score | 556.7267 | 5 | <.0001 |
| Wald | 402.3282 | 5 | <.0001 |

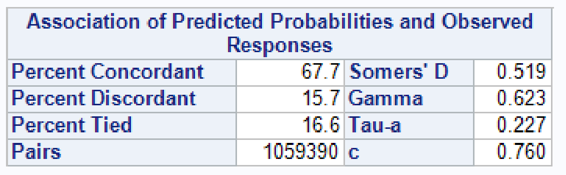
Type 3 Analysis of Effects

| | | Wald | |
|---|---|---|---|
| Effect | DF | Chi-Square | Pr > ChiSq |
| Age | 1 | 18.9236 | <.0001 |
| Gender | 1 | 297.0678 | <.0001 |
| Class | 3 | 108.2432 | <.0001 |

Analysis of Maximum Likelihood Estimates

| | | | | Standard | Wald | |
|---|---|---|---|---|---|---|
| Parameter | | DF | Estimate | Error | Chi-Square | Pr > ChiSq |
| Intercept | | 1 | 2.2477 | 0.2988 | 56.5771 | <.0001 |
| Age | 1 | 1 | -1.0615 | 0.2440 | 18.9236 | <.0001 |
| Gender | 1 | 1 | -2.4201 | 0.1404 | 297.0678 | <.0001 |
| Class | 1 | 1 | 0.8577 | 0.1573 | 29.7149 | <.0001 |
| Class | 2 | 1 | -0.1604 | 0.1738 | 0.8521 | 0.3560 |
| Class | 3 | 1 | -0.9201 | 0.1486 | 38.3441 | <.0001 |

**Table 7.** Model fit statistics and parameter estimates



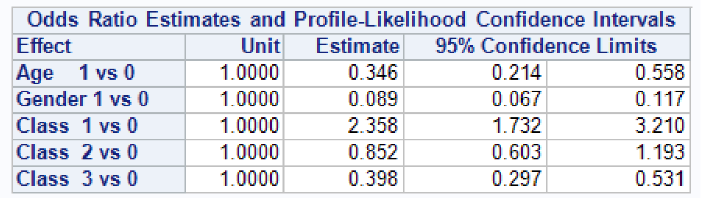
Association of Predicted Probabilities and Observed Responses

| | | | |
|---|---|---|---|
| Percent Concordant | 67.7 | Somers' D | 0.519 |
| Percent Discordant | 15.7 | Gamma | 0.623 |
| Percent Tied | 16.6 | Tau-a | 0.227 |
| Pairs | 1059390 | c | 0.760 |

**Table 8.** Model performance statistics

Odds ratios together with 95% confidence intervals are listed in Table 9 and illustrated in Figure 4. With the exception of the comparison between the second class and the crew, confidence intervals do not contain one, which means that there were significant differences in the chances of survival in all other cases.

The estimated odds ratio for adults compared to children indicates that children had 1/0.346 = 2.89 times higher odds of survival than adults. For females, the estimated odds ratio translates into 1/0.089 = 11.24 times higher odds of survival compared to males. Compared to the crew, first class passengers had 2.36 times higher odds of surviving the disaster. Finally, the odds of survival for third class passengers were 0.40 times lower than for members of the crew.



Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

| Effect | Unit | Estimate | 95% Confidence Limits | |
|---|---|---|---|---|
| Age 1 vs 0 | 1.0000 | 0.346 | 0.214 | 0.558 |
| Gender 1 vs 0 | 1.0000 | 0.089 | 0.067 | 0.117 |
| Class 1 vs 0 | 1.0000 | 2.358 | 1.732 | 3.210 |
| Class 2 vs 0 | 1.0000 | 0.852 | 0.603 | 1.193 |
| Class 3 vs 0 | 1.0000 | 0.398 | 0.297 | 0.531 |

**Table 9.** Odds ratio estimates for main effects of Gender, Age and Class

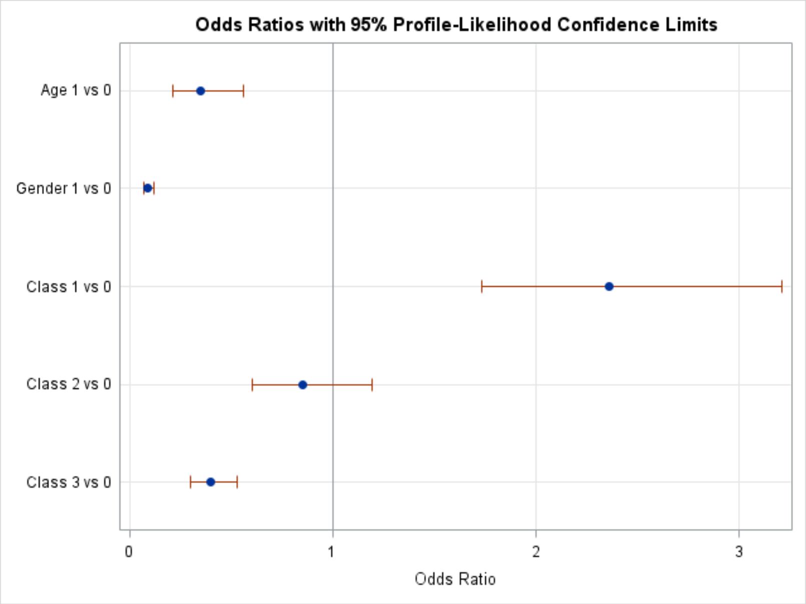
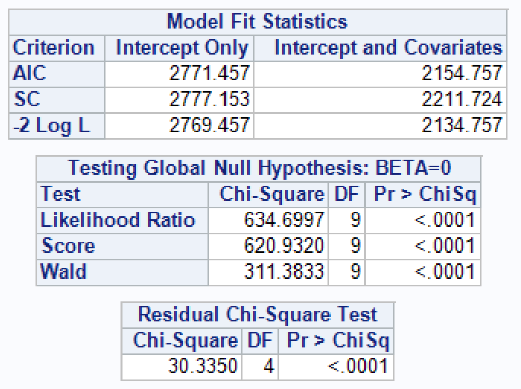


**Figure 4.** Plot of odds ratios for main effects of Gender, Age and Class

(c) Now fit and interpret a model with both main effects and interactions. Specifically, start with a model that includes the same three variables and all possible two-way interactions, and use a backwards elimination technique.

***Note:*** A backwards elimination method will produce what is called a hierarchical model. In this kind of model, main effects cannot be removed from the model if these effects are involved in an interaction that remains in the model.



| Model Fit Statistics | | |
|---|---|---|
| Criterion | Intercept Only | Intercept and Covariates |
| AIC | 2771.457 | 2154.757 |
| SC | 2777.153 | 2211.724 |
| -2 Log L | 2769.457 | 2134.757 |

| Testing Global Null Hypothesis: BETA=0 | | | |
|---|---|---|---|
| Test | Chi-Square | DF | Pr > ChiSq |
| Likelihood Ratio | 634.6997 | 9 | <.0001 |
| Score | 620.9320 | 9 | <.0001 |
| Wald | 311.3833 | 9 | <.0001 |

| Residual Chi-Square Test | | |
|---|---|---|
| Chi-Square | DF | Pr > ChiSq |
| 30.3350 | 4 | <.0001 |

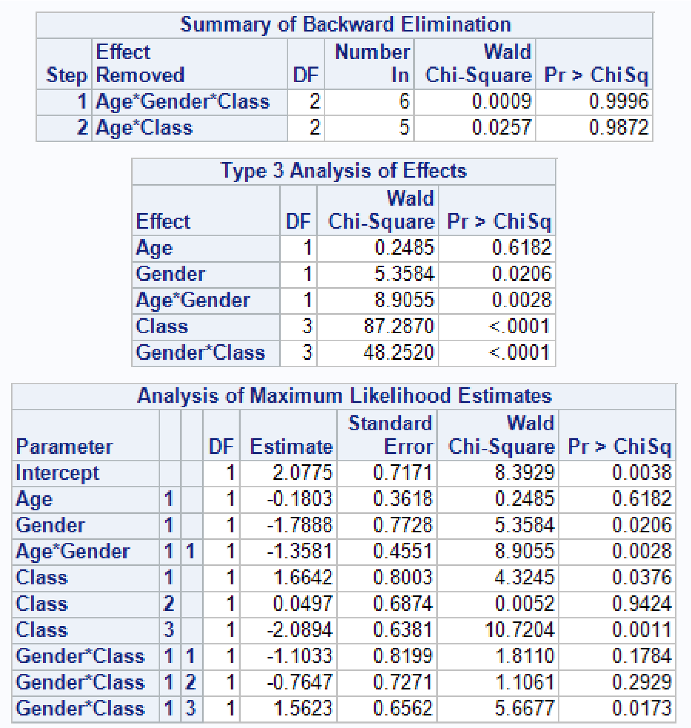
**Table 10.** Model fit statistics

Comparing model fit statistics in Table 10 to those in Table 7, both AIC and SC measures have decreased, indicating that the model that includes interactions between Age and Gender as well as Gender and Class is slightly better than the model without interactions.

Model performance statistics in Table 12 indicate a slight improvement as well; the *c* statistic has increased from 0.76 to 0.766. The percentage of concordant pairs has increased to 16.8% but the number of ties remained the same. As the improvements are small, one may choose to work with the simpler model.

Odds ratio estimates are shown in Table 13 and Figure 5. Of particular note is the estimated odds ratio for female passengers in first class relative to female passengers travelling in third class. At 42.67, the estimate indicates odds of survival 42.67 times higher for female passengers in the first class.

Interpretation of other estimates is left as an exercise.



Summary of Backward Elimination

| Step | Effect<br>Removed | DF | Number<br>In | Wald<br>Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| 1 | Age*Gender*Class | 2 | 6 | 0.0009 | 0.9996 |
| 2 | Age*Class | 2 | 5 | 0.0257 | 0.9872 |

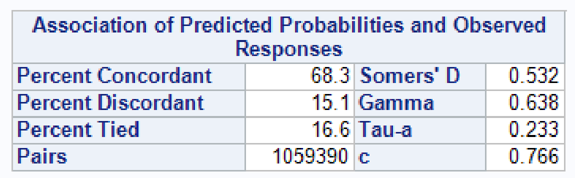
Type 3 Analysis of Effects

| Effect | DF | Wald<br>Chi-Square | Pr > ChiSq |
|---|---|---|---|
| Age | 1 | 0.2485 | 0.6182 |
| Gender | 1 | 5.3584 | 0.0206 |
| Age*Gender | 1 | 8.9055 | 0.0028 |
| Class | 3 | 87.2870 | <.0001 |
| Gender*Class | 3 | 48.2520 | <.0001 |

Analysis of Maximum Likelihood Estimates

| Parameter | | | DF | Estimate | Standard<br>Error | Wald<br>Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| Intercept | | | 1 | 2.0775 | 0.7171 | 8.3929 | 0.0038 |
| Age | 1 | | 1 | -0.1803 | 0.3618 | 0.2485 | 0.6182 |
| Gender | 1 | | 1 | -1.7888 | 0.7728 | 5.3584 | 0.0206 |
| Age*Gender | 1 | 1 | 1 | -1.3581 | 0.4551 | 8.9055 | 0.0028 |
| Class | 1 | | 1 | 1.6642 | 0.8003 | 4.3245 | 0.0376 |
| Class | 2 | | 1 | 0.0497 | 0.6874 | 0.0052 | 0.9424 |
| Class | 3 | | 1 | -2.0894 | 0.6381 | 10.7204 | 0.0011 |
| Gender*Class | 1 | 1 | 1 | -1.1033 | 0.8199 | 1.8110 | 0.1784 |
| Gender*Class | 1 | 2 | 1 | -0.7647 | 0.7271 | 1.1061 | 0.2929 |
| Gender*Class | 1 | 3 | 1 | 1.5623 | 0.6562 | 5.6677 | 0.0173 |

**Table 11.** Parameter estimates



Association of Predicted Probabilities and Observed Responses

| | | | |
|---|---|---|---|
| Percent Concordant | 68.3 | Somers' D | 0.532 |
| Percent Discordant | 15.1 | Gamma | 0.638 |
| Percent Tied | 16.6 | Tau-a | 0.233 |
| Pairs | 1059390 | c | 0.766 |

**Table 12.** Model performance statistics

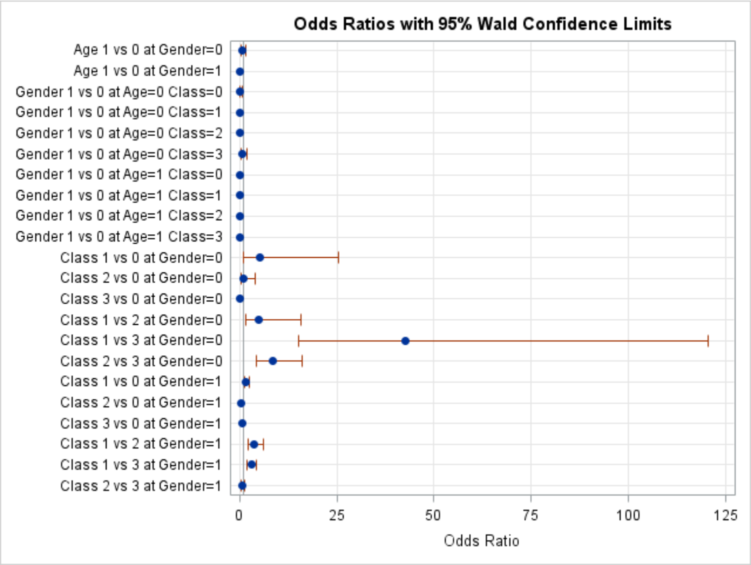
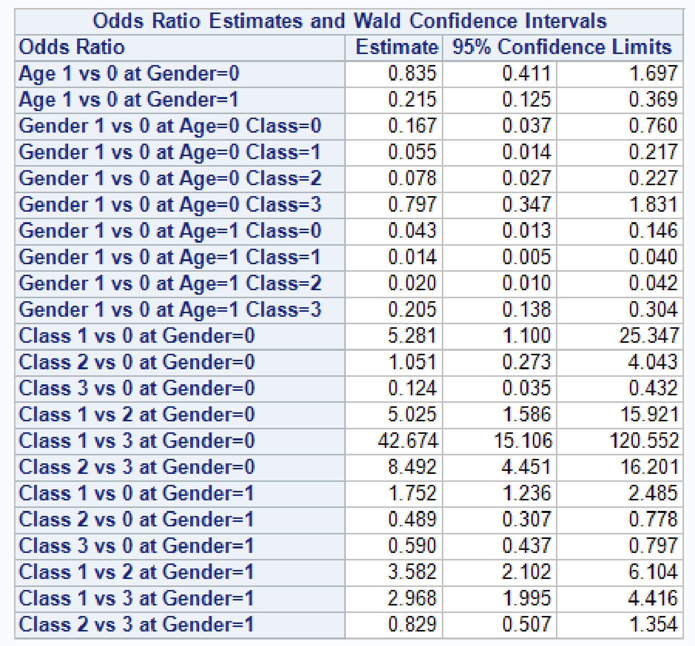


**Figure 5.** Plot of odds ratios for main and interaction effects



| Odds Ratio Estimates and Wald Confidence Intervals | | | |
|---|---|---|---|
| Odds Ratio | Estimate | 95% Confidence Limits | |
| Age 1 vs 0 at Gender=0 | 0.835 | 0.411 | 1.697 |
| Age 1 vs 0 at Gender=1 | 0.215 | 0.125 | 0.369 |
| Gender 1 vs 0 at Age=0 Class=0 | 0.167 | 0.037 | 0.760 |
| Gender 1 vs 0 at Age=0 Class=1 | 0.055 | 0.014 | 0.217 |
| Gender 1 vs 0 at Age=0 Class=2 | 0.078 | 0.027 | 0.227 |
| Gender 1 vs 0 at Age=0 Class=3 | 0.797 | 0.347 | 1.831 |
| Gender 1 vs 0 at Age=1 Class=0 | 0.043 | 0.013 | 0.146 |
| Gender 1 vs 0 at Age=1 Class=1 | 0.014 | 0.005 | 0.040 |
| Gender 1 vs 0 at Age=1 Class=2 | 0.020 | 0.010 | 0.042 |
| Gender 1 vs 0 at Age=1 Class=3 | 0.205 | 0.138 | 0.304 |
| Class 1 vs 0 at Gender=0 | 5.281 | 1.100 | 25.347 |
| Class 2 vs 0 at Gender=0 | 1.051 | 0.273 | 4.043 |
| Class 3 vs 0 at Gender=0 | 0.124 | 0.035 | 0.432 |
| Class 1 vs 2 at Gender=0 | 5.025 | 1.586 | 15.921 |
| Class 1 vs 3 at Gender=0 | 42.674 | 15.106 | 120.552 |
| Class 2 vs 3 at Gender=0 | 8.492 | 4.451 | 16.201 |
| Class 1 vs 0 at Gender=1 | 1.752 | 1.236 | 2.485 |
| Class 2 vs 0 at Gender=1 | 0.489 | 0.307 | 0.778 |
| Class 3 vs 0 at Gender=1 | 0.590 | 0.437 | 0.797 |
| Class 1 vs 2 at Gender=1 | 3.582 | 2.102 | 6.104 |
| Class 1 vs 3 at Gender=1 | 2.968 | 1.995 | 4.416 |
| Class 2 vs 3 at Gender=1 | 0.829 | 0.507 | 1.354 |

**Table 13.** Odds ratio estimates for main and interaction effects of Gender, Age and Class

## Appendix – SAS code

```
ods graphics on;

proc format;
value Surv 0 = 'No' 1 = 'Yes';
value P_Gender 0 = 'Female' 1 = 'Male';
value P_Age 0 = 'Child' 1 = 'Adult';
value P_Class 0 = 'Crew' 1 = 'First' 2 = 'Second' 3 = 'Third';
run;

title ' Mosaic plot (default) Survived vs Gender';

proc freq data=math4044.titanic;
tables Survived*Gender / norow chisq plots=MOSAIC; /* alias for
MOSAICPLOT */
format Survived Surv. Gender P_Gender.;
run;

title ' Mosaic plot (default) Survived vs Age';

proc freq data=math4044.titanic;
tables Survived*Age / norow chisq plots=MOSAIC; /* alias for
MOSAICPLOT */
format Survived Surv. Age P_Age.;
run;

title ' Mosaic plot (default) Survived vs Class';

proc freq data=math4044.titanic;
tables Survived*Class / norow chisq plots=MOSAIC; /* alias for
MOSAICPLOT */
format Survived Surv. Class P_Class.;
run;

title 'Mosaic plot (colour by response) Survived vs Gender';

proc freq data=math4044.titanic;
tables Survived*Gender / norow cellchi2 expected stdres crosslist;
ods output CrossList=FreqList(where=(Expected>0));
format Survived Surv. Gender P_Gender.;
run;

/* colour by response (notice that PROC FREQ reverses Y axis) */

proc template;
define statgraph mosaicPlotParm;
begingraph;
layout region;
MosaicPlotParm category=(Gender Survived) count=Frequency /
colorresponse=StdResidual name="mosaic";
continuouslegend "mosaic" / title="StdRes";
endlayout;
endgraph;
end;
run;
```

```
proc sgrender data=FreqList template=mosaicPlotParm;
run;

title 'Mosaic plot (colour by response) Survived vs Age';

proc freq data=math4044.titanic;
tables Survived*Age / norow cellchi2 expected stdres crosslist;
ods output CrossList=FreqList(where=(Expected>0));
format Survived Surv. Age P_Age.;
run;

proc template;
define statgraph mosaicPlotParm;
begingraph;
layout region;
MosaicPlotParm category=(Age Survived) count=Frequency /
colorresponse=StdResidual name="mosaic";
continuouslegend "mosaic" / title="StdRes";
endlayout;
endgraph;
end;
run;

proc sgrender data=FreqList template=mosaicPlotParm;
run;

title 'Mosaic plot (colour by response) Survived vs Class';

proc freq data=math4044.titanic;
tables Survived*Class / norow cellchi2 expected stdres crosslist;
ods output CrossList=FreqList(where=(Expected>0));
format Survived Surv. Class P_Class.;
run;

proc template;
define statgraph mosaicPlotParm;
begingraph;
layout region;
MosaicPlotParm category=(Class Survived) count=Frequency /
colorresponse=StdResidual name="mosaic";
continuouslegend "mosaic" / title="StdRes";
endlayout;
endgraph;
end;
run;

proc sgrender data=FreqList template=mosaicPlotParm;
run;
```

```
title 'Logistic model with three categorical predictors';

proc logistic data=math4044.titanic;
class Age (ref='0') Gender (ref='0') Class (ref='0') /
param=reference;
model Survived (event='1') = Age Gender Class / clodds=pl;
run;
quit;

title 'Logistic model with three categorical predictors plus
interactions';

proc logistic  data=math4044.titanic;
class Age (ref='0') Gender (ref='0') Class (ref='0') /
param=reference;
model Survived (event='1') = Age | Gender | Class /
selection=backward clodds=pl;
oddsratio Age;
oddsratio Gender;
oddsratio Class;
run;
quit;

ods graphics off;
```