StatR 502 Final Project:

Women and Children First?

Thomas Wright

Introduction

In times of emergency, is it "every man for himself" or "women and chlidren first"?

The Titanic, thought to be "unsinkable", was packed with 1,317 passengers on a voyage from England to North America before striking an iceberg and plummeting to the ocean floor two and a half hours later.

By utilizing a logistic model on a dataset regarding the fates of the passengers of the Titanic, the study will quantify the benefits (or lack of) of being a minor or woman in a time of crisis. Furthermore, an inferential look at the “winning” model will reveal if other factor significantly influence the chance of survival in a mass catastrophe.

In addition to identifying the characteristics most useful in survival situations, this question also presents an opportunity to model a binary response variable with a multitude of categorical explanatory variables and interaction terms. By utilizing stepAIC, a larger breadth of models can be tested. Moreover, the models produced by automated methods out perform those models derived from domain knowledge in this study.

Data

The data have been compiled from the sole voyage of the Titanic. There are 891 passenger observations (or rows) with 9 explanatory variables and one binary response (Survival or Death). Explanatory variables include:

- pclass
    - Factor with levels -- Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)
    - Refers to the Socio-economic class of the passenger
- sex
    - Factor with two levels – Female (1) or Male (0)
- age
    - An integer value for Age
- sibsp
    - An Integer
    - Number of Siblings/Spouses Aboard
- parch
    - An Integer
    - Number of Parents/Children Aboard
- fare
    - An Integer
    - The cost of the ticket
- cabin
    - A factor with three levels
    - Which Cabin the passenger's room was located
- embarked
    - Factor with three levels
    - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)

The dataset also includes two ID columns:

- Name
- Ticket Number

Cleaning and Manipulation

The dataset contains missing values with no way of approximating the value. As such, the study removed all rows with missing values and deleted the ID columns. Next, new variables were constructed

Age

Age was originally a discrete numerical object. The study created two new variables:

Binary Predictor:

A binary predictor was created for for whether a passenger was a child or not - using 14 as a cut off age due to the Fisher Act.

The 1918 Fisher Act after the First World War brought in a standard leaving age for all of 14, against opposition from some employers and many parents. The Government had legislated for part-time continuation education for young workers up to age 16, but this was not well-supported by employers or by young people, and continuation schools fell by the wayside.

Categorical Predictor:

A cateogrical predictor was created by cutting the age variable into groups, converting into a numerical form, and then back to a factor. The ages were grouped from 0-5, 6-12, 13-18, 19-30, 30-60, and 60-80.

The rationale was to split the groups by physical ability - from a small child that needs to be held, to a child that can walk on its on, to a fit adult, to a mature adult, to an elderly individual.

Coding the variable in such a way will allow the study to isolate the impact of being a child or not to a more granular level - avoiding the "ecological fallacy" (to be discuessed more later).

Family

The variables SibSp and Parch are too collapsed for the study:

- SibSp contains information on whether a spouse or multiple siblings are on board
- Parch containing information on how many parents a child has or how many children a parent has on board.

The study created four new variables to capture the following structure:

- **Family**: How many siblings the passenger has
- **spouse**: Binary factor for whether the spouse is on board or not
- **parents**: How many parents does the passenger have on the titanic

- **children**: How many children the passenger has with them

Temporary models

Two temporary models were constructed.

- "Project2" contains all of the variables of the master dataset, except for age, age2 (grouped), SibSp, and Parch.
- "Project3" contains all of the variables of the master dataset excluding age, child (binary), SibSp, and Parch.

The variables Age, Age2, and Child are all linear combination of one another, so only one can be included in a model. Each of the temporary data sets only contains one for ease of modeling. The varibles for the family characteristic and SibSp/Parch are also linear combinations and therefore SibSp/Parch have been dropped for the more detailed variables in the modeling.

Importantly, these models cannot be compared for accuracy due to their now differing data structures.

Final cleaning

The structures of all of the objects in the above mentioned dataframes were coersed into factors with the expection of Fare - the sole continuous predictor variable. The response variable is also a factor (survival or not.)

Methodology

The study conducted a three-step analysis. First, the data is explored with visual plots. Next, the temporary dataset containing the binary indicator for being a child was examined to see if there was a significant impact on survival. Finally, the dataset containing Age coded as a multi-level factor was modeled to see if a more granular approach yielded more insight.

For both stages of the analysis, the study utilized a variety of step-wise (automated) regression techniques to select the winning model. Individual regression coefficients were then examined for significance. The same methodology was applied to each dataset (with only the source data changing) and the only the sample output from the binary Age dataset is shown.

Reordering factor levels:

To aid in interpretation, the base level of several factors were adjusted:

- Pclass
    - Class 2 (middle socio-economic class) is now the base level - illuminating the differences between being poor and rich.
- Age2
    - "Fit adult" (18-40) is now the base level the base level - illuminating the differences between being an able bodied adult and someone who may need benefial treatment due to their advanced or premature age.
- Sex
    - Male the base level - illuminating the benefit of being a woman in the survival situation.

Defining the limits of the Stepwise Function

When running a stepwise regression, one must step the upper parameter limit for the model to test – such as the order of the polynomial and whether to include interaction terms.

After changing the continuous parameter of Age (which could have a quadratic effect) to either a binary predictor or a factor, there are no variables that present an obvious need for a higher order model.

Without Interaction Terms

Looking at the model without interaction terms suggests that Pclass, Sex, and child are signifant predictors.

```r
mod.NoInt <- glm(Survived ~ ., family = binomial (link='logit'),data=Project2
)
summary(mod.NoInt)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial(link = "logit"),
##     data = Project2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0785  -0.6486  -0.4244   0.6108   2.4686
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.253e+01  1.696e+03   0.007 0.994102
## X            2.149e-04  3.960e-04   0.543 0.587345
## Pclass1      8.597e-01  3.196e-01   2.690 0.007136 **
## Pclass3     -8.665e-01  2.585e-01  -3.351 0.000804 ***
## Sexfemale    2.711e+00  2.325e-01  11.660  < 2e-16 ***
## Fare         4.595e-03  3.156e-03   1.456 0.145346
## EmbarkedC   -1.382e+01  1.696e+03  -0.008 0.993497
## EmbarkedQ   -1.430e+01  1.696e+03  -0.008 0.993269
## EmbarkedS   -1.409e+01  1.696e+03  -0.008 0.993372
## spouse1     -1.748e-01  2.713e-01  -0.644 0.519300
## siblings1    1.493e-02  5.274e-01   0.028 0.977420
## siblings2   -8.416e-01  6.272e-01  -1.342 0.179693
## siblings3   -3.331e+00  9.858e-01  -3.379 0.000727 ***
## siblings4   -2.717e+00  9.074e-01  -2.994 0.002755 **
## siblings5   -1.678e+01  1.030e+03  -0.016 0.987004
## children1   -3.795e-01  3.704e-01  -1.024 0.305612
## children2   -5.328e-02  5.319e-01  -0.100 0.920210
## children3    3.704e-01  1.202e+00   0.308 0.757897
## children4   -1.647e+01  1.060e+03  -0.016 0.987612
## children5   -1.545e+00  1.179e+00  -1.310 0.190199
## children6   -1.705e+01  2.400e+03  -0.007 0.994332
## parents1     1.002e+00  6.632e-01   1.511 0.130854
## parents2     6.872e-01  7.689e-01   0.894 0.371495
## parents3    -1.471e+01  2.400e+03  -0.006 0.995109
## child1       1.249e+00  6.151e-01   2.030 0.042346 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 960.90  on 711  degrees of freedom
## Residual deviance: 612.06  on 687  degrees of freedom
## AIC: 662.06
##
## Number of Fisher Scoring iterations: 15
```

Interpreting the preliminary results yields:

- That in two individuals identidical except for Class, moving from middle class to upper class will increase your chance of survival by 70%. Conversely, moving from middle class to lower class will lower your chance of surivival by 70%.
- In two individual identical except for Sex, those that are women are 93% more likely to survive than men.
- Lastly, in those that are identical except for Age, that that are women are 77% more likely to survive than men.

However, with large effects as observed above[1], interaction terms are often needed to deal with confounding factors.

As such, the study utilized an upper boundary of a 1st order polynomial with interaction terms for the stepwise regression methods.

Selecting the Model

To construct a model with interactions terms, the study utilized five approaches:

1. Using stepAIC: Fitting the Null Model and iterating forward.
2. Using stepAIC: Fitting the Full Model and iterating backward.
3. Using stepAIC: Fitting the Null Model and iterating both ways.
4. Using stepAIC: Fitting the Full Model and iterating both ways.
5. Using domain knowledge to construct a best guess model after the stepAIC outputs.

StepAIC is an automated model selection that accepts a preliminarly model, a scope of terms to test and a method of iteration.

Null and Full Models[2]:

```
mod.null <-glm(Survived ~ 1, family = binomial (link='logit'),data=Project2)
mod.full <- glm(Survived ~ .*., family = binomial (link='logit'),data=Project
2)
```

After running the five methods, the most parsimonious model with the approximate lowest AIC was chosen as the winning model and examined for its significance.

---

[1] Using the second dataset (with Age as a factor) yielded similar results suggesting that interaction terms are necessary. Excluded from the report for brevity.

[2] The Null and Full models are identical for the second analysis (Age as a factor), with the data argument inside the glm function being changed to Project3.

Mixed Effects

The analytic goal of a mixed effect model is to represent the general population of groups and is used when the main effect of any single group (the fixed effects) is not important. It is useful to make inferences about the "population of possible batches", and to understand that population distribution

In the model so far, only siblings could be represented by a mixed effect. By doing so, the model could represent the general impact of having multiple children.

However the variance of the random effect in the model is 0, showing that fixed effects should be used instead.

```
lmm <- lmer(formula = Survived ~ 1+Pclass + Sex + Fare + Embarked + spouse +
              (1|siblings) + parents + child + Pclass:Sex + Pclass:spouse +
              Pclass:child +  Sex:parents + Sex:child + Fare:spouse + Fare:pa
rents + Fare:child + parents:child,
            family = binomial(link = "logit"), data = Project2,
            REML = FALSE)

summary(lmm)

## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: Survived ~ 1 + Pclass + Sex + Fare + Embarked + spouse + (1 |
##     siblings) + parents + child + Pclass:Sex + Pclass:spouse +
##     Pclass:child + Sex:parents + Sex:child + Fare:spouse + Fare:parents +
##     Fare:child + parents:child

##
##      AIC      BIC   logLik deviance df.resid
##    595.5    728.0   -268.8    537.5      683
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.9265 -0.3862 -0.2748  0.2830  4.8222
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  siblings (Intercept) 0        0
## Number of obs: 712, groups:  siblings, 6
##
```

Checking with AIC confirms that a fixed effects model is better than a mixed effect for the dataset.[3]

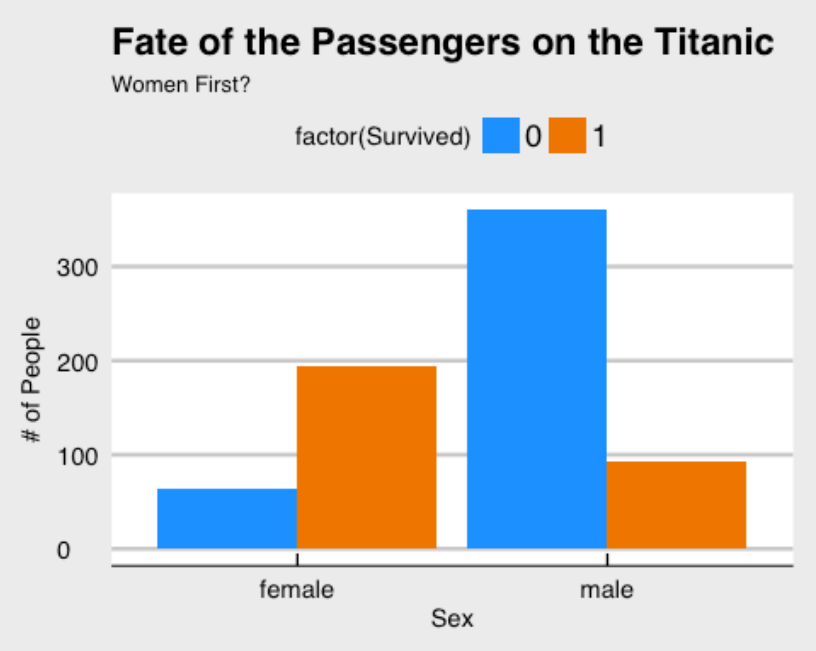
---

[3] Again, the second dataset (with Age as a factor) produced a mixed effect model with variance of 0. As such, a fixed effect model was used for both stages of the analysis.

Results

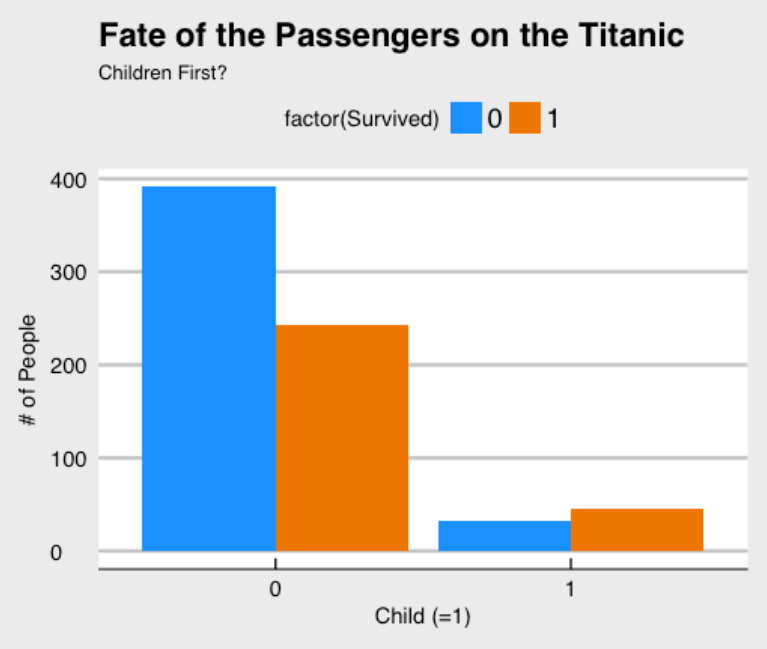
Stage One - Data Visualizations:

Looking first at whether women where more likely to survive or not:



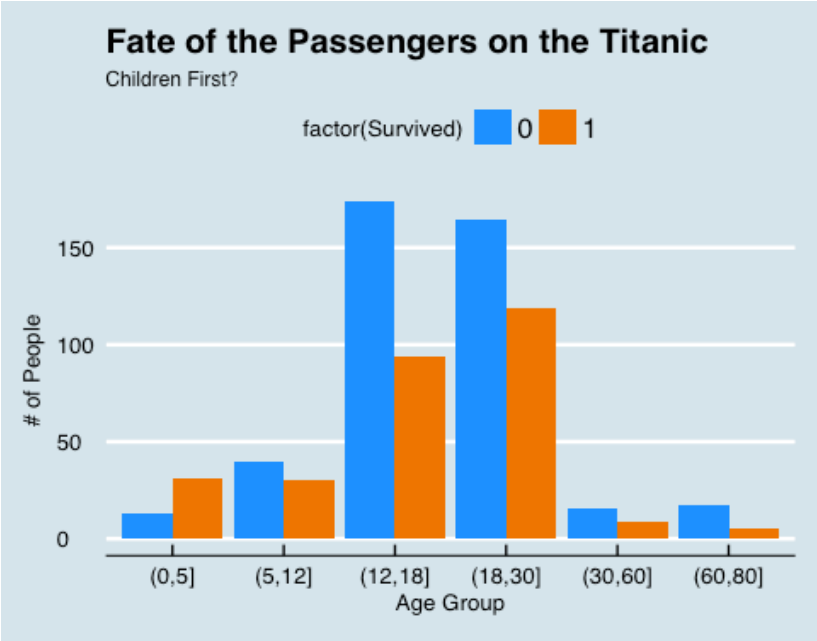
It appears that at first glance, women did in face survive at a much higher rate then men – with a majority of women surviving and the majority of men dying.

Turning to Children shows a similar split – with a majority of children surviving and a majority of adults dying. However, the benefit enjoyed by women appears larger.

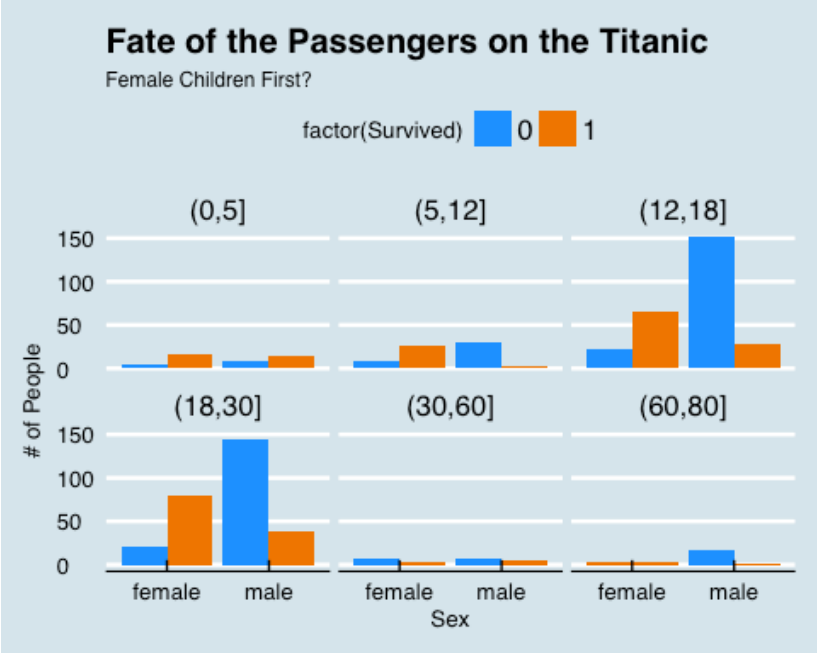


This discrepancy could be due to an ecological fallacy – where inference is drawn on a group at too broad of a level. With a more granular approach, the differences can manifest themselves accurately.

Looking at Age by factor level reveals an interesting pattern – children below five were the only age group where the majority survived. This survival rate is more in-line with the female survival rate.

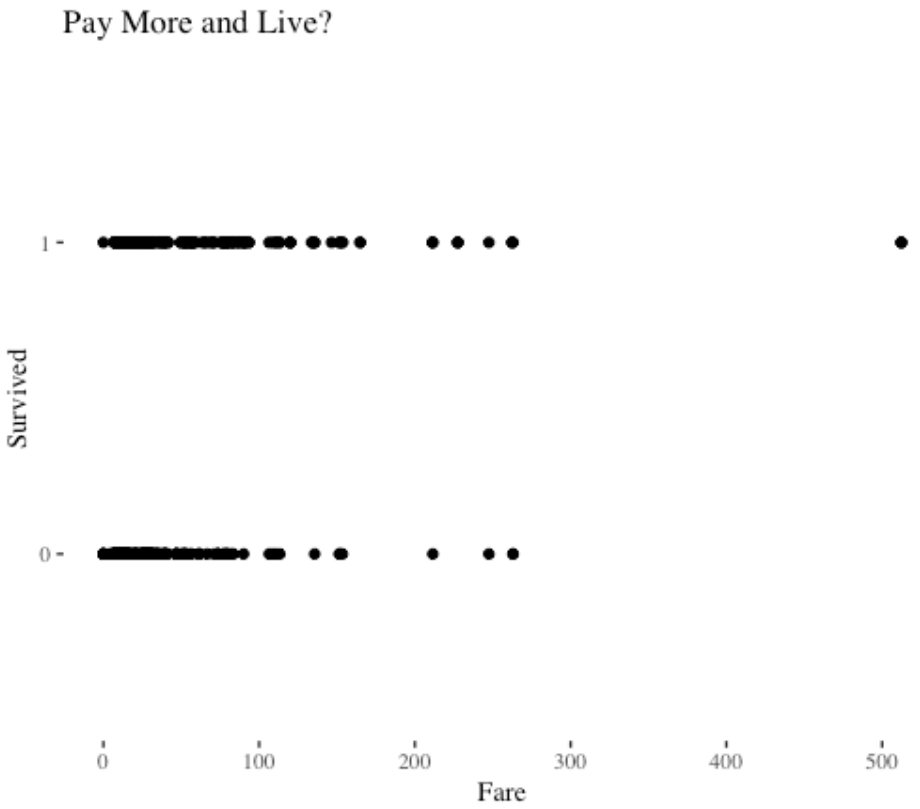


Looking at women who are also children:



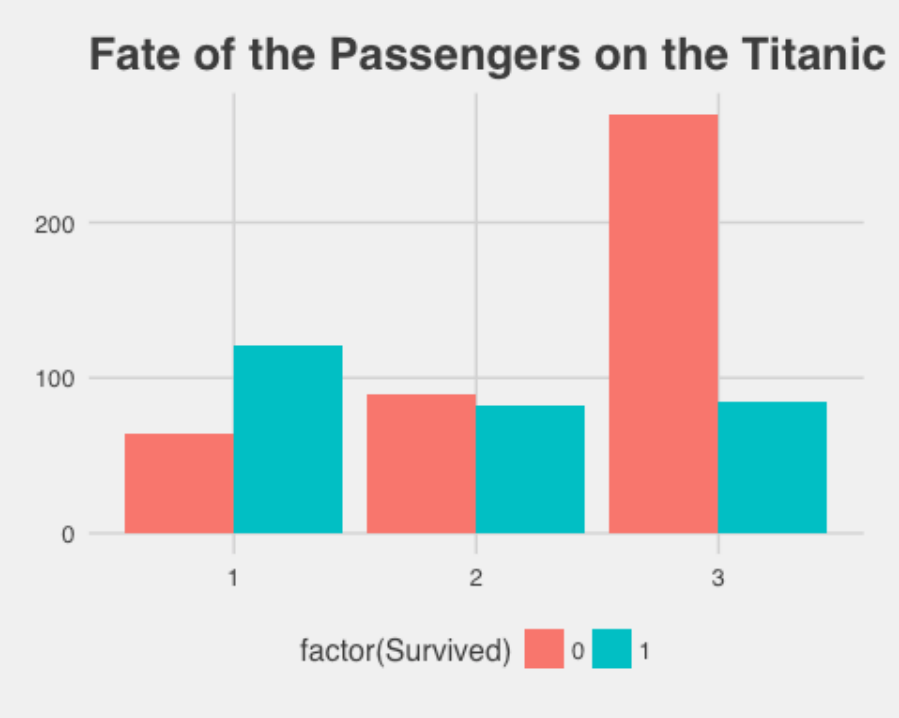
Both male and female children that are under-five survived at similar rate. The protection afforded to women only holds until the age of 30 - whereas the only male group that survived in majority where those under five – perhaps babes in arms.

Looking past physical characteristics, socio-economic class may have a large impact on the chances of survival.



It appears that those that paid more for their ticket survived at a higher rate – as hinted by the clustering of survival and deaths between $100 & $200. Furthermore, every passenger that paid more than $275 survived.

When looking at the fate of the passengers by socio-economic class, it becomes clear that most of the wealthy people survived - regardless of sex or age. Conversely, most of the poor passengers died, suggesting that the "women and children first" protection either didn't extend to them, or that there was an overwhelming majority of men in 3rd class.




```
table(Project$Pclass,Project$Sex)

##
##      female male
##   1     84  101
##   2     73   99
##   3    102  253

table(Project$Pclass,Project$Age2)

##
##      (0,5] (5,12] (12,18] (18,30] (30,60] (60,80]
##   1      3     12      44     111       1      14
##   2     13     12      66      74       4       3
##   3     28     46     158      98      20       5
```

It appears that ~55% of men where in third class, while approximately 40% of women were also. This suggests that “women first” may hold true in the lowest class because if women made up equal numbers of the 3rd class - with the terrible survival rates in the 3rd class -- it would imply that poor women were not getting the benefit of their richer peers.

As a majority of the children in the age groups 5-12 and 18-30 were "poor", it could be that those older than 5 could not garner enough sympathy to displace a richer older person or that they died together with their families.

To quantify those relationships while taking confounding variables into account, the study turns to the results of the modeling.

Stage Two – Age as a Binary:

Following the procedure outlined in methodology section yielded five final models:

AIC Comparison Table		
Method	**DF**	**AIC**
1. Forward	20	594.5597
2. Backward	42	592.7725
3. Null both	20	594.5597
4. Full Both	42	592.7725
5. Domain	27	606.2384

While models #2 and 4 have the lowest AIC, models #1 and 3 have relatively equal AIC and much less parameters to estimate (more parsimonious).

Looking at the "Winning model" – which was fit forward from the Null Model:

```
summary(mod.for)

##
## Call:
## glm(formula = Survived ~ Sex + Pclass + child + siblings + Sex:Pclass +
##     Pclass:child + Sex:child + child:siblings, family = binomial(link = "l
ogit"),
##     data = Project2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6335  -0.5228  -0.3768   0.4385   2.4892
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.6094     0.4232  -6.166 7.00e-10 ***
## Sexfemale            4.9029     0.6035   8.123 4.53e-16 ***
## Pclass1              2.1607     0.4712   4.585 4.54e-06 ***
## Pclass3              0.6881     0.4684   1.469  0.14186
## child1              19.3952   859.3827   0.023  0.98199
## siblings1           -0.4859     0.7098  -0.684  0.49366
## siblings2           -0.7734     0.7729  -1.001  0.31701
## siblings3            0.3212     1.3977   0.230  0.81823
```

```
## siblings4            1.0115     1.5710   0.644  0.51966
## siblings5          -17.4644  3956.1803  -0.004  0.99648
## Sexfemale:Pclass1   -1.0182     0.8747  -1.164  0.24444
## Sexfemale:Pclass3   -3.0832     0.6770  -4.554 5.26e-06 ***
## Pclass1:child1     -17.8703   859.3832  -0.021  0.98341
## Pclass3:child1     -15.5031   859.3828  -0.018  0.98561
## Sexfemale:child1    -3.2230     1.0080  -3.197  0.00139 **
## child1:siblings1     0.5865     1.1810   0.497  0.61948
## child1:siblings2     1.3046     1.5320   0.852  0.39445
## child1:siblings3   -19.1842  1415.3090  -0.014  0.98919
## child1:siblings4    -4.6308     1.9675  -2.354  0.01859 *
## child1:siblings5    -2.0724  4423.1442   0.000  0.99963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 960.90  on 711  degrees of freedom
## Residual deviance: 554.56  on 692  degrees of freedom
## AIC: 594.56
##
## Number of Fisher Scoring iterations: 16
```

While the above is the model that produces the lowest AIC, method test a wide range of models. Below is an approximate ranking of importance – produced by an ANOVA.

```
anova(mod.for)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev
## NULL                              711     960.90
## Sex             1  211.334       710     749.57
## Pclass          2   77.765       708     671.80
## child           1   15.826       707     655.98
## siblings        5   25.915       702     630.06
## Sex:Pclass      2   35.565       700     594.50
## Pclass:child    2   15.826       698     578.67
## Sex:child       1   11.078       697     567.59
## child:siblings  5   13.035       692     554.56
```

As we can see from this model, Sex is the most important determining factor while class is the second most significant indicator of survival. Surprisingly the binary indicator of being

a child or not is the least important of the three indicators identified through visual analysis.

Stage Three – Age as a multi-level Factor

Following the procedure outlined in methodology section yielded six final models (with the Mixed Effects model AIC displayed):

AIC Comparison Table		
Method	**DF**	**AIC**
1. Forward	36	590.085
2. Backward	123	651.0584
3. Null both	36	590.085
4. Full Both	123	651.0584
5. Domain	38	607.5519
6. Mixed Effects	27	615.085

Methods #1 and 3 converge on the same model have the lowest AIC by a significant margin.

Looking at the "Winning model" – which was fit forward from the Null Model:

Looking at the summary

```
summary(mod.for2)

##
## Call:
## glm(formula = Survived ~ Sex + Pclass + Age2 + siblings + Sex:Pclass +
##     Sex:Age2 + Sex:siblings + Pclass:Age2, family = binomial(link = "logit
"),
##     data = Project3)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8460  -0.5748  -0.2666   0.3057   2.5903
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  -2.4248     0.5152  -4.706 2.52e-06 ***
## Sexfemale                     4.8075     0.7028   6.840 7.90e-12 ***
```

```
## Pclass1                          2.1294     0.5718   3.724 0.000196 ***
## Pclass3                          0.4231     0.5935   0.713 0.475899
## Age2Young Child                 52.0752  2827.1997   0.018 0.985304
## Age2Child                       -0.6894     1.5753  -0.438 0.661639
## Age2Teenager                    -0.8943     0.7724  -1.158 0.246916
## Age2Adult                       39.2641  6226.4916   0.006 0.994969
## Age2Elderly                      1.7317     1.3287   1.303 0.192481
## siblings1                       -1.5086     1.3042  -1.157 0.247381
## siblings2                       -0.7711     1.1029  -0.699 0.484457
## siblings3                      -55.6370  4737.1292  -0.012 0.990629
## siblings4                      -38.8392  2071.6548  -0.019 0.985042
## siblings5                      -54.5836  4117.0851  -0.013 0.989422
## Sexfemale:Pclass1               -1.1036     1.0058  -1.097 0.272555
## Sexfemale:Pclass3               -3.7280     0.7543  -4.943 7.71e-07 ***
## Sexfemale:Age2Young Child      -37.1185  2071.6544  -0.018 0.985705
## Sexfemale:Age2Child              1.3466     1.0367   1.299 0.193958
## Sexfemale:Age2Teenager           0.5846     0.6088   0.960 0.336907
## Sexfemale:Age2Adult            -22.4442  3459.7483  -0.006 0.994824
## Sexfemale:Age2Elderly           35.9913  6511.3080   0.006 0.995590
## Sexfemale:siblings1              2.0594     1.5871   1.298 0.194417
## Sexfemale:siblings2              0.7082     1.4536   0.487 0.626106
## Sexfemale:siblings3             54.3648  4737.1293   0.011 0.990843
## Sexfemale:siblings4             38.6722  2071.6551   0.019 0.985107
## Sexfemale:siblings5             34.7361 11515.1720   0.003 0.997593
## Pclass1:Age2Young Child        -34.5460  2441.1627  -0.014 0.988709
## Pclass3:Age2Young Child        -12.8437  1923.8784  -0.007 0.994673
## Pclass1:Age2Child                1.7939     2.2312   0.804 0.421409
## Pclass3:Age2Child                0.5465     1.5938   0.343 0.731686
## Pclass1:Age2Teenager             0.9337     0.9032   1.034 0.301262
## Pclass3:Age2Teenager             1.1794     0.7806   1.511 0.130820
## Pclass1:Age2Adult              -17.8940 12426.5036  -0.001 0.998851
## Pclass3:Age2Adult              -35.1085  6226.4914  -0.006 0.995501
## Pclass1:Age2Elderly             -3.8342     1.7082  -2.245 0.024794 *
## Pclass3:Age2Elderly            -18.7242  4039.9934  -0.005 0.996302
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 960.90  on 711  degrees of freedom
## Residual deviance: 518.08  on 676  degrees of freedom
## AIC: 590.08
##
## Number of Fisher Scoring iterations: 18
```

And again the ANOVA of importance:

```
anova(mod.for2)
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Survived
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev
## NULL                           711     960.90
## Sex           1  211.334       710     749.57
## Pclass        2   77.765       708     671.80
## Age2          5   29.741       703     642.06
## siblings      5   26.520       698     615.54
## Sex:Pclass    2   35.534       696     580.01
## Sex:Age2      5   23.539       691     556.47
## Sex:siblings  5   14.444       686     542.03
## Pclass:Age2  10   23.942       676     518.08
```

Sex is again far-away the most important determining factor in survival. Class maintains its spot in distant second, while the importance of Age has increased with our more granular approach.

Final Model

The study has concluded that its final model is derived from using forward StepAIC from the Null model of the third step. Using the more granular approach to the age variable allows the study to gain greater insight into the significance of Age, Sex and Income in determining survival.

The Model[4]

Survived ~ Sex + Pclass1 + Sex:Pclass3 + Pclass1:Age2Elderly

Significant Coefficients

Pulling out the significant results yeilds:

```
Estimate2 <- coef(mod.t4)[c(1:3,10:11,13,25)]
d2 <- confint(mod.t4, c(c(1:3,10:11,13,25)), level = 0.95)

e2 <- cbind(Estimate2,d2)
g2 <- e2[c(1,4,5,7),]*-1
h2<- e2[c(2,3,6),]
```

Breaking out the positive and negative impacts on probability with 95% Confidence Intervals:

```
positive <- inv.logit(h2)

f2 <- inv.logit(g2)
negative <- f2 * -1

positive
```

```
##                      Estimate2     2.5 %    97.5 %
## Sexfemale           0.9923596 0.9732750 0.9983024
## Pclass1             0.8985951 0.7585392 0.9684020
## Sexfemale:Age2Child 0.8384877 0.4682404 0.9786700
```

```
negative
```

```
##                       Estimate2      2.5 %     97.5 %
## (Intercept)         -0.9210665 -0.9738272 -0.8251737
## Sexfemale:Pclass1   -0.8252040 -0.9666367 -0.4226173
## Sexfemale:Pclass3   -0.9735474 -0.9945125 -0.9028084
## Pclass1:Age2Elderly -0.9798947 -0.9994905 -0.5676078
```

---

[4] Only significant terms included here in the final presentation. See the results section for the full model.

Interpretation

The interpretability of the model is hampered by the large impact of Sex on survival and the few amount of parameters available for the model to test. Traditionally, a study would enter in a 1 for all of the positive values to calculate the percent impacts (after taking the inverse logit as done above).

Ex: A female in first class = 99.23% - 82.5% + 89.8% = 107% chance of survival. An elderly woman in first class would have 107% - 97.9% = 9%.

However, the inferential insights are found in the ANOVA on the winning model, where variables are ranked in importance to explaining the variance in survival rates.

Discussion

In times of disaster, is it women and children first?

From this inferential study – no it is not. While women are overwhelming favorites to survive, the ANOVAs of the winning models revealed a different order of importance.

Socio-economic class was clearly the second most important explanatory variable, with Age of the passenger or a binary indicator of whether they were a child both falling far short. When looking at the data on a granular level, the protection offered to children completely disappears from the significant factors of the final model.

When looking at the visual representations of the data, it appears that those under five have a higher chance of survival than other groups. However, it appears that sex and class are highly confounding factors – with most of those children that survived likely from the upper classes.

In conclusion, in the event of an disaster, it is best to be a woman, almost as good to be rich and fairly useless to be a child.