Douglas Locke
March 31, 2018
Adv. Quant Assignment #3
Logistic Regression

Objective

The objective of this study is to better understand the sinking of the ship HMS Titanic in 1912. Specifically, I will attempt to document and understand the factors that best led to survival. Using logistic regression, factors such as class, age, and gender will be evaluated in how well they can predict survival.

Literature Review

Given the highly publicized nature of the Titanic sinking and its continued embodiment in the general culture, there exists many books and films regarding the tragedy. The famous New York Herald headline from 1912 actually states both the general survivor rate (675 survived out of 1800 on board) and tell us something about who is saved "mostly women and children." For this study I also examined "Gender, Social Norms, and Survival in Maritime Disasters" (Elinder, Erixson, 2012). They found that in studying many maritime disasters, women actually typically were at a survival disadvantage:.

> "Women have a distinct survival disadvantage compared with men. Captains and crew survive at a significantly higher rate than passengers. We also find that: the captain has the power to enforce normative behavior; there seems to be no association between duration of a disaster and the impact of social norms; women fare no better when they constitute a small share of the ship's complement; the length of the voyage before the disaster appears to have no impact on women's relative survival rate; the sex gap in survival rates has declined since World War I; and women have a larger disadvantage in British shipwrecks. "

In "Social Class and Survival on the S.S Titantic" (Hall, 1986) the author remarks that "in third class more women and children survived than did men and persons of unknown sex." The authors offer an extensive discussion of class and survival aboard the Titanic. The key passage is as follows:

> "The factors that seem to be of relevance in explaining the social class differences in survival were: (1) the positioning of the lifeboats on the deck where first and second class passengers were located; (2) a policy of looking after the first and second class passengers first; (3) neglect of third class passengers who were left to fend for themselves, and who could only find their way to the boat deck by trial and error; and (4) some unsystematic exclusion of third class passengers from the boat deck by members of the crew."

Summary of Assumptions

<u>Dependent Variable:</u> Survival

<u>Research based assumptios:</u> Given these prior findings, I expect age, gender, and class to play a significant role in the model. It seem that gender at least historically was known to be a significant factor, and class seems to be a historically under appreciated, at least until the 1980s.

<u>Intuition based assumptions:</u> I would expect proximity to the life boats to be important. Class or ticket may be proxies for this. Point of embarkation I would not expect to matter significantly, but still could be tested for learning.

<u>Omitted Variables:</u>

A variable that would be interesting but impossible to have data would be the strength of each passengers belief that another ship would rescue them. Tickets looks interesting to parse, but I did not given time constraints.

Variable Order of Entry:

Entry	Variable	Definition	Assumption Notes for Survival	Expected Importance
1	sex	Sex, coded as male/female	Assume female	High
2	Age	Age in years, coded in bins of 10 years	Assume youngest	High
3	pclass	Ticket class, coded in 3 dummy variables	Assume class 1, then 2	High
4	sibsp	# of siblings / spouses aboard the Titanic, Code in dummies of "None, Single, Multi"	Assume Multi, then Single	Medium
5	parch	# of parents / children aboard the Titanic Code in dummies of "None, Single, Multi"	Assume Multi, then Single	Medium
6	fare	Passenger fare, coded in dummy variable of upper quartile passenger fare	Assume high fares	Low
7	ticket	Ticket number	Could be of value to parse and test, not enough time	Low
8	cabin	Cabin Number	Could be of value to parse and test, not enough time	None

Summary of Data Acquisition & Preparation

The training data set contains 710 observations. However, only 564 observations have an "age" variable. The observations with missing age variables were omitted from the model building process.

Summary of Model

In my final model, most significant ($p < .001$) variables were Sex (Female) , Passenger Class (1 and 2), and Age 0-10. The next group of variables (where $p < .01$) were Age 31-40 and if the passenger had multiple (2+) siblings. Finally, the next variables (where $p < .05$) were Ages 11-20 and Ages 21-30.

The model overall accuracy was 80.6 % accurate on training data. Testing data was accurate at 80.4%.

I was surprised that the Parents/Children variable (coded as dummy variables for None, Single, Multi) was not significant. All attempts at using the Fare variable also didn't work. I tried it also coded as dummies including for values greater than the 75th and 90th percentiles. Neither was significant.

Data Set Description

Variable	Definition	Key	Notes
survival	Survival	0 = No, 1 = Yes	
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd	A proxy for socio-economic status (SES) 1st = Upper 2nd = Middle 3rd = Lower
sex	Sex		
Age	Age in years		Age is fractional if less than 1. If the age is estimated, is it in the form of xx.5
sibsp	# of siblings / spouses aboard the Titanic		The dataset defines family relations in this way... Sibling = brother, sister, stepbrother, stepsister Spouse = husband, wife (mistresses and fiancés were ignored)
parch	# of parents / children aboard the Titanic		The dataset defines family relations in this way... Parent = mother, father Child = daughter, son, stepdaughter, stepson Some children travelled only with a nanny, therefore parch=0 for them.
ticket	Ticket number		
fare	Passenger fare		
cabin	Cabin number		
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton	

Analysis Process

Data Checks
Removed all observations where age value was missing.

```
train3 = na.omit(train)
```

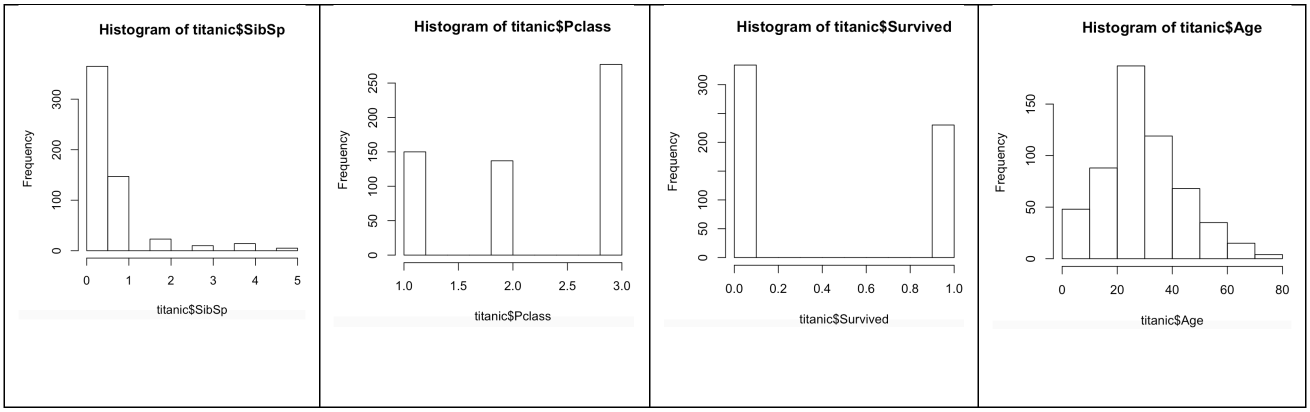
```
> describe(train3)
            vars   n   mean     sd median trimmed    mad  min    max  range  skew kurtosis   se
X              1 564 355.89 205.85  352.0  355.40 264.64 1.00 711.00 710.00  0.02    -1.20 8.67
PassengerId    2 564 355.89 205.85  352.0  355.40 264.64 1.00 711.00 710.00  0.02    -1.20 8.67
Survived       3 564   0.41   0.49    0.0    0.38   0.00 0.00   1.00   1.00  0.37    -1.86 0.02
Pclass         4 564   2.23   0.84    2.0    2.28   1.48 1.00   3.00   2.00 -0.44    -1.45 0.04
Name*          5 564 335.21 209.40  314.0  331.38 269.09 1.00 710.00 709.00  0.15    -1.23 8.82
Sex*           6 564   1.63   0.48    2.0    1.67   0.00 1.00   2.00   1.00 -0.55    -1.70 0.02
Age            7 564  30.02  14.61   28.0   29.52  12.97 0.75  80.00  79.25  0.40     0.11 0.62
SibSp          8 564   0.54   0.96    0.0    0.32   0.00 0.00   5.00   5.00  2.44     6.56 0.04
Parch          9 564   0.44   0.87    0.0    0.25   0.00 0.00   6.00   6.00  2.59     8.62 0.04
Ticket*       10 564 280.25 168.59  273.5  280.14 231.29 1.00 564.00 563.00  0.02    -1.29 7.10
Fare          11 564  35.17  51.61   16.1   23.98  12.76 0.00 512.33 512.33  4.46    29.22 2.17
Embarked*     12 564   2.58   0.79    3.0    2.72   0.00 1.00   3.00   2.00 -1.42     0.11 0.03
```

Several variables are categorical (Name, Sex, Ticket, Embarked, PassengerID, PClass).
Others are binary categorical (Survived).
SibSp, Parch, Age, Fare are quantitative.

We can see from the summary statistics the mean passenger age is 30, 66% of the values (1 standard deviation) are within 14.61 years (assuming a normal distribution, which it is not, because there is positive skew). The youngest passenger is < 1 years old, the oldest is 80.

```
hist(titanic$SibSp)
hist(titanic$Pclass)
```

```
hist(titanic$Survived)
hist(titanic$Age)
```




```
> table(titanic$Sex)

female   male 
   207    357 
> table(titanic$SibSp)

  0   1   2   3   4   5 
365 147  23  10  14   5 
> table(titanic$Pclass)

  1   2   3 
150 137 277 
> table(titanic$Survived)

  0   1 
334 230 
```

Frequencies Across the Data Set:

From these frequencies we can see there are more male passengers (63%) than female (36%).

The SibSp code tells us that 365 (64%) had no sibilings or spouses onboard. 157 (27%) has at least one sibling or spouse, and the rest (9%) had 2-5 siblings/spouses on board.

The P-Class variable tells us 277 or 49% of the passengers were in 3rd class, the rest were split about evenly between 1st and 2nd class (about 26% and 24% respectively).

With cross tables, we can see that in the 3rd class, 70% of the passengers were men.

Frequencies Across the Survival Variable:

The survival variable tells us that 334 (59%) passengers did not survive, while 230 (41%) did survive.

We can also see that only 54% of children 16 and under survived, but this is above the total 41% survival rate.
Across all survivors, 83% are adults.

From pClass, we can see that amongst survivors, there were 41% first class, and 29% each in both 2nd and 3rd class.

```
titanic$isChild[titanic$Age <=16] <- 1 ; titanic$isChild[titanic$Age > 16] <- 0 ;
CrossTable(titanic$isChild, titanic$Survived, expected = TRUE, format="SPSS")


   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
```

```
|              Row Percent |
|           Column Percent |
|            Total Percent |
|-------------------------|

Total Observations in Table:  564

                 | titanic$Survived
titanic$isChild |         0  |         1  | Row Total |
----------------|-----------|-----------|-----------|
              0 |       300  |       190  |       490  |
                |   290.177  |   199.823  |           |
                |     0.333  |     0.483  |           |
                |    61.224% |    38.776% |    86.879% |
                |    89.820% |    82.609% |           |
                |    53.191% |    33.688% |           |
----------------|-----------|-----------|-----------|
              1 |        34  |        40  |        74  |
                |    43.823  |    30.177  |           |
                |     2.202  |     3.197  |           |
                |    45.946% |    54.054% |    13.121% |
                |    10.180% |    17.391% |           |
                |     6.028% |     7.092% |           |
----------------|-----------|-----------|-----------|
   Column Total |       334  |       230  |       564  |
                |    59.220% |    40.780% |           |
----------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  6.214362     d.f. =  1     p =  0.0126718

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  5.59781     d.f. =  1     p =  0.01798294


       Minimum expected frequency: 30.1773



> CrossTable(titanic$Sex, titanic$Pclass, expected = TRUE, format="SPSS")

Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|-------------------------|

Total Observations in Table:  564

             | titanic$Pclass
 titanic$Sex |         1  |         2  |         3  | Row Total |
-------------|-----------|-----------|-----------|-----------|
      female |        64  |        60  |        83  |       207  |
             |    55.053  |    50.282  |   101.665  |           |
             |     1.454  |     1.878  |     3.427  |           |
             |    30.918% |    28.986% |    40.097% |    36.702% |
             |    42.667% |    43.796% |    29.964% |           |
             |    11.348% |    10.638% |    14.716% |           |
-------------|-----------|-----------|-----------|-----------|
        male |        86  |        77  |       194  |       357  |
             |    94.947  |    86.718  |   175.335  |           |
             |     0.843  |     1.089  |     1.987  |           |
             |    24.090% |    21.569% |    54.342% |    63.298% |
             |    57.333% |    56.204% |    70.036% |           |
             |    15.248% |    13.652% |    34.397% |           |
-------------|-----------|-----------|-----------|-----------|
Column Total |       150  |       137  |       277  |       564  |
             |    26.596% |    24.291% |    49.113% |           |
```

```
-------------|-----------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  10.67797     d.f. =  2     p =  0.004800736




> CrossTable(titanic$Pclass, titanic$Survived, expected = TRUE, format="SPSS")

   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|-------------------------|

Total Observations in Table:  564

               | titanic$Survived
titanic$Pclass |        0  |        1  | Row Total |
---------------|-----------|-----------|-----------|
             1 |       55  |       95  |      150  |
               |   88.830  |   61.170  |           |
               |   12.884  |   18.709  |           |
               |   36.667% |   63.333% |   26.596% |
               |   16.467% |   41.304% |           |
               |    9.752% |   16.844% |           |
---------------|-----------|-----------|-----------|
             2 |       69  |       68  |      137  |
               |   81.131  |   55.869  |           |
               |    1.814  |    2.634  |           |
               |   50.365% |   49.635% |   24.291% |
               |   20.659% |   29.565% |           |
               |   12.234% |   12.057% |           |
---------------|-----------|-----------|-----------|
             3 |      210  |       67  |      277  |
               |  164.039  |  112.961  |           |
               |   12.878  |   18.700  |           |
               |   75.812% |   24.188% |   49.113% |
               |   62.874% |   29.130% |           |
               |   37.234% |   11.879% |           |
---------------|-----------|-----------|-----------|
  Column Total |      334  |      230  |      564  |
               |   59.220% |   40.780% |           |
---------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  67.61897     d.f. =  2     p =  0.000000000000002073614



       Minimum expected frequency: 55.86879
```

Multi-Collinearity Concerns:

There may be multi-collinearity between fare and p-class. However p-class I suspect to be a stronger determinant of survival – it speaks more directly to passenger position and the ways the passengers are treated on the ship. I decided to code Fare as a dummy, and first tried the upper quartile of Fare, then the 90th percentile of Fare. A cross table between my upper Quartile of Fare and pClass showed that 80% of high fare were in 1st class. However the rest were split between 2nd and 3rd class.

Final Model Build

Confusion Matrix:

	Predicted Died	Predicted Survived	
Actual Died	276	51	327
Actual Survived	58	179	158
	334	230	

The overall model accuracy was .806. At one point, I had the model accuracy at .812 but this was at the inclusion of the Parent variables (Presence of Multiple Parents/Children), but it was insignificant, so I removed it.

The model showed at least 10 points spread between quartile probabilities.

Interpreting the co-efficients, we can say

The odds of survival are ….

… ~ 12.9 times greater if the passenger is a female vs male
… ~ 11.3 times greater if the passenger is between 0 and 10 years old
… ~ 10.4 times greater if the passenger is in first class vs not in first class
… ~ 3.3 times greater if the passenger is in second class vs not in second class
… ~ 2.8 times greater if the passenger is between 21 and 30 years old
… ~ 2.8 times greater if the passenger is between 31 and 40 years old
… ~ 2.6 times greater if the passenger is between 11 and 20 years old
… ~ .24 times less likely if the person has multiple (2+) siblings

```
exp(coef(titanic_2))
  (Intercept)        female       Class_1       Class_2       Age_0_10     Age_11_20
   0.04701336   12.95817398   10.35702147    3.26626864   11.32706230    2.66223266
    Age_21_30     Age_31_40 SibSp_isMulti
   2.14155349    2.80706544    0.24511120
```

The Nagelkeke Pseudo R-squared tells us that the independent variables explain 49% of the variance of the dependent variable (survival).

Final Model Code & Results

```
titanic_2 <- glm(Survived ~ female + Class_1 + Class_2 + Age_0_10 + Age_11_20 + Age_21_30 + Age_31_40 +
SibSp_isMulti, family=binomial, data=titanic)
> summary(titanic_2)

Call:
glm(formula = Survived ~ female + Class_1 + Class_2 + Age_0_10 +
    Age_11_20 + Age_21_30 + Age_31_40 + SibSp_isMulti, family = binomial,
    data = titanic)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9268  -0.7541  -0.4380   0.5881   2.4913

Coefficients:
              Estimate Std. Error z value             Pr(>|z|)
(Intercept)    -3.0573     0.3419  -8.943 < 0.0000000000000002 ***
female          2.5617     0.2394  10.703 < 0.0000000000000002 ***
Class_1         2.3377     0.3060   7.640   0.0000000000000218 ***
Class_2         1.1836     0.2795   4.235   0.0000228094195617 ***
Age_0_10        2.4272     0.5294   4.585   0.0000045438893759 ***
Age_11_20       0.9792     0.4036   2.426              0.01527 *
Age_21_30       0.7615     0.3250   2.344              0.01910 *
Age_31_40       1.0321     0.3422   3.016              0.00256 **
SibSp_isMulti  -1.4060     0.4581  -3.069              0.00215 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 762.58  on 563  degrees of freedom
Residual deviance: 509.55  on 555  degrees of freedom
AIC: 527.55

Number of Fisher Scoring iterations: 5

> exp(coef(titanic_2))
  (Intercept)        female       Class_1       Class_2      Age_0_10     Age_11_20
   0.04701336   12.95817398   10.35702147    3.26626864   11.32706230    2.66223266
    Age_21_30     Age_31_40 SibSp_isMulti
   2.14155349    2.80706544    0.24511120
>
> prob_2 = predict(titanic_2,type="response")
> titanic$prob_2 <- prob_2
>
> quantile(titanic$prob_2)
        0%        25%        50%        75%       100%
0.02408385 0.11123790 0.32746811 0.63100783 0.98620098
> # (2) Run Diagnostics
>
> pred_2 = rep("Died", 564)
> pred_2[titanic$prob_2>0.50] = "Survived"
> titanic$pred_2 <- pred_2
>
> table(pred_2,titanic$Survived_value)

pred_2     Died Survived
  Died      276       51
  Survived   58      179

> misClassifiError = mean(pred_2 != titanic$Survived_value)
> print(paste('Accuracy', 1 - misClassifiError))
[1] "Accuracy 0.806737588652482"

> logisticPseudoR2s(titanic_2)
Pseudo R^2 for Logistic Regression
Hotitanicer and Lemeshow R^2  0.332
Cox and Snell R^2        0.362
Nagelkerke R^2           0.488
```

Extensions

I ran the test set of data. The same missing age values problem exists. I removed missing age values, for a total of 148 records.
I create a separate code file for this.

The first part of the code preps the variables & rebuilds the model on the train data exactly as before.
The next of the code then creates all the variables on the test set and prepares it for running against the training model.

The test data set accuracy is .804, which is just .002 less what was achieved during the training.
It seems the model is well-fit to the data.

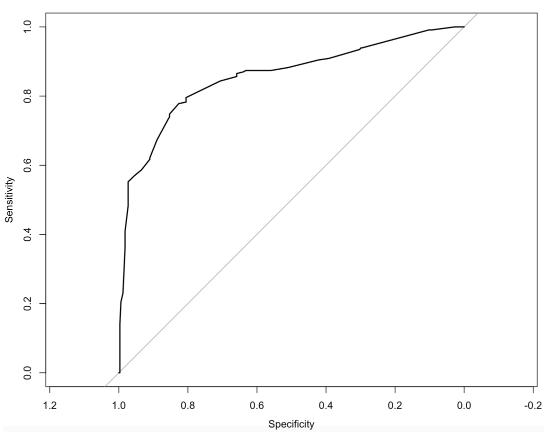
```
pred_final Died Survived
  Died       72       11
  Survived   18       47

> misClassifiError = mean(pred_final != titanic_test$Survived_value)
> print(paste('Accuracy', 1 - misClassifiError))
[1] "Accuracy 0.804054054054054"
```

Rocplot of final training model

```
rocplot <- plot.roc(titanic$Survived_value,titanic$prob_final)
plot(rocplot)
```

TRAINING



CODE APPENDIX

CODE TO BUILD TRAIN MODEL & PREDICT WITH TEST DATA

```
# Douglas Locke
# Adv Quant  3-31-2018
# This file build the final model using the train data
# and then runs the testing data against the trained model

install.packages("gmodels")
install.packages("LogisticDx")
install.packages("psych")
install.packages("car")

# Now, lets load them into our current working session.

library(gmodels)
library(LogisticDx)
library(psych)
library(car)

# We will leverage this function as well. Run the below code to load it into your environment.

logisticPseudoR2s <- function(LogModel) {
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <- length(LogModel$fitted.values)
  R.l <- 1 - dev / nullDev
  R.cs <- 1 - exp ( -(nullDev - dev) / modelN)
  R.n <- R.cs / (1 - ( exp (-(nullDev / modelN))))
  cat("Pseudo R^2 for Logistic Regression\n")
  cat("Hotitanicer and Lemeshow R^2 ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2        ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2           ", round(R.n, 3),     "\n")
}

# Remove scientific notation
options(scipen=999)

#  ---------- LOAD DATA ----------

train <- read.csv("train_set.csv")
titanic_train <- train
View(titanic_train)

test <- read.csv("test_set.csv")
titanic_test <- test
View(titanic_test)


train = na.omit(titanic_train)
describe(train)

test = na.omit(titanic_test)
describe(test)

#no null values
titanic <- train
titanic_test <- test


#build dummies & Variable prep TRAIN
titanic$Embarked_Q[titanic$Embarked == "Q"] <- 1 ; titanic$Embarked_Q[titanic$Embarked == "S"] <- 0 ;
titanic$Embarked_Q[titanic$Embarked == "C"] <- 0
titanic$Embarked_S[titanic$Embarked == "Q"] <- 0 ; titanic$Embarked_S[titanic$Embarked == "S"] <- 1 ;
titanic$Embarked_S[titanic$Embarked == "C"] <- 0
titanic$Embarked_C[titanic$Embarked == "Q"] <- 0 ; titanic$Embarked_C[titanic$Embarked == "S"] <- 0 ;
titanic$Embarked_C[titanic$Embarked == "C"] <- 1
titanic$female[titanic$Sex == "male"] <- 0 ; titanic$female[titanic$Sex == "female"] <- 1
titanic$male[titanic$Sex == "female"] <- 0 ; titanic$male[titanic$Sex == "male"] <- 1
titanic$Class_1[titanic$Pclass == "1"] <- 1 ; titanic$Class_1[titanic$Pclass == "2"] <- 0 ;
titanic$Class_1[titanic$Pclass == "3"] <- 0
titanic$Class_2[titanic$Pclass == "1"] <- 0 ; titanic$Class_2[titanic$Pclass == "2"] <- 1 ;
titanic$Class_2[titanic$Pclass == "3"] <- 0
titanic$Class_3[titanic$Pclass == "1"] <- 0 ; titanic$Class_3[titanic$Pclass == "2"] <- 0 ;
titanic$Class_3[titanic$Pclass == "3"] <- 1
```

```
titanic$SibSp_isMulti[titanic$SibSp > 1] <- 1 ; titanic$SibSp_isMulti[titanic$SibSp <= 1] <- 0 ;
titanic$SibSp_isOne[titanic$SibSp == 1] <- 1 ; titanic$SibSp_isOne[titanic$SibSp != 1] <- 0 ;
titanic$SibSp_isNone[titanic$SibSp == 0] <- 1 ; titanic$SibSp_isNone[titanic$SibSp > 0] <- 0 ;
titanic$Parch_isMulti[titanic$Parch > 1] <- 1 ; titanic$Parch_isMulti[titanic$Parch <= 1] <- 0 ;
titanic$Parch_isOne[titanic$Parch == 1] <- 1 ; titanic$Parch_isOne[titanic$Parch != 1] <- 0 ;
titanic$Parch_isNone[titanic$Parch == 0] <- 1 ; titanic$Parch_isNone[titanic$Parch > 0] <- 0 ;
titanic$Age_0_10[titanic$Age > 0]<- 0 ; titanic$Age_0_10[titanic$Age <=10] <- 1 ;
titanic$Age_11_20[titanic$Age > 0] <- 0 ; titanic$Age_11_20[titanic$Age > 10 & titanic$Age <=20 ] <- 1 ;
titanic$Age_21_30[titanic$Age > 0] <- 0 ; titanic$Age_21_30[titanic$Age > 20 & titanic$Age <=30 ] <- 1 ;
titanic$Age_31_40[titanic$Age > 0] <- 0 ; titanic$Age_31_40[titanic$Age > 30 & titanic$Age <=40 ] <- 1 ;
titanic$Age_41_50[titanic$Age > 0] <- 0 ; titanic$Age_41_50[titanic$Age > 40 & titanic$Age <=50 ] <- 1 ;
titanic$Age_51_60[titanic$Age > 0] <- 0 ; titanic$Age_51_60[titanic$Age > 50 & titanic$Age <=60 ] <- 1 ;
titanic$Age_61_70[titanic$Age > 0] <- 0 ; titanic$Age_61_70[titanic$Age > 60 & titanic$Age <=70 ] <- 1 ;
titanic$Age_71_110[titanic$Age > 0] <- 0 ; titanic$Age_71_110[titanic$Age > 70 & titanic$Age <=110 ] <- 1 ;
titanic$isHighFare[titanic$Fare >=120] <- 1 ; titanic$isHighFare[titanic$Fare < 120 ] <- 0 ;
titanic$Survived_value[titanic$Survived == 0] <- "Died" ; titanic$Survived_value[titanic$Survived == 1] <-
"Survived"

#build dummies & Variable prep TEST
titanic_test$Embarked_Q[titanic_test$Embarked == "Q"] <- 1 ; titanic_test$Embarked_Q[titanic_test$Embarked ==
"S"] <- 0 ; titanic_test$Embarked_Q[titanic_test$Embarked == "C"] <- 0
titanic_test$Embarked_S[titanic_test$Embarked == "Q"] <- 0 ; titanic_test$Embarked_S[titanic_test$Embarked ==
"S"] <- 1 ; titanic_test$Embarked_S[titanic_test$Embarked == "C"] <- 0
titanic_test$Embarked_C[titanic_test$Embarked == "Q"] <- 0 ; titanic_test$Embarked_C[titanic_test$Embarked ==
"S"] <- 0 ; titanic_test$Embarked_C[titanic_test$Embarked == "C"] <- 1
titanic_test$female[titanic_test$Sex == "male"] <- 0 ; titanic_test$female[titanic_test$Sex == "female"] <- 1
titanic_test$male[titanic_test$Sex == "female"] <- 0 ; titanic_test$male[titanic_test$Sex == "male"] <- 1
titanic_test$Class_1[titanic_test$Pclass == "1"] <- 1 ; titanic_test$Class_1[titanic_test$Pclass == "2"] <- 0 ;
titanic_test$Class_1[titanic_test$Pclass == "3"] <- 0
titanic_test$Class_2[titanic_test$Pclass == "1"] <- 0 ; titanic_test$Class_2[titanic_test$Pclass == "2"] <- 1 ;
titanic_test$Class_2[titanic_test$Pclass == "3"] <- 0
titanic_test$Class_3[titanic_test$Pclass == "1"] <- 0 ; titanic_test$Class_3[titanic_test$Pclass == "2"] <- 0 ;
titanic_test$Class_3[titanic_test$Pclass == "3"] <- 1
titanic_test$SibSp_isMulti[titanic_test$SibSp > 1] <- 1 ; titanic_test$SibSp_isMulti[titanic_test$SibSp <= 1] <-
0 ;
titanic_test$SibSp_isOne[titanic_test$SibSp == 1] <- 1 ; titanic_test$SibSp_isOne[titanic_test$SibSp != 1] <- 0
;
titanic_test$SibSp_isNone[titanic_test$SibSp == 0] <- 1 ; titanic_test$SibSp_isNone[titanic_test$SibSp > 0] <- 0
;
titanic_test$Parch_isMulti[titanic_test$Parch > 1] <- 1 ; titanic_test$Parch_isMulti[titanic_test$Parch <= 1] <-
0 ;
titanic_test$Parch_isOne[titanic_test$Parch == 1] <- 1 ; titanic_test$Parch_isOne[titanic_test$Parch != 1] <- 0
;
titanic_test$Parch_isNone[titanic_test$Parch == 0] <- 1 ; titanic_test$Parch_isNone[titanic_test$Parch > 0] <- 0
;
titanic_test$Age_0_10[titanic_test$Age > 0]<- 0 ; titanic_test$Age_0_10[titanic_test$Age <=10] <- 1 ;
titanic_test$Age_11_20[titanic_test$Age > 0] <- 0 ; titanic_test$Age_11_20[titanic_test$Age > 10 &
titanic_test$Age <=20 ] <- 1 ;
titanic_test$Age_21_30[titanic_test$Age > 0] <- 0 ; titanic_test$Age_21_30[titanic_test$Age > 20 &
titanic_test$Age <=30 ] <- 1 ;
titanic_test$Age_31_40[titanic_test$Age > 0] <- 0 ; titanic_test$Age_31_40[titanic_test$Age > 30 &
titanic_test$Age <=40 ] <- 1 ;
titanic_test$Age_41_50[titanic_test$Age > 0] <- 0 ; titanic_test$Age_41_50[titanic_test$Age > 40 &
titanic_test$Age <=50 ] <- 1 ;
titanic_test$Age_51_60[titanic_test$Age > 0] <- 0 ; titanic_test$Age_51_60[titanic_test$Age > 50 &
titanic_test$Age <=60 ] <- 1 ;
titanic_test$Age_61_70[titanic_test$Age > 0] <- 0 ; titanic_test$Age_61_70[titanic_test$Age > 60 &
titanic_test$Age <=70 ] <- 1 ;
titanic_test$Age_71_110[titanic_test$Age > 0] <- 0 ; titanic_test$Age_71_110[titanic_test$Age > 70 &
titanic_test$Age <=110 ] <- 1 ;
titanic_test$isHighFare[titanic_test$Fare >=120] <- 1 ; titanic_test$isHighFare[titanic_test$Fare < 120 ] <- 0 ;
titanic_test$Survived_value[titanic_test$Survived == 0] <- "Died" ;
titanic_test$Survived_value[titanic_test$Survived == 1] <- "Survived"

# (1) Create Model

titanic_final <- glm(Survived ~ female + Class_1 + Class_2 + Age_0_10 + Age_11_20 + Age_21_30 + Age_31_40 +
SibSp_isMulti, family=binomial, data=titanic)
summary(titanic_final)
exp(coef(titanic_final))



prob_final = predict(titanic_final, titanic_test, type="response")
titanic_test$prob_final <- prob_final
quantile(titanic_test$prob_final)
# (2) Run Diagnostics

pred_final = rep("Died", 148)
```

```
pred_final[titanic_test$prob_final>0.50] = "Survived"
titanic_test$pred_final <- pred_final

table(pred_final,titanic_test$Survived_value)
misClassifiError = mean(pred_final != titanic_test$Survived_value)
print(paste('Accuracy', 1 - misClassifiError))
```

TRAINING MODEL ONLY WITH EXPLORATORY ANALYSIS CODE

```
# Doug Locke
# 3-31-2018
# Build & evaluate TRAINING model

install.packages("gmodels")
install.packages("LogisticDx")
install.packages("psych")
install.packages("car")

library(gmodels)
library(LogisticDx)
library(psych)
library(car)

# We will leverage this function as well.

logisticPseudoR2s <- function(LogModel) {
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <- length(LogModel$fitted.values)
  R.l <- 1 - dev / nullDev
  R.cs <- 1 - exp ( -(nullDev - dev) / modelN)
  R.n <- R.cs / (1 - ( exp (-(nullDev / modelN))))
  cat("Pseudo R^2 for Logistic Regression\n")
  cat("Hotitanicer and Lemeshow R^2 ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2       ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2          ", round(R.n, 3),     "\n")
}

# Remove scientific notation
options(scipen=999)

#  ---------- LOAD DATA ----------

train <- read.csv("train_set.csv")
titanic <- train
View(titanic)

# ---------- PRELIMINARY STEPS ----------

summary <- describe(titanic)
View(summary)

train3 = na.omit(train)
describe(train3)
titanic <- train3

# EXPLORATORY ANALYSIS
hist(titanic$SibSp)
hist(titanic$Pclass)
hist(titanic$Survived)
hist(titanic$Age)
hist(titanic$Fare)

quantile(titanic$Fare)
quantile(titanic$Fare, 0.95)
#34.86

table(titanic$Sex)
table(titanic$SibSp)
table(titanic$Pclass)
table(titanic$Survived)
table(titanic$Parch)

# crosstables
```

```r
CrossTable(titanic$Sex, titanic$Pclass, expected = TRUE, format="SPSS")  # run the same with the other
variables, in various combinations. Discuss your findings
CrossTable(titanic$Sex, titanic$Survived, expected = TRUE, format="SPSS")

titanic$isChild[titanic$Age <=16] <- 1 ; titanic$isChild[titanic$Age > 16] <- 0 ;
CrossTable(titanic$isChild, titanic$Survived, expected = TRUE, format="SPSS")

CrossTable(titanic$Pclass, titanic$Survived, expected = TRUE, format="SPSS")
CrossTable(titanic$Pclass, titanic$isHighFare, expected = TRUE, format="SPSS")
CrossTable(titanic$Pclass, titanic$Sex, titanic$Survived, expected = TRUE, format="SPSS")

#  ---------- CORE ASSIGNMENT ----------

# NAIVE MODEL........

titanic$Embarked_Q[titanic$Embarked == "Q"] <- 1 ; titanic$Embarked_Q[titanic$Embarked == "S"] <- 0 ;
titanic$Embarked_Q[titanic$Embarked == "C"] <- 0
titanic$Embarked_S[titanic$Embarked == "Q"] <- 0 ; titanic$Embarked_S[titanic$Embarked == "S"] <- 1 ;
titanic$Embarked_S[titanic$Embarked == "C"] <- 0
titanic$Embarked_C[titanic$Embarked == "Q"] <- 0 ; titanic$Embarked_C[titanic$Embarked == "S"] <- 0 ;
titanic$Embarked_C[titanic$Embarked == "C"] <- 1
titanic$female[titanic$Sex == "male"] <- 0 ; titanic$female[titanic$Sex == "female"] <- 1
titanic$male[titanic$Sex == "female"] <- 0 ; titanic$male[titanic$Sex == "male"] <- 1
titanic$Class_1[titanic$Pclass == "1"] <- 1 ; titanic$Class_1[titanic$Pclass == "2"] <- 0 ;
titanic$Class_1[titanic$Pclass == "3"] <- 0
titanic$Class_2[titanic$Pclass == "1"] <- 0 ; titanic$Class_2[titanic$Pclass == "2"] <- 1 ;
titanic$Class_2[titanic$Pclass == "3"] <- 0
titanic$Class_3[titanic$Pclass == "1"] <- 0 ; titanic$Class_3[titanic$Pclass == "2"] <- 0 ;
titanic$Class_3[titanic$Pclass == "3"] <- 1
titanic$SibSp_isMulti[titanic$SibSp > 1] <- 1 ; titanic$SibSp_isMulti[titanic$SibSp <= 1] <- 0 ;
titanic$SibSp_isOne[titanic$SibSp == 1] <- 1 ; titanic$SibSp_isOne[titanic$SibSp != 1] <- 0 ;
titanic$SibSp_isNone[titanic$SibSp == 0] <- 1 ; titanic$SibSp_isNone[titanic$SibSp > 0] <- 0 ;
titanic$Parch_isMulti[titanic$Parch > 1] <- 1 ; titanic$Parch_isMulti[titanic$Parch <= 1] <- 0 ;
titanic$Parch_isOne[titanic$Parch == 1] <- 1 ; titanic$Parch_isOne[titanic$Parch != 1] <- 0 ;
titanic$Parch_isNone[titanic$Parch == 0] <- 1 ; titanic$Parch_isNone[titanic$Parch > 0] <- 0 ;
titanic$Age_0_10[titanic$Age > 0]<- 0 ; titanic$Age_0_10[titanic$Age <=10] <- 1 ;
titanic$Age_11_20[titanic$Age > 0] <- 0 ; titanic$Age_11_20[titanic$Age > 10 & titanic$Age <=20 ] <- 1 ;
titanic$Age_21_30[titanic$Age > 0] <- 0 ; titanic$Age_21_30[titanic$Age > 20 & titanic$Age <=30 ] <- 1 ;
titanic$Age_31_40[titanic$Age > 0] <- 0 ; titanic$Age_31_40[titanic$Age > 30 & titanic$Age <=40 ] <- 1 ;
titanic$Age_41_50[titanic$Age > 0] <- 0 ; titanic$Age_41_50[titanic$Age > 40 & titanic$Age <=50 ] <- 1 ;
titanic$Age_51_60[titanic$Age > 0] <- 0 ; titanic$Age_51_60[titanic$Age > 50 & titanic$Age <=60 ] <- 1 ;
titanic$Age_61_70[titanic$Age > 0] <- 0 ; titanic$Age_61_70[titanic$Age > 60 & titanic$Age <=70 ] <- 1 ;
titanic$Age_71_110[titanic$Age > 0] <- 0 ; titanic$Age_71_110[titanic$Age > 70 & titanic$Age <=110 ] <- 1 ;

#90th percentile for high fare
titanic$isHighFare[titanic$Fare >=120] <- 1 ; titanic$isHighFare[titanic$Fare < 120 ] <- 0 ;
titanic$Survived_value[titanic$Survived == 0] <- "Died" ; titanic$Survived_value[titanic$Survived == 1] <-
"Survived"
View(titanic)

# (1) Create Model

titanic_final <- glm(Survived ~ female + Class_1 + Class_2 + Age_0_10 + Age_11_20 + Age_21_30 + Age_31_40 +
SibSp_isMulti, family=binomial, data=titanic)
summary(titanic_final)
exp(coef(titanic_final))

prob_final = predict(titanic_final,type="response")
titanic$prob_final <- prob_final

quantile(titanic$prob_final)

pred_final = rep("Died", 564)
pred_final[titanic$prob_final>0.50] = "Survived"
titanic$pred_final <- pred_final

table(pred_final,titanic$Survived_value)
misClassifiError = mean(pred_final != titanic$Survived_value)
print(paste('Accuracy', 1 - misClassifiError))

exp(coef(titanic_final))
logisticPseudoR2s(titanic_final)

library(pROC)
rocplot <- plot.roc(titanic$Survived_value,titanic$prob_final)
plot(rocplot)
```