How to Survive the Titanic Using Logistic Regression Modeling

Michael Streyle

Carly Mester

Matt Foundos

## Abstract

In this data set we are able to examine real life data from the passengers on the RMS Titanic which sank in 1912. The response variable for the data is the survival status of the passenger. In the popular movie *Titanic*, Jack and Rose fall in love despite their differences in economic class. However, it is obvious throughout the movie that it is significantly harder for the lower class passengers to survive. Even though this is a movie, we are interested to see if there were any factors that made it significantly harder for someone to survive. We want to know which combination of predictor variables will give someone the best chance of survival. The data provides us with 10 different variables, some of which include name, passenger class, age, and sex. Since some of the variables were missing for a large number of people, we created new columns to to show whether data was recorded or not for some columns ie. age\_rec(whether the age was recorded or not). Throughout our analysis, we are looking to explore a few questions - what type of passenger has the lowest chance of survival? What is the probability the average Luther student survives? By utilizing logistic regression, we hope to be able to answer these questions as well as find the best possible model for predicting survival.

Logistic Regression - Titanic Survival

#### Michael Streyle, Carly Mester, Matt Foundos

First, we read in the data and did some cleaning.

#importing and cleaning the data   
ship <- read.table('http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic.txt', header = T, sep=',')  
library(base)  
library(Hmisc)

## Warning: package 'Hmisc' was built under R version 3.4.2

## Warning: package 'ggplot2' was built under R version 3.4.2

ship$pclass <- strtoi(substring(ship$pclass, 1, 1))  
colnames(ship) <- c("index", "pclass","survived","name","age","embarked","home.dest","room","ticket","boat","sex")

Below we are adding columns to the dataframe that are categorical based on missing data from other columns, as well as a column of the length of the passengers name, just for fun.

#new column of length of passenger's name  
library(stringr)  
ship$name\_length <- str\_count(ship$name)  
   
#new column of if ticket is reported  
ship$ticket\_rec <- ifelse(ship$ticket == '', yes=0, no=1)  
  
#new column of if room is reported  
ship$room\_rec <- ifelse(ship$room == '', yes=0, no=1)  
  
#new column of if home.dest is reported  
ship$home.dest\_rec <- ifelse(ship$home.dest == '', yes=0, no=1)  
  
#new column for if age was recorded  
ship$age\_rec <- ifelse(ship$age == '', yes=0, no=1)  
ship$age\_rec[is.na(ship$age\_rec)] <- 0  
  
#new column for age groups with 9 referencing missing   
ship$agegroup = ifelse(ship$age\_rec == 1, cut(ship$age,c(0,10, 20, 30, 40, 50, 60, 70, 80)), 9)  
  
  
attach(ship)

### Initial Exploration

The following is some initial exploration and individual variable plots: Some interesting things to point out: Missing age is the largest agegroup by nearly 500 people. The 3rd passenger class is the largest by nearly 400 people. The missing values category in agegroup causes the fitted line to decrease, however, it looks like it decreases slightly before agegroup=missing. The plot of age and survival by sex shows men obviously did not survive nearly as often as women did.

library(MASS)

## Warning: package 'MASS' was built under R version 3.4.2

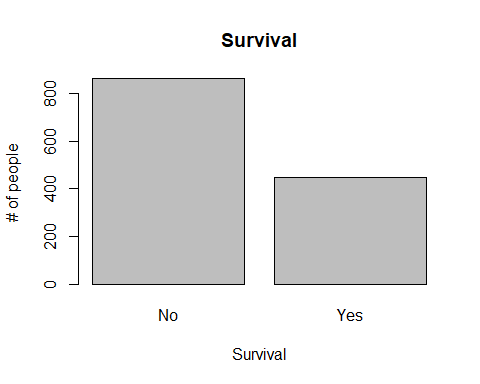
table1= xtabs(~survived+sex)  
table1

## sex  
## survived female male  
## 0 156 708  
## 1 307 142

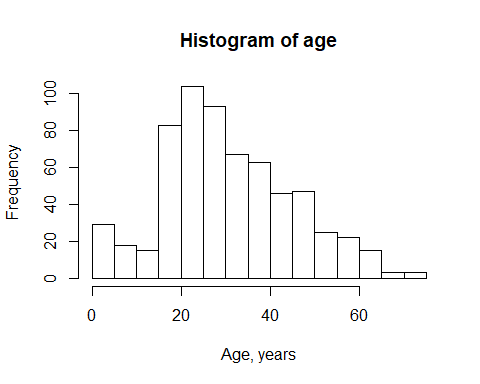
table2 = xtabs(~survived+pclass+sex)  
table2

## , , sex = female  
##   
## pclass  
## survived 1 2 3  
## 0 9 13 134  
## 1 134 94 79  
##   
## , , sex = male  
##   
## pclass  
## survived 1 2 3  
## 0 120 148 440  
## 1 59 25 58

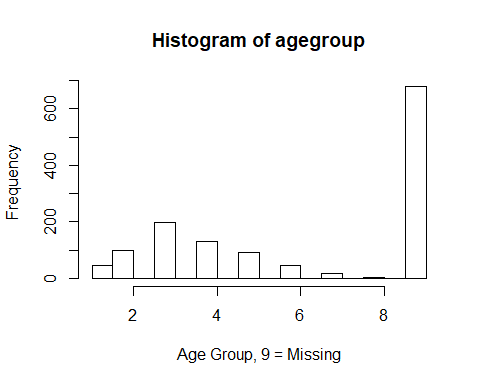
barplot (table (ifelse (survived==1, "Yes", "No")), xlab="Survival", ylab="# of people", main='Survival')



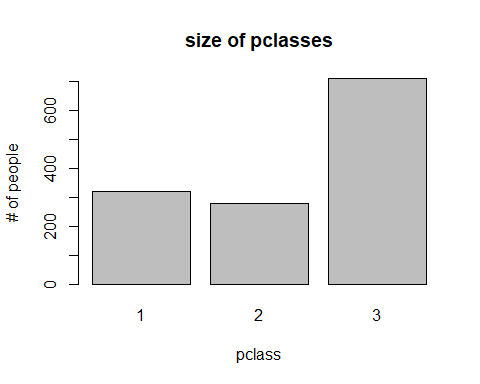
hist (age, xlab="Age, years")



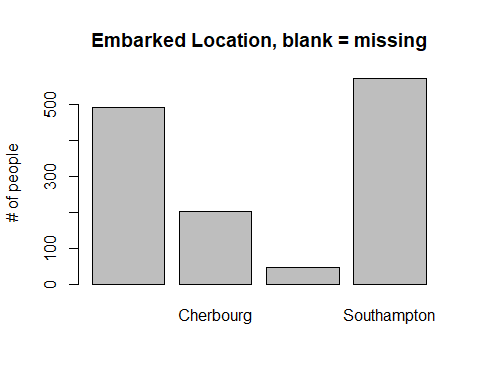
hist (agegroup, xlab="Age Group, 9 = Missing")



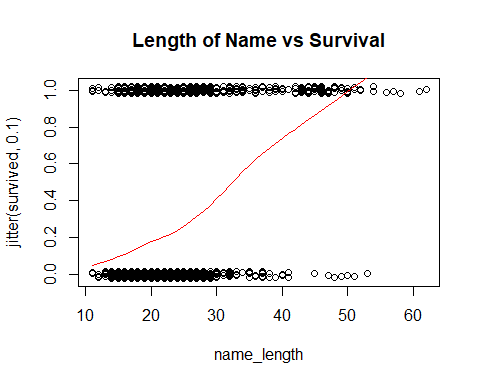
barplot(table(pclass), xlab='pclass', ylab='# of people', main='size of pclasses')



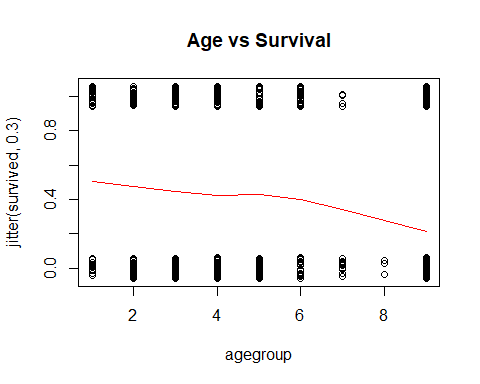
barplot(table(embarked), main='Embarked Location, blank = missing', ylab='# of people')



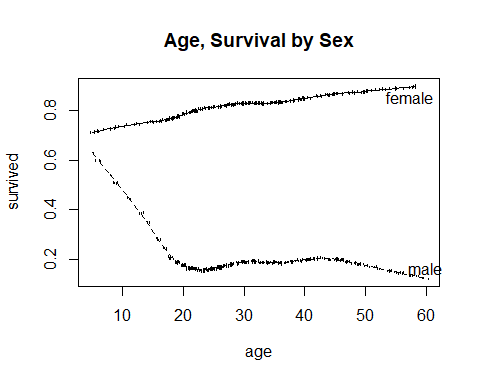
#plot survival vs length of name  
plot (jitter (survived, 0.1) ~ name\_length, main='Length of Name vs Survival')  
lines (lowess (name\_length, survived), col='red')



#plotting survival vs agegroup  
plot (jitter (survived, 0.3) ~ agegroup, main='Age vs Survival')  
lines (lowess(agegroup, survived), col='red')



#pairs plot   
#pairs(ship)  
  
  
  
#this plsmo function is super useful to see interactions  
plsmo(age, survived, group=sex, datadensity=T, main='Survival vs Age grouped by Gender') # or group=pclass  
title('Age, Survival by Sex')



Next, we will create a first order model. We opted to use the ticket\_rec, room\_rec, and home.dest\_rec variables because they are all factor variables with a large number of levels, and we wanted to keep our degrees of freedom as high as possible. Also, these variables did not seem like they would logically be very significant, and home.dest has formatting inconsistencies. We also thought that these variables may not have been very accurately recorded. The coefficients are transformed back from the logit scale, resulting in odds ratios. An example of interpretting is ' for every one unit increase in name\_length, the passengers odds of surviving increases by 2.18 percent'.

attach(ship)  
#creating a first order model  
ship.logit <- glm(survived ~ pclass + name\_length + as.factor(agegroup) + ticket\_rec + room\_rec + as.factor(embarked) + home.dest\_rec + as.factor(sex), data=ship,family=binomial)  
  
summary(ship.logit)

##   
## Call:  
## glm(formula = survived ~ pclass + name\_length + as.factor(agegroup) +   
## ticket\_rec + room\_rec + as.factor(embarked) + home.dest\_rec +   
## as.factor(sex), family = binomial, data = ship)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8504 -0.6462 -0.3420 0.5450 2.4708   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.48189 0.75145 4.634 3.59e-06 \*\*\*  
## pclass -0.88803 0.13334 -6.660 2.74e-11 \*\*\*  
## name\_length 0.02158 0.01093 1.974 0.048403 \*   
## as.factor(agegroup)2 -1.38126 0.45688 -3.023 0.002501 \*\*   
## as.factor(agegroup)3 -1.85789 0.42477 -4.374 1.22e-05 \*\*\*  
## as.factor(agegroup)4 -1.60327 0.43990 -3.645 0.000268 \*\*\*  
## as.factor(agegroup)5 -2.16478 0.46884 -4.617 3.89e-06 \*\*\*  
## as.factor(agegroup)6 -2.29185 0.55853 -4.103 4.07e-05 \*\*\*  
## as.factor(agegroup)7 -4.00577 0.83961 -4.771 1.83e-06 \*\*\*  
## as.factor(agegroup)8 -15.77621 494.36448 -0.032 0.974542   
## as.factor(agegroup)9 -1.67688 0.43391 -3.865 0.000111 \*\*\*  
## ticket\_rec 0.74818 0.36488 2.050 0.040316 \*   
## room\_rec 0.41567 0.34000 1.223 0.221500   
## as.factor(embarked)Cherbourg -0.35570 0.38808 -0.917 0.359362   
## as.factor(embarked)Queenstown -0.68641 0.52501 -1.307 0.191067   
## as.factor(embarked)Southampton -0.71985 0.36823 -1.955 0.050595 .   
## home.dest\_rec 1.17255 0.34816 3.368 0.000758 \*\*\*  
## as.factor(sex)male -2.38233 0.16863 -14.127 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1686.8 on 1312 degrees of freedom  
## Residual deviance: 1120.3 on 1295 degrees of freedom  
## AIC: 1156.3  
##   
## Number of Fisher Scoring iterations: 13

#odds ratios  
expbetas = exp(ship.logit$coefficients)  
expbetas

## (Intercept) pclass   
## 3.252107e+01 4.114675e-01   
## name\_length as.factor(agegroup)2   
## 1.021816e+00 2.512628e-01   
## as.factor(agegroup)3 as.factor(agegroup)4   
## 1.560019e-01 2.012375e-01   
## as.factor(agegroup)5 as.factor(agegroup)6   
## 1.147754e-01 1.010797e-01   
## as.factor(agegroup)7 as.factor(agegroup)8   
## 1.821018e-02 1.407606e-07   
## as.factor(agegroup)9 ticket\_rec   
## 1.869570e-01 2.113145e+00   
## room\_rec as.factor(embarked)Cherbourg   
## 1.515383e+00 7.006806e-01   
## as.factor(embarked)Queenstown as.factor(embarked)Southampton   
## 5.033797e-01 4.868262e-01   
## home.dest\_rec as.factor(sex)male   
## 3.230218e+00 9.233512e-02

Next, we use stepwise elimination with both directions, using AIC criteria, which results in the following model. The process removed the room\_rec variable and the embarked location factor variable

survived ~ pclass + name\_length + agegroup + ticket\_rec + home.dest\_rec + as.factor(sex)

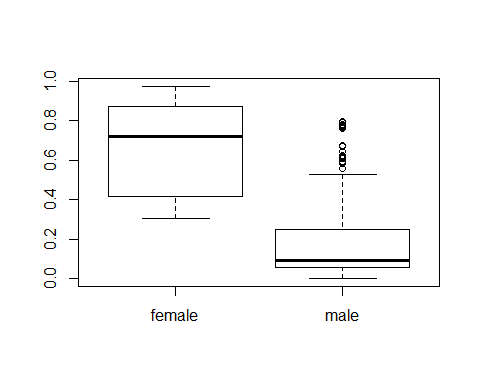
ship.logit.step = step(ship.logit, direction = 'both')

## Start: AIC=1156.34  
## survived ~ pclass + name\_length + as.factor(agegroup) + ticket\_rec +   
## room\_rec + as.factor(embarked) + home.dest\_rec + as.factor(sex)  
##   
## Df Deviance AIC  
## - room\_rec 1 1121.9 1155.9  
## - as.factor(embarked) 3 1126.0 1156.0  
## <none> 1120.3 1156.3  
## - name\_length 1 1124.3 1158.3  
## - ticket\_rec 1 1124.6 1158.6  
## - home.dest\_rec 1 1132.2 1166.2  
## - as.factor(agegroup) 8 1161.5 1181.5  
## - pclass 1 1167.6 1201.6  
## - as.factor(sex) 1 1358.9 1392.9  
##   
## Step: AIC=1155.86  
## survived ~ pclass + name\_length + as.factor(agegroup) + ticket\_rec +   
## as.factor(embarked) + home.dest\_rec + as.factor(sex)  
##   
## Df Deviance AIC  
## - as.factor(embarked) 3 1127.3 1155.3  
## <none> 1121.9 1155.9  
## + room\_rec 1 1120.3 1156.3  
## - name\_length 1 1125.8 1157.8  
## - ticket\_rec 1 1128.2 1160.2  
## - home.dest\_rec 1 1133.0 1165.0  
## - as.factor(agegroup) 8 1163.2 1181.2  
## - pclass 1 1175.3 1207.3  
## - as.factor(sex) 1 1363.7 1395.7  
##   
## Step: AIC=1155.27  
## survived ~ pclass + name\_length + as.factor(agegroup) + ticket\_rec +   
## home.dest\_rec + as.factor(sex)  
##   
## Df Deviance AIC  
## <none> 1127.3 1155.3  
## + as.factor(embarked) 3 1121.9 1155.9  
## + room\_rec 1 1126.0 1156.0  
## - name\_length 1 1132.7 1158.7  
## - ticket\_rec 1 1132.9 1158.9  
## - home.dest\_rec 1 1135.0 1161.0  
## - as.factor(agegroup) 8 1166.6 1178.6  
## - pclass 1 1191.6 1217.6  
## - as.factor(sex) 1 1371.4 1397.4

summary(ship.logit.step)

##   
## Call:  
## glm(formula = survived ~ pclass + name\_length + as.factor(agegroup) +   
## ticket\_rec + home.dest\_rec + as.factor(sex), family = binomial,   
## data = ship)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.7192 -0.6848 -0.3320 0.5424 2.5050   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.19174 0.67689 4.715 2.41e-06 \*\*\*  
## pclass -0.88787 0.11503 -7.718 1.18e-14 \*\*\*  
## name\_length 0.02469 0.01076 2.294 0.021765 \*   
## as.factor(agegroup)2 -1.35448 0.45118 -3.002 0.002682 \*\*   
## as.factor(agegroup)3 -1.80510 0.41770 -4.322 1.55e-05 \*\*\*  
## as.factor(agegroup)4 -1.54144 0.43349 -3.556 0.000377 \*\*\*  
## as.factor(agegroup)5 -2.04926 0.45958 -4.459 8.23e-06 \*\*\*  
## as.factor(agegroup)6 -2.10173 0.54258 -3.874 0.000107 \*\*\*  
## as.factor(agegroup)7 -3.85075 0.82018 -4.695 2.67e-06 \*\*\*  
## as.factor(agegroup)8 -15.53642 501.54133 -0.031 0.975288   
## as.factor(agegroup)9 -1.51672 0.42058 -3.606 0.000311 \*\*\*  
## ticket\_rec 0.80383 0.34445 2.334 0.019615 \*   
## home.dest\_rec 0.70159 0.25268 2.777 0.005493 \*\*   
## as.factor(sex)male -2.37618 0.16610 -14.306 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1686.8 on 1312 degrees of freedom  
## Residual deviance: 1127.3 on 1299 degrees of freedom  
## AIC: 1155.3  
##   
## Number of Fisher Scoring iterations: 13

fitprob0 = predict(object = ship.logit.step, type = 'response')  
  
pred = predict(ship.logit, interval='prediction')  
plot(sex, fitprob0)



Next we fit a model with some interactions between significant predictors Then we used AIC criteria for another step elimination with both directions. This resulted in the following model:

survived ~ pclass + agegroup + ticket\_rec + as.factor(embarked) + home.dest\_rec + as.factor(sex) + pclass:agegroup + pclass:as.factor(sex)

Two interactions were kept: pclass times age group, and pclass times sex.

This model only loses 10 degrees of freedom and nearly all of the variables in the model are significant.

An example of interpretting the odds ratios: If your home\_destination is recorded, your odds of surviving is 4.59 times greater than if it is not recorded.

Interpretting the pclass\*sex(male) interaction: the combination of pclass and sex is significant to determining if you survive. If you are male, your passenger class has a significant effect on your survival.

ship.logit.int <- glm(survived ~ pclass + name\_length + agegroup + ticket\_rec + room\_rec + as.factor(embarked) + home.dest\_rec + as.factor(sex) + pclass\*name\_length + pclass\*agegroup + as.factor(sex)\*agegroup + as.factor(sex)\*agegroup + as.factor(sex)\*pclass, family = binomial)  
  
summary(ship.logit.int)

##   
## Call:  
## glm(formula = survived ~ pclass + name\_length + agegroup + ticket\_rec +   
## room\_rec + as.factor(embarked) + home.dest\_rec + as.factor(sex) +   
## pclass \* name\_length + pclass \* agegroup + as.factor(sex) \*   
## agegroup + as.factor(sex) \* agegroup + as.factor(sex) \* pclass,   
## family = binomial)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9690 -0.6625 -0.4354 0.3790 2.2056   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 4.364642 1.443836 3.023 0.002503 \*\*   
## pclass -1.624680 0.559786 -2.902 0.003704 \*\*   
## name\_length 0.036648 0.036312 1.009 0.312850   
## agegroup -0.185456 0.111795 -1.659 0.097135 .   
## ticket\_rec 0.830667 0.372916 2.227 0.025915 \*   
## room\_rec 0.425512 0.380707 1.118 0.263700   
## as.factor(embarked)Cherbourg -0.567561 0.458278 -1.238 0.215544   
## as.factor(embarked)Queenstown -1.017018 0.570456 -1.783 0.074616 .   
## as.factor(embarked)Southampton -0.992252 0.458909 -2.162 0.030603 \*   
## home.dest\_rec 1.419207 0.392957 3.612 0.000304 \*\*\*  
## as.factor(sex)male -5.506220 0.675924 -8.146 3.76e-16 \*\*\*  
## pclass:name\_length -0.011029 0.014938 -0.738 0.460323   
## pclass:agegroup 0.051670 0.044987 1.149 0.250744   
## agegroup:as.factor(sex)male 0.006261 0.055897 0.112 0.910813   
## pclass:as.factor(sex)male 1.284592 0.244904 5.245 1.56e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1686.8 on 1312 degrees of freedom  
## Residual deviance: 1118.0 on 1298 degrees of freedom  
## AIC: 1148  
##   
## Number of Fisher Scoring iterations: 5

anova(ship.logit.int)

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: survived  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev  
## NULL 1312 1686.8  
## pclass 1 172.190 1311 1514.6  
## name\_length 1 76.772 1310 1437.8  
## agegroup 1 13.296 1309 1424.5  
## ticket\_rec 1 6.710 1308 1417.8  
## room\_rec 1 3.975 1307 1413.8  
## as.factor(embarked) 3 6.875 1304 1406.9  
## home.dest\_rec 1 1.498 1303 1405.4  
## as.factor(sex) 1 248.213 1302 1157.2  
## pclass:name\_length 1 6.049 1301 1151.2  
## pclass:agegroup 1 0.271 1300 1150.9  
## agegroup:as.factor(sex) 1 1.543 1299 1149.4  
## pclass:as.factor(sex) 1 31.406 1298 1118.0

step.int = step(ship.logit.int, direction = 'both')

## Start: AIC=1147.96  
## survived ~ pclass + name\_length + agegroup + ticket\_rec + room\_rec +   
## as.factor(embarked) + home.dest\_rec + as.factor(sex) + pclass \*   
## name\_length + pclass \* agegroup + as.factor(sex) \* agegroup +   
## as.factor(sex) \* agegroup + as.factor(sex) \* pclass  
##   
## Df Deviance AIC  
## - agegroup:as.factor(sex) 1 1118.0 1146.0  
## - pclass:name\_length 1 1118.5 1146.5  
## - room\_rec 1 1119.2 1147.2  
## - pclass:agegroup 1 1119.3 1147.3  
## <none> 1118.0 1148.0  
## - as.factor(embarked) 3 1124.8 1148.8  
## - ticket\_rec 1 1122.9 1150.9  
## - home.dest\_rec 1 1132.6 1160.6  
## - pclass:as.factor(sex) 1 1149.4 1177.4  
##   
## Step: AIC=1145.97  
## survived ~ pclass + name\_length + agegroup + ticket\_rec + room\_rec +   
## as.factor(embarked) + home.dest\_rec + as.factor(sex) + pclass:name\_length +   
## pclass:agegroup + pclass:as.factor(sex)  
##   
## Df Deviance AIC  
## - pclass:name\_length 1 1118.5 1144.5  
## - room\_rec 1 1119.2 1145.2  
## - pclass:agegroup 1 1119.3 1145.3  
## <none> 1118.0 1146.0  
## - as.factor(embarked) 3 1125.0 1147.0  
## + agegroup:as.factor(sex) 1 1118.0 1148.0  
## - ticket\_rec 1 1122.9 1148.9  
## - home.dest\_rec 1 1133.1 1159.1  
## - pclass:as.factor(sex) 1 1150.9 1176.9  
##   
## Step: AIC=1144.52  
## survived ~ pclass + name\_length + agegroup + ticket\_rec + room\_rec +   
## as.factor(embarked) + home.dest\_rec + as.factor(sex) + pclass:agegroup +   
## pclass:as.factor(sex)  
##   
## Df Deviance AIC  
## - name\_length 1 1119.5 1143.5  
## - room\_rec 1 1119.7 1143.7  
## - pclass:agegroup 1 1120.3 1144.3  
## <none> 1118.5 1144.5  
## + pclass:name\_length 1 1118.0 1146.0  
## - as.factor(embarked) 3 1126.3 1146.3  
## + agegroup:as.factor(sex) 1 1118.5 1146.5  
## - ticket\_rec 1 1123.6 1147.6  
## - home.dest\_rec 1 1134.7 1158.7  
## - pclass:as.factor(sex) 1 1156.4 1180.4  
##   
## Step: AIC=1143.47  
## survived ~ pclass + agegroup + ticket\_rec + room\_rec + as.factor(embarked) +   
## home.dest\_rec + as.factor(sex) + pclass:agegroup + pclass:as.factor(sex)  
##   
## Df Deviance AIC  
## - room\_rec 1 1120.7 1142.7  
## - pclass:agegroup 1 1121.4 1143.4  
## <none> 1119.5 1143.5  
## + name\_length 1 1118.5 1144.5  
## + agegroup:as.factor(sex) 1 1119.5 1145.5  
## - as.factor(embarked) 3 1128.0 1146.0  
## - ticket\_rec 1 1124.7 1146.7  
## - home.dest\_rec 1 1138.6 1160.6  
## - pclass:as.factor(sex) 1 1159.5 1181.5  
##   
## Step: AIC=1142.67  
## survived ~ pclass + agegroup + ticket\_rec + as.factor(embarked) +   
## home.dest\_rec + as.factor(sex) + pclass:agegroup + pclass:as.factor(sex)  
##   
## Df Deviance AIC  
## <none> 1120.7 1142.7  
## - pclass:agegroup 1 1122.8 1142.8  
## + room\_rec 1 1119.5 1143.5  
## + name\_length 1 1119.7 1143.7  
## + agegroup:as.factor(sex) 1 1120.6 1144.6  
## - as.factor(embarked) 3 1128.8 1144.8  
## - ticket\_rec 1 1127.1 1147.1  
## - home.dest\_rec 1 1138.9 1158.9  
## - pclass:as.factor(sex) 1 1160.8 1180.8

summary(step.int)

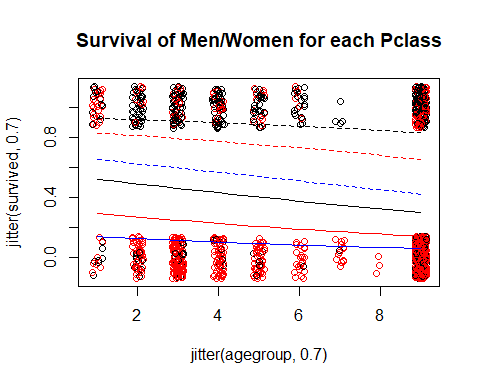
##   
## Call:  
## glm(formula = survived ~ pclass + agegroup + ticket\_rec + as.factor(embarked) +   
## home.dest\_rec + as.factor(sex) + pclass:agegroup + pclass:as.factor(sex),   
## family = binomial)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8102 -0.6880 -0.4323 0.3819 2.1985   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 5.64192 0.85895 6.568 5.09e-11 \*\*\*  
## pclass -2.04536 0.32453 -6.303 2.93e-10 \*\*\*  
## agegroup -0.21154 0.09520 -2.222 0.0263 \*   
## ticket\_rec 0.92693 0.36400 2.546 0.0109 \*   
## as.factor(embarked)Cherbourg -0.63920 0.44881 -1.424 0.1544   
## as.factor(embarked)Queenstown -1.08677 0.56154 -1.935 0.0529 .   
## as.factor(embarked)Southampton -1.07943 0.44438 -2.429 0.0151 \*   
## home.dest\_rec 1.52433 0.37508 4.064 4.82e-05 \*\*\*  
## as.factor(sex)male -5.66767 0.60483 -9.371 < 2e-16 \*\*\*  
## pclass:agegroup 0.06309 0.04318 1.461 0.1440   
## pclass:as.factor(sex)male 1.34630 0.23203 5.802 6.54e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1686.8 on 1312 degrees of freedom  
## Residual deviance: 1120.7 on 1302 degrees of freedom  
## AIC: 1142.7  
##   
## Number of Fisher Scoring iterations: 5

#final model   
fm =glm(formula = survived ~ pclass + agegroup + ticket\_rec + as.factor(embarked) + home.dest\_rec + as.factor(sex) + pclass:agegroup + pclass:as.factor(sex), family = binomial)  
  
exp(fm$coefficients)

## (Intercept) pclass   
## 2.820050e+02 1.293337e-01   
## agegroup ticket\_rec   
## 8.093359e-01 2.526741e+00   
## as.factor(embarked)Cherbourg as.factor(embarked)Queenstown   
## 5.277131e-01 3.373032e-01   
## as.factor(embarked)Southampton home.dest\_rec   
## 3.397894e-01 4.592088e+00   
## as.factor(sex)male pclass:agegroup   
## 3.455891e-03 1.065126e+00   
## pclass:as.factor(sex)male   
## 3.843192e+00

In order to plot our model with the data, we reduced it to a model containing only agegroup, sex, and pclass. In the graph, the solid lines represent men and the dotted lines represent women. For both men and women, black is 1st class, red is 2nd class, and blue is 3rd class.The black data points are female and the red data points are male.

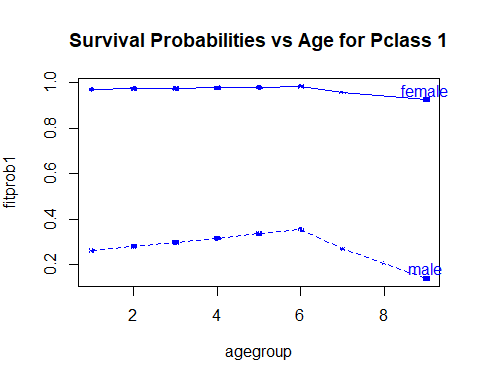
#reduced model to plot final model  
pi1 = glm (survived ~ agegroup + sex + pclass, family=binomial)  
  
agegpseq = seq (1, 9, len=30)  
  
# plot the data, with different plot symbols for male and female  
  
plot (jitter (agegroup, 0.7), jitter(survived, 0.7),  
 col=sex)  
  
# plot solid lines for males in each passenger class  
pi.male.1 = predict (pi1, data.frame (agegroup=agegpseq, sex='male', pclass=1), type='response')  
lines (agegpseq, pi.male.1)  
  
pi.male.2 = predict (pi1, data.frame (agegroup=agegpseq, sex='male', pclass=2), type='response')  
lines (agegpseq, pi.male.2, col='red')  
  
pi.male.3 = predict (pi1, data.frame (agegroup=agegpseq, sex='male', pclass=3), type='response')  
lines (agegpseq, pi.male.3, col='blue')  
  
# Plot dashed lines for females in each passenger class  
pi.female.1 = predict (pi1, data.frame (agegroup=agegpseq, sex='female', pclass=1), type='response')  
lines (agegpseq, pi.female.1, lty=2)  
  
pi.female.2 = predict (pi1, data.frame (agegroup=agegpseq, sex='female', pclass=2), type='response')  
lines (agegpseq, pi.female.2, lty=2, col='red')  
  
pi.female.3 = predict (pi1, data.frame (agegroup=agegpseq, sex='female', pclass=3), type='response')  
lines (agegpseq, pi.female.3, lty=2, col='blue')  
title('Survival of Men/Women for each Pclass')



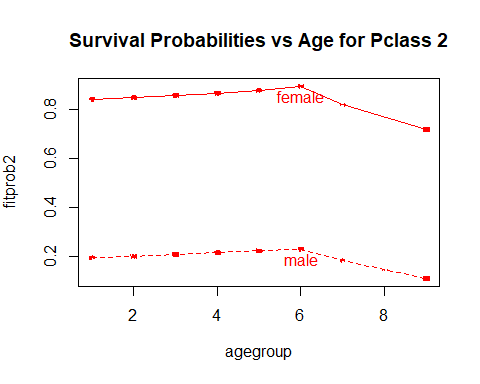
Next, we plot some of the fitted probabilities for each pclass against age and sex. These plots are useful for seeing the differences between men and womens survival amongst the passenger classes. If you want to survive, it is best to be a young woman in the first passenger class.

These plots show roughly the same thing as the survival of men/women for each Pclass plot does, however, these plots represent the whole final model rather than a reduced model. The main difference in the plots is that when the full final model is used, the plot generally increases with age until age group 6, and then drops off for every pclass and gender. These plots use the plsmo package and also show the density of the data for the different agegroups.

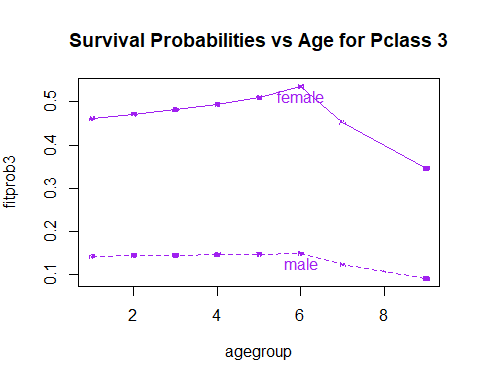
#fitted probabilites for each pclass  
ageseq = seq(min(agegroup), max(agegroup), length=1313)  
fitprob1 = predict (fm, data.frame (agegroup=ageseq, pclass=1), type='response')  
fitprob2 = predict (fm, data.frame (agegroup=ageseq, pclass=2), type='response')  
fitprob3 = predict (fm, data.frame (agegroup=ageseq, pclass=3), type='response')  
  
plsmo(agegroup, fitprob1, group=sex, datadensity=T, col='blue')  
title('Survival Probabilities vs Age for Pclass 1')



plsmo(agegroup, fitprob2, group=sex, datadensity=T, col='red')   
title('Survival Probabilities vs Age for Pclass 2')



plsmo(agegroup, fitprob3, group=sex, datadensity=T, col='purple')  
title('Survival Probabilities vs Age for Pclass 3')



expbetas\_fm = exp (fm$coefficients)  
expbetas\_fm

## (Intercept) pclass   
## 2.820050e+02 1.293337e-01   
## agegroup ticket\_rec   
## 8.093359e-01 2.526741e+00   
## as.factor(embarked)Cherbourg as.factor(embarked)Queenstown   
## 5.277131e-01 3.373032e-01   
## as.factor(embarked)Southampton home.dest\_rec   
## 3.397894e-01 4.592088e+00   
## as.factor(sex)male pclass:agegroup   
## 3.455891e-03 1.065126e+00   
## pclass:as.factor(sex)male   
## 3.843192e+00

### Model Diagnostics and Tests

Deviance test for lack of fit (a.k.a., goodnesss of fit):

: The fitted model fits well

: The fitted model does not fit well

Here, we want a high p-value, because a low p-value would indicate a lack of fit. Our p-values are: initial model: 0.9998 final model: 0.9999 Both our initial model and our final model seem to fit well.

# First model  
pchisq(deviance(ship.logit), df.residual(ship.logit), lower=F)

## [1] 0.9998323

# Final model  
pchisq(deviance(fm), df.residual(fm), lower=F)

## [1] 0.9999004

As seen below, both our initial model and our final model have p-values much less than 0.05 which means they both have significant effect on survival status.

#Getting the LR test statistic and P-value in R (multiple logistic regression):  
  
# First model  
  
1 - pchisq(ship.logit$null.deviance - ship.logit$deviance,   
 ship.logit$df.null - ship.logit$df.residual)

## [1] 0

# Final model  
  
1 - pchisq(fm$null.deviance - fm$deviance,   
 fm$df.null - fm$df.residual)

## [1] 0

Confidence intervals and odds ratios for each variable:

confint.default (fm)

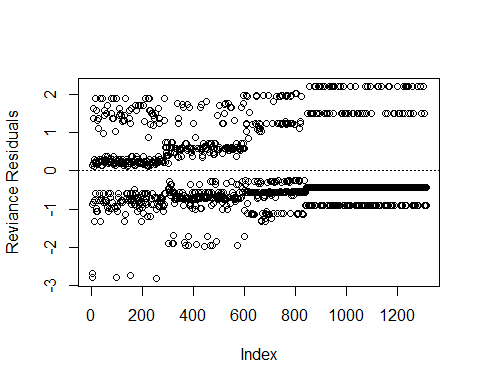
## 2.5 % 97.5 %  
## (Intercept) 3.95840476 7.32544456  
## pclass -2.68142872 -1.40928972  
## agegroup -0.39812078 -0.02496180  
## ticket\_rec 0.21349912 1.64036185  
## as.factor(embarked)Cherbourg -1.51886283 0.24045795  
## as.factor(embarked)Queenstown -2.18736566 0.01381936  
## as.factor(embarked)Southampton -1.95039633 -0.20846209  
## home.dest\_rec 0.78919984 2.25946970  
## as.factor(sex)male -6.85312979 -4.48222017  
## pclass:agegroup -0.02153826 0.14772526  
## pclass:as.factor(sex)male 0.89154236 1.80106405

exp (confint.default (fm))

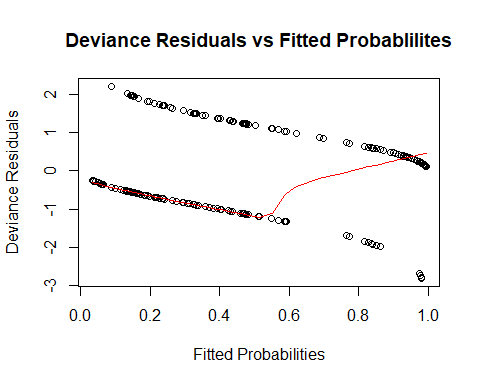
## 2.5 % 97.5 %  
## (Intercept) 52.373710521 1.518449e+03  
## pclass 0.068465267 2.443168e-01  
## agegroup 0.671580911 9.753472e-01  
## ticket\_rec 1.238002405 5.157035e+00  
## as.factor(embarked)Cherbourg 0.218960740 1.271831e+00  
## as.factor(embarked)Queenstown 0.112211964 1.013915e+00  
## as.factor(embarked)Southampton 0.142217696 8.118318e-01  
## home.dest\_rec 2.201634059 9.578009e+00  
## as.factor(sex)male 0.001056145 1.130828e-02  
## pclass:agegroup 0.978692027 1.159194e+00  
## pclass:as.factor(sex)male 2.438888389 6.056088e+00

Residual plots The Deviance Residuals looks really good. The lowess fitted line should be horizontal around zero, and it is not perfect, but not bad.

# Residual plot  
  
dev<-residuals(fm)  
plot(dev, ylab="Reviance Residuals")  
abline(h=0, lty=3)



plot (fm$fitted.values, dev, ylab="Deviance Residuals", xlab="Fitted Probabilities", main='Deviance Residuals vs Fitted Probablilites')  
lines (lowess (fm$fitted, dev), col="Red")



ROC Curve: The ROC curve suggests the predictive ability of this model is better than random guessing, since the AUC is larger than 0.5. By examining the possible cutoff values in the written excel file, the cutoff that minimizes the distance from the curve to the top left corner is 0.326. With this cutoff, we have a sensitivity of 0.7684 and a specificity of 0.7662. These seem like a very good values, and looking at our ROC curve, 0.326 as a cutoff looks good. With the 0.326 cutoff, the false positive rate is 0.2338 and the true positive rate (sensitivity) is 0.7684.

ship$pred.cat = ifelse (fm$fitted.values < 0.326, 0, 1)  
#ship[order(fm$fitted.values),]  
  
attach(ship)  
table1 = table (survived, pred.cat)  
table1

## pred.cat  
## survived 0 1  
## 0 662 202  
## 1 104 345

sensitivity = table1[2,2]/sum(table1[2,])  
sensitivity

## [1] 0.7683742

specificity = table1[1,1]/sum(table1[1,])  
specificity

## [1] 0.7662037

library(ROCR)

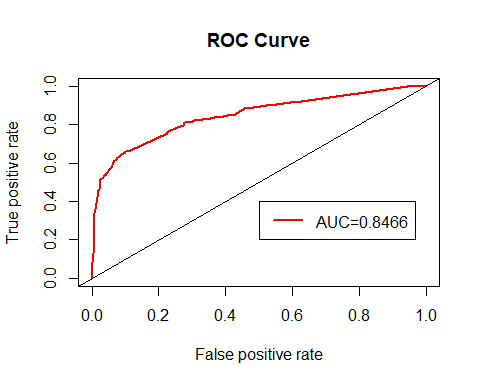
## Warning: package 'ROCR' was built under R version 3.4.2

## Warning: package 'gplots' was built under R version 3.4.2

pred1 <- prediction(fm$fitted.values, fm$y)  
perf1 <- performance(pred1,"tpr","fpr")  
auc1 <- performance(pred1,"auc")@y.values[[1]]  
auc1

## [1] 0.8466487

plot(perf1, lwd=2, col=2, main='ROC Curve')  
abline(0,1)  
legend(0.5,0.4, c(paste ("AUC=", round (auc1, 4), sep="")), lwd=2, col=2)



roc.table = cbind.data.frame(pred1@tn, pred1@fn, pred1@fp, pred1@tp,   
 pred1@cutoffs)  
names (roc.table) = c("TrueNeg", "FalseNeg", "FalsePos", "TruePos", "Cutoff")  
attach (roc.table)  
roc.table$sensitivity = TruePos / (TruePos + FalseNeg) # True positive rate  
roc.table$specificity = TrueNeg / (TrueNeg + FalsePos) # 1 - False pos rate  
roc.table$FalsePosRate = 1 - roc.table$specificity  
roc.table$PctCorrect = (TruePos + TrueNeg) /   
 (TruePos + TrueNeg + FalsePos + FalseNeg)  
#write.csv (roc.table, "ROC table2.csv")

## Influence Diagnostics

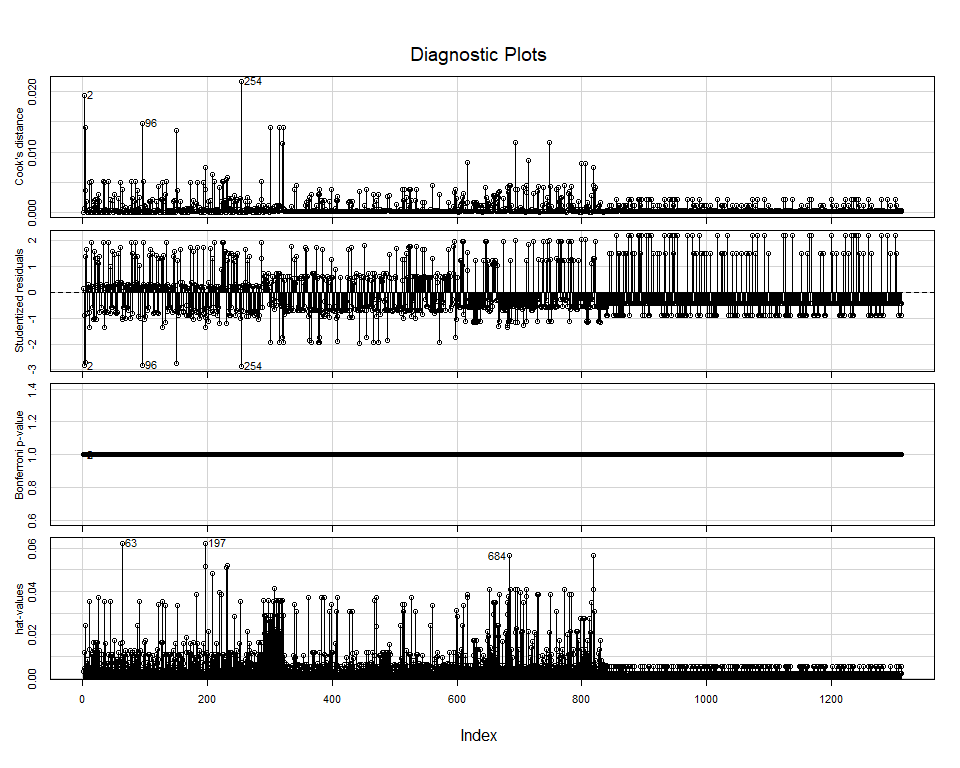
Plotting age and survival as well as pclass and survival shows that the two rows with the largest influence (2 and 254) are because they were young and old women (2yrs and 63yrs) respectively that were both from the highest passenger class. They were both expected to survive, but neither did. The plots agegroup vs survived and pclass vs survived show these two rows in red. A final model was created without rows 2 nd 254, and the ROC curve was created for this model. The AUC increases from .8466 to .8488, which is not a large enough increase to leave out these two rows, so we will stick with our final model.

Rows 197 and 63, which showed up as having high hat-values are men without ages recorded that were part of first class. They were expected to survive, but did not.

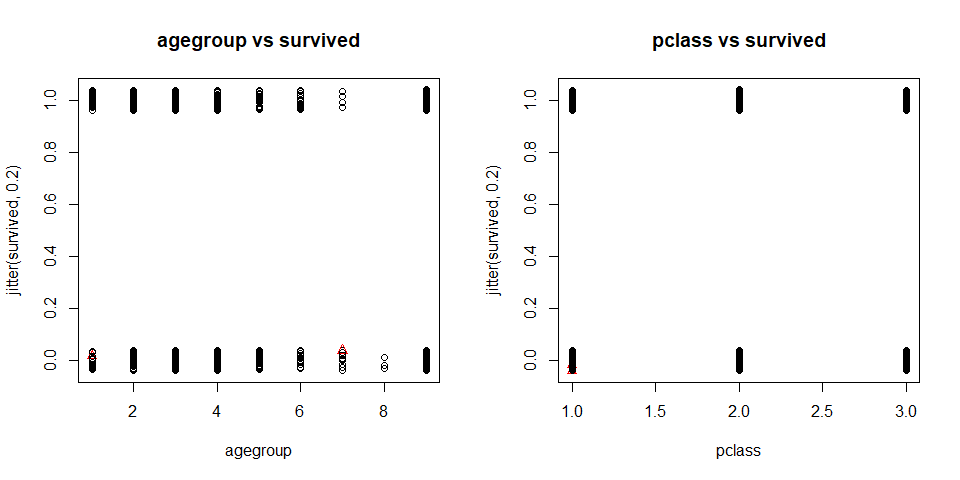
library(car)

## Warning: package 'car' was built under R version 3.4.2

influenceIndexPlot(fm, id.n=3)



# Plot data with rows 2 and 25 in red  
my.colors = rep ('black', length (index))  
my.colors [2] = 'red'  
my.colors [254] = 'red'  
plotsym = ifelse (my.colors == 'black', 1, 2)  
  
par(mfrow=c(1,2))  
  
plot (agegroup, jitter (survived, 0.2), col=my.colors, pch=plotsym, main='agegroup vs survived')  
plot (pclass, jitter (survived, 0.2), col=my.colors, pch=plotsym, main='pclass vs survived')

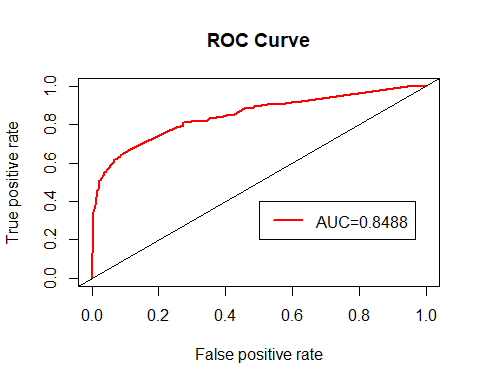


#model without rows 2 and 254  
ship1= ship[-c(2, 254), ]  
fm.d =glm(formula = ship1$survived ~ ship1$pclass + ship1$agegroup + ship1$ticket\_rec + as.factor(ship1$embarked) + ship1$home.dest\_rec + as.factor(ship1$sex) + ship1$pclass\*ship1$agegroup + ship1$pclass\*as.factor(ship1$sex), family = binomial)

par(mfrow=c(1,1))  
library(ROCR)  
pred1 <- prediction(fm.d$fitted.values, fm.d$y)  
perf1 <- performance(pred1,"tpr","fpr")  
auc1 <- performance(pred1,"auc")@y.values[[1]]  
auc1

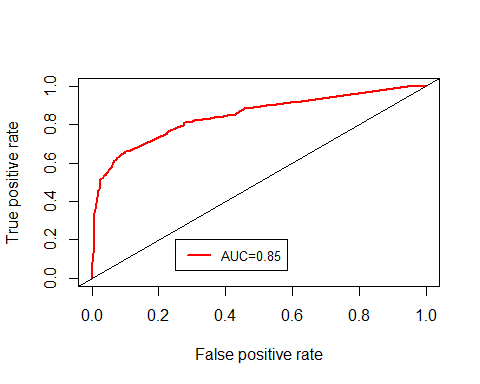
## [1] 0.8488314

plot(perf1, lwd=2, col=2, main='ROC Curve')  
abline(0,1)  
legend(0.5,0.4, c(paste ("AUC=", round (auc1, 4), sep="")), lwd=2, col=2)



The following is the function form Dr. Phil to check our ROC related calculations....they are correct.

roc.logistic = function (fit) {  
 fitvals = fit$fitted.values   
 pred1 <- prediction(fitvals, fit$y)  
 perf1 <- performance(pred1,"tpr","fpr")  
 auc1 <- performance(pred1,"auc")@y.values[[1]]  
 plot(perf1, lwd=2, col=2)  
 abline(0,1)  
 legend(0.25, 0.2, c(paste ("AUC=", round(auc1, 2), sep="")),   
 cex=0.8, lwd=2, col=2)  
 roc.table = cbind.data.frame (pred1@tn, pred1@fp, pred1@fn, pred1@tp,  
 pred1@cutoffs, perf1@x.values, perf1@y.values)  
 roc.table$spec = 1 - perf1@x.values[[1]]  
 roc.table$ppv = pred1@tp[[1]] / (pred1@tp[[1]] + pred1@fp[[1]])  
 roc.table$npv = pred1@tn[[1]] / (pred1@tn[[1]] + pred1@fn[[1]])  
 roc.table$pctcorr = (pred1@tn[[1]] + pred1@tp[[1]]) /   
 (pred1@tn[[1]] + pred1@tp[[1]] + pred1@fn[[1]] + pred1@fp[[1]])  
 roc.table$optdist = sqrt ((perf1@x.values[[1]] - 0)^2 +  
 (perf1@y.values[[1]] - 1)^2)  
 names (roc.table) = c("TN", "FP", "FN", "TP", "Cutoff", "FPR", "TPR", "Spec",  
 "PPV", "NPV", "PctCorr", "OptDist")  
 return (roc.table)  
}  
  
roc.table = roc.logistic (fm)



summary(ship.logit)

##   
## Call:  
## glm(formula = survived ~ pclass + name\_length + as.factor(agegroup) +   
## ticket\_rec + room\_rec + as.factor(embarked) + home.dest\_rec +   
## as.factor(sex), family = binomial, data = ship)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.8504 -0.6462 -0.3420 0.5450 2.4708   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.48189 0.75145 4.634 3.59e-06 \*\*\*  
## pclass -0.88803 0.13334 -6.660 2.74e-11 \*\*\*  
## name\_length 0.02158 0.01093 1.974 0.048403 \*   
## as.factor(agegroup)2 -1.38126 0.45688 -3.023 0.002501 \*\*   
## as.factor(agegroup)3 -1.85789 0.42477 -4.374 1.22e-05 \*\*\*  
## as.factor(agegroup)4 -1.60327 0.43990 -3.645 0.000268 \*\*\*  
## as.factor(agegroup)5 -2.16478 0.46884 -4.617 3.89e-06 \*\*\*  
## as.factor(agegroup)6 -2.29185 0.55853 -4.103 4.07e-05 \*\*\*  
## as.factor(agegroup)7 -4.00577 0.83961 -4.771 1.83e-06 \*\*\*  
## as.factor(agegroup)8 -15.77621 494.36448 -0.032 0.974542   
## as.factor(agegroup)9 -1.67688 0.43391 -3.865 0.000111 \*\*\*  
## ticket\_rec 0.74818 0.36488 2.050 0.040316 \*   
## room\_rec 0.41567 0.34000 1.223 0.221500   
## as.factor(embarked)Cherbourg -0.35570 0.38808 -0.917 0.359362   
## as.factor(embarked)Queenstown -0.68641 0.52501 -1.307 0.191067   
## as.factor(embarked)Southampton -0.71985 0.36823 -1.955 0.050595 .   
## home.dest\_rec 1.17255 0.34816 3.368 0.000758 \*\*\*  
## as.factor(sex)male -2.38233 0.16863 -14.127 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 1686.8 on 1312 degrees of freedom  
## Residual deviance: 1120.3 on 1295 degrees of freedom  
## AIC: 1156.3  
##   
## Number of Fisher Scoring iterations: 13

# Find the row(s) in the ROC table with the largest percent correctly classified  
  
roc.table [which.max (roc.table$PctCorr), ]

## TN FP FN TP Cutoff FPR TPR Spec PPV  
## 767 809 55 178 271 0.4654823 0.06365741 0.6035635 0.9363426 0.8312883  
## NPV PctCorr OptDist  
## 767 0.8196555 0.8225438 0.4015149

# Find the row(s) in the ROC table that are closest to the (0, 1) corner.  
  
roc.table [which.min (roc.table$OptDist), ]

## TN FP FN TP Cutoff FPR TPR Spec PPV  
## 276 662 202 104 345 0.3262392 0.2337963 0.7683742 0.7662037 0.630713  
## NPV PctCorr OptDist  
## 276 0.8642298 0.7669459 0.3291067

## Would the Average Luther Student Survive?

Using data from College Factual (<https://www.collegefactual.com/colleges/luther-college/student-life/diversity/>), the 'average' Luther student is a female in her 20's from either the first or second passenger class. (using the term average very loosely here).

Using our reduced model that we used to plot the model (survived ~ agegroup + pclass + sex), the average Luther student would survive. If the student decided to spend extra and get a first class ticket, our model predicts their probability of survival is 0.908 and if they buy a second class ticket, their probability of survival is 0.793. Both are well above 0.5, as well as far above our 0.326 cutoff.

luther1 = predict (pi1, data.frame (agegroup=3, sex='female', pclass=1), type='response')  
luther2 = predict (pi1, data.frame (agegroup=3, sex='female', pclass=2), type='response')  
  
#luther student from pclass1  
luther1

## 1   
## 0.9083469

#luther student from pclass2  
luther2

## 1   
## 0.7927083