HW4 - Logistic Regression

Olivia Samples

7/1/2020

```
rm(list=ls())
library(e1071)
library(caret)
```

```
## Loading required package: lattice
## Loading required package: ggplot2
```

```
library(ROCR)
```

Let's load the Titanic training data.

```
ship <- read.csv("tit-train.csv")
ship$Survived <- factor(ship$Survived)
ship$Pclass <- factor(ship$Pclass)
ship$Embarked <- factor(ship$Embarked)
```

Let's create a model that does not include "PassengerId", "Name", "Ticket", or "Cabin" because these are each unique or missing too many entries. In this dataset, a 1 corresponds to Survived and a 0 corresponds to Not-Survived.

```
logit1 <- glm(Survived ~ Age + Sex + Pclass + SibSp + Parch + Fare + Embarked, data = ship, family = "bi
summary(logit1)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass + SibSp + Parch +
##     Fare + Embarked, family = "binomial", data = ship)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7220  -0.6455  -0.3770   0.6293   2.4461
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  16.691979 607.920015   0.027 0.978095
## Age          -0.043308   0.008322  -5.204 1.95e-07 ***
## Sexmale      -2.637859   0.223006 -11.829  < 2e-16 ***
## Pclass2      -1.189637   0.329197  -3.614 0.000302 ***
## Pclass3      -2.395220   0.343356  -6.976 3.04e-12 ***
## SibSp        -0.362925   0.129290  -2.807 0.005000 **
## Parch        -0.060365   0.123944  -0.487 0.626233
## Fare          0.001451   0.002595   0.559 0.576143
## EmbarkedC   -12.259048 607.919885  -0.020 0.983911
## EmbarkedQ   -13.082427 607.920088  -0.022 0.982831
```

```
## EmbarkedS   -12.661895 607.919868  -0.021 0.983383
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 632.34  on 703  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 654.34
##
## Number of Fisher Scoring iterations: 13
```

Now we can see that the most significant variables are "Age", "Sex", "Pclass", "SibSp" for determining survivorship. We will make a new model with these variables.

```r
logit2 <- glm(Survived ~ Age + Sex + Pclass + SibSp, data = ship, family = "binomial")
summary(logit2)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex + Pclass + SibSp, family = "binomial",
##     data = ship)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7876  -0.6417  -0.3864   0.6261   2.4539
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.334201   0.450700   9.617  < 2e-16 ***
## Age         -0.044760   0.008225  -5.442 5.27e-08 ***
## Sexmale     -2.627679   0.214771 -12.235  < 2e-16 ***
## Pclass2     -1.414360   0.284727  -4.967 6.78e-07 ***
## Pclass3     -2.652618   0.285832  -9.280  < 2e-16 ***
## SibSp       -0.380190   0.121516  -3.129  0.00176 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 964.52  on 713  degrees of freedom
## Residual deviance: 636.56  on 708  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 648.56
##
## Number of Fisher Scoring iterations: 5
```

What are the odds of surviving the shipwreck? (20 points)

The general odds formula is $odds = \frac{p}{1-p}$, so we will calculate this below.

```r
p = sum(ship$Survived == 1)/NROW(ship$Survived)
odds = p/(1-p)
odds
```

```
## [1] 0.6229508
```

In order to calculate the odds, one must calculate the ratio of probability of surviving the shipwreck versus not surviving the shipwreck. The probability of surviving divided by the probability of not surviving gives you the odds of surviving, which is 0.6229508. In other words, based on the people that survived and didn't survive, then the odds of surviving are 62 to 100. For every 62 people that survive, 100 people do not survive.

Using the logit model, estimate how much lower are the odds of survival for men relative to women? (20 points)

```r
logit3 <- glm(Survived ~ Sex, data = ship, family = "binomial")
summary(logit3)
```

```
##
## Call:
## glm(formula = Survived ~ Sex, family = "binomial", data = ship)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6462  -0.6471  -0.6471   0.7725   1.8256
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0566     0.1290   8.191 2.58e-16 ***
## Sexmale      -2.5137     0.1672 -15.036  < 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: 1186.7  on 890  degrees of freedom
## Residual deviance:  917.8  on 889  degrees of freedom
## AIC: 921.8
##
## Number of Fisher Scoring iterations: 4
```

```r
exp(coef(logit3))
```

```
## (Intercept)     Sexmale
##  2.87654321  0.08096732
```

The ratio of the odds for male relative to the odds for female is 0.08097. In other words, for every 8 men that survive, 100 women survive. So, the odds for male to survive are about 91.9% lower than the odds for female to survive. This could make sense because women and children were asked to fill the boats first, resulting in a larger number of women that survived and a fewer number of women that did not survive. This allows for a favorable odds ratio for women's survival whereas the male survivor ratio was significantly lower, 91.9% to be exact. We can confirm this by the calculations below.

```r
nmale = sum(ship$Sex == "male")
nfem = sum(ship$Sex == "female")
nfsur = sum(ship$Sex == "female" & ship$Survived == 1)
nmsur = sum(ship$Sex == "male" & ship$Survived == 1)

notf = nfem - nfsur
notm = nmale - nmsur
```

```r
femodds = nfsur/notf
maleodds = nmsur/notm

maleodds/femodds
```

```
## [1] 0.08096732
```

Controlling for gender, does age have a statistically significant effect on the odds of survival? (20 points) If so, what is the magnitude of that effect (20 points)?

```r
logit4 <- glm(Survived ~ Age + Sex, data = ship, family = "binomial")
summary(logit4)
```

```
##
## Call:
## glm(formula = Survived ~ Age + Sex, family = "binomial", data = ship)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7405  -0.6885  -0.6558   0.7533   1.8989
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.277273   0.230169   5.549 2.87e-08 ***
## Age         -0.005426   0.006310  -0.860     0.39
## Sexmale     -2.465920   0.185384 -13.302  < 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: 964.52  on 713  degrees of freedom
## Residual deviance: 749.96  on 711  degrees of freedom
##   (177 observations deleted due to missingness)
## AIC: 755.96
##
## Number of Fisher Scoring iterations: 4
```

```r
exp(coef(logit4))
```

```
## (Intercept)         Age     Sexmale
##  3.58684584  0.99458879  0.08493066
```

Holding gender constant, the effect of increasing Age by 1 year decreases the odds of surviving by -0.005 percent. This is not statistically significant since the p-value is $0.39 > 0.05$.

This can begin to make sense when analyzing the age distribution. The prioritized passengers were Women and Children, but when holding gender as a constant, the age did not significantly affect the outcome. As you can see, the majority of the passengers were considered adults (778/891). The survival rates of children vs. the survival rates of adults were also extremely similar. This supports the fact that the age does not significantly affect their survival.

```r
children = sum((ship$Age < 18) & !(is.na(ship$Age)))
surchi = sum((ship$Age < 18) & !(is.na(ship$Age)) & (ship$Survived == 1))
children
```

```
## [1] 113
```

```r
surchi/children
```

```
## [1] 0.539823
```

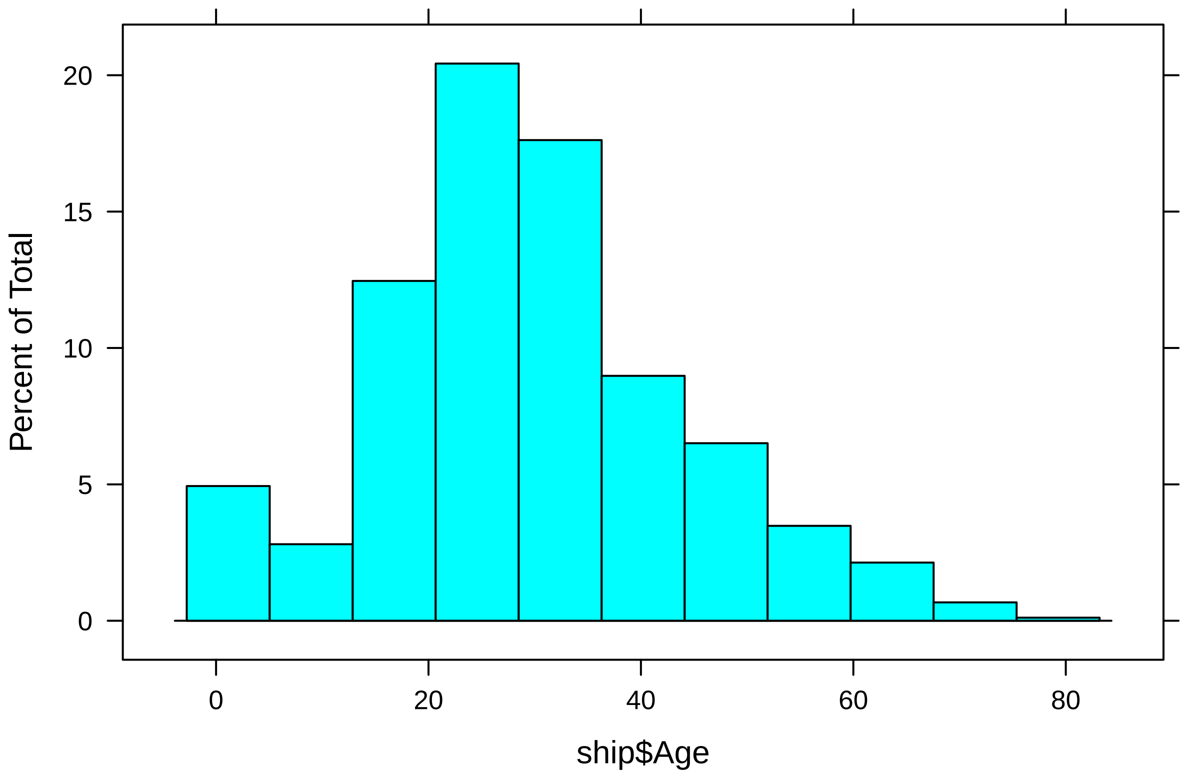
```r
adults = sum((ship$Age >= 18) | is.na(ship$Age))
surad = sum((ship$Age >= 18) & (ship$Survived == 1) | is.na(ship$Age))
adults
```

```
## [1] 778
```

```r
surad/adults
```

```
## [1] 0.5218509
```

```r
histogram(ship$Age)
```



Suppose this report is to someone that does not know statistics/machine learning/analysis, report above questions in a way that people could understand (20 points)

See explanations corresponding to each question above.