# Stat 151A - Homework 6

*Nicholas Lai*

*November 21, 2018*

## Question 1

```
Train <- titanic::titanic_train
Test <- titanic::titanic_test
```

The Data Has missing values for age, so I replaced them with the mean of their respective datasets.

```
Train$Age[is.na(Train$Age)] = mean(Train$Age, na.rm = TRUE)
Test$Age[is.na(Test$Age)] = mean(Test$Age, na.rm = TRUE)
```

There are some variables that have no real useful interpretation for a glm, so I remove them:

```
bad = c("PassengerId","Name","Ticket","Embarked","Cabin")
Train = Train[,!(names(Train) %in% bad)]
```

I perform Backwards Selection based on p-values:

```
TitanicLog1 = glm(Survived~., data = Train, family = binomial)
summary(TitanicLog1)
```

```
##
## Call:
## glm(formula = Survived ~ ., family = binomial, data = Train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7129  -0.6032  -0.4273   0.6191   2.4186
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.960445   0.532937   9.308  < 2e-16 ***
## Pclass      -1.084297   0.139119  -7.794 6.49e-15 ***
## Sexmale     -2.762930   0.199011 -13.883  < 2e-16 ***
## Age         -0.039702   0.007797  -5.092 3.55e-07 ***
## SibSp       -0.350725   0.109552  -3.201  0.00137 **
## Parch       -0.111963   0.117400  -0.954  0.34024
## Fare         0.002852   0.002361   1.208  0.22718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  788.73  on 884  degrees of freedom
## AIC: 802.73
##
## Number of Fisher Scoring iterations: 5
```

```r
TitanicLog2 = glm(Survived ~ . - Parch, data = Train, family = binomial)
summary(TitanicLog2)
```

```
##
## Call:
## glm(formula = Survived ~ . - Parch, family = binomial, data = Train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7458  -0.5948  -0.4170   0.6109   2.4501
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.942367   0.530775   9.312  < 2e-16 ***
## Pclass      -1.098189   0.137969  -7.960 1.72e-15 ***
## Sexmale     -2.726408   0.194561 -14.013  < 2e-16 ***
## Age         -0.039385   0.007773  -5.067 4.05e-07 ***
## SibSp       -0.378646   0.106212  -3.565 0.000364 ***
## Fare         0.002373   0.002250   1.054 0.291707
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  789.65  on 885  degrees of freedom
## AIC: 801.65
##
## Number of Fisher Scoring iterations: 5
```

```r
TitanicLog3 = glm(Survived ~ . - Parch - Fare, data = Train, family = binomial)
summary(TitanicLog3)
```

```
##
## Call:
## glm(formula = Survived ~ . - Parch - Fare, family = binomial,
##     data = Train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6869  -0.6055  -0.4169   0.6111   2.4547
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.191976   0.478346  10.854  < 2e-16 ***
## Pclass      -1.172391   0.119725  -9.792  < 2e-16 ***
## Sexmale     -2.739806   0.194142 -14.112  < 2e-16 ***
## Age         -0.039793   0.007755  -5.131 2.88e-07 ***
## SibSp       -0.357788   0.104033  -3.439 0.000583 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 1186.66  on 890  degrees of freedom
## Residual deviance:  790.84  on 886  degrees of freedom
## AIC: 800.84
##
## Number of Fisher Scoring iterations: 5
```

We arrive at a model with 4 parameters: Passenger Class, Sex, Age, and number of siblings/spouses aboard. All the coefficients of this model are negative, suggesting that an increase in age, number of family members, Class(higher classes are poorer), and Sex(male is 1, so it's bad to be male on the Titanic). In particular, being male and being from a lower class are the most powerful predictors of death, influencing the prediction the most in the 0 direction.

```
predictTest = predict(TitanicLog3, type = "response", newdata = Test)

Test$Survived = as.numeric(predictTest >= 0.5)

Predictions = data.frame(Test[c("PassengerId","Survived")])
write.csv(file = "TitanicPred", x = Predictions, row.names = FALSE)
```

Prediction Error (According to Kaggle): 0.74641

## Question 2

$\hat{p}$ is a function of the projection of Y onto X defined by the logistic model. Therefore, $Y - \hat{p}$ denotes the component of Y that belongs in the orthogonal space of X

## Question 3

### (a)

*Standard Error of the intercept estimate:*

Following from the z-test statistic:

$$\frac{\hat{\beta_0}}{\hat{s_0}} = z_0$$

$$\frac{0.6864}{\hat{s_0}} = 0.313$$

$$\hat{s_0} = \frac{0.6864}{0.313} = 2.192$$

*Standard Error of the estimate of the log(distance) coefficient:*

Analagously to above:

$$\frac{\hat{\beta_1}}{\hat{s_1}} = z_1$$

$$\frac{-0.9050}{\hat{s_1}} = -4.349$$

$$\hat{s_1} = \frac{-0.9050}{-4.349} = 0.208$$

*Null Deviance:*

$$N.Dev = -2(n\bar{y}log(\bar{y}) + n(1-\bar{y})log(1-\bar{y}))$$

From the givens in the question, we know that of the 212 observations, 79 of the responses equal 1 and the rest equal 0. Therefore:

$$\bar{y} = \frac{79}{212} = 0.373$$
$$N.Dev = -2(212(0.373)log(0.373) + 212(1 - 0.373)log(1 - 0.373))$$
$$N.Dev = 280.1$$

*Null Deviance Degrees of Freedom:*

The Null model contains only the intercept, so it has no parameters. Therefore it has $n-1$ degrees of freedom:

$$df = n - 1 = 211$$

*Residual Deviance*

Residual deviance is related to AIC thusly:

$$R.Dev = AIC - 2(1 + p)$$

Where $p$ is the number of parameters in our model. Therefore:

$$R.Dev = 222.18 - 8 = 214.18$$

*Residual Deviance Degrees of Freedom:*

The Residual Deviance is a measure of our model, which has degrees of freedom:

$$df = n - p - 1 = 212 - 3 - 1 = 208$$

*The log(NoOfPools), Intercept cross term*

$X^T W X$ is symmetric, so:

$$X^T W X_{31} = X^T W X_{13} = -0.255928180$$

*The meanmin coefficient variance*

The diagonal entries of $X^T W X$, $d_{ii}$ are the variances of the $i$th coefficients respectively, so

$$d_{44} = \hat{s_3}^2 = 0.098$$

**(b)**

Directly from our logistic model:

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 log(distance) + \beta_2 log(NoOfPools) + \beta_3 meanmin$$
$$log(\frac{\hat{p}}{1-\hat{p}}) = 0.6864 + (-0.9050)log(265) + (0.5027)log(26) + (1.1153)3.5$$
$$log(\frac{\hat{p}}{1-\hat{p}}) = 1.178$$
$$\frac{\hat{p}}{1-\hat{p}} = e^{1.178}$$
$$\hat{p} = \frac{e^{1.178}}{1 + e^{1.178}} = 0.7646$$

**(c)**

The residual deviance would decrease with the inclusion of another variable, `altitude`, as residual deviance depends on how much variance in the data is not explained by the model. The inclusion of an extra parameter will decrease deviance or keep it the same.

The null deviance will stay the same, as it only depends on the data, which remains the same.

## Question 4

**(a)**

The maximum likelyhood is estimated using newton's method:

$$\beta^{(m+1)} = \beta^{(m)} - (H\ell(\beta^{(m)})^{-1}\nabla\ell(\beta^{(m)})$$

The only part of this equation that depends on Y is $\nabla\ell(\beta^{(m)})$, and

$$\nabla\ell(\beta^{(m)}) = X^T(Y - p)$$

Which clearly only depends on $X^TY$, as desired.

**(b)**

By our logistic model:

$$log\frac{\hat{p_i}}{1-\hat{p_i}} = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_p x_{ip}$$

So a rearrangement yields:

$$\frac{\hat{p_i}}{1-\hat{p_i}} = e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_p x_{ip}}$$

$$\hat{p_i} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_p x_{ip}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_p x_{ip}}}$$

**(c)**

Since $\hat{p}$ is an unbiased estimator of $y$, this fact immediately follows.

**(d)**

The residual deviance is equal to minus 2 times the maximized log likelyhood. Therefore:

$$R.Dev = -2(\sum_i y_i log(\hat{p_i}) + \sum_i (1 - y_i) log(1 - \hat{p_i}))$$

## Question 5

### (a)

The subtle addition of s does not actually effect the calculation of any of these parameters.

*z-value of intercept:*

$$z_0 = \frac{\hat{\beta}_0}{\hat{s_0}}$$

$$z_0 = \frac{4.11947}{0.36342} = 11.335$$

*Estimate of beta 4*

$$\beta_4 = \hat{s_4} * z_4 = 0.02800 * 12.345 = 0.34566$$

*Null Deviance:*

$$N.Dev = -2(n\bar{y}log(\bar{y}) + n(1-\bar{y})log(1-\bar{y}))$$

From the givens in the question, we know that of the 4601 observations, 1813 of the responses equal y and the rest equal n. Therefore:

$$\bar{y} = \frac{1813}{4601} = 0.394$$

$$N.Dev = -2(4601(0.394)log(0.394) + 4601(1-0.394)log(1-0.394))$$

$$N.Dev = 6170$$

*Null Deviance Degrees of Freedom:*

The Null model contains only the intercept, so it has no parameters. Therefore it has $n-1$ degrees of freedom:

$$df = n - 1 = 4600$$

*Residual Deviance Degrees of Freedom:*

The Residual Deviance is a measure of our model, which has degrees of freedom:

$$df = n - p - 1 = 4601 - 6 - 1 = 4594$$

*AIC:*

AIC is related to Residual deviance thusly:

$$AIC = R.Dev + 2(1+p)$$

Where $p$ is the number of parameters in our model. Therefore:

$$AIC = 3245.1 + 2(1+6) = 3259.1$$

**(b)**

Directly from our logistic model (even though some variables are zero, we cannot ignore them because os s):

$$log(\frac{\hat{p}}{1-\hat{p}}) = \beta_0 + \beta_1 log(crl.tot) + \beta_2 log(dollar+s) + \beta_3 log(bang+s) + \beta_4 log(money+s) + \beta_5 log(n000+s) + \beta_6 log(make+s)$$

The RHS is just a number. Calculating it yields:

$$log(\frac{\hat{p}}{1-\hat{p}}) = 2.68$$

$$\frac{\hat{p}}{1-\hat{p}} = e^{2.68}$$

$$\hat{p} = \frac{e^{2.68}}{1+e^{2.68}} = 0.936$$

**(c)**

AIC (a criteria for selecting models) for a given number of parameters is a function only of a model's residual deviance. Both candidate models in this question have the same number of parameters, so it suffices to select the one with the smaller deviance, M1.