Stats141XP Discussion 1: HW 2 (Team 2)

Team Members: Charles Liu, Joseph Gallegos, Anshu Mahalley, Jacob Samuels, Jordan Tallman

4/18/2021

Loading Necessary Packages

```
library(readr)
library(car)
library(caret)
library(effects)
library(sjPlot)
library(nnet)
library(dplyr)
library(caTools)
library(MASS)
library(pROC)
library(ROCR)
library(mlbench)
```

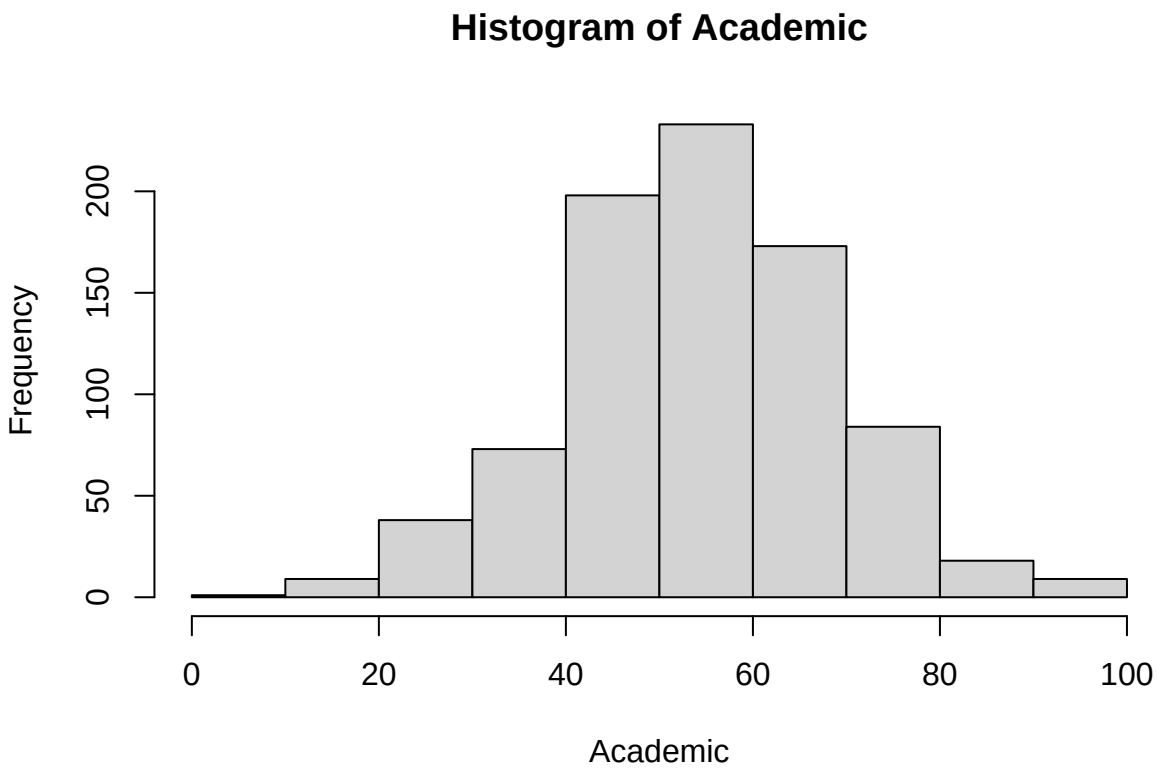
Loading Necessary Data

```
setwd(getwd())
stem_data <- read_csv("stem.csv")
d <- read_csv("diabeticsub.csv")
attach(stem_data)
```

Problem 1

Q1.a

```
bin_academic <- Academic
bin_academic[which(Academic < median(na.omit(Academic)))] <- 0
bin_academic[which(Academic >= median(na.omit(Academic)))] <- 1
hist(Academic)
```



Q1.b

```
table(Q3.2)
```

```
## Q3.2
##              Agree           Disagree           Not Sure    Strongly Agree
##                340                127                156                172
## Strongly Disagree
##                 55
```

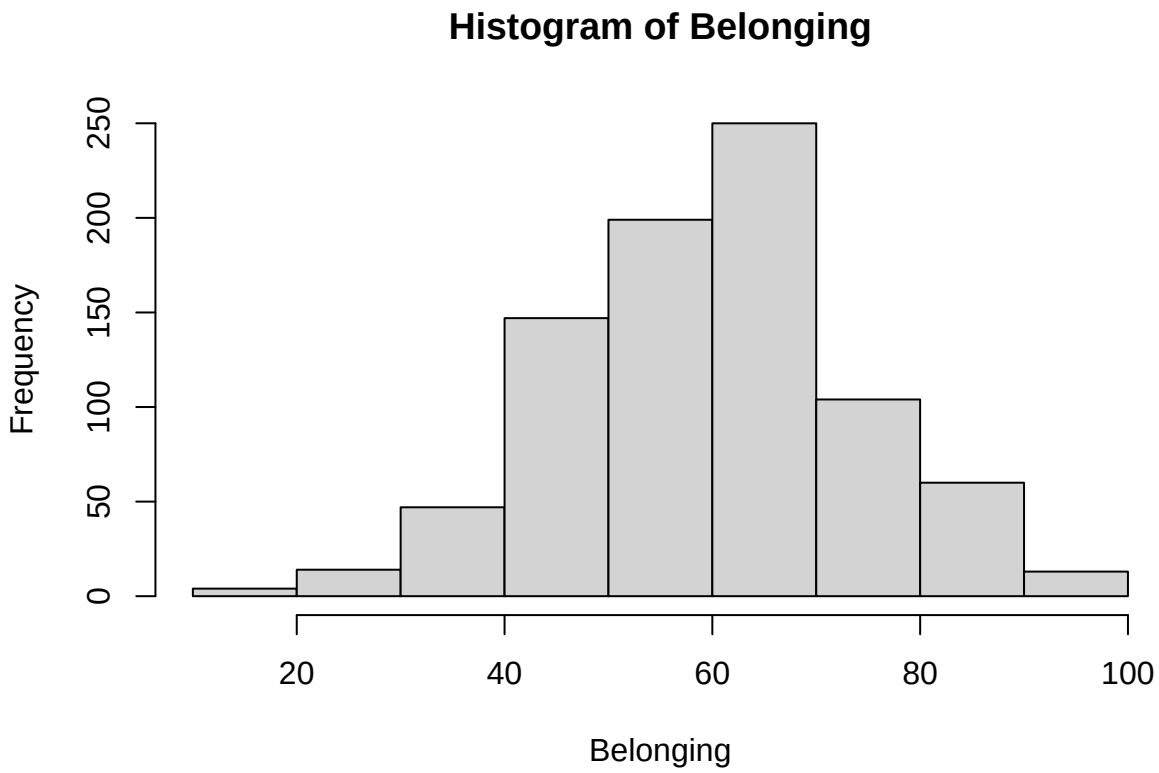
Q1.c

```
table(International)
```

```
## International
## Is English your first language?                          No
##                                1                         694
##                              Yes
##                              414
```

Q1.d

```
hist(Belonging)
```



Q1.e

```
Q3.2r <- recode(Q3.2, 'Strongly Agree' = 'Agree','Strongly Disagree'='Disagree')
Q3.2r <- factor(Q3.2r, levels = c("Agree", "Not Sure", "Disagree"))
unique(Q3.2r)
```

```
## [1] Disagree Not Sure Agree    <NA>
## Levels: Agree Not Sure Disagree
```

```
m1 <- glm(bin_academic~Q3.2r*International+Transfer*Belonging, family = "binomial")
summary(m1)
```

```
##
## Call:
## glm(formula = bin_academic ~ Q3.2r * International + Transfer *
##     Belonging, family = "binomial")
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2330  -0.8508   0.3689   0.8685   2.5506
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -3.779037   0.507403  -7.448 9.49e-14 ***
## Q3.2rNot Sure                  -0.967302   0.264607  -3.656 0.000257 ***
## Q3.2rDisagree                  -2.239536   0.276403  -8.102 5.39e-16 ***
## InternationalYes                0.307986   0.216662   1.422 0.155171
## TransferYes                    -0.594180   1.082200  -0.549 0.582972
## Belonging                       0.072929   0.008157   8.941  < 2e-16 ***
## Q3.2rNot Sure:InternationalYes -0.035347   0.447691  -0.079 0.937069
## Q3.2rDisagree:InternationalYes  0.976974   0.544880   1.793 0.072972 .
## TransferYes:Belonging           0.003633   0.018174   0.200 0.841547
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1138.58  on 821  degrees of freedom
## Residual deviance:  870.32  on 813  degrees of freedom
##   (306 observations deleted due to missingness)
## AIC: 888.32
##
## Number of Fisher Scoring iterations: 4
```

Q1.f

```
round(exp(cbind(Estimate=coef(m1), confint(m1))),2)
```

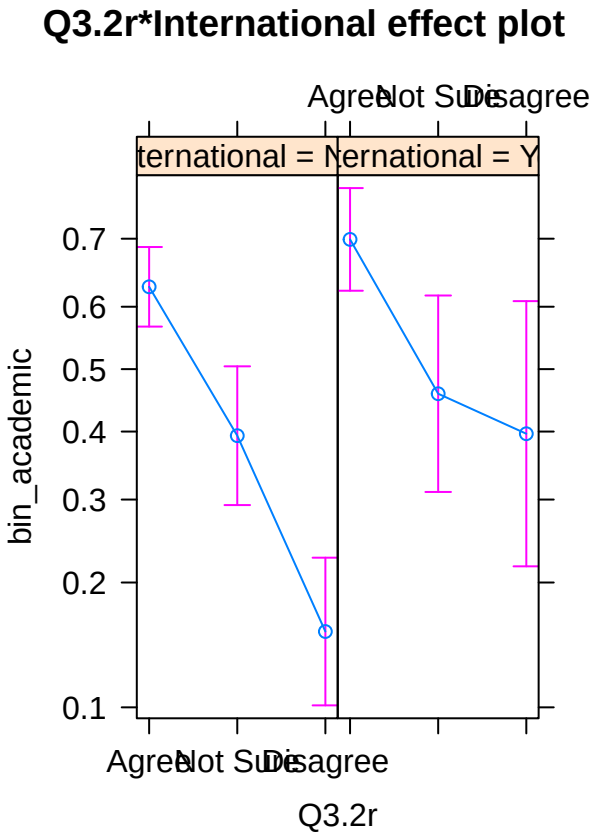
```
##                                Estimate 2.5 % 97.5 %
## (Intercept)                        0.02  0.01   0.06
## Q3.2rNot Sure                      0.38  0.22   0.64
## Q3.2rDisagree                      0.11  0.06   0.18
## InternationalYes                   1.36  0.89   2.09
## TransferYes                        0.55  0.06   4.26
## Belonging                          1.08  1.06   1.09
## Q3.2rNot Sure:InternationalYes     0.97  0.40   2.32
## Q3.2rDisagree:InternationalYes     2.66  0.89   7.63
## TransferYes:Belonging              1.00  0.97   1.04
```

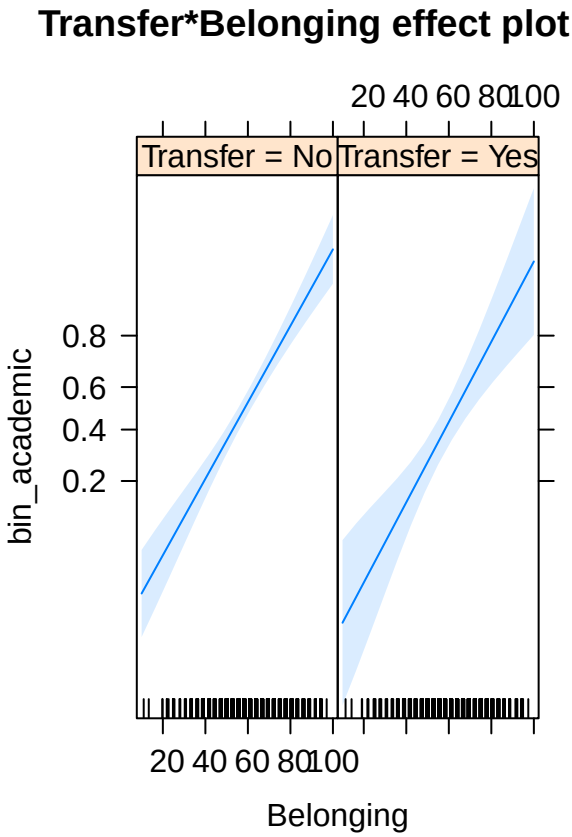
Q1.g

ANSWER: *Null hypothesis:* The predictors of college preparedness, whether a student is international or not and a student's sense of belonging are not significant in determining a student's satisfaction with academics at UCLA. Alternate hypothesis: The predictors of college preparedness, whether a student is international or not and a student's sense of belonging are significant in determining a student's satisfaction with academics at UCLA.

Q1.h

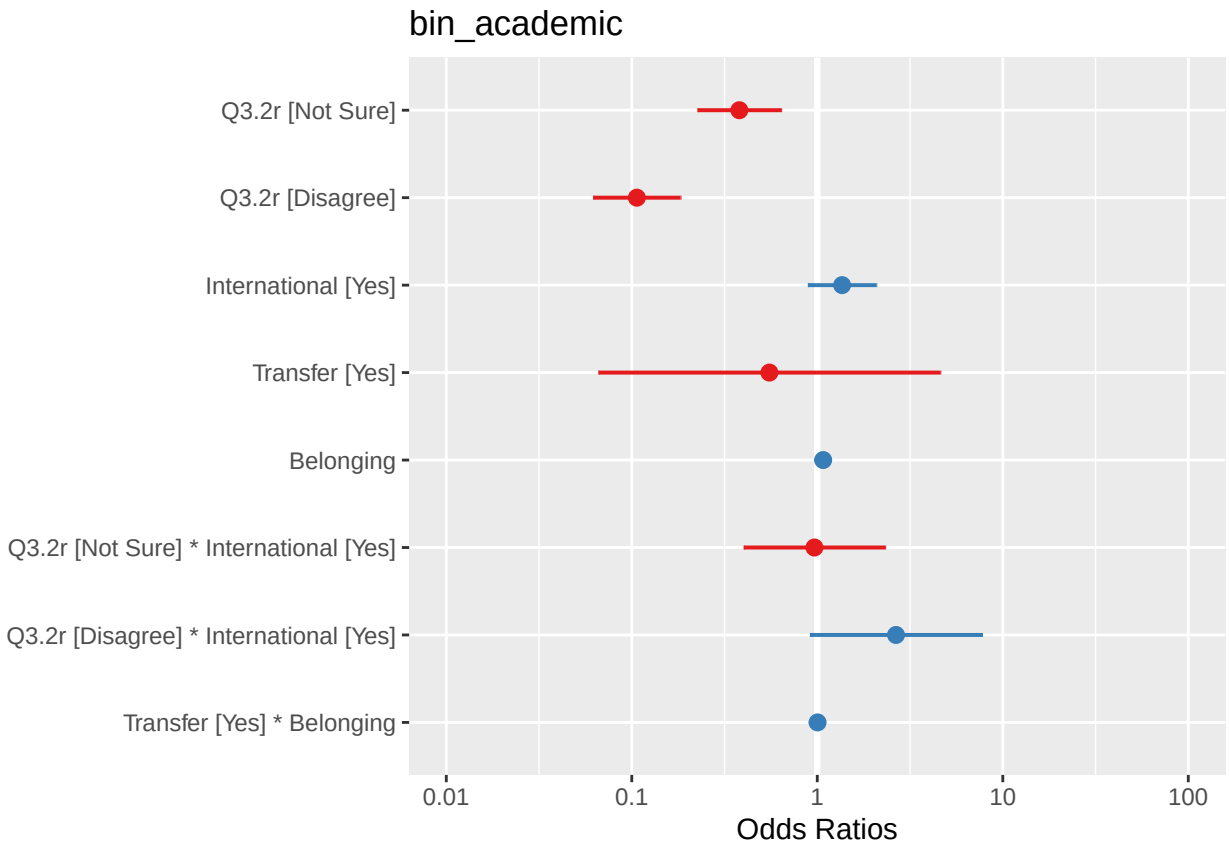
```
Transfer <- factor(Transfer)
bin_academic <- factor(bin_academic)
plot(allEffects(m1), ask=FALSE)
```





Q1.i

```
plot_model(m1)
```



ANSWER: People who disagreed with Q3.2 (thought high school did not prepare them well) did not have high academic confidence. Transfer and international students didn't seem to have a major difference in their academic confidence. Students with a higher sense of belonging seemed to have a higher overall academic confidence.

Q1.j

ANSWER: The null deviance shows how well the model predicts the response with just the intercept.

Q1.k

ANSWER: The residual deviance shows how well the model predicts the response with the predictors.

Q1.l

```
m2 <- glm(bin_academic~1, family = "binomial")
summary(m2)
```

```
##
## Call:
## glm(formula = bin_academic ~ 1, family = "binomial")
##
```

```
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.204  -1.204   1.151   1.151   1.151
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.06222    0.06920   0.899    0.369
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1158.1  on 835  degrees of freedom
## Residual deviance: 1158.1  on 835  degrees of freedom
##   (292 observations deleted due to missingness)
## AIC: 1160.1
##
## Number of Fisher Scoring iterations: 3
```

Q1.m

ANSWER: In the intercept only model, the null and residual deviance are the same meaning that there are no predictors, since the residual deviance takes the intercept and predictors into account when evaluating the model. The null deviance is different in both models because there are differing degrees of freedom.

Q1.n

```
Ps_r2 <- 1 - 870.32/1138.58
Ps_r2
```

```
## [1] 0.2356093
```

Q1.o

```
m3 <- multinom(bin_academic~Q3.2r*International+Transfer*Belonging)
```

```
## # weights:  20 (19 variable)
## initial  value 569.766982
## iter  10 value 436.950814
## final  value 435.158716
## converged
```

```
pred <- predict(m3, bin_academic)
table(pred, bin_academic)
```

```
##     bin_academic
## pred   0   1
##    0 269  90
##    1 128 335
```

```r
(269+335)/(269+335+90+128)
```

```
## [1] 0.7347932
```

Q1.p

```r
m4 <- glm(bin_academic~Q3.2r*International+Transfer+Belonging, family = "binomial")
summary(m4)
```

```
##
## Call:
## glm(formula = bin_academic ~ Q3.2r * International + Transfer +
##     Belonging, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2393  -0.8506   0.3720   0.8690   2.5372
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -3.823513   0.457657  -8.355  < 2e-16 ***
## Q3.2rNot Sure                  -0.966424   0.264662  -3.652 0.000261 ***
## Q3.2rDisagree                  -2.242740   0.276347  -8.116 4.83e-16 ***
## InternationalYes                0.308243   0.216570   1.423 0.154651
## TransferYes                    -0.382267   0.218686  -1.748 0.080461 .
## Belonging                       0.073665   0.007308  10.080  < 2e-16 ***
## Q3.2rNot Sure:InternationalYes -0.032138   0.446800  -0.072 0.942658
## Q3.2rDisagree:InternationalYes  0.974810   0.544864   1.789 0.073600 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1138.58  on 821  degrees of freedom
## Residual deviance:  870.36  on 814  degrees of freedom
##   (306 observations deleted due to missingness)
## AIC: 886.36
##
## Number of Fisher Scoring iterations: 4
```

Q1.q

```r
anova(m1, m4, test = "Chisq")
```

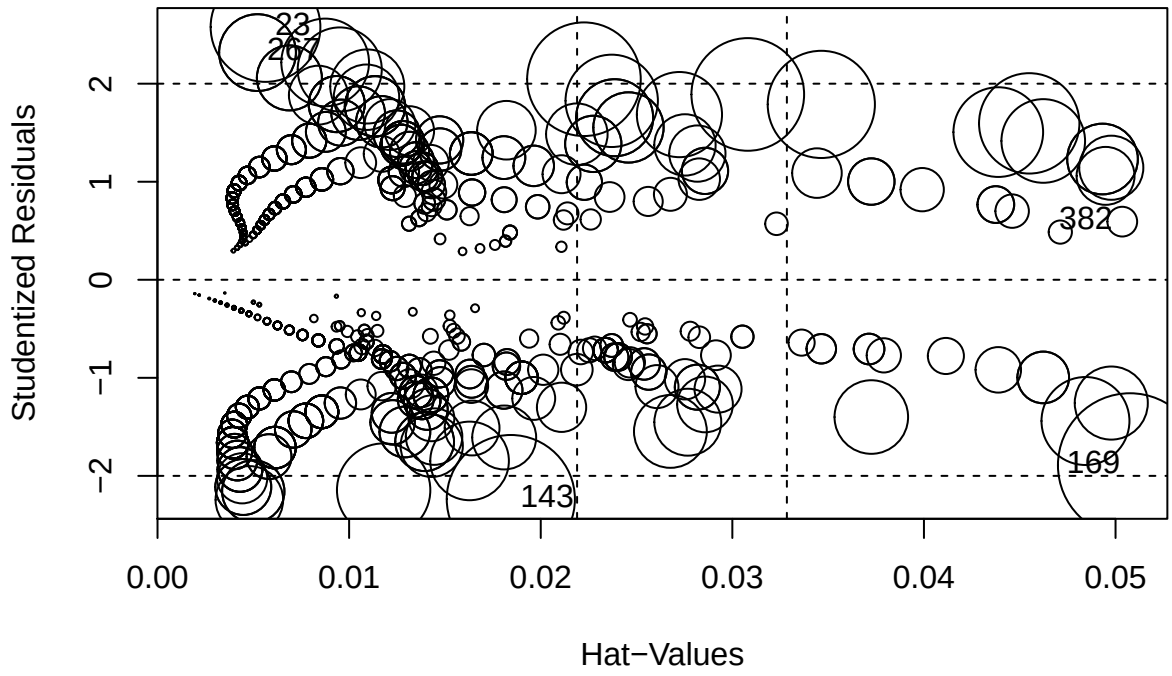
```
## Analysis of Deviance Table
##
## Model 1: bin_academic ~ Q3.2r * International + Transfer * Belonging
## Model 2: bin_academic ~ Q3.2r * International + Transfer + Belonging
```

```
##   Resid. Df Resid. Dev Df  Deviance Pr(>Chi)
## 1       813     870.32
## 2       814     870.36 -1 -0.040339   0.8408
```

ANSWER: With a p-value of .8408, we can say there is no statistically significant difference between the model with and without the interaction effect with Transfer and Belonging.

Q1.r

```
influencePlot(m1)
```




```
##        StudRes         Hat       CookD
## 23   2.5780985 0.005635615 0.015746074
## 143 -2.2368040 0.018438737 0.021473577
## 169 -1.8917433 0.050751359 0.027104865
## 267  2.3169495 0.005218831 0.007704863
## 382  0.5929744 0.050350204 0.001155922
```

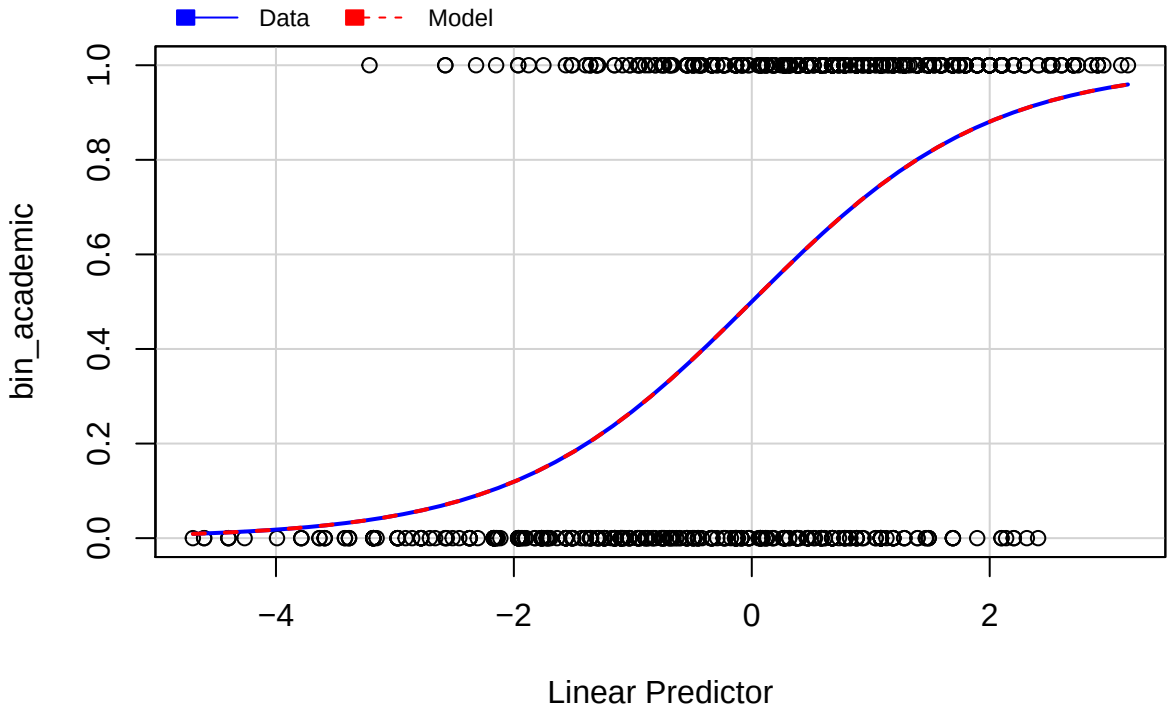
ANSWER: Observation 143, 382, and 169 all have high leverage scores, however their standardized residual isn't that high, so there are no leverage points to worry about.

Q1.s

ANSWER: From the plot above and the information above we see that there are three points with high leverage. Because this is a relatively large data set, a high standardized residual would be +4 or -4. Seeing how no points accomplish this, we can say there are no bad leverages that we need to worry about.

Q1.t

```
mmp(m1)
```



ANSWER: The mmps(...) function tells us how accurate our model is with the data. We can see here that the lines are practically on top of each other, indicating that our model matches our data. This shows it is a good analysis of the data.

Q1.u

ANSWER: From the plot above (in Q1.t), we see that the blue and red lines are basically on top of each other and that it creates an S shape. Both of these give us that logistic regression was a good fit for our data.

Problem 2

Data Set-up

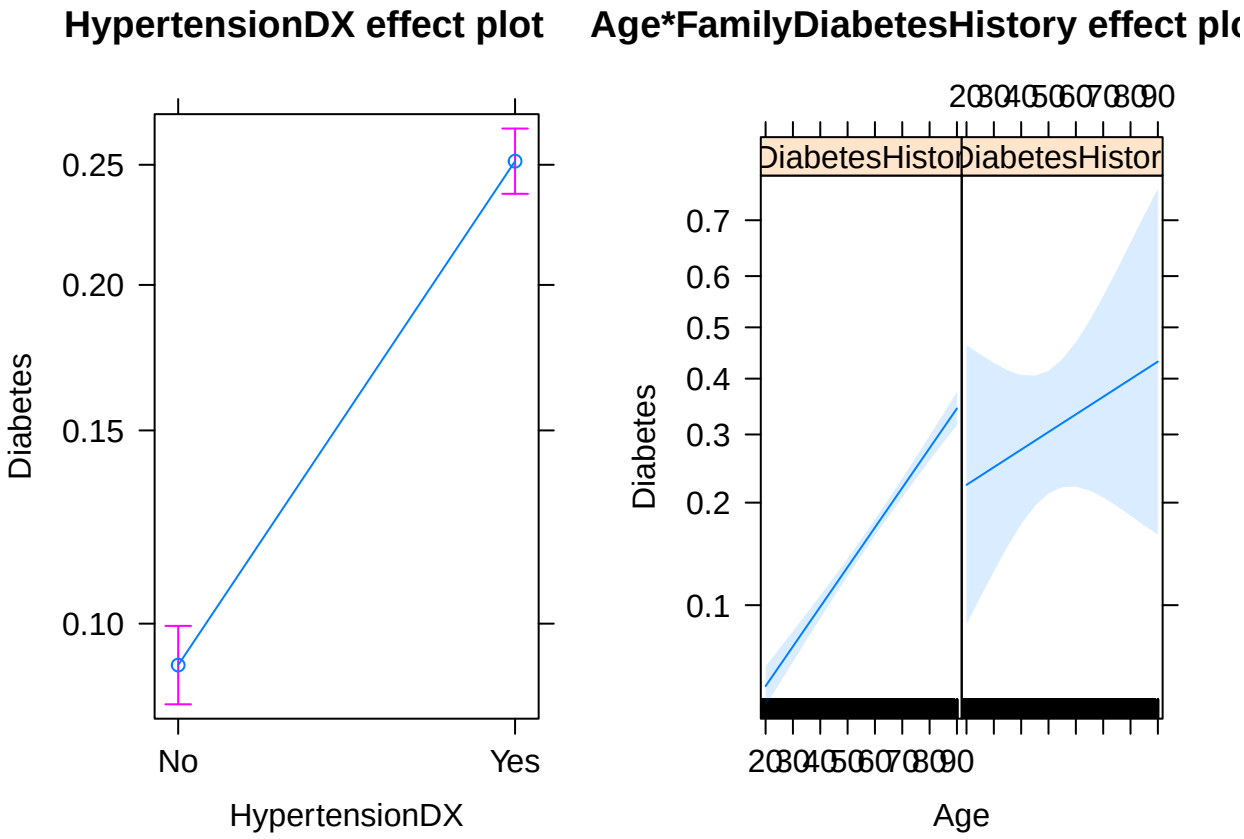
```
# setting up as.factor(...)
d$Diabetes <- as.factor(d$Diabetes)
d$FamilyDiabetesHistory <- as.factor(d$FamilyDiabetesHistory)
```

Q2.1

```
m2.1 <- glm(Diabetes ~
              HypertensionDX +
              Age * FamilyDiabetesHistory,
            data = d, family = "binomial")
summary(m2.1)
```

```
##
## Call:
## glm(formula = Diabetes ~ HypertensionDX + Age * FamilyDiabetesHistory,
##     family = "binomial", data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2909  -0.6877  -0.4101  -0.2889   2.5877
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -3.97103    0.11354 -34.976   <2e-16 ***
## HypertensionDXYes               1.20645    0.06152  19.611   <2e-16 ***
## Age                             0.03137    0.00189  16.602   <2e-16 ***
## FamilyDiabetesHistoryYes        1.94088    0.87990   2.206   0.0274 *
## Age:FamilyDiabetesHistoryYes -0.01744    0.01665  -1.048   0.2947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9422.3  on 9947  degrees of freedom
## Residual deviance: 8270.9  on 9943  degrees of freedom
## AIC: 8280.9
##
## Number of Fisher Scoring iterations: 5
```

```
# Check for Interaction Effect
# For every increase in Age, we can see there is an increase in likely of having Diabetes. It is simila
# HOWEVER, this is NOT statistically significant!
plot(allEffects(m2.1), ask = FALSE)
```




```
# Interpretation of Age
exp(m2.1$coefficients[3] * 10) # Age every 10 years
```

```
##      Age
## 1.368518
```

ANSWER: We can see here that the odds of people with Hypertension (*HypertensionDXYes*) are approximately 20.65% more likely to experience Diabetes. As for Age, we needed to adjust the variable by finding the exponential and multiplying it by 10 (for every 10 years of Age). After the adjustments, we see that the odds of people every 10 years of Age increase will be 36.85% more likely to have Diabetes. The odds of people with a Family History of having Diabetes approximately (*FamilyDiabetesHistoryYes*) is approximately 94.09% more likely to experience Diabetes in their life. Finally, we see that Age and people with a Family History of Diabetes interaction effect is not statistically significant.

Q2.2

```
split <- sample.split(d$Diabetes, SplitRatio = 0.65)
train <- subset(d, split = TRUE)
test <- subset(d, split = FALSE)
p <- predict(m2.1, newdata = test, type = "response")
summary(p)[3] # using the Median as the threshold of 0.1336525
```

```
##    Median
## 0.1336525
```

```r
t1 <- table(d$Diabetes, p > summary(p)[3])
sum(diag(t1))/sum(t1) # accuracy
```

```
## [1] 0.6142943
```

ANSWER: We can see that using the median as threshold, we have an accuracy of approximately 61.43% for our model.

Q2.3

```r
summary(p)[4] # using the Mean as the threshold of 0.1814435
```

```
##      Mean
## 0.1814435
```

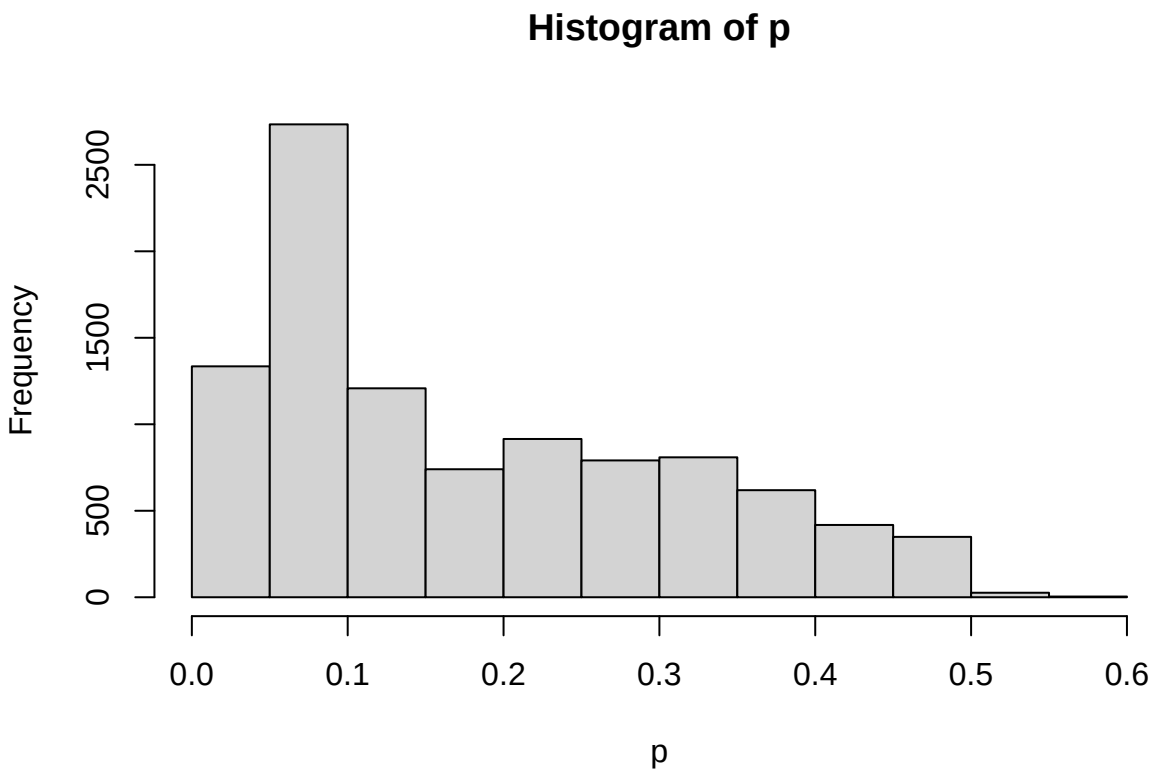
```r
t2 <- table(d$Diabetes, p > summary(p)[4])
sum(diag(t2))/sum(t2) # accuracy
```

```
## [1] 0.6640531
```

ANSWER: We can see that using the mean as threshold, we have an accuracy of approximately 66.41% for our model. By comparison, we can see that using the mean as the threshold will give us about a 5% higher accuracy than using the median as a threshold.

Q2.4

```r
pred <- prediction(p, d$Diabetes) # still using 65% as testing
hist(p) # most fall behind 0.5 on the histogram (potential cut-off)
```

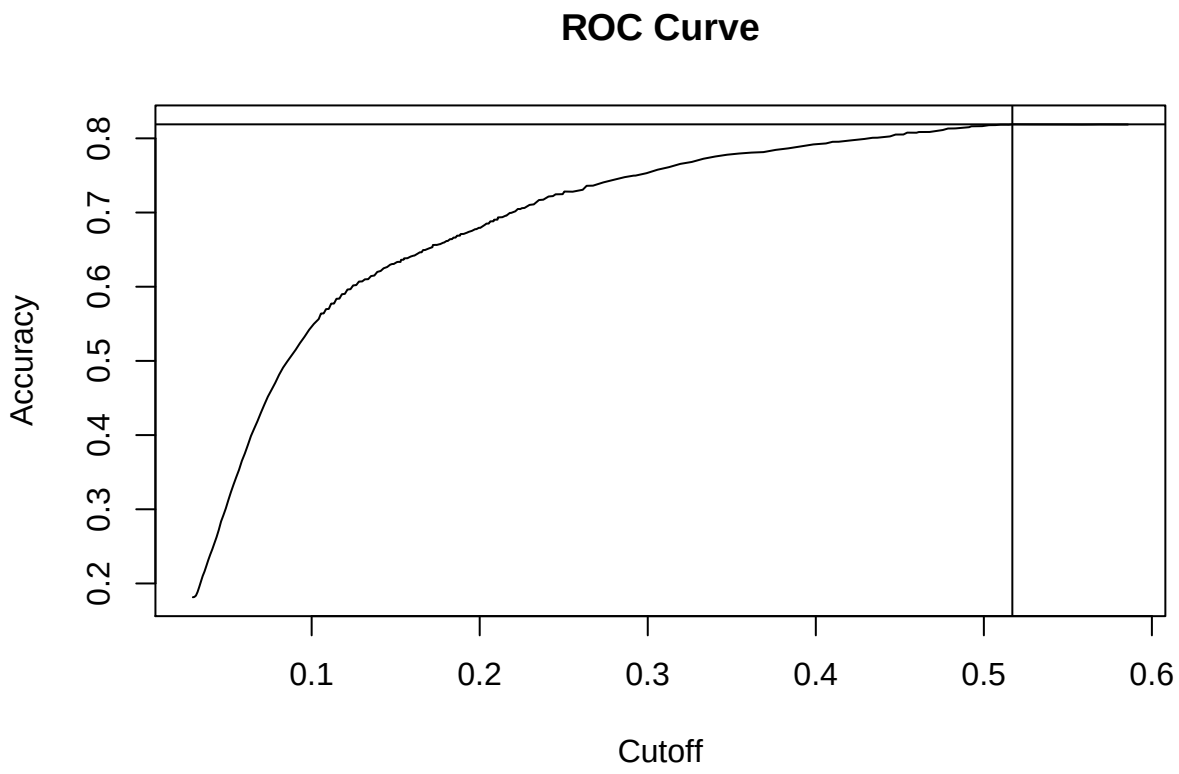



```
eval <- performance(pred, "acc")

# Estimation of the cutoff that creates the maximum accuracy
max <- which.max(slot(eval, "y.values")[[1]])
acc <- slot(eval, "y.values")[[1]][max]
cut <- slot(eval, "x.values")[[1]][max]
print(c(Accuracy=acc, Cutoff = cut))
```

```
##    Accuracy Cutoff.2849
##   0.8188581   0.5169476
```

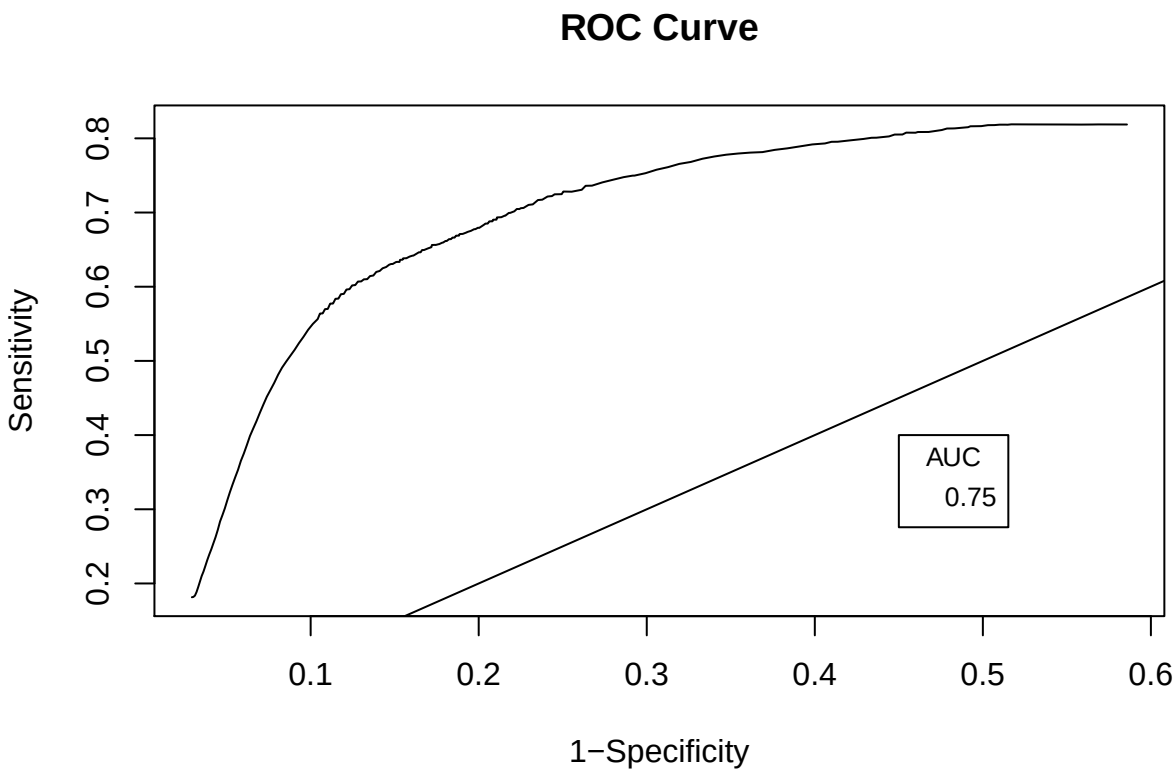
```
# Plot ROC Curve w/ accuracy & cut-off
plot(eval, main = "ROC Curve")
abline(h=0.8188581,v=0.5169476)
```




```
# Find AUC
auc <- performance(pred, "auc")
auc <- unlist(slot(auc,"y.values"))
auc <- round(auc,2)
auc
```

```
## [1] 0.75
```

```
# Plot AUC and ROC Curve
plot(eval, main="ROC Curve", ylab="Sensitivity", xlab="1-Specificity")
abline(0,1)
legend(0.45,0.4,auc,title="AUC",cex=0.8)
```



ANSWER: Using the ROC Curve and plotting it, we can see that the AUC is 75% to our model. We have an accuracy of approximately 81.89% to our model with a cut-off of 51.68%.

Q2.5

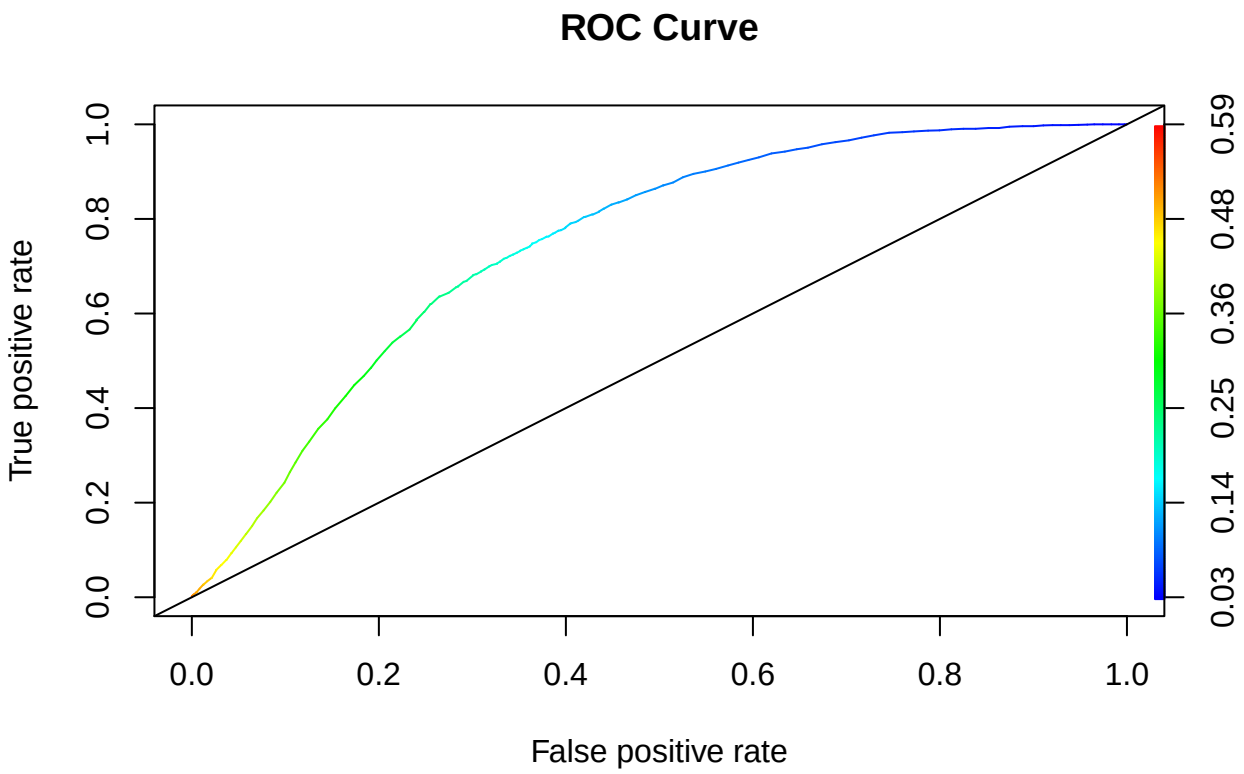
```
t3 <- table(test$Diabetes,p > 0.5168)
sum(diag(t3))/sum(t3)  # Accuracy
```

```
## [1] 0.8188581
```

ANSWER: The accuracy for our model based on the cut-off is equal to 81.88 %.

Q2.6

```
roc_curve <- performance(pred, "tpr", "fpr")
plot(roc_curve, colorize = T, main = "ROC Curve")
abline(0,1)
```



ANSWER: For the plot being shown above, ROC curves show the relationship between sensitivity (true positive rate - tpr) and (1-specificity) or (the false positive rate; fpr).

Q2.7

```
fitControl <- trainControl(method = "cv", number = 5, savePredictions = T)
mod_fitcv <- train(Diabetes ~ HypertensionDX + Age*FamilyDiabetesHistory, data = d, method = "glm", fami

summary(mod_fitcv)
```

```
##
## Call:
## NULL
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2909  -0.6877  -0.4101  -0.2889   2.5877
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -3.97103    0.11354 -34.976   <2e-16 ***
## HypertensionDXYes               1.20645    0.06152  19.611   <2e-16 ***
## Age                             0.03137    0.00189  16.602   <2e-16 ***
## FamilyDiabetesHistoryYes        1.94088    0.87990   2.206   0.0274 *
```

```
## `Age:FamilyDiabetesHistoryYes` -0.01744    0.01665  -1.048   0.2947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9422.3  on 9947  degrees of freedom
## Residual deviance: 8270.9  on 9943  degrees of freedom
## AIC: 8280.9
##
## Number of Fisher Scoring iterations: 5
```

```
mod_fitcv
```

```
## Generalized Linear Model
##
## 9948 samples
##    3 predictor
##    2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 7959, 7958, 7959, 7958, 7958
## Resampling results:
##
##   Accuracy   Kappa
##   0.8169484  0.004483567
```

ANSWER: The accuracy resulting from five-fold cross validation is equal to 81.68% which is less than the cutoff from the ROC curve.

Q2.8

```
confusionMatrix(table((mod_fitcv$pred)$pred,(mod_fitcv$pred)$obs))
```

```
## Confusion Matrix and Statistics
##
##
##         No  Yes
##   No  8116 1794
##   Yes   27   11
##
##                Accuracy : 0.8169
##                  95% CI : (0.8092, 0.8245)
##     No Information Rate : 0.8186
##     P-Value [Acc > NIR] : 0.667
##
##                   Kappa : 0.0045
##
##  Mcnemar's Test P-Value : <2e-16
##
```

```
##             Sensitivity : 0.996684
##             Specificity : 0.006094
##          Pos Pred Value : 0.818971
##          Neg Pred Value : 0.289474
##              Prevalence : 0.818556
##          Detection Rate : 0.815842
##    Detection Prevalence : 0.996180
##       Balanced Accuracy : 0.501389
##
##        'Positive' Class : No
##
```

ANSWER: The sensitivity refers to the percentage of true positives and specificity is the percentage of true negatives. Sensitivity is calculated by dividing number of True Positives by the sum of True positives and False Negatives. Specificity is calculated by dividing the number of True Negatives by the sum of True negatives and false positives. It does better with Sensitivity because we have approximately 99.66% compared to the Specificity of about 0.72%.