STAT 426 Assignment 8

Due Tuesday, October 26, 11:59 pm.

Submit through Moodle.

Name: Brianna Diaz

Netid: bdiaz22

Submit your work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown. Be sure to show your work.

Problem 1. (6 pts)

This problem refers to the horseshoe crab example of Notes_8_1, where various logistic regression models were compared. The data are available as "horseshoe.txt" from the Moodle folder, "Data sets for lecture notes and assignments." In the notes the variable `weight` was eliminated from consideration due to its correlation with `width`. In this problem we allow both `weight` and `width` to be candidate variables for the model.

```
library(readr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v dplyr   1.0.7
## v tibble  3.1.4     v stringr 1.4.0
## v tidyr   1.1.3     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
horseshoe <- read.table("horseshoe.txt", header=TRUE)
```

a) (2 pts) Starting from the intercept only model, try a forward selection that allows `width`, `weight`, and the two factor variables, `color` and `spine`, as candidate variables for the model, but not their interactions. Using AIC as the selection criterion, show the steps and a summary of the final model. Is the final model the same as the final model in the class notes?

```
hsfit <- glm(y~ weight + width + factor(color) + factor(spine), family=binomial, data =
#summary(hsfit)

nullmod <- glm(y~1, family=binomial, data=horseshoe)

formod <- step(nullmod, ~ width + weight + factor(color) + factor(spine), direction="for
```

```
## Start:  AIC=227.76
## y ~ 1
##
##                 Df Deviance    AIC
## + width          1   194.45 198.45
## + weight         1   195.74 199.74
## + factor(color)  3   212.06 220.06
## <none>               225.76 227.76
## + factor(spine)  2   223.23 229.23
##
## Step:  AIC=198.45
## y ~ width
##
##                 Df Deviance    AIC
## + factor(color)  3   187.46 197.46
## <none>               194.45 198.45
## + weight         1   192.89 198.89
## + factor(spine)  2   194.43 202.43
##
## Step:  AIC=197.46
## y ~ width + factor(color)
##
##                 Df Deviance    AIC
## <none>               187.46 197.46
## + weight         1   186.21 198.21
## + factor(spine)  2   186.61 200.61
```

```
summary(formod)
```

```
##
## Call:
## glm(formula = y ~ width + factor(color), family = binomial, data = horseshoe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1124  -0.9848   0.5243   0.8513   2.1413
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -11.38519    2.87346  -3.962 7.43e-05 ***
## width            0.46796    0.10554   4.434 9.26e-06 ***
## factor(color)3   0.07242    0.73989   0.098    0.922
## factor(color)4  -0.22380    0.77708  -0.288    0.773
## factor(color)5  -1.32992    0.85252  -1.560    0.119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 187.46  on 168  degrees of freedom
## AIC: 197.46
##
## Number of Fisher Scoring iterations: 4
```

```
#In comparison to the class notes the numbers are the same.
```

b) (2 pts) Perform backward selection starting from the additive model that includes all four of the variables listed in Part a). Show the steps and final model. Is the selected model the same as the final backward selection model in the notes?

```
fullmod <- glm(y~width + weight + factor(color) + factor(spine), family=binomial, data=h

backmod <- step(fullmod)
```

```
## Start:  AIC=201.2
## y ~ width + weight + factor(color) + factor(spine)
##
##                 Df Deviance    AIC
## - factor(spine)  2   186.21 198.21
## - weight         1   186.61 200.61
## - width          1   187.00 201.00
## <none>               185.20 201.20
```

```
## - factor(color)  3   192.80 202.80
##
## Step:  AIC=198.21
## y ~ width + weight + factor(color)
##
##                 Df Deviance    AIC
## - weight         1   187.46 197.46
## <none>               186.21 198.21
## - width          1   188.54 198.54
## - factor(color)  3   192.89 198.89
##
## Step:  AIC=197.46
## y ~ width + factor(color)
##
##                 Df Deviance    AIC
## <none>               187.46 197.46
## - factor(color)  3   194.45 198.45
## - width          1   212.06 220.06
```

```
summary(backmod)
```

```
##
## Call:
## glm(formula = y ~ width + factor(color), family = binomial, data = horseshoe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1124  -0.9848   0.5243   0.8513   2.1413
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -11.38519    2.87346  -3.962 7.43e-05 ***
## width            0.46796    0.10554   4.434 9.26e-06 ***
## factor(color)3   0.07242    0.73989   0.098    0.922
## factor(color)4  -0.22380    0.77708  -0.288    0.773
## factor(color)5  -1.32992    0.85252  -1.560    0.119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 187.46  on 168  degrees of freedom
## AIC: 197.46
```

```
##
## Number of Fisher Scoring iterations: 4
```

```
#The AIC numbers are different from the notes
```

c) (2 pts) Perform stepwise selection starting from the null model and allowing all four of the variables listed in Part a) as candidate variables for the model, but not their interactions. Is the selected model the same as the final stepwise model in the notes?

```
stepmod <- step(nullmod, ~ width + weight + factor(color) + factor(spine ), direction="b
```

```
## Start:  AIC=227.76
## y ~ 1
##
##                  Df Deviance    AIC
## + width           1   194.45 198.45
## + weight          1   195.74 199.74
## + factor(color)  3   212.06 220.06
## <none>                225.76 227.76
## + factor(spine)  2   223.23 229.23
##
## Step:  AIC=198.45
## y ~ width
##
##                  Df Deviance    AIC
## + factor(color)  3   187.46 197.46
## <none>                194.45 198.45
## + weight          1   192.89 198.89
## + factor(spine)  2   194.43 202.43
## - width           1   225.76 227.76
##
## Step:  AIC=197.46
## y ~ width + factor(color)
##
##                  Df Deviance    AIC
## <none>                187.46 197.46
## + weight          1   186.21 198.21
## - factor(color)  3   194.45 198.45
## + factor(spine)  2   186.61 200.61
## - width           1   212.06 220.06
```

```
summary(stepmod)
```

```
##
## Call:
## glm(formula = y ~ width + factor(color), family = binomial, data = horseshoe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1124  -0.9848   0.5243   0.8513   2.1413
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -11.38519    2.87346  -3.962 7.43e-05 ***
## width            0.46796    0.10554   4.434 9.26e-06 ***
## factor(color)3   0.07242    0.73989   0.098    0.922
## factor(color)4  -0.22380    0.77708  -0.288    0.773
## factor(color)5  -1.32992    0.85252  -1.560    0.119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 187.46  on 168  degrees of freedom
## AIC: 197.46
##
## Number of Fisher Scoring iterations: 4
```

```
#The selected model is the same as the notes.
```

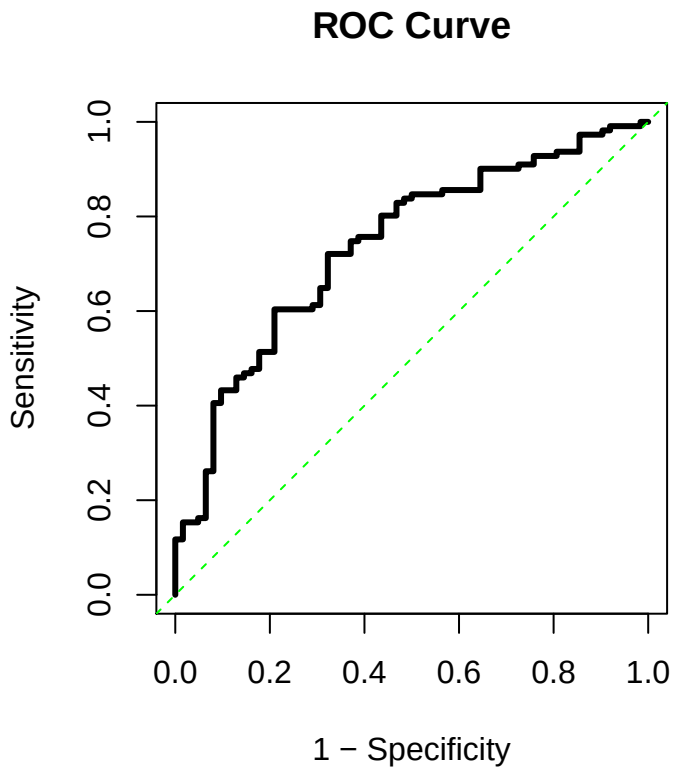
Problem 2. (10 pts)

This problem refers to the horseshoe crab example of Notes_8_3.

a) (2 pts) Consider the additive logistic regression model with `y` as the response and `width` and the factor variable `color` as predictors. In the lecture notes we computed the leave-one-out cross-validation estimates of sensitivity and specificity for this model, with threshold $\pi_0 = 0.5$, but we did not use cross validation for the ROC curve. Redo the ROC curve using the leave-one-out predicted values instead of the fitted values.

```
mod1 <- glm(y ~ width, family=binomial, data=horseshoe)

mod2 <- glm(y ~ factor(color), family=binomial, data=horseshoe)

mod3 <- glm(y ~ width + factor(color), family=binomial, data=horseshoe)
```

```
pi0<- 0.5
pihatcv <- numeric(nrow(horseshoe))
for(i in 1:nrow(horseshoe))
pihatcv[i] <- predict(update(mod3, subset=-i),
newdata=horseshoe[i,],type="response")

#table(y=horseshoe$y, yhat=as.numeric(pihatcv > pi0))

false.neg <- c(0,cumsum(tapply(horseshoe$y,pihatcv,sum)))
true.neg <- c(0,cumsum(table(pihatcv))) - false.neg
par(pty="s")
plot(1-true.neg/max(true.neg), 1-false.neg/max(false.neg), type="l",
main="ROC Curve", xlab="1 - Specificity", ylab="Sensitivity",
xlim=c(0,1), ylim=c(0,1), lwd=3)
abline(a=0, b=1, lty=2, col="green")
```



b) (2 pts) For the ROC curve you computed in Part a), calculate the leave-one-out concordance index (area under the curve). How does the value compare to the concordance index using fitted values that was calculated in the notes?

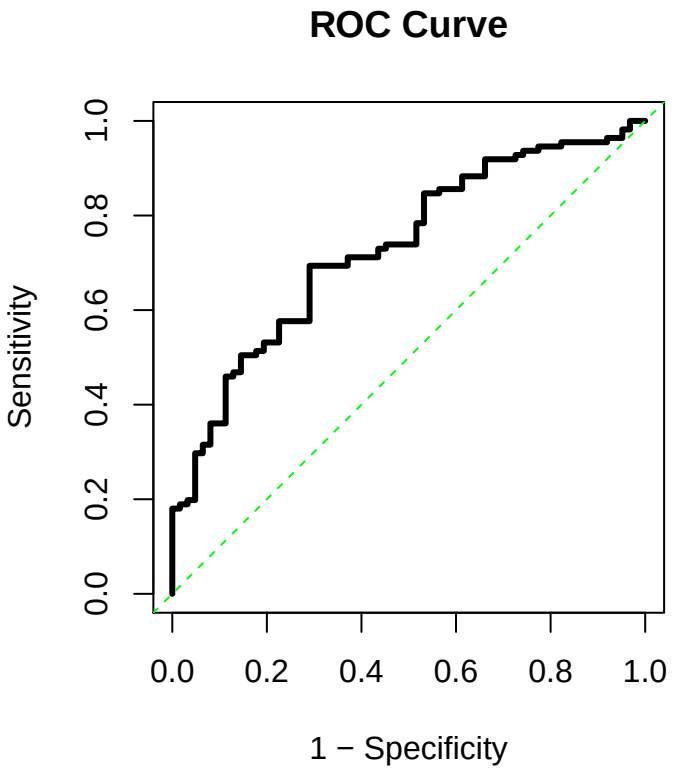
```
mean(outer(pihatcv[horseshoe$y==1], pihatcv[horseshoe$y==0], ">")
+ 0.5 * outer(pihatcv[horseshoe$y==1], pihatcv[horseshoe$y==0], "=="))
```

```
## [1] 0.7375763
```

```
#The index does not match the one in the notes
```

c) (2 pts) Next consider the logistic regression model for `y` that only includes `width` as a predictor. Compute and display its ROC curve using the leave-one-out predictions.

```
pihatcv2 <- numeric(nrow(horseshoe))
for(i in 1:nrow(horseshoe))
pihatcv2[i] <- predict(update(mod1, subset=-i),
newdata=horseshoe[i,],type="response")

#table(y=horseshoe$y, yhat=as.numeric(pihatcv2 > pi0))

false.neg <- c(0,cumsum(tapply(horseshoe$y,pihatcv2,sum)))
true.neg <- c(0,cumsum(table(pihatcv2))) - false.neg
par(pty="s")
plot(1-true.neg/max(true.neg), 1-false.neg/max(false.neg), type="l",
main="ROC Curve", xlab="1 - Specificity", ylab="Sensitivity",
xlim=c(0,1), ylim=c(0,1), lwd=3)
abline(a=0, b=1, lty=2, col="green")
```



d) (2 pts) For the ROC curve you computed in Part c), calculate the leave-one-out concordance index (area under the curve). Is the value much different from the value in b)?

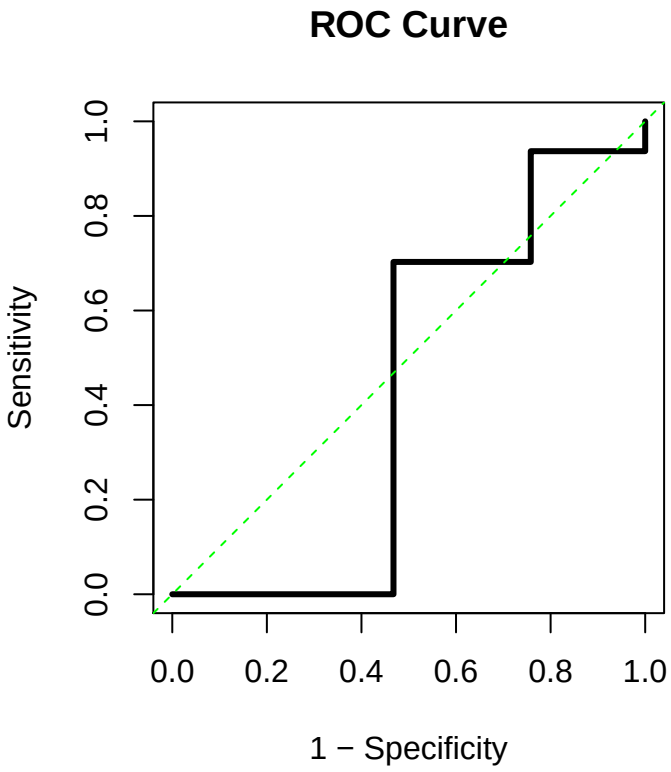
```
mean(outer(pihatcv2[horseshoe$y==1], pihatcv2[horseshoe$y==0], ">")
+ 0.5 * outer(pihatcv2[horseshoe$y==1], pihatcv2[horseshoe$y==0], "=="))
```

```
## [1] 0.731764
```

```
#The index is about the same as b
#Just a difference of 0.0058123
```

e) (2 pts) Finally, consider the logistic regression model for `y` that includes only the factor varaible `color` as a predictor. Compute and display its ROC curve using the leave-one-out predictions. What do you conclude about this model's performance as a classifier?

```
pihatcv3 <- numeric(nrow(horseshoe))
for(i in 1:nrow(horseshoe))
pihatcv3[i] <- predict(update(mod2, subset=-i),
newdata=horseshoe[i,],type="response")

#table(y=horseshoe$y, yhat=as.numeric(pihatcv3 > pi0))

false.neg <- c(0,cumsum(tapply(horseshoe$y,pihatcv3,sum)))
true.neg <- c(0,cumsum(table(pihatcv3))) - false.neg
par(pty="s")
plot(1-true.neg/max(true.neg), 1-false.neg/max(false.neg), type="l",
main="ROC Curve", xlab="1 - Specificity", ylab="Sensitivity",
xlim=c(0,1), ylim=c(0,1), lwd=3)
abline(a=0, b=1, lty=2, col="green")
```




```
#the model had a bad performance.
```

Problem 3 (4 pts)

Once again we consider the horseshoe crab data! The model in Part 2e) that includes the factor variable `color` as the only predictor could be fit using a grouped analysis. Assuming you called the horseshoe data frame "horseshoe", the code below will create grouped data with 0/1 response frequencies for each color level.

```
grouped = data.frame(with(data=horseshoe, table(factor(color), y)))
names(grouped)[1] = "colorlev"
grouped = reshape(grouped, idvar="colorlev", timevar="y", direction="wide")
grouped
```

a) (2 pts) Run/modify the code to create and display the grouped data. Using the grouped data, fit the logistic regression model for `y` that includes `colorlev` as a factor variable. Display the model summary and obtain its residual deviance and residual degrees of freedom.

```
grouped = data.frame(with(data=horseshoe, table(factor(color), y)))
names(grouped)[1] = "colorlev"
grouped = reshape(grouped, idvar="colorlev", timevar = "y", direction="wide")
grouped
```

```
##   colorlev Freq.0 Freq.1
## 1        2      3      9
## 2        3     26     69
## 3        4     18     26
## 4        5     15      7
```

```
pb3a <- glm(cbind(`Freq.0`, `Freq.1`) ~ factor(colorlev), family = binomial, data=groupe

summary(pb3a)
```

```
##
## Call:
## glm(formula = cbind(Freq.0, Freq.1) ~ factor(colorlev), family = binomial,
##     data = grouped)
##
## Deviance Residuals:
## [1]  0  0  0  0
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.0986     0.6667  -1.648   0.0994 .
## factor(colorlev)3   0.1226     0.7053   0.174   0.8620
## factor(colorlev)4   0.7309     0.7338   0.996   0.3192
## factor(colorlev)5   1.8608     0.8087   2.301   0.0214 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.3698e+01  on 3  degrees of freedom
## Residual deviance: 9.3259e-15  on 0  degrees of freedom
## AIC: 23.134
##
## Number of Fisher Scoring iterations: 3
```

```
#residual deviance is 9.3259e-15
#residual degrees of freedom is 0
```

b) (2 pts) For the equivalent ungrouped model from part 2e), show the model summary and obtain the residual deviance and residual degrees of freedom. Summarize how the results compare between the model in Part a) and this model in terms of their coefficient estimates, residual deviance and residual degrees of freedom.

```
summary(mod2)
```

```
##
## Call:
## glm(formula = y ~ factor(color), family = binomial, data = horseshoe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6651  -1.3370   0.7997   0.7997   1.5134
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.0986     0.6667   1.648   0.0994 .
## factor(color)3  -0.1226     0.7053  -0.174   0.8620
## factor(color)4  -0.7309     0.7338  -0.996   0.3192
## factor(color)5  -1.8608     0.8087  -2.301   0.0214 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 225.76  on 172  degrees of freedom
## Residual deviance: 212.06  on 169  degrees of freedom
## AIC: 220.06
##
## Number of Fisher Scoring iterations: 4
```

```
#The coefficient estimates are the same for both
#but the coefficients have different signs.

#2e has a higher residual deviance and degress of freedom than 3a.

#residual deviance is 212.06
#residual degrees of freedom is 169
```