STAT 426 Assignment 9

Due Tuesday, November 2, 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. (10 pts) Variable selection and diagnositc classification

Install faraway package: For this exercise, first make sure you have the `faraway` package installed. Check your R Studio list of packages. If the list includes `faraway`, you have it. If not, click the 'install' button in the packages window, and start typing `faraway` into the packages dialog box. It should auto-complete. Select 'faraway' and hit the **install** button. You will need to issue the command `library(faraway)` to load the data from the package into your environment.

```
library(faraway)
```

The data set `wbca` in the `faraway` package is from a study of breast cancer in Wisconsin. There are 681 cases of potentially cancerous tumors of which 238 are malignant, and the rest are benign. The response variable, `Class`, equals 0 if malignant and 1 if benign. There are 9 possible predictor variables also included in the data. The predictor values were determined by a doctor observing the cells and rating them on a scale from 1 (normal) to 10 (most abnormal) with respect to the particular characteristic. You will want to predict malignant cases, so you might find it helpful to define a new response, `malignant = 1*(Class==0)`. Please see the help file for the data set in the faraway library for more information about the variables.

a) (2 pts) Use logistic regression to fit the model where all predictors are included as additive variables (no interactions). Show the model summary. Which variables are significant at level $\alpha = 0.05$?

```
add <- glm(1*(Class==0) ~ ., family=binomial, data = wbca)
summary(add)
```

```
##
## Call:
## glm(formula = 1 * (Class == 0) ~ ., family = binomial, data = wbca)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.06425  -0.09678  -0.04739   0.01179   2.48282
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.16678    1.41491  -7.892 2.97e-15 ***
## Adhes         0.39681    0.13384   2.965  0.00303 **
## BNucl         0.41478    0.10230   4.055 5.02e-05 ***
## Chrom         0.56456    0.18728   3.014  0.00257 **
## Epith         0.06440    0.16595   0.388  0.69795
## Mitos         0.65713    0.36764   1.787  0.07387 .
## NNucl         0.28659    0.12620   2.271  0.02315 *
## Thick         0.62675    0.15890   3.944 8.01e-05 ***
## UShap         0.28011    0.25235   1.110  0.26699
## USize        -0.05718    0.23271  -0.246  0.80589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 881.388  on 680  degrees of freedom
## Residual deviance:  89.464  on 671  degrees of freedom
## AIC: 109.46
##
## Number of Fisher Scoring iterations: 8
```

```
#Adhes,BNucl,Chrom, NNucl, Thick are all significant at the 0.05 level.
```

b) (2 pts) Use step-wise regression search to find the best subset of variables you can find to minimize AIC. Display the model summary for the selected model. Which variables are significant at level $\alpha = 0.05$?

```
mod1b <- glm(1*(Class==0) ~ 1, family = binomial, data = wbca)

stepmod <- step(mod1b, ~ Adhes + BNucl + Chrom + Epith + Mitos + NNucl + Thick + UShap +
## Start:  AIC=883.39
## 1 * (Class == 0) ~ 1
##
##         Df Deviance    AIC
## + USize  1   251.77 255.77
## + UShap  1   265.32 269.32
## + BNucl  1   331.46 335.46
## + Chrom  1   379.41 383.41
## + Epith  1   450.30 454.30
## + Thick  1   451.69 455.69
## + NNucl  1   454.76 458.76
## + Adhes  1   459.10 463.10
## + Mitos  1   713.09 717.09
## <none>       881.39 883.39
##
## Step:  AIC=255.77
## 1 * (Class == 0) ~ USize
##
##         Df Deviance    AIC
## + BNucl  1   161.41 167.41
## + Thick  1   190.83 196.83
## + Chrom  1   200.26 206.26
## + NNucl  1   216.29 222.29
## + UShap  1   218.28 224.28
## + Adhes  1   224.85 230.85
## + Epith  1   236.32 242.32
## + Mitos  1   237.97 243.97
## <none>       251.77 255.77
## - USize  1   881.39 883.39
##
## Step:  AIC=167.41
## 1 * (Class == 0) ~ USize + BNucl
##
##         Df Deviance    AIC
## + Thick  1   127.76 135.76
## + NNucl  1   143.41 151.41
## + Chrom  1   144.09 152.09
## + UShap  1   148.43 156.43
## + Mitos  1   153.26 161.26
## + Adhes  1   155.48 163.48
## + Epith  1   156.64 164.64
## <none>       161.41 167.41
## - BNucl  1   251.77 255.77
## - USize  1   331.46 335.46
##
```

```
## Step:  AIC=135.76
## 1 * (Class == 0) ~ USize + BNucl + Thick
##
##         Df Deviance    AIC
## + Chrom  1   112.91 122.91
## + NNucl  1   113.34 123.34
## + Adhes  1   117.23 127.23
## + UShap  1   122.43 132.43
## + Epith  1   123.94 133.94
## + Mitos  1   124.98 134.98
## <none>       127.76 135.76
## - Thick  1   161.41 167.41
## - USize  1   184.19 190.19
## - BNucl  1   190.82 196.82
##
## Step:  AIC=122.91
## 1 * (Class == 0) ~ USize + BNucl + Thick + Chrom
##
##         Df Deviance    AIC
## + Adhes  1   102.61 114.61
## + NNucl  1   104.50 116.50
## + Mitos  1   109.64 121.64
## + UShap  1   109.70 121.70
## + Epith  1   110.61 122.61
## <none>       112.91 122.91
## - Chrom  1   127.76 135.76
## - USize  1   129.19 137.19
## - Thick  1   144.09 152.09
## - BNucl  1   148.57 156.57
##
## Step:  AIC=114.61
## 1 * (Class == 0) ~ USize + BNucl + Thick + Chrom + Adhes
##
##         Df Deviance    AIC
## + NNucl  1   95.042 109.04
## + Mitos  1   98.340 112.34
## + UShap  1  100.240 114.24
## <none>      102.608 114.61
## + Epith  1  101.447 115.45
## - USize  1  111.384 121.38
## - Adhes  1  112.913 122.91
## - Chrom  1  117.230 127.23
## - BNucl  1  129.494 139.49
## - Thick  1  139.226 149.23
##
```

```
## Step:  AIC=109.04
## 1 * (Class == 0) ~ USize + BNucl + Thick + Chrom + Adhes + NNucl
##
##         Df Deviance    AIC
## + Mitos  1   90.923 106.92
## - USize  1   96.494 108.49
## <none>       95.042 109.04
## + UShap  1   93.713 109.71
## + Epith  1   94.777 110.78
## - NNucl  1  102.608 114.61
## - Adhes  1  104.496 116.50
## - Chrom  1  105.792 117.79
## - BNucl  1  120.039 132.04
## - Thick  1  129.419 141.42
##
## Step:  AIC=106.92
## 1 * (Class == 0) ~ USize + BNucl + Thick + Chrom + Adhes + NNucl +
##     Mitos
##
##         Df Deviance    AIC
## - USize  1   91.884 105.88
## <none>       90.923 106.92
## + UShap  1   89.613 107.61
## + Epith  1   90.627 108.63
## - Mitos  1   95.042 109.04
## - NNucl  1   98.340 112.34
## - Adhes  1  100.870 114.87
## - Chrom  1  102.551 116.55
## - Thick  1  115.780 129.78
## - BNucl  1  116.676 130.68
##
## Step:  AIC=105.88
## 1 * (Class == 0) ~ BNucl + Thick + Chrom + Adhes + NNucl + Mitos
##
##         Df Deviance    AIC
## + UShap  1   89.662 105.66
## <none>       91.884 105.88
## + USize  1   90.923 106.92
## + Epith  1   91.355 107.36
## - Mitos  1   96.494 108.49
## - NNucl  1  103.711 115.71
## - Adhes  1  105.473 117.47
## - Chrom  1  109.699 121.70
## - BNucl  1  124.813 136.81
## - Thick  1  130.842 142.84
```

```
##
## Step:  AIC=105.66
## 1 * (Class == 0) ~ BNucl + Thick + Chrom + Adhes + NNucl + Mitos +
##     UShap
##
##         Df Deviance    AIC
## <none>       89.662 105.66
## - UShap  1   91.884 105.88
## + Epith  1   89.523 107.52
## + USize  1   89.613 107.61
## - Mitos  1   93.714 107.71
## - NNucl  1   95.853 109.85
## - Adhes  1  100.126 114.13
## - Chrom  1  100.844 114.84
## - BNucl  1  109.762 123.76
## - Thick  1  110.632 124.63
```

```
summary(stepmod)
```

```
##
## Call:
## glm(formula = 1 * (Class == 0) ~ BNucl + Thick + Chrom + Adhes +
##     NNucl + Mitos + UShap, family = binomial, data = wbca)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.08205  -0.09741  -0.04962   0.01119   2.44161
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -11.0333     1.3632  -8.094 5.79e-16 ***
## BNucl         0.4192     0.1020   4.111 3.93e-05 ***
## Thick         0.6216     0.1579   3.937 8.27e-05 ***
## Chrom         0.5679     0.1840   3.085  0.00203 **
## Adhes         0.3984     0.1294   3.080  0.00207 **
## NNucl         0.2915     0.1236   2.358  0.01837 *
## Mitos         0.6456     0.3634   1.777  0.07561 .
## UShap         0.2541     0.1785   1.423  0.15461
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 881.388  on 680  degrees of freedom
## Residual deviance:  89.662  on 673  degrees of freedom
## AIC: 105.66
##
## Number of Fisher Scoring iterations: 8
```

```
#Best subset is 1 * (Class == 0) ~ BNucl + Thick + Chrom + Adhes + NNucl + Mitos + USh

#significant variables: Adhes, BNucl, Chrom, NNucl, Thick
```

c) (2 pts) Based on your selected model, find the predicted probability that a tumor is malignant if each predictor is rated as 5, and also provide a 95% confidence interval for the probability.

```
pred <- predict(stepmod, newdata = data.frame(BNucl=5, Thick=5, Chrom=5, Adhes=5, NNucl=

pred
```

```
## $fit
##         1
## 0.9930239
##
## $se.fit
##          1
## 0.01064508
##
## $residual.scale
## [1] 1
```

```
low <- 0.9930239 - 1*0.01064508*1.96
high <- 0.9930239 + 1*0.01064508*1.96

c(low,high)
```

```
## [1] 0.9721595 1.0138883
```

d) (2 pts) Suppose, based on your selected model, that we decide to classify a tumor as malignant if the predicted probability of malignancy is 0.4 or greater. Calculate the apparent sensitivity and specificity based on the data.

```
pi0 <- 0.4
table(y=wbca$Class, yhat=as.numeric(fitted(stepmod) > pi0))
```

```
##    yhat
## y     0   1
##   0   8 230
##   1 433  10
```

```
#sensitivity
10/(433+10)
```

```
## [1] 0.02257336
```

```
#specificity
8/(8+230)
```

```
## [1] 0.03361345
```

e) (2 pts) Under the same setup as in part d), calculate the leave-one-out cross validation estimates of sensitivity and specificity.

```
pihatcv <- numeric(nrow(wbca))

for(i in 1:nrow(wbca))
pihatcv[i] <- predict(update(stepmod, subset=-i),
newdata=wbca[i,],type="response")

table(y=wbca$Class, yhat=as.numeric(pihatcv > pi0))
```

```
##    yhat
## y     0   1
##   0  11 227
##   1 433  10
```

```
#sensitivity
10/(10+433)
```

```
## [1] 0.02257336
```

```
#specificity
227/(11+227)
```

```
## [1] 0.9537815
```

Problem 2. (8 pts) Revisiting the Death Penalty Study

This problem involves further analysis of the death penalty data. The data are available as "deathpenalty.txt" from the class Moodle site in the folder, "Data sets for lecture notes and assignments." Recall that the data had frequencies of death penalty decisions (yes or no) along with the race of the defendant and victim.

a) (2 pts) After reshaping the data appropriately, use logistic regression to fit the homogeneous association model where the response (Deathpenalty = yes or no) depends only on the race of the defendant and the race of the victim, but not their interaction. Display the model summary.

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

death <- reshape(deathpenalty,varying=list(c("No","Yes")), v.names = "Freq", timevar = "

death
```

```
##   Defendant Victim No Yes
## 1     white  white 53 414
## 3     black  white 11  37
## 5     white  black  0  16
## 7     black  black  4 139
```

```
mod2a <- glm(cbind(No,Yes) ~ Defendant + Victim, family=binomial,
data=death)

summary(mod2a)
```

```
##
## Call:
## glm(formula = cbind(No, Yes) ~ Defendant + Victim, family = binomial,
##     data = death)
##
## Deviance Residuals:
##        1         3         5         7
##  0.02660  -0.06232  -0.60535   0.09379
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -3.5961     0.5069  -7.094 1.30e-12 ***
## Defendantwhite  -0.8678     0.3671  -2.364   0.0181 *
## Victimwhite      2.4044     0.6006   4.003 6.25e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22.26591  on 3  degrees of freedom
## Residual deviance:  0.37984  on 1  degrees of freedom
## AIC: 19.3
##
## Number of Fisher Scoring iterations: 4
```

b) (2 pts) Test goodness of fit of the model in a). What do you conclude?

```
deviance(mod2a)
```

```
## [1] 0.3798378
```

```
df.residual(mod2a)
```

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

```
1 - pchisq(deviance(mod2a),df.residual(mod2a))
```

```
## [1] 0.5376901
```

```
#Based off the P-Value is large so we can conclude to reject the null.
```

c) (2 pts) Based on the model in a), provide a likelihood confidence interval for the conditional odds ratio association between the response and defendant's race. What do you conclude?

```
exp(confint(mod2a)["Defendantwhite",])
```

```
## Waiting for profiling to be done...
```

```
##     2.5 %    97.5 %
## 0.2094370 0.8922188
```

```
#The model shows there is an association between response and race
```

d) (2 pts) Perform a Mantel-Haenszel test for the common conditional odds ratio association between the death penalty decision and defendant's race, stratified on victim's race. How does the result compare with the result you got in c) (similar conclusion, or very different)? Hint: You will need to create the appropriate cross-classified table first.

```
death.array <- xtabs(Freq ~ Defendant + Victim + DeathPenalty, data = deathpenalty)
death.array[,,1:2]
```

```
## , , DeathPenalty = no
##
##          Victim
## Defendant black white
##     black   139    37
##     white    16   414
##
## , , DeathPenalty = yes
##
##          Victim
## Defendant black white
##     black     4    11
##     white     0    53
```

```
mantelhaen.test(death.array, correct=FALSE)
```

```
##
##  Mantel-Haenszel chi-squared test without continuity correction
##
## data:  death.array
## Mantel-Haenszel X-squared = 385.34, df = 1, p-value < 2.2e-16
## alternative hypothesis: true common odds ratio is not equal to 1
## 95 percent confidence interval:
##   54.01461 186.60967
## sample estimates:
## common odds ratio
##          100.3975
```

```
#The results are are the same as part c
#there is an interaction between race in defendant and victim
```

Problem 3. (2 pts) Complete separation

Consider a simple set up where we have n_1 observations with $y = 0$ and $x = 0$ and n_2 observations with $y = 1$ and $x = 1$. We wish to fit a logistic regression model with y as the response and x as the predictor, so

$$P(Y = 1|X = x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}, \quad x = 0, 1.$$

Write down the log-likelihood $L(\alpha, \beta)$, simplifying as much as possible. Then show that for any fixed, finite value of α,

$$\frac{\partial L(\alpha, \beta)}{\partial \beta} > 0.$$

Therefore, for any finite value of α, the likelihood is maximized by letting $\beta \to \infty$.