Homework3

Rachel Holman

2023-07-03

1. For this question, we will use the "nfl.txt" data set. As a reminder, the data are on NFL team performance from the 1976 season. The variables are:

- `y` : Games won (14-game season)
- `x1` : Rushing yards (season)
- `x2` : Passing yards (season)
- `x3` : Punting average (yards/punt)
- `x4` : Field goal percentage (FGs made/FGs attempted)
- `x5` : Turnover differential (turnovers acquired minus turnovers lost)
- `x6` : Penalty yards (season)
- `x7` : Percent rushing (rushing plays/total plays)
- `x8` : Opponents' rushing yards (season)
- `x9` : Opponents' passing yards (season)

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

```
##    y   x1   x2   x3   x4 x5  x6   x7   x8   x9
## 1 10 2113 1985 38.9 64.7  4 868 59.7 2205 1917
## 2 11 2003 2855 38.8 61.3  3 615 55.0 2096 1575
## 3 11 2957 1737 40.1 60.0 14 914 65.6 1847 2175
## 4 13 2285 2905 41.6 45.3 -4 957 61.4 1903 2476
## 5 10 2971 1666 39.2 53.8 15 836 66.1 1457 1866
## 6 11 2309 2927 39.7 74.1  8 786 61.0 1848 2339
```

a. **Use the regsubsets() function from the leaps package to run all possible regressions. Set nbest=1. Identify the model (the predictors and the corresponding estimated coefficients) that is best in terms of**

i. Adjusted R^2
ii. Mallow's C_p
iii. BIC

```
allreg <- leaps::regsubsets(y ~., data=nfl, nbest=1)
#summary(allreg)
coef(allreg, which.max(summary(allreg)$adjr2))
```

```
##  (Intercept)           x2           x7           x8           x9
## -1.821703427  0.003818572  0.216894094 -0.004014887 -0.001634926
```

```
coef(allreg, which.min(summary(allreg)$cp))
```

```
##  (Intercept)           x2           x7           x8
## -1.808372059  0.003598070  0.193960210 -0.004815494
```

```
coef(allreg, which.min(summary(allreg)$bic))
```

```
##  (Intercept)           x2           x7           x8
## -1.808372059  0.003598070  0.193960210 -0.004815494
```

The regression with the highest adj R^2 is:

$$Y = -1.821703427 + 0.003818572(X2) + 0.216894094(X7) + -0.004014887(X8) + -0.001634926(X9)$$

The regression with the best Mallow's C_p is:

$$Y = -1.808372059 + 0.003598070(X2) + 0.193960210(X7) + -0.004815494(X8)$$

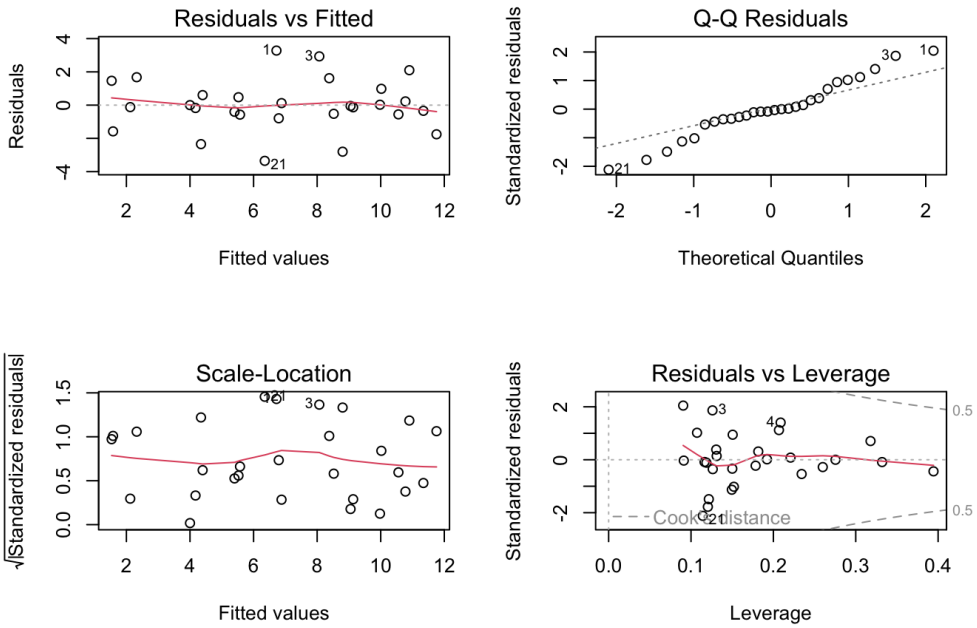
The regression with the best BIC is:

$$Y = -1.808372059 + 0.003598070(X2) + 0.193960210(X7) + -0.004815494(X8)$$

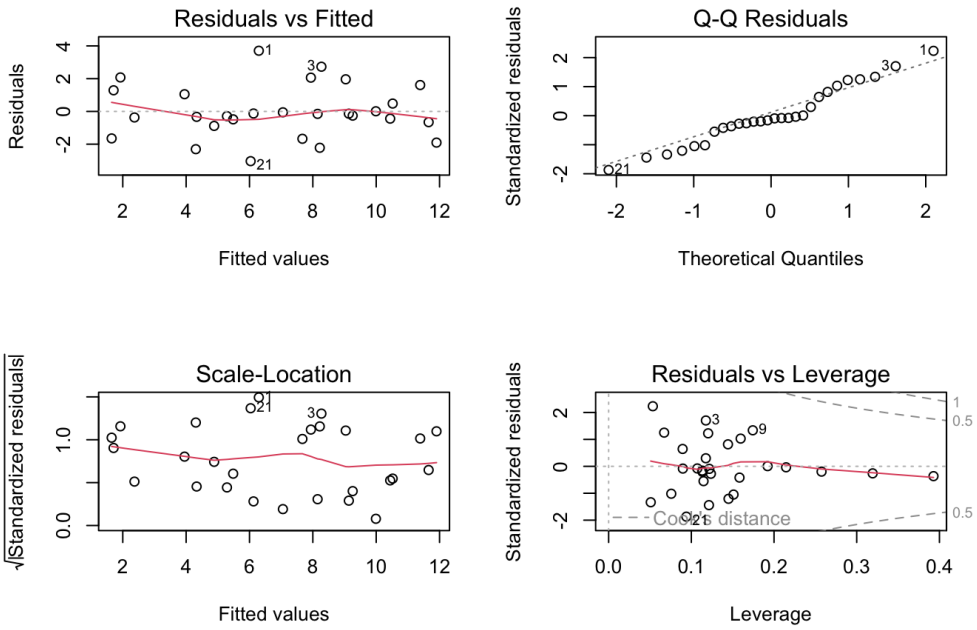
We notice that the model selected using Mallow's C_p is the same as the model selected using BIC.

b. **For the models found in part 1a, use residual plots to assess if the regression assumptions are met, and address if any variables need to be transformed. If needed, transform the appropriate variable, and re-do part 1a using the transformed variables.**

```
adj2_mod <- lm(y~x2+x7+x8+x9, data=nfl)
cp_bic_mod <- lm(y~x2+x7+x8, data=nfl)

par(mfrow = c(2, 2))
plot(adj2_mod)
```




```
par(mfrow = c(2, 2))
plot(cp_bic_mod)
```



All of the assumptions appear to be met in both the model chosen using adjusted R^2 and the model chosen using Mallow's C_p and BIC. This is clear because the residuals are scattered randomly around 0 in the residual plot with constant variance for both models. No transformations are needed.

c. **Run forward selection, starting with an intercept-only model. Report the predictors and the estimated coefficients of the model selected.**

```
##intercept only model
regnull <- lm(y~1, data=nfl)
##model with all predictors
regfull <- lm(y~., data=nfl)

##forward selection, backward elimination, and stepwise regression
step(regnull, scope=list(lower=regnull, upper=regfull), direction="forward")
```

```
## Start:  AIC=70.81
## y ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + x8    1   178.092 148.87 50.785
## + x1    1   115.068 211.90 60.669
## + x7    1    97.238 229.73 62.931
## + x5    1    86.116 240.85 64.255
## + x2    1    76.193 250.77 65.385
## + x9    1    30.167 296.80 70.104
## <none>              326.96 70.814
## + x4    1    21.844 305.12 70.878
## + x6    1    16.411 310.55 71.372
## + x3    1     2.135 324.83 72.631
##
## Step:  AIC=50.78
## y ~ x8
##
##        Df Sum of Sq     RSS    AIC
## + x2    1    64.934  83.938 36.741
## + x5    1    11.607 137.265 50.512
## <none>              148.872 50.785
## + x1    1     6.636 142.236 51.508
## + x3    1     6.368 142.504 51.561
## + x4    1     6.345 142.527 51.565
## + x7    1     0.974 147.898 52.601
## + x6    1     0.487 148.385 52.693
## + x9    1     0.008 148.864 52.783
##
## Step:  AIC=36.74
## y ~ x8 + x2
##
##        Df Sum of Sq    RSS    AIC
## + x7    1   14.0682 69.870 33.604
## + x1    1   11.1905 72.748 34.734
## + x3    1    8.9010 75.037 35.602
## + x5    1    5.8147 78.124 36.730
## <none>              83.938 36.741
## + x9    1    2.0256 81.913 38.057
## + x6    1    1.3216 82.617 38.296
## + x4    1    0.0161 83.922 38.735
##
## Step:  AIC=33.6
## y ~ x8 + x2 + x7
##
##        Df Sum of Sq    RSS    AIC
## + x9    1    4.8657 65.004 33.583
## <none>              69.870 33.604
## + x3    1    1.3873 68.483 35.043
## + x4    1    0.9792 68.891 35.209
## + x1    1    0.9022 68.968 35.240
## + x6    1    0.4879 69.382 35.408
## + x5    1    0.2987 69.571 35.484
##
## Step:  AIC=33.58
## y ~ x8 + x2 + x7 + x9
##
##        Df Sum of Sq    RSS    AIC
## <none>              65.004 33.583
## + x1    1   1.86452 63.140 34.768
## + x4    1   1.74260 63.262 34.822
## + x3    1   0.70148 64.303 35.279
## + x6    1   0.45071 64.554 35.388
## + x5    1   0.32667 64.678 35.442
```

```
##
## Call:
## lm(formula = y ~ x8 + x2 + x7 + x9, data = nfl)
##
## Coefficients:
## (Intercept)           x8           x2           x7           x9
##   -1.821703    -0.004015     0.003819     0.216894    -0.001635
```

The predictors in the model selected by the forward selection are x8, x2, x7, and x9. The estimated regression equation is:

$$Y = -1.821703 + -0.004015(X8) + 0.003819(X2) + 0.216894(X7) + -0.001635(X9)$$

d. **Run backward elimination, starting with the model with all predictors. Report the predictors and the estimated coefficients of the model selected.**

```
step(regfull, scope=list(lower=regnull, upper=regfull), direction="backward")
```

```
## Start:  AIC=41.48
## y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x5    1     0.000  60.293 39.476
## - x1    1     0.549  60.842 39.730
## - x3    1     0.746  61.039 39.821
## - x6    1     0.803  61.096 39.847
## - x4    1     1.968  62.261 40.376
## - x7    1     3.451  63.744 41.035
## <none>                60.293 41.476
## - x9    1     5.348  65.642 41.856
## - x8    1    12.072  72.365 44.587
## - x2    1    62.448 122.741 59.380
##
## Step:  AIC=39.48
## y ~ x1 + x2 + x3 + x4 + x6 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x1    1     0.553  60.846 37.732
## - x3    1     0.750  61.043 37.822
## - x6    1     0.818  61.111 37.854
## - x4    1     2.053  62.346 38.414
## - x7    1     3.859  64.152 39.213
## <none>                60.293 39.476
## - x9    1     5.351  65.644 39.857
## - x8    1    12.086  72.379 42.592
## - x2    1    66.979 127.272 58.395
##
## Step:  AIC=37.73
## y ~ x2 + x3 + x4 + x6 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x6    1     0.690  61.536 36.048
## - x3    1     1.715  62.561 36.510
## - x4    1     3.051  63.897 37.102
## <none>                60.846 37.732
## - x9    1     4.852  65.698 37.880
## - x7    1     8.961  69.807 39.579
## - x8    1    16.599  77.445 42.486
## - x2    1    67.010 127.856 56.524
##
## Step:  AIC=36.05
## y ~ x2 + x3 + x4 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x3    1     1.726  63.262 34.822
## - x4    1     2.767  64.303 35.279
## <none>                61.536 36.048
## - x9    1     4.831  66.367 36.164
## - x7    1     9.390  70.926 38.024
## - x8    1    18.314  79.851 41.343
## - x2    1    66.447 127.984 54.552
##
## Step:  AIC=34.82
## y ~ x2 + x4 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## - x4    1     1.743  65.004 33.583
## <none>                63.262 34.822
## - x9    1     5.629  68.891 35.209
## - x8    1    17.701  80.962 39.730
## - x7    1    18.583  81.845 40.033
## - x2    1    75.598 138.860 54.835
##
## Step:  AIC=33.58
## y ~ x2 + x7 + x8 + x9
##
##        Df Sum of Sq     RSS    AIC
## <none>                65.004 33.583
## - x9    1     4.866  69.870 33.604
```

```
## - x7    1    16.908  81.913 38.057
## - x8    1    23.299  88.303 40.160
## - x2    1    82.892 147.897 54.601
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8 + x9, data = nfl)
##
## Coefficients:
## (Intercept)           x2           x7           x8           x9
##   -1.821703     0.003819     0.216894    -0.004015    -0.001635
```

The predictors in the model selected by the backward elimination are x2, x7, x8, and x9. The estimated regression equation is:

$$Y = -1.821703 + 0.003819(X2) + 0.216894(X7) + -0.004015(X8) + -0.001635(X9)$$

This models chosen by the forward and backward stepwise selection proceedures are the same with the same coefficients!

e. **The PRESS statistic can be used in model validation as well as a criteria for model selection. Unfortunately, the regsubsets() function from the leaps package does not compute the PRESS statistic. The PRESS statistic can be written as**

$$\text{PRESS} = \sum_{i=1}^{n}[y_i - \hat{y}_{i(i)}]^2$$

$$= \sum_{i=1}^{n}(\frac{e_i}{1-h_{ii}})^2$$

where h_{ii} denotes the ith diagonal element from the hat matrix. Write a function that computes the PRESS statistic for a regression model.

Hint: the diagonal elements from the hat matrix can be found using the lm.influence() function.

A function that can be used to compute the PRESS statistic for a regression model is as follows:

```
PRESS <- function(model){
  sum((resid(model)/(1 - lm.influence(model)$hat))^2)
}
```

f. **Using the function you wrote in part 1e, calculate the PRESS statistic for your regression model with x2, x7, x8, x9 as predictors. Calculate the R2 Prediction for this model, and compare this value with its R2. What comments can you make about the likely predictive performance of this model?**

```
newmodel <- lm(y~x2+x7+x8+x9, data=nfl)
anova(newmodel)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x2         1  76.193  76.193 26.9589 2.900e-05 ***
## x7         1 139.501 139.501 49.3585 3.693e-07 ***
## x8         1  41.400  41.400 14.6483  0.000863 ***
## x9         1   4.866   4.866  1.7216  0.202435
## Residuals 23  65.004   2.826
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#PRESS stat
press_stat = PRESS(newmodel)
press_stat
```

```
## [1] 87.65965
```

```
#R^2 Prediction
SSt= (76.193+139.501+41.400+4.866+65.004)
1-(press_stat/SSt)
```

```
## [1] 0.7318982
```

```
#actual R^2
summary(newmodel)$r.squared
```

```
## [1] 0.8011882
```

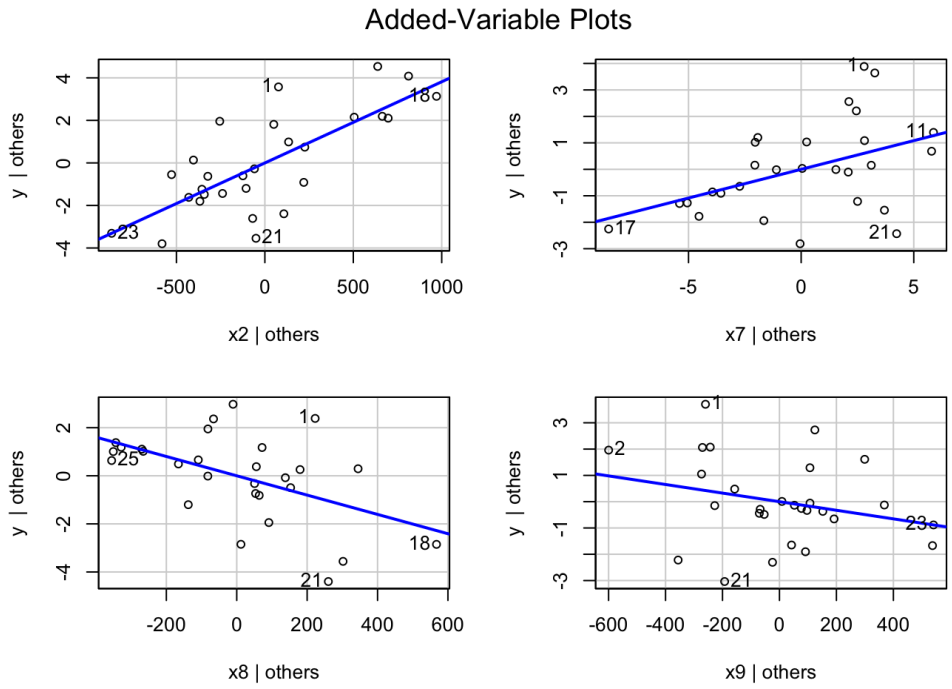
The press statistic is 87.65965.

$R^2_{prediction} = 0.7318982$ which is less than the true $R^2 = 0.8011882$. This may be an indication that there is some overfitting in this model. In other words, there may be more predictors than are needed and helpful for prediction purposes.

For the rest of the parts, we regress the number of games won against four predictors: passing yards, x2, percent rushing, x7, opponents' rushing yards in the season, x8, and opponents' passing yards in the season, x9.

g. **Create partial regression plots for this model. What are these plots telling us?**

```
car::avPlots(newmodel)
```



Based on the partial regression plots for x2, x7, x8, and x9, we see clear linear patterns, as the points are evenly scattered across the blue lines. So x2, x7, x8, and x9 do not need to be transformed.

h. **Using externally studentized residuals, do we have any outliers? What teams are these?**

```
ext.student<-rstudent(newmodel) ##ext studentized res

ext.student[abs(ext.student)>3]
```

```
## named numeric(0)
```

```
sort(abs(ext.student))
```

```
##          23           5          13          11          14          27
## 0.000315582 0.015356592 0.031051813 0.079120509 0.081865691 0.085634289
##          20           8           6          17          19          26
## 0.107708581 0.138984972 0.219873645 0.271115712 0.305893844 0.330278572
##          24          16          18          25           2          22
## 0.347870613 0.377244734 0.430064184 0.530163595 0.698462761 0.943200983
##          12          28           9           7           4          10
## 1.020566939 1.021220969 1.126646341 1.139124789 1.436316628 1.531860339
##          15           3           1          21
## 1.870882721 1.980618602 2.212417494 2.309621577
```

When using externally studentized residuals, we do not have any outliers.

i. **Do we have any high leverage data points for this multiple linear regression? What teams are these?**

```
hii<-lm.influence(newmodel)$hat ##leverages
n<-nrow(nfl)
p<-5

hii[hii>2*p/n]
```

```
##        18
## 0.3944162
```

```
sort(hii)
```

```
##          1         13         12         21         14         20         15
## 0.09050075 0.09122834 0.10729974 0.11434330 0.11641267 0.11865377 0.12046261
##         10          3         24         16          8          7         26
## 0.12162286 0.12606030 0.12639033 0.13061106 0.13111979 0.14955744 0.15013056
##         22         28          6         19          5          9          4
## 0.15060309 0.15203936 0.17844846 0.18169637 0.19226330 0.20677028 0.20879768
##         11         25         17         23          2         27         18
## 0.22086642 0.23433159 0.25998042 0.27543110 0.31796781 0.33199443 0.39441617
```

We do have one high leverage data point for this multiple linear regression. The team with the high leverage point is the 18th team in the nfl dataset.

j. **Use DF F IT Si, DF BET ASj,i, and Cook's distance to check for influential observations. What teams are influential?**

First, we identify influential observations based on DFBETAS.

```
DFBETAS<-dfbetas(newmodel)
abs(DFBETAS)>2/sqrt(n)
```

```
##    (Intercept)    x2    x7    x8    x9
## 1        FALSE FALSE FALSE  TRUE  TRUE
## 2        FALSE FALSE FALSE FALSE FALSE
## 3        FALSE FALSE FALSE FALSE FALSE
## 4        FALSE  TRUE FALSE FALSE FALSE
## 5        FALSE FALSE FALSE FALSE FALSE
## 6        FALSE FALSE FALSE FALSE FALSE
## 7        FALSE FALSE FALSE FALSE FALSE
## 8        FALSE FALSE FALSE FALSE FALSE
## 9        FALSE FALSE FALSE FALSE FALSE
## 10       FALSE FALSE FALSE  TRUE FALSE
## 11       FALSE FALSE FALSE FALSE FALSE
## 12       FALSE FALSE FALSE FALSE FALSE
## 13       FALSE FALSE FALSE FALSE FALSE
## 14       FALSE FALSE FALSE FALSE FALSE
## 15       FALSE FALSE FALSE FALSE  TRUE
## 16       FALSE FALSE FALSE FALSE FALSE
## 17       FALSE FALSE FALSE FALSE FALSE
## 18       FALSE FALSE FALSE FALSE FALSE
## 19       FALSE FALSE FALSE FALSE FALSE
## 20       FALSE FALSE FALSE FALSE FALSE
## 21        TRUE FALSE  TRUE  TRUE FALSE
## 22       FALSE FALSE FALSE FALSE FALSE
## 23       FALSE FALSE FALSE FALSE FALSE
## 24       FALSE FALSE FALSE FALSE FALSE
## 25       FALSE FALSE FALSE FALSE FALSE
## 26       FALSE FALSE FALSE FALSE FALSE
## 27       FALSE FALSE FALSE FALSE FALSE
## 28       FALSE FALSE FALSE FALSE FALSE
```

We see that: • 1 is influential for β_{X8} and β_{X9}.
• 4 is influential for β_{X2}.
• 10 is influential for β_{X8}.
• 15 is influential for β_{X9}.
• 21 is influential for β_0(intercept), β_{X7}, and β_{X8}.

```
##see actual values for DFBETAS of these teams
DFBETAS[1,]
```

```
## (Intercept)          x2          x7          x8          x9
## -0.23398333  0.07487825  0.34213343  0.42984018 -0.44666485
```

```
DFBETAS[4,]
```

```
## (Intercept)          x2          x7          x8          x9
## -0.27871327  0.43201358  0.18051900 -0.08869195  0.35808040
```

```
DFBETAS[10,]
```

```
## (Intercept)          x2          x7          x8          x9
##  0.33495274  0.04735786 -0.31870069 -0.41059884  0.02959332
```

```
DFBETAS[15,]
```

```
##  (Intercept)           x2           x7           x8           x9
## -0.173019474 -0.089826637  0.004907575 -0.019752984  0.526073599
```

```
DFBETAS[21,]
```

```
## (Intercept)          x2          x7          x8          x9
##  0.41666240  0.05177761 -0.54875067 -0.53127105  0.35037972
```

None of the magnitudes in the change of standard errors are larger than 0.54, so the removal of any of these teams does not change the model very significantly.

Next, we identify influential observations based on DFFITS:

```
DFFITS<-dffits(newmodel)
DFFITS[abs(DFFITS)>2*sqrt(p/n)]
```

```
## named numeric(0)
```

```
sort(abs(DFFITS))
```

```
##           23            5           13           14           20           11
## 0.0001945713 0.0074921789 0.0098383996 0.0297151091 0.0395200778 0.0421257802
##            8           27            6           24           26           19
## 0.0539910787 0.0603702604 0.1024736422 0.1323170900 0.1388157811 0.1441406977
##           16           17           25           18           12           22
## 0.1462197742 0.1606953858 0.2932952043 0.3470751337 0.3538245840 0.3971603198
##           28            2            7           10            9           15
## 0.4324241083 0.4769055746 0.4776973425 0.5700149750 0.5752176241 0.6923816414
##            1            4            3           21
## 0.6978980556 0.7378519822 0.7522276400 0.8298765308
```

None of the teams are flagged as influential using DFFITS because all of their DFFITS are smaller than $2\sqrt{\frac{p}{n}}$.

Lastly, let's look at Cook's distance:

```
COOKS<-cooks.distance(newmodel)
COOKS[COOKS>1]
```

```
## named numeric(0)
```

```
sort(COOKS)
```

```
##           23            5           13           14           20           11
## 7.915765e-09 1.173672e-05 2.023788e-05 1.845685e-04 3.263937e-04 3.709433e-04
##            8           27            6           24           26           19
## 6.089729e-04 7.617922e-04 2.190817e-03 3.640698e-03 4.009265e-03 4.325787e-03
##           16           17           25           18           12           22
## 4.441678e-03 5.381376e-03 1.775954e-02 2.497735e-02 2.499321e-02 3.169938e-02
##           28            7            2           10            9            1
## 3.732851e-02 4.505595e-02 4.652375e-02 6.138922e-02 6.540912e-02 8.330549e-02
##           15            3            4           21
## 8.647793e-02 1.004092e-01 1.040750e-01 1.158980e-01
```

None of the observations are flagged as influential for having a Cook's Distance value greater than 1.

Since both DFFITS and Cook's Distance do not flag any teams as being influential, none of the teams are influential enough to require removal in my opinion even though DFBETAS flagged teams 1, 4, 10, 15, and 21.

---

2. The Western Collaborative Group Study (WCGS) is one of the earliest studies regarding heart disease. Data were collected from 3154 males aged 39 to 59 in the San Francisco area in 1960. They all did not have coronary heart disease at the beginning of the study. The data set comes from the faraway package and is called wcgs. We will focus on predicting the likelihood of developing coronary heart disease based on the following predictors:

- `age` : age in years
- `sdp` : systolic blood pressure in mm Hg
- `dbp` : diastolic blood pressure in mm Hg
- `cigs` : number of cigarettes smoked per day
- `dibep` : behavior type, labeled A and B for aggressive and passive respectively.

The response variable is chd, whether the person developed coronary heart disease during annual follow ups in the study. Read the data in. We will also randomly split the data into two: half the data will be the training data set, and the remaining half will be the test data set. We will explore the training-test split in more detail in the next module. For this exercise, perform all analysis on the training data. The code below will randomly split the data into two halves.

```
library(faraway)
```

```
##
## Attaching package: 'faraway'
```

```
## The following objects are masked from 'package:car':
##
##     logit, vif
```

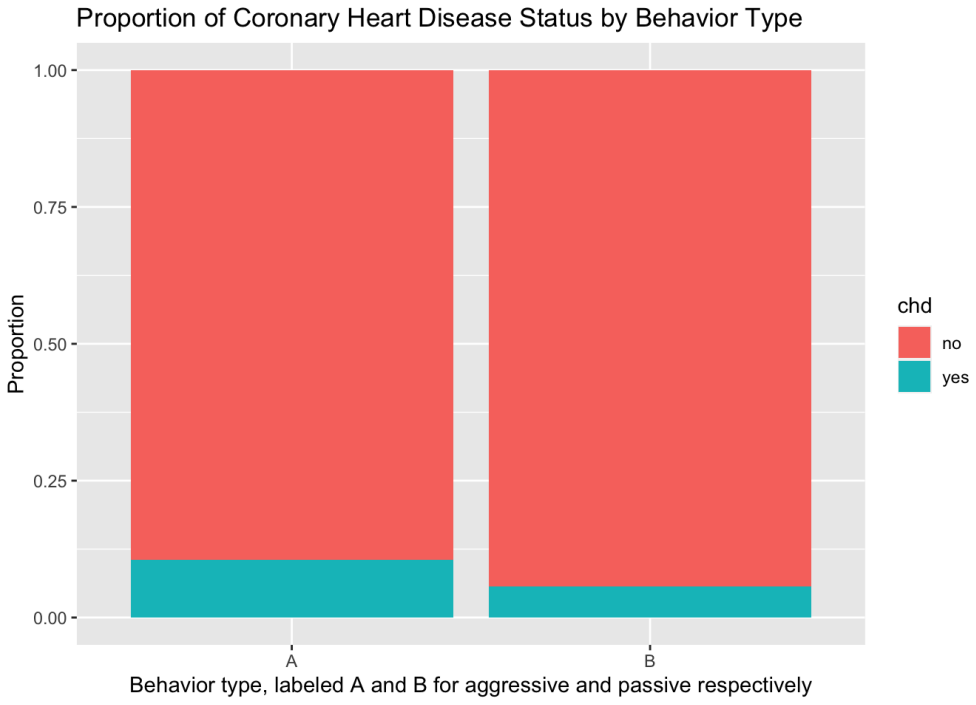
```
## The following objects are masked from 'package:survival':
##
##     rats, solder
```

```
## The following object is masked from 'package:GGally':
##
##     happy
```

```
Data<-faraway::wcgs
set.seed(6021) ##for reproducibility to get the same split
sample<-sample.int(nrow(Data), floor(.50*nrow(Data)), replace = F)
train<-Data[sample, ] ##training data frame
test<-Data[-sample, ] ##test data frame
```

a. **Before fitting a model, create some data visualizations to explore the relationship between these predictors and whether a middle-aged male develops coronary heart disease.**

```
chart1<-ggplot2::ggplot(train, aes(x=dibep, fill=chd))+
  geom_bar(position = "fill")+
  labs(x="Behavior type, labeled A and B for aggressive and passive respectively", y="Proportion",
       title="Proportion of Coronary Heart Disease Status by Behavior Type")
chart1
```



```
dp1<-ggplot2::ggplot(train,aes(x=age, color=chd))+
  geom_density()+
  labs(title="Density Plot of Age",
       subtitle= "by Coronary Heart Disease Status")

dp2<-ggplot2::ggplot(train,aes(x=sdp, color=chd))+
  geom_density()+
  labs(title="Density Plot of Systolic Blood Pressure (mm Hg)",
       subtitle= "by Coronary Heart Disease Status")

dp3<-ggplot2::ggplot(train,aes(x=dbp, color=chd))+
  geom_density()+
  labs(title="Density Plot of Diastolic Blood Pressure (mm Hg)",
       subtitle= "by Coronary Heart Disease Status")

dp4<-ggplot2::ggplot(train,aes(x=cigs, color=chd))+
  geom_density()+
  labs(title="Density Plot of Cigarettes Smoked per Day",
       subtitle= "by Coronary Heart Disease Status")

gridExtra::grid.arrange(dp1, dp2, dp3, dp4, ncol = 2, nrow = 2)
```

- Looking at the bar graph, we see that the proportion of people with coronary heart disease is nearly double in men with aggressive (A) behavior type than those with passive (B) behavior type.
- When looking at age, we see that those older than 47 have higher levels of coronary heart disease than those younger than 47.
- The higher the systolic blood pressure, the higher the levels of coronary heart disease based on the top left plot.
- Diastolic blood pressure seems to have a smaller impact on coronary heart disease because there is not a large difference between the density lines. There may be higher levels of coronary heart disease when the dbp is between 90 and 110 mm Hg.
- Finally there are higher levels of coronary heart disease for those who smoke more cigarettes per day.

Overall, we see that behavior type, smoking, and age seem to play the biggest role in whether a person develops heart disease.

b. **Use R to fit the logistic regression model using all the predictors listed above, and write the estimated logistic regression equation.**

```
chdfull <- glm(chd~age+sdp+dbp+cigs+dibep, family="binomial", data=train)
chdfull
```

```
##
## Call:  glm(formula = chd ~ age + sdp + dbp + cigs + dibep, family = "binomial",
##     data = train)
##
## Coefficients:
## (Intercept)          age          sdp          dbp         cigs       dibepB
##    -8.30877      0.06021      0.01512      0.01203      0.02137     -0.52691
##
## Degrees of Freedom: 1576 Total (i.e. Null);  1571 Residual
## Null Deviance:       893
## Residual Deviance: 837.5     AIC: 849.5
```

The estimated regression equation is:
$$\text{chd} = -8.30877 + 0.06021(\text{age}) + 0.01512(\text{sdp}) + 0.01203(\text{dbp}) + 0.02137(\text{cigs}) + -0.52691(\text{dibepB})$$

c. **Interpret the estimated coefficient for cigs in context.**

```
exp(0.02137)
```

```
## [1] 1.0216
```

The estimated odds of a middle-aged male developing coronary heart disease is multiplied by a factor of exp(0.02137) = 1.0216 for each additional cigarette smoked per day, when controlling for age, systolic blood pressure, diastolic blood pressure, and behavior type.

d. **Interpret the estimated coefficient for dibep in context.**

```
exp(-0.52691)
```

```
## [1] 0.5904266
```

The estimated odds of a middle-aged male developing coronary heart disease is exp(-0.52691) = 0.5904266 times the odds for people with behavior type A (aggressive), when controlling for age, systolic blood pressure, diastolic blood pressure, and cigarettes smoked per day.

e. **What are the estimated odds of developing heart disease for an adult male who is 45 years old, has a systolic blood pressure of 110 mm Hg, diastolic blood pressure of 70 mm Hg, does not smoke, and has type B personality? What is this person's corresponding probability of developing heart disease?**

```
newdata1<-data.frame(age=45, sdp=110, dbp=70,cigs=0,dibep="B")

##predicted log odds for test data
logodds<-predict(chdfull, newdata=newdata1)
logodds
```

```
##         1
## -3.621211
```

```
exp(logodds)
```

```
##          1
## 0.02675027
```

```
##predicted probabilities for test data
probs<-predict(chdfull, newdata=newdata1, type="response")
probs
```

```
##          1
## 0.02605333
```

The estimated odds of developing heart disease for an adult male who is 45 years old, has a systolic blood pressure of 110 mm Hg, diastolic blood pressure of 70 mm Hg, does not smoke, and has type B personality is exp(-3.621211) = 0.02675027.

This persons corresponding predicted probability of developing heart disease is 0.02605333

f. **Carry out the relevant hypothesis test to check if this logistic regression model with five predictors is useful in estimating the odds of heart disease. Clearly state the null and alternative hypotheses, test statistic, and conclusion in context.**

H_0 : $\beta_{age} = \beta_{sdp} = \beta_{dbp} = \beta_{cigs} = \beta_{dibepB} = 0$ (The full model is not useful for estimating the odds of heart disease)
H_A : At least one coefficient in H_0 is nonzero (The full model is useful for estimating the odds of heart disease)

```
#likelihood ratio test
chdnull <- glm(chd~1, family="binomial", data=train)

##test statistic
TS1<-chdnull$deviance-chdfull$deviance
TS1
```

```
## [1] 55.49501
```

```
##critical value
qchisq(1-0.05,5)
```

```
## [1] 11.0705
```

```
##pvalue
1-pchisq(TS1,5)
```

```
## [1] 1.032455e-10
```

Test Statistic, $\triangle G^2$, = 55.49501
Critical value = 11.0705
P-value = $1.032455 * 10^{-10}$

Because the test statistic is larger than the critical value, and the p-value is smaller than $\alpha = 0.05$, we reject H_0 in favor of H_A. In other words, we have enough evidence to conclude that this logistic regression model with five predictors is useful in estimating the odds of heart disease compared to the intercept-only model.

g. **Suppose a co-worker of yours suggests fitting a logistic regression model without the two blood pressure variables. Carry out the relevant hypothesis test to check if this model without the blood pressure variables should be chosen over the previous model with all five predictors.**

H_0 : $\beta_{sdp} = \beta_{dbp} = 0$ (We can drop the two blood pressure variables in the presence of the other predictors)
H_A : At least one coefficient in H_0 is nonzero (We cannot drop the two blood pressure variables in the presence of the other predictors)

```
#likelihood ratio test
chdred <- glm(chd~age+cigs+dibep, family="binomial", data=train)

##test statistic
TS2<-chdred$deviance-chdfull$deviance
TS2
```

```
## [1] 13.70587
```

```
##critical value
qchisq(1-0.05,2)
```

```
## [1] 5.991465
```

```
##pvalue
1-pchisq(TS2,2)
```

```
## [1] 0.00105635
```

Test Statistic, $\triangle G^2$, = 13.70587
Critical value = 5.991465
P-value = 0.00105635

Because the test statistic is larger than the critical value, and the p-value is smaller than $\alpha = 0.05$, we reject H_0 in favor of H_A. In other words, we have enough evidence to conclude that `sdp` and `dbp` are necessary to keep in the model in the presence of the other variables. We cannot drop both `sdp` and `dbp` as they are necessary in this model, so the logistic regression model with five predictors is preferred.

h. **Based on the Wald test, is diastolic blood pressure a significant predictor of heart disease, when the other predictors are already in the model?**

$H_0 : \beta_{dbp} = 0$ (We can drop diastolic blood pressure variables in the presence of the other predictors, it is not a significant predictor)
$H_A : \beta_{dbp} \neq 0$ (We cannot drop diastolic blood pressure variables in the presence of the other predictors, it is a significant predictor)

```
summary(chdfull)
```

```
##
## Call:
## glm(formula = chd ~ age + sdp + dbp + cigs + dibep, family = "binomial",
##     data = train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.308765   1.080141  -7.692 1.45e-14 ***
## age          0.060212   0.016604   3.626 0.000287 ***
## sdp          0.015119   0.008805   1.717 0.085950 .
## dbp          0.012026   0.014345   0.838 0.401818
## cigs         0.021366   0.006095   3.506 0.000456 ***
## dibepB      -0.526914   0.198429  -2.655 0.007921 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 893.04  on 1576  degrees of freedom
## Residual deviance: 837.55  on 1571  degrees of freedom
## AIC: 849.55
##
## Number of Fisher Scoring iterations: 5
```

```
Z = 0.012026-0/0.014345
Z
```

```
## [1] 0.012026
```

Test Statistic Z = 0.838
P-value = 0.401818

Because the p-value is greater than $\alpha = 0.05$, we fail to reject H_0. We do not have enough evidence to support keeping `dbp` in the model in the presence of the other predictors, so we can drop `dbp` from the logistic regression model.

i. **Based on all the analysis performed, which of these predictors would you use in your logistic regression model?**

We found in part 2g that we cannot remove both `sdp` and `dbp` from the model as they are significant predictors in the presence of the other variables. Then, in part 2h, we note that we should drop `dbp` because it alone is not a significant predictor in the presence of the other variables in the model. Using this information along with the visuals created in 2a, I think we should use only `age`, `sdp`, `cigs`, and `dibep` as predictors of coronary heart disease in middle-aged males.

j. **Fit a logistic regression model based on your answer in part 2i. Based on the estimated coefficients of your logistic regression, briefly comment on the relationship between the predictors and the (log) odds of developing heart disease.**

```
chdnew <- glm(chd~age+sdp+cigs+dibep, family="binomial", data=train)
summary(chdnew)
```

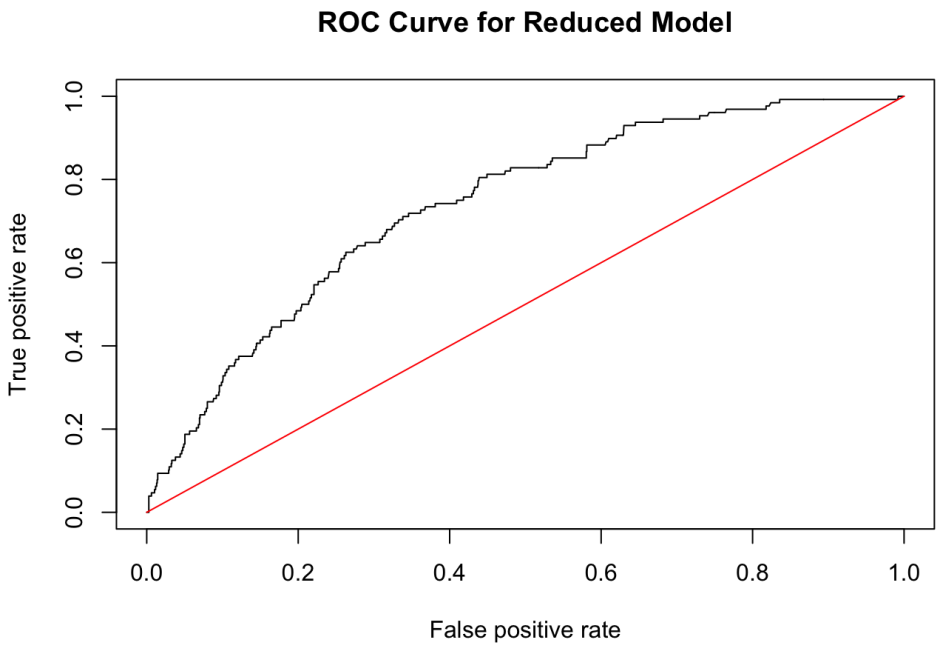
```
##
## Call:
## glm(formula = chd ~ age + sdp + cigs + dibep, family = "binomial",
##     data = train)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.065578   1.036178  -7.784 7.03e-15 ***
## age          0.060880   0.016560   3.676 0.000237 ***
## sdp          0.020757   0.005595   3.710 0.000207 ***
## cigs         0.020642   0.006035   3.421 0.000625 ***
## dibepB      -0.531792   0.198281  -2.682 0.007318 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 893.04  on 1576  degrees of freedom
## Residual deviance: 838.25  on 1572  degrees of freedom
## AIC: 848.25
##
## Number of Fisher Scoring iterations: 5
```

The estimated regression equation is: $\text{chd} = -8.065578 + 0.060880(\text{age}) + 0.020757(\text{sdp}) + 0.020642(\text{cigs}) + -0.531792(\text{dibepB})$

$\beta_{\text{age}} = 0.060880$. The estimated log odds a middle-aged male developing coronary heart disease increases by 0.060880 for each additional year of age, when controlling for the other predictors.
$\beta_{\text{sdp}} = 0.020757$. The estimated log odds a middle-aged male developing coronary heart disease increases by 0.020757 for each additional unit of systolic blood pressure in mm Hg, when controlling for the other predictors.
$\beta_{\text{cigs}} = 0.020642$. The estimated log odds a middle-aged male developing coronary heart disease increases by 0.020642 for each additional cigarette smoked per day, when controlling for the other predictors.
$\beta_{\text{dibepB}} = -0.531792$. The estimated log odds a middle-aged male developing coronary heart disease is 0.531792 lower for people with passive (B) behavior type than aggressive (A) behavior type, when controlling for the other predictors.

k. **Validate your logistic regression model using an ROC curve. What does your ROC curve tell you?**

```
##predicted probs for test data
preds<-predict(chdnew,newdata=test, type="response")

##produce the numbers associated with classification table
rates<-ROCR::prediction(preds, test$chd)

##store the true positive and false positive rates
roc_result<-ROCR::performance(rates,measure="tpr", x.measure="fpr")

plot(roc_result, main="ROC Curve for Reduced Model")
lines(x = c(0,1), y = c(0,1), col="red")
```



Because our ROC curve is above the diagonal line, we see that the logistic regression does better than randomly guessing.

l. **Find the AUC associated with your ROC curve. What does your AUC tell you?**

```
auc<-ROCR::performance(rates, measure = "auc")
auc@y.values
```

```
## [[1]]
## [1] 0.7371679
```

The AUC of our ROC curve is 0.7371679, which means our logistic regression does better than random guessing, but does not classify all observations correctly.

m. **Create a confusion matrix using a cutoff of 0.5. Report the accuracy, true positive rate (TPR), and false positive rate (FPR) at this cutoff.**

```
##confusion matrix with threshold of 0.5
table(test$chd, preds>0.5)
```

```
##
##       FALSE
##   no   1449
##   yes   128
```

Accuracy = $\frac{1449}{1577}$ = 0.9188332
True Positive Rate = $\frac{0}{128}$ = 0
False Positive Rate = $\frac{0}{1449}$ = 0

n. **Based on the confusion matrix in part 2m, a classmate says the logistic regression at this cutoff is as good as a "no information classifier". Do you agree with your classmate's statement? Briefly explain.**

Yes, I do agree with this classmate. At threshold 0.5, this model is "guessing at random" or is a "no information classifier" because no matter what the data shows, the outcome is always reported as false, or not developing coronary heart disease.

o. **Discuss if the threshold should be adjusted. Will it be better to raise or lower the threshold? Briefly explain.**

I do think that the threshold should be adjusted. It would be best to lower the threshold to correctly identify observations of people who have coronary heart disease to increase the true positive rate. The impact of misclassifying all these points is that these people would not be treated for heart disease when they need to be. We would be willing to accept the consequence of increasing the false positive rate because if someone without heart disease was told they do have it, that would not be as devastating.

The best way to determine what value to change the threshold to would be by consulting with an expert in the field.

p. **Based on your answer in part 2o, adjust the threshold accordingly, and create the corresponding confusion matrix. Report the accuracy, TPR, and FPR for this threshold.**

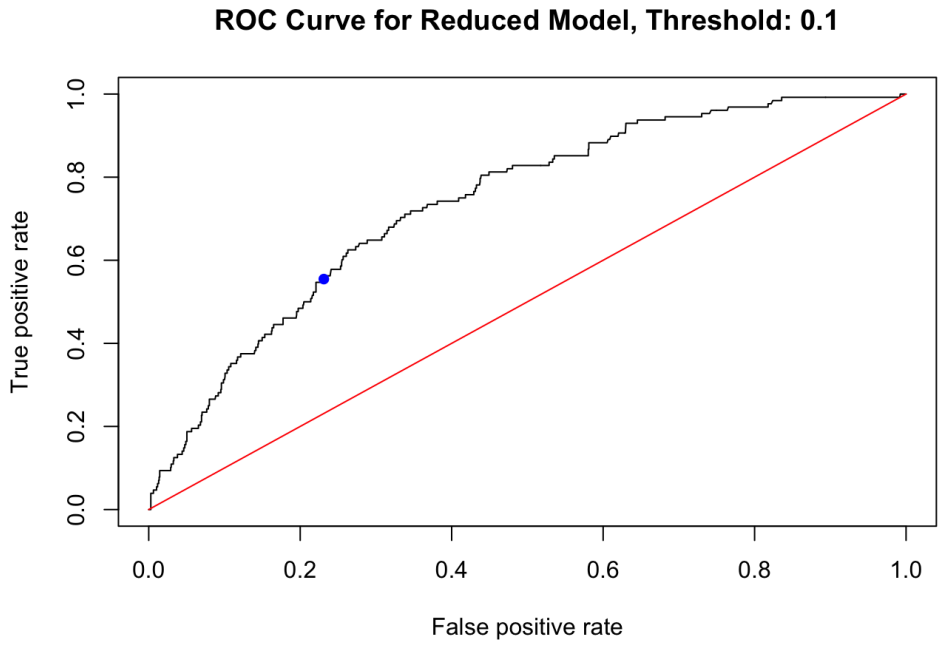
```
##confusion matrix with threshold of 0.1
table(test$chd, preds>0.1)
```

```
##
##       FALSE TRUE
##   no   1114  335
##   yes    57   71
```

```
#0.1
fpr0.1=335/1449
tpr0.1=71/128
accuracy= (1114+71)/1577
accuracy
```

```
## [1] 0.7514268
```

```
plot(roc_result, main="ROC Curve for Reduced Model, Threshold: 0.1")
lines(x = c(0,1), y = c(0,1), col="red")
points(x=fpr0.1, y=tpr0.1, col="blue", pch=16)
```



After exploring the impact of various different threshold values on the accuracy, true positive rate, and false positive rate for classifying coronary heart disease, I decided that a threshold value of 0.1 would be best.

With threshold = 0.1:
Accuracy = $\frac{1114+71}{1577}$ = 0.7514268
True Positive Rate = $\frac{71}{128}$ = 0.5546875
False Positive Rate = $\frac{57}{128}$ = 0.4453125

q. **Comment on the results from the confusions matrices in parts 2m and 2p. What do you think is happening?**

There are so many more negative observations in the data than positive, the model was being inflated toward negatives. In other words, we are modeling a rare event using an unbalanced sample with very unequal proportions between those with and without coronary heart disease. Therefore, at the usual 0.5 threshold, more predictions were classified as negative because there were so few examples of positive in the training and testing data sets.

---

3. For this question, we will revisit the penguins data set from the palmerpenguins package. The data set contains information regarding measurements of adult penguins near Palmer station, Antarctica. We will focus on using the four measurement variables (bill length, bill depth, flipper length, body mass) to model the

gender of the penguins. Since there are three species involved, we also want to control for species in the logistic regression. We will not consider the island and year in this logistic regression.

When you read the data in, notice that there are a number of penguins with missing values for gender. Remove these observations from the data frame. Also remove columns 2 and 8 since we are not considering island and year in this logistic regression.

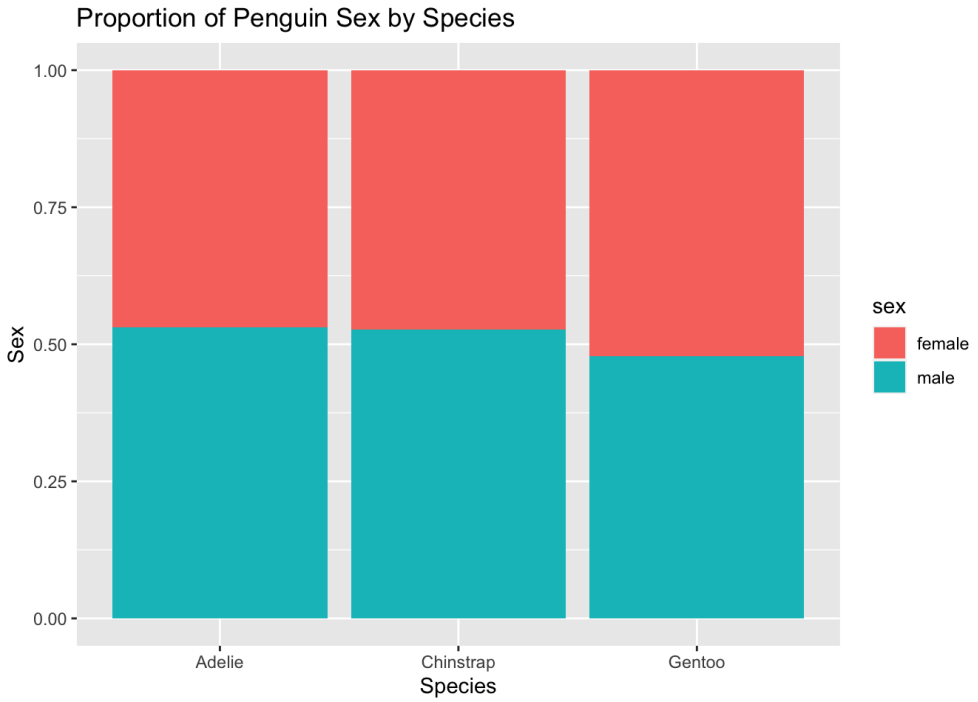
You can run the following block of code to carry out the needed steps.

```
library(palmerpenguins)
Data<-palmerpenguins::penguins
##remove penguins with gender missing
Data<-Data[complete.cases(Data[ , 7]),-c(2,8)]
##80-20 split
set.seed(1)
sample<-sample.int(nrow(Data), floor(.80*nrow(Data)), replace = F)
train<-Data[sample, ]
test<-Data[-sample, ]
head(train)
```

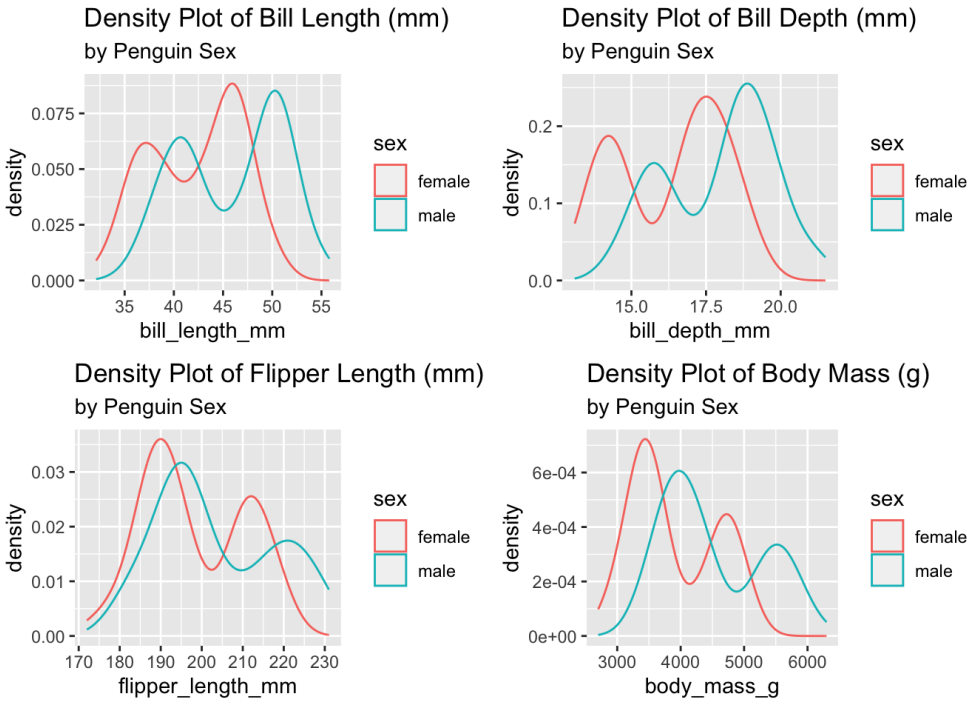
```
## # A tibble: 6 × 6
##   species   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g sex
##   <fct>              <dbl>         <dbl>             <int>       <int> <fct>
## 1 Chinstrap           50.2          18.8               202        3800 male
## 2 Gentoo              50.2          14.3               218        5700 male
## 3 Adelie              38.1          17.6               187        3425 female
## 4 Chinstrap           51            18.8               203        4100 male
## 5 Chinstrap           52.7          19.8               197        3725 male
## 6 Gentoo              49.6          16                 225        5700 male
```

a. **Create some data visualizations to explore the relationship between the various body measurements and the gender of penguins. Be sure to briefly comment on your data visualizations.**

```
chart2<-ggplot2::ggplot(train, aes(x=species, fill=sex))+
  geom_bar(position = "fill")+
  labs(x="Species", y="Sex",
       title="Proportion of Penguin Sex by Species")
chart2
```



```
dp1<-ggplot2::ggplot(train,aes(x=bill_length_mm, color=sex))+
  geom_density()+
  labs(title="Density Plot of Bill Length (mm)",
       subtitle= "by Penguin Sex")

dp2<-ggplot2::ggplot(train,aes(x=bill_depth_mm, color=sex))+
  geom_density()+
  labs(title="Density Plot of Bill Depth (mm)",
       subtitle= "by Penguin Sex")

dp3<-ggplot2::ggplot(train,aes(x=flipper_length_mm, color=sex))+
  geom_density()+
  labs(title="Density Plot of Flipper Length (mm)",
       subtitle= "by Penguin Sex")

dp4<-ggplot2::ggplot(train,aes(x=body_mass_g, color=sex))+
  geom_density()+
  labs(title="Density Plot of Body Mass (g)",
       subtitle= "by Penguin Sex")

gridExtra::grid.arrange(dp1, dp2, dp3, dp4, ncol = 2, nrow = 2)
```



There is roughly an equal proportion of male and female penguins for each species int this data set. Female penguins on average have smaller body mass and shorter bill length, bill depth, and flipper length than male penguins. Each of these quantitative predictors may be related to penguin sex. It looks like shorter bill length, shorter bill depth, shorter flipper length, and smaller body mass is associated with increased likelihood of being female.

b. **Use R to fit the logistic regression model. Based on the results of the Wald tests for the individual coefficients, which predictor(s) appears to be insignificant in the model?**

```
sexmod <- glm(sex~bill_length_mm+bill_depth_mm+flipper_length_mm+body_mass_g+species, family="binomial", data=tra
in)
summary(sexmod)
```

```
##
## Call:
## glm(formula = sex ~ bill_length_mm + bill_depth_mm + flipper_length_mm +
##     body_mass_g + species, family = "binomial", data = train)
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -94.355394  17.638204  -5.349 8.82e-08 ***
## bill_length_mm      1.025200   0.238593   4.297 1.73e-05 ***
## bill_depth_mm       2.287977   0.516595   4.429 9.47e-06 ***
## flipper_length_mm  -0.088318   0.065040  -1.358  0.17450
## body_mass_g         0.008094   0.001662   4.871 1.11e-06 ***
## speciesChinstrap  -10.608813   2.634752  -4.026 5.66e-05 ***
## speciesGentoo     -10.384568   3.565641  -2.912  0.00359 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 368.619  on 265  degrees of freedom
## Residual deviance:  68.297  on 259  degrees of freedom
## AIC: 82.297
##
## Number of Fisher Scoring iterations: 8
```

Looking at the p-values for each individual coefficient, we notice that flipper length has a p-value larger than $\alpha = 0.05$, which means that is is not a significant predictor in this model according to Walds test. All the other variables are significant. Because of this outcome, we can drop `flipper_length_mm` from our model in the presence of the other variables.

c. **Based on your answer in part 3b, drop the predictor(s) and refit the logistic regression. Write out the estimated logistic regression equation. If you did not drop any predictor, write out the logistic regression equation from part 3b.**

```
sexmod2 <- glm(sex~bill_length_mm+bill_depth_mm+body_mass_g+species, family="binomial", data=train)
sexmod2
```

```
##
## Call:  glm(formula = sex ~ bill_length_mm + bill_depth_mm + body_mass_g +
##     species, family = "binomial", data = train)
##
## Coefficients:
##      (Intercept)    bill_length_mm     bill_depth_mm       body_mass_g
##       -1.032e+02         9.513e-01         2.099e+00         7.714e-03
## speciesChinstrap     speciesGentoo
##       -1.042e+01        -1.238e+01
##
## Degrees of Freedom: 265 Total (i.e. Null);  260 Residual
## Null Deviance:       368.6
## Residual Deviance: 70.17     AIC: 82.17
```

After dropping the `flipper_length_mm` predictor, the estimated logistic regression equation is:

$$\text{sex} = -103.2 + 0.9513(\text{bill\_length\_mm}) + 2.099(\text{bill\_depth\_mm}) + 0.007714(\text{body\_mass\_g}) + -10.42(\text{speciesChinstrap}) + -12.38(\text{speciesGe}$$

d. **Based on your estimated logistic regression equation in part 3c, how would you generalize the relationship between some of the body measurement predictors and the (log) odds of a penguin being male?**

```
contrasts(train$sex)
```

```
##        male
## female    0
## male      1
```

The estimated log odds of a penguin being male increases by 0.9513 for each additional millimeter of bill length, when controlling for the other variables.
The estimated log odds of a penguin being male increases by 2.099 for each additional millimeter of bill depth, when controlling for the other variables.
The estimated log odds of a penguin being male increases by 0.007714 for each additional gram of body mass, when controlling for the other variables.

In general, when controlling for species, the log odds of a penguin being male increases with larger body measurements when taking all variables into account.

e. **Based on your estimated logistic regression equation in part 3c, interpret the estimated coefficient for bill length contextually.**

```
exp(0.9513)
```

```
## [1] 2.589073
```

The estimated log odds of a penguin being male increases by 0.9513 for each additional millimeter of bill length, when controlling for the other variables. In other words, the estimated odds of a penguin being male is multiplied by exp(0.9513) = 2.589073 for each additional millimeter of bill length, when controlling for the other variables.

f. **Consider a Gentoo penguin with bill length of 49mm, bill depth of 15mm, flipper length of 220mm, and body mass of 5700g. Based on your estimated logistic regression equation in part 3c, what are the log odds, odds, and probability that this penguin is male?**

```
newdata2<-data.frame(bill_length_mm=49, bill_depth_mm=15, flipper_length_mm=220,body_mass_g=5700, species="Gento
o")

##predicted log odds for test data
logodds<-predict(sexmod2, newdata=newdata2)
logodds
```

```
##        1
## 6.462668
```

```
exp(logodds)
```

```
##        1
## 640.7683
```

```
##predicted probabilities for test data
probs<-predict(sexmod2, newdata=newdata2, type="response")
probs
```

```
##         1
## 0.9984418
```

The log odds that a a Gentoo penguin with bill length of 49mm, bill depth of 15mm, flipper length of 220mm, and body mass of 5700g is male is 6.462668. In other words, the estimated odds that this penguin is make is exp(6.462668) = 640.7683. This penguins corresponding predicted probability of being male is 0.9984418.

g. **Conduct a relevant hypothesis test to assess if the logistic regression in part 3c is a useful model. Be sure to write out the null and alternative hypotheses, report the value of the test statistic, and write a relevant conclusion.**

H_0 : $\beta_{\text{bill\_length\_mm}} = \beta_{\text{bill\_depth\_mm}} = \beta_{\text{body\_mass\_g}} = \beta_{\text{speciesChinstrap}} = \beta_{\text{speciesGentoo}} = 0$ (The full model is not useful for estimating the sex of a penguin)
H_A : At least one coefficient in H_0 is nonzero (The full model is useful for estimating the sex of a penguin)

```
#likelihood ratio test
sexnull <- glm(sex~1, family="binomial", data=train)

##test statistic
TS1<-sexnull$deviance-sexmod2$deviance
TS1
```

```
## [1] 298.4472
```

```
##critical value
qchisq(1-0.05,4)
```

```
## [1] 9.487729
```

```
##pvalue
1-pchisq(TS1,4)
```

```
## [1] 0
```

Test Statistic, $\triangle G^2$, = 298.4472
Critical value = 9.487729
P-value = 0

Because the test statistic is larger than the critical value, and the p-value is smaller than $\alpha = 0.05$, we reject H_0 in favor of H_A. In other words, we have enough evidence to conclude that this logistic regression model with the body weight, bill length, bill depth, and species predictors is useful in estimating the sex of a penguin compared to the intercept-only model.

h. **Validate your model from part 3c on the test data by creating an ROC curve. What does your ROC curve tell you?**

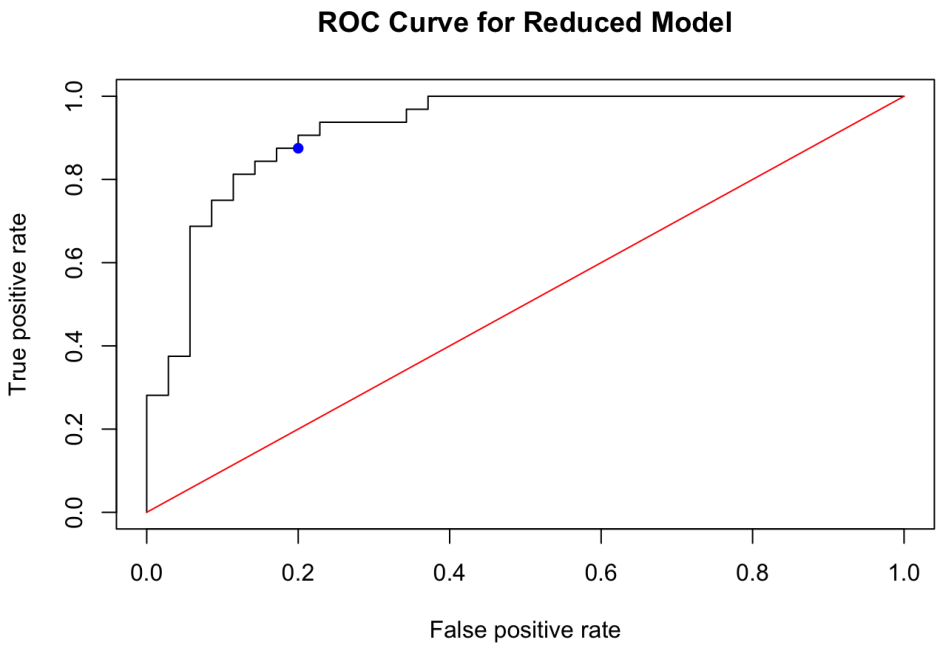
```
##predicted probs for test data
preds<-predict(sexmod2,newdata=test, type="response")

##produce the numbers associated with classification table
rates<-ROCR::prediction(preds, test$sex)

##store the true positive and false positive rates
roc_result<-ROCR::performance(rates,measure="tpr", x.measure="fpr")

table(test$sex, preds>0.5)
```

```
##
##          FALSE TRUE
##   female    28    7
##   male       4   28
```

```
fpr0.5=7/35
tpr0.5=28/32
accuracy= (28+28)/67
accuracy
```

```
## [1] 0.8358209
```

```
plot(roc_result, main="ROC Curve for Reduced Model")
lines(x = c(0,1), y = c(0,1), col="red")
points(x=fpr0.5, y=tpr0.5, col="blue", pch=16)
```



This ROC curve is very far above the diagonal, so it does much better than random guessing.

i. **Find the AUC associated with your ROC curve. What does your AUC tell you?**

```
auc<-ROCR::performance(rates, measure = "auc")
auc@y.values
```

```
## [[1]]
## [1] 0.9214286
```

The AUC of our ROC curve is 0.9214286, which means our logistic regression does better than random guessing. Since the AUC is very close to 1, our logistic regression is very close to classifying all observations correctly.

j. **Create a confusion matrix using a threshold of 0.5. What is the false positive rate? What is the false negative rate? What is error rate?**

```
##confusion matrix with threshold of 0.5
table(test$sex, preds>0.5)
```

```
##
##          FALSE TRUE
##   female    28    7
##   male       4   28
```

FPR = $\frac{7}{28+7} = \frac{7}{35} = 0.2$
FNR = $\frac{4}{28+4} = \frac{4}{32} = 0.125$
Error Rate= $\frac{7+4}{28+7+4+28} = \frac{11}{67} = 0.1641791$
Accuracy = $\frac{28+28}{28+7+4+28} = \frac{56}{67} = 0.8358209$

k. **Discuss if the threshold should be changed. If it should be changed, explain why, and create another confusion matrix with a different threshold.**

I do not think the threshold should be changed. When using a threshold of 0.5, the false negative and false positive rates are both very low. The error rate is very small and there is a fairly high accuracy, so there is nothing major to change. Additionally, there is no advantage to predicting females with more accuracy than males or vise versa so one should not be improved at the expense of the other.

---

4. (You may only use R as a simple calculator or to find p-values or critical values) The data for this question are 36 monthly observations on variables affecting sales of a product. The objective is to determine an efficient model for predicting and explaining market share sales, Share, which is the average monthly market share for the product, in percent. The predictors are average monthly price in dollars, price, amount of advertising exposure based on gross Nielson rating, nielsen, whether a discount price was in effect, discount

(1 if discount, 0 otherwise), whether a package promotion was in effect, promo (1 if promotion, 0 otherwise), and time in months, time.

a. **The output below is obtained after using the step() function using forward selection, starting with a model with just the intercept term. What predictors are selected based on forward selection?**

```
> start<-lm(Share~1, data=data)
> end<-lm(Share~.,data=data)
> result.f<-step(start, scope=list(lower=start,
+ upper=end), direction="forward")

Start: AIC=-94.8
Share ~ 1
           Df  Sum of Sq  RSS     AIC
+ discount  1     1.52953 0.91672 -128.137
+ promo     1     0.22756 2.21870 -96.318
<none>                    2.44626 -94.803
+ price     1     0.08693 2.35933 -94.105
+ nielsen   1     0.01288 2.43337 -92.993
+ time      1     0.00469 2.44156 -92.872

Step: AIC=-128.14
Share ~ discount
           Df    Sum of Sq  RSS     AIC
+ promo     1     0.086097 0.83063 -129.69
+ price     1     0.080864 0.83586 -129.46
+ time      1     0.058506 0.85822 -128.51
<none>                     0.91672 -128.14
+ nielsen   1     0.041559 0.87516 -127.81

Step: AIC=-129.69
Share ~ discount + promo
           Df  Sum of Sq  RSS     AIC
+ price     1    0.112673 0.71795 -132.94
+ time      1    0.075200 0.75543 -131.10
<none>                    0.83063 -129.69
+ nielsen   1    0.025277 0.80535 -128.80

Step: AIC=-132.94
Share ~ discount + promo + price
           Df  Sum of Sq  RSS     AIC
<none>                    0.71795 -132.94
+ time      1   0.0110210 0.70693 -131.49
+ nielsen   1   0.0003132 0.71764 -130.95
```

Based on forward selection `discount`, `promo`, and `price` are selected as predictors for predicting `share`.

b. **Your client asks you to explain what each step in the output shown above means. Explain the forward selection procedure to your client, for this output.**

When trying to predict `share`, the average monthly market share for the product, this forward selection procedure first creates a linear regression model that does not use any predictors (the null model) to predict `share`. It then evaluates all the variables in the dataset to determine which one would improve the model the most if added as a predictor. The most helpful variable is added into the model and then the process is repeated, adding the variable that most improves the prediction model one-by-one until the addition of predictors no longer helps make the model better. In the output above we see that in the first step, `discount` makes the largest positive impact on the model so it is added. Next, when looking at the model of `share` predicted by `discount` we note that the addition of `promo` would further improve the model so this is done. Next, `price` further improves the model so it is added, but the last two variables we have available in our dataset would make the model worse so we end with the final model of `share` being predicted by `discount`, `promo`, and `price`.

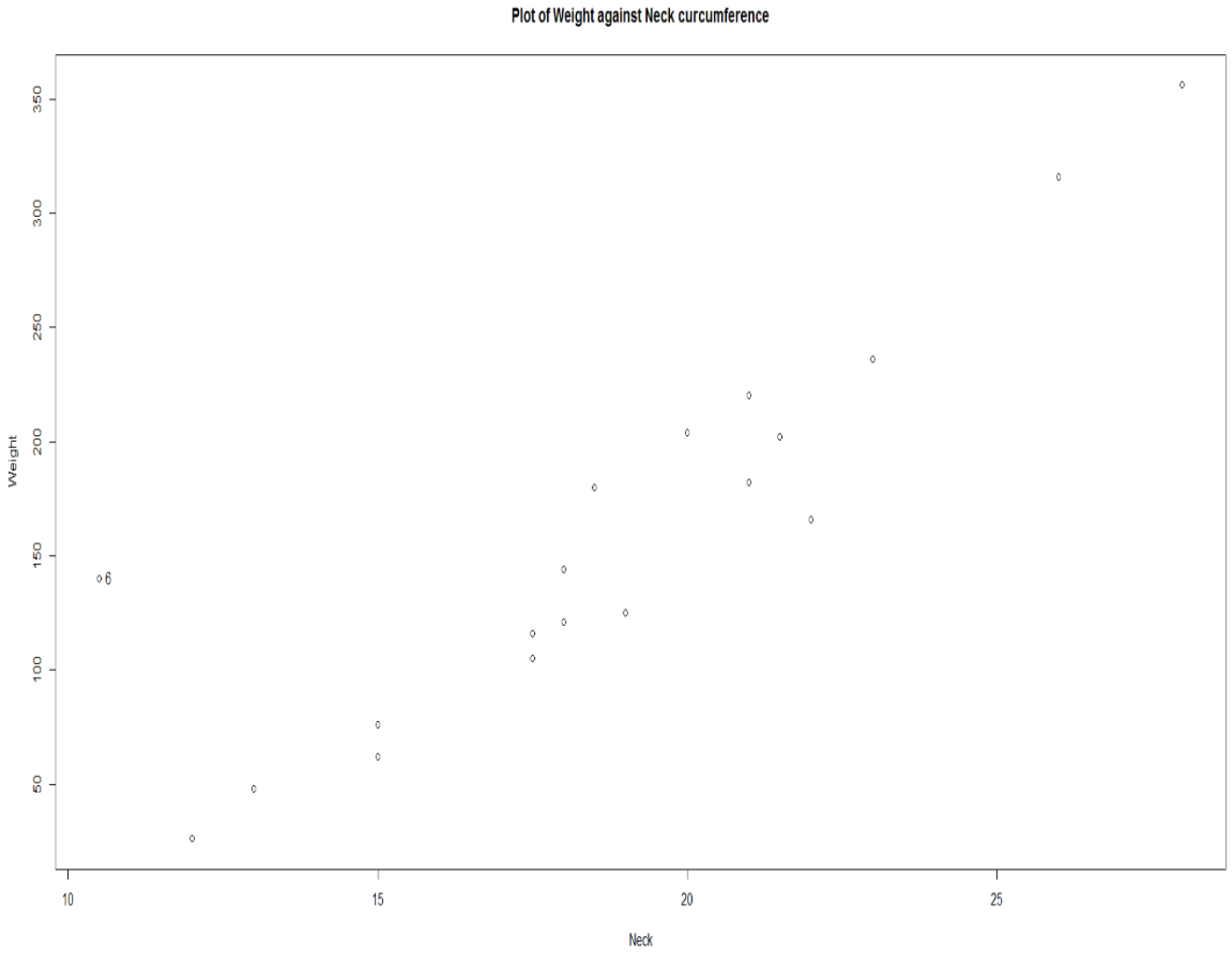
c. **Your client asks if he should go ahead and use the models selected in part 4a. What advice do you have for your client?**

While the model selected in part 4a is a useful model, it would be a good idea to run other selection methods to determine if the model generated is the best one. A backward selection method, which removes variables one-by-one until the best model is left, would be useful to look into as well as a method which works by both adding and subtracting variables to create the bet model. Additionally, while these selection methods are useful for quickly determining some of the best models with the fewest necessary predictors, you should always perform an analysis on the models to assess if the regression assumptions are met and determine if any transformations are necessary before using it further.

---

5. (You may only use R as a simple calculator or to find p-values or critical values)

Data from n = 19 bears of varying ages are used to develop an equation for estimating Weight from Neck circumference. From a visual inspection of the scatterplot, it appears observation 6 may be an outlier.



The output below comes from fitting the linear regression model on the data.

```
##with all 19 bears
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -158.78 40.46 -3.924 0.00109 **
Neck 16.95 2.10 8.071 3.24e-07 ***
Residual standard error: 40.13 on 17 degrees of freedom
Multiple R-squared: 0.793, Adjusted R-squared: 0.7809
F-statistic: 65.14 on 1 and 17 DF, p-value: 3.235e-07
```

The output below comes from fitting the linear regression model on the data, with the outlier removed.

```
##with outlier removed, so 18 bears
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -234.60 25.93 -9.049 1.08e-07 ***
Neck 20.54 1.32 15.562 4.39e-11 ***
Residual standard error: 22.6 on 16 degrees of freedom
Multiple R-squared: 0.938, Adjusted R-squared: 0.9342
F-statistic: 242.2 on 1 and 16 DF, p-value: 4.394e-11
```

The output below displays the values of the predictor and response for the 6th observation.

```
> data[6,]
Neck Weight
6 10.5 140
```

Some additional information from R, regarding ordinary residuals, ei, and leverages, hii shown below, from the full data.

```
> result$residuals ##ordinary residuals
         1          2          3          4          5          6          7
-25.276933 -48.066801 22.880666 23.828133 -2.276933 120.829070 -32.803200
         8          9         10         11         12         13         14
-18.592131 -38.224400 25.249333 -21.803200 -15.119334 40.248397 34.143331
        15         16         17         18         19
-3.593068 -33.434532 4.985732 -19.434532 -13.539598

> tmp$hat ##leverages
         1          2          3          4          5          6          7
0.05422642 0.08132161 0.06633278 0.05682064 0.05422642 0.23960510 0.05700079
         8          9         10         11         12         13         14
0.17788427 0.05278518 0.05282121 0.05700079 0.06633278 0.28626504 0.19604381
        15         16         17         18         19
0.07314261 0.09141025 0.10178713 0.09141025 0.14358291
```

a. **Calculate the externally studentized residual, ti, for observation 6. Will this be considered outlying?**

```
n = 19
p = 2
ei = 120.829070
hii = 0.23960510
MSresi = (22.6)^2
t_i = ei / sqrt(MSresi*(1-hii))
t_i
```

```
## [1] 6.131171
```

The externally studentized residual, ti, for observation 6 is 6.131171. Since this value is greater than 3, this observation is flagged as an outlier.

b. **What is the leverage for observation 6? Based on the criterion that leverages greater than 2p/n are considered outlying in the predictor(s), is this observation high leverage?**

```
hii
```

```
## [1] 0.2396051
```

```
2*p/n
```

```
## [1] 0.2105263
```

Observation 6 is high leverage based on the criterion that leverages greater than 2p/n are considered outlying in the predictor(s) because 2p/n = 0.2105263 which is smaller than the hii of 0.23960510.

c. **Calculate the DFFITS for observation 6.**

```
dffits = t_i*((hii/(1-hii))^(1/2))
dffits
```

```
## [1] 3.441691
```

```
2*sqrt(p/n)
```

```
## [1] 0.6488857
```

```
# dffits > 2*sqrt(p/n) is influential
```

The DFFITS for observation 6 is 3.441691. Because the magnitude of this value is greater than $2\sqrt{\frac{p}{n}}$ (0.6488857), observation 6 is definitely flagged as a n influential observation.

d. **Calculate Cook's distance for observation 6.**

```
MSres = 40.13^2
r_i = ei / sqrt(MSres*(1-hii))
cooks_d = (r_i^2 / p)*(hii/(1-hii))
cooks_d
```

```
## [1] 1.878418
```

Any observation with a Cook's Distance value larger than 1 is flagged as influential. Observation 6 has a Cook's Distance of 1.878418 which exceeds that criteria so it is definitely influential.

e. **Would you say that observation 6 is influential, based on DFFITS and Cook'sdistance?**

Yes, I would say that observation 6 is influential, based on DFFITS and Cook's distance. This is because the DFFITS for observation 6 is greater than $2\sqrt{\frac{p}{n}}$ (0.6488857) and the Cook's distance value is larger than 1. This indicates that observation 6 should be flagged as an influential point.

f. **Briefly describe the difference in what DFFITS and Cook's distance are measuring.**

DFFITS can be interpreted as the number of standard errors $\hat{y}$changes if observation 6 is removed when estimating the model. The numerator of Cook's distance measures the Squared Euclidean distance between the vector of fitted values with all observations, $\hat{y}$, and the vector of fitted values with observation 6 removed, $\hat{y_6}$. So it measures how the fitted values for all observations change, if observation 6 is removed. This is unlike DFFITS, which only measures the change in fitted value for observation 6.