Biostats 546 HW 3

Ronald Buie

February 10, 2019

Due Via Online Submission to Canvas: Sunday, February 24 at 12 PM (Noon)

1. In this problem, you will make use of a (real) dataset of your choice in order to predict a binary response Y using predictors X1; : : : ;Xp. You should have p 5.

```
library(ISLR)

AutoData <- ISLR::Auto

AutoData$yearBinary <- "older"
AutoData[AutoData$year > 76,]$yearBinary <- "newer"
AutoData$yearBinary <- as.factor(AutoData$yearBinary)
AutoData$origin <- as.factor(AutoData$origin)
AutoData$cylinders <- as.factor(AutoData$cylinders)
AutoData$year <- NULL
AutoData$name <- NULL
```

(a) Describe the dataset. What is the response, and what are the predictors?

```
head(AutoData)
```

```
##   mpg cylinders displacement horsepower weight acceleration origin
## 1  18         8          307        130   3504         12.0      1
## 2  15         8          350        165   3693         11.5      1
## 3  18         8          318        150   3436         11.0      1
## 4  16         8          304        150   3433         12.0      1
## 5  17         8          302        140   3449         10.5      1
## 6  15         8          429        198   4341         10.0      1
##   yearBinary
## 1      older
## 2      older
## 3      older
## 4      older
## 5      older
## 6      older
```

```
summary(AutoData)
```

```
##       mpg        cylinders  displacement     horsepower        weight    
##  Min.   : 9.00   3:  4     Min.   : 68.0   Min.   : 46.0   Min.   :1613  
##  1st Qu.:17.00   4:199     1st Qu.:105.0   1st Qu.: 75.0   1st Qu.:2225  
##  Median :22.75   5:  3     Median :151.0   Median : 93.5   Median :2804  
##  Mean   :23.45   6: 83     Mean   :194.4   Mean   :104.5   Mean   :2978  
##  3rd Qu.:29.00   8:103     3rd Qu.:275.8   3rd Qu.:126.0   3rd Qu.:3615  
##  Max.   :46.60             Max.   :455.0   Max.   :230.0   Max.   :5140
##   acceleration   origin  yearBinary
##  Min.   : 8.00   1:245   newer:178
##  1st Qu.:13.78   2: 68   older:214
##  Median :15.50   3: 79
##  Mean   :15.54
##  3rd Qu.:17.02
##  Max.   :24.80
```

The Auto data set contains information about the performance, origin and model of 392 vehicles. We have created a binary indicator variable for the year where “older” is 70 to 76, and “newer” is > 76 and will use this as our outcome. our predictors are mpg, cylinders, displacement, horsepower, weight, acceleration, and origin.

(b) Fit a logistic regression model to predict Y using X1; : : : ;Xp. What is the classification error (i.e. the fraction of incorrectly classified observations) on the training set?

```
LogFitAutoData <- glm(yearBinary ~ ., data = AutoData, family="binomial")

LogFitAutoDataPredictions <- predict(LogFitAutoData, type="response")

LogPredictions <- rep("newer",392)
LogPredictions[LogFitAutoDataPredictions >.5]<- "older"

LogPredictionsErrorRate <- 1-mean(LogPredictions==AutoData$yearBinary)
```

The training error rate is 21%

(c) Use the validation set approach in order to estimate the test classification error. Report the error you obtain.

```
set.seed(1000)

trainSample <- sample(392, 196)

LogFitAutoDataVS <- glm(yearBinary ~ ., data = AutoData, family="binomial", subset = trainSample)

LogFitAutoDataVSPredictions <- predict(LogFitAutoDataVS, newdata = AutoData, type="response")

LogPredictionsVS <- rep("newer",392)
LogPredictionsVS[LogFitAutoDataVSPredictions >.5]<- "older"

LogPredictionsVSErrorRate <- 1-mean(LogPredictionsVS==AutoData$yearBinary)[-trainSample]
```

The training error rate for the Validation set approach is 21%

(d) Use the leave-one-out cross-validation approach in order to estimate the test classification error. Report the error you obtain.

```
library(boot)

LogFitAutoDataCVError <- cv.glm(AutoData, LogFitAutoData)
```

```
LogFitAutoDataCVError$delta
```

```
## [1] 0.1382103 0.1381981
```

The cross validation approach yiels a MSE of 0.1382103 and bias corrected MSE of 0.1381981

(e) Use the 5-fold cross-validation approach in order to estimate the test classification error. Report the error you obtain.

```
library(boot)
set.seed(1000)
attach(AutoData)


LogFitAutoDataCVK5Error <- cv.glm(AutoData, LogFitAutoData, K=5)

LogFitAutoDataCVK5Error$delta[1]
```

```
## [1] 0.1393391
```

The cross validation approach yiels an MSE of 0.1382103 and bias corrected MSE of 0.1381981

(f) Comment on your findings in (b)-(e).

2. In this problem, you will make use of a (real) dataset of your choice in order to predict a continuous response Y using predictors X1; : : : ;X6. You need to choose a dataset with at least 6 predictors. If your dataset has more than 6 predictors, then please choose 6 of them now. In other words, you should have p = 6.

```
AutoData6 <- AutoData[,!(names(AutoData) %in% "origin")]
AutoData6$cylinders <- as.numeric(as.character(AutoData6$cylinders))
```

(a) Describe the dataset. What is the response, and what are the predictors? 1

```
head(AutoData6)
```

```
##   mpg cylinders displacement horsepower weight acceleration yearBinary
## 1  18         8          307        130   3504         12.0      older
## 2  15         8          350        165   3693         11.5      older
## 3  18         8          318        150   3436         11.0      older
## 4  16         8          304        150   3433         12.0      older
## 5  17         8          302        140   3449         10.5      older
## 6  15         8          429        198   4341         10.0      older
```

```
summary(AutoData6)
```

```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0
##      weight      acceleration   yearBinary
##  Min.   :1613   Min.   : 8.00   newer:178
##  1st Qu.:2225   1st Qu.:13.78   older:214
##  Median :2804   Median :15.50
##  Mean   :2978   Mean   :15.54
##  3rd Qu.:3615   3rd Qu.:17.02
##  Max.   :5140   Max.   :24.80
```

For this question we chose the same data set as above, but only included mpg, cylinders, displacement, horsepower, weight, and acceleration as predictors, and our binary year as our outcome. Cylindars, previously a dummy variable, has been converted to a continuous integer to meet the p=6 requirememt

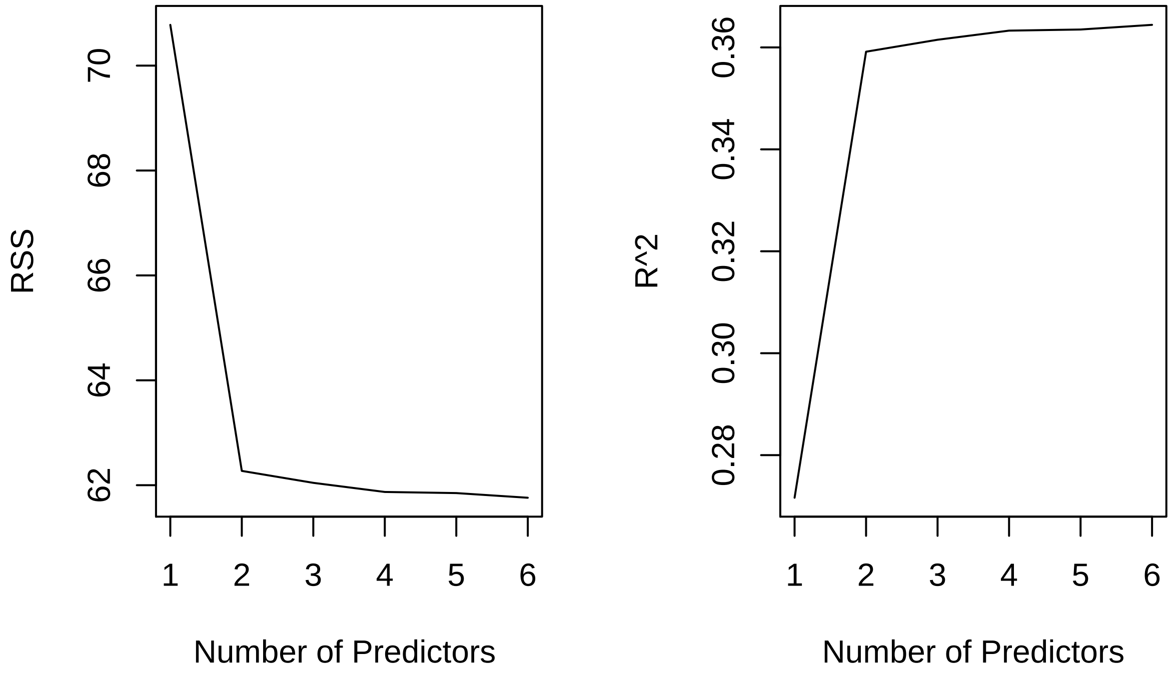
(b) Fit a least squares linear model using every possible subset of the features. How many models did you fit?

```r
library(leaps)
set.seed(1000)
a6Subsets <- regsubsets(yearBinary~., AutoData6)
a6SubsetsSummary <- summary(a6Subsets)
```

63 models are possible.

(c) Re-create Figure 6.1 in the textbook using your data. The left-hand panel should display (training set) RSS on the y-axis, and the right-hand panel should display the R2 on the y-axis. Both panels should display the number of predictors on the x-axis.

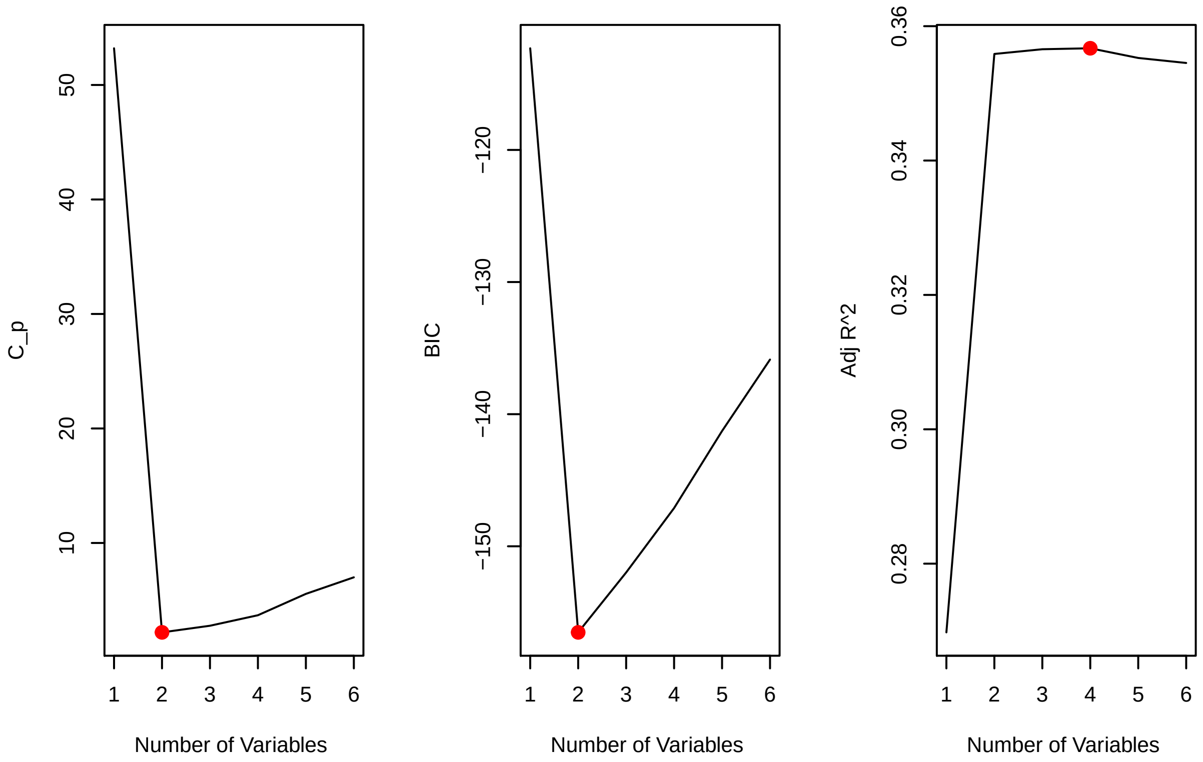
```r
par(mfrow=c(1,2))
plot(a6SubsetsSummary$rss ,xlab="Number of Predictors ",ylab="RSS",type="l")
plot(a6SubsetsSummary$rsq ,xlab="Number of Predictors", ylab="R^2",type="l")
```



(d) Report the predictors in each of the models M0;M1; : : : ;Mp.

(e) Re-create Figure 6.2 in the textbook using your data. The y-axes for the three panels should be Cp, BIC, and adjusted R2; all x-axes should display the number of predictors.

```
par(mfrow=c(1,3))
plot(a6SubsetsSummary$cp ,xlab="Number of Variables ",ylab="C_p", type='l')
points(which.min(a6SubsetsSummary$cp),min(a6SubsetsSummary$cp),col="red",cex=2,pch=20)
plot(a6SubsetsSummary$bic ,xlab="Number of Variables ",ylab="BIC", type='l')
points(which.min(a6SubsetsSummary$bic),min(a6SubsetsSummary$bic),col="red",cex=2,pch=20)
plot(a6SubsetsSummary$adjr2 ,xlab="Number of Variables ",ylab="Adj R^2", type='l')
points(which.max(a6SubsetsSummary$adjr2),max(a6SubsetsSummary$adjr2),col="red",cex=2,pch=20)
```



(f) Based on your results, what is the best" least squares linear model on this dataset? (Your answer should include not only the predictors, but also the coecient estimates.) Explain your answer.

```
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(round(coef(a6Subsets, 2), 6), no.space=TRUE, header=FALSE, float= FALSE, ci = TRUE, table.pla
```

(Intercept)	mpg	weight
3.924	-0.062	-0.0003

Our least biased model contains 2 variables, mpg and weight. Similarly, our best performing models using C_p and BIC is the same. However, performance according to adjusted R^2 is the 4 variable model using mpg, cylinders, displacement, and weight.

3. Consider using the Auto data set to predict mpg using polynomial functions of horsepower in a least squares linear regression.

(a) Perform the validation set approach, and produce a plot like the one in the right-hand panel of Figure 5.2 of the textbook. Your answer won't look exactly the same as the results in Figure 5.2, since you'll be starting with a different random seed. Discuss your findings. What degree polynomial is best, and why?

(b) Perform leave-one-out cross-validation, and produce a plot like the one in the left-hand panel of Figure 5.4 of the textbook. Discuss your findings. What degree polynomial is best, and why?

(c) Perform 10-fold cross-validation, and produce a plot like the one in the right-hand panel of Figure 5.4 of the textbook. Discuss your findings. What degree polynomial is best, and why?

(d) Fit a least squares linear model to predict mpg using polynomials of degrees from 1 to 10, using all available observations. Make a plot showing Degree of Polynomial" on the x-axis, and Training Set Mean Squared Error" on the y-axis. Discuss your findings.

(e) Fit a least squares linear model to predict mpg using a degree-10 poly- nomial, using all available observations. Using the summary command in R, examine the output. Comment on the output, and discuss how this relates to your findings in (a)(d).

4. We will now continue with the Auto data set. Note that the R package class contains the knn function, which can be used to perform k-nearest neighbors classification.

(a) Create a binary variable, HighMPG, that equals 1 if a car's gas mileage is above the median in the Auto data set, and equals 0 if the car's gas mileage is below the median. 2

(b) Make a plot with horsepower on the x-axis, displacement on the y-axis, and with each of the cars in the Auto data set displayed as a point. The cars with gas mileage above the median should be displayed in one color, and the cars with gas mileage below the median should be displayed in another color. Be sure to create a legend and to label the axes appropri- ately.

(c) Use the validation set approach in order to estimate the test error of k- nearest neighbors classification, when using horsepower and displacement to predict HighMPG. Since this is a classification problem, you can define test error as the fraction of test set observations that are incorrectly clas- sified. Make a plot of the estimated test error, as a function of k. What value of k gives you the smallest estimated test error? Comment on your results.

(d) Now perform k-nearest neighbors regression on the full data set, for various values of k. Make a plot displaying the training error rate obtained, as a function of k, for the same values of k considered in (c). Comment on your results, and discuss how they relate to your findings in (c). Hint: In (c), make sure to consider an appropriate range of values for k! I'd like to see values of k that are too small" (in terms of estimated test error) and also values of k that are too large“.

5. Prove the following claim: The (training) RSS of the model y = 0 + is greater than or equal to the (training) RSS of the model y = 0 + 1X + : 3