FE590. Assignment #2.

Kejia Huang

2016-05-01

Instructions

In this assignment, you should use R markdown to answer the questions below. Simply type your R code into embedded chunks as shown above.

When you have completed the assignment, knit the document into a PDF file, and upload *both* the .pdf and .Rmd files to Canvas.

Question 1 (based on JWHT Chapter 2, Problem 9)

Use the `Auto` data set from the textbook's website. When reading the data, use the options `as.is = TRUE` and `na.strings="?"`. Remove the unavailable data using the `na.omet()` function.

```r
library(ISLR)
attach(Auto)
write.csv(Auto, file = "Auto.csv")
data_1 <- read.csv("Auto.csv",as.is = TRUE,na.strings = "?")
data_1 <- na.omit(data_1)
```

1. List the names of the variables in the data set.

```r
names(data_1)
```

```
##  [1] "X"            "mpg"          "cylinders"    "displacement"
##  [5] "horsepower"   "weight"       "acceleration" "year"
##  [9] "origin"       "name"
```

2. The columns `origin` and `name` are unimportant variables. Create a new data frame called `cars` that contains none of these unimportant variables.

```r
cars <- data_1[1:8]
```

3. What is the range of each quantitative variable? Answer this question using the `range()` function with the `sapply()` function (e.g., `sapply(cars, range)`. Print a simple table of the ranges of the variables. The rows should correspond to the variables. The first column should be the lowest value of the corresponding variable, and the second column should be the maximum value of the variable. The columns should be suitably labeled.

```
library(xtable)
cars_range <- sapply(cars,range)
xt <- xtable(data.frame(cars_range),
                 caption = "Range Of Variables",
                 label   =  "label",
                 type= "tex"
            )
print.xtable(xt)
```

% latex table generated in R 3.2.4 by xtable 1.8-2 package % Sun May 01 11:39:11 2016

	X	mpg	cylinders	displacement	horsepower	weight	acceleration	year
1	1.00	9.00	3.00	68.00	46.00	1613.00	8.00	70.00
2	397.00	46.60	8.00	455.00	230.00	5140.00	24.80	82.00

Table 1: Range Of Variables

4. What is the mean and standard deviation of each variable? Create a simple table of the means and standard deviations.

```
get_m_sd <- function(x){
    return(matrix(c(mean(x),sd(x)),nrow = 2, ncol = 1))
}
mean_sd <- apply(cars,2,get_m_sd)
row.names(mean_sd) <- c("Mean","SD")
xt <- xtable(data.frame(mean_sd),
                 caption = "Mean and Standard Deviations Of Variables",
                 label   =  "label",
                 type= "tex"
            )
print.xtable(xt)
```

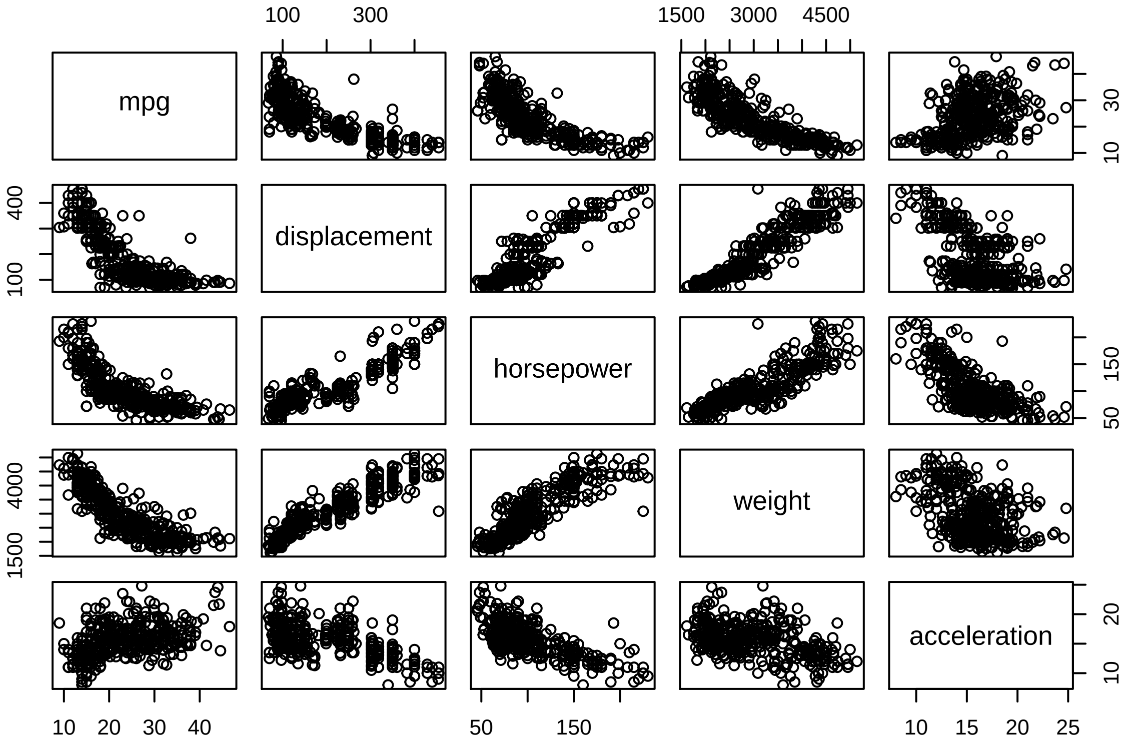
% latex table generated in R 3.2.4 by xtable 1.8-2 package % Sun May 01 11:39:11 2016

	X	mpg	cylinders	displacement	horsepower	weight	acceleration	year
Mean	198.52	23.45	5.47	194.41	104.47	2977.58	15.54	75.98
SD	114.44	7.81	1.71	104.64	38.49	849.40	2.76	3.68

Table 2: Mean and Standard Deviations Of Variables

5. Create a scatterplot matrix that includes the variables `mpg`, `displacement`, `horsepower`, `weight`, and `acceleration` using the `pairs()` function.

```
pair <- cars[-8][-3][-1]
pairs(pair)
```



6. From the scatterplot, it should be clear that has an almost linear relationship to predictors, and higher-order relationships to other variables. Using the `regsubsets` function in the `leaps` library, regress `mpg` onto

- `displacement`,
- `displacement` squared,
- `horsepower`,
- `horsepower` squared,
- `weight`,
- `weight` squared, and
- `acceleration`. Print a table showing what variables would be selected using best subset selection for all model orders.

```
library(leaps)
pair_2 <- cbind(pair,(pair[,2])^2,(pair[,3])^2,(pair[,4])^2)
colnames(pair_2) <- c("mpg","displacement","horsepower","weight","acceleration",
                      "squared_displacement","squared_horsepower","squared_weight")
regfit <- regsubsets(mpg~.,pair_2)
reg <- summary(regfit)
summary(regfit)
```

```
## Subset selection object
## Call: regsubsets.formula(mpg ~ ., pair_2)
## 7 Variables  (and intercept)
##                       Forced in Forced out
## displacement              FALSE      FALSE
## horsepower                FALSE      FALSE
## weight                    FALSE      FALSE
## acceleration              FALSE      FALSE
## squared_displacement      FALSE      FALSE
## squared_horsepower        FALSE      FALSE
## squared_weight            FALSE      FALSE
## 1 subsets of each size up to 7
## Selection Algorithm: exhaustive
##          displacement horsepower weight acceleration squared_displacement
## 1  ( 1 ) " "          " "        "*"    " "          " "                 
## 2  ( 1 ) " "          " "        "*"    " "          " "                 
## 3  ( 1 ) " "          "*"        "*"    " "          " "                 
## 4  ( 1 ) "*"          "*"        " "    "*"          " "                 
## 5  ( 1 ) "*"          "*"        " "    "*"          "*"                 
## 6  ( 1 ) "*"          "*"        "*"    "*"          "*"                 
## 7  ( 1 ) "*"          "*"        "*"    "*"          "*"                 
##          squared_horsepower squared_weight
## 1  ( 1 ) " "                " "           
## 2  ( 1 ) " "                "*"           
## 3  ( 1 ) "*"                " "           
## 4  ( 1 ) "*"                " "           
## 5  ( 1 ) "*"                " "           
## 6  ( 1 ) "*"                " "           
## 7  ( 1 ) "*"                "*"           
```

What is the most important variable affecting fuel consumption?

`weight`

What is the second most important variable affecting fuel consumption?

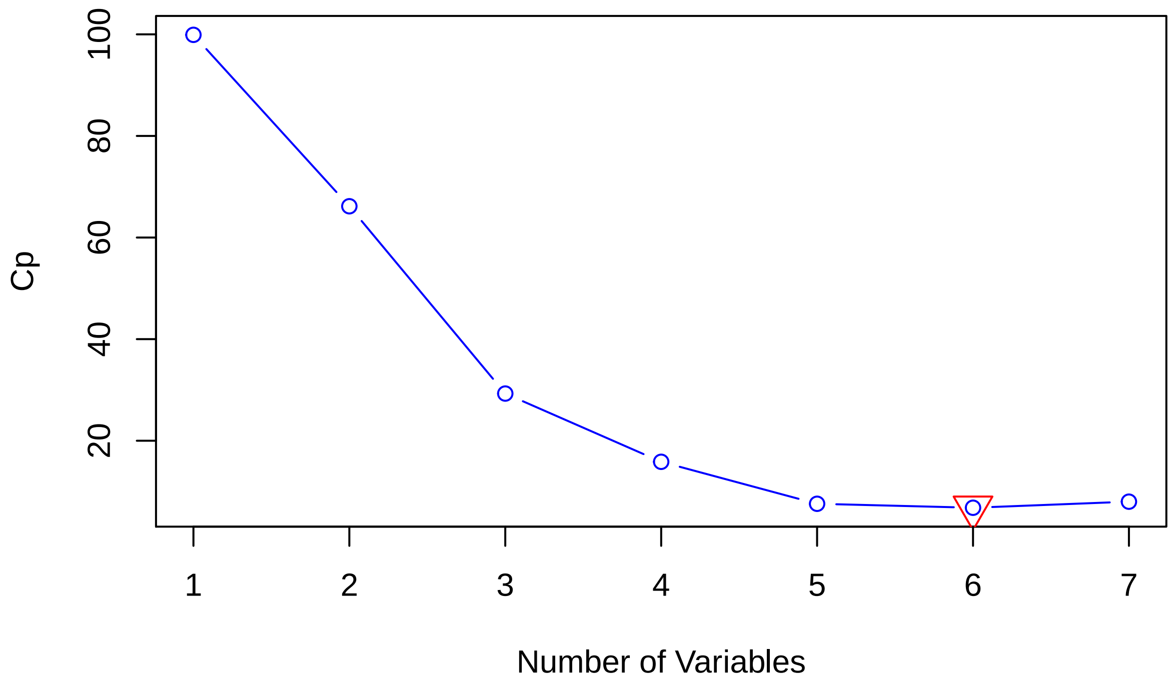
`weight` squared

What is the third most important variable affecting fuel consumption?

`horsepower` or `horsepower` squared

7. Plot a graph showing Mallow's Cp as a function of the order of the model. Which model is the best?

```r
plot(reg$cp,xlab = "Number of Variables",col="blue",ylab = "Cp",type = "b")
n <- which.min(reg$cp)
points(n,reg$cp[n],col="red",cex=2,pch=25)
```



So the model have 6 variables is best

Question 2 (based on JWHT Chapter 3, Problem 10)

This exercise involves the Boston housing data set.

1. Load in the `Boston` data set, which is part of the `MASS` library in R. The data set is contained in the object `Boston`. Read about the data set using the command `?Boston`. How many rows are in this data set? How many columns? What do the rows and columns represent?

```
library(MASS)
attach(Boston)
#?Boston
```

The Boston data frame has 506 rows and 14 columns. 506 rows represent 506 differents observations for different suburbs. 14 columns are:

```
colnames(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

crim: per capita crime rate by town.

zn: proportion of residential land zoned for lots over 25,000 sq.ft.

indus: proportion of non-retail business acres per town.

chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox: nitrogen oxides concentration (parts per 10 million).

rm: average number of rooms per dwelling.

age: proportion of owner-occupied units built prior to 1940.

dis: weighted mean of distances to five Boston employment centres.

rad: index of accessibility to radial highways.

tax: full-value property-tax rate per $10,000.

ptratio: pupil-teacher ratio by town.

black: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town.

lstat: lower status of the population (percent).

medv: median value of owner-occupied homes in $1000s.

2. Do any of the suburbs of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

```
high_rate <- matrix(c(which.max(Boston$crim),which.max(Boston$tax),
                      which.max(Boston$ptratio)),ncol = 1)
ran <- matrix(rbind(range(Boston$crim),range(Boston$tax),
                    range(Boston$ptratio)),nrow = 3, ncol = 2)
bo <- cbind(high_rate,ran)
colnames(bo) <- c("suburbs","range_min","range_max")
rownames(bo) <- c("crim","tax","ptratio")
bo
```

```
##         suburbs range_min range_max
## crim        381  6.32e-03   88.9762
## tax         489  1.87e+02  711.0000
## ptratio     355  1.26e+01   22.0000
```

From the table we can see that, the range of crim rate is very big, some suburbs are very safe, but some place very dangerous. The range of tax rate is big too. But the range of pupil-teacher ratios is not so much compared with other two rates.

3. How many of the suburbs in this data set bound the Charles river?

```
sum(Boston$chas)
```

```
## [1] 35
```

4. What is the median pupil-teacher ratio among the towns in this data set?

```
median(Boston$ptratio)
```

```
## [1] 19.05
```

5. In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the suburbs that average more than eight rooms per dwelling.

```
idx_7 <- Boston$rm > 7
print("number of suburbs whose average number of room more than 7")
```

```
## [1] "number of suburbs whose average number of room more than 7"
```

```
sum(idx_7)
```

```
## [1] 64
```

```
print("number of suburbs whose average number of room more than 8")
```

```
## [1] "number of suburbs whose average number of room more than 8"
```

```
idx_8 <- Boston$rm > 8
sum(idx_8)
```

```
## [1] 13
```

```
which(idx_8)
```

```
##  [1]  98 164 205 225 226 227 233 234 254 258 263 268 365
```

We can see that the suburbs average more than eight room per dwelling mostly in the center of the Boston.

Question 3 (based on JWHT Chapter 4, Problem 10)

This question should be answered using the `Weekly` data set, which is part of the `ISLR` package. This data contains 1,089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

1. What does the data represent?

```
library(ISLR)
attach(Weekly)
```

Weekly: Weekly percentage returns for the S&P 500 stock index between 1990 and 2010. It contains 1089 observation on the following 9 varibles.

Year: The year that the observation was recorded

Lag1: Percentage return for previous week

Lag2: Percentage return for 2 weeks previous

Lag3: Percentage return for 3 weeks previous

Lag4: Percentage return for 4 weeks previous

Lag5: Percentage return for 5 weeks previous

Volume: Volume of shares traded (average number of daily shares traded in billions)

Today: Percentage return for this week

Direction: A factor with levels Down and Up indicating whether the market had a positive or negative return on a given week

2. Use the full data set to perform a logistic regression with `Direction` as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?

```
glm.fit <- glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume,
               data=Weekly, family=binomial)
summary(glm.fit)
```

```
##
## Call:
## glm(formula = Direction ~ Lag1 + Lag2 + Lag3 + Lag4 + Lag5 +
##     Volume, family = binomial, data = Weekly)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6949  -1.2565   0.9913   1.0849   1.4579
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.26686    0.08593   3.106   0.0019 **
## Lag1        -0.04127    0.02641  -1.563   0.1181
## Lag2         0.05844    0.02686   2.175   0.0296 *
## Lag3        -0.01606    0.02666  -0.602   0.5469
## Lag4        -0.02779    0.02646  -1.050   0.2937
## Lag5        -0.01447    0.02638  -0.549   0.5833
## Volume      -0.02274    0.03690  -0.616   0.5377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1496.2  on 1088  degrees of freedom
## Residual deviance: 1486.4  on 1082  degrees of freedom
## AIC: 1500.4
##
## Number of Fisher Scoring iterations: 4
```

The smallest p-value here is associated with `Lag2`, except `intercept`. So `Lag2` appears to be statistically significant than others.

3. Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

```
glm.probs <- predict(glm.fit, type="response")
glm.pred <- rep("Down",length(glm.probs))
glm.pred[glm.probs > 0.5] <- "Up"
table(glm.pred,Direction)
```

```
##         Direction
## glm.pred Down  Up
##     Down   54  48
##     Up    430 557
```

```
print("Overall Fraction Of Correct Predictions")
```

```
## [1] "Overall Fraction Of Correct Predictions"
```

```
mean(glm.pred == Direction)
```

```
## [1] 0.5610652
```

Type one error is 430, it means we predict the direction is `Up`, but actucally is `Down`. Type two error is 48, it means we predict the direction is `Down`, but actucally is `Up`.

4. Fit a logistic regression model using a training data period from 1990 to 2008, with `Lag2` as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).

```
train <- (Year < 2009 )
Weekly.2009 <- Weekly[!train,]
Direction.2009 <- Direction[!train]
glm.fit2 <- glm(Direction~Lag2,
               data=Weekly, family=binomial,subset = train)
glm.probs2 <- predict(glm.fit2,Weekly.2009,type="response")
glm.pred2 <- rep("Down",length(Direction.2009))
glm.pred2[glm.probs2>0.5]="Up"
table(glm.pred2,Direction.2009)
```

```
##          Direction.2009
## glm.pred2 Down Up
##      Down    9  5
##      Up     34 56
```

```
print("Overall Fraction Of Correct Predictions")
```

```
## [1] "Overall Fraction Of Correct Predictions"
```

```
LG_mean <- mean(glm.pred2 == Direction.2009)
LG_mean
```

```
## [1] 0.625
```

5. Repeat Part 4 using LDA.

```r
library(MASS)
lda.fit <- lda(Direction~Lag2, data = Weekly, subset = train)
lda.pred <- predict(lda.fit, Weekly.2009)
lda.class <- lda.pred$class
table(lda.class,Direction.2009)
```

```
##          Direction.2009
## lda.class Down Up
##      Down    9  5
##      Up     34 56
```

```r
print("Overall Fraction Of Correct Predictions")
```

```
## [1] "Overall Fraction Of Correct Predictions"
```

```r
LDA_mean <- mean(lda.class == Direction.2009)
LDA_mean
```

```
## [1] 0.625
```

6. Repeat Part 4 using QDA.

```r
qda.fit <- qda(Direction~Lag2, data = Weekly, subset = train)
qda.class <- predict(qda.fit,Weekly.2009)$class
table(qda.class,Direction.2009)
```

```
##          Direction.2009
## qda.class Down Up
##      Down    0  0
##      Up     43 61
```

```r
print("Overall Fraction Of Correct Predictions")
```

```
## [1] "Overall Fraction Of Correct Predictions"
```

```r
QDA_mean <- mean(qda.class == Direction.2009)
QDA_mean
```

```
## [1] 0.5865385
```

7. Repeat Part 4 using `KNN` with $K = 1, 2, 3$.

```r
library(class)
```

```
## Warning: package 'class' was built under R version 3.2.5
```

```r
train.X <- as.matrix(Weekly$Lag2[train])
test.X <- as.matrix(Weekly$Lag2[!train])
train.Direction <- Direction[train]
set.seed(1)
knn.pred_1 <- knn(train.X,test.X,train.Direction,k=1)
table(knn.pred_1,Direction.2009)
```

```
##           Direction.2009
## knn.pred_1 Down Up
##       Down   21 30
##       Up     22 31
```

```r
print("Overall Fraction Of Correct Predictions")
```

```
## [1] "Overall Fraction Of Correct Predictions"
```

```r
KNN_1_mean <- (21 + 31)/104
KNN_1_mean
```

```
## [1] 0.5
```

```r
set.seed(2)
knn.pred_2 <- knn(train.X,test.X,train.Direction,k=2)
table(knn.pred_2,Direction.2009)
```

```
##           Direction.2009
## knn.pred_2 Down Up
##       Down   20 25
##       Up     23 36
```

```r
print("Overall Fraction Of Correct Predictions")
```

```
## [1] "Overall Fraction Of Correct Predictions"
```

```r
KNN_2_mean <- (20 + 36)/104
KNN_2_mean
```

```
## [1] 0.5384615
```

```r
set.seed(3)
knn.pred_3 <- knn(train.X,test.X,train.Direction,k=3)
table(knn.pred_3,Direction.2009)
```

```
##           Direction.2009
## knn.pred_3 Down Up
##       Down   15 19
##       Up     28 42
```

```
print("Overall Fraction Of Correct Predictions")
```

```
## [1] "Overall Fraction Of Correct Predictions"
```

```
KNN_3_mean <- (16 + 42)/104
KNN_3_mean
```

```
## [1] 0.5576923
```

8. Which of these methods in Parts 4, 5, 6, and 7 appears to provide the best results on this data?

```
library(xtable)
df <- matrix(c(LG_mean,LDA_mean,QDA_mean,KNN_1_mean,KNN_2_mean,KNN_3_mean),nrow = 1,byrow = FALSE)
colnames(df) <- c("LG","LDA","QDA","KNN_1","KNN_2","KNN_3")
df <- data.frame(
    Methods = df
    )
xt <- xtable(df,
                 caption = "Overll Fraction of Correct Predictions",
                 label   =  "label",
                 type= "tex"
            )
print.xtable(xt)
```

% latex table generated in R 3.2.4 by xtable 1.8-2 package % Sun May 01 11:39:12 2016

	Methods.LG	Methods.LDA	Methods.QDA	Methods.KNN_1	Methods.KNN_2	Methods.KNN_3
1	0.62	0.62	0.59	0.50	0.54	0.56

Table 3: Overll Fraction of Correct Predictions

Logistic regression model and LDA model provide the best results on this data.