DNSC 6279 Spring 2020 Homework 2

Ziwei Li

Remark. This homework aims to help you go through the necessary preliminary from linear regression. Credits for **Theoretical Part** and **Computational Part** are in total 100 pt. For **Computational Part**, please complete your answer in the **RMarkdown** file and summit your printed PDF homework created by it.

Computational Part

1. (*35 pt*) Consider the dataset "Boston" in predicting the crime rate at Boston with associated covariates.

```
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

Suppose you would like to predict the crime rate with explantory variables

- `medv` - Median value of owner-occupied homes
- `dis` - Weighted mean of distances to employement centers
- `indus` - Proportion of non-retail business acres

Run with the linear model

```
mod1 <- lm(crim ~ medv + dis + indus, data = Boston)
summary(mod1)
```

```
##
## Call:
## lm(formula = crim ~ medv + dis + indus, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.625  -3.345  -1.242   1.608  78.994
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.67738    2.12190   5.503 5.95e-08 ***
## medv        -0.26061    0.04204  -6.199 1.19e-09 ***
## dis         -0.96320    0.22758  -4.232 2.75e-05 ***
## indus        0.13145    0.07728   1.701   0.0896 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.519 on 502 degrees of freedom
## Multiple R-squared:  0.2404, Adjusted R-squared:  0.2358
## F-statistic: 52.95 on 3 and 502 DF,  p-value: < 2.2e-16
```

Answer the following questions.

i. What do the following quantities that appear in the above output mean in the linear model? Provide a breif description.

  - `t value` and `Pr(>|t|)` of `medv`

  Answer: T value is its coefficients divided by the std error. It is like a measure of the precision with which the regression is measured. Here, the 'medv's t-score is -6.199. P value is used to test the null hypothesis that the coefficient is equal to 0. If the p-value is less than the alpha, then we will reject the null hypothesis and concluded that coefficient is not equal to 0. Otherwise, we will fail to reject null hypothesis and conclude that the coefficient is equal to 0. Here, the p-value for 'medv' is 1.19e-09, if alpha equals to 0.05, then we will reject the null hypothesis and conclude that the coefficient of 'medv' is significant.

  - `Multiple R-squared`

  Answer: R-square is also called coefficient of determination. It measures the the percentage of variation in the dependent variable explained by the independent variables. It is calculated by SSR divided by SST. Here, we can say that 24.04% data can be explained by the model we created.

  - `F-statistic`, `DF` and corresponding `p-value`

  Answer: F-statistic is a result to test the overall significane of this model. This can be calculated using ANOVA table. If F-statisitc is bigger than critical value, then the null hypothesis will be rejected and conclude that the estimated regression model is significant overall. DF is degree of freedom. The number of independent pieces of information that go into the the estimate of a parameter. In F-test, we have two DF, first is the number of variable, which is 3 here. And second is the number of observation minus the number of variable minus 1, here we have 502 df. And the total degree of freedom is number of observation minus 1. P-value is the probability of obtaining the observed results of a test. In F-test, we try to test if the overall model is significant. After we get the value, we can use this to compare with the alpha to reach a result. If the p-value is less than alpha, the null hypothesis will be rejected and we can conclude that the over model is significant. Otherwise, we fail to reject the null hypothesis and conclude that overall model is not significant.

---

ii. Are the following sentences True of False? Briefly justify your answer. + `indus` is not a significant predictor of crim, and we can drop this from the model.

```
**Answer:** False. From the model's perspective, the p-value of 'indus' is bigger tha
n alpha = 0.05, and we fail to reject the null hypothesis and conclude that coefficie
nt of 'indus' is not significant, therefore, we may excluse this variable from this m
odel. However,from other perspectives, we still need more information to decide wheth
er this variable is irrevelant or not.

***
+ `Multiple R-squared` is preferred to `Adjusted R-squared` as it takes into account
all the variables.

**Answer:** No. R-squared increases with more independent vaeiables, regardless of wh
ether they are actually related to the dependen variable. R-squared assumes that ever
y single variable explains the variation in the dependent variable. The adjusted r-sq
uared is the percentage of variation explained by only the independent variables that
actually affect the dependent variables.

***
+ `medv` has a negative effect on the response.

**Answer:** Yes. Because the coefficient of 'medv' is negative. If 'medv' gets bigger
, and the rest of the variables stay constant, then the 'crim' will decrease.

***
+ Our model residuals appear to be normally distributed.

\begin{hint}
  You need to access to the model residuals in justifying the last sentence. The foll
owing commands might help.
\end{hint}

```r
# Obtain the residuals
res1 <- residuals(mod1)

# Normal QQ-plot of residuals
plot(mod1, 2)

# Conduct a Normality test via Shapiro-Wilk and Kolmogorov-Smirnov test
```
```

```
shapiro.test(res1)
ks.test(res1, "pnorm")
```

```
**Answer:** First, we can see from the qqplot. If the qqplot shows a diagonal line, t
hen the data is are normally distributed. However, we cannot say this is a diagonal l
ine, then we may say that the data is not normally distributed.
            Second, we can use the Shapiro-Wilk and Kolmogorov-Smirnov test to check
if the data is normally distributed. In Shapiro-Wilk test, if the p-vaue is larger th
an alpha, then we may fail to reject the null hypothesis and conclude that the data i
s normally distributed. However, here, the p-value is less than 2.2e-16. Therefore, w
e may conclude that the data is not normally distributed.
            In Kolmogorov-Smirnov test, the null hypothesis states that the data foll
ow a specificed distribution. And the alternative hypothesis states that the data do
not follow a specificed distribution. If the p-value is less than the alpha, we will
reject the null hypothesis. Here, in Kolmogorov-Smirnov test, we can get a p-value of
p-value < 2.2e-16, we may reject the null hypothesis and conclude that the data does
not follow a normal distribution.

***
```

2. (*35 pt*, Textbook Exercises 3.10) This question should be answered using the `Carseats` data set.

```
head(Carseats)
```

```
##    Sales CompPrice Income Advertising Population Price ShelveLoc Age
## 1  9.50       138     73          11        276   120       Bad  42
## 2 11.22       111     48          16        260    83      Good  65
## 3 10.06       113     35          10        269    80    Medium  59
## 4  7.40       117    100           4        466    97    Medium  55
## 5  4.15       141     64           3        340   128       Bad  38
## 6 10.81       124    113          13        501    72       Bad  78
##   Education Urban  US
## 1        17   Yes Yes
## 2        10   Yes Yes
## 3        12   Yes Yes
## 4        14   Yes Yes
## 5        13   Yes  No
## 6        16    No Yes
```

a. Fit a multiple regression model to predict `Sales` using `Price` , `Urban` , and `US` .

Answer: mod2 <- lm(Sales~Price+Urban+US,data=Carseats)

---

b. Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

Answer: summary(mod2) If we have zero price and not urban and not an us citizen, then the sales will be 13.043469 thousand. The coefficient of price indicates that for every additional price increase you may expect the sale to decrease by an average of 0.054459 thousand. The coefficient of Urban indicates that if the stores is located in urban the sale will decrease by an average of 0.021916 thousand, if it is located in urban, the sale will remain constant if the rest of variables remain constant. The coefficient for US indicates that if the store is located in US, the sale will increase by an average of 1.200573 thousand. However, if it is located in US, the sale will remain constant if the rest of variables remain constant.

c. Write out the model in equation form, being careful to handle the qualitative variables properly.

Answer: Sales = 13.043469 -0.054459*Price -0.021916*Urban(yes) + 1.200573*Us(yes)

---

d. For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

Answer: The p-value for Price and USYes are all less than 0.05, then the null hypothesis can be rejected and conclude that these variables are significant.

---

e. On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

Answer:mod3 <- lm(Sales~Price+US,data=Carseats)

---

f. How well do the models in (a) and (e) fit the data?

Answer: summary(mod2) summary(mod3) The multiple r-squared is 0.2393 and adjusted r-squared is 0.2335. For adjusted r-squared, that 23.35% data can be explained using model from (a). The multiple r-squared is 0.2393 and adjusted r-squared is 0.2354.For adjusted r-squared, that 23.54% data can be explained using model from (3). If using adjusted r-squared to decide which model works best, it might be model from (3), because the adjusted r-squared is higher.

---

g. Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

Answer:

```
## step 1: calculate the alpha:
alpha = 0.05
## step 2: calculate the p value
p = 1 - alpha/2
## step 3: calculate the degree of freedom
df = nrow(Carseats) - 2
## step 4: calculate the critcal value
critical=qt(p,df)
## step 5: compute the 'Price''s margin of error
### ME = critical * StdError
pme = critical*0.00523
## step 6: calculate the interval
c(-0.05448-pme, -0.05448+pme)
```

```
## [1] -0.06476188 -0.04419812
```

```
## Step 5.1 Then, we can repeat the step 5 to get the 'US''s margin of error
ume =  critical * 0.25846
## Step 6.1 Calculate the interval
c(1.19964-ume,1.19964+ume)
```

```
## [1] 0.6915225 1.7077575
```

h. Using the leave-one-out cross-validation and 5-fold cross-validation techniques to compare the performance of models in (a) and (e). What can you tell from (f) and (h)?

Hint. Functions `update` (with option `subset`) and `predict`.

Answer:

```r
# Split the data into training and test
## Step 1. Shuffle the whole dataset
Carseats2 <- Carseats
Carseats4 <- Carseats2[sample(nrow(Carseats2)),]
## Set the test_size = 0.2
test_loocv = Carseats4[c(1:(nrow(Carseats4)*0.2)),]
train_loocv = Carseats4[-c(1:(nrow(Carseats4)*0.2)),]
# Leave-one-out Cross Validation
## For model in (a)
ssr_1 = c()
for (i in 1:nrow(train_loocv)){
  sub = train_loocv[-i,]
  model = lm(Sales~Price+Urban+US,data=sub)
  fact  = train_loocv[i,c(6,10,11)]
  pred = predict(model,data.frame(fact))
  actual = train_loocv[i,1]
  error = (pred-actual)^2
  ssr_1 = append(ssr_1,error)
}
mean(ssr_1)
```

```
## [1] 6.365374
```

```r
## For model in (e)
ssr_2 = c()
for (s in 1:nrow(train_loocv)){
  sub2 = Carseats[-s,]
  model2 = lm(Sales~Price+US,data=sub2)
  fact2  = train_loocv[s,c(6,11)]
  pred2 = predict(model2,data.frame(fact2))
  actual2 = train_loocv[s,1]
  error2 = (pred2-actual2)^2
  ssr_2 = append(ssr_2,error2)
}
mean(ssr_2)
```

```
## [1] 6.234327
```

```
## Using the leave-one-out method, the model in (a)'s MSE is 6.214935. The model in (
e)'s MSE is 6.0099382. The MSE in (e) is less than MSE in (a). Therefore, we can conc
lude that the model in (e) is better.

# 5-fold Cross Validation
## Data Preprocessing
###shuffle the data
Carseats2 <- Carseats
Carseats3 <- Carseats2[sample(nrow(Carseats2)), ]
###Create 5 subset
s1 = Carseats3[c(1:80),c(1,6,10,11)]
s2 = Carseats3[c(81:160),c(1,6,10,11)]
s3 = Carseats3[c(161:240),c(1,6,10,11)]
s4 = Carseats3[c(241:320),c(1,6,10,11)]
s5 = Carseats3[c(321:400),c(1,6,10,11)]
fold5 = list(s1,s2,s3,s4,s5)

## For model in (a)
x36 = c()
for(y in 1:length(fold5)){
    r=0
    if(r<=80){
      x34 = c()
      model3 = lm(Sales~Price+Urban+US,data=fold5[[y]])
      fact3 = fold5[[y]][c(2,3,4)]
      actual = fold5[[y]][1]
      pred3 = predict(model3,data.frame(fact3))
      error3 = (actual-pred3)^2
      x34 = append(x34,error3)
      r=r+1
    }else{
      break
    }
    x35 = sum(x34$Sales)
    x36 = append(x36,x35)
}
msea = mean(x36)
msea
```

```
## [1] 465.1355
```

1. I think we need more information to justify. However, in my opinion, Polynomial regression will have a less SSR on training data. Linear regression is regarded as inflexible because it is biased towards linear relationships among its variables. And also some outliers may influence the fit of the model. However, polynomial regression is more flexible and may detect some non-linear relationships among data. Therefore, the polynomial regression may have less SSR and can better fit the data.

2. For the test data, if there is a true relationship between X and Y is linear, I believe the linear regression may have less SSR. Because the polynomial may overfit the data and have more errors than linear regression. Therefore, I believe the linear regression may have lower SSR on test data.

3. I think the Polynomial regression will have lower SSR on training data. The answer is almost the same the question 1. Because Polynomial regression has more flexibility and deal with non-linear in a better way than linear regression. Therefore, I believe the polynomial regression will have a lower SSR.

4. I think we need more information to decide which model is better because we do not know how far it is from linear. If it is closer to linear regression, then we may believe that the linear regression may have a lower SSR on test data. However, if it is more closer to polynomial regression, then the polynomial regression may fit the data better and have a lower SSR.

```r
## For model in (e)
x41 = c()
for(w in 1:length(fold5)){
    t=0
    if(t<=80){
      x42 = c()
      model4 = lm(Sales~Price+US,data=fold5[[w]])
      fact4 = fold5[[w]][c(2,4)]
      actual4 = fold5[[w]][1]
      pred4 = predict(model4,data.frame(fact4))
      error4 = (actual4-pred4)^2
      x42 = append(x42,error4)
      t=t+1
    }else{
      break
    }
    x43 = sum(x42$Sales)
    x41 = append(x41,x43)
}
msee = mean(x41)
msee
```

```
## [1] 468.9397
```

```
# Using the 5-fold method, the mse in (a) is 466.506, and the mse in (e) is 462.4411.
The mse in (e) is less than mse in (a). Therefore, we can conclude that the model in
(e) is bette than model in (a).
```

Theoretical Part

1. I think we need more information to justify. However, in my opinion, Polynomial regression will have a less SSR on training data. Linear regression is regarded as inflexible because it is biased towards linear relationships among its variables. And also some outliers may influence the fit of the model. However, polynomial regression is more flexible and may detect some non-linear relationships among data. Therefore, the polynomial regression may have less SSR and can better fit the data.

2. For the test data, if there is a true relationship between X and Y is linear, I believe the linear regression may have less SSR. Because the polynomial may overfit the data and have more errors than linear regression. Therefore, I believe the linear regression may have lower SSR on test data.

3. I think the Polynomial regression will have lower SSR on training data. The answer is almost the same the question 1. Because Polynomial regression has more flexibility and deal with non-linear in a better way than linear regression. Therefore, I believe the polynomial regression will have a lower SSR.

4. I think we need more information to decide which model is better because we do not know how far it is from linear. If it is closer to linear regression, then we may believe that the linear regression may have a lower SSR on test data. However, if it is more closer to polynomial regression, then the polynomial regression may fit the data better and have a lower SSR.