STOR 565 Spring 2020 Homework 2

Yimei Fan

*Remark.* This homework aims to help you go through the necessary preliminary material from linear regression. Credits for **Theoretical Part** and **Computational Part** are in total 110 pt. At most 100 pt can be earned for this homework. For **Computational Part**, please complete your answer in the **RMarkdown** file and submit your printed PDF homework created using **RMarkdown**.

## Computational Part

1. (*25 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 lstat  
## 1 0.00632 18 2.31 0 0.538 6.575 65.2 4.0900 1 296 15.3 396.90 4.98  
## 2 0.02731 0 7.07 0 0.469 6.421 78.9 4.9671 2 242 17.8 396.90 9.14  
## 3 0.02729 0 7.07 0 0.469 7.185 61.1 4.9671 2 242 17.8 392.83 4.03  
## 4 0.03237 0 2.18 0 0.458 6.998 45.8 6.0622 3 222 18.7 394.63 2.94  
## 5 0.06905 0 2.18 0 0.458 7.147 54.2 6.0622 3 222 18.7 396.90 5.33  
## 6 0.02985 0 2.18 0 0.458 6.430 58.7 6.0622 3 222 18.7 394.12 5.21  
## medv  
## 1 24.0  
## 2 21.6  
## 3 34.7  
## 4 33.4  
## 5 36.2  
## 6 28.7

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

* age - Proportion of owner-occupied units built prior to 1940
* dis - Weighted mean of distances to employement centers
* rad - Index of accessibility to radial highways

Run with the linear model

mod1 <- lm(crim ~ age + dis + rad, data = Boston)  
summary(mod1)

##   
## Call:  
## lm(formula = crim ~ age + dis + rad, data = Boston)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.097 -1.572 -0.265 0.699 76.240   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.75072 1.85563 -0.943 0.346   
## age 0.01339 0.01611 0.831 0.406   
## dis -0.25642 0.22056 -1.163 0.246   
## rad 0.56750 0.03979 14.264 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.69 on 502 degrees of freedom  
## Multiple R-squared: 0.3986, Adjusted R-squared: 0.395   
## F-statistic: 110.9 on 3 and 502 DF, p-value: < 2.2e-16

Answer the following questions.

1. What do the following quantities that appear in the above output mean in the linear model? Provide a brief description.
   * t value and Pr(>|t|) of rad

* **Answer:** t-value: it is calculated by estimate/Std.Error. It is used to calculated P-value, which can be used to do hypothesis testing for each variables. Pr(>|t|) of rad: H0: the coefficient of ‘rad’ equals to zero. Ha: the coefficient of ‘rad’ does not equal to zero. Since the p-value for rad is <2e-16, and it is lower than 0.05. We can reject H0, which means the coefficient of rad is not zero and it is significant.
  + Adjusted R-squared
* **Answer:** Adjusted R-squared: It measures how well the model can fit the data. It will not be influenced by the number of variables. The formula is 1-MSE/MST.
  + F-statistic, DF and corresponding p-value
* **Answer:** F-statistic: It is similar to t-statistic but F-statistic is for the whole model. If there is only one variable, then F-statistic should equals to (t-statistic)^2. It compares the fit of the whole model and null model(Y=1) and the larger F-statistic is, the more significant the whole model will be. DF: There are two DF in the table. The first one is the number of variables(p), and the second one is for the degree of fredom for whole model: n-p-1. p-value: It is calculated from F-statistics, and it is used to conduct hypothesis test for H0: all coefficients are zero. Ha: At lease one coefficient is not zero.2.2e-16 is lower than 0.05, therefore, we should reject H0, that means at least one coefficient is not zero.

1. Are the following sentences True of False? Briefly justify your answer. + age is not a significant predictor of crim, and we can drop it from the model.

\*\*Answer:\*\*   
True. P-value is 0.406, and it is higher than 0.05.  
  
\*\*\*  
+ Both `Multiple R-squared` and `Adjusted R-squared` increase with number of variables since they take into account all the variables and more variance is explained as the model becomes more complicated when the number of variables increases.   
  
\*\*Answer:\*\*   
False. Adjusted R-squared will not increase with the number of variables, but it will increase if the model fits data better.  
  
\*\*\*   
+ `rad` has a positive effect on the response.  
  
\*\*Answer:\*\*  
True. The coefficient of 'rad' is 0.56750, and it is positive. Moreover, 'rad' is significant in this model, which means it does have effect on the response.  
  
  
\*\*\*  
+ Our model residuals appear to be normally distributed.  
  
\begin{hint}  
 You need to access to the model residuals in justifying the last sentence. The following 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:\*\*   
  
True. According to the qq-plot, the residuals roughly form a line, which means they are normally distributed.  
According to Normality tests, W and D are larger than 0.05. Therefore, the residuals are normally distributed.  
  
\*\*\*

1. (*30 pt*, Similar to Textbook Exercises 3.10) This question should be answered using the Carseats data set.

head(Carseats)

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

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

**Answer:**

model = lm(Sales~US+Urban+Advertising+Price, data = Carseats)  
print(summary(model))

##   
## Call:  
## lm(formula = Sales ~ US + Urban + Advertising + Price, data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.879 -1.565 -0.027 1.543 6.367   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.011278 0.633326 20.544 < 2e-16 \*\*\*  
## USYes 0.058514 0.345060 0.170 0.865   
## UrbanYes -0.038790 0.264279 -0.147 0.883   
## Advertising 0.120333 0.024837 4.845 1.82e-06 \*\*\*  
## Price -0.054612 0.005099 -10.710 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.405 on 395 degrees of freedom  
## Multiple R-squared: 0.2819, Adjusted R-squared: 0.2747   
## F-statistic: 38.77 on 4 and 395 DF, p-value: < 2.2e-16

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

**Answer:** The coeffieicents of ‘US’ and ‘Urban’ are not significant, and we should not include them in the model. The coefficient of ‘Advertising’ is 0.120333, and it means that with more advertising, the sales will increase. The coefficient of ‘Price’ is -0.054612, and it means if the price of car seats increase, the sales will decrease accordingly.

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

**Answer:** If ‘Yes’ in ‘US’ and ‘Yes’ in ‘Urban’ equals 1: sales = 13.03 + 0.12Advertising -0.05Price

If ‘Yes’ in ‘US’ and ‘Yes’ in ‘Urban’ equals 0: sales = 13.01 + 0.12Advertising -0.05Price

1. For which of the predictors can you reject the null hypothesis ?

**Answer:** In order to reject the null hypothesis, the P-value need to be smaller than 0.05. Therefore, for ‘Advertising’ and ‘Price’, we can reject the null hypothesis.

1. 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:**

summary(lm(Sales~Advertising+Price, data = Carseats))

##   
## Call:  
## lm(formula = Sales ~ Advertising + Price, data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.9011 -1.5470 -0.0223 1.5361 6.3748   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.003427 0.606850 21.428 < 2e-16 \*\*\*  
## Advertising 0.123107 0.018079 6.809 3.64e-11 \*\*\*  
## Price -0.054613 0.005078 -10.755 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.399 on 397 degrees of freedom  
## Multiple R-squared: 0.2819, Adjusted R-squared: 0.2782   
## F-statistic: 77.91 on 2 and 397 DF, p-value: < 2.2e-16

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

**Answer:** The model in (a): Adjusted R-squared is 0.2747, P-value < 2.2e-16 The model in (e): Adjusted R-squared is 0.2782, P-value < 2.2e-16

Compared to model(a), model(e) has slightly higher adjusted R-squared, and both models are significant. Therefore, both model fit the data well and model(e) is slightly better.

1. Using the model from (e), obtain 90% confidence intervals for the coefficient(s).

**Answer:**

#Advertising:  
tstar = qt(.9+0.05,397)  
  
0.123107+ tstar\*0.018079\*c(-1,1)

## [1] 0.09330014 0.15291386

#Price:  
tstarr = qt(.9+0.05,397)  
  
-0.054613 + tstarr\*0.005078\*c(-1,1)

## [1] -0.0629851 -0.0462409

1. 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:**

library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

#leave-one-out cross-validation  
#(a)  
train.control1 <- trainControl(method = "LOOCV")  
model1 <- train(Sales~US+Urban+Advertising+Price, data = Carseats, method = "lm",  
 trControl = train.control1)  
print(model1)

## Linear Regression   
##   
## 400 samples  
## 4 predictor  
##   
## No pre-processing  
## Resampling: Leave-One-Out Cross-Validation   
## Summary of sample sizes: 399, 399, 399, 399, 399, 399, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 2.420802 0.263687 1.914181  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

#(e)  
train.control2 <- trainControl(method = "LOOCV")  
model2 <- train(Sales ~ Advertising + Price, data = Carseats, method = "lm",  
 trControl = train.control2)  
print(model2)

## Linear Regression   
##   
## 400 samples  
## 2 predictor  
##   
## No pre-processing  
## Resampling: Leave-One-Out Cross-Validation   
## Summary of sample sizes: 399, 399, 399, 399, 399, 399, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 2.408753 0.2708067 1.904313  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

#RMSE for (a) is 2.420802, and Rsquared is 0.263687; RMSE for (e) is 2.408753 and Rsquared is 0.2708067. (e) performs better.  
  
  
#5-fold cross-validation  
#(a)  
set.seed(123)   
train.control3 <- trainControl(method = "cv", number = 5)  
model3 <- train(Sales~US+Urban+Advertising+Price, data = Carseats, method = "lm",  
 trControl = train.control3)  
print(model3)

## Linear Regression   
##   
## 400 samples  
## 4 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 320, 321, 319, 320, 320   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 2.436002 0.2746255 1.923115  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

#(e)  
set.seed(123)   
train.control4 <- trainControl(method = "cv", number = 5)  
model4 <- train(Sales ~ Advertising + Price, data = Carseats, method = "lm",  
 trControl = train.control4)  
print(model4)

## Linear Regression   
##   
## 400 samples  
## 2 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold)   
## Summary of sample sizes: 320, 321, 319, 320, 320   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 2.42669 0.2803708 1.916222  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

#RMSE for (a) is 2.436002, and Rsquared is 0.2746255; RMSE for (e) is 2.42669 and Rsquared is 0.2803708. (e) performs better.  
  
  
#(f) and (h) gives us similar result regarding to model performance and they provide similar R-squared value.