ADTA 5410 by Levent Bulut, Week 2 Lab

AUTHOR
Justin Bussey

General Instructions

1. Please note that this Quarto document constitutes 10% of your weekly R Lab assignment grade. The remaining 90% of your grade is determined by your answers to questions in Canvas. Be sure to read each question in Canvas carefully and provide complete and accurate answers.
2. You can create a new folder for this lab assignment and store this Quarto document and the provided data set in the same folder you just created.
3. The first code chunk installs certain R packages that might be useful to answer some of the questions.
4. Unless instructed otherwise, you can choose any R package you want to answer the questions. You are not limited to using the packages listed in this Quarto template.
5. Be sure to include the code that produces each answer, and make sure that your code and output are visible in your knitted HTML document.
6. When you are finished, knit your Quarto document to an HTML document and save it in the folder you created in step 2.
7. Submit your assignment by answering each question in Canvas and uploading the knitted HTML document to the designated course portal in Canvas before the due date and time.

Brief information to help you write your codes for each question

- In this lab assignment, you will first conduct exploratory data analysis, then use multiple linear regression method to predict your variable of interest. Also, you will check the model assumptions, check for outliers and influential factors, and finally do predictions.
- We have state level census data on various socio-economic and demographic data called **mydata**. The data consists of the following variables:

```r
mydata<-read.csv("Data_RLab2.csv", head=T)
names(mydata)
```

```
 [1] "State"                         "OwnComputer"                  
 [3] "CommutePublicTransport"        "TotalPopulation"              
 [5] "MedianAge"                     "WithCashAssistanceIncome"     
 [7] "MeanSocialSecurityIcnome"      "SupplementarySecurityIncome"  
 [9] "WhiteOnly"                     "Latinos"                      
[11] "Asians"                        "AfricanAmerican"              
[13] "Income100K.150K"               "Income75K.100K"               
[15] "Income50K.75K"                 "Income35K.50K"                
[17] "Income25K.35K"                 "PovertyRate"                  
```

- There are 52 observations and 18variables in the data. Some variables are presented in percentage points as a fraction of the total population. Below is a snapshot of our data.

State	OwnComputer	CommutePublicTransport	TotalPopulation	MedianAge	WithCashAssistanceIncome	MeanSocialSecurityIcnome	SupplementarySecurityIncome	Whi
Minnesota	90.3	198984.888	5527358	37.9	3.4	20192	4.2	
Mississippi	81.5	8966.286	2988762	37.2	2.2	17667	7.9	
Missouri	87.3	85260.868	6090062	38.5	1.9	19054	5.5	
Montana	87.3	8333.856	1041732	39.8	2.2	18696	4.4	
Nebraska	88.5	13333.320	1904760	36.4	1.9	19673	3.9	
Nevada	91.2	99376.866	2922849	37.9	3.0	19141	4.1	

- Our target variable is **OwnComputer**, the percentage of people who own a computer. It may not be an interesting question, yet, in this lab assignment, we will try to find the factors that determine our target variable.
- **model1** will be fit to **mydata** and it has the following predictors: **Asians**, **PovertyRate**, and **Income100K.150K**

$$Model\ \ 1: OwnComputer = \beta_0 + \beta_1 Asians + \beta_2 PovertyRate + \beta_3 Income100K.150K + \epsilon$$

- **Cook's distance** is a commonly used measure to identify influential points that have a large impact on the regression model. In this lab assignment, use a threshold Cook's Distance of 1 to identify the row numbers of any outlier and enter your answers in Canvas.
- Filter out the two observations in **mydata** that have a Cook's Distance greater than 1, and create a new dataset named **mydata1a** that excludes these outliers.
- **model1a** will be fit to **mydata1a** and it has the following predictors: **Asians**, **PovertyRate**, and **Income100K.150K**

- **model2** will be fit to **mydata1a** and it has the following predictors: **Asians**, **PovertyRate**, **Income100K.150K, Income25K.35K, SupplementarySecurityIncome, and WhiteOnly.**
- $Model\ \ 2: OwnComputer = \beta_0 + \beta_1 Asians + \beta_2 PovertyRate + \beta_3 Income100K.150K + \beta_4 Income25K.35K + \beta_5 SupplementarySecurityIncome +$
- Multicollinearity occurs when two or more independent variables in a regression model are highly correlated, which can result in unstable and unreliable estimates of the regression coefficients. We can check for multicollinearity by calculating the variance inflation factor (VIF). Any VIF value above 10 can be considered as an evidence of multi-collinearity.
- To construct **model3**, we exclude all predictors from **model2** that have a VIF value greater than 10.
- If you come across any instructions in this QMD file or a question in Canvas that you find confusing or unclear, please post your related questions in the '**Week 2 Questions in here!**' discussion forum.

Your code for Question 1

```
#cor(mydata$state, mydata$OwnComputer)
abs(cor(mydata$TotalPopulation, mydata$OwnComputer))
```

```
[1] 0.1161734
```

```
abs(cor(mydata$MeanSocialSecurityIcnome, mydata$OwnComputer))
```

```
[1] 0.24247
```

```
abs(cor(mydata$Latinos, mydata$OwnComputer))
```

```
[1] 0.1011996
```

```
abs(cor(mydata$Income100K.150K, mydata$OwnComputer))
```

```
[1] 0.5838208
```

```
abs(cor(mydata$Income35K.50K, mydata$OwnComputer))
```

```
[1] 0.5110413
```

```
abs(cor(mydata$MedianAge, mydata$OwnComputer))
```

```
[1] 0.2108768
```

```
abs(cor(mydata$SupplementarySecurityIncome, mydata$OwnComputer))
```

```
[1] 0.6515474
```

```
abs(cor(mydata$Asians, mydata$OwnComputer))
```

```
[1] 0.2683996
```

```
abs(cor(mydata$Income75K.100K, mydata$OwnComputer))
```

```
[1] 0.3322613
```

```
abs(cor(mydata$Income25K.35K, mydata$OwnComputer))
```

```
[1] 0.6500929
```

```
abs(cor(mydata$CommutePublicTransport, mydata$OwnComputer))
```

```
[1] 0.1048506
```

```
abs(cor(mydata$WithCashAssistanceIncome, mydata$OwnComputer))
```

```
[1] 0.3024275
```

```
abs(cor(mydata$WhiteOnly, mydata$OwnComputer))
```

```
[1] 0.1528747
```

```
abs(cor(mydata$AfricanAmerican, mydata$OwnComputer))
```

```
[1] 0.3025683
```

```
abs(cor(mydata$PovertyRate, mydata$OwnComputer))
```

```
[1] 0.3302821
```

Your code for Question 2

```
#Linear Regression Model
model1 <- lm(OwnComputer ~ Asians + PovertyRate + Income100K.150K, data = mydata)
summary(model1)
```

```

Call:
lm(formula = OwnComputer ~ Asians + PovertyRate + Income100K.150K, 
    data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.6545 -1.0858 -0.0344  1.0583  4.3388 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     64.17600    4.92178  13.039  < 2e-16 ***
Asians          -0.03232    0.04379  -0.738   0.4641    
PovertyRate      0.42075    0.12223   3.442   0.0012 ** 
Income100K.150K  1.27555    0.23524   5.422 1.89e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.146 on 48 degrees of freedom
Multiple R-squared:  0.4721,	Adjusted R-squared:  0.4391 
F-statistic: 14.31 on 3 and 48 DF,  p-value: 8.663e-07
```

Your code for Question 3

```
#show model1 coefficient, which is also shown above.
coef(model1)["PovertyRate"]
```

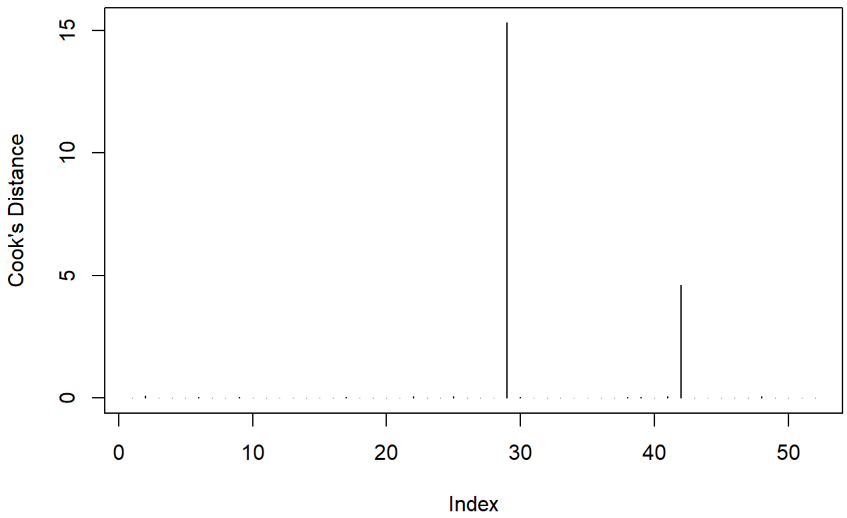
```
PovertyRate 
  0.4207497 
```

Your code for Question 4

```
#residual analysis

model1cookdis <- cooks.distance(model1)

#Use this script to plot the model.
plot(model1cookdis, type="h", ylab="Cook's Distance")
```



```
observation <- as.numeric(names(model1cookdis)[(model1cookdis>1.0)])

#This gives me the outliers
observation
```

```
[1] 29 42
```

Your code for Question 5

```
#This will create a new dataset called mydata1a that excludes outliers.

mydata1a <- mydata[-observation,]

#This will create model1a

#The predictors for model1a are: Asians, PovertyRate, and Income100K.150K

model1a <- lm(OwnComputer~ Asians + PovertyRate + Income100K.150K, data = mydata1a)

summary(model1a)
```

```

Call:
lm(formula = OwnComputer ~ Asians + PovertyRate + Income100K.150K, 
    data = mydata1a)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5527 -1.2783 -0.1263  0.7324  3.5709 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     87.10517    5.00593  17.400   <2e-16 ***
Asians           0.24710    0.09073   2.724   0.0091 ** 
PovertyRate     -0.41406    0.15584  -2.657   0.0108 *  
Income100K.150K  0.40618    0.22052   1.842   0.0719 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.571 on 46 degrees of freedom
Multiple R-squared:  0.7236,	Adjusted R-squared:  0.7055 
F-statistic: 40.14 on 3 and 46 DF,  p-value: 6.759e-13
```

Your code for Question 6

```
#This will Extract the Adjusted R-squared for both models 1 and 1a.
summary(model1)$adj.r.squared
```

```
[1] 0.4390665
```

```
summary(model1a)$adj.r.squared
```

```
[1] 0.7055421
```

```
#According to the adjusted R-square value, which model does a better job to explain OwnComputer? Model1a does a better job because it is higher.
```

Your code for Question 7

```
#This will create a new model called model2. Model2 will be fit to mydata1a and it has the following predictors: Asians, PovertyRate, Income100K.150K,

model2 <- lm(OwnComputer~ Asians + PovertyRate + Income100K.150K + Income25K.35K + SupplementarySecurityIncome + WhiteOnly, data = mydata1a)

summary(model2)
```

```

Call:
lm(formula = OwnComputer ~ Asians + PovertyRate + Income100K.150K + 
    Income25K.35K + SupplementarySecurityIncome + WhiteOnly, 
    data = mydata1a)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.66806 -1.00154 -0.01902  0.59293  2.93675 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 76.423639  12.744798   5.996  3.7e-07 ***
Asians                       0.295970   0.106085   2.790  0.00783 ** 
PovertyRate                 -0.024397   0.257830  -0.095  0.92505    
Income100K.150K              0.780048   0.421143   1.852  0.07087 .  
Income25K.35K                0.403698   0.557218   0.724  0.47269    
SupplementarySecurityIncome -0.888772   0.302438  -2.939  0.00528 ** 
WhiteOnly                    0.007306   0.030155   0.242  0.80972    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.388 on 43 degrees of freedom
Multiple R-squared:  0.7984,	Adjusted R-squared:  0.7702 
F-statistic: 28.38 on 6 and 43 DF,  p-value: 1.924e-13
```

Your code for Question 8

```
#This will show the variance inflation factor(VIF) estimates for the predictors of model2.PovertyRate, Income100K.150K, and Income25K.35K show a sign

vif(model2)
```

```
                     Asians                 PovertyRate 
                   2.725824                   14.291210 
            Income100K.150K               Income25K.35K 
                  23.946871                   16.370464 
SupplementarySecurityIncome                   WhiteOnly 
                   3.274392                    5.136633 
```

Your code for Question 9

```
#The following code will construct model3, and exclude all predictors from model2 that have a VIF value greater than 10

model3 <- lm(OwnComputer~ Asians + SupplementarySecurityIncome + WhiteOnly, data = mydata1a)

summary(model3)
```

```
Call:
lm(formula = OwnComputer ~ Asians + SupplementarySecurityIncome + 
    WhiteOnly, data = mydata1a)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5913 -1.2602 -0.2678  1.2015  3.8231 

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 91.02986    2.04135  44.593  < 2e-16 ***
Asians                       0.54319    0.09169   5.924 3.76e-07 ***
SupplementarySecurityIncome -1.46027    0.19816  -7.369 2.56e-09 ***
WhiteOnly                    0.03948    0.01884   2.096   0.0417 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.61 on 46 degrees of freedom
Multiple R-squared:  0.7098,	Adjusted R-squared:  0.6909 
F-statistic:  37.5 on 3 and 46 DF,  p-value: 2.051e-12
```

Your code for Question 10

```
# The following code will extract all models (1, 1a, 2, & 3) Adjusted-R squared value . Model 2 performs the best befcause it has the highest Adjusted
summary(model1)$adj.r.squared
```

```
[1] 0.4390665
```

```
summary(model1a)$adj.r.squared
```

```
[1] 0.7055421
```

```
summary(model2)$adj.r.squared
```

```
[1] 0.7702387
```

```
summary(model3)$adj.r.squared
```

```
[1] 0.6908607
```

Your code for Question 11

Consider the following scenario: Canada held a referendum to become the 51st state of the United States, and the US accepted their request with pleasure.”

Use **model2** to predict the **OwnComputer** ratio in Canada with a 90% prediction interval.

Hypothetical Data for Canada:

Asians: 18.4

PovertyRate: 5.8

Income100K.150K: 23

Income25K.35K: 13

SupplementarySecurityIncome: 9

WhiteOnly: 75

```
#This is creating a new dataframe called Canada, with six different columns.
Canada<-data.frame(Asians=18.4,PovertyRate=5.8,Income100K.150K=23,Income25K.35K= 13,SupplementarySecurityIncome= 9,WhiteOnly=75)

#This code will predict an outcome variable and a 90% confidence internval.
predict.lm(model2, Canada, interval = "predict", level=0.90)
```

```
       fit      lwr      upr
1 97.46618 87.73019 107.2022
```