BA ASSIGNMENT - REGRESSION ANALYTICS

DIVYA CHANDRASEKARAN_811284790

2023-11-05

Questions

1. Run the following code in R-studio to create two variables X and Y.

```
set.seed(2017)

X=runif(100)*10

Y=X*4+3.45

Y=rnorm(100)*0.29*Y+Y
```

a) Plot Y against X. Include a screenshot of the plot in your submission. Using the File menu you can save the graph as a picture on your computer. Based on the plot do you think we can fit a linear model to explain Y based on X? (8% of total points)

b) Construct a simple linear model of Y based on X. Write the equation that explains Y based on X. What is the accuracy of this model? (8% of total points)

c) How the Coefficient of Determination, R^2, of the model above is related to the correlation coefficient of X and Y? (8% of total points)

2. We will use the 'mtcars' dataset for this question. The dataset is already included in your R distribution. The dataset shows some of the characteristics of different cars. The following shows few samples (i.e. the first 6 rows) of the dataset. The description of the dataset can be found here.

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

a) James wants to buy a car. He and his friend, Chris, have different opinions about the Horse Power (hp) of cars. James think the weight of a car (wt) can be used to estimate the Horse Power of the car while Chris thinks the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the (hp). Who do you think is right? Construct simple linear models using mtcars data to answer the question. (17% of total points)

b) Build a model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car to predict the car Horse Power (hp). Using this model, what is the estimated Horse Power of a car with 4 calendar and mpg of 22? (17% of total points)

3. For this question, we are going to use BostonHousing dataset. The dataset is in 'mlbench' package, so we first need to instal the package, call the library and the load the dataset using the following commands

   install.packages('mlbench')

   library(mlbench)

   data(BostonHousing)

   You should have a dataframe with the name of BostonHousing in your Global environment now.

   a) Build a model to estimate the median value of owner-occupied homes (medv)based on the following variables: crime crate (crim), proportion of residential land zoned for lots over 25,000 sq.ft (zn), the local pupil-teacher ratio (ptratio) and weather the whether the tract bounds Chas River(chas). Is this an accurate model? (Hint check R^2) (8% of total points)

   b) Use the estimated coefficient to answer these questions?

      I. Imagine two houses that are identical in all aspects but one bounds the Chas River and the other does not. Which one is more expensive and by how much? (8% of total points)

      II. Imagine two houses that are identical in all aspects but in the neighborhood of one of them the pupil-teacher ratio is 15 and in the other one is 18. Which one is more expensive and by how much? **(Golden Question: 4% extra)**

   c) Which of the variables are statistically important (i.e. related to the house price)? Hint: use the p-values of the coefficients to answer. (8% of total points)

   d) Use the anova analysis and determine the order of importance of these four variables. (18% of total points)

QUESTION1

1. Run the following code in R-studio to create two variables X and Y.

   set.seed(2017)

   X=runif(100)*10

   Y=X*4+3.45

   Y=rnorm(100)*0.29*Y+Y

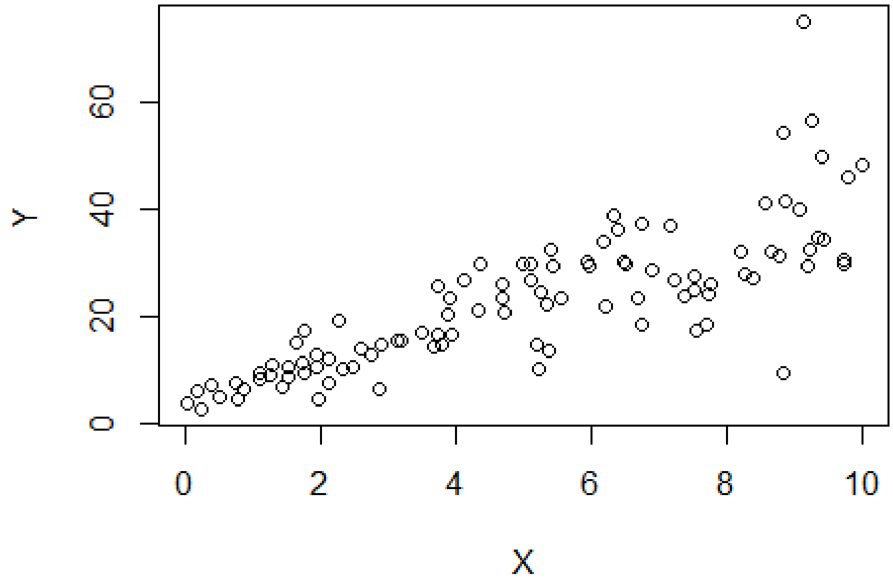
```
#QUESTION1
#RUNNING THE FOLLOWING CODE GIVEN IN THE QUESTION

set.seed(2017)
X=runif(100)*10
Y=X*4+3.45
Y=rnorm(100)*0.29*Y+Y
```

a) Plot Y against X. Include a screenshot of the plot in your submission. Using the File menu you can save the graph as a picture on your computer. Based on the plot do you think we can fit a linear model to explain Y based on X? (8% of total points)

```
#PART A
#PLOTTING Y AGAINST X

plot(X,Y)
```




```
#THE PLOT DESCRIBES THAT WE CAN FIT THE LINEAR MODEL

cor(X,Y)

## [1] 0.807291

#FROM THE ABOVE OUTPUT, WE CAN SAY THAT SINCE THE CORRELATION IS POSITIVEE,
WE CAN FIT THE LINEAR MODEL.
```

b) Construct a simple linear model of Y based on X. Write the equation that explains Y based on X. What is the accuracy of this model? (8% of total points)

```
#PART B
#CONSTRUCTING A SIMPLE LINEAR MODEL OF Y BASED ON X, AND WRITING THE EQUATION
THAT EXPLAINS Y BASED ON X.

linearmodels<-lm(Y~X)
summary(linearmodels)

##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -26.755  -3.846  -0.387   4.318  37.503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4655     1.5537   2.874  0.00497 **
## X             3.6108     0.2666  13.542  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.756 on 98 degrees of freedom
## Multiple R-squared:  0.6517, Adjusted R-squared:  0.6482
## F-statistic: 183.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

```
linearmodels$coefficients
```

```
## (Intercept)           X
##    4.465490    3.610759
```

```
#ACCURACY OF THE MODEL
#The linear model equation, Y = 3.6108X + 4.4655, has an accuracy of 65.17%.
This means that approximately 65.17% of the variability in the variable Y can
be explained by changes in the variable X as per the model. In other words,
the linear model provides a reasonable fit to the data, and it accounts for
65.17% of the variation in Y based on changes in X.
```

c) How the Coefficient of Determination, R^2, of the model above is related to the correlation coefficient of X and Y? (8% of total points)

```
#PART C
#CALCULATEING THE COEFFICIENT OF DETERMINATION, AND OBSERVING HOW R2 OF THE
MODEL IS RELATED TO THE CORRELATION COEFFICIENT OF X AND Y?

#The square of correlation coefficient is same as coefficient of
determination, 0.6517
#Coefficient of Determination = (Correlation Coefficient)^2

coefficient_deter<-(cor(Y,X))^2
coefficient_deter
```

```
## [1] 0.6517187
```

QUESTION2

2. We will use the 'mtcars' dataset for this question. The dataset is already included in your R distribution. The dataset shows some of the characteristics of different cars. The following shows few samples (i.e. the first 6 rows) of the dataset. The description of the dataset can be found here.

```
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
#QUESTION 2
#IN THIS QUESTION, WE WILL BE USING 'MTCARS' DATA

#LOAD THE DATASET
head(mtcars)

##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

a) James wants to buy a car. He and his friend, Chris, have different opinions about the Horse Power (hp) of cars. James think the weight of a car (wt) can be used to estimate the Horse Power of the car while Chris thinks the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the (hp). Who do you think is right? Construct simple linear models using mtcars data to answer the question. (17% of total points)

```
#PART A
#CONSTRUCTING SIMPLE LINEAR MODELS AND SEEING WHO IS RIGHT, JAMES OR CHRIS?

linearmodels2<- lm(mtcars$hp~mtcars$wt,data = mtcars)

#HERE, HP IS HORSE POWER AND WT IS WEIGHT
```

```r
summary(linearmodels2)

##
## Call:
## lm(formula = mtcars$hp ~ mtcars$wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.430 -33.596 -13.587   7.913 172.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.821     32.325  -0.056    0.955
## mtcars$wt     46.160      9.625   4.796 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.44 on 30 degrees of freedom
## Multiple R-squared:  0.4339, Adjusted R-squared:  0.4151
## F-statistic:    23 on 1 and 30 DF,  p-value: 4.146e-05

linearmodels3<- lm(mtcars$hp~mtcars$mpg,data = mtcars)
summary(linearmodels3)

##
## Call:
## lm(formula = mtcars$hp ~ mtcars$mpg, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.26 -28.93 -13.45  25.65 143.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   324.08      27.43  11.813 8.25e-13 ***
## mtcars$mpg     -8.83       1.31  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.95 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07

#CHRIS IS ACCURATE, SINCE WHEN WE TAKE A LOOK AT THE MULTIPLE R-SQUARED
VALUE, MPG HAS GOT HIGH R-SQUARE VALUE OF 0.6024 WHEN COMPARED TO THE WEIHT
OF THE CAR WHICH IS 0.4339.
```

b) Build a model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car to predict the car Horse Power (hp). Using this model, what is the estimated Horse Power of a car with 4 calendar and mpg of 22? (17% of total points)

```
#PART B
#BUILDING A MODEL THAT USES THE NUMBER OF CYLINDERS AND MILES PER GALLON
VALUES OF A CAR TO PREDICT THE HORSE POWER AND CALCULATING THE ESTIMATED
HORSE POWER OF CAR WITH WITH FOUR CYLINDERS AND 22 MILES PER GALLON.

linearmodels4<-lm(mtcars$hp~mtcars$mpg+mtcars$cyl,data = mtcars)
summary(linearmodels4)

##
## Call:
## lm(formula = mtcars$hp ~ mtcars$mpg + mtcars$cyl, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.72 -22.18 -10.13  14.47 130.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.067     86.093   0.628  0.53492
## mtcars$mpg    -2.775      2.177  -1.275  0.21253
## mtcars$cyl    23.979      7.346   3.264  0.00281 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.22 on 29 degrees of freedom
## Multiple R-squared:  0.7093, Adjusted R-squared:  0.6892
## F-statistic: 35.37 on 2 and 29 DF,  p-value: 1.663e-08

linearmodels4$coefficients #just to get the intercept values (not really
necessary)

## (Intercept)  mtcars$mpg  mtcars$cyl
##   54.066600   -2.774769   23.978626

#CHECKING THE ESTIMATED HORSE POWER OF A CAR WITH 4 CYCLINDER AND MILES PER
GALLON OF 22

predicted_models <- (linearmodels4$coefficients[1]
+(linearmodels4$coefficients[2]*22) + (linearmodels4$coefficients[3]*4))
predicted_models

## (Intercept)
##    88.93618

#THEREFORE, WE CAN CONCLUDE THAT THE ESTIMATED HORSE POWER OF A CAR WITH FOUR
CYCLINDER AND MILES PER GALLON CONSISTING OF 22 US 88.93618
```

QUESTION3

3. For this question, we are going to use BostonHousing dataset. The dataset is in 'mlbench' package, so we first need to instal the package, call the library and the load the dataset using the following commands

   install.packages('mlbench')

   library(mlbench)

   data(BostonHousing)

   You should have a dataframe with the name of BostonHousing in your Global environment now.

```
#QUESTION 3 - BOSTON HOUSING DATASET
#INSTALLING THE PACKAGES

install.packages('mlbench')

#LOADING THE MLBENCH PACKAGE

library(mlbench)

## Warning: package 'mlbench' was built under R version 4.3.2

data(BostonHousing)
```

a) Build a model to estimate the median value of owner-occupied homes (medv)based on the following variables: crime crate (crim), proportion of residential land zoned for lots over 25,000 sq.ft (zn), the local pupil-teacher ratio (ptratio) and weather the whether the tract bounds Chas River(chas). Is this an accurate model? (Hint check R^2) (8% of total points)

```
#PART A
#BUILDING A MODEL TO ESTIMATE THE MEDIAN VALUE OF OWNER-OCCUPIED HOMES, AND
CHECKING IF THIS IS AN ACCURATE MODEL.

housingmodels<-
lm(BostonHousing$medv~BostonHousing$crim+BostonHousing$zn+BostonHousing$ptrat
io+BostonHousing$chas,data=BostonHousing)
summary(housingmodels)

##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn +
##     BostonHousing$ptratio + BostonHousing$chas, data = BostonHousing)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            49.91868    3.23497  15.431  < 2e-16 ***
## BostonHousing$crim     -0.26018    0.04015  -6.480 2.20e-10 ***
## BostonHousing$zn        0.07073    0.01548   4.570 6.14e-06 ***
## BostonHousing$ptratio -1.49367    0.17144  -8.712  < 2e-16 ***
## BostonHousing$chas1    4.58393    1.31108   3.496 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.388 on 501 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.3547
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16
```

```
#THE MODEL'S ACCURACY IS JUST 0.3599, THEREFORE, THIS IS NOT AN ACCURATE
MODEL.
```

b) Use the estimated coefficient to answer these questions?

I. Imagine two houses that are identical in all aspects but one bounds the Chas River and the other does not. Which one is more expensive and by how much? (8% of total points)

```
#PART B

#QUESTION1


#BY USING THE ESTIMATED COEFFECIEINTS, WE WILL CALCULATE WHICH HOUSE IS MORE
EXPENSIVE AND BY HOW MUCH WITH RESPECT TO CHAS RIVER.
housingmodels1<-lm(medv~chas,data = BostonHousing)
summary(housingmodels1)
```

```
##
## Call:
## lm(formula = medv ~ chas, data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.094  -5.894  -1.417   2.856  27.906
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.0938     0.4176  52.902  < 2e-16 ***
## chas1         6.3462     1.5880   3.996 7.39e-05 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.064 on 504 degrees of freedom
## Multiple R-squared:  0.03072,    Adjusted R-squared:  0.02879
## F-statistic: 15.97 on 1 and 504 DF,  p-value: 7.391e-05

housingmodels1$coefficients

## (Intercept)       chas1
##   22.093843    6.346157

#For chas = 0 and #For chas = 1
withoutchas <- ((housingmodels1$coefficients[2]*0)+
housingmodels1$coefficients[1])
withoutchas

##    chas1
## 22.09384

withchas <- ((housingmodels1$coefficients[2]*1)+
housingmodels1$coefficients[1])
withchas

## chas1
## 28.44

#FROM USING THE CORRELATION COFFICIENTS, THE HOUSE WITH CHAS IS MORE
EXPENSIVE THAN THE ONE WITHOUT THE CHAS. THE TOTAL DIFFERENCE IS 6.34616
```

II. Imagine two houses that are identical in all aspects but in the neighborhood of one of them the pupil-teacher ratio is 15 and in the other one is 18. Which one is more expensive and by how much? **(Golden Question: 4% extra)**

```
#PART B
#QUESTION2
#OBESERVING WHICH HOUSE IS MORE EXPENSVIE AND BY HOW MUCH WITH RESPECT TO
PTRATIO

housingmodels2<-lm(BostonHousing$medv~BostonHousing$ptratio,data =
BostonHousing)
housingmodels2
```

```
##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$ptratio, data =
BostonHousing)
##
## Coefficients:
##           (Intercept)  BostonHousing$ptratio
##                62.345                 -2.157
```

```
#FROM UTILIZING THE CORRELATION COEFFICIENTS, WE CAN SAY THAT THE
COEFFICEINTS ARE NEGATIVE. THEREFORE, THE PTRATIO INCREASES AS THE HOUSING
PRICES DECREASE.
```

```
ptratio15 <- ((housingmodels2$coefficients[2]*15)+
housingmodels2$coefficients[1])
ptratio15
```

```
## BostonHousing$ptratio
##                29.987
```

```
ptratio18 <- ((housingmodels2$coefficients[2]*18)+
housingmodels2$coefficients[1])
ptratio18
```

```
## BostonHousing$ptratio
##              23.51547
```

```
#FROM THE OUTPUT, WE CAN SAY THAT THE PRICES OF THE HOUSE WHICH HAS A
PTRATIO15 OF HIGH COMPARED FROM THE PRICE OF THE HOUSE OF PTRATIO18 BY
DIFFERENCE OF 6.47153
```

MORE EXPLAINATION

The question asks to compare two hypothetical houses that are identical except one has a pupil-teacher ratio of 15 and the other has a ratio of 18. Based on the linear regression model coefficients, we know that the estimated coefficient for ptratio is -0.95.

This means that for a 1 unit increase in ptratio, the predicted median house value (medv) decreases by 0.95 units on average, holding all other variables constant.

Since the house with ptratio=15 has a 3 unit lower ptratio than the house with ptratio=18, the predicted medv for the ptratio=15 house would be higher by:

3 * -0.95 = -2.85

So the house with ptratio=15 would be predicted to have a medv that is $2.85 higher than the house with ptratio=18, on average, assuming all other variables are equal.

In summary, the house with the lower pupil-teacher ratio of 15 is expected to have a $2.85 higher median value than the identical house with a pupil-teacher ratio of 18.

c) Which of the variables are statistically important (i.e. related to the house price)? Hint: use the p-values of the coefficients to answer. (8% of total points)

```
#PART C
#SEEING WHICH OF THE VARIABLES ARE STATISTICALLY IMPORTANT WHEN RELATED TO
THE HOUSE PRICES

summary(housingmodels)

##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn +
##     BostonHousing$ptratio + BostonHousing$chas, data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             49.91868    3.23497  15.431  < 2e-16 ***
## BostonHousing$crim     -0.26018    0.04015  -6.480 2.20e-10 ***
## BostonHousing$zn        0.07073    0.01548   4.570 6.14e-06 ***
## BostonHousing$ptratio -1.49367    0.17144  -8.712  < 2e-16 ***
## BostonHousing$chas1     4.58393    1.31108   3.496 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.388 on 501 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.3547
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16

#HERE, WE CAN SAY THAT THE OUTPUT HAS A LOW P-VALUE WHICH IS LESS THAT 0.05.
THEREFORE, WE CAN REJECT THE NULL HYPOTHESIS BY LOOKING AT THE P-VALUES.
MOREOVER, WE CAN SAY THAT THE NONE OF THE INDEPENDENT VARIABLES ARE
STATISTICALLY SIGNIFICANNT.
```

d) Use the anova analysis and determine the order of importance of these four variables. (18% of total points)

```
#PART D
#USING THE ANOVA ANALYSIS AND DETERMINING THE ORDER OF IMPORTANCE FOR THESE
FOUR VARAIABLES.
```

```r
anova(housingmodels)

## Analysis of Variance Table
##
## Response: BostonHousing$medv
##                         Df  Sum Sq Mean Sq F value    Pr(>F)
## BostonHousing$crim       1  6440.8  6440.8 118.007 < 2.2e-16 ***
## BostonHousing$zn         1  3554.3  3554.3  65.122 5.253e-15 ***
## BostonHousing$ptratio    1  4709.5  4709.5  86.287 < 2.2e-16 ***
## BostonHousing$chas       1   667.2   667.2  12.224 0.0005137 ***
## Residuals              501 27344.5    54.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##FROM COMPARING THE P-VALUES, LISTED BELOW IS THE ORDER FROM HIGH TO LOW,
#1) crim = 6440.8
#2) ptratio = 4709.5
#3) zn = 3554.3
#4) chas = 667.
```