BA64036\_Assignment-3

Durga Prasad Gandi

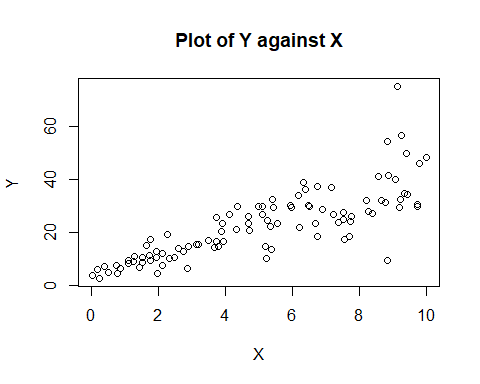
2023-11-03

#### Question1. 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

**1a) Plotting of Y against X**

plot(X ,Y ,main="Plot of Y against X", xlab = "X", ylab = "Y")



**1b) 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?**

Linear model

linear\_m <- lm(Y ~ X)  
  
summary(linear\_m)

##   
## 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

***Explanation of Y based on X is:***

Residual is the difference between observed values of Y and predicted values of the given model i.e error (7.756).

Coefficients: When (X) is zero the value of (Y) intercept is (4.4655) and the coefficient for X is 3.6108

Equation:

Accuracy:Accuracy of the given model is 65.17%

**1c) How the Coefficient of Determination, R2, of the model above is related to the correlation coefficient of X and Y?**

Correlation coefficient between X and Y:

cor(X, Y)

## [1] 0.807291

summary(linear\_m)

##   
## 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

R^2 = 0.6517 and r^2 = (0.807291)^2 = 0.6571 From above R^2 is equal to r^2 where the condition in linear regression model is satisfied.

#### Question 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.

***Load the dataset mtcars***

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

*Construct a simple linear model to estimate hp using wt*

lm\_wt <- lm(hp ~ wt, data = mtcars)  
  
lm\_wt

##   
## Call:  
## lm(formula = hp ~ wt, data = mtcars)  
##   
## Coefficients:  
## (Intercept) wt   
## -1.821 46.160

*Construct a simple linear model to estimate hp using mpg*

lm\_mpg <- lm(hp ~ mpg, data = mtcars)  
  
lm\_mpg

##   
## Call:  
## lm(formula = hp ~ mpg, data = mtcars)  
##   
## Coefficients:  
## (Intercept) mpg   
## 324.08 -8.83

*Compare the summary statistics of the above two models*

summary(lm\_wt)

##   
## Call:  
## lm(formula = hp ~ 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   
## 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

summary(lm\_mpg)

##   
## Call:  
## lm(formula = hp ~ 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 \*\*\*  
## 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

***Which is the Best model of the above:***

It seems that the model that uses Miles Per Gallon (mpg) as a predictor fits the’mtcars’ dataset better for estimating Horsepower (hp) than the model that uses Weight (wt), based on the R-squared values and the statistical significance of the coefficients. A better correlation between horsepower and mpg is indicated by the higher R-squared value (0.6024), and the coefficient for mpg is statistically significant.

Based on the information at this point, Chris appears to be more accurate in this study, since he feels that Miles Per Gallon (mpg) is a better indicator of Horsepower (hp).

#### 2b) 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?

*Create a multiple linear regression model*

multiple\_model <- lm(hp ~ cyl + mpg, data = mtcars)

*Make predictions*

new\_data <- data.frame(cyl = 4, mpg = 22)  
estimated\_hp <- predict(multiple\_model, newdata = new\_data)  
estimated\_hp

## 1   
## 88.93618

The Estimated horse power for 4 cylinders and 22 MPG is 88.93618

#### Question3. 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 and call the mlbench package and load the Bostonhousing data set:**

library(mlbench)

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

data(BostonHousing)  
BostonHousing$chas = factor(BostonHousing$chas, levels = c(0, 1))

#### 3(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?

***Construct multiple linear regression model***

homes\_model <- lm(medv ~ crim + zn + ptratio + chas, data = BostonHousing)  
summary(homes\_model)

##   
## Call:  
## lm(formula = medv ~ crim + zn + ptratio + 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 \*\*\*  
## crim -0.26018 0.04015 -6.480 2.20e-10 \*\*\*  
## zn 0.07073 0.01548 4.570 6.14e-06 \*\*\*  
## ptratio -1.49367 0.17144 -8.712 < 2e-16 \*\*\*  
## 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 R-squared value of the model is roughly 0.3599. This figure indicates the percentage of the variance in median home prices (medv) that can be attributed to the variables in the model. Although the model accounts for a portion of the variance, it indicates that housing values may be better explained.

Statistically significant coefficients for ‘crim,’ ‘zn,’ ‘ptratio,’ and ‘chas1’ are all present (low p-values); these results imply that these variables influence median house values. This is one of the model’s advantages.

At 0.3547, the corrected R-squared value is somewhat less than the unadjusted It implies that, given the amount of predictors, even while the model explains some variance, it might not offer a very good fit.

#### 3(b) 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?

summary(linear\_m)

##   
## 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

According to the result of our model, the coefficient for ‘chas1’ is 4.58393 so that the median house value for houses on the Charles River is estimated to be $4,583.93 higher when compared to other houses not on river.Hence, the house on the Charles River is predicted to be more expensive.

#### 3(b2): 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?

Price\_Comparison = pt-ratio coefficient*(pupil-teacher (18)-pupil-teacher(15)) Price\_Comparison = -1.49367*  (18 -15) Price\_Comparison ≈ -4.48101

\*\*The coefficient of pupil-teacher ratio = -1.49367 in which for every unit change there will be decrease of approximately ($1,494). Hence,there will be increase of 3 units (yielding pupil-teacher ratio of 15 and 18 for the two houses). The values estimated indicates that the pupil-teacher ratio of 18 will be less expensive compared to that of pupil-teacher ratio of 15 ($1,494\*3)≈ $4,482.\*\*

#### 3(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.

You can look at the p-values that correspond to the coefficients of each independent variable in your linear regression model to find out which factors are statistically important in relation to house price. Lower p-values often denote statistical significance and imply a relationship between the respective variables and price of home.

*The p-values and the coefficients are used to rate their importance:*

Crime Rate:For ‘crim’ the p-value is 2.20e-10,that is too low i.e is <0.05. This indicates that the crime rate is statistically significant and related to price of the houses.

zn: For ‘zn’ the p-value is 6.14e-06 which is also less. It can be said that the proportion of land zoned for large lots is statistically significant and related to price of the houses.

Ptratio:For ‘ptratio’ the p-value is < 2e-16, that is very low. This indicates that the pt ratio is more significant and related to price of the houses.

Chas1: For ‘chas1’ the p-value is 0.000514, which is low. This tells us that whether the tract bounds the Charles River is important and related to price of the houses .

In Conclusion,p-values (crim, zn, ptratio, and chas1) for all the variables in this model,suggest that they are less important and related to price of the houses .

#### 3(D): Use the anova analysis and determine the order of importance of these four variables.

anova\_model<-anova(homes\_model)  
  
print(anova\_model)

## Analysis of Variance Table  
##   
## Response: medv  
## Df Sum Sq Mean Sq F value Pr(>F)   
## crim 1 6440.8 6440.8 118.007 < 2.2e-16 \*\*\*  
## zn 1 3554.3 3554.3 65.122 5.253e-15 \*\*\*  
## ptratio 1 4709.5 4709.5 86.287 < 2.2e-16 \*\*\*  
## 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

***Anova Analysis Interpretation***

The order of importance for the variables according to the ANOVA results:

Pupil-Teacher Ratio (ptratio): ptratio is first in place of importance that has the lowest p-value (p < 2.2e-16),and highest F value(118.007) and Mean Square value(6440.8) implying the most statistically significant variable and the most important in explaining house prices.

Crime Rate (crim): Crime rate also has almost near p values(p < 2.2e-16) as of ptratio and F value (86.287) and also mean square is 4709.5 implies second most important variable.

Proportion of Residential Land Zoned for Large Lots (zn): This variable is the third place in importance with F value (65.122) and p-value (p = 5.253e-15).

Whether the Tract Bounds the Charles River (chas1): It has the highest p-value among the four variables with(p = 0.0005137), and less F value(12.224) suggest that it is the less significant of the variables in describing price of the houses.

#### The list of importance of the variables are mentioned below:

1.Pupil-Teacher Ratio

2.Crime Rate

3.Proportion of Residential Land Zoned for Large Lots

4.Whether Tract Bounds the Charles River