Regression Analytics

Melissa Paniagua

11/1/2020

Assignment 2 | Module 5

This assignment integrates linear regression knowledge.

Linear regression is one of the widely used statistical methods to describe the linear relationship between variables, and it is used to predict the value of a variable (called dependent, response, or outcome) based on the values of another variable or set of variables (called independent, explanatory, or input).

The following is the linear regression equation:

$$y = \beta_0 + \beta_1 x + \epsilon$$

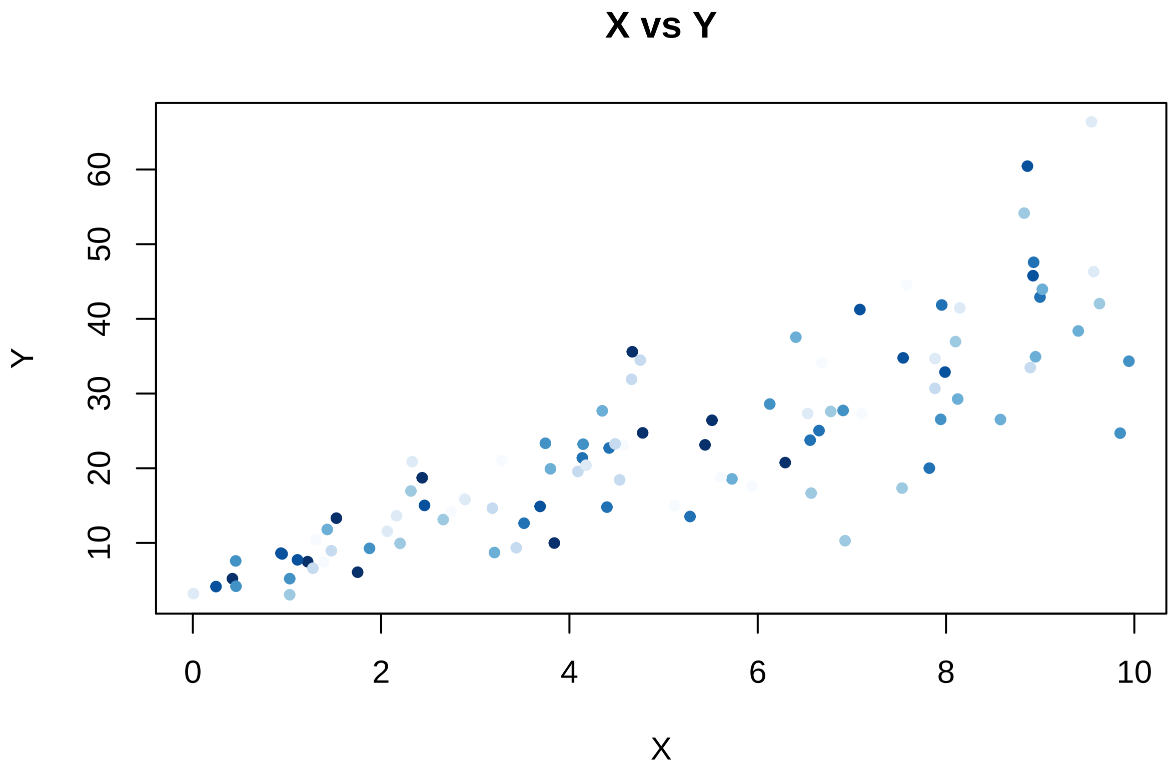
Questions

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

```r
# To get the same random values
set.seed(123)

# Create the following variables
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?

```r
#Plot personal Balance vs Age
plot(X , Y, main="X vs Y",
     xlab="X",
     ylab="Y",
     col = blues9,
     pch = 20)
```



Based on this scatter plot, we definitely can use X values to explain Y. Here we can see there is a positive relationship between X and Y because as the X values increase, Y values tend to increase as well.

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?

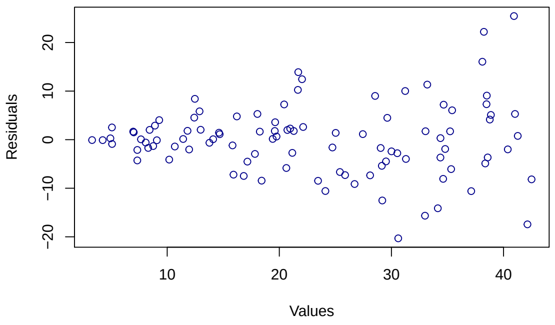
```
# To get the same random values
set.seed(123)

# To run the linear model
linear_model <- lm(Y ~ X)

# To get descriptive statistics of the model
summary(linear_model)
```

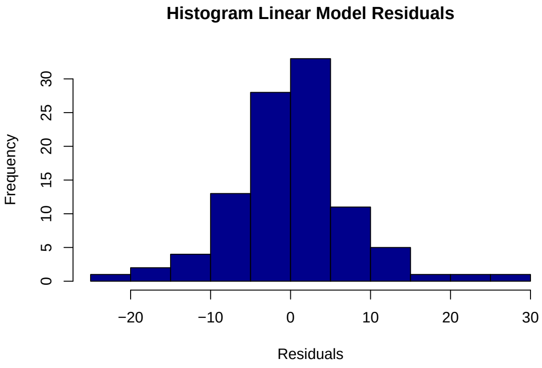
```
Call:
lm(formula = Y ~ X)

Residuals:
     Min       1Q   Median       3Q      Max
-20.3132  -4.0022   0.1144   3.0670  25.4482

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.2746     1.4828   2.208   0.0296 *
```

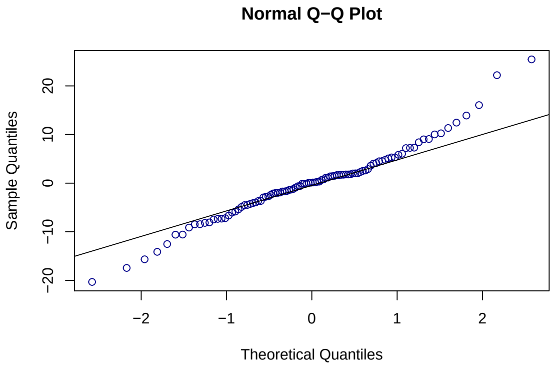
```
X              3.9452     0.2585  15.260   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.331 on 98 degrees of freedom
Multiple R-squared:  0.7038,    Adjusted R-squared:  0.7008
F-statistic: 232.9 on 1 and 98 DF,  p-value: < 2.2e-16
```

This output shows us a lot of valuable information. However, before looking at the details of this output, it is essential to see if the regression assumptions such as normal distribution, residual analysis, etc. are satisfied. To accomplish it, let's plot some valuable visuals such as QQ plot, residuals plot, and histogram.

```r
# Residuals plot
plot(linear_model$fitted.values, linear_model$residuals,
     xlab = "Values",
     ylab= "Residuals",
     col= "darkblue")

# Histogram plot
hist(linear_model$residuals,
     main="Histogram Linear Model Residuals",
     xlab="Residuals",
     col="darkblue")

# QQ plot
qqnorm(linear_model$residuals, col = "darkblue")
qqline(linear_model$residuals)
```







As we can see, the residuals graph might show a little shape. Regarding the QQ Plot, we can see that most of the data points are in the line, so it satisfies the normality assumption. Based on the histogram, we could say that the data is normally distributed.

Now that we can see that the assumptions of the regression model are satisfied, we can make conclusions from the summary table. Here we can see that the p-value for X is very small, which means that the hypothesis is true, so we will reject the hypothesis and we conclude that the X variable is statistically significant.

Regarding the accuracy of the model, we can see that R^2 is 70.38%, which we can affirm it is highly accurate.

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

Here is the coefficient of determination equation:

$$R^2 = 1 - \frac{RSS}{TSS}$$

In other words, $R^2 = Coefficient\,of\,Determination = (Correlation\,Coefficient)^2$.

```r
# R
cor(X,Y)
```

```
[1] 0.8389348
```

```r
# R^2
cor(X,Y)^2
```

```
[1] 0.7038116
```

As we learned in this course, the coefficient of determination shows the proportion of the variability of the dependent variable (y) to the independent one (x), and it is a number between 0 and 1. Regarding the Coefficient of Correlation, it shows the relationship between two variables (x and y), and it is a number between -1 and 1. In our model, R is 83.89% and R^2 is 70.38%, which means the correlation coefficient of X and Y are related.

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.

```r
# To show the first 6 rows of the data frame
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.

Let’s analyze James’s point of view

```
# To get the same random values
set.seed(123)

# To run the linear model
james_model <- lm(mtcars$hp ~ mtcars$wt)

# To get descriptive statistics of the model
summary(james_model)
```

```
Call:
lm(formula = mtcars$hp ~ mtcars$wt)

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
```

Now, let’s study Chris’s opinion

```
# To get the same random values
set.seed(123)

# To run the linear model
chris_model <- lm(mtcars$hp ~ mtcars$mpg)

# To get descriptive statistics of the model
summary(chris_model)
```

```
Call:
lm(formula = mtcars$hp ~ mtcars$mpg)

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
```

Based on these two summaries, we can determine that the P-value if the weight of the car and the mile per gallon are very small, which means those variables are statistically significant. However, the R^2 of Jame’s model is only 43.39%, which is too low. On the other hand, Chris’s statement that miles per gallon are a better predictor of horse car’s power is right because this linear regression model shows higher accuracy than James’ opinion, by giving an accuracy of 60.24% compare to 43.39%.

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?

```r
# To get the same random values
set.seed(123)

# To run the linear model
cyl_mpg_model <- lm(hp ~ cyl + mpg, data = mtcars)

# To get descriptive statistics of the model
summary(cyl_mpg_model)
```

```
Call:
lm(formula = hp ~ cyl + mpg, 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
cyl           23.979      7.346   3.264  0.00281 **
mpg           -2.775      2.177  -1.275  0.21253
---
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
```

Now, let’s build the prediction

```r
# To get the same random values
set.seed(123)

# Predicted model
```

```
predicted_hp <- predict(cyl_mpg_model, newdata =  data.frame(cyl = 4, mpg = 22))

#To see the result
predicted_hp
```

```
       1
88.93618
```

Based on four cylinders and 22 miles per gallon, the predicted of car Horse Power (hp) is 88.93618.

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:

```
# To load the library
library(mlbench)

# To load the dataset
data(BostonHousing)

#To see the values of the data frame
head(BostonHousing)
```

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

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

The dataset contains information about houses in different parts of Boston. Details of the dataset is explained here. Note the dataset is old, hence low house prices!

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 R2)

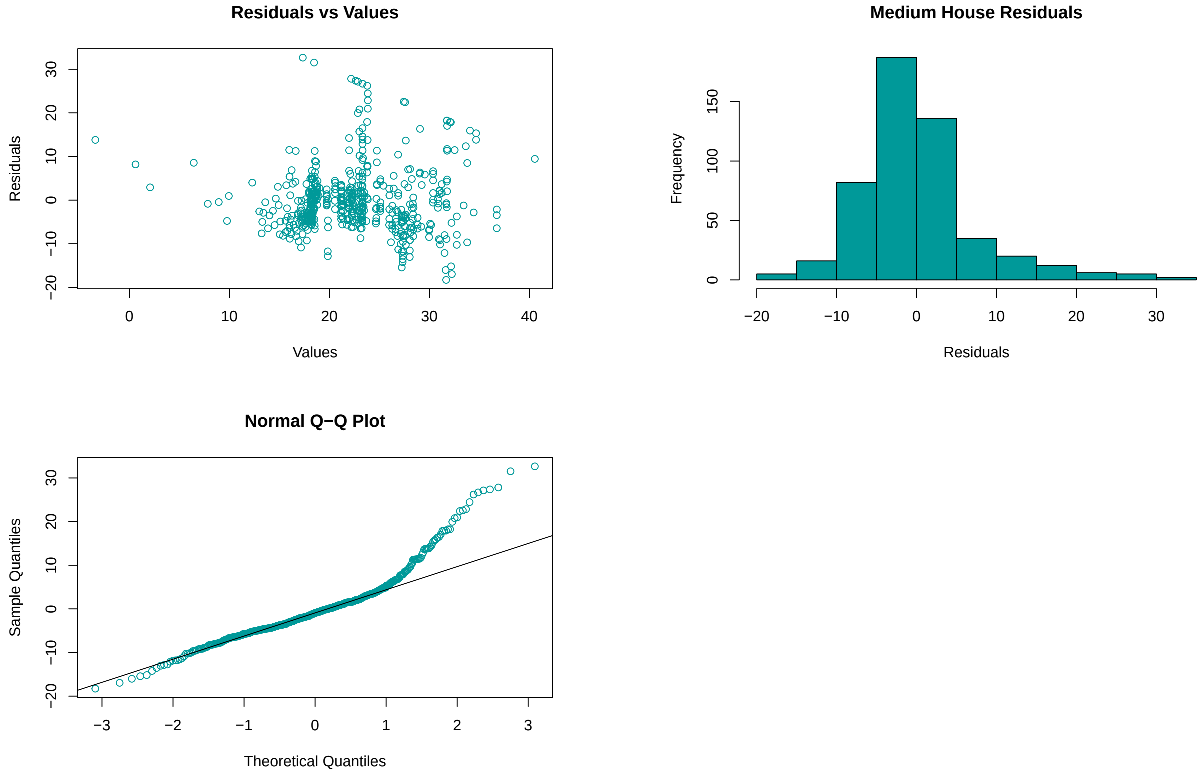
```
# To get the same random values
set.seed(123)

# To run the linear model
medium_house_model <- lm(BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn
                         + BostonHousing$ptratio + BostonHousing$chas)

# To get descriptive statistics of the model
summary(medium_house_model)
```

```
Call:
```

```
lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn +
    BostonHousing$ptratio + BostonHousing$chas)

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
```

Besides this summary output, let’s plot the QQ plot, residuals plot, and histogram to analyze the assumptions of the regression model.

```r
# Residuals plot
plot(medium_house_model$fitted.values, medium_house_model$residuals,
     main="Residuals vs Values",
     xlab = "Values",
     ylab= "Residuals",
     col= "#009999")

# Histogram plot
hist(medium_house_model$residuals,
      main="Medium House Residuals",
      xlab = "Residuals",
      ylab= "Frequency",
      col="#009999")

# QQ plot
qqnorm(medium_house_model$residuals, col = "#009999")
qqline(medium_house_model$residuals)
```



Based on the regression assumptions plots and the summary output we cannot trust this model. We can see how the histogram has a right-skewed tail, the residual scatterplot has its data points concentrated on a specific area, and the QQ Plot is not showing the values on the line. Additionally, even though the p-values are small and show that those variables are statistically significant, the R^2 is performing too bad, with 35.99% accuracy.

b) Use the estimated coefficient to answer these questions?

Extract the coefficients from the model

```
# To get the coefficients
medium_house_model$coefficients
```

```
          (Intercept)     BostonHousing$crim       BostonHousing$zn
          49.91868439            -0.26017612             0.07072809
BostonHousing$ptratio   BostonHousing$chas1
          -1.49367255             4.58392591
```

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?

As we learned in lecture VI, the coefficient values do not tell much, what is important is the sign those values have.

As we can see, the coefficient value for the Chas River(chas) variable is positive. It means that the price of a house that has bounds with the Chas River will have higher price than the house that does not. Its price will increase by 4.58392591.

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?

Under this scenario, the coefficient value for pupil-teacher ratio (ptratio) is negative. It means that the house with a ratio of 15 will be more expensive than the house with a ratio of 18. The house's price will increase -1.49367255 per unit, so in our example, the price will increase by 4.481018 to the total price.

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.

The P-values, shown below, help us to determine if a variable is statistically significant. Based on the regression model of the medium value of an occupied house, we can see that those p-values are very small, which confirms these variables are statistically significant and help to predict the price value of a house. Nevertheless, even though this is a good sign, we cannot completely trust them. We should also evaluate other statistics metrics, but for now: Yes! According to the p-values, these variables are statistically significant.

- Crime crate: 2.20e-10 ***
- Residential land zoned for lots over 25,000 sq.ft 6.14e-06 ***
- Local pupil-teacher ratio < 2e-16 ***
- Weather the whether the tract bounds Chas River 0.000514 ***

d) Use the anova analysis and determine the order of importance of these four variables.

```r
# To get the same random values
set.seed(123)

# Run ANOVA analysis
anova(medium_house_model)
```

```
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
```

This ANOVA analysis help us to determine that all these variables (crime, proportion of residential land zoned for lots over 25,000 sq.ft, the local pupil-teacher ratio, and weather the whether the tract bounds Chas River) are very significant to explain the predicted variable (estimate the median value of owner-occupied homes). As the ANOVA proves, the p-values are very very small which mean are statistically significant, and the square error are big which means all the variables explain the variability.