Linear Regressions

semicolon

Learning Outcomes

- Participants should understand how simple and multiple linear regression works

Introduction

- Regression modeling is method used in explaining the relationship between set of variables from measured data through mathematical models. These models are used to explain an output value given a new set of input values. Linear regression modeling is a specific form of regression modeling that assumes that the output can be explained using a linear combination of the input values

  The overall purpose for developing a regression model is to determine what the output value of a system should be for a new set of input values, given that you have a collection of data about similar systems.

Introduction

- The benefit of using regression analysis is that it identifies the significant relationships between dependent and independent variables and the strength of the impact of multiple independent variables on independent variables.
  Linear regression finds the relationship between one dependent variable and one independent variable using a regression line

$$y = b_0 + b_1 x$$

- *y* is the dependent variable, *x* is the independent variable, *b*0 is the intercept, and *b*1 is the slope

Understanding Your Data

- Having a reliable data is the basis of any sort of regression modeling. If there is a problem with the data or if the authenticity of the data can not be ascertain, the results of your model(s) will be flawed.

- Hence be sure of the reliability of your data before attempting to model it. Most of the data in the field are not always in the right shape for modeling, you may need to do some data cleaning and manipulation, such as fixing the missing values and transforming existing variables as we have learnt in the previous classes.

To calculate the slope, you can use

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

To calculate the intercept, you can use

If $b1 > 0$, x and y have a positive relationship.
If $b1 < 0$, x and y have a negative relationship

Assumptions of Linear Regression

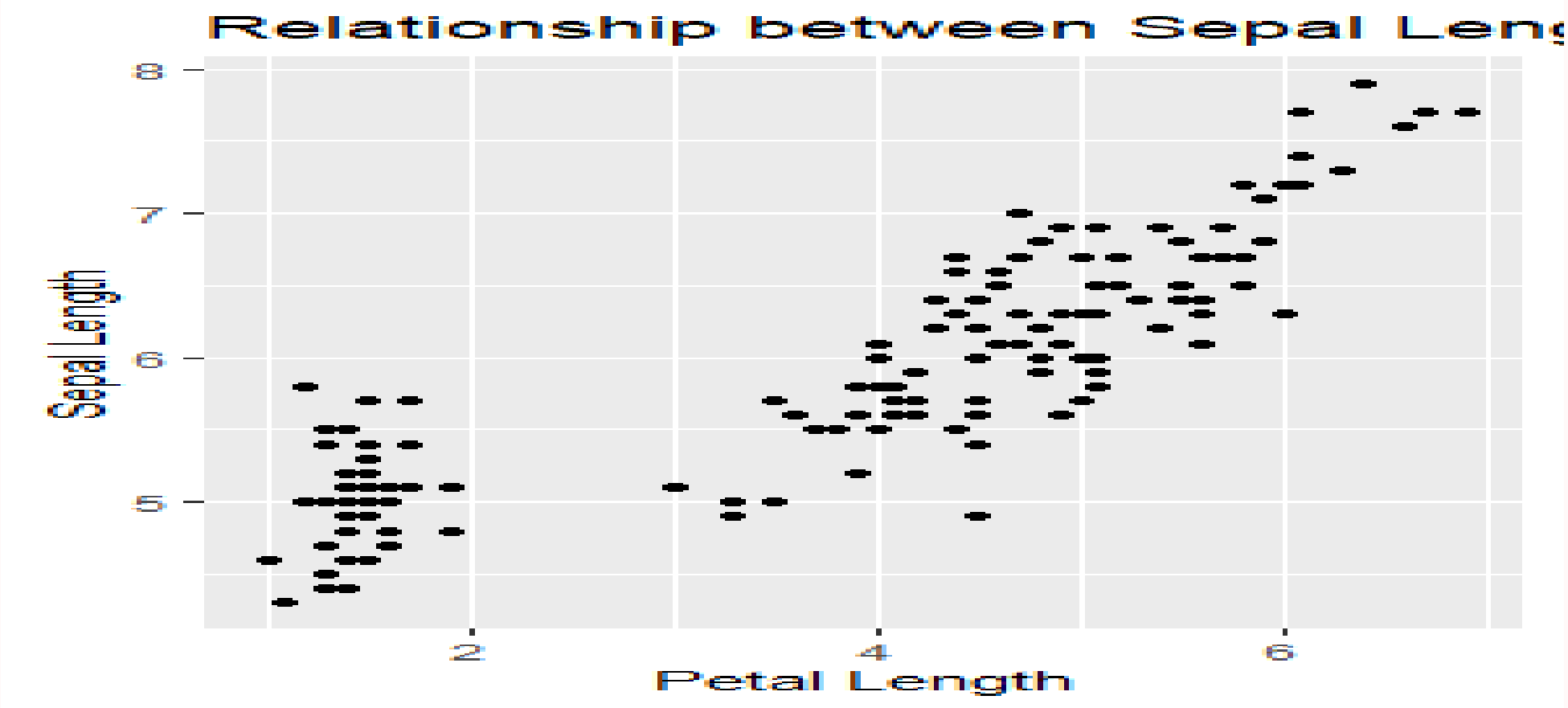
- Constant variance assumption
- Linear relationship dependent and independent variables (This assumption can be tested through scatter plot)
- Normality of residuals (This assumption may be checked by looking at a histogram or a Q-Q-Plot. Normality can also be checked with a goodness of fit test (e.g., the Kolmogorov-Smirnov test), though this test must be conducted on the residuals themselves)
- The data come from a random sample of size n from the population of interest or a randomized experiment

To use linear regression in R, you use the lm() function:

Example : Using the Iris data, is there is significant relationship between Sepal Length and Petal Length? To what extent does the Petal Length affects the Sepal Length?

To answer this question, let’s examine the relationship between Sepal Length and Petal Length Visually

Scatterplot of both variables



- From the scatter plot, there is positive relationship between the dependent variable “Sepal Length” and the independent variable “ Petal Length”

- Hence, to fit a linear regression model in R, we use the lm function.

Fitting Simple linear regression in R

```
m1<-lm(Sepal.Length~Petal.Length, data = iris)
summary(m1)
Call:
lm(formula = Sepal.Length ~ Petal.Length, data = iris)
Residuals:
     Min       1Q   Median       3Q      Max
-1.24675 -0.29657 -0.01515  0.27676  1.00269
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.30660    0.07839   54.94   <2e-16 ***
Petal.Length  0.40892    0.01889   21.65   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Residual standard error: 0.4071 on 148 degrees of freedom
- Multiple R-squared: 0.76, Adjusted R-squared: 0.7583
- F-statistic: 468.6 on 1 and 148 DF, p-value: < 2.2e-16

The output depicts that the linear equation is

$y = 4.30660 + 0.40892(\text{Petal Length})$

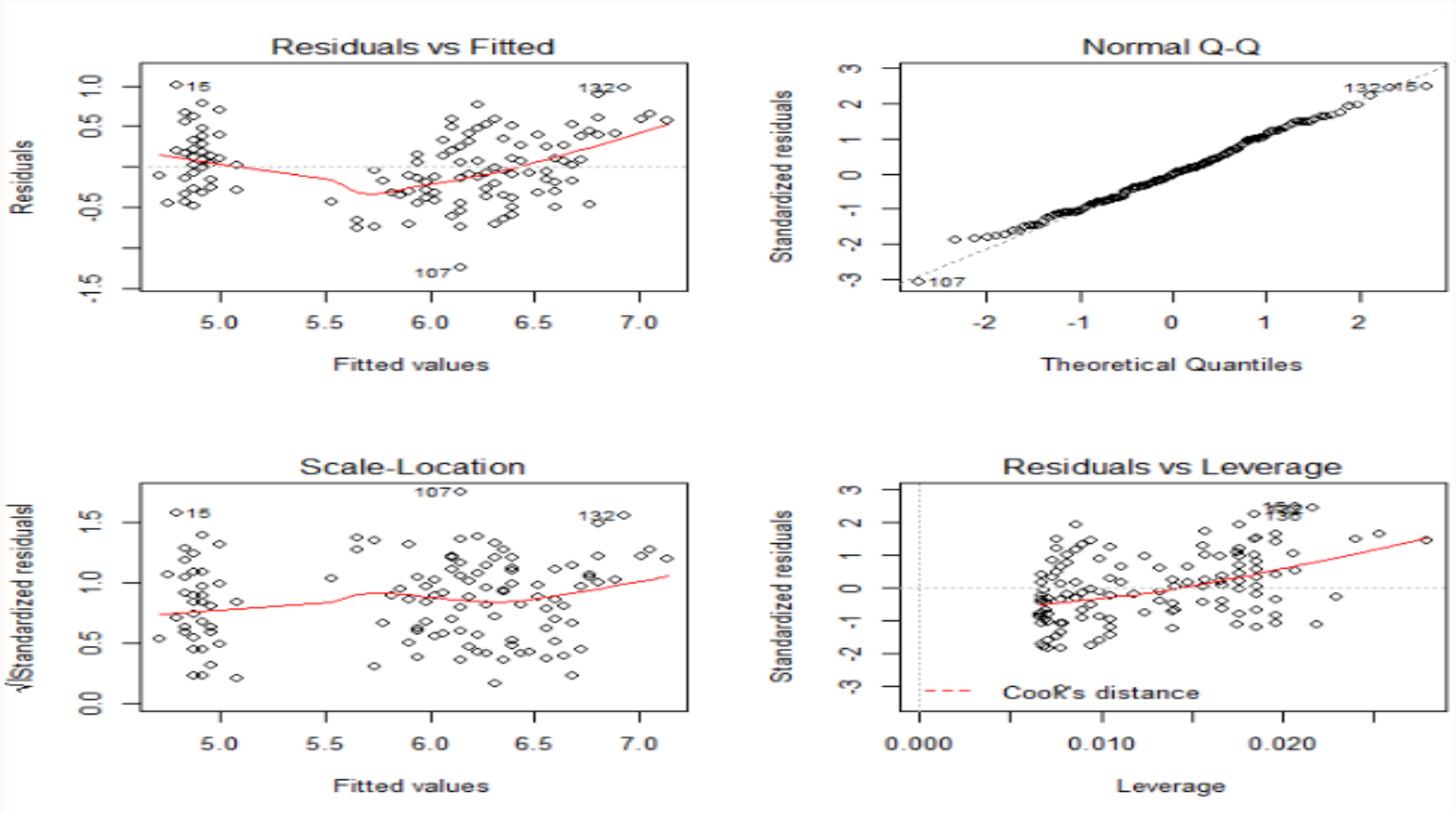
- The P-value of 2.2e-16 indicates that the Petal Length has significant effect on Sepal Length.

The intercept is also significant at 5% level of significance.

To interpret the Slope, for a unit increase in the Petal Length, the Sepal Length increases by 0.40892.

The R-squared is interpreted as the amount of variation of the dependent variable that can be explained by the independent variable. For the model, the 75.83% of the total variation in Sepal Length can be explained by the independent variable “Petal Length”

Is the Any Assumption Violated?



semicolon

- From the Q-Q Plot, Normality of residuals, the assumption of normality of residuals has not been violated as the points on the residual plot lie closely to the line.

- Constant variance assumption is met as there is no pattern of points in the residual vs Fitted plot

- Linear relationship between Sepal Length and Sepal Weigth, the assumption is not violated as there is a linear relationship between both variables

- R-square depicts the proportion of the variation in the dependent variable, and the formula is

$$R^2 = 1 - \frac{SSE}{SST}$$

- where SSE is the sum of squared errors

$$SSE = \sum_i^n \left(y_i - \hat{y}_i \right)^2$$

- and SST is the sum of the squared total

$$SST = \sum_i^n \left(y_i - \bar{y}_i \right)^2$$

$\bar{y}$ is the mean of Y and $\hat{y}$ is the fitted value for row i

Hence, the higher the R-squared and the adjusted R-squared, the better the linear model. The lower the standard error, the better the model

Multiple Linear Regression

- A simple linear regression is for a single response variable, y, and a single independent variable, x. The equation for a simple linear regression is

$$y = b_0 + b_1 x$$

- Multiple linear regression is built from a simple linear regression. Multiple linear regression is used when you have more than one independent variable. The equation of a multiple linear regression is
-

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k + \epsilon$$

- You create a multiple linear regression model using the lm() function
- Example 2: Using the House data set, What are the factors that significantly affect the price of house?

- Let’s start by loading our house data set into R

- hs<-read.csv(file.choose(),header = T)

```
> str(hs)
'data.frame':	781 obs. of  8 variables:
 $ MLS.      : int  132842 134364 135141 135712 136282 136431 137036 137090 137159 137570 ...
 $ Location  : Factor w/ 34 levels "Arroyo Grande",..: 1 24 24 19 30 22 30 30 19 2 ...
 $ Price     : int  795000 399000 545000 909000 109900 324900 192900 215000 999000 319000 ...
 $ Bedrooms  : int  3 4 4 4 3 3 4 3 4 3 ...
 $ Bathrooms : int  3 3 3 4 1 3 2 2 3 2 ...
 $ SQFT      : int  2371 2818 3032 3540 1249 1800 1603 1450 3360 1323 ...
 $ Price.SQFT: num  335 142 180 257 88 ...
 $ Status    : Factor w/ 3 levels "Foreclosure",..: 3 3 3 3 3 3 3 3 3 3 ...
```

- > attach(hs)
- > m2<-lm(Price~Bedrooms+Bathrooms+SQFT+Status)
- > summary(m2)
- Call:
- lm(formula = Price ~ Bedrooms + Bathrooms + SQFT + Status)

- Residuals:
- Min 1Q Median 3Q Max
- -1243571 -104020 -11043 65370 3902344

- Coefficients:
- Estimate Std. Error t value Pr(>|t|)
- (Intercept) 21595.39 38467.52 0.561 0.57469
- Bedrooms -88542.29 13104.93 -6.756 **2.78e-11 *****
- Bathrooms 47260.03 16311.00 2.897 **0.00387 ****
- SQFT 293.32 17.06 17.195 **< 2e-16 *****
- StatusRegular 208722.97 30641.27 6.812 **1.93e-11 *****
- StatusShort Sale -20767.89 21768.32 -0.954 0.34036

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 241100 on 775 degrees of freedom

Multiple R-squared: 0.5259,   Adjusted R-squared: 0.5229

F-statistic: 171.9 on 5 and 775 DF, p-value: < 2.2e-16

The linear model from the output is

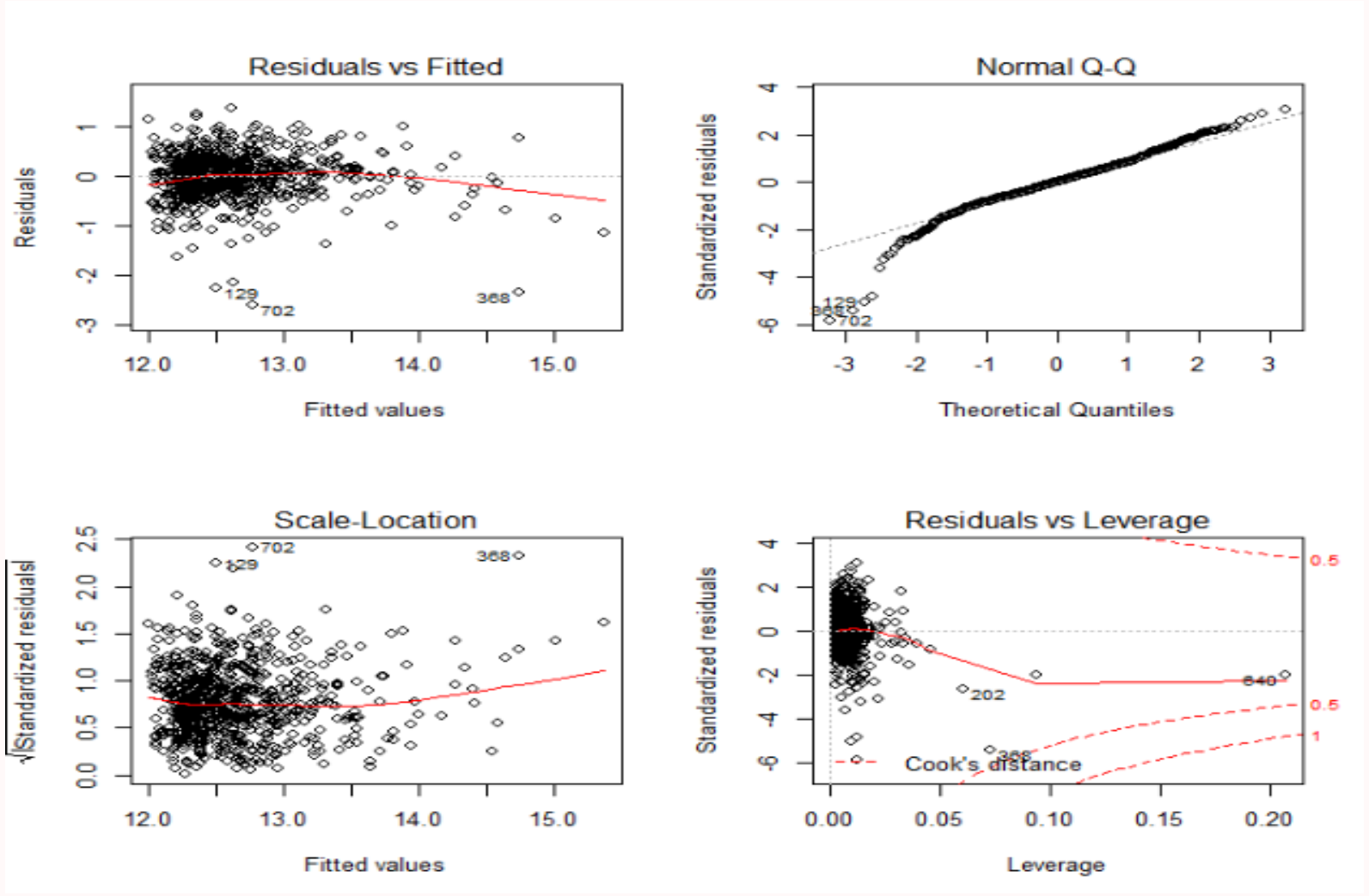
Y = 21595.39 - 88542.29(Bedrooms) + 47260.03(Bathrooms) +293.32(SQFT) + 208722.97(StatusRegular)-20767.89(StatusShort-Sale)

- The model fitted indicate that factors such as “Bedroom”,”Bedroom” “SQFT” and “StatusRegular” have significant effect on the price of house. The Adjusted R-square value of 0.5229 indicates that 52.29% of the total variation in the dependent variable “Price” is explained by the independent variables

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_i \left(y_i - \hat{y}_i\right)^2}{\sum_i \left(y_i - \hat{y}_i\right)^2}$$

- $\bar{y}$ is the mean of Y and y is the fitted value for row I
- $\hat{y}$ is the fitted value, which means that in $\hat{y}$ = 21595.39 - 88542.29(Bedrooms) + 47260.03(Bathrooms) +293.32(SQFT) + 208722.97(StatusRegular) - 20767.89(StatusShort-Sale), you fit in Bedrooms, Bathrooms, SQFT, StatusRegular and StatusShort-Sale to get y .

Any Assumption Violated?



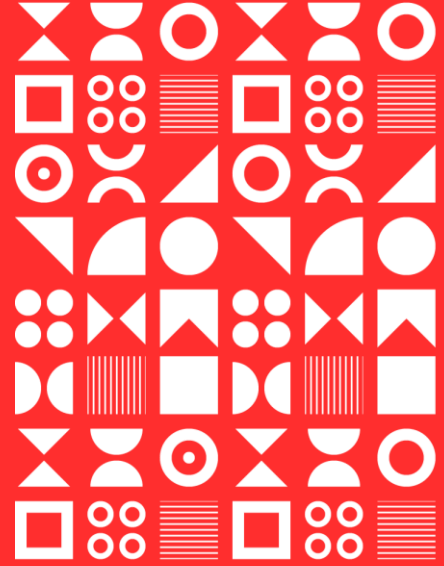
semicolon

- From the Q-Q Plot, Normality of residuals, the assumption of normality of residuals is violated as many points on the residual plot do not lie closely to the line.

- Constant variance assumption is not met as there is are pattern of points in the residual vs Fitted plot

Reference

- All IT eBooks. “Learn R for Applied Statistics - PDF EBook Free Download.” *Allitebooks.In*, 13 Feb. 2019, www.allitebooks.in/learn-r-for-applied-statistics/.



semicolon