3) Checking Heteroscedasticity, Multicollinearity, and Autocorrelation

FOSSEE (R Team)

28 January, 2020

What is Heteroscedasticity, Multicollinearity, and Autocorrelation?

1. An essential assumption of a linear regression model is that the error terms of the model have a constant variance. Unfortunately, it is often the case that we encounter non-constant variances of the error terms. For instance, there may be increasing variances of the error terms with the value of the response. **Heteroscedasticity** refers to these non-constant variances in the errors.
2. **Multicollinearity** is a phenomenon which occurs in the Multiple Regression Model. It exists when collinearity occurs between three or more predictor variables.
3. **Autocorrelation** is the measure of the correlation between a time series and a lagged version of itself over successive time intervals. It is measured to get an idea of whether a variable's current value is in any way related to its past value.

Introduction

The purpose of this experiment is to create a multivariate linear regression model fitted over the dataset "iris". This dataset contains the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. After successfully creating the model, we calculate its Heteroscedasticity, Multicollinearity and Autocorrelation.

Procedure

Step by step procedure to conduct the required experiment -

1. Normalizing data
2. Labelling species
3. Creating a linear regression model for predicting species
4. Checking Heteroscedasticity using the Breusch-Pagan test
5. Checking Multicollinearity
6. Checking Autocorrelation

Note : Please make sure that the following packages are already installed -

- *lmtest*
- *car*

Code and Results

Data used for analysis

```
## R has a predefined dataset with the name "iris"
  # To know more about the dataset type "?iris" in the console
```

Normalizing data

```
  # 1) Normalizing data
    Data <- iris
    Data[,1:4] <- as.data.frame(scale(iris[,1:4]))
```

Labelling species

```
  # 2) Labelling species
    Data$Species <- as.integer(factor(Data$Species))
```

Creating a linear regression model for predicting species

```
  # 3) Creating a Linear Regression model and predicting values
    Model <- lm(Species~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width, Data)
```

Checking Heteroscedasticity using the Breusch-Pagan test

```
  # 4) Checking Heteroscedasticity using the Breusch-Pagan test
    # Run the following command after removing "#" if "lmtest" library is not installed
    # install.packages("lmtest")
    library(lmtest)
    # Test for Heteroskedasticity
    bptest(Model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  Model
## BP = 32.381, df = 4, p-value = 1.599e-06
```

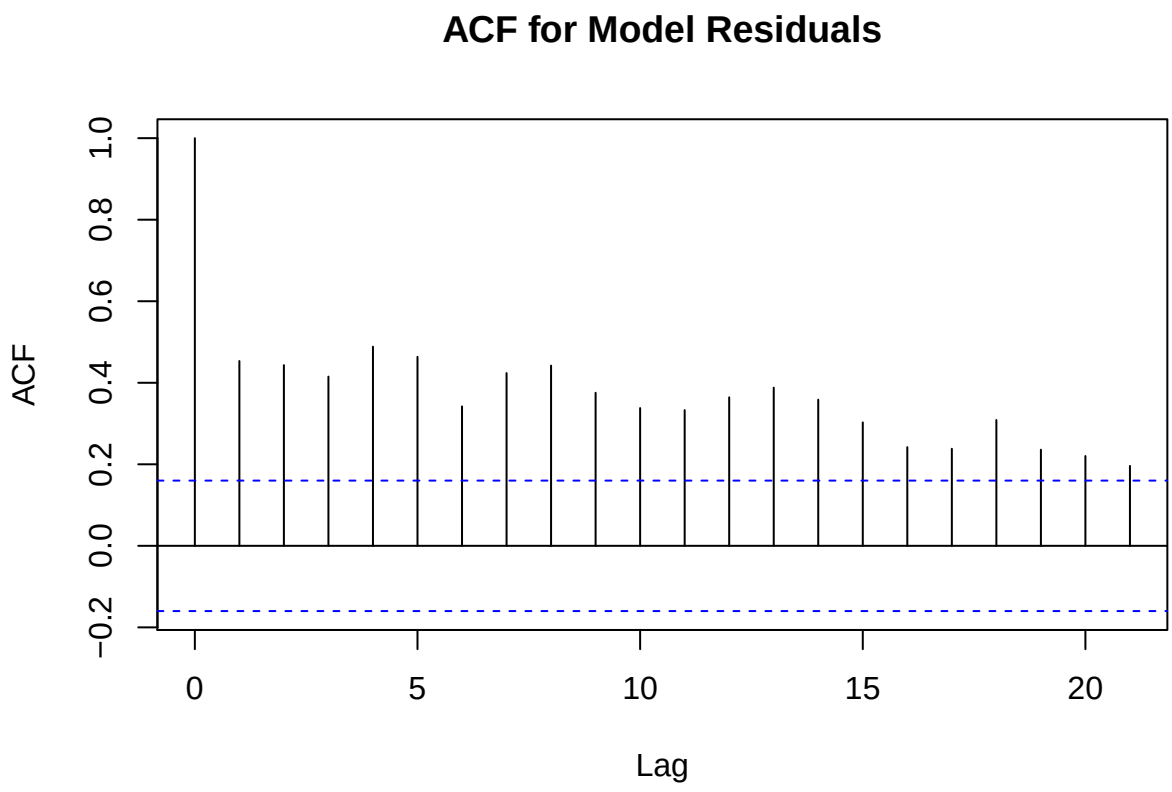
Checking Multicollinearity

```r
# 5) Checking Multicollinearity
  # Run the following command after removing "#" if "car" library is not installed
  # install.packages("car")
  library(car)
  # Calculating Variance Inflation
  vif(Model)
```

```
## Sepal.Length  Sepal.Width Petal.Length  Petal.Width
##     7.072722     2.100872    31.261498    16.090175
```

Checking Autocorrelation

```r
# 6) Checking Autocorrelation
  # Autocorrelation is checked for the residuals of Linear model
  AutoCorrelation <- acf(Model$residuals, plot = FALSE)
  plot(AutoCorrelation, main = "ACF for Model Residuals")
```



Conclusion

The **Breusch-Pagan test** against **Heteroscedasticity** shows that the p-value is very small. Hence, we conclude that heteroskedasticity is present. By checking the **Multicollinearity**, we observed that the variance of **'Petal.Length'** coefficient is inflated the most due to Multicollinearity in the model and the

variance of **'Sepal.Width'** inflated the least. The **Autocorrelation** plot shows that there is a correlation among the residuals of the generated multivariate linear regression model.