2) Estimation of a Multiple Regression Model, Significance, and Confidence Interval

FOSSEE (R Team)

28 January, 2020

What is a Multiple Regression Model, and what do we mean by its Significance and Confidence Interval?

1. **Multiple Regression Model** refers to a predictive modelling technique using which the relationship between multiple independent variables and a single dependent variable can be determined. Let us represent the independent variables by '$X1, X2, X3, ..., Xn$' and the dependent variable by 'Y'. We can represent the Multiple Regression Model by the equation "$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + ... + \beta nXn + \epsilon$". Here, '$\beta 0, \beta 1, \beta 2, \beta 3, ..., \beta n$' represent regression coefficients and 'ϵ' represents error term.
2. To determine the **significance** of the generated model we can analyze the following parameters associated with the model -

- *Residual standard error* - A measure of the deviation of predicted values from the actual values. If residual standard error is zero, then there is probably an overfitting situation.
- *R-squared* - It is a measure of the proportion of variability in predicted values that can be explained using model's inputs. The value should be as high or close to one as possible for best fit.
  - *Multiple R-squared* - It represents the percentage of the variance of the generated model intact after subtracting the error of the model.
  - *Adjusted R-squared* - It gives an adjusted value of R-squared by increasing R-squared only when a newly added term in the model improves the model's fit; otherwise R-squared is reduced.
- *p-value* - Summary function performs an F-test on the generated model and compares it to a model that has fewer parameters. Theoretically, a model with more parameters should fit better, and hence the p-value should be as low as possible and very near to absolute zero.

3. **Confidence Interval** of a regression model represents the difference between the upper and lower limit of the predicted values for each data point.

Null Hypothesis (H_o)

There is a relationship between sepal length, sepal width, petal length & petal width, and species of iris.

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 significance and confidence interval.

Procedure

Step by step procedure to conduct the required experiment -

1. Splitting the dataset for training and testing
2. Normalizing data
3. Labelling species
4. Creating a linear regression model and predicting species
5. Plotting predictions against actual values
6. Calculating the significance of the generated model
7. Calculating the confidence interval of the generated model

Code and Results

Data used for analysis

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

Splitting the dataset for training and testing

```r
  # 1) Creating Training and Testing data from "iris" dataset
    # 1.1) Generating Random Numbers
    set.seed(100)
    # 1.2) Creating sample for splitting dataset
    Sample <- sample(nrow(iris),0.7*nrow(iris))
    # 1.3) Training dataset containing 70% of data
    Training <- iris[Sample,]
    # 1.4) Testing dataset containing 30% of data
    Testing <- iris[-Sample,]
```

Normalizing data

```r
  # 2) Normalizing data
    # 2.1) Normalizing training data
    Training[,1:4] <- scale(Training[,1:4])
    # 2.2) Normalizing testing data
    Testing[,1:4] <- scale(Testing[,1:4])
```

Labelling species

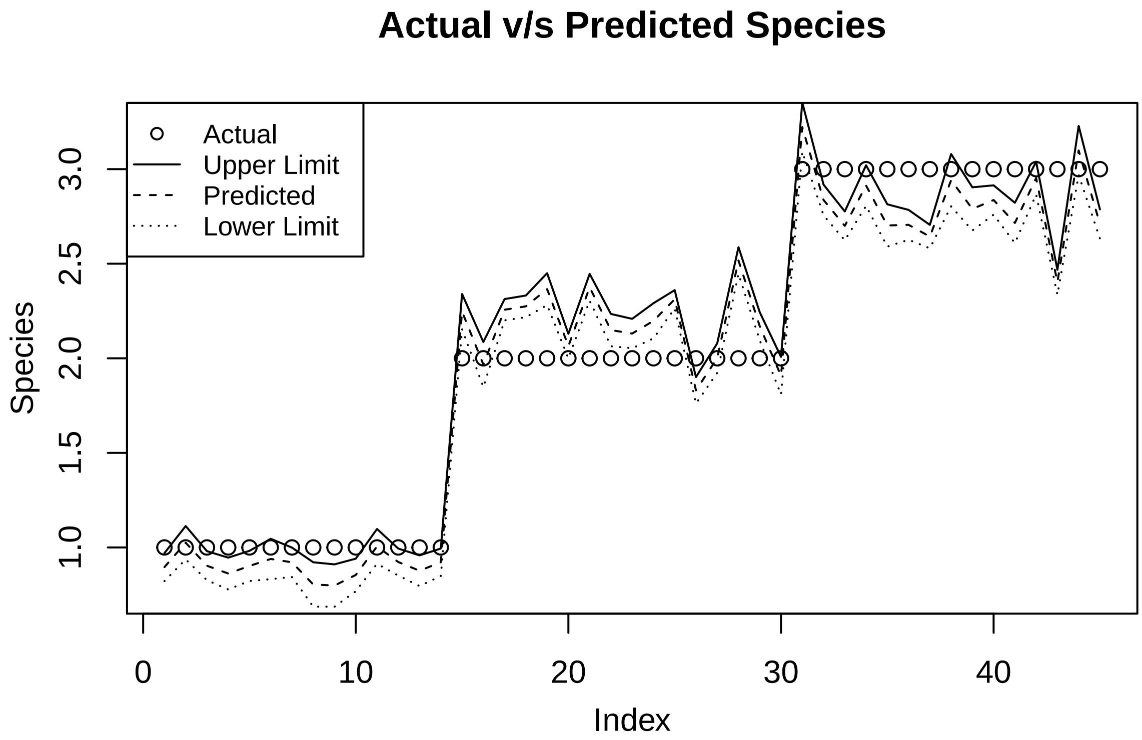
```r
  # 3) Labelling species
    # 3.1) Labelling species for training data
    Training$Species <- as.integer(factor(Training$Species))
    # 3.2) Labelling species for testing data
    Testing$Species <- as.integer(factor(Testing$Species))
```

Creating a linear regression model and predicting species

```
# 4) Creating a Linear Regression model and predicting values
  # 4.1) Applying Linear Regression
  Model <- lm(Species~Sepal.Length+Sepal.Width+Petal.Length+Petal.Width,
              data = Training)
  # 4.2) Predicting species
  Prediction <- as.data.frame(predict(Model,Testing,interval =
                                'confidence'))
```

Plotting predictions against actual values

```
# 5) Plotting Predictions
  plot(Testing$Species,main = "Actual v/s Predicted Species",ylab = "",
       xlab = "",ylim = c(0.75,3.25))
  # 5.1) Adding a line to represent Upper Limit of prediction
  lines(Prediction$upr,lty = 1)
  # 5.2) Adding a line to represent Best Fit of prediction
  lines(Prediction$fit,lty = 2)
  # 5.3) Adding a line to represent Lower Limit of prediction
  lines(Prediction$lwr,lty = 3)
  # 5.4) Clipping legend to the figure region instead of the plot region
  par(xpd=TRUE)
  # 5.5) Adding a legend
  legend("topleft",text.width = 7,legend=c("Actual","Upper Limit",
         "Predicted","Lower Limit"),pch = c(1,NA,NA,NA) ,lty = c(NA,
         1,2,3) ,cex=0.8)
  # 5.6) Setting and adjusting labels of x and y axes
  title(ylab="Species",xlab="Index", line=2.25, cex.lab=1)
```



Calculating the significance of the generated model

```
# 6) The Significance of resultant Linear Model
  summary.lm(Model)
```

```
##
## Call:
## lm(formula = Species ~ Sepal.Length + Sepal.Width + Petal.Length +
##     Petal.Width, data = Training)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61283 -0.14650  0.01088  0.10227  0.49154
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.99048    0.02132  93.353  < 2e-16 ***
## Sepal.Length -0.11995    0.05421  -2.213 0.029196 *
## Sepal.Width  -0.02640    0.02975  -0.887 0.377148
## Petal.Length  0.39106    0.11279   3.467 0.000777 ***
## Petal.Width   0.50396    0.08335   6.046 2.57e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

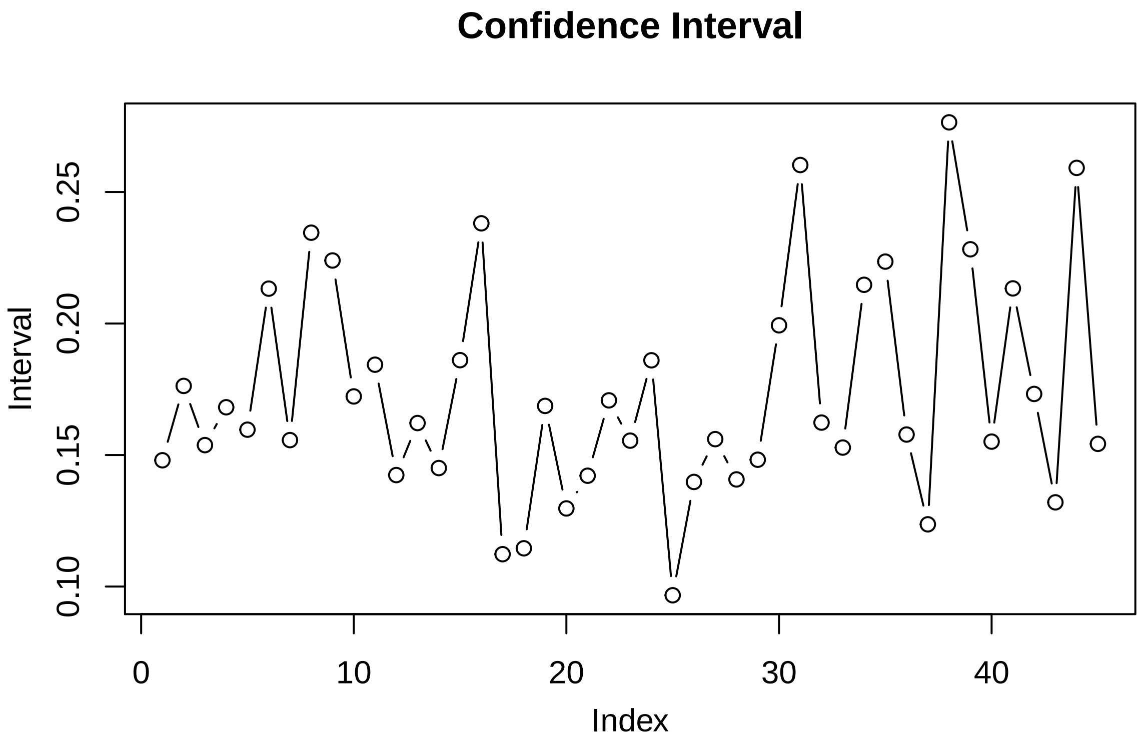
```
## Residual standard error: 0.2185 on 100 degrees of freedom
## Multiple R-squared:  0.9328, Adjusted R-squared:  0.9301
## F-statistic: 346.8 on 4 and 100 DF,  p-value: < 2.2e-16
```

Calculating the confidence interval of the generated model

```
# 7) Confidence Interval
  # 7.1) Printing Confidence Interval values
  print(Prediction$upr-Prediction$lwr)
```

```
##  [1] 0.14801001 0.17627413 0.15377171 0.16816653 0.15963880 0.21326906
##  [7] 0.15567929 0.23453247 0.22397448 0.17228212 0.18435260 0.14238072
## [13] 0.16216059 0.14505279 0.18607380 0.23810463 0.11230462 0.11454578
## [19] 0.16867483 0.12970364 0.14213106 0.17079017 0.15546856 0.18601734
## [25] 0.09669062 0.13976311 0.15604767 0.14073431 0.14821041 0.19932557
## [31] 0.26026791 0.16233498 0.15283196 0.21472740 0.22354795 0.15779732
## [37] 0.12365780 0.27648776 0.22822536 0.15509858 0.21335685 0.17322462
## [43] 0.13201575 0.25917330 0.15424942
```

```
  # 7.2) Plotting Confidence Interval
  plot(Prediction$upr-Prediction$lwr,main = "Confidence Interval",type =
       "b",ylab = "",xlab = "")
  # 7.3) Setting and adjusting labels of x and y axes
  title(ylab = "Interval",xlab="Index", line=2.25, cex.lab=1)
```



Conclusion

The *generated model is the best model for the prediction of iris species* as the **multiple R-squared and adjusted R-squared values are close to one.** Also, the **p-value is significantly low and near to zero.** The **residual standard error is non-zero;** hence there is no overfitting.

We **accept our Null Hypothesis** as we can observe a clear relationship between sepal length, sepal width, petal length & petal width, and species of iris.