Building a Multiple Linear Regression Model for Students Marks Dataset

Nithya S

2023-11-17

Introduction

The multiple regression model helps us in the prediction of the independent variable with the help of the explanatory variables. The relationship between the independent variables should not be linear or simple, the independent variables are not supposed to be correlated with each-other.

The dataset taken here is the marks of the students. We have the number of courses, study time, and marks as the components in the dataset.

Objective

To choose a data set of choice and perform a multiple linear regression with at least two regressors and to analyze it with the following questions:

1. To plot a matrix of scatter diagrams between the variables of interest and to find the matrix of coefficient of correlations and interpret it. To see if the regressors are independent of each other?
2. To fit a multiple linear regression model and interpret the estimated coefficients.
3. To test the significance of regression parameters using the t-test and interpret it.
4. Obtain the prediction and Confidence interval and interpret the results.

To Prepare a report based on the above questions with introduction, analysis, and conclusions.

URl for the dataset:

https://drive.google.com/file/d/1XkzjIz9JFUg20ynbQeTLG6h4nhogr8WN/view?usp=sharing

Data description

The dataset taken here is the marks of the students. We have the number of courses, study time and marks as the components in the dataset.

Procedure and Analysis

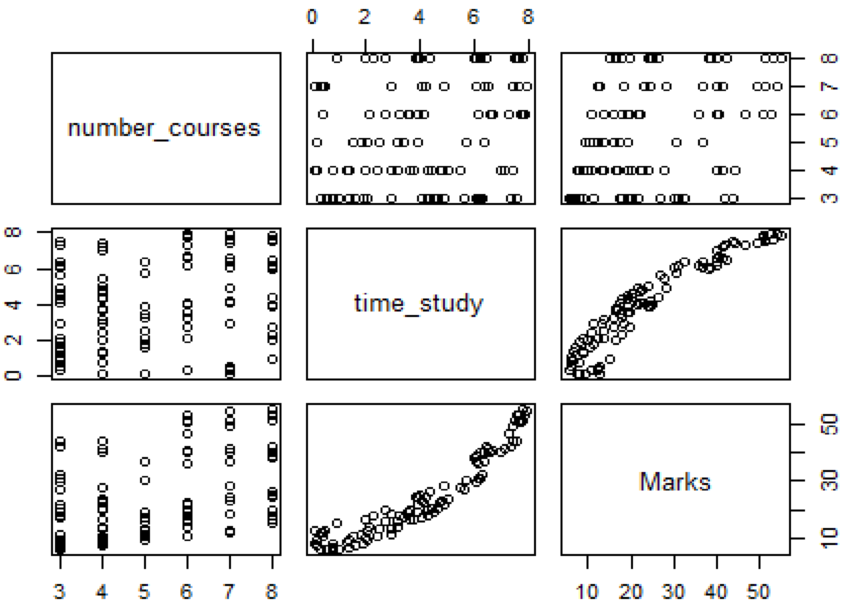
First we import the dataset as follows:

```
library(readr)
Student_Marks_dataset <- read_csv("G:/My Drive/Linear
Regression/Datasets/Student_Marks_dataset.csv")

## Rows: 100 Columns: 3
## ── Column specification
────────────────────────────────────────────────────────

## Delimiter: ","
## dbl (3): number_courses, time_study, Marks
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.

View(Student_Marks_dataset)
attach(Student_Marks_dataset)
```

1. To plot a matrix of scatter diagrams between the variables of interest and to find the matrix of coefficient of correlations and interpret it. To see if the regressors are independent of each other?

After importing the dataset, we check for the relationship between the variables of interest by a matrix of scatter diagrams. We use "pairs" to do so.

```
pairs(Student_Marks_dataset[1:3])
```



The matrix of coefficient of correlations can be obtained by importing the package “Hmisc” We use “rcorr” to get the correlation coefficient.

```
library(Hmisc)

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units

rcorr(as.matrix(Student_Marks_dataset))

##                number_courses time_study Marks
## number_courses           1.00       0.20  0.42
## time_study               0.20       1.00  0.94
## Marks                    0.42       0.94  1.00
##
## n= 100
##
##
## P
##                number_courses time_study Marks
## number_courses                0.0409     0.0000
## time_study     0.0409                    0.0000
## Marks          0.0000         0.0000
```

From the plot and the correlation coefficients we build our model as: The model for the dataset is,

$$Y = B_0 + B_1X_1 + B_2X_2 + U$$

where,

Y is the dependent variable. In the considered dataset it is “Marks” of the students.

B_0 is the intercept parameter

B_1 is the slope parameter.

X_1 and X_2 are the two dependent variables. Here they are “Number of courses” and “Study Time” respectively.

From the plot we observe that “Marks” of a student depends on “Number of courses” and “Study Time”. We also observe that there is no strong relationship between the dependent variables also the relationship between the dependent and independent variables are strongly and moderately present.

We see that the dependent variable “Marks” is positively strongly correlated with the independent variable “Study time” with 0.94 also “Marks” and the “Number of courses” has a moderate strong correlation of 0.42.

However, the correlation between the two variables under study being "Number of courses" and "Study time" has a weak positive correlation of 0.20 which makes them the two independent variables. And, yes, the regressors number of courses and study time are independent of each other.

2. To fit a multiple linear regression model and interpret the estimated coefficients.

On these lines, we can build our model as follows:

```
modell=lm(Marks~.,Student_Marks_dataset)
modell

##
## Call:
## lm(formula = Marks ~ ., data = Student_Marks_dataset)
##
## Coefficients:
##    (Intercept)  number_courses      time_study
##         -7.456           1.864           5.399
```

From the model, the estimated coefficients are as follows:

The intercept term (B_0) depicts that if X_1 ("Number of courses") and X_2 ("Study time") sometimes becomes zero, the intercept is simply the expected value of Y at that value. Here, B_0=-7.456, which can be inferred as the average marks of the student when the Number of courses and Study time is zero will be -7.456 ~ 0 marks, which simply says that the student will fail.

The slope term (B_1) depicts that for every 1 unit increase in X, the value of Y increases by B_1. Here, when study time is fixed and B_0 is given, B_1= 1.864 which says that the marks of the student increases by 1.864 ~ 2 marks. Similarly, when number of courses if fixed and B_0 is given, B_2=5.399 which says that the marks of the student increases by 5.399 ~ 5 marks.

Hence the model after estimation becomes,

$$Y^{hat} = -7.456 + 1.864X_1 + 5.399X_2$$

which is the estimated model. Also note that the sum of squares of error or the residual error can be obtained by the deviation of the model with the estimated model.$(Y - Y^{hat})$

3. To test the significance of regression parameters using the t-test and interpret it.

We give the hypotheses as follows:

H_0: B_1=B_{10}[zero] and B_2=B_{20}[zero]

v/s

H_1: B_1=!B_{10}[zero] and B_2=!B_{20}[zero]

We quickly check the summary of the model as follows:

```
summary(modell)

##
## Call:
## lm(formula = Marks ~ ., data = Student_Marks_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5617 -3.1023 -0.8361  3.6051  7.2158
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -7.4563     1.1745  -6.349 6.98e-09 ***
## number_courses    1.8641     0.2017   9.243 5.78e-15 ***
## time_study        5.3992     0.1529  35.303  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.534 on 97 degrees of freedom
## Multiple R-squared:  0.9404, Adjusted R-squared:  0.9391
## F-statistic: 764.8 on 2 and 97 DF,  p-value: < 2.2e-16
```

Here, t value for B_1 which is 9.243 is significantly large than t (table value) = 1.9643[two tailed test with {n-3=100-3} 97 degrees of freedom at 5% significance level], then there exist a very strong linear relation. Here t calculated value is 9.243 is greater than t table value [1.9643]

Here, t value for B_2 which is 35.303 is significantly large than t(table)=1.9643[two tailed test with {n-3=100-3} 97 degrees of freedom at 5% significance level], then there exists a very strong linear relation. Here t calculated value is 35.303 is greater than t table value [1.9643]

4. Obtain the prediction and Confidence interval and interpret the results.

The confidence intervals for the model as be obtained as follows: For our understanding we find the intervals at 99% confidence

```
confint(modell,level=0.99)

##                      0.5 %    99.5 %
## (Intercept)    -10.542197 -4.370495
## number_courses   1.334165  2.393936
## time_study       4.997335  5.801023
```

We take the significance level of 0.01. The reason being the p-value between the independent variables is 0.0409. So we consider 99% confidence as p-value will be greater than the significance level.

From the 99% confidence interval we infer that when the experiment/model is repeated 100 times say, we are confident that 95 times the true value of the parameters will lie in the obtained intervals.

For the intercept, the interval is [-10.542197, -4.370495]

For number of courses, it is [1.334165, 2.393936]

For study time it is [4.997335, 5.801023]

The confidence intervals width depends on the confidence level and the sample size of the data.

In order to get the prediction intervals, we first get the new observation as

```
new_data1=data.frame(number_courses=5,time_study=6.05)
new_data1

##   number_courses time_study
## 1              5       6.05
```

Now, the predicted value will be

```
predict(modell,new_data1)

##        1
## 34.52894
```

We infer from the predicted value as the marks of the student will be 34.52894 if he/she has 5 courses and spends 6.05 hours for study.

The prediction interval will be

```
PI=predict(modell,new_data1,interval="confidence",level=0.99)
PI

##        fit      lwr     upr
## 1 34.52894 33.27818 35.7797
```

From the prediction interval we infer that, the marks of a student if he/she has 5 courses and spends 6.05 hours on study, he/she gets marks as 34.52894.

Also, the predicted value of the marks will lie in the range of [33.27818, 35.7797]

Conclusion

Here, we fitted a Multiple linear regression model as,

$$Marks[Y] = B_0 + B_1 * Numberofcourses[X_1] + B_2 * Studytime[X_2] + U$$

where,

Y is the dependent variable being marks, X_1 and X_2 are the independent variables being the number of courses and study time respectively and U is the error component.

B_0, B_1, B_2 are the regression coefficients of the model that are to be estimated.

The estimated model was

$$Marks[Y^{hat}] = -7.456 + 1.864 * Numberofcourses[X_1] + 5.399 * Studytime[X_2]$$

We have chosen a significance level of 1% as the p-value of the independent variables are supposed to be greater than the significance level.

We have obtained the 99% confidence interval and also for a new data which was taken arbitrarily, we predicated the prediction interval.