MULTIPLE LINEAR REGRESSION

A.Mayuri(2348133)

Multiple Linear regression

A Regression Model for Predicting Marks from Hours of Study and Number of Courses

A.Mayuri(2348133)

2023-11-17

Assignment:

Choose a data set of your choice and perform a multiple linear regression with at least two regressors and do the analysis

Objective:

1. Plot a matrix of scatter diagrams between the variables of interest and also find the matrix of coefficient of correlations and interpret it. Are the regressors independent of each other? Justify your answer.
2. Fit a multiple linear regression model and interpret the estimated coefficients.
3. Test the significance of regression parameters using the t-test and interpret it.
4. Obtain the prediction and Confidence interval and interpret the results.

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

Variable of Interest:

-*Marks* : Marks is the dependent variable in our study. we are interested in understanding if marks can be influenced by the other two variables.

-*Number of courses and Hours of study* : Are the Independent variable in our study. we are interested in observing the relationship between these two variables and the marks.

Dataset(Source):

The data set was collected from kaggle website for multiple linear regression.

Model :

$Y=b0+b1X1+b2X2+E$

where X1-number of courses

X2- Hours of study

b1,b0,b2 are coefficients which are to be estimated E-error/residual

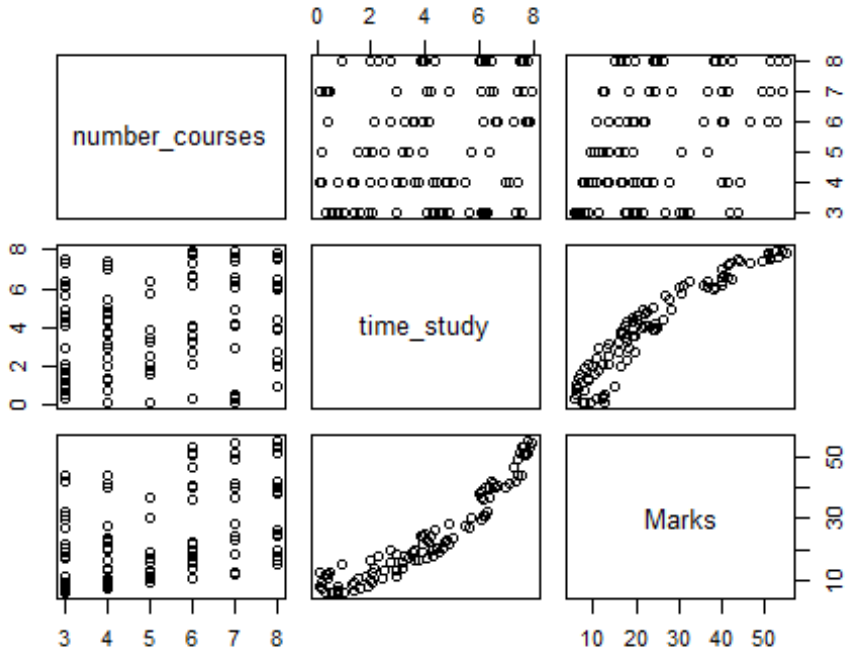
Prcedure And Analyses:

Step1: Import Data set

```
library(readr)
Marks <- read_csv("C:/Users/mayur/Desktop/Mstat/Semesters/Tri-sem2/Regression
/Dataset/Marks.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 m
essage.

View(Marks)
attach(Marks)

## The following object is masked _by_ .GlobalEnv:
##
##     Marks
```

Step2: Scatterplot across all variables

```
#scatter plot matrice between the independent and dependent variables
pairs(Marks[1:3])
```



####

Interpretation: In the above scatter plot we observe that there is a linear relationship between marks-time to study and marks -Number of courses.

Limitation:

this graph does not provide us the information that if the independent variables have a linear relationship jointly with the two variables.

Step3: Correlation test between the variables-Check dependency(Hypothesis testing)

```
library(Hmisc) #Loading the package
```

```
##
## Attaching package: 'Hmisc'
```

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

```
rcorr(as.matrix(Marks))
```

```
##                 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

model=lm(Marks~.,data=Marks)
```

0.20 is the correlation between the two independent variable which is very low.This helps us to deduce that the independent variables are sufficiently independent with each other.

0.42 is the correlation between the number of courses and marks, while 0.94 is the correlation between marks and hours of study both are sufficiently enough at this stage to say that there exsist some positive correlation between the independent and dependent variable seperately.

Conclusion from hypothsis Testing

At 5% confidence level using the p value 0.000 for marks and number of courses we reject null hypothsis and thus there exist a significant linear relationship between the variables.

similarly for marks and time required to study we can have a similar conclusion. thus there exist a significant correlation between dependent variable and no correlation between the independent variables

Step4: Building a Model

```
summary(model)

##
## Call:
## lm(formula = Marks ~ ., data = Marks)
##
## 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

confint(model,level=0.95)
```

```
##                     2.5 %    97.5 %
## (Intercept)    -9.787321 -5.125372
## number_courses  1.463789  2.264313
## time_study      5.095636  5.702721
```

If x2 variable is fixed y changes b1 units = 1.8641. similarly if x1 variable is fixed then y changes by b2 units=5.3992. from the estimate column of the table. Also since corresponding p values of the estimates are less than 0.05 thus significant.ie, we reject null hypothesis thus there is a significance between marks and courses ie(beta1 =!0), marks and hours of study(beta2 =!0).

we observe the adjusted r^2 than the r metric as it gives a better information about the reliability of the model 93.91% of the variation of Y can be predicted by two independent variable. it is a good model since (R^2 > 0.5)

Confidence interval : 95% sure that the b1 is between (1.4637,2.2643) and b2~ (5.095636,5.702721), b0 ~ (-9.787321,-5.125372) y(hat)=b0+b1*x(1)+b2*x(2)

Step5 : Prediction Data

Let us try to predict the marks (dependent variable) upon our estimated beta values given we assume a student studies for 8 hours and completes 9 courses

```
newdata=data.frame(number_courses=9,time_study=8)
newdata
```

```
##   number_courses time_study
## 1              9          8
```

```
predict(model,newdata)
```

```
##        1
## 52.51354
```

Thus our model predicts that the particular student may obtain a 52.5 marks approximately if he follows his study plan.

Step6: Prediction interval

```
pi=predict(model,newdata,interval="confidence",level=0.85)
pi
```

```
##        fit      lwr      upr
## 1 52.51354 51.16732 53.85977
```

Also if we assume a confidence level of 0.85 then his marks would be between the range (51.16,53.85) whose fit is 52.51 marks for the 9 courses and 8 hour study plan.

Conclusion:

Thus, we have formulated a linear model Y=b0+b1*X1+b2*X2

where X1-number of courses X2- Hourse of study b1,b0,b2 are coeffecients which are to be estimated E-error/residual

from the scatter plot,

we observe that there is a linear relationship between marks-time to study and marks - Number of courses.

from correlation matrice,

0.20 is the correlation between the two independent variable which is very low.This helps us to deduce that the independent variables are sufficiently independent with each other.

0.42 is the correlation between the number of courses and marks, while 0.94 is the correlation between marks and hours of study both are sufficiently enough at this stage to say that there exsist some positive correlation between the independent and dependent variable seperately.

Hypothesis testing,

at 5% confidence level using the p value 0.000 for marks and number of courses we reject null hypothsis and thus there exist a significant linear relationship between the variables.

similarly for marks and time required to study we can have a similar conclusion. thus there exist a significant correlation between dependent variable and no correlation between the independent variables

from the model,

if x2 variable is fixed y changes b1 units = 1.8641. similarly if x1 variable is fixed then y changes by b2 units=5.3992. from the estimate column of the table. Also since corresponding p values of the estimates are less than 0.05 thus significant.ie, we reject null hypothesis thus there is a significance between marks and courses ie(beta1 =!0), marks and hours of study(beta2 =!0).

we observe the adjusted r^2 than the r metric as it gives a better information about the reliability of the model 93.91% of the variation of Y can be predicted by two independent variable. it is a good model since ($R^2 > 0.5$)

from prediction interval,

if we assume a confidence level of 0.85 then his marks would be between the range (51.16,53.85) whose fit is 52.51 marks for the 9 courses and 8 hour study plan.