Predicting the percentage of an studdent based on the no.of study hours using linear regression model

Naveen.A

03/02/2021

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.

#Reading the packages required for analysis  
library(readr)  
library(dplyr)

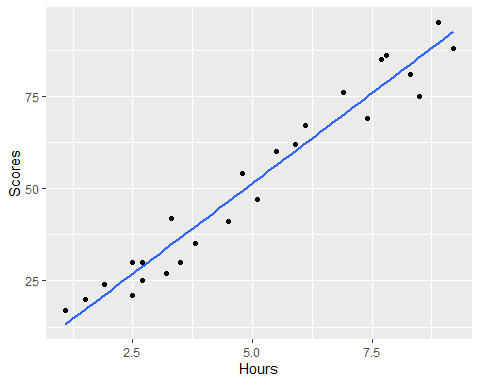
#Importing the Dataset and storing it into a variable.  
student\_data <- read\_csv("student\_scores.csv")

head(student\_data) #showing the top 5 data from the table.

## # A tibble: 6 x 2  
## Hours Scores  
## <dbl> <dbl>  
## 1 2.5 21  
## 2 5.1 47  
## 3 3.2 27  
## 4 8.5 75  
## 5 3.5 30  
## 6 1.5 20

#loading the ggplot2 libraries for plotting the graph  
library(ggplot2)  
  
#plotting the graph where the independent variable on the x #axis(hours) and dependent variable on the y axis(scores).  
ggplot(data = student\_data,  
 aes( x = Hours, y = Scores)) +   
 geom\_point() +   
 geom\_smooth(method = "lm", se = FALSE)

## `geom\_smooth()` using formula 'y ~ x'



#finding the coefficients of the data using linear model function   
  
student\_lm\_fit <- lm(Scores ~ Hours, data = student\_data)  
  
summary(student\_lm\_fit)

##   
## Call:  
## lm(formula = Scores ~ Hours, data = student\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.578 -5.340 1.839 4.593 7.265   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.4837 2.5317 0.981 0.337   
## Hours 9.7758 0.4529 21.583 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5.603 on 23 degrees of freedom  
## Multiple R-squared: 0.9529, Adjusted R-squared: 0.9509   
## F-statistic: 465.8 on 1 and 23 DF, p-value: < 2.2e-16

coef(student\_lm\_fit) #calling the coefficient value from the summary. that is the intercept and slope.

## (Intercept) Hours   
## 2.483673 9.775803

#finding the expected values i.e predicted value using the observed value.  
student\_data <- student\_data %>%   
 mutate(perc\_predictions = fitted(student\_lm\_fit))  
head(student\_data)

## # A tibble: 6 x 3  
## Hours Scores perc\_predictions  
## <dbl> <dbl> <dbl>  
## 1 2.5 21 26.9  
## 2 5.1 47 52.3  
## 3 3.2 27 33.8  
## 4 8.5 75 85.6  
## 5 3.5 30 36.7  
## 6 1.5 20 17.1

#predicted score if a student studies for 9.25 hours  
student\_score <- coef(student\_lm\_fit)[[1]] + coef(student\_lm\_fit)[[2]] \* 9.25  
student\_score

## [1] 92.90985

#The status of students depending upon their scores using case\_when function  
student\_data <- student\_data %>% mutate(status = case\_when(perc\_predictions < 35 ~ "Fail", perc\_predictions >= 35 & perc\_predictions <= 60 ~ "Pass", perc\_predictions > 60 & perc\_predictions < 85 ~ "first\_class", perc\_predictions >= 85 ~ "Distinction"))  
student\_data

## # A tibble: 25 x 4  
## Hours Scores perc\_predictions status   
## <dbl> <dbl> <dbl> <chr>   
## 1 2.5 21 26.9 Fail   
## 2 5.1 47 52.3 Pass   
## 3 3.2 27 33.8 Fail   
## 4 8.5 75 85.6 Distinction  
## 5 3.5 30 36.7 Pass   
## 6 1.5 20 17.1 Fail   
## 7 9.2 88 92.4 Distinction  
## 8 5.5 60 56.3 Pass   
## 9 8.3 81 83.6 first\_class  
## 10 2.7 25 28.9 Fail   
## # ... with 15 more rows

confint(student\_lm\_fit)

## 2.5 % 97.5 %  
## (Intercept) -2.753470 7.720817  
## Hours 8.838823 10.712784

#We can see that there is 97.5% probability that the true unknown value of the coefficient will fall within the specified range.and the intercept falls between the range ( -2.75 to 7.720 ) that is the score if a student studies #for 0 hours. and the hours falls between the range (8.83 - 10.71) that mean there will increase in score between 8.83 - 10.71 if the average of hour increases by 1 hour.

##conclusion: We can see that there is 97.5% probability that the true unknown value of the coefficient will fall within the specified range.variables are linear and hence i used linear regression model to predict the dependent/ response variable and we were successful in predicting the score of student studying for 9.25 hours.