spark-task-1-regression.R

SURABHI. S

2021-02-06

```
#to import the given data
#In  the given data, number of hours a student study is the
#independent variable, and scores are the dependent variable
hours=c(2.5,5.1,3.2,8.5,3.5,1.5,9.2,5.5,8.3,2.7,7.7,5.9,
        4.5,3.3,1.1,8.9,2.5,1.9,6.1,7.4,2.7,4.8,3.8,6.9,
        7.8)
scores=c(21,47,27,75,30,20,88,60,81,25,85,62,41,42,17,
          95,30,24,67,69,30,54,35,76,86)
mydata=data.frame(hours,scores)
names(mydata)=c("hours","scores")
mydata
```

```
##    hours scores
## 1    2.5     21
## 2    5.1     47
## 3    3.2     27
## 4    8.5     75
## 5    3.5     30
## 6    1.5     20
## 7    9.2     88
## 8    5.5     60
## 9    8.3     81
## 10   2.7     25
## 11   7.7     85
## 12   5.9     62
## 13   4.5     41
## 14   3.3     42
## 15   1.1     17
## 16   8.9     95
## 17   2.5     30
## 18   1.9     24
## 19   6.1     67
## 20   7.4     69
## 21   2.7     30
## 22   4.8     54
## 23   3.8     35
## 24   6.9     76
## 25   7.8     86
```

```
#regression equation
relation <- lm(scores~hours)
print(relation)
```

```
##
## Call:
## lm(formula = scores ~ hours)
##
## Coefficients:
## (Intercept)        hours
##       2.484        9.776

#to know the average error in prediction
print(summary(relation))

##
## Call:
## lm(formula = scores ~ hours)
##
## 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

#What will be predicted score if a student
#studies for 9.25 hrs/ day?
a <- data.frame(hours = 9.25)
result <-  predict(relation,a)
print(result)

##        1
## 92.90985

plot(scores,hours,col = "blue",main
     = "percentage of an student based on the no. of study hours",
     abline(lm(hours~scores)),cex = 1.3,pch = 16,xlab
     = "number of hours student study",ylab = "scores")
```

ɔercentage of an student based on the no. of study h