BUSINESS ANALYTICS ASSIGNMENT -3

LOKESH JETANGI

2023-03-29

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

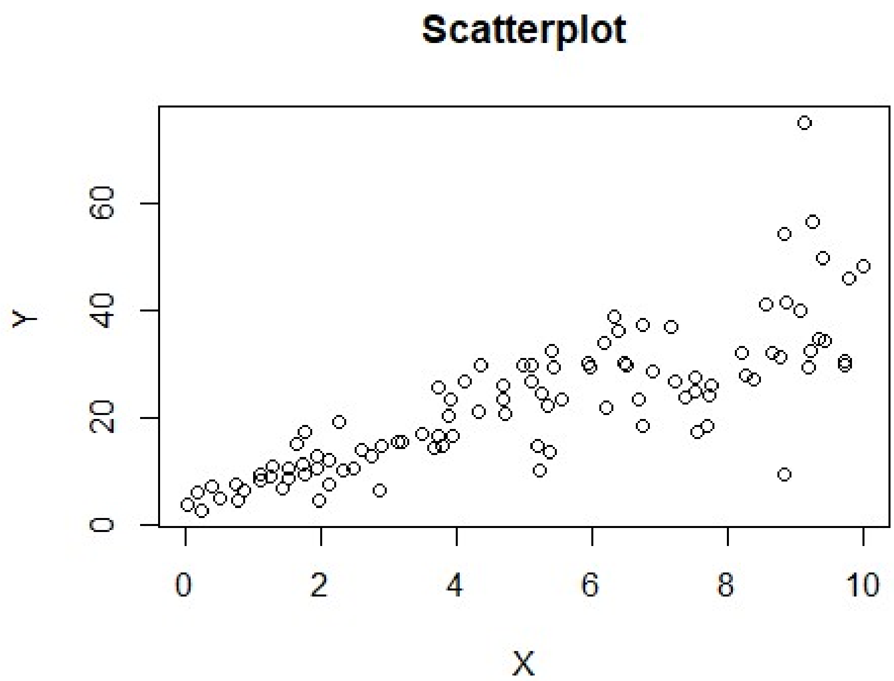
```
#TASK-1
set.seed(2017)
X=runif(100)*10
Y=X*4+3.45
Y=rnorm(100)*0.29*Y+Y
head(X)

## [1] 9.242426 5.371764 4.691956 2.886262 7.700882 7.727687

head(Y)

## [1] 56.33934 13.48174 23.49129 14.90010 18.42442 24.14282

#(A)Plot Y against X. Include a screenshot of the plot in your submission.
Using the File menu you can save the graph as a picture on your computer.
Based on the plot do you think we can fit a linear model to explain Y based
on X?
plot(X, Y, main = "Scatterplot",
     xlab = "X",
     ylab = "Y")
```




```
# Yes,based on the plot we can fit a linear model to explain Y based on X

#(B)
#Construct a simple linear model of Y based on X. Write the equation that
explains Y based on X. What is the accuracy of this model?

model= lm(Y ~ X)
summary(model)

##
## Call:
## lm(formula = Y ~ X)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.755  -3.846  -0.387   4.318  37.503
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.4655     1.5537   2.874  0.00497 **
## X             3.6108     0.2666  13.542  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.756 on 98 degrees of freedom
## Multiple R-squared:  0.6517, Adjusted R-squared:  0.6482
## F-statistic: 183.4 on 1 and 98 DF,  p-value: < 2.2e-16
```

```r
#The equation that explains Y based on X.
#    Y = 4.4655 + 3.6108X
#    The accuracy of the model is 0.6517 i.e, 65.17% which is good fit

#(C)
# How the Coefficient of Determination,R2, of the model above is related to
the correlation coefficient of X and Y?

cor(X,Y)^2

## [1] 0.6517187

#From the above we notice that it’s exact the same value of R-squared 0.6517

#TASK-2
# We will use the ‘mtcars’ dataset for this question. The dataset is already
included in your R distribution. The dataset shows some of the
characteristics of different cars.The following shows few samples (i.e. the
first 6 rows) ofthe dataset.The description of the dataset can be found here.

head(mtcars)

##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1

#(A)James wants to buy a car.He and his friend, Chris, have different
opinions about the Horse Power(hp) of cars. James think the weight of a car
(wt) can be used to estimate the Horse Power of the car while Chris thinks
the fuel consumption expressed in Mile Per Gallon (mpg), is a better
estimator of the (hp). Who do you think is right? Construct simple linear
models using mtcars data to answer the question.

model_James <- lm(hp ~ wt, data = mtcars)
summary(model_James)

##
## Call:
## lm(formula = hp ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.430 -33.596 -13.587   7.913 172.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.821     32.325  -0.056    0.955
```

```
## wt            46.160      9.625   4.796 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 52.44 on 30 degrees of freedom
## Multiple R-squared:  0.4339, Adjusted R-squared:  0.4151
## F-statistic:    23 on 1 and 30 DF,  p-value: 4.146e-05
```

```
model_chris <- lm(hp ~ mpg, data = mtcars)
summary(model_chris)
```

```
##
## Call:
## lm(formula = hp ~ mpg, data = mtcars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -59.26 -28.93 -13.45  25.65 143.36
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   324.08      27.43  11.813 8.25e-13 ***
## mpg            -8.83       1.31  -6.742 1.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.95 on 30 degrees of freedom
## Multiple R-squared:  0.6024, Adjusted R-squared:  0.5892
## F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07
```

```
#Based on the results, the model that predicts horsepower from mpg appears to
be superior because it has a higher R-squared value (0.6024) and a smaller
residual standard error (43.95) than the model that predicts horsepower from
weight (R-squared = 0.4339, residual standard error = 52.44). Hence, Chris's
claim that mpg is a better indicator of horsepower than James's claim that
weight is a better estimator seems to be more correct.
#Therefore,Chris model’s accuracy is 0.6024 which is very high than that of
James i.e, 0.4339.

#(B) Build a model that uses the number of cylinders (cyl) and the mile per
gallon (mpg) values of a car to predict the car HorsePower (hp).Using this
model, what is the estimated Horse Power of a car with 4 calendar and mpg of
22?
HP<- lm(hp ~ cyl + mpg, data = mtcars)
summary(HP)
```

```
##
## Call:
## lm(formula = hp ~ cyl + mpg, data = mtcars)
##
## Residuals:
```

```
##    Min     1Q Median     3Q    Max
## -53.72 -22.18 -10.13  14.47 130.73
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   54.067     86.093   0.628  0.53492
## cyl           23.979      7.346   3.264  0.00281 **
## mpg           -2.775      2.177  -1.275  0.21253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.22 on 29 degrees of freedom
## Multiple R-squared:  0.7093, Adjusted R-squared:  0.6892
## F-statistic: 35.37 on 2 and 29 DF,  p-value: 1.663e-08
```

```
predict(HP, data.frame(cyl = 4, mpg =22))
```

```
##        1
## 88.93618
```

```
HP$coefficients
```

```
## (Intercept)         cyl         mpg
##   54.066600   23.978626   -2.774769
```

```
predict_HP <- (HP$coefficients[2]*22) + (HP$coefficients[3]*4)
+ HP$coefficients[1]
```

```
## (Intercept)
##     54.0666
```

```
print(paste('The estimated horse power of a car with 4 calender and mpg of 22
is ', predict_HP))
```

```
## [1] "The estimated horse power of a car with 4 calender and mpg of 22 is
516.430701894008"
```

```
#The estimated horse power of a car with 4 calender and mpg of 22 is 88.93618

# TASK-3
# For this question, we are going to use Boston Housing data set.The data set
is in ‘mlbench’ package, so we first need to install the package, call the
library and the load the data set using the following commands

#install.packages("mlbench")
library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.2.3
```

```
data(BostonHousing)
str(BostonHousing)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524
0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : num  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ b      : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...

#You should have a data frame with the name of Boston Housing in your Global
environment now.The data set contains information about houses in different
parts of Boston. Details of the data set is explained here. Note the data set
is old, hence low house prices!

#TASK-3(A)
#Build a model to estimate the median value of owner-occupied homes
(medv)based on the following variables: crime crate (crim), proportion of
residential land zoned for lots over 25,000 sq.ft(zn), the local pupil-
teacher ratio (ptratio) and weather the whether the tract bounds Chas
River(chas). Is this an accurate model?

Model_Boston<-lm(formula = BostonHousing$medv ~ BostonHousing$crim +
BostonHousing$zn
        + BostonHousing$ptratio +BostonHousing$chas,data = BostonHousing)

summary(Model_Boston)

##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn +
##     BostonHousing$ptratio + BostonHousing$chas, data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           49.91868    3.23497  15.431  < 2e-16 ***
## BostonHousing$crim    -0.26018    0.04015  -6.480 2.20e-10 ***
## BostonHousing$zn       0.07073    0.01548   4.570 6.14e-06 ***
## BostonHousing$ptratio -1.49367    0.17144  -8.712  < 2e-16 ***
```

```
## BostonHousing$chas1    4.58393    1.31108   3.496 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.388 on 501 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.3547
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16
```

```
#The R-squared value for the model is 0.3599,which is low and indicates that
it is not very accurate.

# TASK-3(B1)
#Imagine two houses that are identical in all aspects but one bounds the Chas
River and the other does not. Which one is more expensive and by how much?
Model_Boston1 <- lm(formula = BostonHousing$medv ~ BostonHousing$chas,data=
BostonHousing)
Model_Boston1
```

```
##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$chas, data =
BostonHousing)
##
## Coefficients:
##        (Intercept)  BostonHousing$chas1
##             22.094                6.346
```

```
#using the coefficient of the above model we can calculate the values of both
the houses
#House 0 without chas and House1 with chas
House0<- Model_Boston1$coefficients[1]+Model_Boston1$coefficients[2]*0
House1<- Model_Boston1$coefficients[1]+Model_Boston1$coefficients[2]*1
print(paste('House with chas is more expensive and by',House1-House0))
```

```
## [1] "House with chas is more expensive and by 6.34615711252662"
```

```
#TASK-3(b2)
#Imagine two houses that are identical in all aspects but in the neighborhood
of one of them the pupil-teacher ratio is 15 and in the other one is 18.
Which one is more expensive and by how much?
Model_Boston2 <- lm(medv ~ ptratio, data = BostonHousing)
Model_Boston2
```

```
##
## Call:
## lm(formula = medv ~ ptratio, data = BostonHousing)
##
## Coefficients:
## (Intercept)      ptratio
##      62.345       -2.157
```

```r
House15 <- predict(Model_Boston2, newdata = data.frame(ptratio = 15))
House18 <- predict(Model_Boston2, newdata = data.frame(ptratio = 18))
price_diff <- House15 - House18
print(paste('House in which pupil-teacher ratio of two houses is 15,18 and is
more expensive and by ',House15-House18))

## [1] "House in which pupil-teacher ratio of two houses is 15,18 and is more
expensive and by  6.47152588818295"

#TASK-3(c)
#Which of the variables are statistically important (i.e. related to the
house price)? Hint: use the p-values of the coefficients to answer.

summary(Model_Boston)

##
## Call:
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn +
##     BostonHousing$ptratio + BostonHousing$chas, data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.282  -4.505  -0.986   2.650  32.656
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)            49.91868    3.23497  15.431  < 2e-16 ***
## BostonHousing$crim    -0.26018    0.04015  -6.480 2.20e-10 ***
## BostonHousing$zn       0.07073    0.01548   4.570 6.14e-06 ***
## BostonHousing$ptratio -1.49367    0.17144  -8.712  < 2e-16 ***
## BostonHousing$chas1    4.58393    1.31108   3.496 0.000514 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.388 on 501 degrees of freedom
## Multiple R-squared:  0.3599, Adjusted R-squared:  0.3547
## F-statistic: 70.41 on 4 and 501 DF,  p-value: < 2.2e-16

#As their p-values are less than 0.05.Looking at the p-values we can say that
non of the independent variables are statistically significant

#TASK-3(d)
#Use the anova analysis and determine the order of importance of these four
variables

anova(Model_Boston)

## Analysis of Variance Table
##
## Response: BostonHousing$medv
##                         Df  Sum Sq Mean Sq F value    Pr(>F)
```

```
## BostonHousing$crim      1  6440.8  6440.8 118.007 < 2.2e-16 ***
## BostonHousing$zn        1  3554.3  3554.3  65.122 5.253e-15 ***
## BostonHousing$ptratio   1  4709.5  4709.5  86.287 < 2.2e-16 ***
## BostonHousing$chas      1   667.2   667.2  12.224 0.0005137 ***
## Residuals             501 27344.5    54.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# To determine the order of importance, we can look at the magnitude of the
F-values. Based on the F-values, the order of importance from highest to
lowest is:

#crim
#ptratio
#zn
#chas
```