Assignment 03- Business Analytics

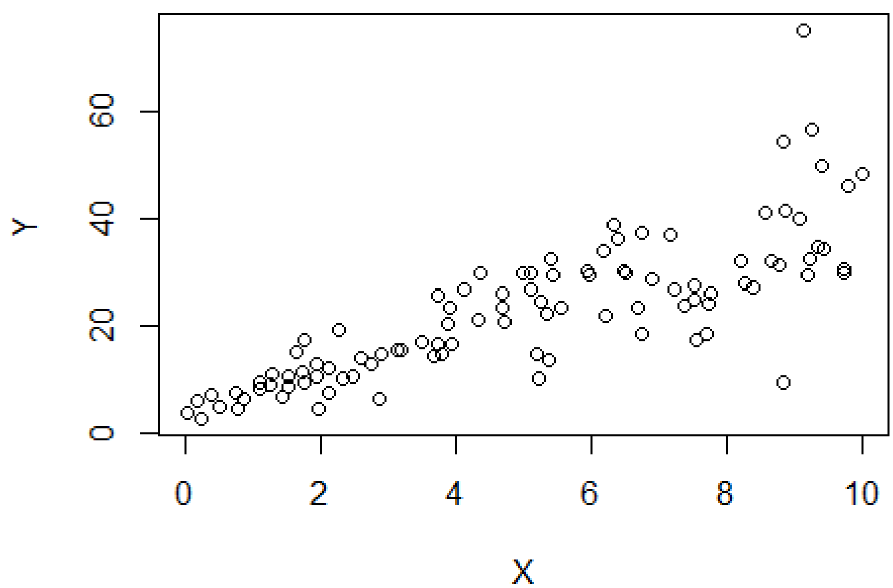
Chathurani Ekanayake

2023-04-02

```r
# 01.Run the following code in R-studio to create two variables X and Y.

set.seed(2017)
X=runif(100)*10
Y=X*4+3.45
Y=rnorm(100)*0.29*Y+Y


# 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)
```




```r
# Answer:
# According to the plot, it seems that a linear model can be used to show the
relationship between X and Y.It seems to be a positive linear relationship
```

```r
between X and Y and some scatter around the line. There are some points which
seem to be outliers and may affect the fit of the model.

# 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)
model

##
## Call:
## lm(formula = Y ~ X)
##
## Coefficients:
## (Intercept)            X
##       4.465        3.611

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

# Y Equation
# Y = 4.4655 + 3.6108*X
# Based on the R-squared and the residual standard error value the model
accuracy can be explained. According to the model summary, the R squared
value is 0.6517, which means that model explains 65.17% of the variability in
the data. The residual standard error is 7.756 which means the average
distance that the data points fall from the regression line. The overall
model appears to be a good fit for the data, as evidenced by the R-squared
value and the low residual standard error.

# C) How the Coefficient of Determination, R2, of the model above is related
to the correlation coefficient of X and Y?
# Answer:
# The coefficient of determination(R squared) measures how well the linear
```

```r
regression model fits the data.It measures the proportion of the variance in
the dependent variable (Y) according to the independent variable (X) in the
linear regression model.The correlation coefficient measures the strength and
direction of linear relationship between X and Y variable. Generally, the
correlation coefficient value should be a value in between -1 and +1.
According to the above summary, the R-squared value is 0.6482 means 64.82% of
variance in Y is explained by the linear relationship with X. It shows
somewhat good fit of the model to data. The correlation between X and Y can
be calculated as the square root of R-squared value. In here it is
approximately 0.807
cor(X,Y)

## [1] 0.807291

sqrt(summary(model)$r.squared)

## [1] 0.807291

# 02. 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.
# 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.
data("mtcars")

# Model1 - Horsepower(hp) and weight(wt)
model_wt <- lm(hp~wt, data=mtcars)

# Model2 - Horsepower(hp) and Miles Per Gallon (mpg)
model_mpg <- lm(hp~mpg, data=mtcars)

summary(model_wt)

##
## 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
```

```
summary(model_mpg)
```

```
##
## 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
```

```
# According to the summary statistics, both models are significant because
the significant coefficients (p< 0.05). The multiple R-squared value of
model_wt 0.4339 and adjusted R squared value is 0.4151.The multiple R-squared
value of model_mpg 0.6024 and adjusted R squared value is 0.5892. Based on
the R-squared and adjusted R-squared values the model_mpg can be identified
as the better model. Hence that, as Chris says to estimate horse power, Miles
per Gallon is better variable than wight of a car.

# 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 Horse Power (hp). Using this
model, what is the estimated Horse Power of a car with 4 calendar and mpg of
22?
#Build linear regression model
model3 <- lm(hp~ cyl + mpg, data= mtcars)

# predict the Horse Power (hp) of a car with 4 cylinders and mpg 22
new_data <- data.frame(cyl=4,mpg=22)
predicated_hp <- predict(model3, newdata = new_data)
predicated_hp
```

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

```r
# Answer:
# According to the above calculation the estimated Horse Power of a car with 
4 cylinders and 22mpg is 88.93618.

# 03. For this question, we are going to use BostonHousing dataset. The 
dataset is in ‘mlbench’ package, so we first need to instal the package, call 
the library and the load the dataset using the following commands

library(mlbench)

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

data("BostonHousing")

# 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? (Hint check R2 )
model4 <- lm(medv ~ crim + zn + ptratio + chas, data= BostonHousing)
summary(model4)

## 
## Call:
## lm(formula = medv ~ crim + zn + ptratio + 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 ***
## crim        -0.26018    0.04015  -6.480 2.20e-10 ***
## zn           0.07073    0.01548   4.570 6.14e-06 ***
## ptratio     -1.49367    0.17144  -8.712  < 2e-16 ***
## 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

summary(model4)$r.squared

## [1] 0.359859

# According to the above summary the R-Squared value is 0.3599. Therefore, 
this model indicates relatively low R-squared value. It means the model is
```

```
not accurately predicting the median value of owner-occupied homes based on
the predictor variables.

# B) Use the estimated coefficient to answer these questions?
## I) 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?
coefs <- coef(model4)

difference <- coefs["chas1"] * (1 - 0) # this is based on the assumption that
if a house bounds Chas river (chas=1), if a house dooes not bounds Chas river
(chas=0)
difference

##    chas1
## 4.583926

# Answer:
# According to the above explanation the difference is 4.583926. Because the
difference is positive, a house that bounds the Charles river is expected to
be $4,583.93 approximately than a house that does not.

## II) 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?
diff_ptratio <- 15-18
coefs <- coef(model4)
diff_medv <- coefs["ptratio"]*diff_ptratio
diff_medv

##  ptratio
## 4.481018

# Answer:
# Due to the above difference value is positive, the house with ptratio of 15
is expected to be more expensive than the  house with ptratio of 18.
Approximately $ 4,4810.18.

# 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(model4)

##
## Call:
## lm(formula = medv ~ crim + zn + ptratio + 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 ***
## crim        -0.26018    0.04015  -6.480 2.20e-10 ***
## zn           0.07073    0.01548   4.570 6.14e-06 ***
## ptratio     -1.49367    0.17144  -8.712  < 2e-16 ***
## 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

# To decide which variables are important, we have to consider the P-values
of the coefficients in the coefficient table. A small P-value (generally less
than 0.05) implies a strong evidence against the null hypothesis of no
relationship and suggests that the predictor is statistically significant. In
the given model P-values of,intercept: < 2e-16 ***, crim: 2.20e-10 ***,
zn:6.14e-06 ***, ptratio: < 2e-16 ***,chas1: 0.000514 ***. All of the
predictors have p-values less than 0.05 which indicates that all variables in
the model are important in explaining the variation in the house prices.

# D) Use the anova analysis and determine the order of importance of these
four variables
# fit the linear regression models
model5 <- lm(medv ~ crim + zn + ptratio + chas, data = BostonHousing)
model6 <- lm(medv ~ crim + zn + ptratio, data = BostonHousing)
model7 <- lm(medv ~ crim + zn + chas, data = BostonHousing)
model8 <- lm(medv ~ zn + ptratio + chas, data = BostonHousing)
model9 <- lm(medv ~ crim + ptratio + chas, data= BostonHousing)

# perform anova analysis to compare the models
anova(model6, model5)

## Analysis of Variance Table
##
## Model 1: medv ~ crim + zn + ptratio
## Model 2: medv ~ crim + zn + ptratio + chas
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    502 28012
## 2    501 27345  1    667.19 12.224 0.0005137 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(model7, model5)

## Analysis of Variance Table
##
## Model 1: medv ~ crim + zn + chas
## Model 2: medv ~ crim + zn + ptratio + chas
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    502 31487
```

```
## 2    501 27345  1    4142.9 75.906 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model8, model5)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ zn + ptratio + chas
## Model 2: medv ~ crim + zn + ptratio + chas
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    502 29636
## 2    501 27345  1    2291.7 41.987 2.2e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model9, model5)
```

```
## Analysis of Variance Table
##
## Model 1: medv ~ crim + ptratio + chas
## Model 2: medv ~ crim + zn + ptratio + chas
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1    502 28485
## 2    501 27345  1    1140.1 20.889 6.138e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Answer:
# According to the ANOVA explanation,  by comparing the P-value we can
identify the importance of four variables. The smaller value is the most
significant variable in explaining the variation in the responsible variable.
When compare the model 6 and 5, the difference of the variable is chas. After
adding chas the p value is 0.0005137 ***. When compare the model7 and 5 the
difference is ptratio. After adding ptratio the p value is < 2.2e-16 ***.
When compare the model 8 and 5, the difference of the variable is crim. After
adding crim the p value is  2.2e-10 ***. When compare the model 9 and 5, the
difference of the variable is zn. After adding zn the p value is 6.138e-06
***. According to that to find the order of importance of each variable we
can arrange the P-values of each from lowest to the highest. Ptratio = 2.2e-
16 ***, crim= 2.2e-10 ***, zn = 6.138e-06 ***, chas = 0.0005137 *** likewise
we can order the most important variable to the least important variable
according to the ANOVA analysis.
```