regression_saisree

sai sree pulimamidi

2022-11-13

```
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)
```

Loading the required packages

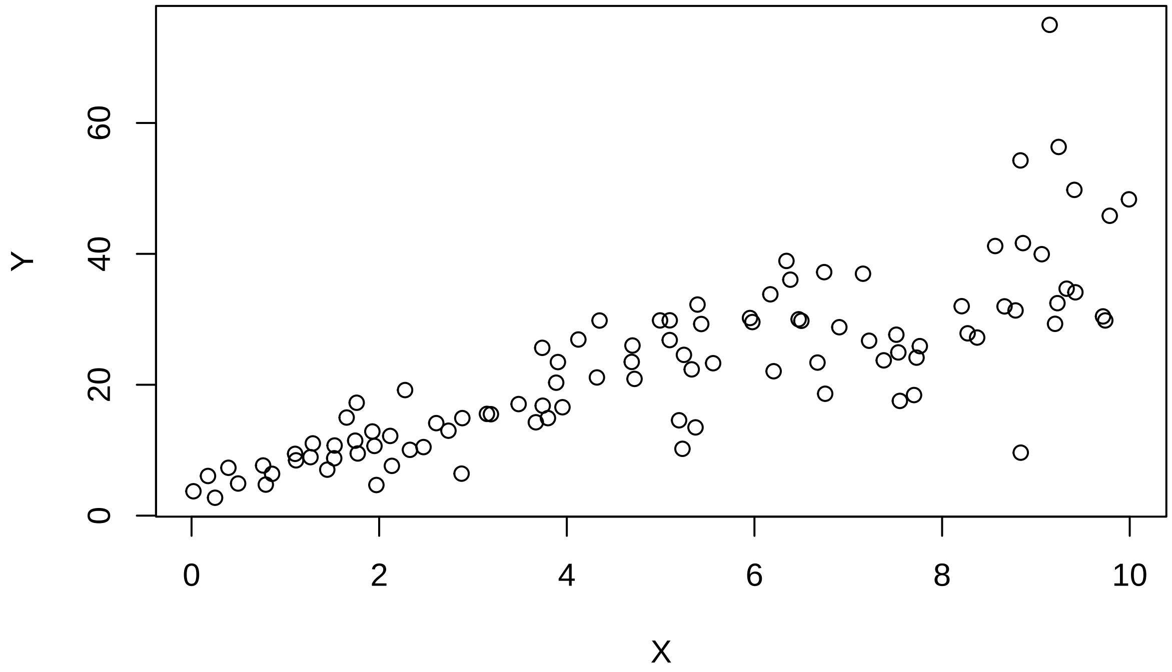
```
library("dplyr")
library("tidyverse")
library("mlbench")
library("tinytex")
```

1. Loading the given data

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

1(a). Plotting Y against X

```
plot(Y~X)
```



Based on the plot do you think we can fit a linear model to explain Y based on X?

Yes, I think that a linear model would be a good choice here to explain Y based on X because of the correlation which can be seen between the variables. As X tends to increase Y is seen to increase as well in the above graph, this indicates a positive correlation across the attributes.

1(b) & 1(c). Constructing a simple linear model of Y based on X

```
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
```

The equation that explains Y based on X is

$$\mathrm{Y} = 3.6108 \ast X + 4.4655$$

What is the accuracy of this model?

The accuracy of the model is explained by the Multiple R - Squared Value which is 0.6517 indicating that the model is 65.17% accurate. This also tells us that 65.17% of variation in Y can be explained by X.

Explain the relation between the Coefficient of Determination - R Squared of the model above with that to the correlation coefficient of X and Y

```
#Correlation coefficient of X and Y
cor(Y,X)^2
```

```
## [1] 0.6517187
```

```
#Coefficient of Determination - R Squared
summary(model)$r.squared
```

```
## [1] 0.6517187
```

In simple linear regression models which consist of only one independent variable and one dependent variable, the coefficient of determination is equal to the square of the correlation coefficient of the variable. Here in this case Y and X. Both the values of the coefficient of determination (r2) and the correlation coefficient of Y and X would be same.

2. Using the mtcars dataset

```
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
```

2(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.

Determining the Horse Power Basis the Weight

```
model_1 <- lm(hp~wt,data=mtcars)
summary(model_1)
```

```
##
## 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
```

James thinks that horse power (hp) of a car can be determined based on the weight of the car (wt), based on his thoughts we built the linear model to understand the predictive power of weight over horse power and got to see that 43.39% of the variability in horse power (hp) can be determined by the weight (wt).

Determining the Horse Power Basis the Mile Per Gallon

```r
model_2 <- lm(hp~mpg,data=mtcars)
summary(model_2)
```

```
##
## 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
```

While Chris thinks that horse power (hp) of a car can be determined based on the mile per gallon (mpg), based on his thoughts we built the linear model to understand the predictive power of mile per gallon over horse power and got to see that 60.24% of the variability in horse power (hp) can be determined by the mile per gallon (mpg).

Logically, speaking in terms of car and it's functionality horse power is a calculative measure which can be determined by the engine and it's size. So, miles per gallon actually makes more sense in determining the *horse power of a car. This can help us know that Chris's thoughts were right when compared to that with James.*

2(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).

```
model_3 <- lm(hp~cyl+mpg,data=mtcars)
summary(model_3)
```

```
##
## 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
```

Using this model, what is the estimated Horse Power of a car with 4 cylinder and mpg of 22?

```
predict(model_3,data.frame(cyl=c(4),mpg=c(22)))
```

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

The estimated horsepower (hp) with 4 cylinders (cyl) and with a mpg of 22 is "88.93618 hp".

$$Formula = \quad 23.979(4) - 2.775(22) + 54.067 = 88.933$$

3. Using the Boston Housing dataset from the "mlbench" package
Note: the dataset is old, hence low house prices

```
data(BostonHousing)
head(BostonHousing)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio      b lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
## 6 28.7
```

3(a). Building 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 of the tract bounds Chas River(chas).

```r
model_4 <- lm(medv~crim+zn+ptratio+chas,data=BostonHousing)
summary(model_4)
```

```
##
## 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
```

Is this an accurate model?

The multiple r squared value is 0.3599 which accounts to nearly 36%. This tells us that model is able to define 36% of the variability in owner-occupied homes (medv) based on the crime crate (crim), proportion of residential land zoned for lots over 25,000 sq.ft (zn), the local pupil-teacher ratio (ptratio) and the tract bounds along the Charles River (chas) thereby making it 36% accurate.

This won't be considered as a good model because the predictive power on the variability of the median value of occupied home is just 36% thereby falling short with a huge backlog in making predictions that are reasonably precise and this also doesn't solve the requirements of the business problem.

3(b)(i). Imagine two houses that are identical in all aspects but one bounds the Charles River and the other does not. Which one is more expensive and by how much?

Note: chas is a factorial variable, if the house is bound to Charles river then the value is going to be 1 and if the house is not bound to the Charles river the the value is going to be 0

Based on the above model built, if the house is bound to the Charles River then the price of that house is going to rise by 4.58393 in 1000$,

$$medv = intercept\ \ + crim\ \ + zn\ \ + ptratio\ \ + chas$$

$$\mathrm{med}v = 49.91868\ \ + (-0.26018)\ \ + 0.07073\ \ + (-1.49367)\ \ + 4.58393(1)$$

$$\mathrm{med}v = 52.81949\ \ in\ \ 1000\ \ dollars\ \ (if\ \ the\ \ house\ \ is\ \ bound\ \ to\ \ the\ \ charles\ \ river)$$

If the house is not going to be bound onto the Charles River then the price of the house is not going to rise by any value based on the chas variable.

$$\mathrm{med}v = 49.91868\ \ + (-0.26018)\ \ + 0.07073\ \ + (-1.49367)\ \ + 4.58393(0)$$

$$\mathrm{med}v = 48.23556\ \ in\ \ 1000\ \ dollars\ \ (if\ \ the\ \ house\ \ is\ \ not\ \ bound\ \ to\ \ the\ \ charles\ \ river)$$

The expensive house is going to be the one which is bound to the charles river by 4.58393 in 1000$ when compared to that with the house which is not bound to the charles river

3(b)(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?

Pupil to Teacher ratio is basically to how many students a teacher has been allocated and the lower this is the expensive the communities or the houses are going to be because people who are economically stable and above the median income class would want to have their children study in less crowded school.

For 15 units of change in the pupil teacher ratio the price of the house is going to change by 15(-1.49367) = -22.40505 in 1000$.

$$medv = intercept\ \ + crim\ \ + zn\ \ + ptratio\ \ + chas$$

$$\mathrm{med}v = 49.91868\ \ + (-0.26018)\ \ + 0.07073\ \ + 15(-1.49367)\ \ + 4.58393$$

$$\mathrm{med}v = 31.90811\ \ in\ \ 1000\ \ dollars\ \ (if\ \ the\ \ pupil\ \ teacher\ \ ratio\ \ is\ \ 15)$$

If the units change to 18 then the price of the house is going to change by 18(-1.49367) = -26.88606 in 1000$.

$$\mathrm{med}v = 49.91868\ \ + (-0.26018)\ \ + 0.07073\ \ + 18(-1.49367)\ \ + 4.58393$$

$$\mathrm{med}v = 27.4271\ \ in\ \ 1000\ \ dollars\ \ (if\ \ the\ \ pupil\ \ teacher\ \ ratio\ \ is\ \ 18)$$

The expensive house is going to be where the pt ratio is 15 with a difference of 4.48101 in 1000$ with that to the pt ratio of 18

3(c). Which of the variables are statistically important (i.e. related to the house price)?

It's interesting to see that all the independent variables which are helping determine the median value of the owner occupied home (medv) are shown to be statistically significant between 0 and 0.001. These variables are crime rate (crim), the local pupil-teacher ratio (ptratio), proportion of residential land zoned for lots over 25,000 sq.ft (zn) and the tract bounds along the Charles River (chas).

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

```
imp_var <- anova(model_4)
imp_var
```

```
## Analysis of Variance Table
##
## Response: medv
##            Df  Sum Sq Mean Sq F value    Pr(>F)    
## crim        1  6440.8  6440.8 118.007 < 2.2e-16 ***
## zn          1  3554.3  3554.3  65.122 5.253e-15 ***
## ptratio     1  4709.5  4709.5  86.287 < 2.2e-16 ***
## 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
```

Based on the anova analysis the order of importance is determined by using Sum Sq (variability) and the order is as follows:

1. crime crate (crim) - 6440.8

2. the local pupil-teacher ratio (ptratio) - 4709.5

3. proportion of residential land zoned for lots over 25,000 sq.ft (zn) - 3554.3

4. the tract bounds along the Charles River (chas) - 667.2