Regression

Karthik Badiganti

2022-11-11

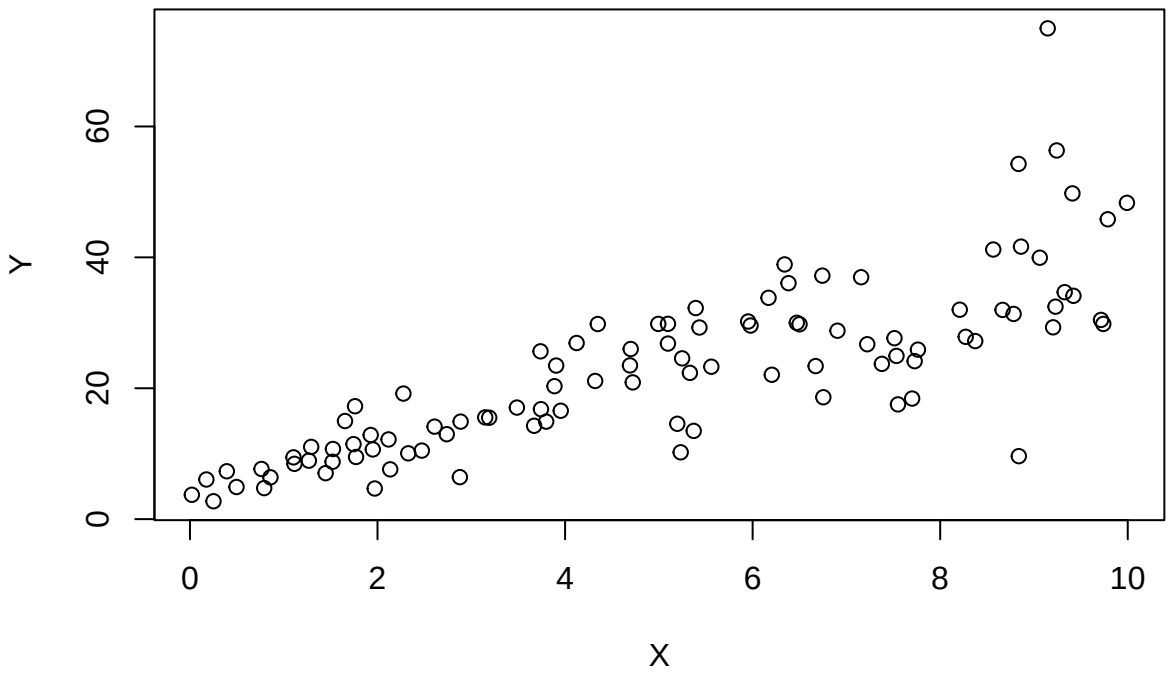
Loading Packages

```
library(mlbench)
```

```
## Warning: package 'mlbench' was built under R version 4.2.2
```

Q1

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




```
# b
Model=lm(Y~X)
s<-summary(Model)
s
```

```
##
## 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
```

```
#c
print(paste('The r square of the above model is',s$r.squared))
```

```
## [1] "The r square of the above model is 0.651718678875602"
```

```
print(paste('The correlation of the model is',cor(X,Y)))
```

```
## [1] "The correlation of the model is 0.807290950572098"
```

```
print(paste('When we take a square root of r.squared we get',sqrt(s$r.squared)))
```

```
## [1] "When we take a square root of r.squared we get 0.807290950572098"
```

From the above we consider square root of r-squared as positive because it has a positive slope which we observe from the plot which in turn illustrates that X increases with respect to Y. **Hence we can say that, 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.**

Q2

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.

```
summary(lm(mtcars$hp~mtcars$wt))$call
```

```
## lm(formula = mtcars$hp ~ mtcars$wt)
```

```
print(paste
      ('The accuracy by using James opinion of weight as independent variable is'
        ,round(((summary(lm(mtcars$hp~mtcars$wt))$r.squared)*100),2),'%'))
```

```
## [1] "The accuracy by using James opinion of weight as independent variable is 43.39 %"
```

```
summary(lm(mtcars$hp~mtcars$mpg))$call
```

```
## lm(formula = mtcars$hp ~ mtcars$mpg)
```

```
print(paste
      ('The accuracy by using Chris opinion of mpg as independent variable is'
        ,round(((summary(lm(mtcars$hp~mtcars$mpg))$r.squared)*100),2),'%'))
```

```
## [1] "The accuracy by using Chris opinion of mpg as independent variable is 60.24 %"
```

We can see that Chris opinion that horse power depends on mpg is more accurate with **60.24%** than James opinion of weight which has only **43.39% accuracy.**

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?***

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

print(paste('Accuracy of the model with cyl and mpg as independent variables is'
            ,round((summary(mcars)$r.squared)*100,2),'%'))
```

```
## [1] "Accuracy of the model with cyl and mpg as independent variables is 70.93 %"
```

```
hp_predict<-predict(mcars,data.frame(cyl=4,mpg=22))
print(paste('The predicted horse power for cyl=4 and mpg=22 is',hp_predict[[1]]))
```

```
## [1] "The predicted horse power for cyl=4 and mpg=22 is 88.9361796789223"
```

Q3

```
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
```

a. ***Estimate the median value of owner-occupied homes (medv)based on the following variables:***

crime rate (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)

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

```
##
## 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
```

From the above we can observe based on multiple r-squared value that the model explains with **accuracy of 35.99%.** Based on the accuracy r.square value we can say that the model can explain with nearly 36% which might not be considered as a good model.

b

i. *When all the parameters are constant and only the buildings differ by tract bounds?*

Let us Consider a real time example where two buildings with same crime rate = 0.02732, zn=17,ptratio of 15.5 and one build tracts the river and one doesn't. then below are the house prices for them.

```
# house does not tract river
chas_0 <-predict(bh,data.frame(crim=0.02732,zn=17,ptratio=15.5,chas='0'))
print(paste('The price of the house that is not by the river is',chas_0[[1]],'in 1000 dollars'))
```

```
## [1] "The price of the house that is not by the river is 27.9620295293422 in 1000 dollars"
```

```
# house is along side river
chas_1 <-predict(bh,data.frame(crim=0.02732,zn=17,ptratio=15.5,chas='1'))
print(paste('The price of the house that is by the river is',chas_1[[1]],'in 1000 dollars'))
```

```
## [1] "The price of the house that is by the river is 32.5459554392892 in 1000 dollars"
```

We can observe that the house by the river has **increase of price by $4583.926** than the house that is not by the river.

ii. ***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?***

Let us Consider a real time example where two buildings with same crime rate = 0.02732,zn=17, chas =1 and one building is in place where ptratio is 15 and other with 18

```
ptratio_15<-predict(bh,data.frame(crim=0.02732,zn=17,ptratio=15,chas='1'))
print(paste('The house with ptratio 15 has price of ',ptratio_15[[1]],'in 1000 dollars'))
```

```
## [1] "The house with ptratio 15 has price of  33.2927917120431 in 1000 dollars"
```

```
ptratio_18<-predict(bh,data.frame(crim=0.02732,zn=17,ptratio=18,chas='1'))
print(paste('The house with ptratio 18 has price of ',ptratio_18[[1]],'in 1000 dollars'))
```

```
## [1] "The house with ptratio 18 has price of  28.8117740755195 in 1000 dollars"
```

We can observe that house with low ptratio has higher price than house with high ptratio. In the above example we can see that there is an **decrease of price by $4481.01** when ptratio is increased.

c. *Which of the variables are statistically important ?*

```
summary(bh)
```

```
##
## 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
```

Based on the p-values of the coefficients above, we can see that p-values for all the independent variables are **statistically significant between 0 and 0.001**

d. ***Use the anova analysis and determine the order of importance of these four variables.***

```
anova(bh)
```

```
## 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 sum squares values of the independent variables, we can determine the order of their importance as below,

1. Crime rate - 6440.8
2. 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