Regression Analytics

Shivani Haridas Pitla

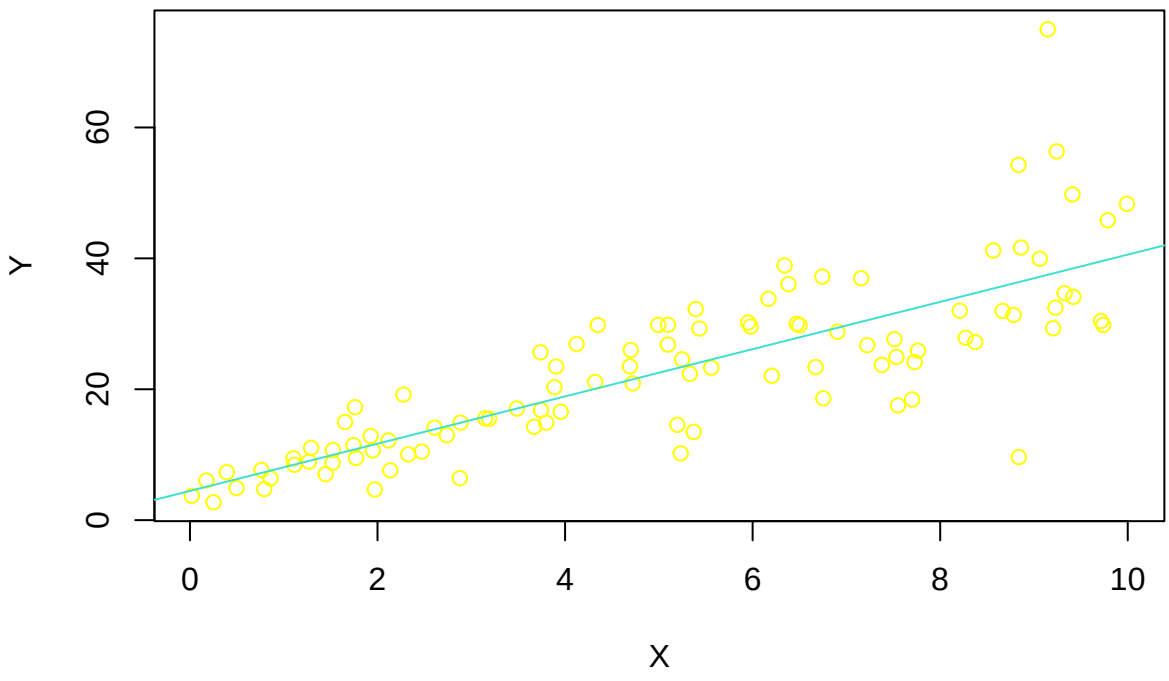
2022-11-13

1. Run the following code in R-studio to create two variables X and Y.set.seed(2017) X=runif(100)10 Y=X4+3.45 Y=rnorm(100)0.29Y+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 modfit a linear model to explain Y based on X?

```
plot(Y~X,xlab='X',ylab='Y',col='yellow')
abline(lsfit(X, Y),col = "turquoise")
```



From the plot it can be seen there exists correlation between the variables "x" and "y", hence linear model would be a good fit. ***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?*** The linear model of Y based on X is given by the equation Y=4.4655+3.6108*X and the accuracy of this model is 0.6517 or 65% .This also reveals that X can account for 65.17 percent of the variation in Y.

```
linearmodel <- lm(Y ~ X)
summary(linearmodel)
```

```
##
## 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) 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
```

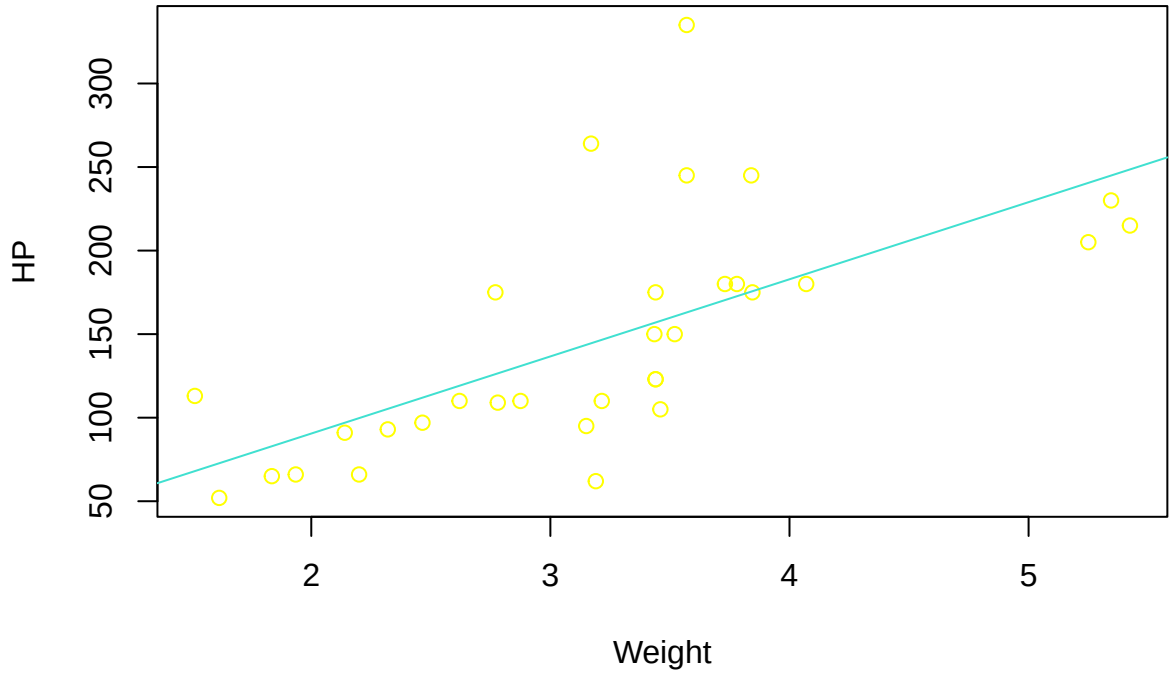
solution:The coefficient of determination is equal to the square of the correlation coefficient. The correlation coefficient between Y and X and the coefficient of determination (r2) would both have the same values. ***2. We will use the 'mtcars' dataset for this question. The dataset is already included 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) of the 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. constructing a model from James' estimation

```
plot(mtcars$hp~mtcars$wt,xlab='Weight',ylab='HP',col='yellow')
abline(lsfit(mtcars$wt,mtcars$hp),col = "turquoise")
```




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

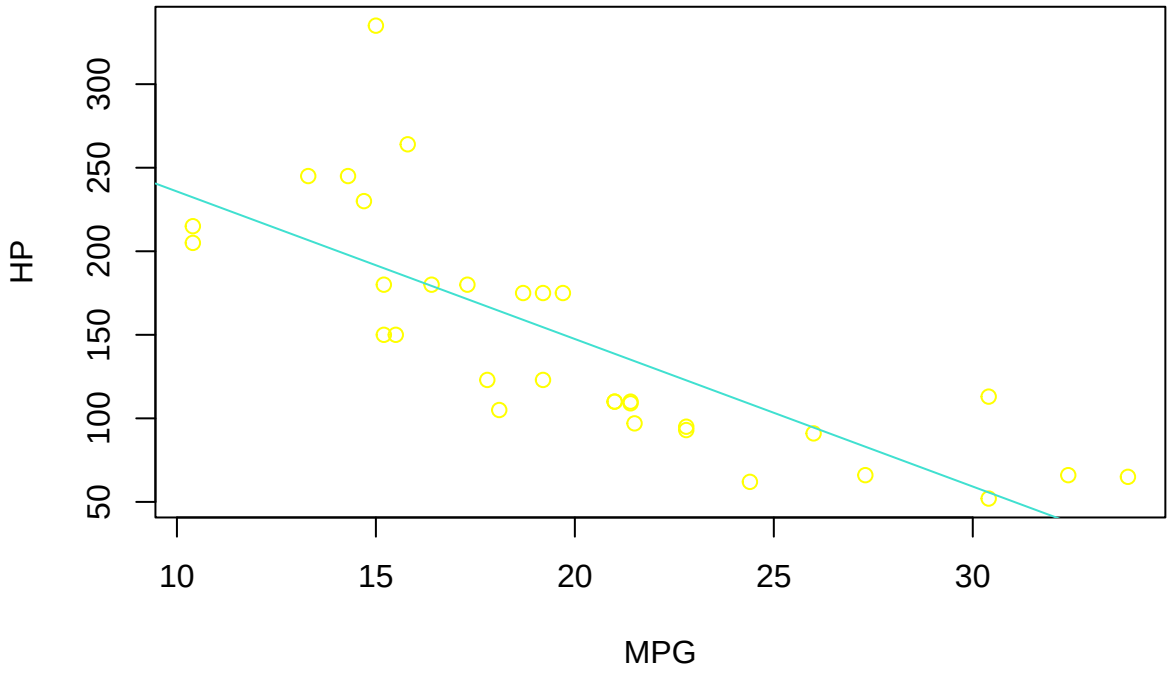
```
##
## Call:
## lm(formula = hp ~ wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.430 -33.596 -13.587   7.913 172.030
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.821     32.325  -0.056    0.955
## wt            46.160      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
```

Jamesmodel has an accuracy of 0.4339

Constructing a model using Chris estimate

```
plot(mtcars$hp~mtcars$mpg,xlab='MPG',ylab='HP',col='yellow')
abline(lsfit(mtcars$mpg, mtcars$hp),col = "turquoise")
```




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

```
##
## 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
```

The Chrismodel has an accuracy of 0.6024. Results: Chris The estimate is fairly correct. Chris is therefore right. ***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?***

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

```
##
## 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
```

```
est.hp<-predict(CHP,data.frame(cyl=4,mpg=22))
est.hp
```

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

The estimated Horse Power is 88.93618 ***3. For this question, we are going to use BostonHousing dataset. The dataset is in 'mlbench' package, so we first need to install the package, call the library and load the dataset using the following commands***

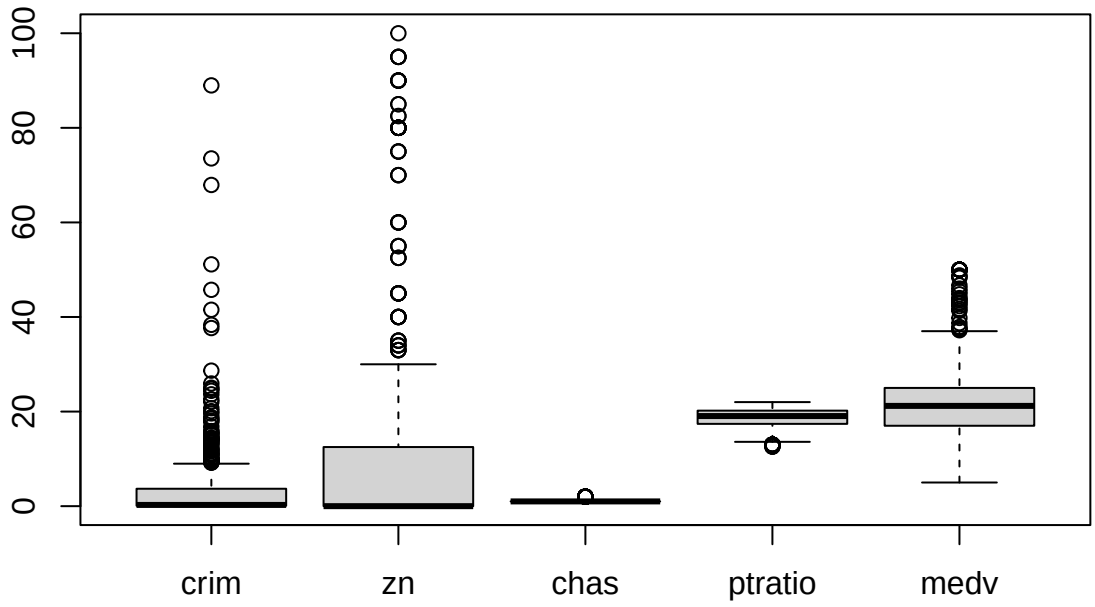
```
#installing required packages
library(mlbench)
```

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

```
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 ...
```

```
#Let's examine how the values of the various variables in the dataset have changed over time.
#Plotting a boxplot of all the variables is done to achieve this.
boxplot(BostonHousing[,c(1,2,4,11,14)])
```



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?

```
set.seed(310)
OH<-lm(medv~crim+zn+ptratio+chas,data = BostonHousing)
summary(OH)
```

```
##
## 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
```

The model's accuracy is 0.3599. This model is quite inaccurate. ***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? Answer:The factor Chas has two levels, "0" and "1". The number "1" stands for those who border the Chas River, while the number "0" is for those who do not. The data description states that the median value of owner-occupied homes is $1,000, and the chas1 coefficient is 4.58393. The result of multiplying by the coefficient is the pricey number 4583.93. ***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?*** Answer:For every unit increase in the ptratio, the price of a property falls by 1.49367(in thousands). The fall will be 15 * 1493.67=22405.05., if ptratio is 15. The reduction will be 18 * 1493.67=26886.06 if ptratio is 18, for example. Therefore, if the pt ratio is 15, it will cost $4481.01 more than if it is 18. ***c) Which of the variables are statistically important (i.e. related to the house price)?*** Answer:Indicating that we can safely reject the default null hypothesis as there is no link between house price and other factors in the model—the p-values for all variables are not equal to zero. Therefore, each variable has statistical significance. ***d) Use the anova analysis and determine the order of importance of these four variables.***

```
anova(OH)
```

```
## 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
```

Answer:It is clear that the crim variable explains a proportionally larger amount of variability (sum squared) than other variables. We can generalize that including the crim made the model significantly better. Residuals, however, show that a significant chunk of the variability is unexplained. The rankings are crim, ptratio, zn, and chas.