Regression Analytics

Nikhil Reddy Addula

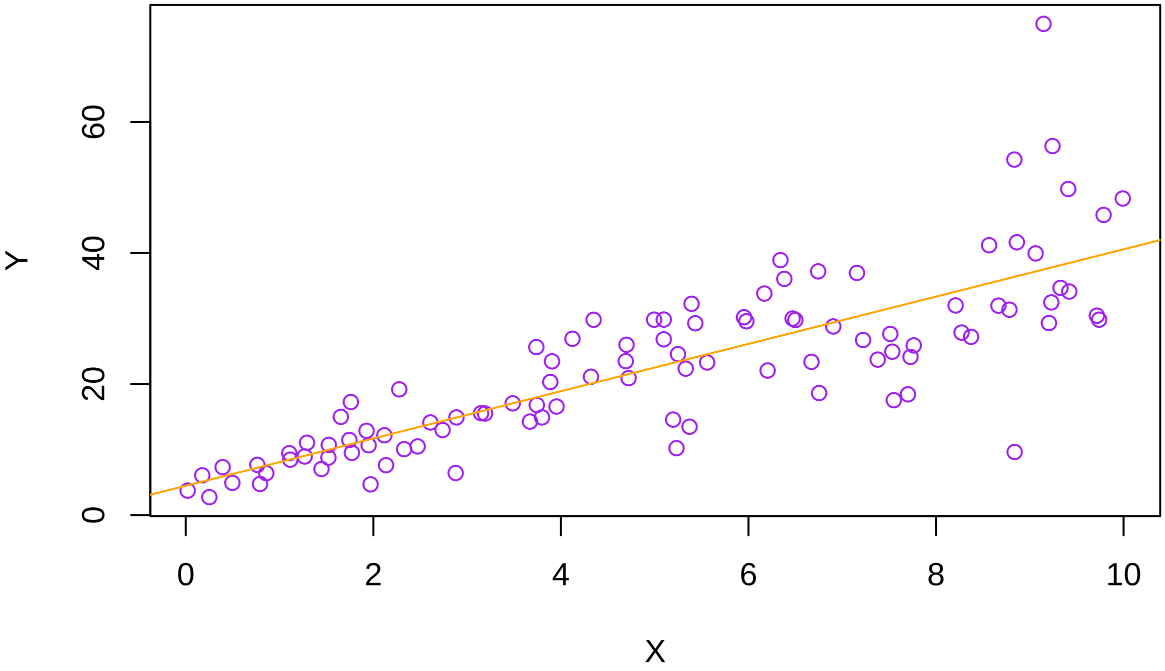
2022-11-13

1)Run the following code in R-studio to create two variables X and Y.

```
set.seed(2017)
X=runif(100)*10
Y=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(Y~X,xlab ='X',ylab = 'Y',col='purple')
abline(lsfit(X,Y),col='orange')
```



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?

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

```
##
## 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=4.4655+3.6108*X Accuracy is 0.6517 or 65%

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
```

The square of correlation coefficient is same as coefficient of determination 65.17% #Coefficient of Determination= (Correlation Coefficient)2

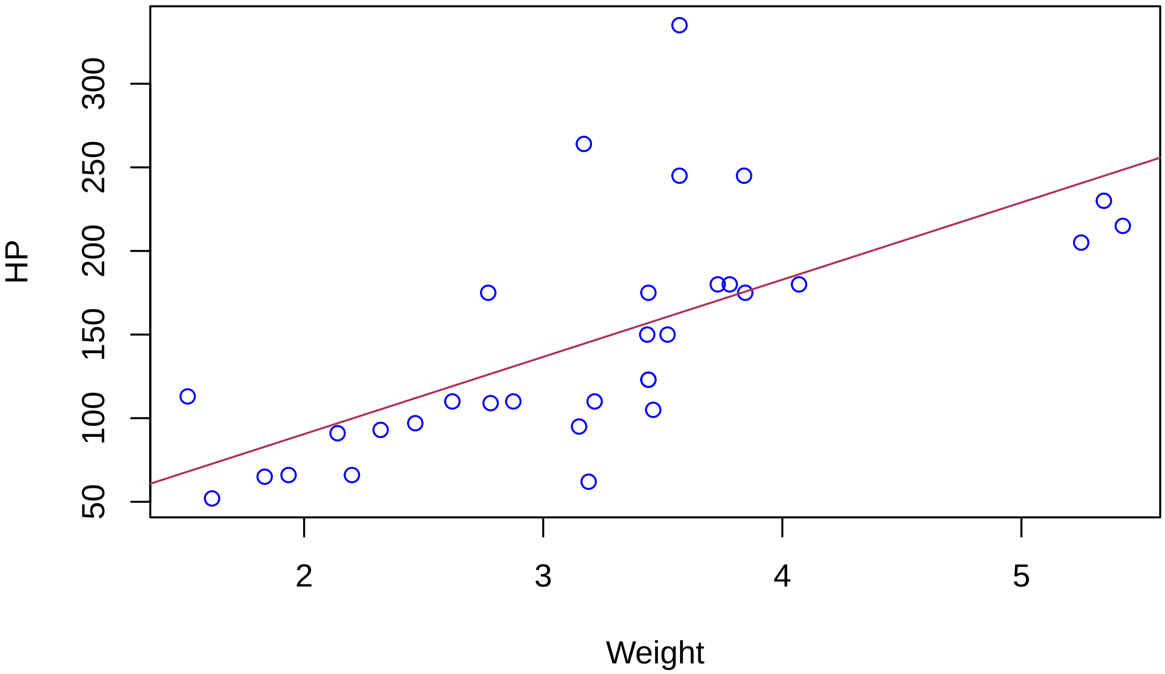
2) We will use the 'mtcars' dataset for this question. The dataset is already included in your R distribution. The dataset shows some of the characteristics of different cars. The following shows few samples (i.e. the first 6 rows) of the dataset.

```
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.

```
# Creates a linear model for weight vs horsepower and displays a plot of the points
plot(mtcars$hp~mtcars$wt,xlab ='Weight',ylab = 'HP', col='blue')
abline(lsfit(mtcars$wt,mtcars$hp),col= 'maroon')
```

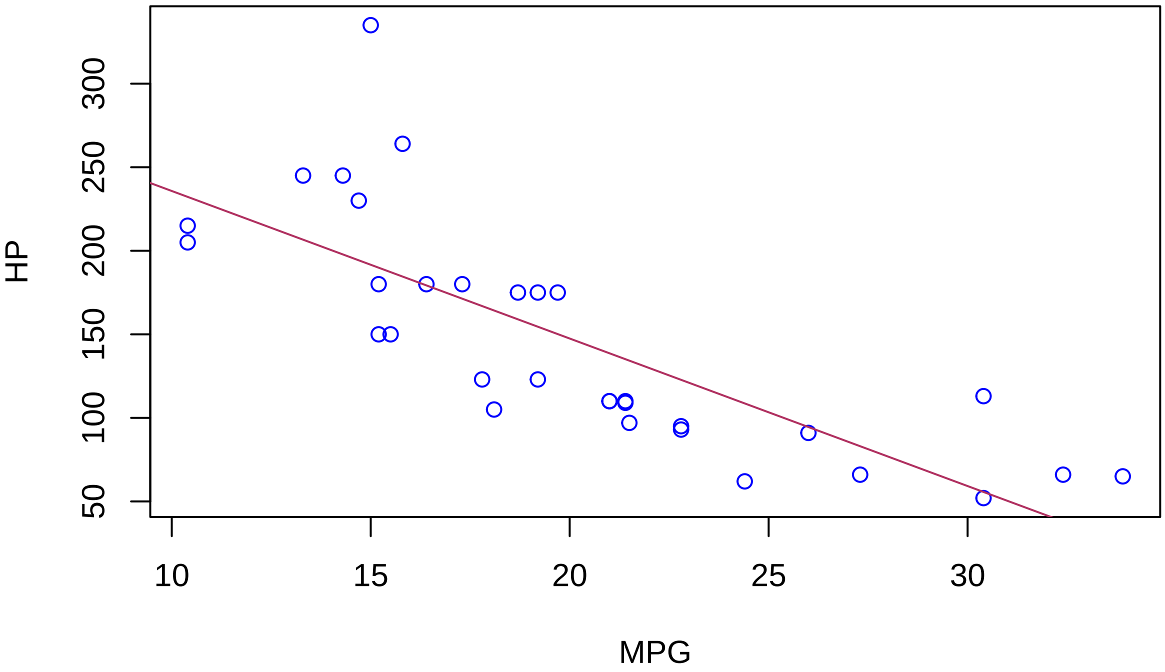


```
lp_model1 <- lm(formula = hp~wt, data = mtcars)
summary(lp_model1)
```

```
##
## 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 we can see that weight results in a model that accounts for 43.39% of the variation in horsepower.

```
# Creates a linear model for mpg vs horsepower and displays a plot of the points
plot(mtcars$hp~mtcars$mpg,xlab ='MPG',ylab = 'HP', col='blue')
abline(lsfit(mtcars$mpg,mtcars$hp),col= 'maroon')
```



```
lp_model2 <- lm(formula = hp~mpg, data = mtcars)
summary(lp_model2)
```

```
##
## 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
```

We can see that a model that incorporates fuel efficiency yields one that explains 60.24% of the variation in horsepower. Fuel economy (mpg) is therefore regarded in this model as statistically significant.

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?

```
LP_Model3 <-lm(hp~cyl+mpg,data=mtcars)
summary(LP_Model3)
```

```
##
## 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
```

```r
# Predict the estimated horse power of a car with 4 cylinders and 22 mpg
predict(LP_Model3, data.frame(mpg = 22, cyl = 4 ))
```

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

The estimated Horse Power of a car with 4 calendar and mpg of 22 is 88.93%

3) 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

```r
#install.packages('mlbench')
library(mlbench)
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)

```r
# Create a linear model for median value based on crim, zn, ptratio, and chas.
set.seed(123)
LP_Model4<-lm(medv~crim+zn+ptratio+chas,data = BostonHousing)
summary(LP_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
```

The variability in the median house value was 35.99% of the R2 value in this model (crim, zn, ptratio, and chas). In terms of accuracy, this is a weak model that might be strengthened by including more 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?

Answer:Based on the coefficients, the resulting formula from our model is: medv = 49.91868 - 0.26018crim + 0.07073zn - 1.49367ptratio + 4.58393chas1 Therefore, if the only difference between two houses is that one borders the Chas River, then we would only focus on the chas variable coefficient. The house that borders the river would be $4,583.93 more than the one that does not. 4.58393 (coeff of chas) * 1 (value of chas) * 1000 (medv in $1,000 units) = $4,583.93

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: Based on the coefficients, the resulting formula from our model is: medv = 49.91868 - 0.26018crim + 0.07073zn - 1.49367ptratio + 4.58393chas1 Therefore, if the only difference between two houses is the pupil-teacher ratio, then we would only focus on the ptratio variable coefficient. As a result, the house with the smaller pupil-teacher ratio value would be more expensive, because the coefficient is found to be negative in our model. The difference in values between the houses would be: -1.49367 (coeff of ptratio) * 0.03 (difference between ptratio values) * 1000 (medv in $1,000 units) = $44.81 Therefore, the house with the lower pupil-teacher ratio would be $44.81 more expensive based on our model.

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.

Based on the model developed from these variables, it was shown that all of the variables (crim, zn, ptratio, and chas) are statistically significant. This is true since no p-values from our model's results went over the 0.05 level of significance.

d) Use the anova analysis and determine the order of importance of these four variables.(5 marks)

```r
# Returns the ANOVA results for the model used in this problem
anova(LP_Model4)
```

```
## 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
```

The significance of these factors is ranked in the following order according to the ANOVA values:

1. “Crim” - 15.08% of the model’s variability is explained by this term.
2. “ptratio” - accounts for 11.02% of the model’s variability.
3. “Zn” - accounts for 8.32% of the model’s variability.
4. “chas” - accounts for 1.56% of the model’s variability.

Furthermore, the residuals in this model still contribute 64.01 percent of its variability, indicating that there is still much space for this model’s accuracy to be enhanced.