Assignment\_3\_Regression

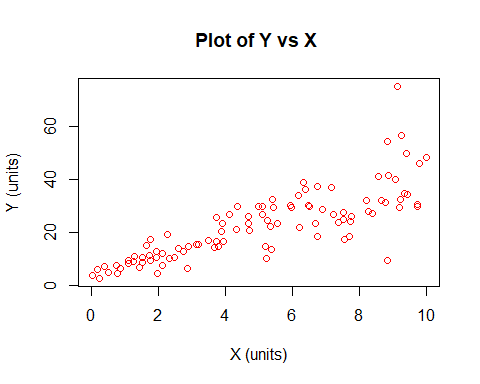
2023-11-05

#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(X,Y, xlab = "X (units)", ylab = "Y (units)", main = "Plot of Y vs X ", col = "red")

 ##There is a positive linear trend between X variables and Y variables.

#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?

linearModel <- lm(Y~X)  
linearModel$coefficients

## (Intercept) X   
## 4.465490 3.610759

# c)How the Coefficient of Determination, R 2, of the model above is related to the correlation coefficient of X and Y?

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

#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)

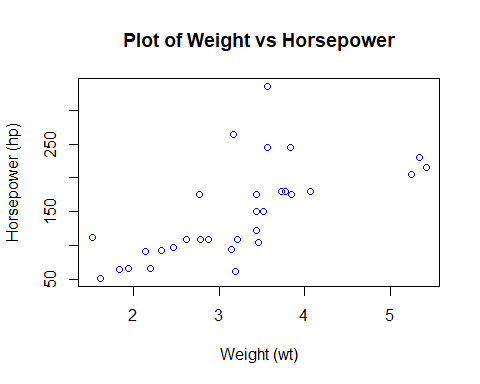
## mpg cyl disp hp drat wt qsec vs am gear carb  
## Mazda RX4 21.0 6 160 110 3.90 2.620 16.46 0 1 4 4  
## Mazda RX4 Wag 21.0 6 160 110 3.90 2.875 17.02 0 1 4 4  
## Datsun 710 22.8 4 108 93 3.85 2.320 18.61 1 1 4 1  
## Hornet 4 Drive 21.4 6 258 110 3.08 3.215 19.44 1 0 3 1  
## Hornet Sportabout 18.7 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 HorsePower 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.

linearmodel2 = lm(hp~wt, data = mtcars)  
summary(linearmodel2)

##   
## 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

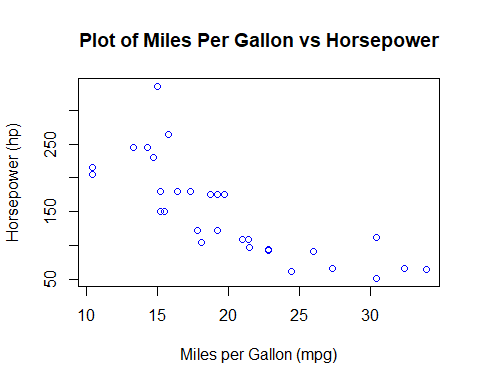
plot(mtcars$wt,mtcars$hp, xlab = "Weight (wt)", ylab = "Horsepower (hp)", main = "Plot of Weight vs Horsepower", col = "blue")



# Creating a linear model for mpg vs horsepower and displays a plot of the points  
linearModel3 = lm(hp~mpg, data = mtcars)  
summary(linearModel3)

##   
## 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

plot(mtcars$mpg,mtcars$hp, xlab = "Miles per Gallon (mpg)", ylab = "Horsepower (hp)", main = "Plot of Miles Per Gallon vs Horsepower", col = "blue")



#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?

# Shows which variables are factor or numeric  
str(mtcars)

## 'data.frame': 32 obs. of 11 variables:  
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
## $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...  
## $ disp: num 160 160 108 258 360 ...  
## $ hp : num 110 110 93 110 175 105 245 62 95 123 ...  
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...  
## $ qsec: num 16.5 17 18.6 19.4 17 ...  
## $ vs : num 0 0 1 1 0 1 0 1 1 1 ...  
## $ am : num 1 1 1 0 0 0 0 0 0 0 ...  
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...  
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...

# Converting cylinder into a factor  
mtcars$cyl = as.factor(mtcars$cyl)  
str(mtcars)

## 'data.frame': 32 obs. of 11 variables:  
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...  
## $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...  
## $ disp: num 160 160 108 258 360 ...  
## $ hp : num 110 110 93 110 175 105 245 62 95 123 ...  
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...  
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...  
## $ qsec: num 16.5 17 18.6 19.4 17 ...  
## $ vs : num 0 0 1 1 0 1 0 1 1 1 ...  
## $ am : num 1 1 1 0 0 0 0 0 0 0 ...  
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...  
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...

linearModel4 = lm(hp~cyl+mpg, data = mtcars)  
summary(linearModel4)

##   
## Call:  
## lm(formula = hp ~ cyl + mpg, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -58.882 -20.904 -6.261 7.043 125.453   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 171.349 57.946 2.957 0.00625 \*\*  
## cyl6 16.623 23.197 0.717 0.47955   
## cyl8 88.105 28.819 3.057 0.00487 \*\*  
## mpg -3.327 2.133 -1.560 0.12995   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 37.01 on 28 degrees of freedom  
## Multiple R-squared: 0.7368, Adjusted R-squared: 0.7086   
## F-statistic: 26.12 on 3 and 28 DF, p-value: 2.888e-08

# Predict the estimated horse power of a car with 4 cylinders and 22 mpg  
predict(linearModel4, data.frame(mpg = c(22), cyl = c("4")))

## 1   
## 98.15275

#With four calendar months and 22 mpg, the estimated horsepower of the car is 98.15%.

#3) Using BostonHousing Dataset

library(mlbench)

## Warning: package 'mlbench' was built under R version 4.3.2

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 tractbounds Chas River(chas). Is this an accurate model?

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 ...

# Creating a linear model for median value based on crim, zn, ptratio, and chas.  
linearModel5 = lm(medv~crim+zn+ptratio+chas, data = BostonHousing)  
summary(linearModel5)

##   
## 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

This model’s (crim, zn, ptratio, and chas) R2 value is 35.99% of the variance in the median home value. The accuracy of this model is poor, and it 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? Based on the coefficients, the resulting formula from our model is:-medv = 49.91868 - 0.26018*crim + 0.07073*zn - 1.49367*ptratio + 4.58393*chas1.

Therefore, if the only difference between two houses is that one borders the Chas River, then we 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 Based on the coefficients, the resulting formula from our model is: medv = 49.91868 - 0.26018*crim + 0.07073*zn - 1.49367*ptratio + 4.58393*chas1.

Thus, we concentrate on the ptratio variable coefficient if the pupil-teacher ratio is the only distinction between two houses. The dwelling with the lower pupil-teacher ratio number is therefore more costly, as our model’s coefficient turns out to be negative. The houses’ valuations diverge to the extent that

-1.49367 (coeff of ptratio) \* 0.03 (difference between ptratio values) \* 1000 (medv in $1,000 units) = $44.81

According to our calculations, the home with the lower student-teacher ratio is therefore $44.81 more expensive.

#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.

All of the variables—crim, zn, ptratio, and chas—were determined to be statistically significant based on the model built using these data. This is confirmed by the fact that every p-value derived from our model was less than the significance threshold of 0.05.

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

anova(linearModel5)

## 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 returned ANOVA values we can determine the order of importance, for these variables;

1. The variable “crim” explains 15.08% of the variability in the model.
2. Following that the variable “ptratio” accounts for 11.02% of the models variability.
3. The “zn” variable contributes 8.32% to the models variability.
4. Lastly we have the “variable, which explains 1.56% of the models variability.

Despite these findings it is important to note that residuals still account for a portion (around 64.01%) of variability within this model. This suggests that there is room for improvement, in terms of accuracy.