Business Analytics Assignment\_3

ALLEN RICHARDS

2023-11-05

library(ggplot2)

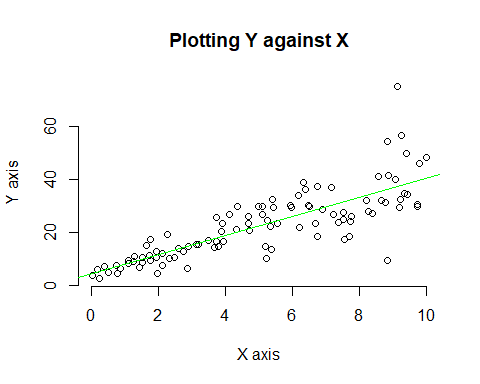
## Warning: package 'ggplot2' was built under R version 4.3.2

# 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 against X.  
plot(X,Y, main = "Plotting Y against X", xlab = "X axis", ylab = "Y axis",frame = FALSE)  
# Add regression line  
abline(lm(Y~X), col = "green")



# Interpretation:

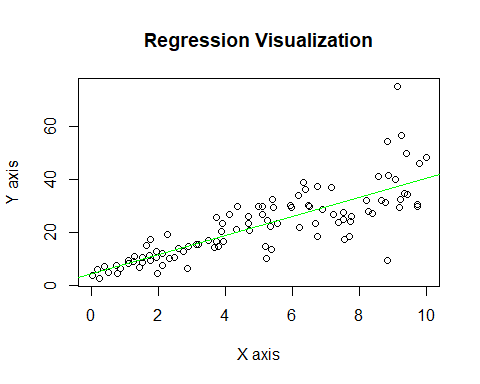
As per my view the exists correlation between the variables “x” and “y” from the plot. So the linear model would be a good fit for this.

# 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?

# Y=4.4655+3.6108\*X  
# Accuracy is 0.6517 or 65%  
linear\_mod <- lm(Y~X)  
summary(linear\_mod)

##   
## 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

# Regression visualization  
plot(X, Y, xlab = "X axis",   
 ylab = "Y axis",  
 main = "Regression Visualization")  
abline(4.4655, 3.6108, col = "green")



# 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

# Interpretation:

The coefficient of X and Y is similar to the coefficient of determination, , of the above model is around 65%

# Question 2:

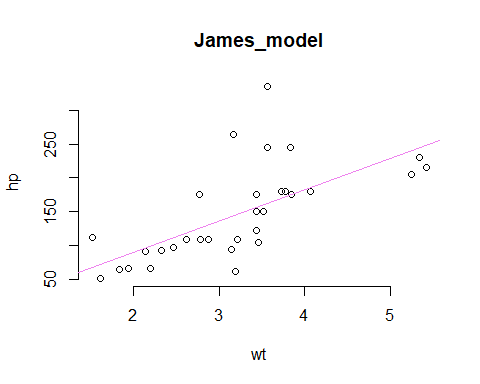
# a)James wants to buy a car. He and his friend, Chris, have different opinions about the Horse Power (hp) of cars. James think the weight of a car (wt) can be used to estimate the Horse Power of the car while Chris thinks the fuel consumption expressed in Mile Per Gallon (mpg), is a better estimator of the (hp). Who do you think is right? Construct simple linear models using mtcars data to answer the question.

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

Simple Linear model according to James

plot(mtcars$wt, mtcars$hp, main = "James\_model",  
 xlab = "wt", ylab = "hp",frame = FALSE)  
# Add regression line  
abline(lm(mtcars$hp ~ mtcars$wt), col = "violet")



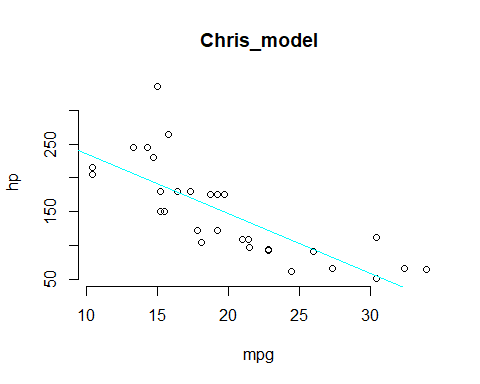
James\_model <- lm(formula = hp~wt, data = mtcars)  
summary(James\_model)

##   
## 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

# Accuracy of James\_model is 0.4339

Simple Linear model according to Chris

plot(mtcars$mpg, mtcars$hp, main = "Chris\_model",  
 xlab = "mpg", ylab = "hp",frame = FALSE)  
# Add regression line  
abline(lm(mtcars$hp ~ mtcars$mpg), col = "cyan")



Chris\_model <- lm(formula = hp~mpg, data = mtcars)  
summary(Chris\_model)

##   
## 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

# Accuracy of Chris\_model is 0.6024

# Interpretation:

However, the value of Chris\_model is accurate. So, Chris’s opinion is correct because the fuel consumption expressed in Mile Per Gallon (mpg) and it is a better estimation of the (hp).

# 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?

H\_model <- lm(formula = hp~cyl+mpg, data = mtcars)  
summary(H\_model)

##   
## 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

Esti\_hp <- predict(H\_model,data.frame(cyl=4,mpg=22))  
Esti\_hp

## 1   
## 88.93618

# Interpretation:

88.93 is the estimated Horse Power of a car with cylinder = 4 and mpg = 22

# Question 3

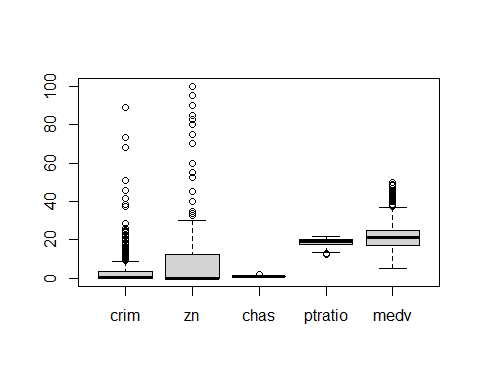
library(mlbench)

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

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

# Plotting all variables using box plot to observe how the values of the various variables in the dataset have changed over time.   
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?(Hint check R2 )

set.seed(125)  
owner\_model <- lm(formula = medv~crim+zn+ptratio+chas,data = BostonHousing)  
summary(owner\_model)

##   
## 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

owner\_model

##   
## Call:  
## lm(formula = medv ~ crim + zn + ptratio + chas, data = BostonHousing)  
##   
## Coefficients:  
## (Intercept) crim zn ptratio chas1   
## 49.91868 -0.26018 0.07073 -1.49367 4.58393

# Interpretation:

The model is not accurate because the accuracy of owner\_model is 0.3599.

# 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?

4.58393 is the estimated coefficient of Chas. Chas river is the factor of two variable 0 and 1, It is given that the median value of owner-occupied homes is 1000 dollars and one bound Chas River is 1 and if it doesn’t it is 0. when multiplied with coefficient(4.58393\*1000 = 4583.93$) So it is expensive.

# 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? (Golden Question)

Generally every single unit increase in pupil-teacher ratio, price of houses is decreased by 1.49367 (i.e) 1493.67 (in thousands).If pupil-teacher ratio is 15, then it will be decrease of 15 \* 1493.67= 22405.05. Moreover, if pupil-teacher ratio is 18 then it will be a decrease of 18\*1493.67 = 26886.06. Finally, if pupil-teacher ratio of 15 is expensive by $4481.01 while it is compared to pupil-teacher ratio of 18.

# 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.

The P-values are not equal to 0. So the null hypothesis rejected and concluded that there is no relationship between house price and other factors in this model. Also, each variable has statistical significance.

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

anova(owner\_model)

## 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

# Interpretation:

We can see that the “crim” variable explains substantially more variability (sum squared) than the other variables. So this could be explained by this model being enhanced. However, residuals demonstrate that a significant fraction of the variability is unaccounted. The order of importance is crim,ptratio,zn,chas So we can fulfill the requirement with these values.