Assignment.3-Regression

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.5   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

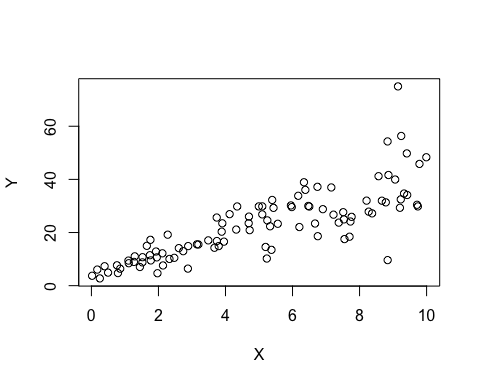
library(mlbench)  
library(tinytex)

#1. 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

#1.a)Based on the plot do you think we can fit a linear model to explain Y based on X?

plot(Y~X)



#From above plot we can observe, the relationship between X and Y shows linear regression pattern. As X increases, Y also increases which states X and Y variables have a positive relationship.

#1.b)Simple linear model of Y based on X

Model=lm(Y ~X)  
  
summary(Model)

##   
## 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

#What is the accuracy of this model? #According above summary, accuracy of model can be derived from R-squared value. R-squared value of above summary is 0.6517. Therefore the accuracy of model is 65.17%.

#Write the equation that explains Y based on X? #Y=3.6108\*X+4.4655

#1.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

#2. Using mtcars 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

#2.a) Building a model based on James assumption

james\_model <- lm(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

#2.a)Building a model based on Chris assumption

chris\_model <- lm(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

#Is james assumption better or chris is better? #From above we can observe, as per James assumption the model shows accuracy of 43.39% and as per Chris assumption the model shows accuracy of 60.24%. We can conclude that Chris assumption of comparing Horse power to mpg is better.

#2.b) what is the estimated Horse Power of a car with 4 cyl and mpg of 22?

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

##   
## Call:  
## lm(formula = hp ~ mpg + cyl, 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   
## mpg -2.775 2.177 -1.275 0.21253   
## cyl 23.979 7.346 3.264 0.00281 \*\*  
## ---  
## 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

predict\_model <- predict(calc\_model, data.frame(cyl=c(4), mpg=c(22)))  
predict\_model

## 1   
## 88.93618

#The estimated Horse power of a car for 4 cyl and 22 mpg is 88.93618.

#3.Viewing data from mlbench library for current question

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 ...

#3.a) Build a model to estimate the median value of owner-occupied homes (medv)based on thefollowing 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).

boston\_model<- lm(medv~crim+zn+ptratio+chas, data = BostonHousing)  
summary(boston\_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

#Is this an accurate model? - Based on R-squared which is 0.3599 i.e., The accuracy of model is 35%. The model is not accurate enough.

#3.b).i)Imagine two houses that are identical in all aspects but one bounds the Chas Riverand the other does not. Which one is more expensive and by how much?

#->Factors of Chas is in factors of 1 and 0. Identical houses is 1 and who dont have identical houses have 0. And Estimate Std. of chas1 in relation to medv is 4.58393.

#Medv = 49.91868 + (-0.26018) + 0.07073 + (-1.49367) + 4.58393(1) = 52.81949.

#Medv = 49.91868 + (-0.26018) + 0.07073 + (-1.49367) + 4.58393(0) = 48.23556.

#Comparing from above, Identical houses for chas River for which factor is 1, the value as per estimated std 52.81949. When factor is 0 the value as per estimated std. is 48.23556. By comparison of factors 1 and 0, chas River is expensive by 4.58393 for factor 1 in $1000.

#3.b.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?

#Medv = 49.91868 + (-0.26018) + 0.07073 + (-1.49367)(15) + 4.58393= 31.9081.

#Medv = 49.91868 + (-0.26018) + 0.07073 + (-1.49367)(18) + 4.58393= 27.4271.

#Difference between Pupil-teacher ratio for 15 and 18 is 4.48101.Therefore, Pupil-teacher ratio for 15 is expensive by 4.48101 when compared to pupil-teacher ratio of 18.

#3.c)Which of the variables are statistically important? #when comparing dependent and independent variables to show statistical importance, we would like to see p values being as small as possible.From above, p values of crim, zn, ptratio,chas when compared to medv are lowest. We can conclude that all the values are statistically significant.

#3.d)Anova analysis and determine the order of importance of these four variables.?

anova\_model <- anova(boston\_model)  
anova\_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

#In Annova analysis, the importance of variables is defined by the sum squared values. From above the values of sum squared, in order of importance are as follows: crim,zn,ptratio,chas.