BA\_ASSIGNMENT 3

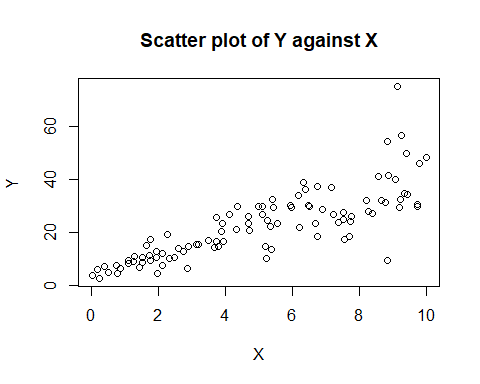
Yeswanth Siripurapu

2023-11-06

#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

set.seed(2017)  
X <- runif(100) \* 10  
Y <- X \* 4 + 3.45  
Y <- rnorm(100) \* 0.29 \* Y + Y  
  
plot(X, Y, main = "Scatter plot of Y against X", xlab = "X", ylab = "Y")



#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

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

#C)How the Coefficient of Determination, R 2, of the model above is related to #the correlation coefficient of X and Y? R-squared (R2) and the correlation coefficient (r) have the following #relationships:The square of the correlation coefficient (r) is known as R-squared (R²). #In this instance, the linear relationship with X accounts for roughly 65.17% of #the variance in Y, as indicated by the R2 of approximately 0.6517.

#The degree and direction of the linear link between X and Y are measured by the #correlation coefficient (r). The value of r in a basic linear regression is #obtained under “Estimate” and is equal to the coefficient for X (3.6108 in #the output).

#Accordingly, in this instance, the coefficient of determination (R2) is roughly #0.6517 and the correlation coefficient (r) is roughly 3.6108. The percentage of #Y’s volatility that can be predicted from X is represented by R².

#QUESTION 2 #a) James wants to buy a car. He and his friend, Chris, have different #opinions about the HorsePower (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 usingmtcars #data to answer the question.

#Building a model based on James estimation:

# Load the mtcars dataset   
data(mtcars)  
  
# Construct a linear model using weight (wt) to estimate horsepower (hp)  
model\_wt\_hp <- lm(hp ~ wt, data = mtcars)  
summary(model\_wt\_hp)

##   
## 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

# Construct a linear model using mileage per gallon (mpg) to estimate horsepower (hp)  
model\_mpg\_hp <- lm(hp ~ mpg, data = mtcars)  
summary(model\_mpg\_hp)

##   
## 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 link between horsepower and fuel efficiency (mpg) appears to have a greater #R-squared value than the association between horsepower and vehicle weight (wt), #according to the R-squared values obtained from the linear regression models. #Since fuel consumption (mpg) is a better predictor of horse power (hp) and fits #the dataset better for predicting hp, it appears that Chris’s point of view is #more valid.

#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?

# Build the model using cyl and mpg to predict hp  
model\_cyl\_mpg <- lm(hp ~ cyl + mpg, data = mtcars)  
  
# Summary of the model  
summary(model\_cyl\_mpg)

##   
## 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

#Based on this, the estimated Horse Power of a car with 4 cylinders and mpg of #22 is approximately 88.93618

#Question 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

# Predicting hp for a car with 4 cylinders and mpg of 22  
new\_data <- data.frame(cyl = 4, mpg = 22)  
predict(model\_cyl\_mpg, newdata = new\_data)

## 1   
## 88.93618

# Load the mlbench library  
library(mlbench)

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

# Load the Boston Housing dataset  
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?

# Build the linear regression model  
model\_medv <- lm(medv ~ crim + zn + ptratio + chas, data = BostonHousing)  
  
# Summary of the model  
summary(model\_medv)

##   
## 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

#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: Chas has two levels: ‘0’ and ‘1’. Those who cross the Chas River are #denoted by a “1,” while those who do not are denoted by a “0.”The data #description states that the median value of owner-occupied dwellings is $1,000, #and the coefficient of chas1 is 4.58393. The product of coefficient and #multiplier is 4583.93$. this costs a lot.

#II. Imagine two houses that are identical in all aspects but in the neighborhood #of one of them the pupilteacher ratio is 15 and in the other one is 18. Which #one is more expensive and by how much?

#Answer: In thousands, the price of a house decreases by 1.49367, or 1493.67, for #If ptratio is 15. Likewise, there will be a drop of 18 \* 1493.67 = 26886.06 if #ptratio is 18.Therefore, if the pt ratio is 15 and the pt ratio is 18, the #difference is $4481.01.

#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.

#Answer: Indeed, we can safely reject the default null hypothesis—that is, the #idea that there is no relationship between house price and other factors in the #model—because the p-values of all the variables do not equal zero. All of the #factors are therefore statistically significant.

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

# ANOVA analysis for the model  
anova\_result <- anova(model\_medv)  
  
# Display the ANOVA table  
anova\_result

## 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 evident that the crim variable explains a substantially larger #amount of variability (sum squared) than the other factors. We could surmise #that the model was greatly enhanced by the addition of the crim. However, #residuals indicate that a significant amount of variability remains unexplained. #The significance is as follows: crim, p tratio, zn, chas