Assignment 3

Keerthi Tiyyagura

2023-10-29

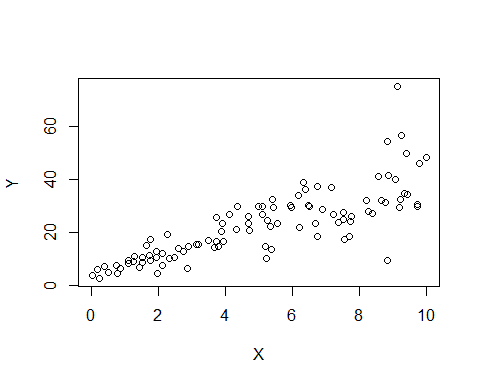
# To create two variables X and Y  
set.seed(2017)  
X <- runif(100)\*10  
Y <- X\*4+3.45  
Y <- rnorm(100)\*0.29\*Y+Y

1.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  
cor(X,Y)

## [1] 0.807291

plot(X,Y)



# Yes,we can fit a linear model to explain Y based on X with a positive correlation.

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?

# Fit a linear 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

coefficients(model)

## (Intercept) X   
## 4.465490 3.610759

# Y=3.6108X+4.4655 is the model equation  
# The accuracy of the above model is 65.17 percent.

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

# Coefficient of determination  
cor(X,Y)^2

## [1] 0.6517187

# (Correlation Coefficient)ˆ2=Coefficient of Determination

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)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.

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

##   
## 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

summary(lm(hp~mpg,data = mtcars))

##   
## 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

# By looking at the values, Chris is right, mpg had a high r square value of 60% compared to weight.

# Fit simple linear models using mtcars dataset.  
data(mtcars)  
model.wt <- lm(hp~wt,data = mtcars)  
model.mpg <- lm(hp~mpg,data = mtcars)

summary(model.wt)

##   
## 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

summary(model.mpg)

##   
## 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

2.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?

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

##   
## 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

(model$coefficients[2]\*4)+(model$coefficients[1])+(model$coefficients[3]\*22)

## cyl   
## 88.93618

predict(model,data.frame(cyl=4,mpg=22),interval = "prediction",level = 0.85)

## fit lwr upr  
## 1 88.93618 28.53849 149.3339

# Installing package "mlbench"  
library(mlbench)  
data(BostonHousing)

3.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 R 2 )

model1 <- lm(medv~crim+zn+ptratio+chas,data = BostonHousing)  
summary(model1)

##   
## 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

# Due to the model's extremely low R square value of 36%, it is not particularly accurate.

3.b.1)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?

summary(model2 <- lm(medv~chas,data = BostonHousing))

##   
## Call:  
## lm(formula = medv ~ chas, data = BostonHousing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.094 -5.894 -1.417 2.856 27.906   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22.0938 0.4176 52.902 < 2e-16 \*\*\*  
## chas1 6.3462 1.5880 3.996 7.39e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.064 on 504 degrees of freedom  
## Multiple R-squared: 0.03072, Adjusted R-squared: 0.02879   
## F-statistic: 15.97 on 1 and 504 DF, p-value: 7.391e-05

model2$coefficients

## (Intercept) chas1   
## 22.093843 6.346157

(model2$coefficients[2]\*0)+model2$coefficients[1]

## chas1   
## 22.09384

(model2$coefficients[2]\*1)+model2$coefficients[1]

## chas1   
## 28.44

# The home with chas of 1 is more expensive than the house without chas of 0 with a value of 4.3 utility

3.b.2)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?

summary(model3 <- lm(medv~ptratio,data = BostonHousing))

##   
## Call:  
## lm(formula = medv ~ ptratio, data = BostonHousing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.8342 -4.8262 -0.6426 3.1571 31.2303   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 62.345 3.029 20.58 <2e-16 \*\*\*  
## ptratio -2.157 0.163 -13.23 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.931 on 504 degrees of freedom  
## Multiple R-squared: 0.2578, Adjusted R-squared: 0.2564   
## F-statistic: 175.1 on 1 and 504 DF, p-value: < 2.2e-16

(model3$coefficients[2]\*15)+model3$coefficients[1]

## ptratio   
## 29.987

(model3$coefficients[2]\*18)+model3$coefficients[1]

## ptratio   
## 23.51547

# The cost of the house with the ptratio of 15 is more expensive than the cost of the house with the ptratio 18 with a value of 6.47.

3.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.

summary(model1)

##   
## 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

# If your p-value is low ( 0.05), you can reject the null hypothesis.

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

anova(model1)

## 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 order of importance of these four variables by comparing the p values are  
# 1) crim  
# 2) ptratio  
# 3) zn  
# 4) chas