Assignment - 2

Vamshee Deepak Goud Katta

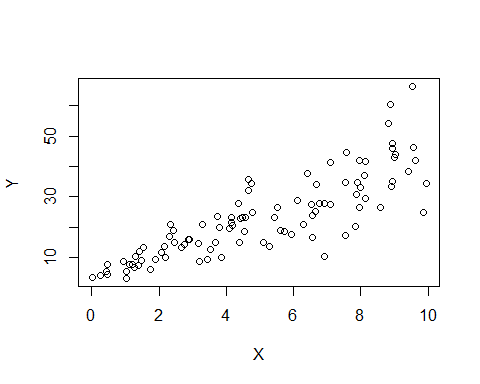
11/17/2021

## 1. Creating variables.

set.seed(123)  
X=runif(100)\*10  
Y=X\*4+3.45  
Y=rnorm(100)\*0.29\*Y+Y

### a) Plotting Y against X.

plot(X,Y)



# Based on the data in graph, we can fit a linear model to explain Y based on X.

### b) Constructing a 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   
## -20.3132 -4.0022 0.1144 3.0670 25.4482   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.2746 1.4828 2.208 0.0296 \*   
## X 3.9452 0.2585 15.260 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.331 on 98 degrees of freedom  
## Multiple R-squared: 0.7038, Adjusted R-squared: 0.7008   
## F-statistic: 232.9 on 1 and 98 DF, p-value: < 2.2e-16

Model$coefficients

## (Intercept) X   
## 3.274645 3.945235

The following equation explains Y based on X:

Y = (3.945235\*X) + 3.274645

The accuracy of the model is 0.7038.

### c) Relation of Coefficient of Determination to the correlation coefficient of X and Y.

A. R2 is the squared value of the coorelation cofficient between x and y where as, regression is based on one variable.

cor(X,Y)^2

## [1] 0.7038116

# From the above result, we can notice that it's the same value as R-squared '0.7038'.

## 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

### a) Constructing simple linear models using mtcars data to determine better estimator of Horse Power (hp) or cars.

James <- lm(mtcars$hp~mtcars$wt, data = mtcars)  
summary(James)

##   
## Call:  
## lm(formula = mtcars$hp ~ mtcars$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   
## mtcars$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

Chris <- lm(mtcars$hp~mtcars$mpg, data = mtcars)  
summary(Chris)

##   
## Call:  
## lm(formula = mtcars$hp ~ mtcars$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 \*\*\*  
## mtcars$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

As per the linear models constructed using mtcars dataset, Chris is right to think that fuel consumption expressed in miles per gallon (mpg) is a better estimator of a car’s horse power (hp).

Accuracy of Chris’ model is 0.6024 which is higher compared to James’ model which is only 0.4339

### b) Building linear model that uses the number of cylinders (cyl) and the mile per gallon (mpg) values of a car to predict the car HorsePower (hp).

Model2 <- lm(mtcars$hp~mtcars$mpg+mtcars$cyl, data = mtcars)  
summary(Model2)

##   
## Call:  
## lm(formula = mtcars$hp ~ mtcars$mpg + mtcars$cyl, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -53.72 -22.18 -10.13 14.47 130.73   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 54.067 86.093 0.628 0.53492   
## mtcars$mpg -2.775 2.177 -1.275 0.21253   
## mtcars$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

Model2$coefficients

## (Intercept) mtcars$mpg mtcars$cyl   
## 54.066600 -2.774769 23.978626

HP <- (Model2$coefficients[2]\*22) + (Model2$coefficients[3]\*4) + Model2$coefficients[1]  
HP

## mtcars$mpg   
## 88.93618

The estimated Horse Power of a car with 4 cylinders giving 22 miles per gallon (mpg) is 88.93618

## 3.Boston Housing

### a) Building linear model to estimate the median value of owner-occupied homes (medv).

library(mlbench)  
data(BostonHousing)  
   
Model3 <- lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn + BostonHousing$ptratio + BostonHousing$chas, data = BostonHousing)  
summary(Model3)

##   
## Call:  
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn +   
## BostonHousing$ptratio + BostonHousing$chas, data = BostonHousing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.282 -4.505 -0.986 2.650 32.656   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 49.91868 3.23497 15.431 < 2e-16 \*\*\*  
## BostonHousing$crim -0.26018 0.04015 -6.480 2.20e-10 \*\*\*  
## BostonHousing$zn 0.07073 0.01548 4.570 6.14e-06 \*\*\*  
## BostonHousing$ptratio -1.49367 0.17144 -8.712 < 2e-16 \*\*\*  
## BostonHousing$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

The accuracy of this linear model is very low as the value of R2 is very low (0.3599).

### b) Using the estimated coefficient.

### I. Determining more expensive of the identical houses based on relation to the Chas River.

summary(Model3)

##   
## Call:  
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn +   
## BostonHousing$ptratio + BostonHousing$chas, data = BostonHousing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.282 -4.505 -0.986 2.650 32.656   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 49.91868 3.23497 15.431 < 2e-16 \*\*\*  
## BostonHousing$crim -0.26018 0.04015 -6.480 2.20e-10 \*\*\*  
## BostonHousing$zn 0.07073 0.01548 4.570 6.14e-06 \*\*\*  
## BostonHousing$ptratio -1.49367 0.17144 -8.712 < 2e-16 \*\*\*  
## BostonHousing$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

Based on the above summary, the estimate of the variation of identical houses in expense is $4.58k as the change in value of house that bounds chas river and that doesn’t is 1.

### II. Determining the more expensive of the identical houses along with the value based on pupil-teacher ratio in the neighborhoods.

# Estimate of the variation of identical houses in expense can be found by multiplying the estimate coefficient with the difference in pupil-teacher ratio among them.  
  
X <- -1.49367\*(15-18)  
X

## [1] 4.48101

### c) Determining statistically important variables.

summary(Model3)

##   
## Call:  
## lm(formula = BostonHousing$medv ~ BostonHousing$crim + BostonHousing$zn +   
## BostonHousing$ptratio + BostonHousing$chas, data = BostonHousing)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.282 -4.505 -0.986 2.650 32.656   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 49.91868 3.23497 15.431 < 2e-16 \*\*\*  
## BostonHousing$crim -0.26018 0.04015 -6.480 2.20e-10 \*\*\*  
## BostonHousing$zn 0.07073 0.01548 4.570 6.14e-06 \*\*\*  
## BostonHousing$ptratio -1.49367 0.17144 -8.712 < 2e-16 \*\*\*  
## BostonHousing$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

Based on the above result, P-value for the model is <0.05. Hence, none of the variables are statistically important.

### d) Using the anova analysis to determine the order of importance of variables.

anova(Model3)

## Analysis of Variance Table  
##   
## Response: BostonHousing$medv  
## Df Sum Sq Mean Sq F value Pr(>F)   
## BostonHousing$crim 1 6440.8 6440.8 118.007 < 2.2e-16 \*\*\*  
## BostonHousing$zn 1 3554.3 3554.3 65.122 5.253e-15 \*\*\*  
## BostonHousing$ptratio 1 4709.5 4709.5 86.287 < 2.2e-16 \*\*\*  
## BostonHousing$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

According to the analysis, the order of importance of the variables can be interpreted as

crim > ptratio > zn > chas