**Linear Modelling**

Jitesh Rawat

2/12/2021

**library**("tidyverse")

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.0.4 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

**library**("caret")

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

**library**("MVN")

## Registered S3 method overwritten by 'GGally':  
## method from   
## +.gg ggplot2

## sROC 0.1-2 loaded

**library**("ISLR")

data <- datasets**::**mtcars  
**head**(data)

## 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

**plot**(data <- mtcars,col= "blue" )*#basic plotting to see some pattern in data*



**featurePlot**(x = data[],y= data**$**mpg)*#plotting mpg with other variables this is very important.*



*#First we can see that am is for Transmission (0 = automatic, 1 = manual) so here automatic transmission of gears give less mpg comparatively than manual.*  
  
*#2nd in mpg vs gear plot we cam see that 4 gear cars give more mpg as compared to both 3 and 5 gears*  
  
*#3rd carburetor is part of engine for air–fuel ratio for combustion. We can see that as number of carburetors increase there is decrease in mpg*  
  
*#4th drat is driver axle ratio which is directly proportional with mpg*  
  
*#5th the weight vs mpg plot directly shows that its inversely proportional*  
  
*#6th the qsec i.e performance measure shows that average performing cars i.e. cars with qsec ratio 18 to 20 gives most mpg*  
  
*#7th*   
  
*#8th the cyl vs mpg plot shows that cars with 4 cylinders gives most mpg*  
  
*#9th Here dist is engine displacement measured in cu in. From the plot we see that it is inversely proportional to mpg and dist depends on no. of cylinders*  
  
*#10th the horse power vs mpg plot is clear that the more is hp the less mpg.*

cormat <- **round**(**cor**(data),2)*#Trying to infer something from correlation matrix*  
**head**(cormat)

## mpg cyl disp hp drat wt qsec vs am gear carb  
## mpg 1.00 -0.85 -0.85 -0.78 0.68 -0.87 0.42 0.66 0.60 0.48 -0.55  
## cyl -0.85 1.00 0.90 0.83 -0.70 0.78 -0.59 -0.81 -0.52 -0.49 0.53  
## disp -0.85 0.90 1.00 0.79 -0.71 0.89 -0.43 -0.71 -0.59 -0.56 0.39  
## hp -0.78 0.83 0.79 1.00 -0.45 0.66 -0.71 -0.72 -0.24 -0.13 0.75  
## drat 0.68 -0.70 -0.71 -0.45 1.00 -0.71 0.09 0.44 0.71 0.70 -0.09  
## wt -0.87 0.78 0.89 0.66 -0.71 1.00 -0.17 -0.55 -0.69 -0.58 0.43

*#Getting the 5 point summary of data*  
**summary**(data)

## mpg cyl disp hp   
## Min. :10.40 Min. :4.000 Min. : 71.1 Min. : 52.0   
## 1st Qu.:15.43 1st Qu.:4.000 1st Qu.:120.8 1st Qu.: 96.5   
## Median :19.20 Median :6.000 Median :196.3 Median :123.0   
## Mean :20.09 Mean :6.188 Mean :230.7 Mean :146.7   
## 3rd Qu.:22.80 3rd Qu.:8.000 3rd Qu.:326.0 3rd Qu.:180.0   
## Max. :33.90 Max. :8.000 Max. :472.0 Max. :335.0   
## drat wt qsec vs   
## Min. :2.760 Min. :1.513 Min. :14.50 Min. :0.0000   
## 1st Qu.:3.080 1st Qu.:2.581 1st Qu.:16.89 1st Qu.:0.0000   
## Median :3.695 Median :3.325 Median :17.71 Median :0.0000   
## Mean :3.597 Mean :3.217 Mean :17.85 Mean :0.4375   
## 3rd Qu.:3.920 3rd Qu.:3.610 3rd Qu.:18.90 3rd Qu.:1.0000   
## Max. :4.930 Max. :5.424 Max. :22.90 Max. :1.0000   
## am gear carb   
## Min. :0.0000 Min. :3.000 Min. :1.000   
## 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:2.000   
## Median :0.0000 Median :4.000 Median :2.000   
## Mean :0.4062 Mean :3.688 Mean :2.812   
## 3rd Qu.:1.0000 3rd Qu.:4.000 3rd Qu.:4.000   
## Max. :1.0000 Max. :5.000 Max. :8.000

*#Generating a model with other variables than mpg*  
model <- **lm**(mpg**~**., data= data)  
**par**(mfrow=**c**(2,2),mar=**c**(4,4,2,0.5))  
**plot**(model)



**summary**(model)

##   
## Call:  
## lm(formula = mpg ~ ., data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.4506 -1.6044 -0.1196 1.2193 4.6271   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 12.30337 18.71788 0.657 0.5181   
## cyl -0.11144 1.04502 -0.107 0.9161   
## disp 0.01334 0.01786 0.747 0.4635   
## hp -0.02148 0.02177 -0.987 0.3350   
## drat 0.78711 1.63537 0.481 0.6353   
## wt -3.71530 1.89441 -1.961 0.0633 .  
## qsec 0.82104 0.73084 1.123 0.2739   
## vs 0.31776 2.10451 0.151 0.8814   
## am 2.52023 2.05665 1.225 0.2340   
## gear 0.65541 1.49326 0.439 0.6652   
## carb -0.19942 0.82875 -0.241 0.8122   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.65 on 21 degrees of freedom  
## Multiple R-squared: 0.869, Adjusted R-squared: 0.8066   
## F-statistic: 13.93 on 10 and 21 DF, p-value: 3.793e-07

*#From this model where mpg vs all we can see that*  
*#The qqplot shows its normal*  
*#the Residual vs fitted plot shows a downwards bend curve which is not anticipated*  
*#the summary is not able to show significant variable that can affect the model, although it shows weight has a significance*  
*#the R squared value is .869 but it can be made better*

*#Generating model with only cyl,dis,hp and drat*  
  
model2 <- **lm**(mpg**~**., data= data[1**:**5])  
**par**(mfrow=**c**(2,2),mar=**c**(4,4,2,0.5))  
**plot**(model2)



**summary**(model2)

##   
## Call:  
## lm(formula = mpg ~ ., data = data[1:5])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.5660 -1.8161 -0.6469 1.4094 6.5749   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 23.98524 7.98905 3.002 0.00571 \*\*  
## cyl -0.81402 0.84368 -0.965 0.34318   
## disp -0.01390 0.01089 -1.276 0.21287   
## hp -0.02317 0.01576 -1.470 0.15299   
## drat 2.15405 1.59866 1.347 0.18905   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.012 on 27 degrees of freedom  
## Multiple R-squared: 0.7825, Adjusted R-squared: 0.7503   
## F-statistic: 24.29 on 4 and 27 DF, p-value: 1.314e-08

*#From summary of this model we see that*  
*#the model is not so normal*   
*#In Residuals vs Fitted the curve is reduced but not straight*  
*#the R squared value is decreased that means we have certainly ignored an important variable from previous model*  
*#p-value is close to 0*

*#Generating model by adding variables wt,hp,disp*  
model3 <- **lm**(mpg**~**wt**+**hp**+**disp,data=mtcars)  
**par**(mfrow=**c**(2,2),mar=**c**(4,4,2,0.5))  
**plot**(model3)



**summary**(model3)

##   
## Call:  
## lm(formula = mpg ~ wt + hp + disp, data = mtcars)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.891 -1.640 -0.172 1.061 5.861   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37.105505 2.110815 17.579 < 2e-16 \*\*\*  
## wt -3.800891 1.066191 -3.565 0.00133 \*\*   
## hp -0.031157 0.011436 -2.724 0.01097 \*   
## disp -0.000937 0.010350 -0.091 0.92851   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.639 on 28 degrees of freedom  
## Multiple R-squared: 0.8268, Adjusted R-squared: 0.8083   
## F-statistic: 44.57 on 3 and 28 DF, p-value: 8.65e-11

*#from this model we can see that*  
*#the weight is certainly significant and horse power too is important factor affecting the mpg*  
*#the R squared value is .8268 that means there is still some important variable we are missing to consider*  
*#but residual error increased*

*#in this model we will try different computations to reduce the residual error*  
model4 <- **lm**(mpg **~** hp**\***gear**+**wt**\***cyl , data = data)  
**par**(mfrow=**c**(2,2),mar=**c**(4,4,2,0.5))  
**plot**(model4)



**summary**(model4)

##   
## Call:  
## lm(formula = mpg ~ hp \* gear + wt \* cyl, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1794 -1.5443 -0.4438 1.3412 4.9153   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 54.290726 11.660718 4.656 9.09e-05 \*\*\*  
## hp -0.045505 0.044620 -1.020 0.317577   
## gear 0.188602 1.858302 0.101 0.919970   
## wt -9.479367 2.430757 -3.900 0.000641 \*\*\*  
## cyl -3.342087 1.102167 -3.032 0.005587 \*\*   
## hp:gear 0.003565 0.009658 0.369 0.715162   
## wt:cyl 0.991454 0.361346 2.744 0.011074 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.293 on 25 degrees of freedom  
## Multiple R-squared: 0.8833, Adjusted R-squared: 0.8553   
## F-statistic: 31.53 on 6 and 25 DF, p-value: 1.702e-10

*#from this model we see that the as the residual is reduced but model is getting away from being normal*  
*#the R squared value has increased*

*#experimenting with computations here and there we get*  
model5 <- **lm**(mpg**~**wt**\***cyl**+**hp**\***qsec, data = data)  
**par**(mfrow=**c**(2,2),mar=**c**(4,4,2,0.5))  
**plot**(model5)



**summary**(model5)

##   
## Call:  
## lm(formula = mpg ~ wt \* cyl + hp \* qsec, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.5123 -1.6422 -0.0097 1.2036 3.8048   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.478329 11.341826 2.775 0.010285 \*   
## wt -10.748455 2.312210 -4.649 9.27e-05 \*\*\*  
## cyl -3.553022 0.931694 -3.814 0.000799 \*\*\*  
## hp 0.142970 0.077191 1.852 0.075846 .   
## qsec 1.499280 0.635579 2.359 0.026451 \*   
## wt:cyl 1.198213 0.320916 3.734 0.000978 \*\*\*  
## hp:qsec -0.011047 0.005412 -2.041 0.051927 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.105 on 25 degrees of freedom  
## Multiple R-squared: 0.9016, Adjusted R-squared: 0.878   
## F-statistic: 38.18 on 6 and 25 DF, p-value: 2.087e-11

*#This a model certainly is good one*  
*#the residual error is almost reduced such that its almost straight line*   
*#the R squared value is 0.9016 which is good enough*  
*#p-value very close to 0*

**shapiro.test**(model5**$**residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: model5$residuals  
## W = 0.97829, p-value = 0.7484

*#the p-value is more than 0.05 that means no need to normalize model*