Assignment 5

Durga Pokharel

18/12/2021

# Fit multiple linear regression on “mtcars” data using mpg variable as dependent variable and rest of the variables as independent variables and interpret the result carefully in terms of model fit and the multicollinearity

library(car)

## Loading required package: carData

mlr <- lm(mpg~.,data = mtcars)  
summary(mlr)

##   
## Call:  
## lm(formula = mpg ~ ., data = mtcars)  
##   
## 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

vif(mlr)

## cyl disp hp drat wt qsec vs am   
## 15.373833 21.620241 9.832037 3.374620 15.164887 7.527958 4.965873 4.648487   
## gear carb   
## 5.357452 7.908747

multicollinearity: is the occurrence of high intercorrelations among two or more independent variables in a multiple regression model. Here in our data mtcars disp have high multicollinearity i.e, 21.620241 also cyl and wt variable have nearly equal multi-colinearity l.e, 15. If we need to drop independent variables with VIF> 10 to fit multilinear regression model.

# Split the “mtcars” data into two random datasets (training and testing sets) with 70:30 partition

library(caTools)

## Warning: package 'caTools' was built under R version 4.1.2

library(caret)

## Warning: package 'caret' was built under R version 4.1.2

## Loading required package: ggplot2

## Loading required package: lattice

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':  
##   
## recode

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

set.seed(123)  
split = sample.split(mtcars$mpg, SplitRatio = 0.7)  
training\_set = subset(mtcars, split == TRUE)  
test\_set = subset(mtcars, split == FALSE)

# Fit the multiple linear regression in the training set and validate its results with testing set

# Fitting Multiple Linear Regression to the Training set  
regressor = lm(formula = mpg ~ .,  
 data = training\_set)  
regressor

##   
## Call:  
## lm(formula = mpg ~ ., data = training\_set)  
##   
## Coefficients:  
## (Intercept) cyl disp hp drat wt   
## 41.430587 -1.546871 -0.003602 -0.004480 1.789690 -1.306315   
## qsec vs am gear carb   
## -0.432134 0.691055 0.535486 -0.892592 -0.659287

# Fit the multiple linear regression in the training set with LOOCV control and validate its results with testing set

train.control<- trainControl(method = "LOOCV")  
model1 <- train(mpg~.,data = mtcars , method= "lm",trControl = train.control)  
print(model1)

## Linear Regression   
##   
## 32 samples  
## 10 predictors  
##   
## No pre-processing  
## Resampling: Leave-One-Out Cross-Validation   
## Summary of sample sizes: 31, 31, 31, 31, 31, 31, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 3.490209 0.6818575 2.743759  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

Our train maodel gives value of cofficent of determination 0.6818575 similarly Root Mean Square Error 3.490290 and Mean Absolute Error is 2.743759.

## Prediction with LOOVC

prediction1 <- model1 %>% predict(test\_set)  
data.frame(R2 = R2(prediction1,test\_set$mpg),RMSE <- RMSE(prediction1, test\_set$mpg),  
MAE <- MAE(prediction1,test\_set$mpg))

## R2 RMSE....RMSE.prediction1..test\_set.mpg.  
## 1 0.9003166 2.016565  
## MAE....MAE.prediction1..test\_set.mpg.  
## 1 1.556515

When we apply the model1 to prediction using LOOCV we get R2 = 0.9003166, RMSE = 2.016565, MAE = 1.556516.

# Fit the multiple linear regression in the training set with 10-folds cross-validation control and validate its results with testing set

set.seed(123)  
train.control <- trainControl(method = "CV",number = 10)  
model2 <- train(mpg~.,data = mtcars, method= "lm", trControl = train.control)  
print(model2)

## Linear Regression   
##   
## 32 samples  
## 10 predictors  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 28, 28, 29, 29, 29, 30, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 3.286242 0.8588116 2.715828  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

While fitting model using k folds cross validation approach I got Root Mean Square Error 3.286242 similarly value of R square 0.8588116 and value of MAE 2.715828.

## Prediction with k-folds cv.

prediction2 <- model2 %>% predict(test\_set)  
data.frame(R2 = R2(prediction2,test\_set$mpg),RMSE <- RMSE(prediction1, test\_set$mpg),  
MAE <- MAE(prediction1,test\_set$mpg))

## R2 RMSE....RMSE.prediction1..test\_set.mpg.  
## 1 0.9003166 2.016565  
## MAE....MAE.prediction1..test\_set.mpg.  
## 1 1.556515

Hence, value of R2 after prediction is 0.9003166 similarly RMSE is 2.016565 and MAE is 1.556515. # Fit the multiple linear regression in the training set with 10-folds and 3 repeats control and validate its results with testing set

set.seed(123)  
train.control <- trainControl(method = "repeatedcv", number = 10, repeats = 3)  
model3<- train(mpg~.,data = mtcars,method = "lm",trControl = train.control)  
print(model3)

## Linear Regression   
##   
## 32 samples  
## 10 predictors  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 3 times)   
## Summary of sample sizes: 28, 28, 29, 29, 29, 30, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 3.333145 0.8608139 2.902094  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

## Prediction of model

prediction3 <- model3 %>% predict(test\_set)  
data.frame(R2 = R2(prediction3,test\_set$mpg),RMSE <- RMSE(prediction3, test\_set$mpg),  
MAE <- MAE(prediction3,test\_set$mpg))

## R2 RMSE....RMSE.prediction3..test\_set.mpg.  
## 1 0.9003166 2.016565  
## MAE....MAE.prediction3..test\_set.mpg.  
## 1 1.556515

Value of R2 is 0.9003166 and RMSE is 2.016565 and MAE 1.556515.

# Which model is the best model? Why? Describe carefully.

Among all above model I found K folds cross validation model was the best model because It gives maximum value for r square and minimum value for error function while training the model. While predicting the value, all model gave same value.

# Predict the weight using the best model identified above.

new.wt <- data.frame(wt = 6 ,cyl=7,disp = 170.0,hp = 115, drat = 4.00,qsec = 17.30, vs = 0, am = 1,gear = 5, carb =5)  
predict(model2, newdata = new.wt)

## 1   
## 11.18066

# Change all the independent variables as standardized variable using “scale” command in R/R Studio

library(standardize)

## Warning: package 'standardize' was built under R version 4.1.2

##   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*   
## Loading standardize package version 0.2.2   
## Call standardize.news() to see new features/changes   
## \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

scale\_wt<- scale(mtcars$wt)[,1]  
 scale\_wt

## [1] -0.610399567 -0.349785269 -0.917004624 -0.002299538 0.227654255  
## [6] 0.248094592 0.360516446 -0.027849959 -0.068730634 0.227654255  
## [11] 0.227654255 0.871524874 0.524039143 0.575139986 2.077504765  
## [16] 2.255335698 2.174596366 -1.039646647 -1.637526508 -1.412682800  
## [21] -0.768812180 0.309415603 0.222544170 0.636460997 0.641571082  
## [26] -1.310481114 -1.100967659 -1.741772228 -0.048290296 -0.457097039  
## [31] 0.360516446 -0.446876870

scale\_cyl <- scale(mtcars$cyl)  
scale\_cyl

## [,1]  
## [1,] -0.1049878  
## [2,] -0.1049878  
## [3,] -1.2248578  
## [4,] -0.1049878  
## [5,] 1.0148821  
## [6,] -0.1049878  
## [7,] 1.0148821  
## [8,] -1.2248578  
## [9,] -1.2248578  
## [10,] -0.1049878  
## [11,] -0.1049878  
## [12,] 1.0148821  
## [13,] 1.0148821  
## [14,] 1.0148821  
## [15,] 1.0148821  
## [16,] 1.0148821  
## [17,] 1.0148821  
## [18,] -1.2248578  
## [19,] -1.2248578  
## [20,] -1.2248578  
## [21,] -1.2248578  
## [22,] 1.0148821  
## [23,] 1.0148821  
## [24,] 1.0148821  
## [25,] 1.0148821  
## [26,] -1.2248578  
## [27,] -1.2248578  
## [28,] -1.2248578  
## [29,] 1.0148821  
## [30,] -0.1049878  
## [31,] 1.0148821  
## [32,] -1.2248578  
## attr(,"scaled:center")  
## [1] 6.1875  
## attr(,"scaled:scale")  
## [1] 1.785922

scale\_disp <- scale(mtcars$disp)  
scale\_disp

## [,1]  
## [1,] -0.57061982  
## [2,] -0.57061982  
## [3,] -0.99018209  
## [4,] 0.22009369  
## [5,] 1.04308123  
## [6,] -0.04616698  
## [7,] 1.04308123  
## [8,] -0.67793094  
## [9,] -0.72553512  
## [10,] -0.50929918  
## [11,] -0.50929918  
## [12,] 0.36371309  
## [13,] 0.36371309  
## [14,] 0.36371309  
## [15,] 1.94675381  
## [16,] 1.84993175  
## [17,] 1.68856165  
## [18,] -1.22658929  
## [19,] -1.25079481  
## [20,] -1.28790993  
## [21,] -0.89255318  
## [22,] 0.70420401  
## [23,] 0.59124494  
## [24,] 0.96239618  
## [25,] 1.36582144  
## [26,] -1.22416874  
## [27,] -0.89093948  
## [28,] -1.09426581  
## [29,] 0.97046468  
## [30,] -0.69164740  
## [31,] 0.56703942  
## [32,] -0.88529152  
## attr(,"scaled:center")  
## [1] 230.7219  
## attr(,"scaled:scale")  
## [1] 123.9387

scale\_hp <- scale(mtcars$hp)  
scale\_hp

## [,1]  
## [1,] -0.53509284  
## [2,] -0.53509284  
## [3,] -0.78304046  
## [4,] -0.53509284  
## [5,] 0.41294217  
## [6,] -0.60801861  
## [7,] 1.43390296  
## [8,] -1.23518023  
## [9,] -0.75387015  
## [10,] -0.34548584  
## [11,] -0.34548584  
## [12,] 0.48586794  
## [13,] 0.48586794  
## [14,] 0.48586794  
## [15,] 0.85049680  
## [16,] 0.99634834  
## [17,] 1.21512565  
## [18,] -1.17683962  
## [19,] -1.38103178  
## [20,] -1.19142477  
## [21,] -0.72469984  
## [22,] 0.04831332  
## [23,] 0.04831332  
## [24,] 1.43390296  
## [25,] 0.41294217  
## [26,] -1.17683962  
## [27,] -0.81221077  
## [28,] -0.49133738  
## [29,] 1.71102089  
## [30,] 0.41294217  
## [31,] 2.74656682  
## [32,] -0.54967799  
## attr(,"scaled:center")  
## [1] 146.6875  
## attr(,"scaled:scale")  
## [1] 68.56287

scale\_drat <- scale(mtcars$drat)  
scale\_drat

## [,1]  
## [1,] 0.56751369  
## [2,] 0.56751369  
## [3,] 0.47399959  
## [4,] -0.96611753  
## [5,] -0.83519779  
## [6,] -1.56460776  
## [7,] -0.72298087  
## [8,] 0.17475447  
## [9,] 0.60491932  
## [10,] 0.60491932  
## [11,] 0.60491932  
## [12,] -0.98482035  
## [13,] -0.98482035  
## [14,] -0.98482035  
## [15,] -1.24665983  
## [16,] -1.11574009  
## [17,] -0.68557523  
## [18,] 0.90416444  
## [19,] 2.49390411  
## [20,] 1.16600392  
## [21,] 0.19345729  
## [22,] -1.56460776  
## [23,] -0.83519779  
## [24,] 0.24956575  
## [25,] -0.96611753  
## [26,] 0.90416444  
## [27,] 1.55876313  
## [28,] 0.32437703  
## [29,] 1.16600392  
## [30,] 0.04383473  
## [31,] -0.10578782  
## [32,] 0.96027290  
## attr(,"scaled:center")  
## [1] 3.596563  
## attr(,"scaled:scale")  
## [1] 0.5346787

scale\_qsec <- scale(mtcars$qsec)  
scale\_qsec

## [,1]  
## [1,] -0.77716515  
## [2,] -0.46378082  
## [3,] 0.42600682  
## [4,] 0.89048716  
## [5,] -0.46378082  
## [6,] 1.32698675  
## [7,] -1.12412636  
## [8,] 1.20387148  
## [9,] 2.82675459  
## [10,] 0.25252621  
## [11,] 0.58829513  
## [12,] -0.25112717  
## [13,] -0.13920420  
## [14,] 0.08464175  
## [15,] 0.07344945  
## [16,] -0.01608893  
## [17,] -0.23993487  
## [18,] 0.90727560  
## [19,] 0.37564148  
## [20,] 1.14790999  
## [21,] 1.20946763  
## [22,] -0.54772305  
## [23,] -0.30708866  
## [24,] -1.36476075  
## [25,] -0.44699237  
## [26,] 0.58829513  
## [27,] -0.64285758  
## [28,] -0.53093460  
## [29,] -1.87401028  
## [30,] -1.31439542  
## [31,] -1.81804880  
## [32,] 0.42041067  
## attr(,"scaled:center")  
## [1] 17.84875  
## attr(,"scaled:scale")  
## [1] 1.786943

scale\_vs <- scale(mtcars$vs)  
scale\_vs

## [,1]  
## [1,] -0.8680278  
## [2,] -0.8680278  
## [3,] 1.1160357  
## [4,] 1.1160357  
## [5,] -0.8680278  
## [6,] 1.1160357  
## [7,] -0.8680278  
## [8,] 1.1160357  
## [9,] 1.1160357  
## [10,] 1.1160357  
## [11,] 1.1160357  
## [12,] -0.8680278  
## [13,] -0.8680278  
## [14,] -0.8680278  
## [15,] -0.8680278  
## [16,] -0.8680278  
## [17,] -0.8680278  
## [18,] 1.1160357  
## [19,] 1.1160357  
## [20,] 1.1160357  
## [21,] 1.1160357  
## [22,] -0.8680278  
## [23,] -0.8680278  
## [24,] -0.8680278  
## [25,] -0.8680278  
## [26,] 1.1160357  
## [27,] -0.8680278  
## [28,] 1.1160357  
## [29,] -0.8680278  
## [30,] -0.8680278  
## [31,] -0.8680278  
## [32,] 1.1160357  
## attr(,"scaled:center")  
## [1] 0.4375  
## attr(,"scaled:scale")  
## [1] 0.5040161

scale\_am <- scale(mtcars$am)  
scale\_am

## [,1]  
## [1,] 1.1899014  
## [2,] 1.1899014  
## [3,] 1.1899014  
## [4,] -0.8141431  
## [5,] -0.8141431  
## [6,] -0.8141431  
## [7,] -0.8141431  
## [8,] -0.8141431  
## [9,] -0.8141431  
## [10,] -0.8141431  
## [11,] -0.8141431  
## [12,] -0.8141431  
## [13,] -0.8141431  
## [14,] -0.8141431  
## [15,] -0.8141431  
## [16,] -0.8141431  
## [17,] -0.8141431  
## [18,] 1.1899014  
## [19,] 1.1899014  
## [20,] 1.1899014  
## [21,] -0.8141431  
## [22,] -0.8141431  
## [23,] -0.8141431  
## [24,] -0.8141431  
## [25,] -0.8141431  
## [26,] 1.1899014  
## [27,] 1.1899014  
## [28,] 1.1899014  
## [29,] 1.1899014  
## [30,] 1.1899014  
## [31,] 1.1899014  
## [32,] 1.1899014  
## attr(,"scaled:center")  
## [1] 0.40625  
## attr(,"scaled:scale")  
## [1] 0.4989909

scale\_gear<- scale(mtcars$gear)  
scale\_gear

## [,1]  
## [1,] 0.4235542  
## [2,] 0.4235542  
## [3,] 0.4235542  
## [4,] -0.9318192  
## [5,] -0.9318192  
## [6,] -0.9318192  
## [7,] -0.9318192  
## [8,] 0.4235542  
## [9,] 0.4235542  
## [10,] 0.4235542  
## [11,] 0.4235542  
## [12,] -0.9318192  
## [13,] -0.9318192  
## [14,] -0.9318192  
## [15,] -0.9318192  
## [16,] -0.9318192  
## [17,] -0.9318192  
## [18,] 0.4235542  
## [19,] 0.4235542  
## [20,] 0.4235542  
## [21,] -0.9318192  
## [22,] -0.9318192  
## [23,] -0.9318192  
## [24,] -0.9318192  
## [25,] -0.9318192  
## [26,] 0.4235542  
## [27,] 1.7789276  
## [28,] 1.7789276  
## [29,] 1.7789276  
## [30,] 1.7789276  
## [31,] 1.7789276  
## [32,] 0.4235542  
## attr(,"scaled:center")  
## [1] 3.6875  
## attr(,"scaled:scale")  
## [1] 0.7378041

scale\_carb <- scale(mtcars$carb)  
scale\_carb

## [,1]  
## [1,] 0.7352031  
## [2,] 0.7352031  
## [3,] -1.1221521  
## [4,] -1.1221521  
## [5,] -0.5030337  
## [6,] -1.1221521  
## [7,] 0.7352031  
## [8,] -0.5030337  
## [9,] -0.5030337  
## [10,] 0.7352031  
## [11,] 0.7352031  
## [12,] 0.1160847  
## [13,] 0.1160847  
## [14,] 0.1160847  
## [15,] 0.7352031  
## [16,] 0.7352031  
## [17,] 0.7352031  
## [18,] -1.1221521  
## [19,] -0.5030337  
## [20,] -1.1221521  
## [21,] -1.1221521  
## [22,] -0.5030337  
## [23,] -0.5030337  
## [24,] 0.7352031  
## [25,] -0.5030337  
## [26,] -1.1221521  
## [27,] -0.5030337  
## [28,] -0.5030337  
## [29,] 0.7352031  
## [30,] 1.9734398  
## [31,] 3.2116766  
## [32,] -0.5030337  
## attr(,"scaled:center")  
## [1] 2.8125  
## attr(,"scaled:scale")  
## [1] 1.6152