# A6 Supervised Learning with Multiple Linear Regression

Dipesh Poudel

12/22/2021

## 1. 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

```
lm1<-lm(mpg~.,data = mtcars)
summary(lm1)

##
## 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
```

Since p-value is 3.793e-07 we can say that the model is significant. The predictor variable wt is significant and others variables are not.

Now we will use `vif()` function from `car` package to calculate variance influence factor(VIF) to check for Multicollinearity.

```
library(car)
```

```
## Loading required package: carData
```

```
vif(lm1)
```

```
##       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
```

When there is occurrence of high inter correlations among two or more independent variables then it is called multicollinearity. We calculate VIF and from variables having VIF>10 we remove the variable with highest VIF value while fitting the model. In our case disp has highest VIF with value 21.06. We need to remove this variable.

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

### Splitting Data into train and test

```
set.seed(1234)
ind<-sample(2,nrow(mtcars),replace = T,prob = c(0.7,0.3))
train_data<-mtcars[ind==1,]
test_data<-mtcars[ind==2,]
```

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

### Training the Model with train data

```
library(caret)
lm2<-train(mpg~.,data = train_data,method="lm")
```

```
## Warning in predict.lm(modelFit, newdata): prediction from a rank-
deficient fit
## may be misleading
```

```
lm2
```

```
## Linear Regression
##
## 26 samples
## 10 predictors
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 26, 26, 26, 26, 26, 26, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   5.377669  0.5820896  4.319041
```

```
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

### Making Predictions on test data

```
predict1<-predict(lm2,newdata = test_data)
```

### Calculation of Evaluation Metrices

```
R2<-R2(predict1,test_data$mpg)
RMSE <- RMSE(predict1,test_data$mpg)
MAE <- MAE(predict1,test_data$mpg)
R2

## [1] 0.7521138

RMSE

## [1] 3.703895

MAE

## [1] 2.610213
```

The value of R-squre has increased for test data and error has decreased compared to the training.

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

```
set.seed(1234)
train_control_1<-trainControl(method = "LOOCV")
lm3<-train(mpg~.,data =
train_data,method="lm",trControl=train_control_1)

lm3

## Linear Regression
##
## 26 samples
## 10 predictors
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 25, 25, 25, 25, 25, 25, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   3.750265  0.6370264  2.961882
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

predict2<-predict(lm3,newdata = test_data)
R2<-R2(predict2,test_data$mpg)
RMSE <- RMSE(predict2,test_data$mpg)
```

```
MAE <- MAE(predict2,test_data$mpg)
R2

## [1] 0.7521138

RMSE

## [1] 3.703895

MAE

## [1] 2.610213
```

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

```
set.seed(1234)
train_control_2<-trainControl(method = "cv",number = 10)
lm4<-
train(mpg~.,data=train_data,method="lm",trControl=train_control_2)

lm4

## Linear Regression
##
## 26 samples
## 10 predictors
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 23, 24, 23, 23, 23, 24, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4.208412  0.9540613  3.705621
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

predict3<-predict(lm4,newdata = test_data)
R2<-R2(predict3,test_data$mpg)
RMSE <- RMSE(predict3,test_data$mpg)
MAE <- MAE(predict3,test_data$mpg)
R2

## [1] 0.7521138

RMSE

## [1] 3.703895

MAE

## [1] 2.610213
```

## 6. 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(1234)
train_control_3<-trainControl(method = "repeatedcv", number = 3,
repeats = 3)
lm5<-train(mpg~.,data =
train_data,method="lm",trControl=train_control_3)
lm5

## Linear Regression
##
## 26 samples
## 10 predictors
##
## No pre-processing
## Resampling: Cross-Validated (3 fold, repeated 3 times)
## Summary of sample sizes: 17, 17, 18, 18, 17, 17, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4.093981  0.7407847  3.272774
##
## Tuning parameter 'intercept' was held constant at a value of TRUE

predict4<-predict(lm5,newdata = test_data)
R2<-R2(predict4,test_data$mpg)
RMSE <- RMSE(predict4,test_data$mpg)
MAE <- MAE(predict4,test_data$mpg)
R2

## [1] 0.7521138

RMSE

## [1] 3.703895

MAE

## [1] 2.610213
```

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

The best model was one with 10 fold cross validation as it has highest R-squred valued and lowest RMSE value. These values in the test data remained same.

## 8. Predict the weight using the best model identified above.

### Creating a dataframe with new value

```
new_data_p<-
data.frame(cyl=4,disp=110,hp=95,drat=3.25,wt=2.50,qsec=19.50,vs=1,am=1
```

```
,gear=4,carb=1)
predict(lm4,newdata = new_data_p)

##        1
## 24.84269
```

The predicted MPG for given new data is 24.84.

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

```
df<-as.data.frame(mtcars)

library(dplyr)
col_names<-c(names(df))
col_names<-col_names[!col_names %in% c('mpg')]
df<-df%>%mutate_at(vars(col_names),scale)
```

### 10. Fit the multiple linear regression on “mtcars” data using mpg as dependent variable and all the standardized variable as the independent variable and interpret the results carefully in terms of model fit and the multicollinearity

```
lm6<-lm(mpg~.,data = df)

summary(lm6)

##
## Call:
## lm(formula = mpg ~ ., data = df)
##
## 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)  20.0906     0.4685  42.884   <2e-16 ***
## cyl          -0.1990     1.8663  -0.107   0.9161
## disp          1.6528     2.2132   0.747   0.4635
## hp           -1.4729     1.4925  -0.987   0.3350
## drat          0.4209     0.8744   0.481   0.6353
## wt           -3.6353     1.8536  -1.961   0.0633 .
## qsec          1.4672     1.3060   1.123   0.2739
## vs            0.1602     1.0607   0.151   0.8814
## am            1.2576     1.0262   1.225   0.2340
## gear          0.4836     1.1017   0.439   0.6652
## carb         -0.3221     1.3386  -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
```

```
library(car)
vif(lm6)
```

```
##       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
```

The value of R-squared is 0.869. There are two variables with VIF>10. The disp have VIF value 21.62 so this variable should be removed.