

Project 4 Part III

Kaushal Khatiwada

2024-05-02

Part 3

Using “mtcars” Dataset

A) Divide the mtcars data into train and test datasets with 80:20 random splits

```
set.seed(13)
index <- sample(2,nrow(mtcars),replace = T,prob = c(0.8,0.2)) #Random sampling into two independent c=v
train.mtcars <- mtcars[index==1,] #Training set
test.mtcars <- mtcars[index==2,] #Test set
```

B) Fit a supervised logistic regression model classification models on train data with “am” as dependent variable and all other variables as independent variable

```
train.suplogr <- glm(am ~., data=train.mtcars,family=binomial)
```

C) Check multicollinearity of this model and finalize it using appropriate VIF cut-off value for logistic regression

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.3.3
```

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

```
## Warning: package 'carData' was built under R version 4.3.3
```

```
vif(train.suplogr)
```

```
##      mpg      cyl      disp      hp      drat      wt      qsec      vs
## 49.51759 42.57840 188.00714 47.15783 22.69969 80.11815 179.19129 66.29596
##      gear      carb
## 31.51372 25.91335
```

Remove “Disp” because Independent variables with VIF >2 means presence of multicollinearity.

```
train.suplogr1 <- glm(am ~ mpg+cyl+hp+drat+wt+qsec+vs+gear+carb, data=train.mtcars,family=binomial)
vif(train.suplogr1)
```

```
##      mpg      cyl      hp      drat      wt      qsec      vs
## 15.416554 37.559992 124.828137 22.856113  7.170175 430.629580 196.850574
##      gear      carb
## 27.353017 50.982486
```

Remove “Qsec” because Independent variables with VIF >2 means presence of multicollinearity.

```
train.suplogr1 <- glm(am ~ mpg+cyl+hp+drat+wt+vs+gear+carb, data=train.mtcars,family=binomial)
vif(train.suplogr1)
```

```
##      mpg      cyl      hp      drat      wt      vs      gear      carb
## 12.288096 44.107880 10.325424  2.582748  6.965767 11.737476 15.083121 33.874432
```

Remove “Cyl” because Independent variables with VIF >2 means presence of multicollinearity

```
train.suplogr1 <- glm(am ~ mpg+hp+drat+wt+vs+gear+carb, data=train.mtcars,family=binomial)
vif(train.suplogr1)
```

```
##      mpg      hp      drat      wt      vs      gear      carb
## 12.747852 14.043675  4.417086  4.954625 23.143845 23.317164 70.590141
```

Remove “Carb” because Independent variables with VIF >2 means presence of multicollinearity

```
train.suplogr1 <- glm(am ~ mpg+hp+drat+wt+vs+gear, data=train.mtcars,family=binomial)
vif(train.suplogr1)
```

```
##      mpg      hp      drat      wt      vs      gear
## 16.490903 25.412813  2.516824  7.947455 13.415795  5.343935
```

Remove “Mpg” because Independent variables with VIF >2 means presence of multicollinearity

```
train.suplogr1 <- glm(am ~ hp+drat+wt+vs+gear, data=train.mtcars,family=binomial)
vif(train.suplogr1)
```

```
##      hp      drat      wt      vs      gear
## 11.185028 2.157473 2.199362 10.003068 2.826436
```

Remove “Hp” because Independent variables with VIF >2 means presence of multicollinearity

```
train.suplogr1 <- glm(am ~ drat+wt+vs+gear, data=train.mtcars,family=binomial)
vif(train.suplogr1)
```

```
##      drat      wt      vs      gear
## 4.185473 2.394115 1.919650 6.046151
```

Remove “Gear” because Independent variables with VIF >2 means presence of multicollinearity

```
train.suplogr1 <- glm(am ~ drat+wt+vs, data=train.mtcars,family=binomial)
vif(train.suplogr1)
```

```
##      drat      wt      vs
## 23.99001 166.23183 250.67951
```

Remove “Vs” because Independent variables with VIF >2 means presence of multicollinearity

```
train.suplogr1 <- glm(am ~ drat+wt, data=train.mtcars,family=binomial)
vif(train.suplogr1)
```

```
##      drat      wt
## 1.040363 1.040363
```

Now, VIF<2 so, “Drat” and “wt” remains