Testing of Multicollinearity Assumption

Srikar

2021-10-28

Introduction

Multicollinearity is the phenomenon when a number of the explanatory variables are strongly correlated. If two independent variables are correlated, they explain the same information. The model will not be able to know which of the two variables is actually responsible for a change in the dependent variable. Test for multicollinearity problems using the Variance Inflation Factor, or VIF in short. The VIF indicates for an independent variable how much it is correlated to the other independent variables

VIF starts from 1 and has no upper limit. A VIF of 1 is the best you can have as this indicates that there is no multicollinearity for this variable. A VIF of higher than 5 or 10 indicates that there is a problem with the independent variables in your model.

Objective: To find if there is multicollinearity in the dataset between the independent variables

Data Description:

The National Family Health Surveys (NFHS) programme, initiated in the early 1990s, has emerged as a nationally important source of data on population, health, and nutrition for India and its states NFHS-3 was designed to provide estimates of important indicators on family welfare, maternal and child health, and nutrition. The dataset provides 10 dependent variables that help in finding out the total fertility rate(TFR)

The data file contains 29 observations on 11 variables sampled from NFHS 2005-06.
Y=TFR
X1=HDI
X2=Infant mortality rate
X3=contraceptive use(any method)

X4=Female Age at marriage
X5=Median number of months since preceding the birth
X6=female literacy in percentage
X7=maternal care
X8=Male age at marriage
X9=percent of population with improved water supply
X10=male literacy in percentage

```
library(readxl)
dat=read_excel("C:/Users/Srikar/Desktop/SS/R/Sem 5/Linear
Regression/Practical 9/data.xlsx")
head(dat)
```

```
##       y    x1    x2    x3    x4    x5    x6    x7    x8    x9   x10
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  2.13 0.789  39.8  66.9  23.4  33.4  77.3  79.2  15.3  92.1  90.2
## 2  2.69 0.644  41.7  63.4  41.4  30.4  60.4  42.1  27.7  95.6  83.4
## 3  1.94 0.681  36.1  72.6  14.4  29.9  79.5  66    10.1  88.4  94
## 4  2.38 0.601  44.7  52.6  16.1  32    53.9  77.2  14.4  80.8  78.1
## 5  1.99 0.679  41.7  63.3  21.6  29.7  68.7  56.1  25.3  99.5  82.9
## 6  3.21 0.537  65.3  47.2  58.4  30.2  36.2  33.9  49.2  81.8  73.9
```

Procedure:

1) Building the regression model

```
mod=lm(y~.,data=dat)
summary(mod)
```

```
##
## Call:
## lm(formula = y ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43837 -0.14250 -0.04833  0.19676  0.35463
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.686737   1.970421   2.379 0.028660 *
## x1           0.613751   1.488450   0.412 0.684958
## x2          -0.001258   0.009745  -0.129 0.898732
## x3          -0.033710   0.007148  -4.716 0.000172 ***
## x4           0.019574   0.008449   2.317 0.032514 *
## x5          -0.024824   0.024090  -1.030 0.316427
## x6           0.005273   0.011015   0.479 0.637931
```

```
## x7          -0.017476   0.006283  -2.782 0.012310 *
## x8          -0.007767   0.011157  -0.696 0.495198
## x9          -0.006798   0.006318  -1.076 0.296125
## x10          0.012862   0.018141   0.709 0.487409
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2751 on 18 degrees of freedom
## Multiple R-squared:  0.8994, Adjusted R-squared:  0.8434
## F-statistic: 16.08 on 10 and 18 DF,  p-value: 5.185e-07
```

We see that that range of residuals is slighlty arge which means that the values will differe a bit more from the observed (y) value

We observe that the intecept p-value is below the significance value (0.05) and hence we can say that the intecept is significant in the prediction.
This means that if the values of the regressors were all zero, the intercept would tell us the mean estimate of the dependent variable (Total fertility).
rate).SInce age can not be 0, intercept has no real meaning.

Only x3,x4 and x7 shows significance i.e. their p-value is lesser than 0.05(significance level).A better fit of the model would be the one where only the significant variables exist.Also the overall p-value is also lesser than significane level (0.05) and hence we can say that the model is significant.

The R-squared value 0.8434 which means that 89.94% of the variation is explained by the regressors. The adjusted R-Squared shows the variation explained by the regressors that truly contribute to the known variation.

#2) Checking the multicolinearity assumption

```
library(car)
```

```
## Warning: package 'car' was built under R version 3.6.3
```

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

```
vif(mod)
```

```
##        x1        x2        x3        x4        x5        x6        x7
x8
## 10.342818  8.857510  3.097699  6.585016  2.115003 11.024608  6.006303
6.787283
##        x9       x10
##  2.254842  7.401606
```

From this we observe that x1 and x6 are the cayses of multicolinearity as their VIF is above 10.We remove these variables to fulfill the multicolinearity assumption

```
new_dat=dat[,c(-2,-7)]
head(new_dat)
```

```
## # A tibble: 6 x 9
##       y    x2    x3    x4    x5    x7    x8    x9   x10
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1  2.13  39.8  66.9  23.4  33.4  79.2  15.3  92.1  90.2
## 2  2.69  41.7  63.4  41.4  30.4  42.1  27.7  95.6  83.4
## 3  1.94  36.1  72.6  14.4  29.9  66    10.1  88.4  94
## 4  2.38  44.7  52.6  16.1  32    77.2  14.4  80.8  78.1
## 5  1.99  41.7  63.3  21.6  29.7  56.1  25.3  99.5  82.9
## 6  3.21  65.3  47.2  58.4  30.2  33.9  49.2  81.8  73.9
```

```
mod1=lm(y~.,data=new_dat)

vif(mod1)
```

```
##       x2       x3       x4       x5       x7       x8       x9      x10
## 6.262698 2.938764 5.783690 2.064378 5.109683 5.447680 2.200527 5.492744
```

We find that there is no multicolinearity between the varibales as the varibales that caused it have been removed.

Conclusion

The mutilcolonearity assumption has been fulfilled