In-Class Lab 8

ECON 425 (Justin Heflin, West Virginia University)

March 10, 2023

The purpose of this lab is to practice using R to detect the severity of multicollinearity in a regression equation. The lab may be completed as a group. To receive credit, upload your .R script to the appropriate place on eCampus ("In-Class Labs'' folder).

For starters

Open a new R script (named `ICL8_XYZ.R`, where `XYZ` are your initials)

Clean out/"Sweep'' R Studio

Click the broom in the Environment panel (top-right), it is directly below the Tutorial button. Also, in the bottom-right panel, click the Plots button and then click the broom in that panel. This should help with loading things into R.

Multicollinearity

Install a New R Package

To illustrate how to calculte VIF for a regression model in R, we will use a built-in data-set in the R package `car`. We installed the `car` R package in the In-Class Lab 5. If you were unable to complete that particular lab then please go ahead and install the R package car using the following line of code:

```
install.packages("car")
```

If you have already installed that R package then you can go ahead and load the `car` package like so:

```
library(car)
```

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

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

Each observation represents a different type of car. Since there are 32 observations this implies there are data on 32 different types of cars. This would be an example of a cross-sectional data-set.

Brief Variable Description

- mpg: miles per gallon
- cyl: number of cylinders
- disp: displacement cubic inches
- hp: gross horsepower
- wt: weight
- drat: rear axle ratio
- qsec: 1/4 mile time (for more information on this variable watch the first Fast and Furious movie)

Regression Model

```
regression <- lm(mpg ~ disp + hp + wt + drat, data = car_data)
summary(regression)
```

Comment out the returned value for R^2. Do you consider this a "good" R^2?

VIF

Now, we will use the `vif()` function from the `car` package to calculate the VIF for each indepedent variable in our model:

```
vif(regression)
```

Comment out the VIF for each independent variable.

Of the variance inflation factors for each independent variable, are there any independent variables we should be concerned about? If so, comment out which independent variable(s) show severe signs of multicollinearity.

Visualizing VIF Values

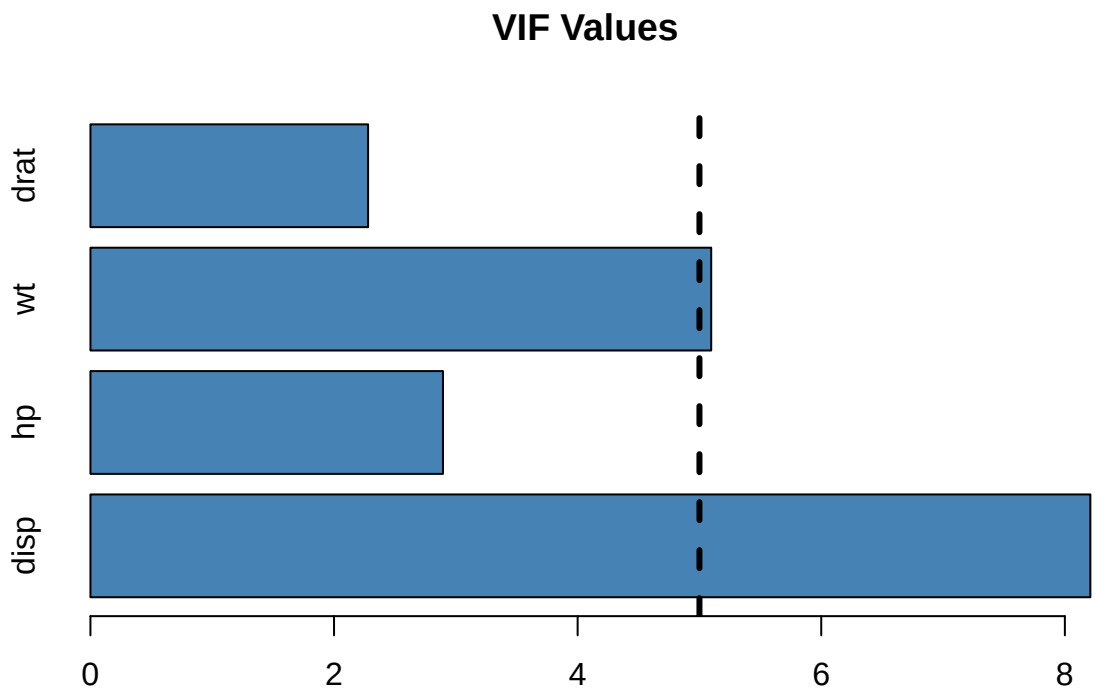
To visualize the VIF values for each independent variable, we can create a simple plot.

First, let's create a vector of our VIF values

```
vif_values <- vif(regression)
```

Now, let's plot our VIF values

```
barplot(vif_values, main = "VIF Values", horiz = TRUE, col = "steelblue") +
  abline(v = 5, lwd = 3, lty = 2)
```



Simple Correlation Coefficient

To gain a better understanding of why one independent variable may have a high VIF value, we can create a correlation matrix to view the linear correlation coefficients between each pair of variables. A simple correlation coefficient, notated by r, is a value from +1 to -1, where the sign indicates the direction of the correlation between the two variables.

First, let's subset and index our data-set to only include the four independent variables in our original regression model. The second line of code creates our correlation matrix

```
car_data_2 <- mtcars[,c("disp", "hp", "wt", "drat")]
cor(car_data_2)
```

```
##            disp         hp         wt       drat
## disp  1.0000000  0.7909486  0.8879799 -0.7102139
## hp    0.7909486  1.0000000  0.6587479 -0.4487591
## wt    0.8879799  0.6587479  1.0000000 -0.7124406
## drat -0.7102139 -0.4487591 -0.7124406  1.0000000
```

Recall that the variable `disp` had a VIF value over 8, which was the largest VIF value among all of the independent variables in the model. From the correlation matrix we can see that `disp` is strongly correlated with all three of the other independent variables, which explains why it has such a high VIF value.

In this case, we may want to remove `disp` from the model because it has a high VIF value **and** it was not statistically significant at the 0.05 significance level.

Dropping `disp` from our model and re-running our regression

```
regression2 <- lm(mpg ~ hp + wt + drat, data = car_data)
summary(regression2)
```

Compare the standard errors for `hp`, `wt`, and `drat` from `regression` to `regression2`. Which regression model has smaller standard errors? Comment out your answer.

Does our R^2 drastically change when compared to the R^2 for `regression`? Does this make sense? Comment out your response.

Re-calculate VIF values

Using our second regression model, let's calculate the VIF values for each independent variable:

```
vif(regression2)
```

Comment out the VIF for each independent variable.

Of the variance inflation factors for each independent variable, are there any independent variables we should be concerned about? If so, comment out which independent variable(s) show severe signs of multicollinearity.