GroupI_HM3

Cvetinovic, Stromieri, Savarin, Cortinovis

2023-11-20

Contents

FSDS - Chapter 6

Ex 6.12

For the UN data file at the book's website (see Exercise 1.24), construct a multiple regression model predicting Internet using all the other variables. Use the concept of multicollinearity to explain why adjusted R^2 is not dramatically greater than when GDP is the sole predictor. Compare the estimated GDP effect in the bivariate model and the multiple regression model and explain why it is so much weaker in the multiple regression model. **Solution** Let's start by creating the models, the first includes all the available predictors while, the second one, only the variable GDP:

```
UN <- read.table("http://stat4ds.rwth-aachen.de/data/UN.dat",header=T)
UN$Nation<-NULL
lm_full <- lm(Internet ~ GDP + HDI + GII + Fertility + CO2 + Homicide + Prison,data=UN)
lm_GDP <- lm(Internet ~ GDP,data=UN)
```

By analysing the output of the `summary` function, we get:

$$\bar{R}^2_{full} = 0.8164 \qquad \bar{R}^2_{GDP} = 0.7637$$

Considering that the *full model* is exploiting five more variables than the other model, these results are pretty strange. We suspect that this low difference is due to the fact that some variables of the full model are not very relevant while others are correlated.

```
library(ggplot2)
library(patchwork)
```

```
## Warning: il pacchetto 'patchwork' è stato creato con R versione 4.3.2
```

```r
library(car)
```

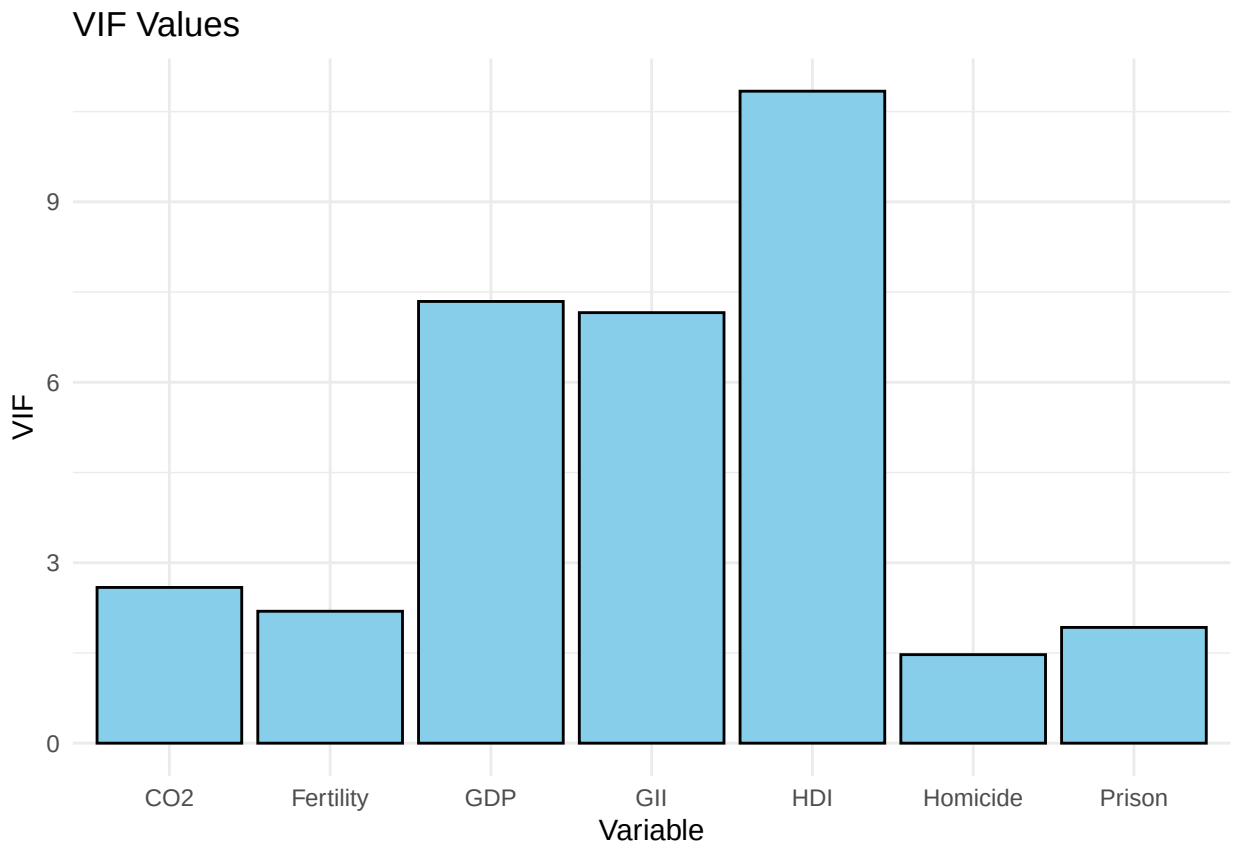
```
## Warning: il pacchetto 'car' è stato creato con R versione 4.3.2

## Caricamento del pacchetto richiesto: carData

##
## Caricamento pacchetto: 'carData'

## Il seguente oggetto è mascherato _da_ '.GlobalEnv':
##
##     UN
```

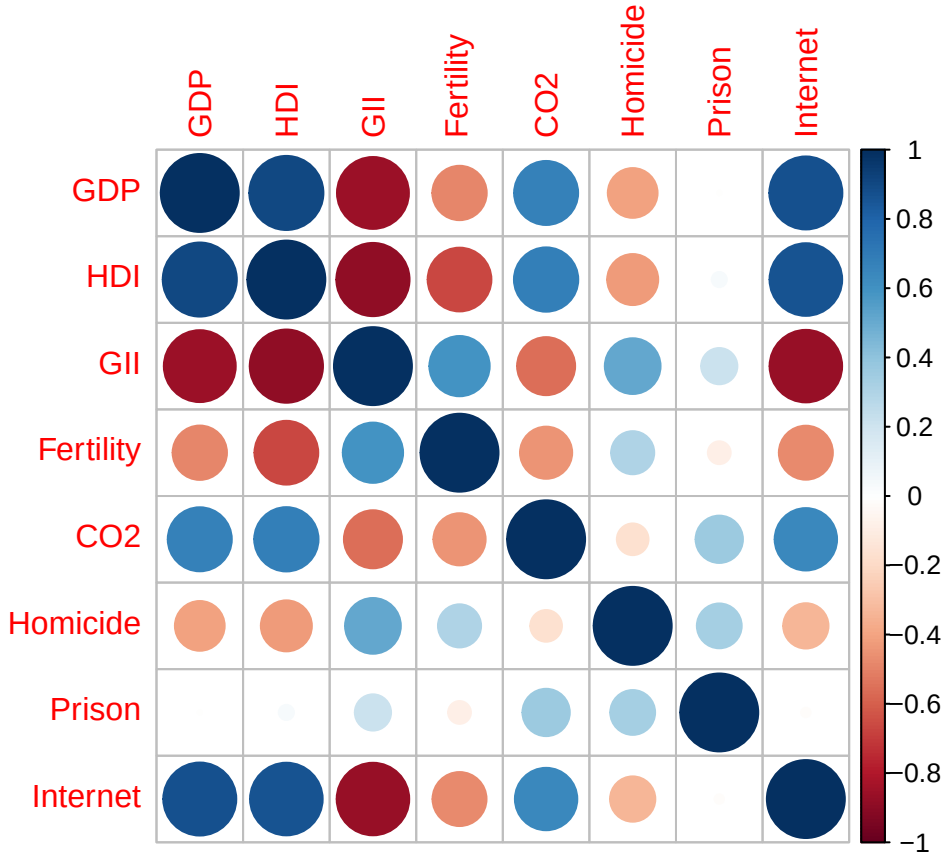
```r
vif_data <- data.frame(variable = c("GDP","HDI","GII","Fertility","CO2","Homicide","Prison"), vif = vif
ggplot <- ggplot(vif_data, aes(x = variable, y = vif)) +
          geom_bar(stat = "identity", fill = "skyblue", color = "black") +
          labs(title = "VIF Values", x = "Variable", y = "VIF") +
          theme_minimal()
print(ggplot)
```




```r
library(corrplot)
```

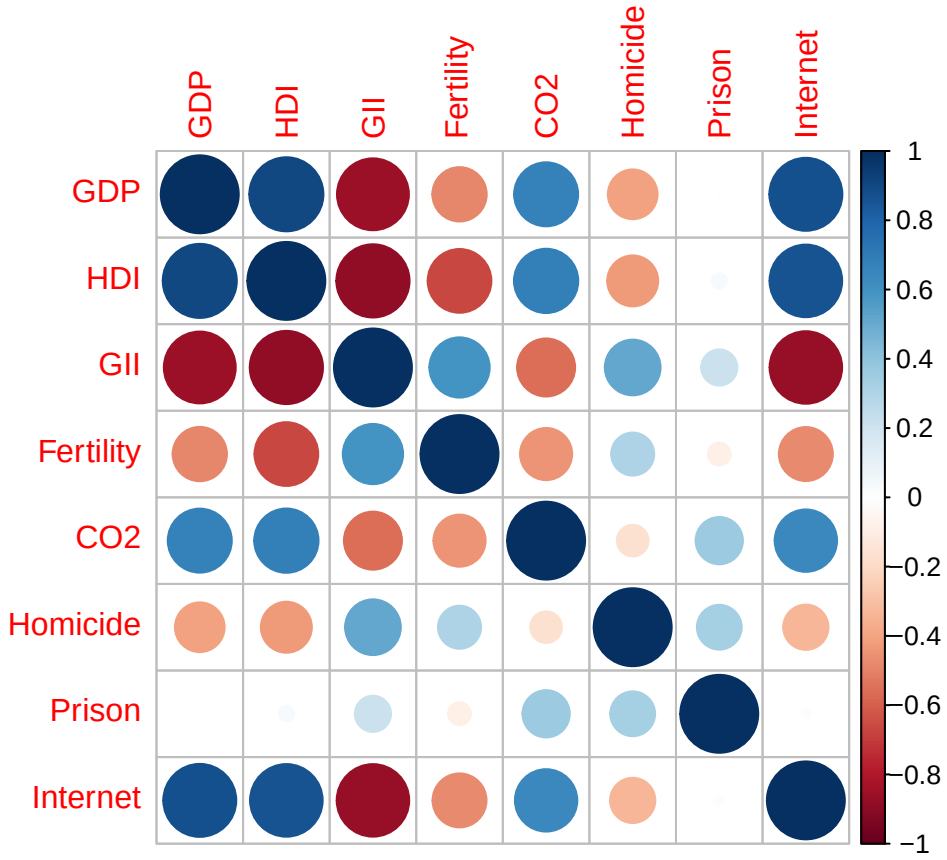
```
## corrplot 0.92 loaded
```

```
cor_matrix <- cor(UN)
cor_plot <- corrplot(cor_matrix, method = "circle")
```



After having analysed the output of the `summary()` function, we suspect that there are some collinearities among the data, therefore we plot both the correlation matrix and the VIF for every variable:

```
library(corrplot)
cor_matrix <- cor(UN)
corrplot(cor_matrix, method = "circle")
```



Ex X.x

Here the text of the second exercise.

Solution

Add comments to the solution.

$$Y \sim N(0, 1)$$

CS - Chapter X

Ex X.x

Here the text of the first exercise from Wood's textbook.

Solution

Add comments to the solution.

$$X \sim N(0, 1)$$