CS5701(Quantitative Data Analysis)

Lab 6: Part 2

The aim of this part of the lab exercise is to give you practical experience in doing multiple regression analysis using R Studio and an R Notebook.

1 Multiple Regression

In this lab we will be using data on Crime Rates in the US. The dependent variable here is the Crime Rate (in offences per million population) and the aim is to understand which of the explanatory variables should be part of a multiple regression model. The data set is called `crime-analysis-data.csv`.

The data includes the following variables:

- CrimeRate Crime rate (number of offences per million population)
- Youth Young males (number of males aged 18-24 per 1000)
- Education Education time (average number of years schooling up to 25)
- ExpenditureYear0 Expenditure (per capita expenditure on police)
- LabourForce Youth labour force (males employed 18-24 per 1000)
- StateSize State size (in hundred thousands)
- YouthUnemployment Youth Unemployment (number of males aged 18-24 per 1000)
- MatureUnemployment Mature Unemployment (number of males aged 35-39 per 1000)
- HighYouthUnemploy High Youth Unemployment 1 = yes, 0 = no
- Wage Wage (median weekly wage)

1. Load the `crime-analysis-data.csv` data into an R notebook.
2. Explore the data numerically and graphically. Confirm the variables that are categorical and numerical/continuous and that R has read them in appropriately
3. Focusing only on the continuous explanatory variables - check their correlations with the CrimeRate.
4. Using the continuous explanatory variables decide on a maximal model for CrimeRate and run it. Consider the need for transformations.
5. Use a model selection approach to achieve a minimal adequate model (you can use the notebook to document your assumptions).
6. Once you have the minimal adequate model, explain its findings.

7. OPTIONAL - model the relationship between the crime rate and the explanatory variables (including the ones that are not continuous).

8. OPTIONAL - If the average education time in the population is 14 years. Compute the mean education time in this sample of 48 rows of data and test the hypothesis that the education time is significantly lower than the population education time.