Introduction to Econometrics

Summer Term 2017 - In-Class Exercise, Lecture 2

We start with analyzing a dataset which "... combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR". It can be downloaded from `http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime`.

(i) Find the documentation for the dataset and take a quick look at which variables are present. Get a subset of the dataset which includes the following variables: household size, percentage of people between the ages of 16 to 24, percentage of people living in urban cities, percentage of people living with social security, per capita income, percent of unemployed, police budget per population and number of violent crimes per 100k population. Understand what each of these variables mean.
(ii) Do histograms of each variable, and get summary statistics of each of them. Can you detect any errors in the data?
(iii) Compute the correlation of each variable with the variable *violent crimes per population*.
(iv) Do a scatterplot of each variable versus the variable *violent crimes per population*.
(iv) Which variables do you think are main determinants of the burnt area? After you decide on a few, check whether they are highly correlated with the other variables.
(vi) From the remaining set of variables, try to choose a variable yourself, which you think would be highly correlated with *violent crimes per population* and do the same analyses with that variable.