Analysis of Crime Rates

This is the final piece of graded work for the half-semester. It is due on Friday, March 6. Only printed copy will be accepted!

1 Outline of the assignment

Download the data file `crime2.gdt`. This is available in the `docs` folder on the ECN 201 home page. The full URL is

`http://ricardo.ecn.wfu.edu/~cottrell/ecn201/docs/crime2.gdt`

The idea is to analyse these date with a view to producing a model that explains as much as possible of the variation in property crime rates across US cities. The basic dependent variable is `prop2005`, the rate of property crime per 100,000 population in the year 2005. The data are from 244 US cities and "places" with populations of 100,000 or more. Please read the description (menu item `/Data/Read info` in gretl) for more details.

2 Comments on the data

You will find that the dataset includes several possible explanatory variables — plus some "raw" variables that were used in constructing variables of interest, but that probably should not be used in their own right, e.g., `genexpend` and `policefrac`, which were used to construct `policepc`, per-capita expenditure on police protection.

Note that some of the potential explanatory variables are alternatives rather than complements. For example `medhhinc` and `pcincome` are alternative measures of income levels; and `fampovpc`, `totpovpc` and `povpcu18` are alternative measures of poverty rates. In these cases you should determine which variant has the greatest explanatory power over crime.

A comment on the education-related variables: `nohischool` is the percentage of the population aged 25 and over who do not have a high school diploma, and `hischool` is the percentage with a high school diploma *but no higher educational attainment*.

And a comment on the age-structure variables such as `pop18_24` (population between the ages of 18 and 24): these are in "raw" numerical form: to get the percentage in each of the age groups you'd need to divide by population.

3 "Explore" the data

The first thing to do is explore the data, getting a sense of the numbers involved, and the distributions. This might also tell you whether there are any "odd cases" among the cities that might need special treatment. Your basic tools here are summary statistics (means, medians, standard deviations, etc.) and graphical methods (boxplots, frequency plots, pairwise X-Y scatter plots). You'll have to be selective here, homing in on what you reckon might be most important: I don't want 20 pages of plots and summary stats! Also in this context: you might find it helpful to look at some sorted tables (as

we did with the OECD health spending exercise), but you don't need to print such tables; there are too many data points

4 Modeling

Once you have some sense of the data, proceed to modeling by means of multiple regression analysis. I suggest you start with a somewhat inclusive model, and then whittle it down by eliminating variables that appear to be insignificant. Note that I say "somewhat" inclusive: a model that includes *all* of the potential explanatory variables would be far too unwieldy. You'll have to use your judgment here.

The special role of the `policepc` variable: In a causal sense, while it is possible that spending more on police does little to *reduce* crime, it is surely nonsense to suppose that more police spending would *raise* crime rates. Therefore, if you end up with a positive sign on `policepc` in your favored model, this is presumably a biased estimate (the bias stemming from simultaneity). In that case, since we are aiming for a causal interpretation here, you are best to drop that variable — even if that lowers the adjusted R^2.

If you estimate several models, you don't have to print out full results from all of them, but please give a clear "trail" as to how you arrived at your final variant.

5 Your write-up

Please pay attention to presentation. The idea is to give the reader the clearest possible picture of "what's going on" in the data, leading up to an explanatory regression model. By now I think you know some of the things you should comment on, in relation to a model. One thing is statistical significance (are we reasonably sure the effect is not zero?), but another is practical importance (is the effect "large" in practical terms?).

6 Extra credit

There are a couple of additional points you might look at.

First, you could see how well your property crime model works for violent crime. Which sort of crime is easier to explain/predict? If you were to build a distinct model for violent crime, how might if differ (if it does) from the property crime model?

Second, you could look at the stability or otherwise of crime rates between 2000 and 2005. How highly correlated are the rates from the two years? Taking property crime for 2005 as the dependent variable, how much "explanation" do you get out of the corresponding rate for 2000 and the population growth rate between the years? Can you interpret the result?

7 Questions?

I will give some more comments/hints and invite questions on the assignment, in class on Monday, March 2.