

GSE 524 Assignment – Crime Regression

The purpose of this assignment is to get experience importing and modifying data in a Pandas DataFrame, and to “replicate” an interesting article about the economics of crime. (I put replicate in quotes because we will not replicate the full analysis.) The article is: “Panic on the Streets of London: Police, Crime, and the July 2005 Terror Attacks” by Draca, Machin, and Witt, *American Economic Review*, 2011.

The purpose of the analysis is to test estimate the effectiveness of police as a crime deterrent. Our regressions will have crime rate as the dependent variable and police deployment as the independent variable of interest (plus some additional controls). A simple cross-sectional analysis is likely to lead to a positive coefficient on police because of reverse causation – neighborhoods with high crime are likely to deploy more police. In other words, high crime causes police deployment to go up but it is not necessarily true that more police cause the crime rate to go up.

The data set has weekly crime data for 2 years and for 32 boroughs of London ($2 \times 52 \times 32 = 3328$ total observations).

The variables included in the data set are:

pop: the population of the borough (constant over time)

pol: police hours worked in the borough in a given week

crime: the number of crimes in the borough in a given week

week: the week (1-104)

borough: 1-32

un: the unemployment rate

emp: the labor force participation rate

ymale: percentage of population that is male under 25 years old

white: percentage of population that is white

Make sure to use the same data frame names and same column names as specified in bold in the assignment below. If you use different names, the grader will not give you full credit. Please name your script crime.py.

Part 1:

Import the provided data into a DataFrame called **fulldata**.

You need to create a column called **crimerate**, equal to the total number of crimes in a borough in a given week divided by the population of the borough.

You need to create a column called **policerate**, equal to the total number of police hours in a borough in a given week divided by the population of the borough.

Our first specification is to use a log-log specification where the crime rate (total crimes / population) depends on the police rate (police hours / population), the unemployment rate, the labor force participation rate, the percentage of young males and the percentage of population that is white. A log-log regressions means that we take the natural log of all of these variables and then run a linear regression on the logged values. The reason to take the logs is that we suspect that the addition of extra police hours does not have a constant effect on the crime rate but probably depends on how many police we are already using. For example, the addition of the first police hour has a bigger effect than the 1000th police hour.

So, you need to create six columns called **lcrime**, **lpolice**, **lemp**, **lun**, **lymale**, and **lwhite** in the DataFrame that equal the log of each of the six variables (crime rate, police rate, emp, un, ymale, and white).

Then run the log-log regression using the statsmodels module and save the RegressionResults object (returned by fit()) as **model1**.

Part 2:

The simple regression in part 1 ignores a lot of important factors. It treats our data as a cross section of 3328 observations on crime and police hours. In fact, we have “panel data”, data that has multiple observations over time for each borough. There are 32 boroughs with 104 weekly observations for each borough. It is likely that, for reasons that are not explained by any of the variables in our data set, the average crime rate in some boroughs is higher than that in other boroughs. It is also likely that those boroughs with higher crime rates deploy more police to deal with the high crime rates. This will affect our regression results. One way to deal with this problem is to take differences in the (logged) crime rate between a given week and that same week 1 year earlier. For example, we compute the difference between the crime rate in week 53 and week 1 for each borough, the difference between week 54 and week 2, etc. To see how this helps, suppose that borough 1 has a higher crime rate and more police hours in all weeks compared to borough 2. In the second year, borough 1 increased its police hours and saw a drop in crime compared to year 1 (though still higher than borough 2). Borough 2 decreased its police hours and experienced an increase in crime rates (though still lower than borough 1). Our first regression will find a positive coefficient on police hours because all of the observations with more police hours (in borough 1) also have higher crime rate. Our differenced regression will have a negative coefficient because increases in police hours lead to decreases in crime rate (and vice versa).

Create a new DataFrame called **diffdata** with “differenced” versions of each (logged) variable (**dlcrime**, **dlpolice**, **dlun**, **dlemp**, **dlymale**, and **dlwhite**). These differenced versions are the values in a given week in year 2 minus the values in the same week in year 1. Note that this cuts our sample size in half – there are now only 52 observations per borough instead of 104.

Run the differenced regression of dlcrime on the other variables. Save the RegressionResults object (returned by fit()) as **model2**.

Part 3

The fundamental problem that we are having difficulty addressing is that the boroughs respond to expectations about crime when deploying police. If a borough experiences an increase crime compared to the previous year and has deployed more police, does this mean the police led to an increase in crime or that the borough deployed more police in expectation of higher crime? To the extent that different

boroughs experience different trends in crime and that these boroughs have better information than we do about these trends, our regressions so far cannot account for this problem.

A Natural Experiment: In July 2005, there were a series of terrorist attacks in London. The terrorist attacks were focused on 5 of the 32 boroughs (boroughs 1, 2, 3, 6, and 14). The city responded by deploying more police in these 5 boroughs. They used overtime to do this so that police hours in the other boroughs was unchanged (i.e. they didn't just move police from the other boroughs to the 5 affected boroughs). This increase in police is an exogenous change in police that is not in response to changes in the crime rate (it is in response to selected high profile crimes, but not to general changes in the crime rate).

The increased police deployment in the five boroughs lasted for 6 weeks and then things returned to normal. For this part, run a regression with logged differenced crime rate (`dlcrime`) as dependent variable and explanatory variables logged differenced `emp`, `un`, `ymale`, and `white` (`dlemp`, `dlun`, `dlymale`, and `dlwhite`). This time, omit the `dlpolice` variable but instead add the following 2 dummy variables (add them as columns in your DataFrame and also include them in your regression):

sixweeks: 1 for the 6 weeks with the increased police (weeks 80-85), 0 otherwise

sixweeks_treat: 1 for the 6 weeks for the 5 boroughs, 0 otherwise

The variable of interest in this regression is `sixweeks_treat` which tells us how crime changed in the 5 boroughs that had increased police hours over the 6 weeks relative to the other boroughs over the same six weeks.

Save the RegressionResults object (returned by `fit()`) as **model3**.

What We Learned:

The takeaway from these three regressions is that the coefficient on the police variable is positive in the first two regressions, possibly due to the reverse causation story. When using the natural experiment with an exogenous increase in police, we instead observe a negative effect (`sixweeks_treat` coef < 0): that the boroughs that saw the increase in police deployment following the attacks saw a decrease in crime relative to other boroughs that did not have extra police. Thus, this is some evidence that putting more police on the streets leads to a decrease in crime.