University of St. Andrews
School of Economics & Finance

EC3301 Econometrics
Martinmas semester 2022

Project Report (The determinants of the crime rate in North Carolina in 1985)

Your mission is to build a regression model to explain the crime rate (for 1985) in the US State of North Carolina. You should build a good econometric model that 'explains' the dependent variable. Build your model carefully and methodically, using Stata and according to the procedures outlined and discussed in laboratories and lectures. Document what you do. It is important you develop your model according to acceptable economic practices. All necessary materials (data, readings, etc.) will be posted to the Module Moodle page (under the "Project" Section – at the bottom of the page).

Variables / Data
These data refer to the US state of North Carolina in 1985. There are 89 observations (by county). Your dependent variable is 'crimerate', it is the number of crimes committed per-person in each county. The rest of the data are as follows:

county	county identifier (don't use as explanatory variable)
prbarr	'probability' of arrest
prbconv	'probability' of conviction if arrested
prbpris	'probability' of prison sentence if convicted
avgsen	avg. sentence, days
polpc	police per capita
density	people per sq. mile
taxpc	tax revenue per capita
west	=1 if in western N.C.
central	=1 if in central N.C.
urban	=1 if in SMSA
pctmin80	percentage of population ethnic minority in 1980
wcon	weekly wage in construction
wtuc	weekly wage in transport, utilities and communications
wtrd	weekly wage, wholesale, retail trade
wfir	weekly wage, financial, insurance, real estate
wser	weekly wage, service industry
wmfg	weekly wage, manufacturing
wfed	weekly wage, federal employees
wsta	weekly wage, state employees
wloc	weekly wage, local govt employees
pctymle	percent of population young male
crmrtelag	crimerate variable in 1984 (previous year)

Note these data have come via the Wooldridge textbook. I don’t have further information on variable definitions, but they are reasonably intuitive.

Reading and theory

Becker, G., 1968, "Crime and Punishment: An Economic Approach." *Journal of Political Economy*, 76, 2, 169-217.

Chalfin, A. and J McCrary, 2017, “Criminal Deterrence: A Review of the Literature ”, *Journal of Economic Literature*, 55, 1, 5–48.

Cornwell, C. and W. N. Trumbull, 1994, "Estimating and Economic Model of Crime with Panel Data”, *The Review of Economics and Statistics*, 76, 2, 360-66. (Note: the data was used – and collected – by the authors of this paper).

Ehrlich, I., 1996, “Crime, Punishment and the Market for Offences”, *The Journal of Economic Perspectives*, 10, 1, 43-67.

Levitt, S. D., 2004, “Understanding Why Crime Fell in the 1990s: Four Factors that Explain the Decline and Six that Do Not”, *Journal of Economic Perspectives*, 18, 1, 163-190.

Mocan, H. N. and D. I. Rees, 1999, “Economic Conditions, Deterrence and Juvenile Crime: Evidence from Microdata”, NBER Working Paper #7405.

This is not an extensive reading list. I am not expecting you to read these articles in depth. However, please peruse them to get a feel for the literature. In particular, you should think about the following issues: (1) which variables from your data set do you think are most likely to be important in modelling the dependent variable – and why? (2) Which variables do you think might not be important? This may be for theoretical reasons or because you think there may be problems including them in a model. (3) What functional form do you believe is the best to use in your regression? (4) What signs and/or magnitudes do you expect your estimated parameters to have? (5) Previous literature may also give some guidance regarding potential problems you might face with your model (or it may not). In sum it is important you convey a good *a priori* understanding of the model and what you expect from it.

Note that section 1.2 of Wooldridge provides a discussion of a basic form of this model (as do the lecture slides). Note also that existing literature may use econometric techniques we have not covered in the course (e.g. time series techniques, simultaneous equation methods, FIML, 2 stage least squares, panel estimation, etc.). This is because much literature tries to apply fairly sophisticated techniques to existing data. Please don’t be put off by this as you don’t need know these techniques, extract the main points you need for your work.

Data

It is a good idea to get to know your data. This would include looking at the summary statistics for these data, check to see if there are any very small or large values of the data (outliers) that might impact upon your results. Perhaps looking at the correlations across some of the variables you think might be important. Also looking at scatterplots of your dependent variables and others think might be important explanatory variables. If this influences the choice of your initial model this should be discussed. Please note – do not go and seek out new data, even if you think there are important variables missing. You must use the data set provided.

Specify and estimate your first model

Based on what you think is theoretically important, and from your review of the data, select your explanatory variables and decide on your functional form. Your literature review and data analysis should mean you have identified important RHS variables. You may not have the exact variable(s) theory suggests. You will need to use the ones 'closest' to those suggested by theory. You may also have to 'chose between' variables you have been given. Do not use all the variables unless you really think theory suggests you should. If you feel there is an important variable missing – you will have to do your best and be aware of possible omitted variable bias. Bear in mind any potential problems with your specification (literature may help guide you on this). It is probably easiest to think about at this stage is the possibility of multicollinearity. Note that the data includes number of dummy variables, think carefully what these mean and how (if at all) they might be included in your specification.

Evaluate your model

Once you have run your model read and report your regression output. It is important you report your regression results in an understandable way. I would recommend you do this in the same way as Wooldridge (and lecture slides), that is present your coefficient estimates alongside the variables, and put standard errors in brackets below this (you don't need to this – but it is a standard format). For example:

$$\widehat{crimerate} = 0.231 - 1.223x_1 + 0.887x_2 + 0.0029x_3$$

$$(0.221) \quad (0.369) \quad (1.225) \quad (0.0025)$$

You should also normally also report the sample size (*n*), R-squared and R(bar)-squared. Also report any other information that is important for any discussion.

Check the signs of the estimated parameters on your model – are they what you expected? Look at the magnitude of your coefficients. Look at the t-stats of the individual coefficients – are they significant? Is this what you expected? Check the overall significance of the regression, using R-squared and R(bar)-squared (or other goodness of fit measure you might think relevant). Could there be any problems with your model including any violations of our MLR assumptions 1-5? For example, might there be multicollinearity or heteroscedasticity? You may want to conduct formal tests for these and discuss what you find (present figures, test-statistics and p-values when relevant). You might also want to look at a histogram of your residuals. If there are problems don't hide them, be open. Remember you can adjust standard errors for the presence of heteroskedasticity. If you think the model still has shortcomings, discuss these.

Specify and estimate your 'improved' Model

Identify and explain how you plan to improve your regression model. Then re-estimate it. You may need to do this a number of times until you get a model you are happy with. That is, you may need to estimate a number of iterations of your model before you get a preferred version. You are unlikely to have space to fully report on each iteration you go through, but the reader will need a good idea of how your model ‘evolves’, the problems you encountered and how you dealt with them.

Write up your preferred specification

Once you have a final preferred model, carefully report your output. Discuss why it is your preferred model. Interpret your model in terms of economic theory – i.e. what economic relationships does your regression model suggest are important. Are there still any problems or issues? Summarize what you believe you have found. Note it is very unlikely your preferred model will be ‘perfect’ – it is highly likely some (hopefully small) issues are likely to persist. Often you find a ‘trade-off’, for example one model might give you high R(bar) squared but give you a problem with multicollinearity, changing your model to deal with the multicollinearity might worsen in it in other ways (this is not a hint – just an example). Please also note there is no ‘correct’ model. If your friend has a different final specification from you, that doesn’t mean they are right (wrong) and you are wrong (right). I am much more interested in how you develop your model, the rationale for your specification, if you properly test it, how you interpret it, etc. than I am in your final model.

Your project report

Please stick to the word limit, it is 1500 words. How you organise your report is up to you. You might consider 5 sections: (1) A discussion of your initial specification, drawing on the literature and your initial data analysis; (2) A discussion of the output of your initial model, its shortcomings and your strategy to improve it; (3) A presentation and discussion of your ‘improved’ model; (4) presentation and discussion of your final model, including a discussion of the process that took you from (2) to (3); (5) A conclusion including any problems that might persist with your model. Please note if you feel you have found a good enough model in (2) and this doesn’t need improving, I will expect more detail on how you get to it from the initial model.

Please make sure you present the relevant output from Stata that you base your discussion on and refer to it in your report. This includes any relevant plots, test stats and p-values. You can put these in an appendix if you wish as long as you carefully direct the reader to the relevant material. It is fine to use your generated log file for this purpose. Please also append your report with a copy of your *do.file.* Your Stata output and *do.file.* do not count towards your word limit.