ECON 308: Econometrics
Assignment 5

Complete each problem to the best of your ability and submit in class on Thursday, December 9. You are encouraged to collaborate with other students, but you should turn in the problem solutions individually. Your writeup should include 1) written/typed responses to the questions, including regression tables where needed, 2) the code you ran to generate them (your do-file), and 3) any graphs produced.

1. Suppose a researcher wishes to determine how changes in property tax rates affect housing values. The researcher uses differences-in-differences, exploiting property tax rate changes across counties in a single state.

   (a) For this type of research design, what assumption must hold if they are to accurately estimate the causal effect of property tax rate changes?

   (b) Give two examples of factors that you reasonably expect would invalidate this assumption in this specific context.

2. Castle Doctrine laws and homicides: The objective of this question is to replicate the main results of Cheng & Hoekstra (2013). We will focus on the homicide results in Table 5, Panels A-B of the paper (which you can find on Blackboard). The authors use a difference-in-difference strategy to examine the impact of expanded self-defense laws on crime rates. It may be helpful to look over the paper before attempting to complete the analysis below.

   (a) Create a do-file and load the data, which you can download from Blackboard. Be sure to preface the do-file with `clear all`. You should have 550 observations, one for each state and year over the period 2000 to 2010. This is a panel dataset, tracking the evolution of outcomes for each state over this time period.

   (b) Run `ssc install estout` to install the `estout` package. This will allow you to store regression results and export them into a nicely formatted table, which you can include in your writeup.

   (c) The raw data require some transformations. Generate log versions of `homicide`, `police`, `prisoner`, `lagprisoner`, `income`, `exp_subsidy`, `exp_pubwelfare`, `larceny`, `motor`. You will use these as outcome and predictor variables. In your writeup, define what each of these variables are (using the discussion in the paper).

   (d) Begin by replicating the first column of Panel B, which presents the results without weighting states by population. This means that each state will be treated as an equal unit. The only covariates here are the the state and year indicators (which we'll call *fixed effects*) and the indicator for the presence of a Castle Doctrine law (`cdl`); this variable gives the share of

months in each year where a Castle Doctrine law was in effect.[1] Regress your logged homicide variable on the Castle Doctrine indicator and state/year fixed effects. Use the numeric state id variable (`sid`) to generate the state fixed effects, since `state` is a string variable. Add `eststo:` to the beginning of your regression, i.e., it should look like `eststo: reg ...`. This will store the estimates you generate. Also, add `, vce(cluster sid)` to the end of your regression - this will compute standard errors that account for correlation in the error terms over time within state. You will get slightly different values than they do for the standard errors, but the coefficients should match.

(e) What is the interpretation of the coefficient on `cdl`?

(f) What purpose do the year and state fixed effects serve?

(g) Replicate column 2 of Panel B. This requires you to include region-by-year fixed effects, i.e., an indicator that equals 1 if and only if the observation is in a particular region in a particular year (for every region-year combination). There are four regions in the data. You can generate these manually, but that would be time consuming. This code will generate region-by-year fixed effects using a loop:

```
forvalues i=2000/2010{
gen year‘i’=(year==‘i’)
gen r1_‘i’=year‘i’*northeast
gen r2_‘i’=year‘i’*midwest
gen r3_‘i’=year‘i’*south
gen r4_‘i’=year‘i’*west
drop year‘i’
}
```

(h) What purpose do these fixed effects serve? How do they change the results?

(i) Continue and replicate columns 3-5; column 6 requires that you generate state-specific linear time trends. You can do that by creating, for each state, a “trend” variable that starts at 1 in the year 2000 and increases by 1 in each year. The following code will be helpful:

```
forvalues i=1/51{
sort state year
by state: gen trend_‘i’=_n
replace trend_‘i’=0 if ‘i’!=sid
}
```

(j) What variables are added for the regression in column 3, and why do you think they might be useful?

(k) What is the point of the variable added in column 4? What does the estimated coefficient tell you?

(l) Why do the authors include contemporaneous crime rates in column 5?

(m) What flexibility do the additional trends in column 6 add? Why does it make sense to include these?

(n) Use the following code to export your stored results to a table:

[1] It’s zero when no law was present that year, and 1 when it was present all year. So it’s the analog of the treatment × post variable we discussed in class.

```
esttab using results.csv, keep(cdl pre2_cdl) se replace
```

(o) The models estimated in Panel A are identical to those in Panel B - the only difference is that states are *weighted* by population size. The states are weighted analytically, meaning that results from larger states are expected to be more precise. This makes sense - we might expect year to year fluctuations in crime (relative to the mean) would be smaller in big states like California, but more variable in small states with few murders, like Hawaii. Thus, larger states are allowed to contribute more to the result. To generate a population weight for each state, average its population over the 11 years observed:

```
sort sid
by sid: egen pop_weight=mean(population)
```

You can then use this weight in your regressions by adding `[aweight=pop_weight]` after the last variable in each model (and before the comma). Replicate all six columns, and store and export the results as in part (n). Include this table in your writeup. How do the conclusions you draw from Panel A differ from those in Panel B?

3. Class sizes and student performance: An important question in education policymaking is the relationship between class sizes and learning. Smaller classes are thought to promote more one-on-one interactions between teacher and student, and also allow for more effective monitoring. However, they also require more teachers to serve a given student population, which entails additional costs. So it would be useful to know just how much benefit students receive from smaller class sizes. Here, we'll think about what we might encounter when attempting to study this question, and explore an IV strategy that could help.

(a) Suppose you ran a regression of student test performance against class sizes and found a negative relationship. Can you conclude that larger classes inhibit learning? What omitted variables might cause a failure of the exogeneity assumption in this situation?

(b) Create a do-file and load the dataset `maimonides.dta`, which is posted on Blackboard. Be sure to preface the do-file with `clear all`. This is a cross-sectional dataset containing data on average math and verbal test scores for 1,185 fifth grade classes across 733 schools in Israel. The variable `enrollment` gives the total fifth grade enrollment for the school, `classize` gives the size of each class, `avgmath` is the average math score for the class, `avgverb` the average verbal score, and `perc_disadvantaged` gives the share of students in the class who are classified as being from a disadvantaged background.

(c) Regress math scores on `classize`, `perc_disadvantaged`, `enrollment`, and `enrollment`^{2}; this will give you the relationship between class sizes and math scores, controlling for % disadvantaged and a quadratic for total enrollment. Report your results and interpret; store these estimates by prefacing the regression with `eststo:`.

(d) Because of the problems you discussed in part (a), it would be nice to have some variation in class sizes that we could plausibly say was exogenous. Fortunately, such variation can be found in this data due to an unusual feature of the Israeli education system: The application of *Maimonides' rule*. Maimonides was a medieval rabbinic scholar who was also interested in education policy, and like many, he believed smaller class sizes are more conducive to learning. Based on his interpretation of the Talmud, he proposed a rule that a class should have no more than 40 students; if there are more than 40 students, the class should be evenly divided into two classes. It turns out that this rule has been employed by Israeli schools

in modern times. Generate a scatterplot using `scatter`, with class size on the y-axis and enrollment on the x-axis, and report what you find.

(e) To investigate the true impact of class size on test scores, we'll use the Maimonides cutoff as an instrumental variable. Generate a new variable that is equal to one if total enrollment is strictly greater than 40, and zero otherwise. This will be our enrollment instrument - regress `classize` on the instrument as well as `perc_disadvantaged`, `enrollment`, and `enrollment`^{2}. Does it pass the relevance test?

(f) Now, we'll use two-stage least squares to replicate the regression from part (c), except this time, we will instrument for class size using the enrollment cutoff. Rerun the regression from part (c), but use `ivregress 2sls` instead of `reg`, and instrument for class size using the instrument constructed in part (e); the IV notes posted on Blackboard may be helpful for structuring the code. Store these results.

(g) How do the OLS and IV results differ? Interpret what this new result tells you. Export the stored results from both (c) and (f) to a table using `estout`, and include it in your writeup.