

ECO 348K, Problem Set #4, due 3/25/22. Please submit answers + Stata code + .log files on Canvas.

Q1. Open ClassSizeRKD.dta, this data should look very familiar. It is the exact same as ClassSizeRD.dta that you used in Problem Set #3 except it now contains a new variable named num_classrooms. That variable is equal to the number of classrooms in each grade cohort. You should not need to use the data for parts (a)-(c). Remember to cluster standard errors at the level of coh_id, like you did in problem set #3.

a. Assume that Maimonides Rule is applied as follows. If a grade cohort contains 40 students or less, they are all put together in the same classroom. If a grade cohort contains 41-80 students, they are split into classes of equal size. Based on this, draw a graph of the relationship between cohort size and number of classes in the grade cohort. In this theoretical relationship, what is the discontinuity in number of classes at the cutoff and what is the change in slope at the cutoff?

b. Based on the description in (a), draw a graph of the relationship between cohort size and class size in the grade cohort. In this theoretical relationship, what is the discontinuity in class size at the cutoff and what is the change in slope at the cutoff?

c. When number of classes in a grade cohort increases, schools need to hire more teachers to teach those classes. Assume that school budgets are limited, so when they need to hire more teachers, they pay each teacher less, and the quality of teachers employed for the grade cohort is lower. If this is true, then based on your answers to (a) and (b), does the RDD that you estimated in problem set #4 identify the causal effect of class size on test scores? Why or why not? If not, explain whether this introduces positive or negative bias to your estimate. In other words, does this cause you to underestimate or overestimate the effect of class size on test scores using the RDD in problem set #3? Why?

d. Limit the data to observations where cohsize < 80. As you did in problem set #3, create a binary variable equal to one if the cohort size is greater than 40, and equal to zero if the cohort size is less than or equal to 40. Also, create a variable that centers cohsize at zero when it is equal to 40. Now, using a bandwidth of +/- 40 in cohort size and a linear polynomial, estimate a standard RKD regression with num_classrooms as the dependent variable. What do the three main coefficients mean? Next, using a bandwidth of +/- 40 in cohort size and a linear polynomial, estimate a standard RKD regression with class size as the dependent variable. What do the three main coefficients mean?

e. Using a bandwidth of +/- 40 in cohort size and a linear polynomial, estimate a standard RKD regression with avgscore as the dependent variable. Using the estimates from the regressions in part (d) as needed, what does this regression tell us about: (1) the direction of the combined effect of class size and number of classes on student test scores (i.e. is it positive or negative?) and (2) the direction and size of the effect of class size alone on student test scores? Please explain your answer to both.

Q2. Open SpeedDiscount.dta. This is a random 10% sample of the data from the paper “A Few Bad Apples? Racial Bias in Policing” by Felipe Goncalves and Steven Mello.

a. Replicate the main results of the paper in this sample by regressing discount on white, lenient, and the interaction between the two variables. Include county fixed effects and year fixed effects. Note: this is slightly different from the regression that they use in the paper, but we will use this version for

convenience. Describe what the interaction term in this regression means and how it relates to officer discrimination.

b. Now, instead of estimating officer discrimination on the basis of race, estimate it based on the following characteristics: (1) whether the driver is female or not, (2) the age of the driver, (3) whether the driver is in a car that has a price > \$25,000, (4) whether the driver has any tickets in the prior year, and (5) whether the driver has an in-state license or not. Finally, estimate officer discrimination by driver race/ethnicity and driver sex. Specifically, estimate the difference in speed discounting from lenient officers for black male drivers, black female drivers, Hispanic male drivers, Hispanic female drivers, and white female drivers all relative to the omitted category of white male drivers. As before, include county fixed effects and year fixed effects in all of your regressions.

Stata Guide

These are the Stata commands & functions that I used to get the answers for all of the questions above. Although this is how I did it, you can use any commands you see fit unless stated otherwise in the problem.

keep if ... -- used to keep observations if a condition is met (you would need to extend the if statement as needed).

gen – used to generate various variables as needed.

reg y x, cl(clustervar) – used to run a regression of y on x, adjusting standard errors for clustering based on the variable clustervar (insert appropriate variables as needed)

areg y x i.year, a(countynum) r – used to run a regression of y on x with year fixed effects and county fixed effects, using robust standard errors.