ECON 626: Empirical Microeconomics

Problem Set 4

Department of Economics
University of Maryland
Fall 2019

Problem Set 4 is due at 5pm on Thursday, October 31.

For this problem set, you will use Stata for each question. You should write a `.do` file that generates your answers; please submit this file via email when you turn in your problem set. Insert comments in your `.do` file to indicate where you are answering each part of each question.

1. **Clusters.** Consider an error term, ϵ_i, under two scenarios.

   In the first scenario, $\epsilon_i \sim \mathcal{N}(0,1)$, with independent and identically distributed draws: $\epsilon_i \perp\!\!\!\perp \epsilon_j \; \forall i \neq j$. There are 800 observations of $Y_i = \beta T_i + \epsilon_i$, of which 400 are randomly assigned to treatment ($T_i = 1$) and the other 400 of which are randomly assigned to comparison ($T_i = 0$).

   In the second scenario, $\epsilon_i = \nu_g + \eta_i$. In this case, there are groups of observations within which there is a common component of the error term, ν_g, and there is an independent component, η_i. Let $\eta_i \sim \mathcal{N}(0, 0.64)$ (meaning SD=0.8), with independent and identically distributed draws: $\eta_i \perp\!\!\!\perp \eta_j \; \forall i \neq j$. Similarly, $\nu_g \sim \mathcal{N}(0, 0.36)$ (meaning SD=0.6), with independent and identically distributed draws: $\nu_g \perp\!\!\!\perp \nu_h \; \forall g \neq h$. Let groups be of size 16. There are 800 observations of $Y_i = \beta T_i + \epsilon_i$, of which 400 (that is, 25 groups of 16) are randomly assigned to treatment ($T_i = 1$) and the other 400 of which (again, the other 25 groups of 16) are randomly assigned to comparison ($T_i = 0$).

   Recall that the standard error of the coefficient on treatment should be (approximately) the standard deviation of the difference between the sample mean for treatment observations and the sample mean for comparison observations.

   (a) In the first scenario, what is the standard deviation of the sample mean of the 400 treatment observations? (Show analytically what it should be.)

   (b) In the first scenario, what is the standard deviation of the *difference* between treatment and comparison sample means? (Show analytically what it should be.)

   (c) In Stata, generate data according to the first DGP, and with $\beta = 0$ or any other value of β you choose, run the regression, showing that the standard error is very close to the value you worked out analytically.

   (d) What is the approximate MDE for power of 0.8 or 0.9, analytically (using the formula in the Lecture 7 slides)?

   (e) Use the `sampsi` command to confirm that the sample size required for the power and MDE you found above is roughly 800 (400 T, 400 C).

(f) In the second scenario, what is the standard deviation of the sample mean of the 400 treatment observations? (Show analytically what it should be.)

(g) In the second scenario, what is the standard deviation of the *difference* between treatment and comparison sample means? (Show analytically what it should be.)

(h) In Stata, generate data according to the second DGP, and with $\beta = 0$ or any other value of β you choose, run the regression, clustering standard errors at the group level, and showing that the standard error is very close to the value you worked out analytically.

(i) What should the intra-cluster correlation (ρ) be, according to the formula given in Lecture 7?

(j) Using the `loneway` command in Stata, confirm that ϵ has an intra-cluster correlation approximately equal to the value you calculated.

(k) What is the "design effect," following the Lecture 7 formula?

(l) What is the approximate MDE for power of 0.8 or 0.9, analytically (using the formula in the Lecture 7 slides)?

(m) Use the `sampsi` and `sampclus` commands to confirm that the sample size required for the power and MDE you found above is roughly 800 (400 T, 400 C, 50 clusters total, 25 per arm).

2. **IV: First Stage F-Statistic with Robust Standard Errors.** As shown at the end of the lecture on instrumental variables, conduct "experiment 1" and "experiment 2," then a third variant described below, being sure to use **robust** standard errors. That is:

   (a) First, create a dataset of 100 observations where the outcome x is a standard normal random variable, distributed $\mathcal{N}(0,1)$, and the right-hand-side variables are two random variables z_1 and z_2 each with a uniform distribution $\mathcal{U}(0,1)$, all independent of one another (so the null of no relationship is true). Repeat this 100 or 1000 times, each time regressing the outcome on the two right-hand-side variables, **being sure to include the , `robust` option**, and testing their joint significance. You may wish to use `Ftail(e(df_m),e(df_r),e(F))` after running a regression in Stata to obtain the p-value from the joint F-test. Show your code, and the resulting histogram of p-values.

   (b) Repeat the exercise above, but with 10 observations.

   (c) Repeat the exercise in problem 2b above, but with the right-hand-size variables z_1 and z_2 distributed according to a standard normal distribution; do this by generating each of them as the square root of a variable that is distributed uniformly $\mathcal{U}(0,1)$.

   (d) Repeat the exercises in parts 2a, 2b, and 2c above, but **without** the robust option in the regression. How do results change?

   (e) What do you conclude about when the first-stage F-statistic with multiple instruments might be misleading? How does this relate to anything written by Alwyn Young in the first two pages of the introduction to his paper, "Consistency without Inference?"

   Programming tip, in case it is useful to some of you: Be careful when recording p-values when there are more iterations than observations in your dataset. If you have Stata 16,

you could do this with frames; if not, you might write p-values to a file that you load at the conclusion of the loops, or you might save the p-values to local macros or a matrix that you then load into data at the conclusion of the loops.

3. **Post-Double-Lasso.** The do file `econ626-2019-L6-A3-post-double-lasso.do` generates standord normal random variables $A1$–$A9$, $B1$–$B9$, and $C1$–$C9$ and then uses them to generate a treatment dummy, T and an outcome variable Y. The A variables predict treatment, the B variables predict both treatment and Y, and the C variables predict Y but not T. The program is a loop, but the local iters is set to 1, so the loop will only run once.

   (a) Run the do file to generate all the variables, and then regress Y on T. What does the coefficient suggest about the impact of T? Now regress Y on T controlling for the Bvariables. How does the coefficient on T change? Why?

   (b) Review the lasso estimation output (from running the code the first time). Which variables does lasso select as predictors of Y? Which variables does lasso select as predictors of T? Are these controls the ones you would expect?

   (c) Set iters to 50 and run the code. Compare the average estimate of beta from post single lasso estimation and post double lasso estimation to the true β.

   (d) Now set scalefactor to 0.1 and rerun the code. How do your results differ from (3)?

   (e) Generate a new treatment dummy, P, that is randomly-assigned (and hence not correlated with any of the other variables, in expectation). How do PSL and PDL compare under the null and when there is a true treatment effect (for example, if $Y2 = Y + P$)?