ECON 626: Empirical Microeconomics

Problem Set 3

Department of Economics
University of Maryland
Fall 2019

Problem Set 3 is due at 5pm on Thursday, October 10.

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. **Power.** In Lecture 2, we showed the variance-covariance matrix for a simple regression, in which outcome was regressed on an indicator for treatment and a constant. The setting was a study in which fraction p of the N participants had the indicator for treatment equal to one. The simplest case was the assumption of homoskedastic error, for which the main descriptions of the resulting variance-covariance matrix appeared on slides 19, 30, and 31 from Lecture 2.

   (a) Suppose $\frac{1}{4}$ of the N observations in the sample are treated. What are $X'X$ and $(X'X)^{-1}$?

   (b) Suppose $\frac{1}{k}$ of the N observations in the sample are treated. What are $X'X$ and $(X'X)^{-1}$?

   (c) Suppose that the error terms are *i.i.d.* draws following a $\mathcal{N}(0,1)$ distribution. Further assume that $\hat{\sigma}^2 = \sigma^2$ (so we will not be concerned with the degrees-of-freedom adjustment). Taking the square root of the upper left element in the variance-covariance matrix, what is the standard error if $N = 10,000$ and ...

      i. $\frac{1}{2}$ of the observations are treated?

      ii. $\frac{1}{4}$ of the observations are treated?

   (d) Simulate eight variations on this arrangement in Stata, going through all combinations of the following options: with error terms having standard deviation 1 or 2; with true treatment effects 0 or 1; and with either $\frac{1}{2}$ or $\frac{1}{4}$ of the observations are treated. In each instance (simulating each possibility just once), what is the standard error for the coefficient on the treatment indicator? How close is it to the analytical calculation you did above? Which of the three factors (the error term SD, the true treatment effect, and the fraction treated) does the standard error depend on, and how?

2. **Instrumental variables.** Begin by generating a dataset following the DGP below:

```
clear all
set obs 10000
version 11.2
version 11.2: set seed 2345
gen instrSum=0
```

```
forvalues i=1/10 {
 qui gen instr‘i’=invnorm(uniform())
 qui replace instrSum=instrSum+instr‘i’
}
gen exogenous1=invnorm(uniform())
gen exogenous2=invnorm(uniform())
gen commonError=3*invnorm(uniform())
gen error1=0.1*invnorm(uniform())+commonError
gen error2=0.1*invnorm(uniform())-commonError
gen endog=0.05*instrSum+exog1-exog2+error1
gen outcome=endog-2*exog1+3*exog2+error2 // True (causal) model
```

(a) Estimate the (naïve) regression: `reg outcome endog exog1 exog2`
Is the coefficient on `endog` biased upwards or downwards in relation to the last line of the data generating process (the true causal model)? Why? What value do you get?

(b) Estimate 2SLS using only one variable, `instrSum`, as an excluded instrument for `endog`. What is the first-stage F-statistic on the excluded instrument? Use two different sets of Stata commands to arrive at this F statistic: first, with `ivregress` followed by `estat firststage`; then, with `regress` (for the first stage only) followed by `test` with the appropriate argument(s).

(c) Estimate 2SLS using ten variables, `instr1` through `instr10` (but not `instrSum`), as excluded instruments for `endog`. What is the first-stage F-statistic on this set of excluded instruments? Again, use the same two sets of Stata commands to confirm this F statistic (though the argument for the `test` command will be longer this time).

(d) Comparing the two F statistics, which is more likely to suffer from a weak instrument problem?

(e) Comparing the 2SLS estimates, which is more positive? Based on the discussion in class and in Mostly Harmless Econometrics on the bias brought about by weak instruments, and the direction of OLS bias you found in the first part of this problem, is this the answer you expect? Why?

(f) If we add "garbage" instruments that have no actual predictive value in the first stage, what should happen to the first-stage F-statistic on all the excluded instruments, and to the bias of the 2SLS estimate?

(g) Try adding the "garbage" instruments yourself using the following commands. Are the results consistent with your answer above?

```
forvalues i=11/20 {
 qui gen instr‘i’=invnorm(uniform())
}
ivregress 2sls outcome (endog=instr1-instr10 instr11-instr20) exog1 exog2
```

3. **Regression discontinuity (centering).** Consider the following data generating process and four possible regression specifications.

```
clear all
set obs 10000
gen r=2*uniform()-1
gen t=cond(r>=0,1,0)
gen itr=t*r
gen yA=2*t+0.1*invnorm(uniform())+0.1*r
gen yB=2*t+0.1*invnorm(uniform())+0.1*r +0.1*itr

gen r10=r+10
gen itr10 = r10*t

reg yA t r itr
reg yB t r itr
reg yA t r10 itr10
reg yB t r10 itr10
```

In the Stata commands above, the running variable is either `r`, with cutoff zero, or `r10`, with cutoff 10. Treatment is `t`, and the actual change in both `yA` and `yB` at the discontinuity is 2. We usually center regression discontinuity specifications at zero because of a relatively simple problem that arises if we don't. That problem is evident in the fourth regression, but none of the other three.

(a) Why does it arise there?

(b) What one-character change to the data generating line for `yB` would make the fourth regression specification produce a coefficient of approximately zero on `t`, and why?

4. **Regression discontinuity (manipulation).** Complete the questions in the .do file associated with this problem set, `PS3-rd-manipulation-questions.do` .