ECON 626: Empirical Microeconomics

Problem Set 2

Department of Economics
University of Maryland
Fall 2019

Problem Set 2 is due at 5pm on Thursday, September 19.

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. **Regression coefficients.** (This problem has lots of short parts.)

   (a) Using any reasonable combination of Stata functions (such as `uniform()`, `invnorm()`, or `rnormal()`), generate a dataset of 10,000 observations with variables as follows. (Be sure to `set seed` so that your code produces identical results when run twice.)

   Generate three variables that are independently drawn from the distribution $\mathcal{N}(0,1)$: `ability`, `epsilon1`, and `epsilon2`. Generate another independent variable, `smallepsilon`, that is drawn from the normal with mean 0 and standard deviation 0.1 (variance 0.01). Generate yet another independent variable, `wideuniform`, that is $\mathcal{U}(-3,3)$ (drawn from the uniform distribution, distributed between -3 and 3).

   Next:
   Generate `education`, so that `education = ability + epsilon1`.
   Generate `wage1`, so that `wage1 = education + ability + epsilon2`.
   Generate `wage2`, so that `wage2 = education + smallepsilon`.
   Generate `noisywage2`, so that `noisywage2 = wage2 + wideuniform`.
   Generate `noisyeducation`, so that `noisyeducation = education + wideuniform`.

   (b) Summarize the variables using `sum`. The standard deviations of `ability`, `epsilon1`, and `epsilon2` should all be close to 1. The standard deviations of `education` and `wage1` are close to other square roots of integers. Which integers? Formally demonstrate why this is the case in terms of the expectations or variances of normal distributions.

   (c) Type `correlate, covariance` to see the covariance matrix. What integer is the covariance of `ability` and `education` close to? Formally demonstrate why this is the case in terms of the expectations or variances of normal distributions.

   (d) In your sample, what is the value of the following ratio:
   Cov(`ability`,`education`)/Var(`ability`)?

   (e) Type `reg education ability`. What is the coefficient on `ability`?

   (f) What integer is the covariance of `wage1` and `education` close to? In terms of expectations or variances of normal distributions, formally demonstrate why this is the case.

(g) In your sample, what is the value of the following ratio:
Cov(`wage1`,`education`)/Var(`education`)?

(h) Type `reg wage1 education`. What is the coefficient on `education`?

(i) Type `reg wage1 education ability`. What is the coefficient on `education`?

(j) Up to here, this problem dealt with basics of regression and potential biases when not including important controls in regressions with observational data. Next we consider attenuation bias. Type `reg wage2 education`. What is the coefficient on `education`? Why?

(k) Make a scatter plot scattering `wage2` against `education`, overlaying the linear fit given by the regression. This is relevant for visual comparison to the graphs you will make in the next parts of the problem.

(l) Now, type `reg noisywage2 education`. What is the coefficient on `education`? Why?

(m) Make a scatter plot scattering `noisywage2` against `education`, overlaying the linear fit given by the regression. This should provide some visual intuition about why the regression coefficient did or did not change between the previous parts of the problem (parts 1j and 1l).

(n) Now, type `reg wage2 noisyeducation`. What is the coefficient on `education`? What simple fraction is that coefficient close to? In calculations you can do by hand, using the data generating process, variances, and covariances, why is this the regression coefficient?

(o) Make a scatter plot scattering `wage2` against `noisyeducation`, overlaying the linear fit given by the associated regression. This should provide some visual intuition about why the regression coefficient did or did not change between parts 1j and 1n.

(p) How do parts 1j, 1l, and 1n relate to the "Attenuation bias" exercise and associated formula from that handout in class (Lecture 2)?

2. **Hospital example.**

   (a) This problem simulated the hospital example that we discussed on the first day of class. Using any reasonable combination of Stata functions (such as `uniform()`, `invnorm()`, or `rnormal()`), generate a dataset of 1,000 observations with variables as follows. Generate three variables that are independently drawn: let `z` be drawn from the normal distribution $\mathcal{N}(0, 4)$ (variance=4), while `u1` and `u2` are drawn from $\mathcal{U}(0, 1)$. Be sure to `set seed` so that your code produces identical results when run twice.

   (b) Generate a variable `illness` that is a dummy (indicator) variable equal to one if `u1` $>$ 0.5. This should be roughly half the observations.

   (c) Consider the cost of sickness, $s = 5$; the benefit of treatment, $b = 4$; and the cost of going to the hospital, $c = 1$. Generate potential outcomes: `y0h` is the outcome without hospitalization with good health, so simply `z`; `y0s` is the outcome without hospitalization when sick, so `z` $- s$; `y1h` is the outcome for hospitalizing the healthy, `z` $- c$; and `y1s` is outcome for hospitalizing those who are sick, `z` $- s + b - c$.

   (d) Scenario 1: treat the sick. Generate a variable, `dS` that is an indicator for going to the hospital in this scenario: `dS` = `illness`. Generate the outcome, `yS`, based on the potential outcomes. If someone is ill (`illness` == 1), and they go to the hospital (`dS` == 1), then `yS` == `y1s`, and so on. Regress `yS` on `dS`. Is hospitalization associated with a better or worse outcome? How does the answer relate to the formula $b - c - s$ given in slides on the first day of class?

   (e) Scenario 2: randomize. Use `u2` to determine whether someone goes to the hospital. Generate a variable, `dR` that is an indicator for going to the hospital in this scenario: `dR` = 1 only for the first 5000 observations when observations are sorted by `u2`. Generate the outcome, `yR`, based on the potential outcomes. If someone is ill (`illness` == 1), and they go to the hospital (`dR` == 1), then `yR` == `y1s`, and so on. Regress `yR` on `dR`. Is hospitalization associated with a better or worse outcome? How does the answer relate to the formula $\lambda b - c$ given in slides on the first day of class?

   (f) Scenario 3: randomize only among the sick. Regress `yR` on `dR`, but only on those for who `illness` == 1. Is hospitalization associated with a better or worse outcome? How does the answer relate to the formula $b - c$ given in slides on the first day of class?

   (g) Scenario 4: The slides on the first day of class also include an endogenous take-up scenario in which instead of randomizing hospitalization, the study randomizes access to the hospital. Construct this scenario (commenting carefully as you explain how you do this), and run the regression. What is the result? How does it align with the formula in the slides?

3. **Diff-in-Diff.** This problem builds on the "How Much Should We Trust Difference-in-Differences Estimates?" exercise that you did in class. Simulate a data-generating process with serially correlated errors. Specifically, generate a sample of 100 individual units ("people") that you will observe over time. For each unit i, draw $\alpha_i \sim \mathcal{U}(0, 100)$. Then expand the data set so that you have 30 observations per person; generate a `time` variable that ranges from 1 to 30 for each observation.

   Now generate an error term $\varepsilon_{i,t} \sim \mathcal{N}(0, 4)$. Construct the outcome variable $y_{i,t}$ according to the following process:
$$y_{i,t} = \alpha_i + \varepsilon_{i,t} + 0.5 * \varepsilon_{i,t-1}$$

   (a) First, consider the case where treatment is randomly assigned across people and time periods. Randomly choose half of your person-time observations to be assigned to treatment. Write a `.do` that simulates such a process 250 times and then estimates two OLS regressions of the outcome on treatment controlling for individual fixed effects and time fixed effects. Estimate one specification with robust standard errors and another specification with standard errors clustered at the person level. Store the p-values associated with each test of the hypothesis that the treatment effect is equal to zero. Do your results suggest that the two tests are correctly sized?

   (b) Next, consider a setting where treatment is randomly phased in across people. In other words, randomly choose half of the people in the sample to receive treatment; then, for each treated individual, randomly select a start time between $t = 11$ and $t = 20$. Once the treatment starts for a given individual, it remains active in all future periods. Simulate this process 250 times, estimating the treatment effect of the randomly-generated treatment while controlling for person FEs and time FEs; estimate a specification with robust standard errors and another specification with standard errors clustered at the person level. Store the p-values associated with each test of the hypothesis that the treatment effect is equal to zero. Do your results suggest that the two tests are correctly sized?

   (c) Modify your program so that the treatment starts in the same time period for all treated observations. Estimate the specification with robust standard errors and the specification with clustered standard errors 250 times (for 250 randomly-generated treatments). Report the observed rejection probabilities. Do your results suggest that variation in treatment timing mitigates the problem of serial correlation?

4. **Bad control.** This problem follows the example in *Mostly Harmless*, section 3.2.3.

   (a) Using any reasonable combination of Stata functions (such as `uniform()`, `invnorm()`, or `rnormal()`), generate a dataset of 10,000 observations with variables as follows. Be sure to `set seed` so that your code produces identical results when run twice. All four of the variables above should be independent from one another.

      i. Generate an `ability` variable that is $\mathcal{U}(-1, 1)$ (uniformly distributed between -1 and 1);

      ii. Generate a `college` indicator variable that is equal to one for a random half of the observations;

      iii. Generate `epsilon1` and `epsilon2` so that they each have a standard normal distribution.

   (b) Generate `w0`, a variable indicating whether someone would be a white-collar worker in the absence of college. Interpreting `epsilon1` as inclination to be a white-collar worker, `w0` should be 1 whenever `epsilon1` is greater than zero, and it should be zero otherwise.

   (c) Next, generate two versions of the (potential outcome) white-collar indicator in the presence of college that represent two scenarios. In the first scenario, college causes all low-ability workers (`ability`$<$ 0) to become white-collar, but doesn't affect the behavior of high-ability workers. Generate an indicator `w1v1` for this scenario. In the second scenario, college causes all high-ability workers (`ability`$\geq$ 0) to become white-collar, but doesn't affect the behavior of low-ability workers. Generate an indicator `w1v2` for this second scenario. Note that these are just potential outcomes, so they don't yet depend on whether the individual actually went to college; they just depend on `ability` and `w0`.

   (d) Next, generate two versions of the actual white-collar indicator. Generate `wv1` so that it equals `w0` for those who don't go to college, and `w1v1` for those who do. Likewise, generate `wv2` so that it equals `w0` for those who don't go to college, and `w1v2` for those who do.

   (e) Next, generate the earnings potential outcome in the absence of college. Generate `y0` so that it equals three times `ability` plus `epsilon2`.

   (f) Next, generate the earnings potential outcome in the presence of college. Generate `y1` so that it equals `y0` plus one.

   (g) Next, generate actual earnings. Generate `y` so that it equals `y0` when `college` $==$ 0, and `y1` when `college` $==$ 1.

   (h) We are ready to run a regression. First, regress `wv1` on `college`; the coefficient should be about 0.25. Why?

   (i) Next, regress `wv2` on `college`; the coefficient should also be about 0.25. Why?

   (j) Finally, following the example in *Mostly Harmless Econometrics*, regress `y` on `college`, but only in the sample where `wv1` $==$ 1. As the book suggests, this should be the sum of a causal effect and a selection bias term.

      i. What should the causal effect be, in expectation? Why?

ii. What should each of the expectations in the selection bias term equal, in expectation? Why?

iii. How close does your regression coefficient come to the value you expect? Is the value you calculated for the sum of the causal effect and the selection bias terms within the confidence interval for the estimated coefficient?

(k) Now, regress `y` on `college`, but only in the sample where `wv2` == 1. As the book suggests, this should be the sum of a causal effect and a selection bias term. How do the results change, and why?

(l) Lastly, regress `y` on `college` without restricting the sample or including any other controls. What coefficient do you get, and why?