ECON 626: Empirical Microeconomics

Problem Set 6

Department of Economics
University of Maryland
Fall 2019

Problem Set 6 is due at 5pm on Tuesday, December 3.

1. **Lee and Manski Bounds.** Consider a dataset of the following form. Ten units have been randomly assigned: five to treatment, and five to comparison. Of these, we observe the endline outcome data for four of the five comparison units, and three of the five treatment units. The outcome is a fraction that could theoretically be anything from zero to one. What we observe is that the four non-missing comparison outcome values are 0.1, 0.2, 0.3, and 0.4. The three non-missing treatment outcome values are 0.5, 0.6, and 0.7.

   Thus the naïve estimate for the mean in the treatment group is 0.6; in the comparison group it is 0.25.

   (a) What is the naïve treatment effect?

   (b) Which mean(s) is/are adjusted for calculation of Lee upper and lower bounds - treatment, comparison, or both?

   (c) What are these adjusted means for the Lee upper and lower bounds of the (treatment-comparison) difference?

   (d) Thus, what are the Lee upper and lower bounds for the treatment effect?

   (e) Which mean(s) is/are adjusted for calculation of Manski upper and lower bounds - treatment, comparison, or both?

   (f) What are these adjusted means for the Manski upper and lower bounds of the (treatment-comparison) difference?

   (g) Thus, what are the Manski upper and lower bounds for the treament effect?

   (h) You should have been able to do parts 1a through 1g above by hand. But in this step, use Stata (in particular, the `regress` and `leebounds` commands) to confirm that your answers to parts 1a and 1d are correct. You may find it convenient to type this small dataset directly into Stata using the `input` command.

2. **Power: RCT, RD.** In Stata, generate a dataset as follows. There are 10,000 observations. A random error term is distributed $\mathcal{N}(0, 1)$.

   (a) Randomly assign half the observations to treatment, and half to comparison. Generate `y1` so that it is the random error term plus, for treated units, a treatment effect of 0.1. When you regress `y1` on treatment, what is the standard error? How wide is the 95% confidence interval?

(b) Statistical power can be approximated by taking the true effect size, dividing by the standard error, subtracting 1.96, and computing the normal CDF at the result. For example, if the true effect size divided by the standard error yielded a result of 3, the approximate power would be $\Phi(3 - 1.96) = \Phi(1.04) = 0.85$. What is the approximate statistical power for the true effect size of 0.1, given the standard error you calculated in part 2a above?

(c) Randomly create a new variable `r`, uniformly distributed between -1 and 1. Create a dummy variable, `d`, for whether `r` is zero or greater. Create an interaction between the two, `dr`. Generate `y2` equal to the random error term plus, for units with `d`==1, a treatment effect of 0.1. For a simple regression discontinuity specification (ignoring bandwidth and kernel concerns), regress `y2` on `d`, `r`, and `dr`. When you do so, what is the standard error for the coefficient on `d`? How wide is the 95% confidence interval?

(d) What is the approximate ratio between the standard error from the RD design and the standard error from the RCT design, in this simple case?

(e) Use the `rd` or `rdrobust` command to estimate the effect in part 2c, but now with the optimal bandwidth and kernel selected by whichever command you use. Is the standard error larger or smaller than it was in part 2c? Why?

(f) As in part 2b above, what is the approximate statistical power for the true effect size of 0.1, given the RD standard error you calculated in part 2c above?

(g) Repeat the data generating processes in parts 2a and 2c 100 times, and count the number of rejections of the null at the 5% level. How many do you get in each case? Is this in accord with your approximate power calculations?

3. **Tobit likelihood function.** Consider one observation from the model in which:

$$y_i^* = x_i'\beta + \epsilon_i$$
$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

(Assume ϵ_i are iid and have the distribution given above, conditional on any value of x_i.) The likelihood function was given in lecture: it is based on the normal CDF for censored values, and based on the normal density for uncensored values. Use the label θ to denote the parameters, β and σ^2. For parts 3a through 3f, provide a symbolic answer in terms of normal CDF or density (Φ or ϕ) as well as an approximate numerical answer, correct to a few decimal places, with the help of Stata or any other program to calculate the answer.

(a) Suppose that $x_i'\beta = 0$, $y_i = 0$, and $\sigma^2 = 1$. What is $\mathcal{L}(\theta)$?

(b) Suppose that $x_i'\beta = 0$, $y_i = 0$, and $\sigma^2 = 4$. What is $\mathcal{L}(\theta)$?

(c) Suppose that $x_i'\beta = 0$, $y_i = 0.1$, and $\sigma^2 = 1$. What is $\mathcal{L}(\theta)$?

(d) Suppose that $x_i'\beta = 0$, $y_i = 0.1$, and $\sigma^2 = 4$. What is $\mathcal{L}(\theta)$?

(e) Suppose that $x_i'\beta = 0$, $y_i = 2$, and $\sigma^2 = 1$. What is $\mathcal{L}(\theta)$?

(f) Suppose that $x_i'\beta = 0$, $y_i = 2$, and $\sigma^2 = 4$. What is $\mathcal{L}(\theta)$?

(g) Which is greater, the answer to 3a or 3b, or are they equal? Intuitively, why?

(h) Which is greater, the answer to 3c or 3d, or are they equal? Intuitively, why?

(i) Which is greater, the answer to 3e or 3f, or are they equal? Intuitively, why?

4. **Estimating risk preference parameters.** Consider a population of agents whose risk preferences take the constant relative risk aversion (CRRA) form

$$u(x) = \frac{x^{1-\rho}}{1-\rho}$$

where the individual CRRA coefficient, ρ_i, depends on i's education level (i.e. years of schooling):

$$\rho_i = \rho_0 + \rho_{ed} \cdot educ_i.$$

Simulate a population of 100 such individuals. Let $\rho_0 = 0.1$ and $\rho_{ed} = 0.05$. Let the level of education take integer values 0 through 12 (years of schooling); assume all possible education levels are equally likely.

Each individual in your simulated population decides how much of a 10 dollar endowment to invest in a risky security. With probability one half, the amount invested in the security is lost (so, for investment level x, the final payoff is $10 - x$). With probability one half, the amount invested is multiplied by 6 (so the final payoff is $10 + 5x$).

(a) Solve for the optimal interior solution to this individual's utility maximization problem, and use this to write down the log likelihood function for the model parameters, including the required censoring adjustments.

(b) Estimate ρ_0 and ρ_{ed} via non-linear least squares using Stata's `nl` command.

(c) Estimate ρ_0 and ρ_{ed} via maximum likelihood using Stata's `ml` command(s). Compare your results to those reported above.

5. **Multiple hypothesis testing.** As in the example table presented in Lecture 9, in this problem, take a set of p-values that came from multiple hypothesis tests, and produce the associated Bonferroni-adjusted p-values, p-values from the Holm procedure, and q-values from the Anderson procedure. Specifically, consider that you have conducted four hypothesis tests. The p-values from those tests are 0.01, 0.02, 0.03, and 0.04. What are the Bonferroni-adjusted p-values, the p-values from the Holm procedure, and the q-values from the Anderson procedure?