Problem Set 0

Since we have yet to begin the main discussion of estimators, this problem set explores and reviews concepts and theory. Subsequent problem sets will introduce real-world data sets.

Parts with an asterisk (*) are "optional". As a practical matter, doing these parts is unnecessary for passing the course, but would likely be necessary for getting an A.

1. You have a randomized controlled trial (RCT) with potential outcomes $Y_i(0), Y_i(1)$ and treatment D_i. Let $P(D_i = 1) = 0.5$ and let $i = 1, ..., N$. You observe $Y_i = Y_i(0) \cdot (1 - D_i) + Y_i(1) \cdot D_i$.

   (a) Consider a regression of Y_i on D_i. Prove that the regression of Y_i on D_i estimates ATE. Does it estimate TOT too? If so, prove that.

   (b) Does setting $P(D_i = 1) = 0.5$ maximize the power (i.e. minimize the standard error of the estimated treatment effect) in your RCT? Prove. (You may take the formula for $\text{Var}(\hat{\beta}_{OLS})$ as given; feel free to look it up if you do not recall it.)

   (c) * Suppose you also observe covariates X_i. You are concerned about heteroskedaticity and estimate a regression of $\hat{\varepsilon}_i^2$ on X_i, where $\hat{\varepsilon}_i$ are OLS residuals from regressing Y_i on X_i. Using the fitted values from this regression you form weights $w_i = 1/\hat{\sigma}_i$ and reweight your data. Prove that a regression of Y_i^w on D_i^w, where a w superscript signifies a reweighted variable, does not necessarily estimate ATE. Does it estimate TOT?

   (d) Assuming your model of heteroskedasticity is correct, name one advantage and one disadvantage of the weighted regression (versus OLS). You may rely on results you've learned from previous econometrics courses.

   (e) Suppose that it costs \$1 to generate each control observation and \$4 to generate each treated observation (i.e. the treatment itself is expensive). You have a total budget of \$150. To maximize power in your RCT, how many observations will you assign to treatment, and how many will you assign to control? (To be clear,

your estimator here will be an OLS regression of Y_i on D_i, not the reweighted regression.)

2. We continue to use the RCT discussed above, but you no longer have any covariates X_i (and you can forget about the weights and costs). You may assume $P(D_i = 1) = 0.5$. Suppose Y_i is missing for some observations (but D_i is complete for all observations). This might happen because, for example, some participants in the RCT did not respond to the survey collecting the outcome data.

   (a) First assume Y_i are missing at random. Propose a simple test for this hypothesis. Does a regression of Y_i on D_i (using complete observations) estimate ATE if your assumption is correct?

   (b) If the Y_i are missing at random, does your test have the correct size? That is, does it reject at a rate $\alpha = 0.05$? Briefly explain.

   (c) If the null hypothesis is false (i.e. the Y_i are *not* missing at random), will your test necessarily reject as $N \to \infty$?

3. We now consider a gentle introduction to the selection model known as the Roy Model. We discussed this model briefly in class in the context of choosing fishing versus hunting. The key assumption in the Roy Model is that individuals select into treatment based on their gains (or expected gains) from treatment. This assumption comes naturally from a framework in which individuals maximize utility, but, to be clear, it is an assumption (i.e. there is no guarantee it holds in your real-world data).

   To fix ideas, assume a binary treatment D_i and potential outcomes $Y_i(0)$ and $Y_i(1)$. Let the treatment gain for individual i be equal to i's treatment effect, $\tau_i = Y_i(1) - Y_i(0)$. Assume i selects into treatment if and only if $\tau_i > 0$ (i.e. $D_I = \mathbf{1}(\tau_i > 0)$).

   For part (d) of this question you may take as known a result we refer to as the inverse Mills ratio — if $A \sim N(0,1)$, then $E[A|A > c] = \frac{\phi(c)}{1-\Phi(c)}$ (we will show this in lecture).

   (a) First assume $Y_i(0) = c$. Show that a regression of Y_i on D_i estimates TOT.

(b) * Now let $Y_i(0)$ and $Y_i(1)$ be independent with marginal Uniform(0,1) distributions. Analytically derive $E[Y(1)_i|D_i = 1]$ and $E[Y(0)_i|D_i = 0]$. Does a regression of Y_i on D_i estimate TOT?

(c) Write some Monte Carlo simulation code in Stata, R, or another package of your choice to confirm your answer to part (b) (or, if you skipped part (b), to determine the answer numerically). What is your intuition as to why the regression does or does not estimate TOT now?

(d) * Now let $Y_i(0)$ and $Y_i(1)$ be independent with marginal normal distributions with $\mu = 0$ and $\sigma^2 = 0.5$. Analytically derive $E[Y(1)_i|D_i = 1]$ and $E[Y(0)_i|D_i = 0]$. Does a regression of Y_i on D_i estimate TOT?

(e) Write Monte Carlo simulation code to confirm your answer to part (d) (or, if you skipped part (d), to determine the answer numerically). What happens if you generate an error ε_i that's normally distributed with $\mu = 0$ and $\sigma^2 = 1$ and add it to both $Y_i(0)$ and $Y_i(1)$ to create a positive correlation between the two. Does the regression estimate TOT, and if not, is it upwardly biased or downwardly biased?

(f) If there is selection bias in a regression of Y_i on D_i, is it positive or negative? How does the sign accord with your intuition about the general form of selection bias into, for example, getting a college degree (i.e. a case in which Y_i is later-life earnings and D_i is whether the individual got a college degree). What does this say about the most basic form of the Roy Model?