ECON 488, Pset #4: Estimation of TEs with Observational Data

Melissa Tartari

Premise

Objective In Pset #2 you used NSW experimental data for control and treated units to estimate the ATE of the offer of training on post-training earnings. You reported an estimate of ATE equal to 1,794 US$. In PSet#3 you used datasets created to mimic observational data and reported estimates of the ATE of the offer of training obtained from various regression-based methodologies e.g. "Difference in Differences". In this pset you continue to use oneof the datasets uses in PSet#3: nswpsid.csv.

Part 1: Stratification Matching on the PScore (37 p)

Objective Regression-adjutment approaches such as those in PSet #3 generate counterfactual earnings that rely on strong restrictions. For instance, PSet #3's approach called "Regression-Adjusted Treatment-Control Comparison" rests on the assumption that the conditional mean of post-treatment earnings is linear in the pre-treatment variables, that is, $E\left[re78_i|\mathbf{x}_i, D_i\right] = \mathbf{x}_i'\boldsymbol{\beta} + \alpha D_i$. Only if this is the case then $E\left[re78_i|\mathbf{x}_i, D_i = 1\right] - E\left[re78_i|\mathbf{x}_i, D_i = 0\right] = \alpha$ and the regression-adjusted comparison identifies the ATE of interest. Not surprisingly we would like to base our inference on counterfactuals that do not rely on restrictive assumptions about the form of $E\left[re78_i|\mathbf{x}_i, D_i = 1\right]$. An obvious way of achieving this goal is to compare treated and untreated individuals with the same exact value of $\mathbf{x}$. However, "exact matching on covariates" is not feasible when there are many pretreatment variables and/or the pretreatment variables take many values (this is the so called "curse of dimensionality problem"). In seminal work, the statisticians Rosenbaum and Rubin (1983) proposed *pscore*-matching as a method to reduce the bias in the estimation of TEs based on observational datasets. What makes this method particularly appealing is that it is implementable even when exact matching on pretreatment characteristics is not feasible. Here is the basic idea. As we know, in observational studies assignment of subjects to the treatment and control groups is not random, hence the estimation of the TE may be biased by the existence of confounding factors. Rosenbaum and Rubin's *pscore* matching is a way to "correct" the estimation of TEs controlling for the existence of these confounding factors. It is based on the idea that the bias is reduced (and ideally eliminated) when the comparison of outcomes is performed using treated and control subject who are *as similar as possible but not necessarily identical* in their pretreatment characteristics. Since matching subjects on a K-dimensional vector of pretreatment characteristics is typically unfeasible, this method proposes to summarize the pre-treatment characteristics of each subject by means of a single-index variable, called the *propensity score (pscore)*, and then match control and treatment units based on their pscores. The pscore is the probability that a subject with pretreatment characteristics $\mathbf{x}_i$ is assigned to be treated; formally, the pscore is $\Pr\left(D_i = 1|\mathbf{x}_i\right)$. Scholars have developed several matching algorithms based on the pscore. Here you are asked to code up your own matching algorithm to implement so called "strata-based matching on balancing scores". The idea of stratification is to divide the range of variation of the *pscore* into intervals such that within each interval there are both treatment and control units and they have, on average, the same *pscore*.

1. (6 p) In RCTs we typically know the pscore of each subject because the scientist who designed the experiment also chose the assignment rule. In observational studies the *pscores* of the subjects are typically unknown. Accordingly, the first step in any *pscore* matching approach requires us to estimate the value of the *pscore* for each subject in the sample. Broadly speaking, there are two approaches for doing so:

a) a parametric approach, and b) a non-parametric approach. Here you adopt a parametric approach. Start by recalling the *linear* probability model (LPM). The LPM provides a parametric specification of $\Pr(D_i = 1|\mathbf{x}_i)$ that is linear-in-parameters. Specifically, for a set of pre-determined characteristics $\mathbf{x}_i$ and a vector of parameters $\boldsymbol{\theta}$, we posit that:

$$\Pr(D_i = 1|\mathbf{x}_i; \boldsymbol{\theta}) = \mathbf{x}_i'\boldsymbol{\theta}. \tag{1}$$

Specification (1) implies (2) because $\Pr(D_i = 1|\mathbf{x}_i) = E[D_i|\mathbf{x}_i]$:

$$D_i = \mathbf{x}_i'\boldsymbol{\theta} + v_i \text{ with } E[v_i|\mathbf{x}_i] \overset{\text{by construction}}{=} 0. \tag{2}$$

One advantage of the LPM specification of the pscore is that its parameters can be estimated by OLS. Denote the OLS estimate of $\boldsymbol{\theta}$ by $\widehat{\boldsymbol{\theta}}$. With $\widehat{\boldsymbol{\theta}}$ in hands, we get predicted probabilities by calculating fitted values of the dependent variable, namely, $\widehat{p}_i = \mathbf{x}_i'\widehat{\boldsymbol{\theta}}$. Thus, $\widehat{p}_i$ is the LPM-based estimate of the pscore for subject i. One of the disadvantages of the LPM specification of the pscore is that $\widehat{p}_i$ may not lie in $(0, 1)$. Pscore-matching breaks down when $\widehat{p}_i$ is not in $(0, 1)$. To obviate to this problem it is common to employ a parametric specification of $\Pr(D_i = 1|\mathbf{x}_i)$ that guarantees that predicted probabilities fall in $(0, 1)$. For the same set of pre-determined characteristics $\mathbf{x}_i$, and a vector of parameters $\boldsymbol{\gamma}$, one such specification is:

$$\Pr(D_i = 1|\mathbf{x}_i; \boldsymbol{\gamma}) = e^{\mathbf{x}_i'\boldsymbol{\gamma}} / \left(1 + e^{\mathbf{x}_i'\boldsymbol{\gamma}}\right). \tag{3}$$

Remark that for any estimate $\widehat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$, $\widehat{p}_i = e^{\mathbf{x}_i'\widehat{\boldsymbol{\gamma}}} / \left(1 + e^{\mathbf{x}_i'\widehat{\boldsymbol{\gamma}}}\right)$ is guaranteed to be in $(0, 1)$. Remark also that there is no transformation of (3) that yields a linear-in-parameters specification. This means that we cannot use OLS to estimate $\boldsymbol{\gamma}$ in (3). Specification (3) is called a Logit model and we estimate $\boldsymbol{\gamma}$ by an approach called Maximum Likelihood Estimation (MLE).

(a) (3 p) Use the `nswpsid.csv` dataset. Estimate both the LPM and the Logit specifications of the *pscore*. For both specifications use the following variables as $\mathbf{x}_i$: {age, agesq, edu, edusq, married, nodegree, black, hisp, re74, re75, re74sq, re75sq, u74black} where e.g. `re74sq` is the square of `re74` and `u74black` is the interaction b/w `u74` and `black`.

(b) (3 p) Graphically compare $\widehat{p}_i$ between the two specifications and briefly comment.

2. (4 p) Use the Logit-based estimates obtained in question **1.b**. Let $\left(\underline{\widehat{p}}^c, \overline{\widehat{p}}^c\right)$ denote the smallest and largest $\widehat{p}_i$ among control units and $\left(\underline{\widehat{p}}^t, \overline{\widehat{p}}^t\right)$ the smallest and largest $\widehat{p}_i$ among treated units. Range $\left[\max\left\{\underline{\widehat{p}}^c, \underline{\widehat{p}}^t\right\}, \min\left\{\overline{\widehat{p}}^c, \overline{\widehat{p}}^t\right\}\right]$ is called the common support. Drop all sample units (control and treated alike) whose $\widehat{p}_i$ falls outside of the common support. Why do you think that we ask you to drop these observations? How many observations do you drop amongst controls and amongst treated subjects? Are you surprised?

3. (5 p) The attached Figure plots `re78` against the Logit-based $\widehat{p}_i$, separately for the treated and the control groups. Each panel also includes a fitted non-parametric regression of `re78` on $\widehat{p}_i$: think of this as a curve that smoothly connects *local averages* of `re78`, where by "local" we mean averages within small intervals of $\widehat{p}_i$ values.

(a) (2 p) Reproduce the plot. **Hint**: You can obtain the curves using a procedure called lowess.

(b) (3 p) Can you eyeball estimate**s** of the TE of the offer of traning from these figures? Explain.

4. (22 p) We should do better than eyeballing TEs. Luckily, you are now ready to implement stratification matching.

(a) (2 p) Generate a variable called `strata` that takes values from 1 to 10 corresponding to 10 equally spaced intervals with `strata=1` if $0 < \widehat{p}_i \leq 0.1$, `strata=2` if $0.1 < \widehat{p}_i \leq 0.2$, and so on. What is the distribution of control and treated observation across strata? Are all strata populated?

(b) (11 p) Is there a reason why you may want to test that the average of each pretreatment variable is the same for treated and control units within each stratum? Run the test and comment on your findings. **Hint:** Recall SUR estimation from PSet #2.

(c) (9 p) Let N_s^T denote the number of treated observations in stratum s with $s = 1, ..., 10$. Let $N^T = \sum_{s=1}^{10} N_s^T$. Let $\overline{re78}_s^{D=1}$ (respectively, $\overline{re78}_s^{D=0}$) denote average post-treatment earnings for treated (respectively, control) units in stratum s. Estimate the ATT of the offer of training using stratification matching estimator (4). Explain in plain words why this is an estimator of the ATT of the offer of training. **Hint:** The answer is 1,563 US$.

$$\widehat{ATT}^{strata} = \sum_{s=1}^{10} \left(\frac{N_s^T}{N^T} \right) \left(\overline{re78}_s^{D=1} - \overline{re78}_s^{D=0} \right). \tag{4}$$

Part 2: "Fancier" Matching on the PScore (18 p)

Objective In Part 2 you noticed that one of the pitfalls of the (naive) stratification method is that it discards observations in strata where either treated or control units are absent. This suggests an alternative way to match treated and control units, which consists of taking each treated unit and searching for the control unit with the closest *pscore* i.e. the unit's nearest control neighbor. Once each treated unit is matched with a control unit, the difference between the outcomes of each treated unit and the outcome of the matched control unit is computed. The Nearest-Neighbor Matching (NM) estimate of ATT is obtained by averaging these difference. While in Stratification Matching there may be treated units that are discarded because no control is available in their stratum, in NM matching all treated units find a match. However, some of these matches may be fairly poor because for some treated units the nearest neighbor may have a very different *pscore*; nevertheless, he contributes to the estimation of the TE. The Radius Matching (RM) and Kernel Matching (KM) methods offer a solution to this problem. With RM matching, each treated unit is matched only with the control units whose pscore falls into a predefined neighborhood of the *pscore* of the treated unit. If the dimension of the neighborhood (the radius) is set very small, it is possible that some treated units are not matched because the neighborhood does not contain control units. On the other hand the smaller the size of the neighborhood, the better the qualify of the resulting matches. With KM each treated unit is matched with a weighted average of all control units with weights that are inversely proportional to the distance between the *pscore* of treated unit and control units. Clearly, these 3 methods reach different points on the frontier of the trade-off between qualify and quantity of the matches, and none of them is a priori superior. In the questions below you first look closely at the estimators produced by these methods, then you use them to obtain estimates of the ATT of the offer of training. Always use the `nswpsid.csv` dataset and the Logit-based estimates of the pscore from Part 1.

Let C denote the collection of control units and T denote the collection of treated units. Let N^C denote the number of units in C and N^T denote the number of units in T. Let $C(i)$ denote treated unit i's matching set, that is, the set of control units matched to i. Let N_i^C denote the number of units in $C(i)$. Also, denote the estimate of i's *pscore* by $\widehat{p}_i$. NM, RM and KM define i's matching set as, respectively:

$$C^{NM}(i) = \{j \in C : |\widehat{p}_j - \widehat{p}_i| \leq |\widehat{p}_{j'} - \widehat{p}_i| \ \ \forall j' \in C\}, \tag{5}$$

$$C^{RM}(i) = \{j \in C : |\widehat{p}_j - \widehat{p}_i| < r\} \text{ (given a choice of radius } r), \tag{6}$$

$$C^{KM}(i) = C. \tag{7}$$

Given (5)-(7), the NM, RM, and KM estimators of the ATT have a common form:

$$\widehat{ATT}^m = \frac{1}{N^T} \sum_{i \in T} \left[y_i - \sum_{j \in C^m(i)} w_{ij}^m y_j \right], \ m \in \{NM, RM, KM\}, \tag{8}$$

where $C^m(i)$ is defined in (5)-(7) and w_{ij}^m is a matching method-specific weight:[1]

$$w_{ij}^m = \begin{cases} \frac{1}{N_i^C} \text{ if } j \in C^m(i) \text{ and zero otherwise} & \text{for } m \in \{NM, RM\} \\ K\left(\frac{\widehat{p}_j - \widehat{p}_i}{h}\right) / \sum_{k \in C} K\left(\frac{\widehat{p}_k - \widehat{p}_i}{h}\right) & \text{for } m = KM \text{ (given kernel function } K(.)) \end{cases}. \quad (9)$$

1. (2 p) Denote the generic outcome variable by y. Show that under unconditional RA, $\widehat{ATT}^m = \overline{y}^{D=1} - \overline{y}^{D=0}$, $\forall m \in \{NM, RM, KM\}$. **Hint:** Assign treatment by the flip of a balanced coin, that is, work with $\widehat{p}_i = 0.5$ for all i.

2. (4 p) Show that $\widehat{ATT}^m$ in (8) rewrites as (10):

$$\begin{aligned} \widehat{ATT}^m &= \overline{y}^{D=1} - \overline{y}_m^{D=0}, \, m \in \{NM, RM, KM\}, \text{ with} \\ \overline{y}_m^{D=0} &= \frac{1}{N^C} \sum_{j \in C} \pi_j^m y_j. \end{aligned} \quad (10)$$

   Use (10) to interpret in plain English the essence of each matching approach. **Hint:** Determine the expression for π_j^m in (10) for all $m \in \{NM, RM, KM\}$.

3. (10 p) Use the R package Matching to estimate ATT using NM. Report the estimate in Table (11). Observe that the table also reports ATT estimates based on 3 other pscore matching approaches (you are not asked to implement them). The "Adaptive Stratification" approach is a refined version of the "Naive Stratification" approach that you implemented in Part 1: strata are created iteratively (as opposed to being fixed ahead of time) so as to ensure balance within each stratum. **Hint:** Use the function Match with the following options: `M=1`, `estimand="ATT"`, and `CommonSupport = T`.

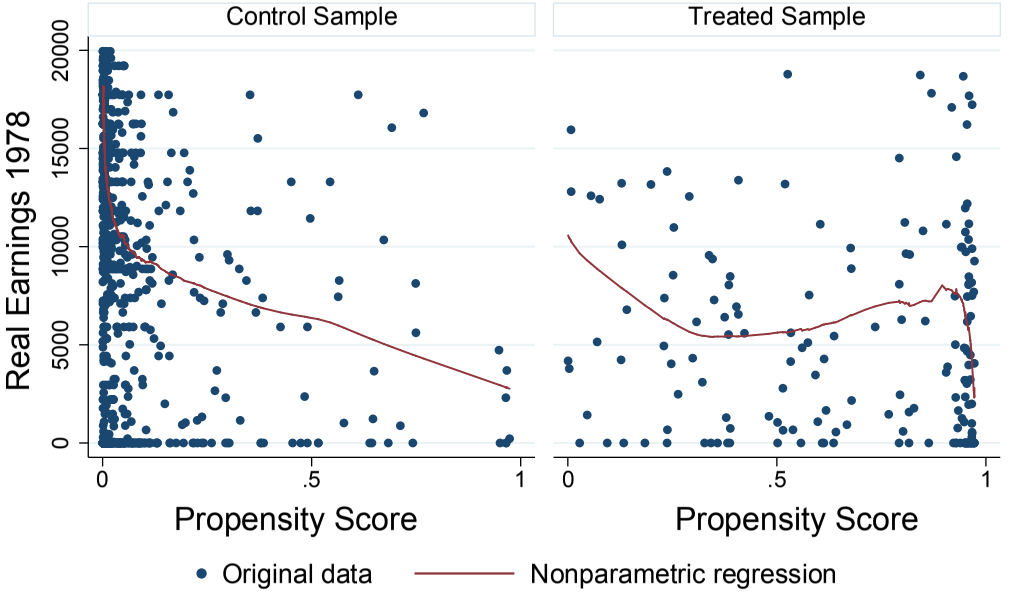
Matching Estimator	**Parameter**	**Estimate**	**SE**
Nearest Neighbor (NM)	ATT		
Radius (RM) with $r = 0.0001$	ATT	-5546.139	4614.773
(Adaptive) Stratification	ATT	2208.603	855.168
Gaussian Kernel (KM)	ATT	1537.947	861.329

(11)

4. (2 p) Take a break, clear your mind, then come back and list the takeaways from PSet #2 and #3. Finally, pat yourself on the back for all that you have accomplished :-).

[1] A kernel function is a real-valued function with the following properties: 1) $\int \psi K(\psi) \, d\psi = 0$ i.e $K(.)$ is symmetric around zero; 2) $\int K(\psi) \, d\psi = 1$; 3) $\int \psi^2 K(\psi) \, d\psi \neq 0$ and finite; and 4) $\int K^2(\psi) \, d\psi < \infty$. An example of a kernel function is the standard normal (aka Gaussian) pdf: $K(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right)$.

Post-treatment Earnings against Propensity Score



Data Source: NSW-PSID1.