ECON 488, Pset #3: Estimation of TEs with Observational Data

Melissa Tartari

Part 1: A First Look at Observational Data (10 p)

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$. Alas, often variation in the cause/treatment is observational in nature, as opposed to resulting from an RCT. Accordingly, in this pset you apply several methods to estimate the effect of the offer of training on post-training earnings using observational data. Here you have a look at two datasets that were put together to mimic observational data.

Consider the two files nswcps.csv and nswpsid.csv. Each file contains a dataset. Each dataset combines two samples: 1) the treated sample from the Dehajia and Wahba's NSW data (i.e. 185 males who were offered NSW training in 1976-1977)[1]; and 2) a sample extracted from a large survey: a) for nswcps.csv this is the Current Population Survey (CPS); b) for nswpsid.csv this is the Panel Study of Income Dynamics (PSID). The samples in 2) contain data on a comparison /control group, that is, on subjects who (as far as we know) did not receive the NSW offer of training.[2] Specifically, the PSID sample consists of 2,490 male household heads under the age of 55 who are not retired (Dehajia and Wahba call this sample the PSID-1 sample); and, the CPS sample consists of 15,992 male household heads under the age of 55 who are not retired (Dehajia and Wahba call this sample the CPS-1 sample). The data file nswcps.csv (respectively, nswpsid.csv) contains the treated individuals (from NSW-treated) along with the PSID (respectively, CPS) comparison individuals. The treatment indicator variable `treat` equals 1 for individuals in the NSW-treated sample and zero for the PSID (respectively, CPS) comparison individuals. Table (1) reports summary statistics. You filled the 3rd and 4th columns of Table (1) in PSet #2.

Variable	**Definition**	**NSW**		**PSID-1**	**CPS-1**
		Treated	**Control**	**Control**	**Control**
age	Age in years	25.82	25.05		
edu	Education in years	10.35	10.09		
nodegree	1 if edu<12	0.71	0.83		
black	1 if race is black	0.84	0.83		
hisp	1 if Hispanic	0.06	0.11		
married	1 if married	0.19	0.15		
u74	1 if unemployed in '74	0.71	0.75		
u75	1 if unemployed in ‘75	0.60	0.68		
re74	Real earnings in '74 (in '82 $)	2,096	2,107		
re75	Real earnings in '75 (in '82 $)	1,532	1,267		
re78	Real earnings in '78 (in '82 $)	6,349	4,555		
treat	1 if offered training	1.00	0.00	0.00	0.00
Sample Size (N)		185	260	2,490	15,992

(1)

[1] Dehejia and Wahba (1999) Causal Effects in Nonexperimental Studies: reevaluating the Evaluation of Training Programs, *JASA*, pp. 1053-1062. Dehejia and Wahba (2002) Propensity-score Matching Methods for Nonexperimental Causal Studies, *ReStat*, pp. 151-161.

[2] When working with observational data the untreated subsample is more properly called a comparison group. Nevertheless it is common to use the terms *control* and *comparison* interchangeably, irrespective of whether the variation in the treatment indicator is induced by RA or not.

1. (2 p) Fill the 5th column of Table (1) using the data in nswpsid.csv with `treat=0`.
2. (2 p) Fill the 6th column of Table (1) using the data in nswcps.csv with `treat=0`.
3. (2 p) Briefly comment on the completed Table (1). **Hint**: Are the PSID-1 and CPS-1 samples "good" control groups? Explain.
4. (4) Why do you think that Dehajia and Wahba constructed their "observational databases" by bringing together the treated sample from NSW and a sample of individuals drawn from either the PSID or the CPS data? **Hint:** Both the PSID and the CPS surveys include information on whether an individual enrolled in a training course during the previous 12 months. Accordingly, Dehajia and Wahba could have exploited exclusively observational variation in whether an individual enrolled in a training program and not used at all the NSW data. Why do you think that they chose not to follow this approach?

Part 2: Regression-based Estimation of TEs (35 p)

Objective Here you use the nswpsid.csv dataset to estimate the ATE of the offer of training by means of 4 regression-based approaches which we reference as specifications (2) through (6). We use: 1) $re78_i$ to represent the variable `re78`; 2) D_i to represent the variable `treat`; 3) $\mathbf{x}_i$ to represent a $K \times 1$ vector of regression covariates; 3) $D78_t$ to represent an indicator variable that equals 1 in the post-treatment year (1978) and zero otherwise. The subscript i denotes a specific individual. The subscript t denotes a specific calendar year. Table (7) lists the 4 specifications. You will complete columns 2 through 4 with your estimation results. Specifically, you will insert in column 2 the reference number of the specification along with the TE-relevant regression coefficient; and, you will insert in columns 3 and 4 the estimate and the standard error (SE) of the TE-relevant regression coefficient.

$$
\begin{aligned}
re78_i &= \phi + \alpha D_i + u_i,\ i = 1, ..., 2675, &(2)\\
re78_i &= \phi + \alpha D_i + \mathbf{x}_i'\boldsymbol{\beta} + u_i,\ i = 1, ..., 2675, &(3)\\
earns_{it} &= \phi + \alpha D78_t + u_{it},\ i = 1, ..., 185,\ t = 75, 78 &(4)\\
earns_{it} &= \phi + \delta D78_t + \gamma D_i + \alpha D78_t \times D_i + u_{it},\ i = 1, ..., 2675,\ t = 75, 78 &(5)\\
earns_{it} &= \phi + \delta D78_t + \gamma D_i + \alpha D78_t \times D_i + \mathbf{x}_i'\boldsymbol{\beta} + u_{it},\ i = 1, ..., 2675,\ t = 75, 78, &(6)
\end{aligned}
$$

Estimator	**Parameter & Specification**	**Estimate**	**SE**
Treatment-Control Comparison			
Regression-Adjusted Treatment-Control Comparison			
Before-after Comparison			
Differences-in-Differences			

(7)

1. (5 p) Use the appropriate specification among those listed in Table (7) to obtain the Treatment-Control Comparison Estimatorof the ATE of the offer to training. As the name suggests, this estimator is simply the difference between the average post-training earnings of the treated and of the control individuals, namely, $\widehat{\alpha} = (\overline{re78}^{D=1} - \overline{re78}^{D=0})$. Compute heteroschedasticity-robust SEs. Report your estimates in Table (7). Explain why this estimator may not be an unbiased/consistent estimator of the ATE of interest.
2. (5 p) Use the appropriate specification among those listed in Table (7) to obtain the Regression-Adjusted Treatment-Control Comparison Estimator of the ATE of the offer of training. Do so by adding the following pre-treatment characteristics as regression covariates, in linear form: `age`, `agesq`, `edu`, `nodegree`, `black`, `hisp`, `re74`, and `re75` (i.e. $K = 8$). Compute heteroschedasticity-robust SEs. Report your estimates in Table (7).

3. (5 p) Use the appropriate specification among those listed in Table (7) to obtain the Before-After (BA) Estimator of the ATE of the offer of training. As the name suggests, the BA estimator is the difference between average post-treatment and average pre-treatment earnings using only the 185 males in the treatment sample, namely, $\widehat{\alpha} = \left(\overline{earn}^{D=1}_{t=78} - \overline{earns}^{D=1}_{t=75}\right)$. Compute heteroschedasticity-robust SEs. Report your estimates in Table (7). Explain why the BA-estimator may be "misleading", that is, why it may not be a consistent estimator of the ATE (nor of the ATT) of interest.

4. (10 p) Here you try to overcome the possible flaws of the BA Estimator by using an estimator that "mixes" before-after comparisons and matching. We will formally review this approach at a later time, for now you implement it and start thinking about it.

   (a) (5 p) Use specification (5) in Table (7) to obtain the so called Differences-in-Differences Estimator (DD) of the ATE of the offer of training. Report your estimate in Table (7). Compute heteroschedasticity-robust SEs. **Hint:** You first need to reshape the data so that there are two separate years of data (1975 and 1978) for each i. We named the resulting earning variable `earns`. As a check, verify that the DD regression uses $N = 5,350$ observations.

   (b) (2 p) Verify that $\widehat{\alpha} = \left(\overline{earn}^{D=1}_{t=78} - \overline{earn}^{D=1}_{t=75}\right) - \left(\overline{earn}^{D=0}_{t=78} - \overline{earn}^{D=0}_{t=75}\right)$.

   (c) (3 p) Intuitively explain why the DD estimator improves on the BA estimator**.** **Hint:** In what sense does the DD estimator attempts to control for both overt (due to observable confounders) and hidden bias (due to unobservable confounders)?

5. (5 p) Use specification (6) to obtain the DD Estimator of the ATE of the offer of training. Compare to question **4.b** and comment.

6. (2 p) Could you obtain the DD estimator with repeated cross-sectional data? Explain.

7. (3 p) Table (7) is by now completely filled. Take stock.