ECON 488, PSet #2: Causal Inference with Data From a RCT

Melissa Tartari

Premise: A First Look at The NSW Experiment (10 p)

Objective In this Pset you use experimental data from the National Supported Work (NSW) demonstration project to estimate the effect of (the offer of) training on earnings. The NSW was conducted in the 1970s with the goal of measuring the impact of on-the-job training on earnings by means of a RCT that assigned some individuals to receive training (a treatment or experimental group) and others to receive no training (a control group).

The NSW project was designed as a transitional, subsidized work experience program for people with longstanding "employment problems". Eligible applicants were randomly assigned (by the "flip of a coin") either to the experimental group or to control group. In this pset you work with Dehejia and Wahba (1999, 2002)'s extract of the NSW original data.[1] In this data extract, the treated sample contains 185 males who were offered employment *cum* training in 1976-1977. The control sample contains 260 males. The treatment is therefore the offer of employment *cum* training.

1. (10 p) Merge the NSW data for control and treatment units: nswre74_control.csv and nswre74_treated.csv. Fill columns 3 and 4 of Table (1) by inserting the sample averages of the variables listed in column 1. Variable `re78` is the (realized) outcome variable, variable `treat` is the indicator of treatment status, and all the other variables are *predetermined*, that is, they capture characteristics of the units that are determined at or before the time when the units are assigned to the treated or control groups. Some of these variables are background characteristics, others capture features of a subject's pre-RCT labor market experience.

Variable	**Description**	**Sample Averages**	
		Treated	**Control**
age	Age in years		
edu	Education in years		
nodegree	=1 if edu<12, =0 otherwise		
black	=1 if race is black, =0 otherwise		
hisp	=1 if Hispanic, =0 otherwise		
married	=1 if married, =0 otherwise		
u74	=1 if unemployed in '74, =0 otherwise		
u75	=1 if unemployed in ‘75, =0 otherwise		
re74	Real earnings in '74 (in '82 $)		
re75	Real earnings in '75 (in '82 $)		
re78	Real earnings in '78 (in '82 $)		
treat	=1 if offered training, =0 otherwise	1.00	0.00
Sample Size		185	260

(1)

[1] Dehejia and Wahba (1999) Causal Effects in Nonexperimental Studies: reevaluating the Evaluation of Training Programs, *JASA*, 1053-1062 and Dehejia and Wahba (2002) Propensity-score Matching Methods for Nonexperimental Causal Studies, *ReStat*, 151-161. The original data (not necessary for this pset) is available at the ICPSR page.

Part 1: Testing Balance (55 p)

Objective You use the NSW data to ascertain whether subjects were indeed randomly assigned to treatment.

Proper randomization ensures that the treatment is assigned randomly (by the "flip of a coin"). As a consequence, it balances all the determinants of the outcome variable, both observed and unobserved. That is, if the randomization is done properly there should be no systematic differences in pretreatment characteristics between the units in the control and in the treatment groups. It is therefore *always* a good idea to check whether pretreatment variables are indeed *balanced* across the two groups. After all, even if randomization took place it may not have been done correctly; and, if we find that the control and treatment groups are systematically different in their pretreatment characteristics, we may suspect that randomization failed and that treatment and control units differ also in their unobserved characteristics. Such difference (or *imbalances*) raise concerns because they may lead us to conclude that the difference in average earnings across the treated and control groups is a biased estimator of the causal effect of the treatment. Below you examine two ways of gauging the "degree of randomness" of treatment assignment. The first approach compares the average and standard deviations (SD) of observed pretreatment variables across the two groups (questions **1.** and **2.**). The second approach ascertains whether treatment status is explained by the observed pretreatment variables (question **3.**).

1. (5 p) Consider the 10 pre-treatment variables in Table (1). With reference to *each* variable, test covariate balance. Do so by testing that sample averages are not different across control and treatment groups. Do so by running 10 separate simple linear regressions. Comment on the results.

2. (40 p) Checking covariate balance as done in question **1.** suffers of the so called "multiple comparisons" or "multiple testing" problem which occurs when one considers a set of statistical inferences simultaneously. The problem emerges because as more variables are compared, it becomes more likely that the treatment and control groups appear to differ on at least one attribute *by random chance alone*. To deal with this problem we can use an estimation methodology called SUR estimation and then use the estimates to *jointly* test that all covariates (the pre-treatment variables in Table (1)) are balanced. SUR stands for seemingly unrelated regression and it is a special case of feasible Generalized Least Squares (GLS) estimation. The basic idea is as follows. We specify an equation for each of the 10 pre-treatment variable as a function of a constant and the treatment indicator. That is, each pre-treatment variable is a dependent variable in its own equation and the treatment indicator is the only regression covariate present in each equation. Then, instead of estimating the coefficients *equation-by-equation* by OLS (and done in question **1.**) we estimate the coefficients present in all the equations jointly accounting for the fact that the unobservables may be correlated across equations within an individuals (we continue to assume that they are uncorrelated across units). After the estimation is completed we can use standard testing to test the *joint* hypothesis that the slope coefficients in all the equations of the SUR system are zero. This joint test is a test of covariate balance that does not suffer of the "multiple testing" problem.

   (a) (25 p) Use the R package `systemfit` (link) to estimate the SUR system with the NSW data. Are the estimated coefficients and their SEs different from those you obtained in question **1.**? Comment. **Hint:** In the following two situations there is no efficiency payoff to GLS versus OLS: 1) when the unobservables are uncorrelated across equations within an individual; and 2) when the equations in the SUR system have identical covariates.

   (b) (15 p) Test *joint* covariate balance. Comment on your findings. **Hint:** You want to test the joint hypothesis that the coefficients of `treat` are zero in all the equations of the SUR system.

3. (10 p) Test that the pretreatment variables do not predict treatment assignment. Report your results and explain in plain English why this is a test that a "sensible" scientist may want to carry out when working with data from a RCT. **Hint:** Use a linear regression model and recall the so called overall test of significance of the regression.

Part 2: The ATE of the Offer of Training (35 p)

Objective You use the NSW data to obtain an estimate of the causal effect of being offered on-the-job training on post-intervention earnings. The outcome variable of interest is `re78`.

1. (5 p) Consider the ATE of receiving a training offer. Here you obtain an estimate of this estimand in two ways. Do describe your findings.

   (a) (2 p) By computing group-specific sample averages of post-treatment earnings and taking their difference (treatment average minus control average).

   (b) (3 p) Specification #0: By estimating the coefficients of a linear regression model.

2. (20 p) Estimate by OLS the coefficients of the following 3 specifications. Report the estimate of the ATE of receiving a training offer for all specifications.

   (a) (3 p) Specification #1: In light of the imbalance in educational attainment (as documented in Part 1 of this pset), modify the regression specification in question **1.b** by adding `nodegree` and `edu` as regression covariates in linear form. **Side note**: This regression-based approach to account for the presence of observable confounders (imbalance in pre-treatment variables) is typically called the regression adjument approach.

   (b) (3 p) Specification #2: Add to specification #1 the other pretreatment variables as covariates in linear form (this yields a total of 11 regression covariates).

   (c) (3 p) Specification #3: Add to specification #2 the interaction between the treatment indicator and the variable `age` in deviation from its sample mean (this yields a total of 12 regression covariates).

   (d) (5 p) Are there reasons to add pre-treatment variables as covariates when covariates are balanced across treatment and control groups? If so, what are they? Comment on the estimation results from specifications #1 and #2.

   (e) (1 p) With reference to specification #2: Do you think that it is problematic to use lagged / past values of the dependent variable as regression covariates? Explain.

   (f) (5 p) Are there reasons to add interactions between pre-treatment variables and the treatment indicator? If so, what are they? With reference to specification #3 test the following two hypothesis: i) the ATE of treatment is zero; ii) the effect of the treatment does not vary by the age of the subject. **Note:** If any additional assumption needs to be spelled out for your to carry out i) and ii) please do so.

3. (5 p) Brainstorm on the possible mechanisms for the estimated ATE of being offered training. **Hint:** Can you think of the possible pathways through which on-the-job training may cause an increase in earnings?

4. (5 p) In this question you brainstorm on why the ATE of the offer of training may be different from the ATE of training. Termonology-wise, the ATE of the offer of training is a so called Intent to Treat Effect (ITT). To think about the relationship b/w these two treatment effects in a systematic way you consider the following setup. Let Z_i denote a binary variable that takes the value 1 if individual i is offered traning, and zero otherwise. Let D_i be a binary variable that takes the value 1 if individual i undergoes training, and zero otherwise. Thus, for instance, and invidual how is offered training but does not take it up has $(Z_i, D_i) = (1, 0)$. Assume that when an individual receives an offer of traning he flips a coin, if it comes up Head he enrols in the training program while if it comes up Tail he does not. Assume also that individuals not offered training cannot take it up. Finally, assume that the offer of training does not *per se* affect future earnings e.g. it is not the case that by virtue of receiving a training offer an individual becomes more optimistic about his future labor market prospects and searches for a well paying job harder than he would have had he not received the offer. Each individual has two pairs of potential outcomes, one associated with treatment defined as being offered training and the other associated with the treatment defined as undergoing training. Specifically, let (y_{1i}, y_{0i}) denote potential re78 earnings *with* training and *without* training. Similarly, let $\left(y_{1i}^{off}, y_{0i}^{off}\right)$ denote

potential re78 earnings *with, and respectively without, the offer of training in hands.* In questions **1.** thorough **3.** your goal was to estimate the ATE of being offered training, let us denote it by $ATE^{off} \equiv E\left[y_{1i}^{off} - y_{0i}^{off}\right]$. In this question you are asked to compare ATE^{off} to $ATE \equiv E\left[y_{1i} - y_{0i}\right]$ (which is the ITT effect) and show/discuss why they may be different. To answer this question start by relating the two pairs of potential outcomes, specifically, express $\left(y_{1i}^{off}, y_{0i}^{off}\right)$ in terms of (y_{1i}, y_{0i}). Then go from there. **Hint #1**: You have all the information necessary for obtaining an exact relationship b/w the ATE of being offered training and the ATE of training. **Hint #2**: The important lession to take from this question is that the ATE of the offer of training is typically different from the ATE of the training even in the absence of self-selection into the training program.