ECMA 31360, PSet 4: Causal Inference with Observational Data

Melissa Tartari, University of Chicago

Objective: In Pset3 you estimated the ATE of the offer of training on post-training earnings using NSW experimental data. The Treatment-Control difference in sample averages indicates that the offer of training causes an additional $1,794 in terms of 1978 earnings. Variation in the cause/treatment is often observational in nature, instead of resulting from an RCT. In this Pset you utilize methods developed to estimate the effect of the offer of training using "observational data" and apply them to two datasets that Dehajia and Wahba constructed to mimic observational data.

Background: Consider the 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 offered NSW training in 1976-1977)[1]; and 2) a sample extracted from a large survey: a) in `nswcps.csv`, such sample is the Current Population Survey (CPS); b) in `nswpsid.csv`, such sample is the Panel Study of Income Dynamics (PSID). The samples in 2) contain data on a comparison group, that is, on subjects who (as far as we know) did not receive the NSW offer of training.[2] Specifically, the PSID sample (called **PSID-1**) consists of 2,490 male household heads under the age of 55 who are not retired; and, the CPS sample (called **CPS-1**) consists of 15,992 male household heads under the age of 55 who are not retired. The 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.

Part 1: Describe the Data (10 p)

Q1. (4 p) Fill Table 1's columns 5 and 6 using, respectively, the data in `nswpsid.csv` and in `nswcps.csv`. **Notes**: You want to limit attention to observations with `treat=0`. You filled columns 3 and 4 in PSet 3.

Variable	Definition	NSW		PSID-1	CPS-1
		Treated	Control	Control	Control
[1]	[2]	[3]	[4]	[5]	[6]
`age`	Age in years	25.82	25.05		
`edu`	Education in years	10.35	10.09		
`nodegree`	1 if education < 12	0.71	0.83		
`black`	1 if 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 received offer of training	1	0		
Sample Size		185	260	2,490	15,992

Table 1: Sample averages for the NSW data (treated and control groups), PSID-1 data, and CPI-1 data.

Q2. (4 p) Briefly comment on the completed Table 1. **Hint**: Are the PSID-1 and CPS-1 samples "good" control groups?

Q3. (2 p) Why do you think that Dehajia and Wahba constructed their "observational datasets" by pulling together the treated sample from NSW and a sample of individuals drawn from either the PSID or the CPS data? **Hint:** Both PSID and CPS include information on whether an individual enrolled in a training course during the previous 12 months. Thus, Dehajia and Wahba could have exploited exclusively observational variation in whether an individual enrolled in a training program. Why do you think that they chose not to follow this approach?

[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 sample 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.

Part 2: Regression-based Estimation of TEs (90 p)

Objective: You use the `nswpsid.csv` dataset to estimate the treatment effect (TE) of the offer of training via regression-based approaches associated with the following three specifications of the outcome equation:

$$re78_i = \alpha + \rho D_i + u_i,\ i = 1, ..., 2675, \tag{1}$$
$$re78_i = \alpha + \rho D_i + \mathbf{x}_i'\beta + u_i,\ i = 1, ..., 2675, \tag{2}$$
$$re78_i = \rho D_i + g(\mathbf{x}_i) + u_i,\ i = 1, ..., 2675, \tag{3}$$

Subscript i denotes an individual. Also: 1) $re78_i$ represents the data field `re78`; 2) D_i represents the data field `treat`; 3) $\mathbf{x}_i$ represents a $K \times 1$ vector of observed pre-determined variables (OPVs); and, 4) $g(\cdot)$ is an unknown and possibly non-linear function (i.e., a generalization of $\alpha + \beta'\mathbf{x}_i$). Table 2's column [1] references the regression specification. Column [2] gives the name of the approach. Column [3] indicates the regression coefficient of interest. You complete columns [4] and [5] with the estimate of the regression coefficient and its standard error (SE).

Reference Model	**Name of the Estimation Approach**	**Parameter of Interest**	**Estimate**	**SE**
[1]	[2]	[3]	[4]	[5]
expression (1)	Treatment-Control Comparison (TCC)	ρ		
expression (2)	Regression-Adjusted Treatment-Control Comparison (Adj. TCC)	ρ		
expression (3)	Double Machine Learning (DML)	ρ		

Table 2: Treatment Effect Estimates Based on Three Regression-Based Approaches Applied to Observational Data.

Background: Heteroschedasticity-Robust Standard Errors. In econometrics, the conditional variance is called the skedastic function. Homoschedasticity obtains when the unobservable in a regression specification has the same conditional variance for all values of the explanatory variable(s). For example, in specification (1) there is only one explanatory variable D_i, and it takes only two values, therefore homoshedasticity obtains if $Var[u_i|D_i = 1] = Var[u_i|D_i = 0]$. If this assumption fails, we say that the model exhibits heteroschedasticity. As a rule, we are better off reporting heteroschedasticity-robust SEs, i.e., SEs computed in a way that allows for heteroschedasticity, because they are valid whether or not homoschedasticity holds.

Background: "Partialling-Out" Interpretation of OLS in a MLRM. Simple linear-in-parameter regression models (SLRM) are of the form

$$y_i = \alpha + \beta x_i + u_i \tag{4}$$

where x_i is a single regression covariate. MLRMs are of the form:

$$y_i = \alpha + \beta_1 x_{1,i} + \cdots + \beta_K x_{K,i} + u_i \text{ with } K > 1. \tag{5}$$

In PSet1 you derived the form of the OLS estimator of the slope coefficient in SLRM (4), namely

$$\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \underbrace{=}_{\text{also equivalent to}} \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{6}$$

Note that if you regress x_i on a constant, the fitted value is $\hat{x}_i = \bar{x}$, thus the regression residuals are $\hat{r}_i \equiv x_i - \hat{x}_i = x_i - \bar{x}$. Similarly, if you regress y_i on a constant, the fitted value is $\hat{y}_i = \bar{y}$, thus the regression residuals are $\hat{v}_i \equiv y_i - \hat{y}_i = y_i - \bar{y}$. Accordingly, we can rewrite $\hat{\beta}$ in expression (6) as:

$$\hat{\beta} = \frac{\sum_{i=1}^n \hat{r}_i y_i}{\sum_{i=1}^n \hat{r}_i^2} \underbrace{=}_{\text{also equivalent to}} \frac{\sum_{i=1}^n \hat{r}_i \hat{v}_i}{\sum_{i=1}^n \hat{r}_i^2}. \tag{7}$$

Similar steps yield a very compact representation of the OLS estimator of the slope coefficient**s** in a MLRM. For example, the OLS estimator of β_1 in MLRM (5) can be written as:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \hat{r}_{1,i} y_i}{\sum_{i=1}^n \hat{r}_{1,i}^2} \underbrace{=}_{\text{also equivalent to}} \frac{\sum_{i=1}^n \hat{r}_{1,i} \hat{v}_{1,i}}{\sum_{i=1}^n \hat{r}_{1,i}^2}, \tag{8}$$

where $\hat{r}_{1,i}$ denotes the residuals from regressing $x_{1,i}$ on a constant and all remaining regression covariates, i.e., $\{x_{2,i}, \ldots, x_{K,i}\}$ and $\hat{v}_{1,i}$ denotes the residuals from regressing y_i on a constant and all remaining regression covariates, i.e., $\{x_{2,i}, \ldots, x_{K,i}\}$. Similar expressions hold for $\hat{\beta}_2$, $\hat{\beta}_3$, etc.

Q4. (30 p) These questions pertain to the specification in expression (1) thus you obtain the Treatment-Control Comparison (TCC) Estimator of the treatment effect of the offer of training.

(a) (8 p) Estimate ρ. **Programming Guidance:** Use `stats::lm( )`. Say that your linear model is `m1 <- lm(re78 ~ treat, data = df)`. View the SEs of estimator $\hat{\rho}$ by using `summary(m1)$coefficients["treat", c("Estimate", "Std. Error"]`. View all SEs by using `lmtest::coeftest(m1, vcov. = vcov(m1))` which runs t-tests for each of the coefficients using the variance-covariance matrix estimated assuming homoschedasticity. Package `lmtest` allows you to perform z and t tests on estimated coefficients from, among others, method `lm( )`. It returns a coefficient matrix with columns containing the estimates, associated SEs, test statistics, and p-values.

(b) (10 p) Compute heteroschedasticity-robust SEs. **Programming Guidance:** There are multiple R packages to estimate the variance-covariance matrix of $(\hat{\alpha}, \hat{\rho})$ under general heteroschedasticity. Here are two ways. Option 1: Use `sandwich::vcovHC(m1, type = ‘‘HC0")` from package `sandwich`. Option 2: Use `car::hccm(m1, type = ‘‘hc0")` from package `car`. In both cases, the argument `type = "hc0"` (or `"HC0"`) tells R that you want to use the variance covariance matrix estimated using White's (1980) estimator, often referred to as HCE (heteroscedasticity-consistent estimator).[3] Display robust SEs by typing, e.g., `lmtest::coeftest(m1, vcov. = sandwich::vcovHC(m1, type = "HC0"))`.

(c) (2 p) Verify that $\hat{\rho}$ in **Q4a** equals $(\overline{re78}^{D=1} - \overline{re78}^{D=0})$, i.e., the difference between the average post-training earnings of the treated and of the control individuals. This fact explains the name of the estimator, and is consistent with what you derived in previous Psets.

(d) (10 p) Intuitively explain why the TCC approach may not deliver a credible estimate of the average effect of the treatment of interest. **Hint**: Use the result in **Q4c** to think about what this approach uses to proxy for the missing data, i.e., for the control units' mean of the potential outcome w/ treatment, and for the treated units' mean of the potential outcome w/out treatment.

Q5. (20 p) These questions pertain to the specification in expression (2) thus you obtain the Regression-Adjusted Treatment-Control Comparison (Adj. TCC) Estimator of the treatment effect of the offer of training.

(a) (10 p) Add to the model estimated in **Q4** the following OPVs as regression covariates: `age`, `agesq`, `edu`, `nodegree`, `black`, `hisp`, `re74`, and `re75`. Report $\hat{\rho}$ and its heteroschedasticity-robust SE. **Programming Guidance:** Add column `agesq` (`age` squared) to your dataframe using, e.g., `dplyr::mutate( )`.

(b) (10 p) Intuitively explain why the Adj. TCC approach may be regarded as an improvement over the TCC approach when it comes to credible identification/estimation of average treatment effects.

Q6. (20 p) Consider again the specification in expression (2) estimated in **Q5**. Here you implement two procedures, as detailed below, to verify the "partialling-out" interpretation of OLS coefficients in MLRM.

(a) (8 p) Procedure A:

i. (4 p) First Stage: Regress `treat` on a constant and the OPVs listed in **Q5a**; obtain the residuals. **Programming Guidance:** If you run `s1 <- lm(treat ~ x1 + x2, data = dt)`, retrieve the residuals as `s1$residuals`.

ii. (4 p) Second Stage: Regress `re78` on a constant and the residuals from **Q6(a)i**.

(b) (8 p) Procedure B:

i. (0 p) First Stage: Same as **Q6(a)i**.

ii. (4 p) First Stage: Regress `re78` on a constant and the OPVs listed in **Q5a**; obtain the residuals.

iii. (4 p) Second Stage: Regress the residuals from **Q6(b)ii** on the residuals from **Q6(b)i**.

(c) (4 p) Verify that the estimates of the slope coefficient from **Q6(a)ii** and **Q6(b)iii** are numerically identical to $\hat{\rho}$ obtained in **Q5a**. Use this fact to give meaning to the expression "partialling-out" interpretation of OLS in a MLRM. **Hint**: Think about what steps **Q6(a)i** and **Q6(b)ii** accomplish.

Q7. (20 p) Consider the partially-linear specification in expression (3). Here you estimate ρ via the the Double Machine Learning (DML) estimation procedure of Robinson (1988)[4], as detailed below.

(a) (2 p) Install four R packages: `DoubleML`, `data.table`, `mlr3`, and `mlr3learners`.

(b) (2 p) If your data is not already a `data.table` object convert it. **Programming Guidance:** Assuming that your dataframe is called `df`, use `dt <- data.table::as.data.table(df)`. `data.table` is an extension of `data.frame` and allows for fast manipulation of very large data.

[3] To dive deeper, read wikipedia page or Mixtape Section 2.26.

[4] Robinson, P. M. (1988). Root-N-consistent semi-parametric regression. Econometrica 56, 931-54. doi:10.2307/1912705

(c) (2 p) Collect all the original OPVs in a list named, for example, `pretreat_colnames`. Note: Henceforth when we refer to these OPVs in mathematical expressions we use the notation $\mathbf{x}_i$.

(d) (2 p) Specify data and variables for the causal model by running the script:

```
dml_data_psid <- DoubleML::DoubleMLData$new(dt,
                                            y_col = "re78",
                                            d_cols = "treat",
                                            x_cols = pretreat_colnames)
```

Look at the resulting object.

(e) (2 p) Suppress messages from the `mlr3` package by adding `lgr::get_logger("mlr3")$set_threshold("warn")` to your script.

(f) (2 p) Here you mimic the first stage of Procedure B in **Q6b**. Namely, you specify the model for the two regression functions $l(\mathbf{x}) = E[\texttt{re78}_i|\mathbf{x}_i = \mathbf{x}]$ and $m(\mathbf{x}) = E[\texttt{treat}_i|\mathbf{x}_i = \mathbf{x}]$. In **Q6b** you used a linear-in-parameter model and a priori decided which OPVs to include and which transformations to apply to the OPVs to include (e.g., you excluded `u74`, you used both `age` and `agesq`, you left as-is the other included OPVs). Instead here you do not a priori exclude any OPVs, and you use flexible models, which accommodate complex non-linearities. Run the script:

```
# Specify a RF model as the learner model for l(x)=E[re78|X=x]
ml_l_rf <- mlr3::lrn("regr.ranger")

# Specify a RF model as the learner model for m(x)=E[treat|X=x]
ml_m_rf <- mlr3::lrn("classif.ranger")
```

The above script uses a Random Forest (RF) model for both conditional expectations functions.[5]

(g) (2 p) Here you initialize & parametrize the model object which you later use to perform estimation. Run the script:

```
# Set seeds for cross-fitting
set.seed(3141)

# Set the DML specification
obj_dml_plr <- DoubleML::DoubleMLPLR$new(dml_data_psid,
                                         ml_l = ml_l_rf, ml_m = ml_m_rf,
                                         n_folds = 2,
                                         score = "partialling out",
                                         apply_cross_fitting = TRUE)
```

The above script: (i) utilizes the data object generated in **Q7d**, namely `dml_data_psid`; (ii) utilizes the models for the first stage regressions picked in **Q7f**, namely `ml_l_rf` and `ml_m_rf`; (iii) specifies that we want to split the sample into 2 parts (`n_folds = 2`), and (iv) that we want to use the "partialling out" approach to estimate causal impacts (`score = "partialling out"`), and (v) that we want to apply cross-fitting (`apply_cross_fitting = TRUE`).

(h) (2 p) Here you fit the DML model defined in **Q7g**. Run the script:

```
obj_dml_plr$fit()
obj_dml_plr
```

At a high level the above script implements all of the following operations: (i) fits the two models for the first stage selected in **Q7f**, (ii) gets residuals, (iii) regresses the residuals for the outcome variables onto the residuals for the treatment indicator to obtain the DML estimate of ρ in expression (3). Note: You specified `n_folds = 2` and requested `apply_cross_fitting = TRUE` in **Q7g** thus the 2-stage estimation procedure proceed as follows. First the entire data is split into two sub-samples, call them A and B (hence the term "2 folds"). Sample A is used to fit the 1st stage models. These fitted models are used to compute residuals in sample B and these residuals are used to fit the 2nd stage model using only data in sample B. Denote the resulting estimate $\hat{\rho}_{AB}$. Then the samples are swapped (hence the term "cross fitting").[6] That is, sample B is used to fit the 1st stage models. Sample A is used to fit the 2nd stage model. Denote the resulting estimate $\hat{\rho}_{BA}$. The DML estimate is the average of $\hat{\rho}_{AB}$ and $\hat{\rho}_{BA}$.

(i) (4 p) Take a look at the output, i.e., at the object `obj_dml_plr`. How does the DML estimate of average treatment effect compare to the estimates based on specifications (1) and (2)?

[5] You do not need to know what a RFM is. Think of this approach as a way to flexibly estimate the form of a function of many variables. If you want to learn more about these approaches consider taking ECMA 31350 in Winter 2024.

[6] Cross-fitting is implemented to eliminate the bias from overfitting resulting from the fact that the two conditional mean functions $l(\cdot)$ and $m(\cdot)$ are estimated via ML models, in our case the RF models specified in **Q7f**.