Economics 675: Applied Microeconometrics – Fall 2018

Assignment 4 – Due date: Mon 12-Nov

Last updated: October 17, 2018

Contents

Guidelines:

- You may work in (small) groups while solving this assignment.
- Submit individual solutions via `http://canvas.umich.edu` in one PDF file collecting everything (e.g., derivations, figures, tables, computer code).
- Start each question on a separate page. Always add a reference section if you cite other sources.
- Clearly label all tables and figures, and always include a brief footnote with useful information.
- Always attach your computer code as an appendix, with annotations/comments as appropriate.
- Please provide as much detail as possible in your answers, both analytical and empirical.

1 Question 1: Estimating Equations

Consider a multi-valued treatment effect model with a generic treatment taking $J+1$ distinct values, which we denote $T_i \in \mathcal{T} = \{0, 1, 2, \cdots, J\}$ for each unit i in the sample. For each unit i, we let $\{Y_i(0), Y_i(1), \cdots, Y_i(J)\}$ be the collection of $J+1$ potential outcomes. We assume the observed data is a random sample from a large population, denoted $\mathbf{Z}_i = (Y_i, T_i, \mathbf{X}_i')'$ with $i = 1, 2, \cdots, n$, where $\mathbf{X}_i \in \mathbb{R}^d$ is a vector of observed pre-treatment (i.e., pre-determined) covariates and

$$Y_i = \sum_{t=0}^{J} D_i(t) Y_i(t) = \begin{cases} Y_i(0) & \text{if } T_i = 0 \\ \vdots & \vdots \\ Y_i(J) & \text{if } T_i = J \end{cases}$$

with $D_i(t) = \mathbf{1}(T_i = t)$, $t \in \mathcal{T}$. Treatment effect parameters are functions of specific features of the potential outcomes, such as the means ($\mu_t = \mathbb{E}[Y(t)]$) or distribution functions ($F_t(y) = \mathbb{E}[\mathbf{1}(Y_i(t) \leq y)]$), which could also be defined for a specific subpopulation (e.g., for those units with $T_i = t$).

This question discusses identification under selection-on-observables of the generic class of parameters:

$$\theta_t(g) = \mathbb{E}[g(Y_i(t))], \qquad g \in \mathcal{G}, \qquad t \in \mathcal{T},$$

where $\mathcal{G}$ denotes a class of functions (e.g., $\mathcal{G} = \{\mathbf{1}(\cdot \leq y) : y \in \mathbb{R}\}$). Define the regression functions:

$$p_t(\mathbf{X}_i) = \mathbb{P}[T_i = t | \mathbf{X}_i], \qquad e_t(g; \mathbf{X}_i) = \mathbb{E}[g(Y_i(t)) | \mathbf{X}_i] = \mathbb{E}[g(Y_i) | \mathbf{X}_i, T_i = 1], \qquad g \in \mathcal{G}, \qquad t \in \mathcal{T}.$$

Throughout this question we assume *Ignorability*: $Y_i(t) \perp\!\!\!\perp D_i(t) \mid \mathbf{X}_i$ and $0 < c < p_t(\mathbf{X}_i)$, for all $t \in \mathcal{T}$ and for some fixed positive constant c.

1. Define, for all $g \in \mathcal{G}$ and $t \in \mathcal{T}$,

$$\psi_{\texttt{IPW},t}(\mathbf{Z}_i; \theta_t(g)) = \frac{D_i(t) \cdot g(Y_i(t))}{p_t(\mathbf{X}_i)} - \theta_t(g),$$
$$\psi_{\texttt{RI1},t}(\mathbf{Z}_i; \theta_t(g)) = e_t(g; \mathbf{X}_i) - \theta_t(g), \qquad \psi_{\texttt{RI2},t}(\mathbf{Z}_i; \theta_t(g)) = \frac{D_i(t) \cdot e_t(g; \mathbf{X}_i)}{p_t(\mathbf{X}_i)} - \theta_t(g),$$
$$\psi_{\texttt{DR},t}(\mathbf{Z}_i; \theta_t(g)) = \frac{D_i(t) \cdot g(Y_i(t))}{p_t(\mathbf{X}_i)} - \theta_t(g) - \frac{e_t(g; \mathbf{X}_i)}{p_t(\mathbf{X}_i)} \cdot (D_i(t) - p_t(\mathbf{X}_i)).$$

   Show that these functions are valid moment conditions for $\theta_t(g)$, with $g \in \mathcal{G}$ and $t \in \mathcal{T}$.

2. Using each of the estimating equations above, propose plug-in semiparametric estimators for $\theta_t(g)$, with $g \in \mathcal{G}$ and $t \in \mathcal{T}$. Discuss how they are implemented, including their asymptotic properties and corresponding methods for inference (i.e., standard errors and resampling methods).

3. Apply the methods above to construct causal estimators of $\sigma_t^2 = \mathbb{V}[Y_i(t)]$, with $t \in \mathcal{T}$. Explain in detail how you would conduct the hypothesis test: $\mathsf{H}_0 : \sigma_t^2 = \sigma^2 \ \forall t \in \mathcal{T}$ using the results above.

4. (Extra credit.) Suppose now that you are interested in testing stochastic dominance between treatment levels $\{1, 2\} \in \mathcal{T}$. Explain in detail how you would conduct such a hypothesis test using the methods above (properly extended).

2 Question 2: Estimating Average Treatment Effects

LaLonde (1986) examined the impact of the National Supported Work (NSW) Demonstration on the post-training income. In this question we investigate to what extend program evaluation econometric techniques could be used to conduct causal inference when there is no experimental data available. Read the appendices carefully before working on this question.

Experimental Data. The program was a federally and privately funded program implemented in the mid-1970s to provide work experience for a period of 6-18 months to individuals who had faced economic and social problems prior to enrollment in the program. Candidates eligible for the NSW were randomized into the program between March 1975 and July 1977. A subset of the file `LaLonde_all.csv` is the data used in Dehejia and Wahba (1999), which is a subset of the initial LaLonde (1986), where earnings in 1974 is available. More specifically, the *experimental treatment group* is denoted by $\texttt{treat}_i = 1$, and the *experimental control group* is denoted by $\texttt{treat}_i = 0$ (number of observations: $185 + 260 = 445$).

PSID Control Group. The rest of the dataset `LaLonde_all.csv` contains individuals from the Panel Study of Income Dynamics (PSID), which is regarded as the *nonexperimental (PSID) control group*, which is labeled as $\texttt{treat}_i = 2$ (number of observations: 2490). To save space, variable descriptions and summary statistics are left to the appendix.

Empirical Exercise. Use `Stata` commands (see appendix), or `R` commands, to estimate the ATE and ATT with each of the following estimating approaches: **[1]** Difference-in-Means (no covariates needed, and the same for ATE and ATT), **[2]** Linear Least-Squares (same for ATE and ATT), **[3]** Regression Imputation (RI, different for ATE and ATT), **[4]** Inverse Probability Weighting (IPW, different for ATE and ATT), **[5]** Doubly Robust (DR, different for ATE and ATT), **[6]** Nearest Neighbor Matching (different for ATE and ATT), and **[7]** Propensity Score Matching (different for ATE and ATT).

Notice that for those procedures involving propensity score estimation, `Stata` (or `R`) may give you an error message that some observations have estimated propensity scores too close to 0 (or 1). If this happens you should exclude those observations from the final estimation. Also the logit (or probit) regression may not converge. In such cases you could set a maximum number of iterations (for example 50). [For the current dataset this arbitrary choice does not matter too much, but in general you should be very careful when numerical procedure does not converge.]

For each estimator use either the experimental control group or the PSID control group, whenever possible. In addition, for those estimators requiring covariates, consider three different specifications:

$$
\begin{aligned}
\mathbf{z}_i^{(a)} &= \texttt{age}_i \;,\; \texttt{educ}_i \;,\; \texttt{black}_i \;,\; \texttt{hisp}_i \;,\; \texttt{married}_i \;,\; \texttt{nodegr}_i \;,\; \log(\texttt{re74}_i + 1) \;,\; \log(\texttt{re75}_i + 1) \\
\mathbf{z}_i^{(b)} &= \texttt{age}_i \;,\; \texttt{educ}_i \;,\; \texttt{black}_i \;,\; \texttt{hisp}_i \;,\; \texttt{married}_i \;,\; \texttt{nodegr} \;,\; \log(\texttt{re74} + 1) \;,\; \log(\texttt{re75} + 1) \\
&\quad\; \texttt{age}_i^2 \;,\; \texttt{educ}_i^2 \;,\; \texttt{u74}_i \;,\; \texttt{u75}_i \\
\mathbf{z}_i^{(c)} &= \texttt{age} \;,\; \texttt{educ}_i \;,\; \texttt{black}_i \;,\; \texttt{hisp}_i \;,\; \texttt{married}_i \;,\; \texttt{nodegr}_i \;,\; \log(\texttt{re74}_i + 1) \;,\; \log(\texttt{re75}_i + 1) \\
&\quad\; \texttt{age}_i^2 \;,\; \texttt{educ}_i^2 \;,\; \texttt{u74}_i \;,\; \texttt{u75}_i \\
&\quad\; \texttt{age}_i^3 \;,\; \texttt{black}_i \times \texttt{u74}_i \;,\; \texttt{educ}_i \times \log(\texttt{re74}_i + 1)
\end{aligned}
$$

In the end, you should be able to complete Table 1 below. Do not forget to give a very brief discussion of your findings.

Table 1: Estimation and Inference on ATE and ATT

(a) ATE

	Experimental Data			PSID Control		
	$\widehat{\tau}$	s.e.	C.I.	$\widehat{\tau}$	s.e.	C.I.
Mean Diff.						
	1794.34			−15204.78		
OLS						
a	1582.17					
b	1506.90					
c	1501.37					
Reg. Impute						
a	1462.27			−11195.04		
b						
c						
IPW						
a						
b						
c						
D. Robust						
a						
b						
c						
N1 Match						
a	1829.80			−15619.49		
b						
c						
p Match						
a						
b						
c						

(b) ATT

	Experimental Data			PSID Control		
	$\widehat{\tau}$	s.e.	C.I.	$\widehat{\tau}$	s.e.	C.I.
Mean Diff.						
	1794.34			−15204.78		
OLS						
a	1582.17			6302.40		
b	1506.90					
c	1501.37					
Reg. Impute						
a	1726.60			8543.16		
b						
c						
IPW						
a						
b						
c						
D. Robust						
a						
b						
c						
N1 Match						
a	1558.16			1671.33		
b						
c	1137.43			2003.80		
p Match						
a						
b						
c						

Columns. (i) Experimental Data: experimental treatment group and experimental control group. (ii) PSID Control: experimental treatment group and PSID control group. (iii) $\widehat{\tau}$: point estimator of the treatment effect. (iv) s.e.: standard error. (v) C.I.: 95% confidence interval.
Rows. (i) Mean Diff.: mean difference estimator. (ii) OLS: regression estimator. (iii) Reg. Impute: regression imputation estimator. (iv) IPW: inverse probability weighting (with logistic propensity score estimation). (v) D. Robust: doubly robust estimator. (vi) N1 Match: nearest-neighbor matching estimator. (vii) p Match: propensity score matching estimator.
a, **b** and **c** are different covariates specifications. See the text for details.

3 Question 3: Post-model Selection Inference

This question is about model selection and its implications for inference procedures. Consider the following linear regression model:

$$y_i = \alpha_0 + \beta_0 \cdot x_i + \gamma_0 \cdot z_i + \varepsilon_i, \qquad \varepsilon_i \perp\!\!\!\perp (x_i, z_i), \qquad \mathbb{E}[\varepsilon_i] = 0. \tag{1}$$

Assume the parameter of interest is β_0, then a common practice is to first estimate the long regression (1), which gives

$$y_i = \widehat{\alpha} + \widehat{\beta}_{(\mathsf{se}(\widehat{\beta}))} \cdot x_i + \widehat{\gamma}_{(\mathsf{se}(\widehat{\gamma}))} \cdot z_i + \widehat{\varepsilon}_i, \tag{2}$$

and if $|\widehat{\gamma}/\mathsf{se}(\widehat{\gamma})| \geq 1.96$, say, one will report $\widehat{\beta}$ and make inference based on (2). On the other hand if $|\widehat{\gamma}/\mathsf{se}(\widehat{\gamma})| < 1.96$, then the researcher will drop z_i from the regression model and reestimate β, which gives

$$y_i = \widetilde{\alpha} + \widetilde{\beta}_{(\mathsf{se}(\widetilde{\beta}))} \cdot x_i + \widetilde{\varepsilon}_i, \tag{3}$$

and the researcher will report $\widetilde{\beta}$ and the inference will be based on (3).

A key, empirically relevant question is: if such model selection procedure is used systematically, is inference based on the usual asymptotic approximation still valid?

To understand the implications of model/covariate selection in empirical work, consider the following simple DGP:

$$\begin{bmatrix} x_i \\ z_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix} \right), \qquad \varepsilon_i | (x_i, z_i) \sim \mathcal{N}(0,1), \qquad y_i = 1 + 0.5 \cdot x_i + z_i + \varepsilon_i.$$

Set the sample size to $n = 50$, and simulate the above model 1000 times. For each simulation, construct the following quantity: (i) $\widehat{\beta}$ as in (2); (ii) $\widetilde{\beta}$ as in (3); and (iii) $\check{\beta}$ defined as

$$\check{\beta} = \begin{cases} \widehat{\beta} & \text{if } |\widehat{\gamma}/\mathsf{se}(\widehat{\gamma})| \geq 1.96 \\ \widetilde{\beta} & \text{if } |\widehat{\gamma}/\mathsf{se}(\widehat{\gamma})| < 1.96, \end{cases}$$

which represents the estimate of β_0 if the model selection rule is used.

1. Give summary statistics of the distribution of the estimators (i)–(iii), and present their kernel density plots.

2. Discuss whether inference based on the conventional asymptotic approximation for linear least-square estimators is valid when a model/covariate selection rule has been used systematically. In particular, report the empirical coverage rate (across simulations) of asymptotic 95% confidence intervals for β_0 constructed using each of the three alternative estimation/inference approaches, and explain why they give different answers.

4 Appendix: `Lalonde_all.csv` Data Description

Variable Description

The following table gives meaning of the variables. `treat` is the treatment indicator (pay special attention to the definition of the two control groups), and `re78` is the outcome variable.

Variable	Description
treat	1 = experimental treatment; 0 = experimental control; 2 = PSID control
re78	earnings in 1978 (post-program outcome), measured in $
age	age, measured in years
educ	educational attainment, measured in years
black	dummy, 1 = black
hisp	dummy, 1 = Hispanic
married	dummy, 1 = married
nodegr	dummy, 1 = high school dropouts (i.e. *educ*>= 12)
u74	unemployed in 1974 (i.e. *earn74*= 0)
u75	unemployed in 1975 (i.e. *earn75*= 0)
re74	earnings 13-24 months prior to the training, measured in $
re75	earnings in 1975, measured in $

Summary Statistics

The following are summary statistics, separately for the experimental treatment group, the experimental control group, and the nonexperimental (PSID) control group.

Columns: `mean`: sample average; `sd`: sample standard deviation; `q()`: sample quantile; `min`: sample minimum; `median`: sample median; `max`: sample maximum.

	Experimental Treatment (obs=185)						
	mean	sd	min	q(.1)	median	q(.9)	max
Treatment Indicator							
treat	1.000						
Outcome							
re78	6349.144	7867.402	0	0	4232.309	14553.088	60307.93
Covariates							
age	25.816	7.155	17	18	25.000	37.000	48.00
educ	10.346	2.011	4	8	11.000	12.000	16.00
black	0.843	0.365	0	0	1.000	1.000	1.00
hisp	0.059	0.237	0	0	0.000	0.000	1.00
married	0.189	0.393	0	0	0.000	1.000	1.00
nodegr	0.708	0.456	0	0	1.000	1.000	1.00
u74	0.708	0.456	0	0	1.000	1.000	1.00
u75	0.600	0.491	0	0	1.000	1.000	1.00
re74	2095.574	4886.620	0	0	0.000	8474.058	35040.07
re75	1532.055	3219.251	0	0	0.000	5407.925	25142.24

	Experimental Control (obs=260)						
	mean	sd	min	q(.1)	median	q(.9)	max
Treatment Indicator							
treat	0.000						
Outcome							
re78	4554.801	5483.836	0	0	3138.796	11355.290	39483.53
Covariates							
age	25.054	7.058	17	18	24.000	34.000	55.00
educ	10.088	1.614	3	8	10.000	12.000	14.00
black	0.827	0.379	0	0	1.000	1.000	1.00
hisp	0.108	0.311	0	0	0.000	1.000	1.00
married	0.154	0.361	0	0	0.000	1.000	1.00
nodegr	0.835	0.372	0	0	1.000	1.000	1.00
u74	0.750	0.434	0	0	1.000	1.000	1.00
u75	0.685	0.466	0	0	1.000	1.000	1.00
re74	2107.027	5687.906	0	0	0.000	7628.054	39570.68
re75	1266.909	3102.982	0	0	0.000	4493.002	23031.98

	PSID Control (obs=2490)						
	mean	sd	min	q(.1)	median	q(.9)	max
Treatment Indicator							
treat	2.000						
Outcome							
re78	21553.921	15555.346	0	0.000	20688.172	38420.891	121173.58
Covariates							
age	34.851	10.441	18	22.000	33.000	50.000	55.00
educ	12.117	3.082	0	8.000	12.000	16.000	17.00
black	0.251	0.433	0	0.000	0.000	1.000	1.00
hisp	0.033	0.177	0	0.000	0.000	0.000	1.00
married	0.866	0.340	0	0.000	1.000	1.000	1.00
nodegr	0.305	0.461	0	0.000	0.000	1.000	1.00
u74	0.086	0.281	0	0.000	0.000	0.000	1.00
u75	0.100	0.300	0	0.000	0.000	0.100	1.00
re74	19428.746	13406.877	0	1716.318	18417.108	35266.802	137148.68
re75	19063.338	13596.955	0	64.452	17903.226	34176.005	156653.23

5 Appendix: Average Treatment Effect Estimators

Treatment Effect

Here we keep using the Rubin causal framework, in which for each individual we define the following

- $(Y_i(1), Y_i(0))$: potential outcomes given the treatment status;
- T_i: treatment status $(= 1$)if treated;
- $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$: the observed outcome;
- $\mathbf{X}_i$: predetermined covariates.

The following are several treatment effects which are commonly seen in the literatures:

- ATE (average treatment effect): $\tau_{\texttt{ATE}} = \mathbb{E}[Y_i(1) - Y_i(0)]$;
- ATT (average treatment effect on the treated): $\tau_{\texttt{ATT}} = \mathbb{E}[Y_i(1) - Y_i(0)|T_i = 1]$;
- ATU (average treatment effect on the untreated): $\tau_{\texttt{ATU}} = \mathbb{E}[Y_i(1) - Y_i(0)|T_i = 0]$.

Estimators

Mean Comparison

Mean comparison is the simplest way to estimate the average treatment effect, which is typically used for experimental data. The estimator is

$$\widehat{\tau}_{\texttt{ATE}} = \frac{1}{n_1} \sum_{i:T_i=1} Y_i - \frac{1}{n_0} \sum_{i:T_i=0} Y_i.$$

Note that for the experimental data all the mean treatment effects (ATE, ATT, ATU) are the same and the above estimator applies. For nonexperimental data, however, the simple mean comparison cannot be used to make inference on ATT or ATU.

In Stata, this could easily be done by regressing the outcome variable on the treatment indicator and a constant. Note that for valid inference, one typically needs the heteroskedastic robust standard error:

```
reg y t, hc2
```

OLS

For the experimental data, the treatment indicator is randomly assigned hence should be orthogonal to the covariates. Then sometimes it is suggested to include covariates in a regression, which could help to reduce the estimation variability (i.e. standard error). This is done by regressing Y_i on the treatment indicator and some covariates $\mathbf{Z}_i$, and the covariates could be $\mathbf{X}_i$, or a subset, or some transformation, or higher order interactions to achieve flexibility.

Again note that this method cannot be used to make inference on ATT or ATU for nonexperimental data.

Regression Imputation

Let $\mathbf{Z}_i$ be a set of covariates, which are chosen to include transformations and interactions of the original variables in $\mathbf{X}_i$, to achieve flexible functional form. Then the regression imputation approach is typically done in the following steps:

1. Regress Y_i on $\mathbf{Z}_i$ for the treated group to obtain $\widehat{\boldsymbol{\beta}}^1$;
2. Regress Y_i on $\mathbf{Z}_i$ for the control group to obtain $\widehat{\boldsymbol{\beta}}^0$;
3. For each individual impute the "individual treatment effect" $\widehat{\tau}_i = \mathbf{Z}_i^{\mathsf{T}}(\widehat{\boldsymbol{\beta}}^1 - \widehat{\boldsymbol{\beta}}^0)$;
4. The treatment effects are estimated by

$$\widehat{\tau}_{\texttt{ATE}} = \frac{1}{n}\sum_i \widehat{\tau}_i \qquad \widehat{\tau}_{\texttt{ATT}} = \frac{1}{n_1}\sum_{i:T_i=1} \widehat{\tau}_i \qquad \widehat{\tau}_{\texttt{ATU}} = \frac{1}{n_0}\sum_{i:T_i=0} \widehat{\tau}_i.$$

Effectively by utilizing flexible functional form, this method estimates the two conditional expectation functions (of Y_i given $\mathbf{X}_i$). The Stata command for this estimator is

```
* ATE
teffects ra (y covariates) (t), ate
* ATT
teffects ra (y covariates) (t), atet
```

Inverse Probability Weighting

Again let $\mathbf{Z}_i$ be a set of covariates which is "rich" enough, the IPW estimator takes the following steps:

1. Estimate the propensity score for each individual, by regressing T_i on the covariates $\mathbf{Z}_i$ (probit, logit, etc.). Denote the estimated propensity score by $\widehat{p}(\mathbf{Z}_i)$ and $\widehat{p} = \frac{1}{n}\sum_i [T_i]$ be the proportion of the treated;
2. Data trimming (more on this later);
3. The treatment effects are estimated by

$$\begin{aligned}
\widehat{\tau}_{\texttt{ATE}} &= \frac{1}{n}\sum_i \left[\frac{T_i Y_i}{\widehat{p}(\mathbf{Z}_i)}\right] - \frac{1}{n}\sum_i \left[\frac{(1-T_i)Y_i}{1-\widehat{p}(\mathbf{Z}_i)}\right] \\
\widehat{\tau}_{\texttt{ATT}} &= \frac{1}{n}\sum_i \left[\frac{T_i Y_i}{\widehat{p}}\right] - \frac{1}{n}\sum_i \left[\frac{(1-T_i)Y_i}{1-\widehat{p}(\mathbf{Z}_i)}\frac{\widehat{p}(\mathbf{Z}_i)}{\widehat{p}}\right] \\
\widehat{\tau}_{\texttt{ATU}} &= \frac{1}{n}\sum_i \left[\frac{T_i Y_i}{\widehat{p}(\mathbf{Z}_i)}\frac{1-\widehat{p}(\mathbf{Z}_i)}{1-\widehat{p}}\right] - \frac{1}{n}\sum_i \left[\frac{(1-T_i)Y_i}{1-\widehat{p}}\right].
\end{aligned}$$

In Stata, this could be done by

```
* ATE
teffects ipw (y) (t covariates, probit), ate
* ATT
teffects ipw (y) (t covariates, probit), atet
```

Note that the second step requires data trimming, which happens quite often in applied work. The reason is that the estimated propensity score could be very close to 0 or 1 for some observations, which makes the estimator perform poorly in some cases. Therefore people typically use some *ad hoc* trimming rule to drop observations whose estimated propensity scores are too extreme. For ATE, the identification requires the propensity score being bounded away from 0 and 1. For ATT, it is only required to be bounded away from 1. Similarly for ATU one needs the propensity score being bounded away from 0. Different estimators will require different data trimming methods. (For example, if ATT is of interest, then we could tolerate propensity scores very close to zero.)

Doubly Robust Estimation

The doubly robust estimator combines both the regression imputation as well as the inverse probability weighting. For notational convenience, we define the imputed regression functions at a particular observation by $\widehat{e}_1(\mathbf{Z}_i) = \mathbf{Z}_i^{\intercal}\widehat{\boldsymbol{\beta}}^1$ and $\widehat{e}_0(\mathbf{Z}_i) = \mathbf{Z}_i^{\intercal}\widehat{\boldsymbol{\beta}}^0$, then ATE is estimated by

$$\widehat{\tau}_{\texttt{ATE}} = \frac{1}{n}\sum_i \left[\frac{T_i\left(Y_i - \widehat{e}_1(\mathbf{Z}_i)\right)}{\widehat{p}(\mathbf{Z}_i)} + \widehat{e}_1(\mathbf{Z}_i)\right] - \frac{1}{n}\sum_i \left[\frac{(1-T_i)\left(Y_i - \widehat{e}_0(\mathbf{Z}_i)\right)}{1-\widehat{p}(\mathbf{Z}_i)} + \widehat{e}_0(\mathbf{Z}_i)\right],$$

and for ATT

$$\widehat{\tau}_{\texttt{ATT}} = \frac{1}{n}\sum_i \left[\frac{T_i Y_i}{\widehat{p}}\right] - \frac{1}{n}\sum_i \left[\frac{(1-T_i)\left(Y_i - \widehat{e}_0(\mathbf{Z}_i)\right)}{1-\widehat{p}(\mathbf{Z}_i)}\frac{\widehat{p}(\mathbf{Z}_i)}{\widehat{p}} + \widehat{e}_0(\mathbf{Z}_i)\frac{\widehat{p}(\mathbf{Z}_i)}{\widehat{p}}\right],$$

and finally for ATU

$$\widehat{\tau}_{\texttt{ATU}} = \frac{1}{n}\sum_i \left[\frac{T_i\left(Y_i - \widehat{e}_1(\mathbf{Z}_i)\right)}{\widehat{p}(\mathbf{Z}_i)}\frac{1-\widehat{p}(\mathbf{Z}_i)}{1-\widehat{p}} + \widehat{e}_1(\mathbf{Z}_i)\frac{1-\widehat{p}(\mathbf{Z}_i)}{1-\widehat{p}}\right] - \frac{1}{n}\sum_i \left[\frac{(1-T_i)Y_i}{1-\widehat{p}}\right].$$

In Stata the doubly robust estimator is

```
* ATE
teffects ipwra (y covariates) (t covariates, probit), ate
* ATT
teffects ipwra (y covariates) (t covariates, probit), atet
```

Note that the covariates for the outcome equation could be different from that for the selection equation. Also since inverse probability weighting is used, one also needs to drop observations with extreme propensity scores.

Matching

The matching estimator starts with a measure of distance between two observations. Let it be ρ, then for each observation i, define the matched outcome as

$$m(i) = \arg\min_{j:\ T_j \neq T_i} \rho(\mathbf{Z}_i, \mathbf{Z}_j),$$

where the requirement $T_j \neq T_i$ simply says we match a treated individual with someone in the control group and vice versa. Then we could define the following "individual treatment effect"

$$\widehat{\tau}_i = \begin{cases} Y_i - Y_{m(i)} & T_i = 1 \\ Y_{m(i)} - Y_i & Y_i = 0. \end{cases}$$

Then similar as the regression imputation method, the treatment effects are estimated by

$$\widehat{\tau}_{\texttt{ATE}} = \frac{1}{n}\sum_i \widehat{\tau}_i \qquad \widehat{\tau}_{\texttt{ATT}} = \frac{1}{n_1}\sum_{i:T_i=1} \widehat{\tau}_i \qquad \widehat{\tau}_{\texttt{ATU}} = \frac{1}{n_0}\sum_{i:T_i=0} \widehat{\tau}_i.$$

Note that the above is the *nearest-neighbor matching* and requires *match with replacement*; that is, two individuals could be matched to the same one in the comparison group. The above method could be generalized to *nearest-K matching* where in most cases K is a pre-determined integer. Commonly used distance measures include the Mahalanobis metric, which is the Euclidean distance weighted by the inverse covariance matrix of the covariates, and matching based on estimated propensity scores.

Nearest-neighbor matching is implemented in Stata as

```
* ATE
teffects nnmatch (y covariates) (t), ate nneighbor(1) metric(maha)
* ATT
teffects nnmatch (y covariates) (t), atet nneighbor(1) metric(maha)
```

and propensity score matching is

```
* ATE
teffects psmatch (y) (t covariates, probit), ate
* ATT
teffects psmatch (y) (t covariates, probit), atet
```

References

Dehejia, R. H., and S. Wahba (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94(448), 1053–1062.

LaLonde, R. J. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76(4), 604–620.