"New Developments in Econometrics"

Lecture 1

Estimation of Average Treatment Effects

Under Unconfoundedness, Part I

Guido Imbens

Cemmap Lectures, UCL, June 2009

Outline

1. Introduction

2. Potential Outcomes

3. Estimands and Identification

4. Estimation and Inference

1. Introduction

We are interested in estimating the average effect of a program or treatment, allowing for heterogenous effects, assuming that selection can be taken care of by adjusting for differences in observed covariates.

This setting is of great applied interest.

Long literature, in both statistics and economics. Influential economics/econometrics papers include Ashenfelter and Card (1985), Barnow, Cain and Goldberger (1980), Card and Sullivan (1988), Dehejia and Wahba (1999), Hahn (1998), Heckman and Hotz (1989), Heckman and Robb (1985), Lalonde (1986). In stat literature work by Rubin (1974, 1978), Rosenbaum and Rubin (1983).

Unusual case with many proposed (semi-parametric) estimators (matching, regression, propensity score, or combinations), many of which are actually used in practice.

We discuss implementation, and assessment of the critical assumptions (even if they are not testable).

In practice concern with overlap in covariate distributions tends to be important.

Once overlap issues are addressed, choice of estimators is less important. Estimators combining matching and regression or weighting and regression are recommended for robustness reasons.

Key role for analysis of the joint distribution of treatment indicator and covariates prior to using outcome data.

2. Potential Outcomes (Rubin, 1974)

We observe N units, indexed by $i = 1, \ldots, N$, viewed as drawn randomly from a large population.

We postulate the existence for each unit of a pair of potential outcomes,
$Y_i(0)$ for the outcome under the control treatment and
$Y_i(1)$ for the outcome under the active treatment
$Y_i(1) - Y_i(0)$ is unit-level causal effect

Covariates X_i (not affected by treatment)
Each unit is exposed to a single treatment; $W_i = 0$ if unit i receives the control treatment and $W_i = 1$ if unit i receives the active treatment. We observe for each unit the triple (W_i, Y_i, X_i), where Y_i is the realized outcome:

$$Y_i \equiv Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

Several additional pieces of notation.

First, the propensity score (Rosenbaum and Rubin, 1983) is defined as the conditional probability of receiving the treatment,

$$e(x) = \mathsf{Pr}(W_i = 1 | X_i = x) = \mathbb{E}[W_i | X_i = x].$$

Also the two conditional regression and variance functions:

$$\mu_w(x) = \mathbb{E}[Y_i(w) | X_i = x], \quad \sigma^2_w(x) = \mathbb{V}(Y_i(w) | X_i = x).$$

3. Estimands and Identification

Population average treatments

$$\tau_P = \mathbb{E}[Y_i(1) - Y_i(0)] \quad \tau_{P,T} = \mathbb{E}[Y_i(1) - Y_i(0)|W = 1].$$

Most of the discussion in these notes will focus on τ_P, with extensions to $\tau_{P,T}$ available in the references.

We will also look at the sample average treatment effect (SATE):

$$\tau_S = \frac{1}{N} \sum_{i=1}^{N} (Y_i(1) - Y_i(0))$$

τ_P versus τ_S does not matter for estimation, but matters for variance.

4. Estimation and Inference

Assumption 1 (Unconfoundedness, Rosenbaum and Rubin, 1983a)

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid X_i.$$

"conditional independence assumption," "selection on observables." In missing data literature "missing at random."

To see the link with standard exogeneity assumptions, assume constant effect and linear regression:

$$Y_i(0) = \alpha + X_i'\beta + \varepsilon_i, \quad \Longrightarrow \quad Y_i = \alpha + \tau \cdot W_i + X_i'\beta + \varepsilon_i$$

with $\varepsilon_i \perp\!\!\!\perp X_i$. Given the constant treatment effect assumption, unconfoundedness is equivalent to independence of W_i and ε_i conditional on X_i, which would also capture the idea that W_i is exogenous.

Motivation for Unconfoundeness Assumption (I)

The first is a statistical, data descriptive motivation.

A natural starting point in the evaluation of any program is a comparison of average outcomes for treated and control units.

A logical next step is to adjust any difference in average outcomes for differences in exogenous background characteristics (exogenous in the sense of not being affected by the treatment).

Such an analysis may not lead to the final word on the efficacy of the treatment, but the absence of such an analysis would seem difficult to rationalize in a serious attempt to understand the evidence regarding the effect of the treatment.

Motivation for Unconfoundeness Assumption (II)

A second argument is that almost any evaluation of a treatment involves comparisons of units who received the treatment with units who did not.

The question is typically not whether such a comparison should be made, but rather which units should be compared, that is, which units best represent the treated units had they not been treated.

It is clear that settings where some of necessary covariates are not observed will require strong assumptions to allow for identification. E.g., instrumental variables settings Absent those assumptions, typically only bounds can be identified (e.g., Manski, 1990, 1995).

Motivation for Unconfoundeness Assumption (III)

Example of a model that is consistent with unconfoundedness: suppose we are interested in estimating the average effect of a binary input on a firm's output, or $Y_i = g(W, \varepsilon_i)$.

Suppose that profits are output minus costs,

$$W_i = \arg\max_w \mathbb{E}[\pi_i(w)|c_i] = \arg\max_w \mathbb{E}[g(w, \varepsilon_i) - c_i \cdot w|c_i],$$

implying

$$W_i = 1\{\mathbb{E}[g(1, \varepsilon_i) - g(0, \varepsilon_i) \geq c_i|c_i]\} = h(c_i).$$

If unobserved marginal costs c_i differ between firms, and these marginal costs are independent of the errors ε_i in the firms' forecast of output given inputs, then unconfoundedness will hold as

$$(g(0, \varepsilon_i), g(1, \varepsilon_i)) \perp\!\!\!\perp c_i.$$

Overlap

Second assumption on the joint distribution of treatments and covariates:

Assumption 2 (Overlap)

$$0 < \Pr(W_i = 1|X_i) < 1.$$

Rosenbaum and Rubin (1983a) refer to the combination of the two assumptions as "stongly ignorable treatment assignment."

Identification Given Assumptions

$$\tau(x) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x] = \mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x]$$

$$= \mathbb{E}[Y_i(1)|X_i = x, W_i = 1] - \mathbb{E}[Y_i(0)|X_i = x, W_i = 0]$$

$$= \mathbb{E}[Y_i|X_i, W_i = 1] - \mathbb{E}[Y_i|X_i, W_i = 0].$$

To make this feasible, one needs to be able to estimate the expectations $\mathbb{E}[Y_i|X_i = x, W_i = w]$ for all values of w and x in the support of these variables. This is where overlap is important.

Given identification of $\tau(x)$,

$$\tau_P = \mathbb{E}[\tau(X_i)]$$

Alternative Assumptions

$$\mathbb{E}[Y_i(w)|W_i, X_i] = \mathbb{E}[Y_i(w)|X_i],$$

for $w = 0, 1$. Although this assumption is unquestionably weaker, in practice it is rare that a convincing case can be made for the weaker assumption without the case being equally strong for the stronger Assumption.

The reason is that the weaker assumption is intrinsically tied to functional form assumptions, and as a result one cannot identify average effects on transformations of the original outcome (e.g., logarithms) without the strong assumption.

If we are interested in $\tau_{P,T}$ it is sufficient to assume

$$Y_i(0) \perp\!\!\!\perp W_i \mid X_i,$$

Propensity Score

Result 1 *Suppose that Assumption 1 holds. Then:*

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i \mid e(X_i).$$

Only need to condition on scalar function of covariates, which would be much easier in practice if X_i is high-dimensional.

(Problem is that the propensity score $e(x)$ is almost never known.)

Efficiency Bound

Hahn (1998): for any regular estimator for τ_P, denoted by $\widehat{\tau}$, with

$$\sqrt{N} \cdot (\widehat{\tau} - \tau_P) \xrightarrow{d} \mathcal{N}(0, V),$$

the variance must satisfy:

$$V \geq \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1 - e(X_i)} + (\tau(X_i) - \tau_P)^2\right]. \tag{1}$$

Estimators exist that achieve this bound.

Estimators

A. Regression Estimators

B. Matching

C. Propensity Score Estimators

D. Mixed Estimators **(recommended)**

A. Regression Estimators

Estimate $\mu_w(x)$ consistently and estimate τ_P or τ_S as

$$\widehat{\tau}_{\text{reg}} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)).$$

Simple implementations include

$$\mu_w(x) = \beta' x + \tau \cdot w,$$

in which case the average treatment effect is equal to τ. In this case one can estimate τ simply by least squares estimation using the regression function

$$Y_i = \alpha + \beta' X_i + \tau \cdot W_i + \varepsilon_i.$$

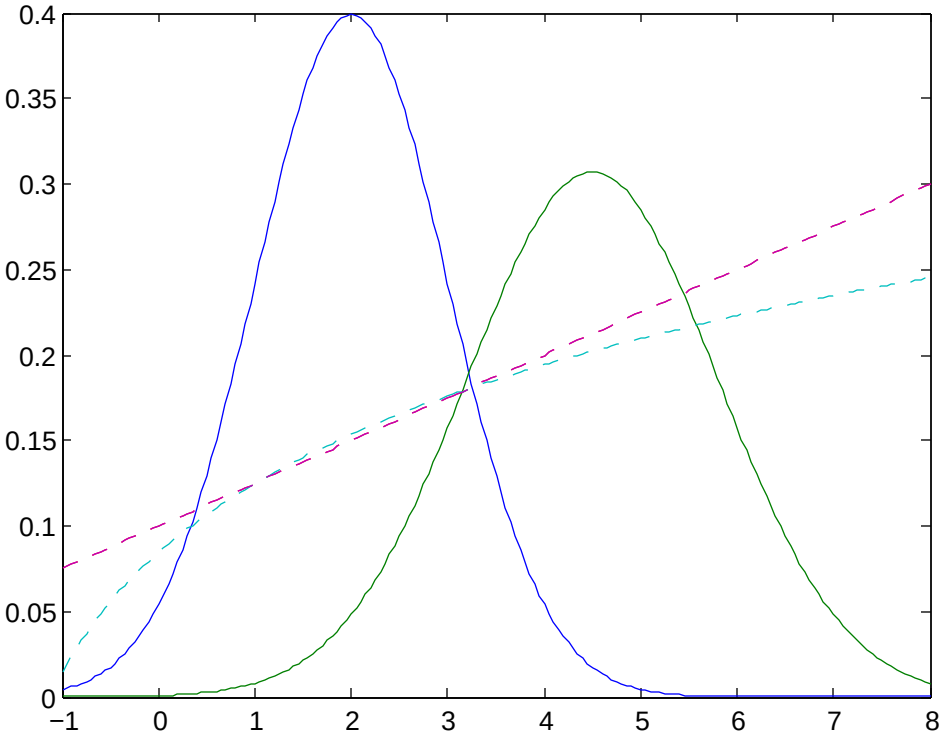
More generally, one can specify separate regression functions for the two regimes, $\mu_w(x) = \beta_w' x$.

These simple regression estimators can be sensitive to differences in the covariate distributions for treated and control units.

The reason is that in that case the regression estimators rely heavily on extrapolation.

Note that $\mu_0(x)$ is used to predict missing outcomes for the treated. Hence on average one wishes to use predict the control outcome at $\overline{X}_T = \sum_i W_i \cdot X_i / N_T$, the average covariate value for the treated. With a linear regression function, the average prediction can be written as $\bar{Y}_C + \widehat{\beta}'(\overline{X}_T - \overline{X}_C)$.

If $\overline{X}_T$ and $\overline{X}_C$ are close, the precise specification of the regression function will not matter much for the average prediction. With the two averages very different, the prediction based on a linear regression function can be sensitive to changes in the specification.


0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−1
0
1
2
3
4
5
6
7
8

B. Matching

let $\ell_m(i)$ is the mth closest match, that is, the index l that satisfies $W_l \neq W_i$ and

$$\sum_{j|W_j \neq W_i} 1\{\|X_j - X_i\| \leq \|X_l - X_i\|\} = m,$$

Then

$$\widehat{Y}_i(0) = \begin{cases} Y_i & \text{if} \quad W_i = 0, \\ \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if} \quad W_i = 1, \end{cases} \qquad \widehat{Y}_i(1) = \begin{cases} \frac{1}{M}\sum_{j \in \mathcal{J}_M(i)} Y_j & \text{if} \\ Y_i & \text{if} \end{cases}$$

The simple matching estimator is

$$\widehat{\tau}_M^{sm} = \frac{1}{N}\sum_{i=1}^{N} \left(\widehat{Y}_i(1) - \widehat{Y}_i(0)\right). \tag{2}$$

Issues with Matching

Bias is of order $O(N^{-1/K})$, where K is dimension of covariates. Is important in large samples if $K \geq 2$ (and dominates variance asymptotically if $K \geq 3$)

Not Efficient (but efficiency loss is small)

Easy to implement, robust.

C.1 Propensity Score Estimators: Weighting

$$\mathbb{E}\left[\frac{WY}{e(X)}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{WY_i(1)}{e(X)}\middle| X\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{e(X)Y_i(1)}{e(X)}\right]\right] = \mathbb{E}[Y_i(1)],$$

and similarly

$$\mathbb{E}\left[\frac{(1-W)Y}{1-e(X)}\right] = \mathbb{E}[Y_i(0)],$$

implying

$$\tau_P = \mathbb{E}\left[\frac{W\cdot Y}{e(X)} - \frac{(1-W)\cdot Y}{1-e(X)}\right].$$

With the propensity score known one can directly implement this estimator as

$$\tilde{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{W_i\cdot Y_i}{e(X_i)} - \frac{(1-W_i)\cdot Y_i}{1-e(X_i)}\right). \tag{3}$$

Implementation of Horvitz-Thompson Estimator

Estimate $e(x)$ flexibly (Hirano, Imbens and Ridder, 2003)

$$\widehat{\tau}_{\text{weight}} = \sum_{i=1}^{N} \frac{W_i \cdot Y_i}{\widehat{e}(X_i)} \Big/ \sum_{i=1}^{N} \frac{W_i}{\widehat{e}(X_i)}$$

$$- \sum_{i=1}^{N} \frac{(1-W_i) \cdot Y_i}{1-\widehat{e}(X_i)} \Big/ \sum_{i=1}^{N} \frac{(1-W_i)}{1-\widehat{e}(X_i)}$$

Is efficient given nonparametric estimator for $e(x)$.

Potentially sensitive to estimator for propensity score.

Matching or Regression on the Propensity Score

Not clear what advantages are.

Large sample properties not known.

Simulation results not encouraging.

D.1 Mixed Estimators: Weighting and Regression

Interpret Horvitz-Thompson estimator as weighted regression estimator:

$$Y_i = \alpha + \tau \cdot W_i + \varepsilon_i, \quad \text{with weights } \lambda_i = \sqrt{\frac{W_i}{e(X_i)} + \frac{1 - W_i}{1 - e(X_i)}}.$$

This weighted-least-squares representation suggests that one may add covariates to the regression function to improve precision, for example as

$$Y_i = \alpha + \beta' X_i + \tau \cdot W_i + \varepsilon_i,$$

with the same weights λ_i. Such an estimator is consistent as long as either the regression model or the propensity score (and thus the weights) are specified correctly. That is, in the Robins-Ritov terminology, the estimator is doubly robust.

Matching and Regression

First match observations.

Define

$$\widehat{X}_i(0) = \begin{cases} X_i & \text{if } W_i = 0, \\ X_{\ell(i)} & \text{if } W_i = 1, \end{cases} \qquad \widehat{X}_i(1) = \begin{cases} X_{\ell(i)} & \text{if } W_i = 0, \\ X_i & \text{if } W_i = 1. \end{cases}$$

Then adjust within pair difference for the within-pair difference in covariates $\widehat{X}_i(1) - \widehat{X}_i(0)$:

$$\widehat{\tau}_M^{adj} = \frac{1}{N} \sum_{i=1}^{N} \left(\widehat{Y}_i(1) - \widehat{Y}_i(0) - \widehat{\beta} \cdot \left(\widehat{X}_i(1) - \widehat{X}_i(0) \right) \right),$$

using regression estimate for β.

Can eliminate bias of matching estimator given flexible specification of regression function.

Estimation of the Variance

For efficient estimator of τ_P:

$$V_P = \mathbb{E}\left[\frac{\sigma_1^2(X_i)}{e(X_i)} + \frac{\sigma_0^2(X_i)}{1-e(X_i)} + (\mu_1(X_i) - \mu_0(X_i) - \tau)^2\right],$$

Estimate all components nonparametrically, and plug in.

Alternatively, use bootstrap.

(Does not work for matching estimator)

Estimation of the Variance

For all estimators of τ_S, for some known $\lambda_i(\mathbf{X}, \mathbf{W})$

$$\widehat{\tau} = \sum_{i=1}^{N} \lambda_i(\mathbf{X}, \mathbf{W}) \cdot Y_i,$$

$$V\left(\widehat{\tau}|\mathbf{X}, \mathbf{W}\right) = \sum_{i=1}^{N} \lambda_i(\mathbf{X}, \mathbf{W})^2 \cdot \sigma^2_{W_i}(X_i).$$

To estimate $\sigma^2_{W_i}(X_i)$ one uses the closest match within the set of units with the same treatment indicator. Let $v(i)$ be the closest unit to i with the same treatment indicator.

The sample variance of the outcome variable for these 2 units can then be used to estimate $\sigma^2_{W_i}(X_i)$:

$$\widehat{\sigma}^2_{W_i}(X_i) = \left(Y_i - Y_{v(i)}\right)^2 /2.$$