AVERAGE TREATMENT EFFECT ESTIMATION: IGNORABLE (OR UNCONFOUNDED) TREATMENT ASSIGNMENT

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Introduction

- What kinds of questions can we answer using a “modern” approach to treatment effect estimation? Here are some examples:

1. What are the effects of a job training program on employment or labor earnings?
2. What are the effects of a school voucher program on student performance?
3. Does a certain medical intervention increase the likelihood of survival?

- The main issue in program evaluation concerns the assignment of the binary intervention, or “treatment.”

- For example, is the “treatment” randomly assigned? (Hardly ever in business and economics, and problematical even in clinical trials because those chosen to be eligible can and do opt out.)
- A more reasonable possibility is that the treatment is effectively randomly assigned conditional on observable covariates. (“Ignorability” of treatment, “unconfoundedness,” or “selection on observables.” Sometimes called “exogenous treatment.”)

- Or, does assignment depend fundamentally on unobservables, where the dependence cannot be broken by controlling for observables? ("Nonignorable" treatment, "confounded" assignment, "selection on unobservables," or "endogenous treatment.")
- Often there is a component of self-selection in program evaluation.

• Nevertheless, start with unconfoundedness because it is often all we have (and is a good starting point in any case). A key point is that, under the ignorability or unconfoundedness assumption, regression methods with the covariates as controls have the same ability – at least in theory – of identifying the treatment effect parameters. Therefore, propensity score methods and/or matching methods are not a panacea for the self-selection problem.

2. Basic Concepts

Counterfactual Outcomes and Parameters of Interest

- For each population unit, two possible outcomes: y_0 (the outcome without treatment) and y_1 (the outcome with treatment). The binary "treatment" indicator is w, where $w = 1$ denotes "treatment." The nature of y_0 and y_1 – discrete, continuous, some mix – is, for now, unspecified. (The generality this affords is one of the attractions of the **Rubin Causal Model**.)
- The gain from treatment is

$$y_1 - y_0. \tag{1}$$

• For a particular unit i, the gain from treatment is

$$y_{i1} - y_{i0}.$$

If we could observe these gains for a random sample, the problem would be easy: just average the gain across the random sample.

• Problem: For each unit i, only one of y_{i0} and y_{i1} is observed.

• In effect, we have a missing data problem (even though we will eventually assume a random sample of units).

• Two parameters are of primary interest. The **average treatment effect** (**ATE**) is

$$\tau_{ate} = E(y_1 - y_0). \tag{2}$$

The expected gain for a randomly selected unit from the population. This is sometimes called the *average causal effect*.

• The **average treatment effect on the treated** (**ATT**) is the average gain from treatment for those who actually were treated:

$$\tau_{att} = E(y_1 - y_0 | w = 1) \tag{3}$$

- With heterogeneous treatment effects, (2) and (3) can be very different. The ATE might average across the gain from units that would be very unlikely to be subject to treatment (but this depends how the population is defined).
- τ_{ate} has "external validity" in that it tells us something about a randomly drawn unit from the population. τ_{att} is specific to the particular program assignment mechanism.

- Important point: τ_{ate} and τ_{att} are defined without reference to a model or a discussion of the nature of the treatment. In particular, these definitions hold when whether assignment is randomized, unconfounded, or endogenous.
- Not suprisingly, how we estimate τ_{ate} and τ_{att} depends on what we assume about treatment assignment.

Sampling Assumptions

- Assume independent, identically distributed observations from the underlying population. The data we would like to have is $\{(y_{i0}, y_{i1}) : i = 1, \ldots, N\}$, but we only observe w_i and

$$y_i = (1 - w_i)y_{i0} + w_i y_{i1}. \tag{4}$$

- Random sampling rules out treatment status of one unit having an effect on other units.
- Also implies that the outcome for unit i does not affect the outcome for other members of the population.

Estimation under Random Assignment

- With $y = (1 - w)y_0 + wy_1$ we can always write

$$E(y|w) = (1 - w)E(y_0|w) + wE(y_1|w)$$

- Strongest form of random assignment: (y_0, y_1) is independent of w.

Then $E(y_0|w) = E(y_0)$ and $E(y_1|w) = E(y_1)$, and so

$$E(y|w) = (1 - w)E(y_0|w) + wE(y_1|w).$$

• It follows that $E(y|w = 1) = E(y_1)$ and $E(y|w = 0) = E(y_0)$, and so

$$E(y|w = 1) - E(y|w = 0) = E(y_1) - E(y_0) = \tau_{ate} = \tau_{att}. \quad (5)$$

• An unbiased and consistent estimator of $E(y|w = 1)$ is the sample average on the treated subsample and similarly for $E(y|w = 0)$. The estimator $\hat{\tau}_{ate}$ is just the simple difference-in-means estimator.

• The randomization of treatment needed for the simple comparison-of-means estimator to consistently estimate the ATE is rare in practice but not unheard of. (Eligibility is sometimes randomly assigned, but actual participation need not be.)

3. The Key Assumptions: Ignorability and Overlap

• Rather than assume random assignment, for each unit i we also draw a vector of covariates, $\mathbf{x}_i$. Let $\mathbf{x}$ be the random vector with a distribution in the population.

A.1. Ignorability (Unconfoundedness): Conditional on a set of covariates $\mathbf{x}$, the pair of counterfactual outcomes, (y_0, y_1), is independent of w, which is often written as

$$(y_0, y_1) \perp w \mid \mathbf{x}, \tag{6}$$

where the symbol "$\perp$" means "independent of" and "$\mid$" means "conditional on."

- w and (y_0, y_1) might be correlated but not once we control for characteristics $\mathbf{x}$. For example, the probability of being chosen for a job training program differs by education levels but is the same at a given level of education.
- A useful way to express ignorability (conditional on $\mathbf{x}$):

$D(w|y_0, y_1, \mathbf{x}) = D(w|\mathbf{x})$, where $D(\cdot|\cdot)$ denotes conditional distribution.

- Unconfoundedness is controversial. In effect, it underlies standard regression methods to estimating treatment effects (via a “kitchen sink” regression that includes covariates, the treatment indicator, and possibly interactions).

- Can show unconfoundedness is generally violated if **x** includes variables that are themselves affected by the treatment. For example, in evaluating a job training program, **x** should not include post-training schooling because that might have been chosen in response to being assigned or not assigned to the program. We would not want to hold post-training schooling fixed.

• In fact, suppose (y_0, y_1) is independent of w but $D(\mathbf{x}|w) \neq D(\mathbf{x})$. In other words, assignment is randomized with respect to (y_0, y_1) but not with respect to $\mathbf{x}$. (Think of random assignment but then $\mathbf{x}$ is defined to include other outcomes affected by w.) Then ignorability generally fails unless $E(y_g|\mathbf{x}) = E(y_g), j = 0, 1$.

- To see this, by iterated expectations,

$$E(y_g|w) = E[E(y_g|w, \mathbf{x})|w], \; g = 0, 1$$

But, because w is independent of y_g, the left-hand-side does not depend on w, and $E(y_g|w, \mathbf{x})$ does not depend on w if unconfoundedness is supposed to hold.

• Write $\mu_g(\mathbf{x}) \equiv E(y_g|\mathbf{x})$, $g = 0, 1$. Then, if $E(y_g|w) = E(y_g)$ and $E(y_g|w, \mathbf{x}) = \mu_g(\mathbf{x})$ we must have

$$E(y_g) = E[\mu_g(\mathbf{x})|w],$$

which is impossible if the right-hand-side depends on w.

• Most often, $\mathbf{x}$ includes variables that are measured prior to treatment assignment, such as previous labor market history. Of course, gender, race, and other demographic variables can be included.

• A weaker version of ignorability (but still pretty strong) is

A.1′. Ignorability in Conditional Mean:

$$E(y_g|w,\mathbf{x}) = E(y_g|\mathbf{x}),\ g = 0,1. \tag{7}$$

• Seems unlikely that this weaker version of the assumption holds without the stronger version, but, technically, (7) allows things like variances depending on w.

• From above discussion, have to think about what should be included in $\mathbf{x}$. Use the same reasoning as in multiple regression analysis in deciding whether to "hold something fixed."

A.2. **Overlap**: For all **x** in its support $\mathcal{X}$,

$$0 < P(w = 1|\mathbf{x}) < 1. \tag{8}$$

In other words, each unit in the defined population has some chance of being treated and some chance of not being treated.

- We define the **propensity score** as

$$p(\mathbf{x}) = P(w = 1|\mathbf{x}),\ \mathbf{x} \in \mathcal{X}. \tag{9}$$

- **Strong Ignorability** [Rosenbaum and Rubin (1983)] = Ignorability + Overlap.

- We now turn to identification of τ_{ate} and τ_{att}. It turns out that, under strong ignorability, we can identify the ATE conditional on $\mathbf{x}$ and therefore τ_{ate} and τ_{att}.
- In fact, for τ_{att} we can get by with a weaker version of ignorability,

$$y_0 \perp w \mid \mathbf{x}, \tag{6'}$$

which allows w to be correlated with the (unobserved) gain, $y_1 - y_0$.

- A weaker overlap assumption suffices, too:

$$p(\mathbf{x}) < 1, \mathbf{x} \in \mathcal{X}.$$

4. Identification of Average Treatment Effects

- Use two ways to show the treatment effects are identified under ignorability.
- First is based on regression functions. Define the **average treatment effect conditional on x** as

$$\tau(\mathbf{x}) = E(y_1 - y_0|\mathbf{x}) = E(y_1|\mathbf{x}) - E(y_0|\mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}).$$

- The function $\tau(\mathbf{x})$ is of interest in its own right, as it provides the mean effect for different segments of the population described by the observables, **x**.

- By iterated expectations,

$$\tau_{ate} = E(y_1 - y_0) = E[\tau(\mathbf{x})] = E[\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})]$$

It follows that τ_{ate} is identified if $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are identified because we observe a random sample on $\mathbf{x}$ and can average across its distribution.

• To see $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are identified under ignorability,

$$\begin{aligned} E(y|\mathbf{x}, w) &= (1-w)E(y_0|\mathbf{x}, w) + wE(y_1|\mathbf{x}, w) \\ &= (1-w)E(y_0|\mathbf{x}) + wE(y_1|\mathbf{x}) \\ &\equiv (1-w)\mu_0(\mathbf{x}) + w\mu_1(\mathbf{x}), \end{aligned} \tag{10}$$

where the second equality holds by ignorability and $\mu_g(\mathbf{x}) \equiv E(y_g|\mathbf{x})$, $g = 0, 1$. So

$$\begin{aligned} E(y|\mathbf{x}, w = 0) &= \mu_0(\mathbf{x}) \\ E(y|\mathbf{x}, w = 1) &= \mu_1(\mathbf{x}) \end{aligned}$$

• The functions $E(y|\mathbf{x}, w = 0)$, $E(y|\mathbf{x}, w = 1)$ are consistently estimable from the data because we have a random sample on $(y, \mathbf{x}, w)$. But overlap is critical. We need to estimate $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. But by defintion $E(y|\mathbf{x}, w = 0)$ will be estimated only using the control group and $E(y|\mathbf{x}, w = 1)$ will be estimated only using the treatment group. (More on the overlap issue when we consider estimation.)

- For ATT note that

$$E(y_1 - y_0|w) = E[E(y_1 - y_0|\mathbf{x}, w)|w] = E[E(y_1 - y_0|\mathbf{x})|w]$$
$$= E[\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})|w],$$

where the second equality holds by ignorability (in the mean), that is, $E(y_1 - y_0|\mathbf{x}, w) = E(y_1 - y_0|\mathbf{x})$.

- So

$$\tau_{att} = E[\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})|w = 1],$$

and we know $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are identified.

• Define the estimable regression functions for the control and treatment groups as

$$m_0(\mathbf{x}) = E(y|\mathbf{x}, w = 0),\ m_1(\mathbf{x}) = E(y|\mathbf{x}, w = 1). \tag{11}$$

• Under ignorability, $m_1(\mathbf{x}) = \mu_1(\mathbf{x})$ and $m_0(\mathbf{x}) = \mu_0(\mathbf{x})$. (If ignorability fails, $m_g(\mathbf{x}) \neq \mu_g(\mathbf{x})$, so it is important to generally keep these means separate.)

- In terms of the estimable mean functions,

$$\tau_{ate} = E[m_1(\mathbf{x}) - m_0(\mathbf{x})]. \tag{12}$$

$$\tau_{att} = E[m_1(\mathbf{x}) - m_0(\mathbf{x})|w = 1]. \tag{13}$$

- See the text for verification that (13) still holds under the weaker ignorability assumption $E(y_0|\mathbf{x}, w) = E(y_0|\mathbf{x})$.

• We can also establish identification using propensity score weighting. Ignorability implies that w and y_g are uncorrelated conditional on $\mathbf{x}$, and so, by iterated expectations,

$$E\left[\frac{wy}{p(\mathbf{x})}\right] = E\left[\frac{wy_1}{p(\mathbf{x})}\right] = E\left[\left(\frac{wy_1}{p(\mathbf{x})}\,\middle|\,\mathbf{x}\right)\right] \tag{14}$$
$$= E\left[\frac{E(w|\mathbf{x})E(y_1|\mathbf{x})}{p(\mathbf{x})}\right] = E[E(y_1|\mathbf{x})] \;\; = E(y_1)$$

A similar argument shows

$$E\left[\frac{(1-w)y}{1-p(\mathbf{x})}\right] = E(y_0). \tag{15}$$

• Putting the two expressions together gives

$$\tau_{ate} = E\left[\frac{wy}{p(\mathbf{x})} - \frac{(1-w)y}{1-p(\mathbf{x})}\right] = E\left\{\frac{[w-p(\mathbf{x})]y}{p(\mathbf{x})[1-p(\mathbf{x})]}\right\}. \quad (16)$$

• Clear from (16) that the overlap assumption is needed: $p(\mathbf{x})$ and $1 - p(\mathbf{x})$ must both be different from zero for all $\mathbf{x}$.

• Intuitively, if we want an average effect over the stated population, then at each $\mathbf{x}$ there must be units in the control and treatment groups.

• The text contains a more general result concerning conditional treatment effects.

• Can also show

$$\tau_{att} = E\left\{ \frac{[w - p(\mathbf{x})]y}{\rho[1 - p(\mathbf{x})]} \right\}, \tag{17}$$

where $\rho = P(w = 1)$ is the unconditional probability of treatment.

• Now, we only need to keep $p(\mathbf{x})$ away from unity. Makes intuitive sense because τ_{att} is an average effect for those eventually treated. Therefore, it does not matter if some units have no chance of being treated; they are excluded from the averaging anyway.

5. Estimating ATEs

- When we assume ignorable treatment and overlap, there are three general approaches to estimating the treatment effects (although they can be combined): (i) regression-based methods; (ii) propensity score methods; (iii) matching methods.
- Sometimes regression or matching are done on the propensity score. We will discuss the pros and cons of such methods.

• Why do many have a preference for PS methods over regression methods?

1. Estimating the PS requires only a single parametric or nonparametric estimation. Regression methods require estimation of $E(y|w = 0, \mathbf{x})$ and $E(y|w = 1, \mathbf{x})$ as well as accounting for the nature of y (continuous, discrete, some mixture?)

2. We have good binary response models for estimating $P(w = 1|\mathbf{x})$. Do not need to worry about the nature of y.

3. Simple propensity score methods have been developed that are asymptotically efficient (although the estimators may not be practically the best, or need some adjustment).
4. PS methods seem more exotic compared with regression.

Regression Adjustment

- First step is to obtain $\hat{m}_0(\mathbf{x})$ from the "control" subsample, $w_i = 0$, and $\hat{m}_1(\mathbf{x})$ from the "treated" subsample, $w_i = 1$. Can be as simple as (flexible) linear regression or full nonparametric regression.
- Compute fitted values in each case for *all* units in sample. (This is made easy in Stata using the "predict" command because a fitted value is computed for all units with nonmissing $\mathbf{x}_i$, even if a unit was not used in estimation.)

- The regression-adjustement estimates are

$$\hat{\tau}_{ate,reg} = N^{-1} \sum_{i=1}^{N} [\hat{m}_1(\mathbf{x}_i) - \hat{m}_0(\mathbf{x}_i)] = N^{-1} \sum_{i=1}^{N} \hat{\tau}(\mathbf{x}_i) \tag{18}$$

$$\hat{\tau}_{att,reg} = N_1^{-1} \sum_{i=1}^{N} w_i [\hat{m}_1(\mathbf{x}_i) - \hat{m}_0(\mathbf{x}_i)] = N_1^{-1} \sum_{i=1}^{N} w_i \hat{\tau}(\mathbf{x}_i) \tag{19}$$

where $N_1 = \sum_{i=1}^{N} w_i$ is the number of treated units.

• Notice that we must observe the same set of covariates for the treated and untreated groups. While we can think of the counterfactual setting as being a missing data problem on (y_{i0}, y_{i1}), we assume we do not have missing data on $(w_i, \mathbf{x}_i)$.

- How does overlap affect estimation of τ_{ate} and τ_{att}? Note that $\hat{\tau}_{ate,reg}$ requires two kinds of extrapolation: we must evaluate $\hat{m}_0(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}_i$ for treated i and we must evaluate $\hat{m}_1(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}_i$ for untreated i.
- When we use parametric models for $m_g(\cdot)$, extrapolation is easy. But it may be hiding a problem. The estimates of the mean functions where data are scarce may be very sensitive to functional form.
- Nonparametric methods that use local averaging will reveal overlap problems.

- Because the ATE as a function of $\mathbf{x}$ is consistently estimated by

$$\hat{\tau}_{reg}(\mathbf{x}) = \hat{m}_1(\mathbf{x}) - \hat{m}_0(\mathbf{x}),$$

we can easily estimate the ATE for subpopulations described by functions of $\mathbf{x}$.

- If there is not sufficient overlap, $\hat{\tau}_{reg}(\mathbf{x})$ can be a poor estimator for certain values of $\mathbf{x}$.

- Let $\mathcal{R} \subset \mathcal{X}$ be a subset of the possible values of $\mathbf{x}$. We can estimate

$$\tau_{ate,\mathcal{R}} = E(y_1 - y_0 | \mathbf{x} \in \mathcal{R})$$

as

$$\hat{\tau}_{ate,\mathcal{R}} = N_{\mathcal{R}}^{-1} \sum_{\mathbf{x}_i \in \mathcal{R}} [\hat{m}_1(\mathbf{x}_i) - \hat{m}_0(\mathbf{x}_i)]$$

where $N_{\mathcal{R}}$ is the number of observations with $\mathbf{x}_i \in \mathcal{R}$.

- If both functions are linear, so $\hat{m}_g(\mathbf{x}) = \hat{\alpha}_g + \mathbf{x}\hat{\boldsymbol{\beta}}_g$ for $g = 0, 1$, then

$$\hat{\tau}_{ate,reg} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \bar{\mathbf{x}}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0), \tag{20}$$

where $\bar{\mathbf{x}}$ is the row vector of sample averages. (To get the ATE, average any nonlinear functions in $\mathbf{x}$, rather than inserting the averages into the nonlinear functions.)

• Easiest way to obtain standard error for $\hat{\tau}_{ate,reg}$ is to ignore sampling error in $\bar{\mathbf{x}}$ and use the coefficient on w_i in the regression

$$y_i \text{ on } 1,\ w_i,\ \mathbf{x}_i,\ w_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}}),\ \ i = 1, \ldots, N. \tag{21}$$

$\hat{\tau}_{ate,reg}$ is the coefficient on w_i.

• Accounting for the sampling error in $\bar{\mathbf{x}}$ [as an estimator of $\boldsymbol{\mu}_{\mathbf{x}} = E(\mathbf{x})$] is possible, too, but unlikely to matter much.

- Note how $\mathbf{x}_i$ is demeaned before forming interaction. This is critical because we want to estimate $\tau_{ate} = (\alpha_1 - \alpha_0) + \boldsymbol{\mu}_{\mathbf{x}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$, not $\alpha_1 - \alpha_0$ (unless we impose $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$).
- Demeaning the covariates before constructing the interactions is known to "solve" the multicollinearity problem in regression. But it "solves" the problem because it redefines the parameter we are trying to estimate, and we can more easily estimate an ATE than the treatment effect at $\mathbf{x} = \mathbf{0}$ which is only of interest in special cases.

- The linear regression estimate of τ_{att} is

$$\hat{\tau}_{att,reg} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \bar{\mathbf{x}}_1(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)$$

where $\bar{\mathbf{x}}_1$ is the average of the $\mathbf{x}_i$ over the treated subsample. $\hat{\tau}_{att,reg}$ can be close to $\hat{\tau}_{ate,reg}$ if (1) $\hat{\boldsymbol{\beta}}_1 \approx \hat{\boldsymbol{\beta}}_0$ or (2) $\bar{\mathbf{x}} \approx \bar{\mathbf{x}}_1$.

• More generally, if we want to use linear regression to estimate $\hat{\tau}_{ate,\mathcal{R}} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \bar{\mathbf{x}}_{\mathcal{R}}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)$, where $\bar{\mathbf{x}}_{\mathcal{R}}$ is the average over some subset of the sample, then the regression

$$y_i \text{ on } 1,\ w_i,\ \mathbf{x}_i,\ w_i \cdot (\mathbf{x}_i - \bar{\mathbf{x}}_{\mathcal{R}}),\ \ i = 1, \ldots, N$$

can be used. Note that it uses all the data to estimate the parameters; it simply centers about $\bar{\mathbf{x}}_{\mathcal{R}}$ rather than $\bar{\mathbf{x}}$.

• If common slopes are imposed, $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_0$, $\hat{\tau}_{ate,reg} = \hat{\tau}_{att,reg}$ is just the coefficient on w_i from the regression across all observations:

$$y_i \text{ on } 1,\ w_i,\ \mathbf{x}_i,\ \ i = 1,\ldots,N. \tag{22}$$

• If linear models do not seem appropriate for $E(y_0|\mathbf{x})$ and $E(y_1|\mathbf{x})$, we can exploit the specific nature of y_g.

• If y is a binary response, or a fractional response, estimate logit or probit separately for the $w_i = 0$ and $w_i = 1$ subsamples and average differences in predicted values:

$$\hat{\tau}_{ate,reg} = N^{-1} \sum_{i=1}^{N} [G(\hat{\alpha}_1 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_1) - G(\hat{\alpha}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_0)]. \tag{23}$$

• Each summand in (23) is the difference in estimate probabilities under treatment and nontreatment for unit i, and the ATE just averages those differences. Use the same approach even if $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_0$ is imposed.

• For general $y \geq 0$, Poisson or gamma regression with exponential mean is attractive:

$$\hat{\tau}_{ate,reg} = N^{-1} \sum_{i=1}^{N} [\exp(\hat{\alpha}_1 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_1) - \exp(\hat{\alpha}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_0)]. \tag{24}$$

• In nonlinear cases, can use delta method or bootstrap to get $se(\hat{\tau}_{ate,reg})$.

• General formula for asymptotic variance of $\hat{\tau}_{ate,reg}$ in the parametric case. Let $m_0(\cdot, \boldsymbol{\delta}_0)$ and $m_1(\cdot, \boldsymbol{\delta}_1)$ be general parametric models of $\mu_0(\cdot)$ and $\mu_1(\cdot)$; as a practical matter, m_0 and m_1 would have the same structure but with different parameters. Assuming that we have consistent, $\sqrt{N}$-asymptotically normal estimators $\hat{\boldsymbol{\delta}}_0$ and $\hat{\boldsymbol{\delta}}_1$,

$$\hat{\tau}_{ate,reg} = N^{-1} \sum_{i=1}^{N} [m_1(\mathbf{x}_i, \hat{\boldsymbol{\delta}}_1) - m_0(\mathbf{x}_i, \hat{\boldsymbol{\delta}}_0)]$$

will be such that $Avar\sqrt{N}(\hat{\tau}_{ate,reg} - \tau_{ate})$ is asymptotically normal with zero mean.

• Using Wooldridge (2010, Problem 12.17), it can be shown that

$$\begin{aligned} Avar\sqrt{N}(\hat{\tau}_{ate,reg} - \tau_{ate}) = {} & E\{[m_1(\mathbf{x}_i, \boldsymbol{\delta}_1) - m_0(\mathbf{x}_i, \boldsymbol{\delta}_0) - \tau_{ate}]^2\} \\ & + E[\nabla_{\boldsymbol{\delta}_0} m_0(\mathbf{x}_i, \boldsymbol{\delta}_0)]\mathbf{V}_0 E[\nabla_{\boldsymbol{\delta}_0} m_0(\mathbf{x}_i, \boldsymbol{\delta}_0)]' \\ & + E[\nabla_{\boldsymbol{\delta}_1} m_1(\mathbf{x}_i, \boldsymbol{\delta}_1)]\mathbf{V}_1 E[\nabla_{\boldsymbol{\delta}_1} m_1(\mathbf{x}_i, \boldsymbol{\delta}_1)]', \end{aligned}$$

where $\mathbf{V}_0$ is the asymptotic variance of $\sqrt{N}(\hat{\boldsymbol{\delta}}_0 - \boldsymbol{\delta}_0)$ and similarly for $\mathbf{V}_1$.

• Clearly better to use more efficient estimators of $\boldsymbol{\delta}_0$ and $\boldsymbol{\delta}_1$ as that makes the quadratic forms smaller.

• Each of the quantities above is easy to estimate by replacing expectations with sample averages and replacing unknown parameters with estimates:

$$N \cdot \widehat{Avar(\hat{\tau}_{ate,reg})} = N^{-1} \sum_{i=1}^{N} [m_1(\mathbf{x}_i, \hat{\boldsymbol{\delta}}_1) - m_0(\mathbf{x}_i, \hat{\boldsymbol{\delta}}_0) - \hat{\tau}_{ate,reg}]^2$$

$$+ \left[N^{-1} \sum_{i=1}^{N} \nabla_{\boldsymbol{\delta}_0} m_0(\mathbf{x}_i, \hat{\boldsymbol{\delta}}_0) \right] \hat{\mathbf{V}}_0 \left[N^{-1} \sum_{i=1}^{N} \nabla_{\boldsymbol{\delta}_0} m_0(\mathbf{x}_i, \hat{\boldsymbol{\delta}}_0) \right]'$$

$$+ \left[N^{-1} \sum_{i=1}^{N} \nabla_{\boldsymbol{\delta}_1} m_1(\mathbf{x}_i, \hat{\boldsymbol{\delta}}_1) \right] \hat{\mathbf{V}}_1 \left[N^{-1} \sum_{i=1}^{N} \nabla_{\boldsymbol{\delta}_1} m_1(\mathbf{x}_i, \hat{\boldsymbol{\delta}}_1) \right]'$$

• The first part of the asymptotic variance formula would result in the naive standard error treating $\{\hat{m}_{i1} - \hat{m}_{i0} : i = 1, 2, \ldots, N\}$ as a random sample of data, ignoring the estimation of $\hat{\boldsymbol{\delta}}_0$ and $\hat{\boldsymbol{\delta}}_1$. The second and third terms account for the sampling errors in $\hat{\boldsymbol{\delta}}_0$ and $\hat{\boldsymbol{\delta}}_1$.

• Regardless of the mean function, without good overlap in the covariate distribution, we must extrapolate a parametric model – linear or nonlinear – into regions where we do not have much or any data. For example, suppose, after defining the population of interest for the effects of job training, those with better labor market histories are unlikely to be treated. Then, we have to estimate $E(y|\mathbf{x}, w = 1)$ only using those who participated – where **x** includes variables measuring labor market history – and then extrapolate this function to those who did not participate. This leads to sensitive estimates if nonparticipants have very different values of **x**.

- Nonparametric methods are not helpful in overcoming poor overlap because they are either based on flexible parametric models (and so require extrapolation) or use local averaging (in which case we cannot estimate $m_1(\mathbf{x})$ for $\mathbf{x}$ values far away from those in the treated subsample).
- The most common local smoothing method, based on kernel estimation, would at least let you know there is very little data to estimate the regression function for values of $\mathbf{x}$ with poor overlap.

• Using τ_{att} has advantages because its estimation requires only one extrapolation:

$$\hat{\tau}_{att,reg} = N_1^{-1} \sum_{i=1}^{N} w_i [\hat{m}_1(\mathbf{x}_i) - \hat{m}_0(\mathbf{x}_i)].$$

Therefore, we only need to estimate $m_1(\mathbf{x})$ for values of $\mathbf{x}$ taken on by the treated group, which we can do well. Unlike with the ATE, we do not need to estimate $m_1(\mathbf{x})$ for values of $\mathbf{x}$ in the untreated group. But we need to estimate $\hat{m}_0(\mathbf{x}_i)$ for treated individuals i, and this can be difficult if we have units in the treated group very different from all units in the control group.

- A “solution” is to “balance” the sample by dropping observations that are either very unlikely or very likely to receive treatment, based on the values of **x**. This is often done based on the propensity score, which we cover below. This effectively changes the population that we are studying.
- It also makes sense to think more carefully about the population ahead of time. If high earners are not going to be eligible for job training, why include them in the analysis at all? The notion of a population is not immutable.

Should We use Regression Adjustment with Randomized Assignment?

- If the treatment w_i is independent of (y_{i0}, y_{i1}), then we know that the simply difference in means is an unbiased and consistent estimator of $\tau_{ate} = \tau_{att}$. But if we have covariates, should we add them to the regression?
- If we focus on large-sample analysis, the answer is yes, provided the covariates help to predict (y_{i0}, y_{i1}). Remember, randomized assignment means w_i is also independent of $\mathbf{x}_i$.

• Consider the case where the treatment effect is constant, so $y_{i1} - y_{i0} = \tau$ for all i. Then we can write

$$y_i = y_{i0} + \tau w_i \equiv \mu_0 + \tau w_i + v_{i0}$$

and w_i is independent of y_{i0} and therefore v_{i0}.

• Simple regression of y_i on $1, w_i$ is unbiased and consistent for τ.

• But writing the linear projection

$$y_{i0} = \alpha_0 + \mathbf{x}_i\boldsymbol{\beta}_0 + u_{i0}$$
$$E(u_{i0}) = 0,\ E(\mathbf{x}_i'u_{i0}) = \mathbf{0}$$

we have

$$y_i = \alpha_0 + \tau w_i + \mathbf{x}_i\boldsymbol{\beta}_0 + u_{i0}$$

where, by randomized assignment, w_i is uncorrelated with $\mathbf{x}_i$ and u_{i0}.

So multiple regression is consistent for τ. If $\boldsymbol{\beta}_0 \neq \mathbf{0}$,

$Var(u_{i0}) < Var(v_{i0})$, and so adding $\mathbf{x}_i$ reduces the error variance.

• Under the constant treatment effect assumption and random assignment, the asymptotic variances of the simple and multiple regression estimators are, respectively,

$$\frac{Var(v_{i0})}{N\rho(1-\rho)}, \frac{Var(u_{i0})}{N\rho(1-\rho)}$$

where $\rho = P(w_i = 1)$.

• The only caveat is that if $E(y_{i0}|\mathbf{x}) \neq \alpha_0 + \mathbf{x}_i\boldsymbol{\beta}_0$, the OLS estimator of τ is only guaranteed to be consistent, not unbiased. This distinction can be relevant in small samples (as often occurs in true experiments).

- With nonconstant treament effect, add the linear projecton

$y_{i1} = \alpha_1 + \mathbf{x}_i\boldsymbol{\beta}_1 + u_{i1}$, so that $\tau_{ate} = \tau = (\alpha_1 - \alpha_0) + \boldsymbol{\mu}_{\mathbf{x}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$.

- Now we can write

$$\begin{aligned} y_i &= \alpha_0 + \tau w_i + \mathbf{x}_i\boldsymbol{\beta}_0 + (\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}})(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0) + u_{i0} + w_i(u_{i1} - u_{i0}) \\ &\equiv \alpha_0 + \tau w_i + \mathbf{x}_i\boldsymbol{\beta}_0 + w_i \cdot (\mathbf{x}_i - \boldsymbol{\mu}_{\mathbf{x}})\boldsymbol{\delta} + u_{i0} + w_i e_i \end{aligned}$$

with $\boldsymbol{\delta} \equiv \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$ and $e_i \equiv u_{i1} - u_{i0}$.

• Under random assignment of treatment, $(e_i, \mathbf{x}_i)$ is independent of w_i, so w_i is uncorrelated with all other terms in the equation. OLS is consistent for τ but it is generally biased unless the equation represents $E(y_i|w_i, \mathbf{x}_i)$.

• Further,

$$E(\mathbf{x}_i' w_i e_i) = E(w_i)E(\mathbf{x}_i' e_i) = \mathbf{0}$$

and so $\mathbf{x}_i$ and $w_i \cdot (\mathbf{x}_i - \boldsymbol{\mu}_\mathbf{x})$ are uncorrelated with $u_{i0} + w_i e_i$ (and this term has zero mean). So OLS consistently estimates all parameters: α_0, τ, $\boldsymbol{\beta}_0$, and $\boldsymbol{\delta}$.

• As a bonus from including covariates interacted with the treatment, we can estimate ATEs as a function of $\mathbf{x}$:

$$\hat{\tau}(\mathbf{x}) = \hat{\tau} + (\mathbf{x} - \bar{\mathbf{x}})\hat{\boldsymbol{\delta}}.$$

• If the $E(y_g|\mathbf{x})$ are not linear, $\hat{\tau}(\mathbf{x})$ is not a consistent estimator of $\tau(\mathbf{x}) = E(y_1 - y_0|\mathbf{x})$, but it should be a reasonable approximation in many cases.

Propensity Score Weighting

• The formula that establishes identification of τ_{ate} base on population moments suggests an imediate estimator of τ_{ate}:

$$\tilde{\tau}_{ate,psw} = N^{-1} \sum_{i=1}^{N} \left[\frac{w_i y_i}{p(\mathbf{x}_i)} - \frac{(1 - w_i) y_i}{1 - p(\mathbf{x}_i)} \right]. \tag{25}$$

• $\tilde{\tau}_{ate,psw}$ is not feasible because it depends on the propensity score $p(\cdot)$.

• Interestingly, we would not use it if we could! Even if we know $p(\cdot)$, $\tilde{\tau}_{ate,psw}$ is not asymptotically efficient. It is *better* to estimate the propensity score!

• Two approaches: (1) Model $p(\cdot)$ parametrically, in a flexible way. Can show estimating the propensity score leads to a *smaller* asymptotic variance when the parametric model is correctly specified. (2) Use an explicit nonparametric approach, as in Hirano, Imbens, and Ridder (2003, *Econometrica*) or Li, Racine, and Wooldridge (2009, *JBES*).

$$\hat{\tau}_{ate,psw} = N^{-1}\sum_{i=1}^{N}\left[\frac{w_i y_i}{\hat{p}(\mathbf{x}_i)} - \frac{(1-w_i)y_i}{1-\hat{p}(\mathbf{x}_i)}\right] = N^{-1}\sum_{i=1}^{N}\frac{[w_i - \hat{p}(\mathbf{x}_i)]y_i}{\hat{p}(\mathbf{x}_i)[1-\hat{p}(\mathbf{x}_i)]}. \quad (26)$$

• Very simple to compute given $\hat{p}(\cdot)$.

- τ_{att} is estimated using identical reasoning:

$$\hat{\tau}_{att,psw} = N^{-1} \sum_{i=1}^{N} \frac{[w_i - \hat{p}(\mathbf{x}_i)]y_i}{\hat{\rho}[1 - \hat{p}(\mathbf{x}_i)]}, \tag{27}$$

where $\hat{\rho} = (N_1/N)$ is the fraction of treated in the sample.

• To exploit estimation error in $\hat{p}(\mathbf{x})$ for reducing the asymptotic variance of $\hat{\tau}_{ate,psw}$, write

$$\hat{\tau}_{ate,psw} \equiv N^{-1} \sum_{i=1}^{N} \hat{k}_i \tag{28}$$

where

$$\hat{k}_i \equiv \frac{[w_i - \hat{p}(\mathbf{x}_i)]y_i}{\hat{p}(\mathbf{x}_i)[1 - \hat{p}(\mathbf{x}_i)]}.$$

• The adjustment for estimating $\boldsymbol{\gamma}$ by MLE is a regression "netting out" of the score for the binary choice MLE. Let

$$\hat{\mathbf{d}}_i = \mathbf{d}(w_i, \mathbf{x}_i, \hat{\boldsymbol{\gamma}}) = \frac{\nabla_\gamma p(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})'[w_i - p(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})]}{p(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})[1 - p(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})]} \tag{29}$$

be the score for the propensity score binary response estimation. Let $\hat{e}_i$ be the OLS residuals from the regression

$$\hat{k}_i \text{ on } 1, \ \hat{\mathbf{d}}_i', \ i = 1, \ldots, N. \tag{30}$$

• Then the asymptotic standard error of $\hat{\tau}_{ate,psw}$ is

$$\left[N^{-1} \sum_{i=1}^{N} \hat{e}_i^2 \right]^{1/2} \Big/ \sqrt{N}. \tag{31}$$

This follows from Wooldridge (2007, *Journal of Econometrics*).

- For logit PS, estimation,

$$\hat{\mathbf{d}}_i' = \mathbf{x}_i(w_i - \hat{p}_i) \tag{32}$$

where $\mathbf{x}_i$ is the $1 \times R$ vector of covariates (including unity) and $\hat{p}_i = \Lambda(\mathbf{x}_i\hat{\boldsymbol{\gamma}}) = \exp(\mathbf{x}_i\hat{\boldsymbol{\gamma}})/[1 + \exp(\mathbf{x}_i\hat{\boldsymbol{\gamma}})]$.

• As noted by Robins and Rotnitzky (1995, JASA), one never does worse by adding functions of $\mathbf{x}_i$ to the PS model, even if they do not predict treatment! If the functions are correlated with

$$k_i = \frac{[w_i - p(\mathbf{x}_i)]y_i}{p(\mathbf{x}_i)[1 - p(\mathbf{x}_i)]},$$

including them in the logit reduces the error variance in e_i.

• Hirano, Imbens, and Ridder (2003) show that the efficient estimator keeps adding terms as the sample size grows – that is, when we think of the PS estimation as being nonparametric.

• A straightforward alternative is to use bootstrapping, where the binary response estimation and averaging (to get $\hat{\tau}_{ate,psw}$) are included in each bootstrap iteration.

• It is conservative to ignore the estimation error in the $\hat{k}_i$ and simply treat it as randomly sampled data. Then, just compute the standard error for a sample average:

$$se(\hat{\tau}_{ate,psw}) = \left[N^{-1} \sum_{i=1}^{N} (\hat{k}_i - \hat{\tau}_{ate,psw})^2 \right]^{1/2} \Big/ \sqrt{N}.$$

This is always larger than (31) and is gotten by the regression $\hat{k}_i$ on 1.

- Similar remarks hold for $\hat{\tau}_{att,psw}$; adjustment to standard error somewhat different. See text.

- Can see directly from $\hat{\tau}_{ate,psw}$ and $\hat{\tau}_{att,psw}$ that the inverse probability weighted (IPW) estimators can be very sensitive to extreme values of $\hat{p}(\mathbf{x}_i)$. $\hat{\tau}_{att,psw}$ is sensitive only to $\hat{p}(\mathbf{x}_i) \approx 1$, but $\hat{\tau}_{ate,psw}$ is also sensitive to $\hat{p}(\mathbf{x}_i) \approx 0$.
- Imbens and coauthors have provided a rule-of-thumb: only use observations with $.10 \leq \hat{p}(\mathbf{x}_i) \leq .90$ (for ATE).
- Sometimes the problem is $\hat{p}(\mathbf{x}_i)$ "close" to zero for many units, which suggests the original population was not carefully chosen.

- After using the PS to choose a new “population,” redo the analysis (regression, matching, or PS weighting) where all estimates are based on the new, smaller sample. Of course, because the PS has been estimated, our new “population” is depends on the sample from the original population.

Regression on the Propensity Score

• The motivation is that one can show, given ignorability, that ignorability actually holds conditional only on $p(\mathbf{x})$:

$$(y_0, y_1) \perp w \mid p(\mathbf{x}),$$

which, of course, implies

$$E[y_g | p(\mathbf{x}), w] = E[y_g | p(\mathbf{x})], g = 0, 1.$$

• In other words, it is sufficient to condition only on the propensity score so break the dependence between w and (y_0, y_1). We need not condition on $\mathbf{x}$.

- By iterated expectations,

$$\tau_{ate} = E(y_1 - y_0) = E\{E[y_1|p(\mathbf{x})] - E[y_0|p(\mathbf{x})]\}.$$

- Now we can obtain a conditional expectation for the observable y:

$$\begin{aligned} E[y|p(\mathbf{x}), w] &= (1-w)E[y_0|p(\mathbf{x}), w] + wE[y_1|p(\mathbf{x}), w] \\ &= (1-w)E[y_0|p(\mathbf{x})] + wE[y_1|p(\mathbf{x})] \end{aligned}$$

where the second equality follows by the previous result.

- We have shown that

$$E[y|p(\mathbf{x}), w = 0] = E[y_0|p(\mathbf{x})]$$
$$E[y|p(\mathbf{x}), w = 1] = E[y_1|p(\mathbf{x})]$$

- So, after estimating $p(\mathbf{x})$ using, say, flexible logit, we estimate $E[y|p(\mathbf{x}), w = 0]$ and $E[y|p(\mathbf{x}), w = 1]$ using the subsamples of nontreated and treated, respectively. Could use nonparametric methods.

- In the linear case, $E[y_g|p(\mathbf{x})] = \alpha_g + \gamma_1 p(\mathbf{x})$, $g = 0, 1$, and we use

$$y_i \text{ on } 1, \hat{p}(\mathbf{x}_i) \text{ for } w_i = 0 \text{ and } y_i \text{ on } 1, \hat{p}(\mathbf{x}_i) \text{ for } w_i = 1, \tag{33}$$

which gives fitted values $\hat{\alpha}_0 + \hat{\gamma}_0 \hat{p}(\mathbf{x}_i)$ and $\hat{\alpha}_1 + \hat{\gamma}_1 \hat{p}(\mathbf{x}_i)$, respectively.

- A consistent estimator of τ_{ate} is

$$\hat{\tau}_{ate,regps} = N^{-1} \sum_{i=1}^{N} [(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\gamma}_1 - \hat{\gamma}_0)\hat{p}(\mathbf{x}_i)]. \tag{34}$$

• Conservative inference: ignore estimation of the propensity score.

Same as using usual statistics on w_i in the regression

$$y_i \text{ on } 1, w_i, \hat{p}(\mathbf{x}_i), w_i \cdot [\hat{p}(\mathbf{x}_i) - \hat{\mu}_{\hat{p}}], i = 1, \ldots, N \tag{35}$$

where $\hat{\mu}_{\hat{p}} = N^{-1} \sum_{i=1}^{N} \hat{p}(\mathbf{x}_i)$. Or, use bootstrap, which will provide the smaller (valid) standard errors.

• Somewhat more common (and less desirable) is to drop the interaction term.

$$y_i \text{ on } 1, \; w_i, \; \hat{p}(\mathbf{x}_i), \; i = 1, \ldots, N. \tag{36}$$

• Because $0 < p(\mathbf{x}) < 1$, linearity of $E[y_g|p(\mathbf{x})]$ can be unrealistic. For a better fit, might use functions of the log-odds ratio,

$$\hat{r}_i \equiv \log\left[\frac{\hat{p}(\mathbf{x}_i)}{1-\hat{p}(\mathbf{x}_i)}\right],$$

as regressors when y has a wide range. So, regress y_i on $1, \hat{r}_i, \hat{r}_i^2, \ldots, \hat{r}_i^Q$ for some Q using both the control and treated samples, and then average the difference in fitted values to obtain $\hat{\tau}_{ate,regprop}$.

• On balance, regression on the propensity score (or functions of it) has little to offer compared with weighting by the propensity score, provided the overlap issue is attended to. The PS weighted estimator does not require us to model $E[y_g|p(\mathbf{x})]$, and PS weighting can be asymptotically efficient.

Matching

- Matching estimators are based on imputing a value on the counterfactual outcome for each unit. That is, for a unit i in the control group, we observe y_{i0}, but we need to impute y_{i1}. For each unit i in the treatment group, we observe y_{i1} but need to impute y_{i0}.
- For τ_{ate}, matching estimators take the general form

$$\hat{\tau}_{ate,match} = N^{-1} \sum_{i=1}^{N} (\hat{y}_{i1} - \hat{y}_{i0})$$

- Looks like regression adjustment but the imputed values are not fitted values from regression.

- For τ_{att},

$$\hat{\tau}_{att,match} = N_1^{-1} \sum_{i=1}^{N} w_i (y_i - \hat{y}_{i0})$$

where this form uses the fact that y_{i1} is always observed for the treated subsample. (In other words, we never need to impute y_{i1} for the treated subsample.)

• Abadie and Imbens (2006, Econometrica) consider several approaches. The simplest is to find a single match for each observation. Suppose i is a treated observation ($w_i = 1$). Then $\hat{y}_{i1} = y_i, \hat{y}_{i0} = y_h$ for h such that $w_h = 0$ and unit h is "closest" to unit i based on some metric (distance) in the covariates. In other words, for the treated unit i we find the "most similar" untreated observation, and use its response as y_{i0}. Similarly, if $w_i = 0$, $\hat{y}_{i0} = y_i$, $\hat{y}_{i1} = y_h$ where now $w_h = 1$ and $\mathbf{x}_h$ is "closest" to $\mathbf{x}_i$.

• Abadie and Imbens matching has been programmed in Stata in the command "nnmatch." The default is to use the single nearest neighbor.

• The default matrix in defining distance is the inverse of the diagonal matrix with sample variances of the covariates on the diagonal. [That is, diagonal Mahalanobis.]

• More generally, we can impute the missing values using an average of M nearest neighbors. If $w_i = 1$ then

$$\hat{y}_{i1} = y_i$$
$$\hat{y}_{i0} = M^{-1} \sum_{h \in \aleph_M(i)} y_h$$

where $\aleph_M(i)$ contains the M untreated nearest matches to observation i, based on the covariates. So for all $h \in \aleph_M(i)$, $w_h = 0$.

- With ties, there can be more than M elements in $\aleph_M(i)$, and then M is replaced with the number of elements in $\aleph_M(i)$.
- Similarly, if $w_i = 0$,

$$\begin{aligned}\hat{y}_{i0} &= y_i \\ \hat{y}_{i1} &= M^{-1} \sum_{h \in \Im_M(i)} y_h\end{aligned}$$

where $\Im_M(i)$ contains the M treated nearest matches to observation i.

• Remarkably, in the general M case can write the matching estimator as

$$\hat{\tau}_{ate,match} = N^{-1} \sum_{i=1}^{N} (2w_i - 1)[1 + K_M(i)]y_i,$$

where $K_M(i)$ is the number of times observation i is used as a match. (See Abadie and Imbens.)

• $K_M(i)$ is a function of the data on $(w, \mathbf{x})$, which is important for variance calculations. Under ignorability, $(w, \mathbf{x})$ are effectively "exogenous."

- The conditional variance of the matching estimator is

$$Var(\hat{\tau}_{ate,match}|\mathbf{W},\mathbf{X}) = N^{-2}\sum_{i=1}^{N}[\{(2w_i-1)[1+K_M(i)]\}^2 \cdot Var(y_i|,w_i,\mathbf{x}_i)].$$

- The unconditional variance is more complicated because of a conditional bias (see Abadie and Imbens), but estimators are programmed in nnmatch.

• For the conditional variance, need to "estimate" $Var(y_i|, w_i, \mathbf{x}_i)$, but they do not have to be good pointwise estimates. (Analagous to the situation with heteroskedasticity-robust variance matrix estimator.)

• Could use models for $E(y|w, \mathbf{x})$ and $Var(y|w, \mathbf{x})$ that exploit the nature of y. This is against the spirit of matching, which does not require parametric mean or variance assumptions. But we would only be doing it to get a standard error; we still use the matching estimator.

- AI suggest a nonparametric estimator:

$$\widehat{Var(y_i|, w_i, \mathbf{x}_i)} = (y_i - y_{h(i)})^2/2$$

where $h(i)$ is the closest match to observation i with $w_{h(i)} = w_i$. (That is, we now match within treatment group.)

• There is a subtle point in all this. The variance matrix estimator is actually for the This is actually the variance estimator for the *sample average treatment effect*, τ_{sate}, which is

$$\tau_{sate} = N^{-1} \sum_{i=1}^{N} (y_{i1} - y_{i0})$$

• Notice that τ_{sate} is not a population parameter; it changes across random samples. But the estimator of τ_{ate} and τ_{sate} are the same. The way we estimate the asymptotic variance depends on τ_{ate} versus τ_{sate}.

- The matching estimators have a large-sample bias if $\mathbf{x}_i$ has dimension greater than one. The estimator is not $\sqrt{N}$-consistent. Computation is an issue when the dimension of $\mathbf{x}_i$ is even moderate.
- It is also possible to match on the estimated propensity score. This is computationally easier because it is a single variable with range in $(0, 1)$.

- Matching without smoothing using estimated propensity scores does not produce valid inference. The technical problem is that matching (without smoothing) is not smooth in $\hat{p}(\mathbf{x}_i)$. If $\hat{p}(\mathbf{x}_i)$ increases a little, that can change the match.
- One can use kernel smoothing and then apply the bootstrap. Stata's command "psmatch2" allows this, along with a variety of other options.
- No optimality results are known for PS matching, but it is simple and fairly common.

6. Combining Regression Adjustment and PS Weighting

- First consider regression adjustment combined with PS weighting.

Why should we use a combined method?

- Answer: With **x** having large dimension, still common to rely on parametric methods for regression and PS estimation. Even if we make functional forms flexible, still might worry about misspecification.

• Idea: Let $m_0(\cdot, \boldsymbol{\delta}_0)$ and $m_1(\cdot, \boldsymbol{\delta}_1)$ be parametric functions for $E(y_g|\mathbf{x}), g = 0, 1$. Let $p(\cdot, \boldsymbol{\gamma})$ be a parametric model for the propensity score. In the first step we estimate $\boldsymbol{\gamma}$ by Bernoulli maximum likelihood and obtain the estimated propensity scores as $p(\mathbf{x}_i, \hat{\boldsymbol{\gamma}})$ (probably logit or probit).

• In the second step, we use regression or a quasi-likelihood method, where we weight by the inverse probability. For example, to estimate $\boldsymbol{\delta}_1 = (\alpha_1, \boldsymbol{\beta}_1')'$, we might solve the weighted linear least squares problem

$$\min_{\alpha_1, \beta_1} \sum_{i=1}^{N} w_i (y_i - \alpha_1 - \mathbf{x}_i \boldsymbol{\beta}_1)^2 / p(\mathbf{x}_i, \hat{\boldsymbol{\gamma}}); \tag{37}$$

for $\boldsymbol{\delta}_0$, we weight by $1/[1 - \hat{p}(\mathbf{x}_i)]$ and use the $w_i = 0$ sample.

- ATE is estimated as

$$\hat{\tau}_{ate,pswreg} = N^{-1} \sum_{i=1}^{N} [(\hat{\alpha}_1 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_1) - (\hat{\alpha}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_0)]. \quad (38)$$

- Same as regression adjustment, but different estimates of $\alpha_g, \boldsymbol{\beta}_g$!

- Scharfstein, Rotnitzky, and Robins (1999, JASA) showed that $\hat{\tau}_{ate,psreg}$ has a "double robustness" property: only one of the models [mean or propensity score] needs to be correctly specified *provided* the the mean and objective function are properly chosen [see Wooldridge (2007, Journal of Econometrics)].
- y_g continuous, negative and positive values: linear mean, least squares objective function, as above.
- y_g binary or fractional: logit mean (not probit!), Bernoulli quasi-log likelihood.

$$\min_{\alpha_1,\boldsymbol{\beta}_1} \sum_{i=1}^{N} w_i\{(1-y_i)\log[1-\Lambda(\alpha_1+\mathbf{x}_i\boldsymbol{\beta}_1)] \\ +y_i\log[\Lambda(\alpha_1+\mathbf{x}_i\boldsymbol{\beta}_1)]\}/p(\mathbf{x}_i,\hat{\boldsymbol{\gamma}}). \tag{39}$$

- That is, probably use logit for w_i and y_i (for each subset, $w_i = 0$ and $w_i = 1$).
- The ATE is estimated as before:

$$\hat{\tau}_{ate,pswreg} = N^{-1}\sum_{i=1}^{N}[\Lambda(\hat{\alpha}_1+\mathbf{x}_i\hat{\boldsymbol{\beta}}_1)-\Lambda(\hat{\alpha}_0+\mathbf{x}_i\hat{\boldsymbol{\beta}}_0)].$$

If $E(y_g|\mathbf{x}) = \Lambda(\alpha_g+\mathbf{x}\boldsymbol{\beta}_g)$, $g = 0,1$ or $P(w=1|\mathbf{x}) = p(\mathbf{x},\boldsymbol{\gamma})$, then $\hat{\tau}_{ate,pswreg} \xrightarrow{p} \tau_{ate}$.

- Of course, if we want $\tau(\mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x})$, then the conditional mean models must be correctly specified. But the approximation may be good under misspecification.

• y_g nonnegative, including count, continuous, or corners at zero: exponential mean, Poisson QLL.

• In each case, must include a constant in the index models for $E(y|w, \mathbf{x})$!

• Asymptotic standard error for $\hat{\tau}_{ate,pswreg}$: bootstrapping is easiest.

• How does the double robustness result work? Consider the linear case where $m_g(\mathbf{x}, \boldsymbol{\delta}_g) = \alpha_g + \mathbf{x}\boldsymbol{\beta}_g$. Now, if we write

$$y_g = \alpha_g + \mathbf{x}\boldsymbol{\beta}_g + u_g$$
$$E(u_g) = 0, E(\mathbf{x}'u_g) = \mathbf{0}$$

then we know

$$\mu_g \equiv E(y_g) = \alpha_g + E(\mathbf{x})\boldsymbol{\beta}_g \equiv \alpha_g + \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\beta}_g$$

- If we have consistent estimators of α_g and $\boldsymbol{\beta}_g$ then a consistent estimator of μ_g is

$$\hat{\mu}_g \equiv N^{-1} \sum_{i=1}^{N} (\hat{\alpha}_g + \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}_g).$$

- Of course, if $E(y_g|\mathbf{x}) = \alpha_g + \mathbf{x}\boldsymbol{\beta}_g$ then $L(y_g|1, \mathbf{x}) = \alpha_g + \mathbf{x}\boldsymbol{\beta}_g$.

• So, the key is determining why combining regression and PSW allows us to consistently estimate $(\alpha_g, \boldsymbol{\beta}_g)$, $g = 0, 1$ when

$$E(y_g|\mathbf{x}) = \alpha_g + \mathbf{x}\boldsymbol{\beta}_g, g = 0, 1$$

or

$$P(w = 1|\mathbf{x}) = p(\mathbf{x}, \boldsymbol{\gamma})$$

(or, of course, both).

• Suppose first that the conditional means are indeed linear. Then we know from our general treatment of PS weighting that we can weight by any positive function of $\mathbf{x}_i$ and the estimator is still consistent for $(\alpha_g, \boldsymbol{\beta}_g)$. It does not matter that our model for $P(w = 1|\mathbf{x})$ is misspecified; we just need to ensure that $p(\mathbf{x}, \boldsymbol{\gamma}^*) > 0$ for all $\mathbf{x} \in \mathcal{X}$ where $\boldsymbol{\gamma}^* \equiv \text{plim}(\hat{\boldsymbol{\gamma}})$. This is the first half of "double robustness."

• Now suppose that $P(w = 1|\mathbf{x}) = p(\mathbf{x}, \boldsymbol{\gamma})$ but the conditional means are not linear: $E(y_g|\mathbf{x}) \neq \alpha_g + \mathbf{x}\boldsymbol{\beta}_g$. We know that with the selection probabilities correctly specified that IPW consistently estimates the linear projection parameters in $L(y_g|1, \mathbf{x}) = \alpha_g + \mathbf{x}\boldsymbol{\beta}_g$ (because these parameters solve the population least squares problem).

• So the estimator of τ based on $\tau = (\alpha_1 + \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\beta}_1) - (\alpha_0 + \boldsymbol{\mu}_{\mathbf{x}}\boldsymbol{\beta}_0)$, given earlier as

$$\hat{\tau}_{ate,pswreg} = (\hat{\alpha}_1 + \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}_1) - (\hat{\alpha}_0 + \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}_0) = (\hat{\alpha}_1 - \hat{\alpha}_0) + \bar{\mathbf{x}}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)$$

is consistent. This is the second half of “double robustness.”

7. Assessing Ignorability

- As mentioned earlier, ignorability is not directly testable. So any assessment of this key assumption is necessarily via indirect means.
- There are several possibilities. With multiple control groups, can establish that a “treatment effect” comparing two different control groups, say, is not statistically different from zero.

- For example, as in Heckman, Ichimura, and Todd (1997), we ,might have ineligibles and eligible nonparticipants. If there is no treatment effect using, say, ineligibles as the “control” and eligible nonparticipants as the “treatment,” we have more faith in unconfoundedness for the actual treatment. (We can conclude there is no self-selection into eligibility.)
- But, of course, unconfoundeness of treatment and of eligibility are potentially different.

• Formalize by having three treatment values, $w_i \in \{-1, 0, 1\}$, with $w_i = -1$, $w_i = 0$ representing two different controls. (For example, $w_i = -1$ means ineligible, $w_i = 0$ means eligible nonparticipants, $w_i = 1$ means treated.) If ignorability holds with respect to w_i, that is,

$$D(y_{i,-1}, y_{i0}, y_{i1} | \mathbf{x}_i, w_i) = D(y_{i,-1}, y_{i0}, y_{i1} | \mathbf{x}_i),$$

and $D(y_{i,-1} || \mathbf{x}_i) = D(y_{i0} | \mathbf{x}_i)$ – then

$$y_i \perp w_i \mid \mathbf{x}_i, w_i \in \{-1, 0\}.$$

- Easy to see for conditional means:

$$\begin{aligned} E(y_i|\mathbf{x}_i, w_i) &= 1[w_i = -1]E(y_{i,-1}|\mathbf{x}_i, w_i) + 1[w_i = 0]E(y_{i0}|\mathbf{x}_i, w_i) \\ &\quad + 1[w_i = 1]E(y_{i1}|\mathbf{x}_i, w_i) \\ &= 1[w_i = -1]E(y_{i,-1}|\mathbf{x}_i) + 1[w_i = 0]E(y_{i0}|\mathbf{x}_i) \\ &\quad + 1[w_i = 1]E(y_{i1}|\mathbf{x}_i) \end{aligned}$$

and so

$$E(y_i|\mathbf{x}_i, w_i \in \{-1, 0\}) = 1[w_i = -1]E(y_{i,-1}|\mathbf{x}_i) + 1[w_i = 0]E(y_{i0}|\mathbf{x}_i).$$

• It follows that if $E(y_{i,-1}|\mathbf{x}_i) = E(y_{i0}|\mathbf{x}_i)$ – conditional on $\mathbf{x}$ there is no difference, on average, between the ineligibles and eligible nonparticipants – then

$$E(y_i|\mathbf{x}_i, w_i = -1) = E(y_i|\mathbf{x}_i, w_i = 0).$$

• This is a testable restriction. It says that, if we focus on the two nontreated groups, we should not see any systematic difference in the observed response conditional on $\mathbf{x}_i$.

- We can estimate separate regression models for $w_i = -1$ and $w_i = 0$ and test whether they are the same (like a Chow test). Another implication is that if $w_i = -1$ is the "control" group and $w_i = 0$ is the "treated" group, the estimated ATE should not be statistically different from zero.
- Problem is that the implication only goes one way. It could be that y_i and w_i are independent conditional on $\mathbf{x}_i$ when we restrict attention to $w_i \in \{-1, 0\}$, but selection into actual treatment ($w_i = 1$ versus $w_i \in \{-1, 0\}$) need not be ignorable.

• If have several pre-treatment outcomes, can construct a treatment effect on a pseudo outcome and establish that it is not statistically different from zero.

• For concreteness, suppose controls consist of time-constant characteristics, $\mathbf{z}_i$, and three pre-assignment outcomes on the response, $y_{i,-1}, y_{i,-2}$, and $y_{i,-3}$. Let the counterfactuals be for time period zero, $y_{i0}(0)$ and $y_{i0}(1)$, where the term in $(\cdot)$ represents control or tretment. Suppose we are willing to assume unconfoundedness given two lags:

$$y_{i0}(0), y_{i0}(1) \perp w_i \mid y_{i,-1}, y_{i,-2}, \mathbf{z}_i$$

• If the process generating $\{y_{is}(g)\}$ is appropriately stationary and exchangeable, it can be shown that

$$y_{i,-1} \perp w_i, \mid y_{i,-2}, y_{i,-3}, \mathbf{z}_i,$$

and this of course is testable. (Again, one can use a Chow type test where the nature of y_i is appropriately accounted for; or a nonparametric test is used.) Conditional on $(y_{i,-2}, y_{i,-3}, \mathbf{z}_i)$, $y_{i,-1}$ should not differ systematically for the treatment and control groups.

• Alternatively, can try to assess sensitivity to failure of ignorability by using a specific alternative mechanism. For example, suppose unconfoundedness holds conditional on an unobservable, v, in addition to $\mathbf{x}$:

$$y_{i0}, y_{i1} \perp w_i \mid \mathbf{x}_i, v_i$$

If we parametrically specify $E(y_{ig}|\mathbf{x}_i, v_i)$, $g = 0, 1$, specify $P(w_i = 1|\mathbf{x}_i, v_i)$, and assume (typically) that v_i and $\mathbf{x}_i$ are independent, then τ_{ate} can be obtained in terms of the parameters of all specifications.

- In practice, we consider the version of ATE conditional on the covariates in the sample, τ_{cate} – the "conditional" ATE – so that we only have to integrate out v_i. Often, v_i is assumed to be very simple, such as a binary variable (indicating two "types" of individuals, say).
- Even for rather simple schemes, approach is complicated. One set of parameters are "sensitivity" parameters, other set is estimated. Then, evaluate how τ_{cate} changes with the sensitivity parameters.
- See Imbens (2003, *REStat*) or Imbens and Wooldridge (2009, *JEL*) for details.

• Altonji, Elder, and Taber (2005, *JPE*) propose a different strategy. In a constant treatment effect case, write the observed response as

$$y_i = \alpha + \tau w_i + \mathbf{x}_i\boldsymbol{\gamma} + u_i$$
$$E(u_i) = 0,\ E(\mathbf{x}_i' u_i) = \mathbf{0}.$$

So w_i is potentially endogenous.

• Linearly project a latent variable w_i^* determining w_i onto the observables part, $\mathbf{x}_i\boldsymbol{\gamma}$, and unobservable part, u_i:

$$w_i^* = \pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \omega u_i + e_i$$
$$E(e_i) = 0,\ Cov(\mathbf{x}_i\boldsymbol{\gamma}, e_i) = Cov(u_i, e_i) = 0.$$

• Suppose $\omega,\ \eta \geq 0$. AET argue that $\omega \leq \eta$ is reasonable: the observables are at least as important as the unobservables in determining assignment. They view estimates with $\omega = \eta$ as a lower bound on τ (assuming positive selection and $\tau > 0$) and estimates with $\omega = 0$ (OLS in this case) as an upper bound. In a counterfactual setting, $y_{i0} = \alpha + \mathbf{x}_i\boldsymbol{\gamma} + u_i$ and the variable determining treatment status, $w_i^*,$ is related to $\mathbf{x}_i\boldsymbol{\gamma}$ and u_i.

• How do we use the restriction $\eta = \omega$? Here is one possibility, used by AET. Write the linear outcome equation with a binary selection equation:

$$y_i = \alpha + \tau w_i + \mathbf{x}_i\boldsymbol{\gamma} + u_i$$
$$w_i = 1[\psi + \mathbf{x}_i\boldsymbol{\beta} + v_i \geq 0] = 1[w_i^* \geq 0]$$

where $D(v_i|\mathbf{x}_i) = Normal(0, 1)$. Then we have

$$w_i^* = \psi + \mathbf{x}_i\boldsymbol{\beta} + v_i$$
$$w_i^* = \pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i + e_i$$

where we impose $\eta = \omega$ in the linear projection.

• The two equations for w_i^* imply

$$v_i = (\pi - \psi) + \eta(\mathbf{x}_i\boldsymbol{\gamma}) - \mathbf{x}_i\boldsymbol{\beta} + \eta u_i + e_i$$

Now use the fact that u_i has zero mean and is uncorrelated with $\mathbf{x}_i$ and e_i:

$$Cov(u_i, v_i) = \eta Var(u)$$
$$\sigma_{uv} = \eta\sigma_u^2$$

But η is the slope from regressing $w_i^* = \psi + \mathbf{x}_i\boldsymbol{\beta} + v_i$ on $1, \mathbf{x}_i\boldsymbol{\gamma}$. Because v_i is uncorrelated with $\mathbf{x}_i$, this is the same as regressing $\mathbf{x}_i\boldsymbol{\beta}$ on $1, \mathbf{x}_i\boldsymbol{\gamma}$.

Therefore,

$$\eta = \frac{Cov(\mathbf{x}_i\boldsymbol{\beta}, \mathbf{x}_i\boldsymbol{\gamma})}{Var(\mathbf{x}_i\boldsymbol{\gamma})}$$

- Using $\sigma_{uv} = \eta\sigma_u^2$, we have an extra restriction,

$$\sigma_{uv} = \frac{\sigma_u^2 Cov(\mathbf{x}_i\boldsymbol{\beta}, \mathbf{x}_i\boldsymbol{\gamma})}{Var(\mathbf{x}_i\boldsymbol{\gamma})}$$

• The AET estimation with $\eta = \omega$ imposes

$$
\begin{aligned}
y_i &= \alpha + \tau w_i + \mathbf{x}_i\boldsymbol{\gamma} + u_i \\
w_i &= 1[\psi + \mathbf{x}_i\boldsymbol{\beta} + v_i \geq 0] \\
\sigma_{uv} &= \frac{\sigma_u^2 Cov(\mathbf{x}_i\boldsymbol{\beta}, \mathbf{x}_i\boldsymbol{\gamma})}{Var(\mathbf{x}_i\boldsymbol{\gamma})}
\end{aligned}
$$

where (u_i, v_i) has zero mean, is normally distributed (with $\sigma_v^2 = 1$), and is independent of $\mathbf{x}_i$.

• The last restriction recognizes that while σ_{uv} is technically identified without the restriction, it would only be identified off of the nonlinear model for w_i – a poor identification strategy.

• If we ignore the last restriction and set $\sigma_{uv} = 0$, then we are just estimating the first equation by OLS.

• If we replace the model for y_i with a probit model, then

$$y_i = 1[\alpha + \tau w_i + \mathbf{x}_i\boldsymbol{\gamma} + u_i \geq 0]$$
$$w_i = 1[\psi + \mathbf{x}_i\boldsymbol{\beta} + v_i \geq 0]$$

and $\sigma_u^2 = \sigma_v^2 = 1$, $\sigma_{uv} = \rho = Corr(u_i, v_i)$, and the additional restriction is

$$\rho = \frac{Cov(\mathbf{x}_i\boldsymbol{\beta}, \mathbf{x}_i\boldsymbol{\gamma})}{Var(\mathbf{x}_i\boldsymbol{\gamma})}$$

(which needs to be kept between –1 and 1).

• A different kind of restriction would be to assume equality of the population R-squareds from regressing w_i^* on 1, $\mathbf{x}_i\boldsymbol{\gamma}$ and from w_i^* on 1, u_i. (Because $\mathbf{x}_i\boldsymbol{\gamma}$ and u_i are uncorrelated, the R-squared from regressing w_i^* on 1, $\mathbf{x}_i\boldsymbol{\gamma}$, u_i is the sum of the separate R-squareds.)

• A different approach to the problem under $\eta = \omega$, and one that could be computationally simpler, is to use the two equations

$$y_i = \alpha + \tau w_i + \mathbf{x}_i\boldsymbol{\gamma} + u_i$$
$$w_i = 1[\pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i + e_i \geq 0]$$

where, without loss of generality, we take $Var(e_i) = 1$. Now impose the parametric model on $D(w_i|\mathbf{x}_i\boldsymbol{\gamma}, u_i)$ – say, probit – and omit the model for $D(w_i|\mathbf{x}_i)$ entirely.

• We can use a set of moment conditions that just identifies the parameters. Write $u_i(\boldsymbol{\theta}) \equiv y_i - \alpha - \tau w_i - \mathbf{x}_i\boldsymbol{\gamma}$.

$$
\begin{aligned}
E(y_i - \alpha - \tau w_i - \mathbf{x}_i\boldsymbol{\gamma}) &= 0 \\
E[\mathbf{x}_i'(y_i - \alpha - \tau w_i - \mathbf{x}_i\boldsymbol{\gamma})] &= \mathbf{0} \\
E\left\{ \frac{\phi(\pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i(\boldsymbol{\theta}))[w_i - \Phi(\pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i(\boldsymbol{\theta}))]}{\Phi(\pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i(\boldsymbol{\theta}))[1 - \Phi(\pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i(\boldsymbol{\theta}))]} \right\} &= 0 \\
E\left\{ \frac{(\mathbf{x}_i\boldsymbol{\gamma})\phi(\pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i(\boldsymbol{\theta}))[w_i - \Phi(\pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i(\boldsymbol{\theta}))]}{\Phi(\pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i(\boldsymbol{\theta}))[1 - \Phi(\pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i(\boldsymbol{\theta}))]} \right\} &= 0 \\
E\left\{ \frac{u_i(\boldsymbol{\theta})\phi(\pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i(\boldsymbol{\theta}))[w_i - \Phi(\pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i(\boldsymbol{\theta}))]}{\Phi(\pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i(\boldsymbol{\theta}))[1 - \Phi(\pi + \eta(\mathbf{x}_i\boldsymbol{\gamma}) + \eta u_i(\boldsymbol{\theta}))]} \right\} &= 0
\end{aligned}
$$

- The first two sets of conditions – $K + 1$ – are from $E(u_i) = 0$, $E(\mathbf{x}_i' u_i) = \mathbf{0}$. The last three conditions are the first order conditions for a probit with explanatory variables $[1, (\mathbf{x}_i\boldsymbol{\gamma}), u_i(\boldsymbol{\theta})]$. The notation $u_i(\boldsymbol{\theta})$ makes it clear that all of the parameters in the outcome equation appear also in the probit FOCs.
- We have $K + 4$ parameters to estimate and $K + 4$ moments.

8. Assessing Overlap

- Simple, first step is to compute normalized differences for each covariate. Let $\bar{x}_{1j}$ and $\bar{x}_{0j}$ be the means of covariate j for the treated and control subsamples, respectively, and let s_{1j} and s_{0j} be the estimated standard deviations. Then the normalized difference is

$$normdiff_j = \frac{(\bar{x}_{1j} - \bar{x}_{0j})}{\sqrt{s_{1j}^2 + s_{0j}^2}}$$

- Imbens and Rubin discuss rules-of-thumb. Normalized differences above about .25 should raise flags.

- $normdiff_j$ is not the t statistic for comparing the means of the distribution. The t statistic depends fundamentally on the sample size. Here interested in difference in population distributions, not statistical significance.
- Limitation of looking at the normalized differences: they only consider each marginal distribution. There can still be areas of weak overlap in the support $\mathcal{X}$ even if the normalized differences are all similar.

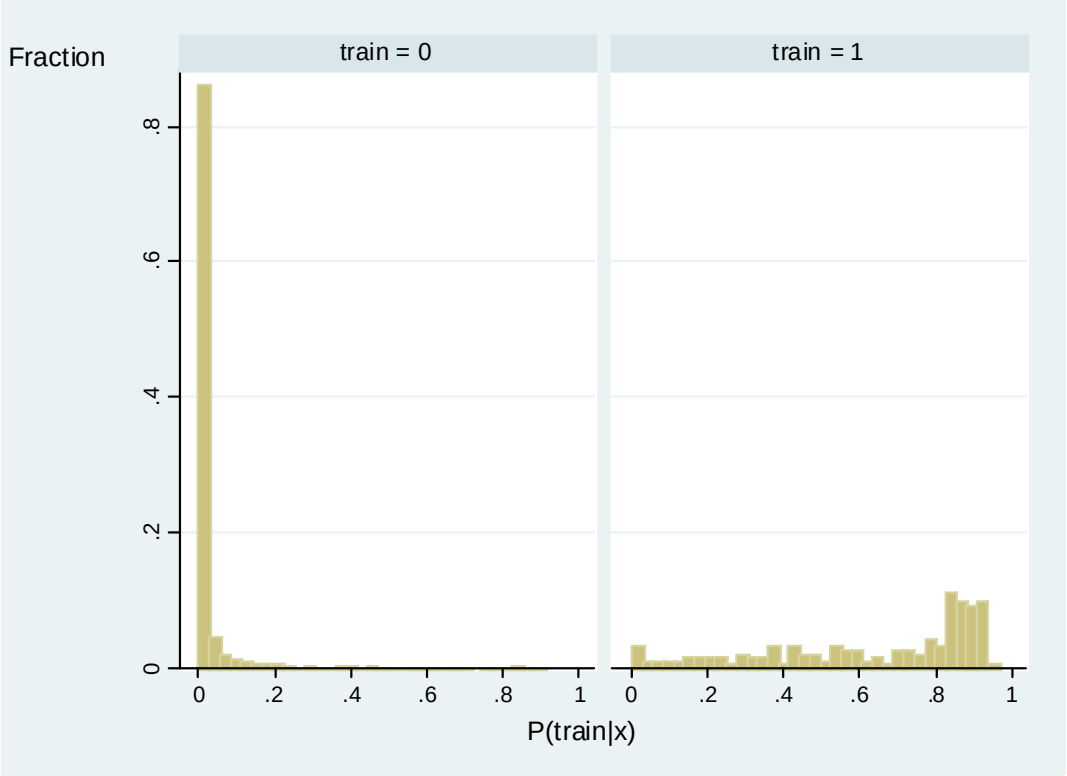
• Recall the key Rosenbaum and Rubin result that justifies both matching and regression on the propensity score: ignorability holds conditional on $p(\mathbf{x})$ if it holds conditional on $\mathbf{x}$. Thus, we need overlap in the distribution of $p(\mathbf{x})$.

• Therefore, look directly at the distributions (histograms) of estimated propensity scores for the treated and control groups. These histograms should show sufficient overlap.

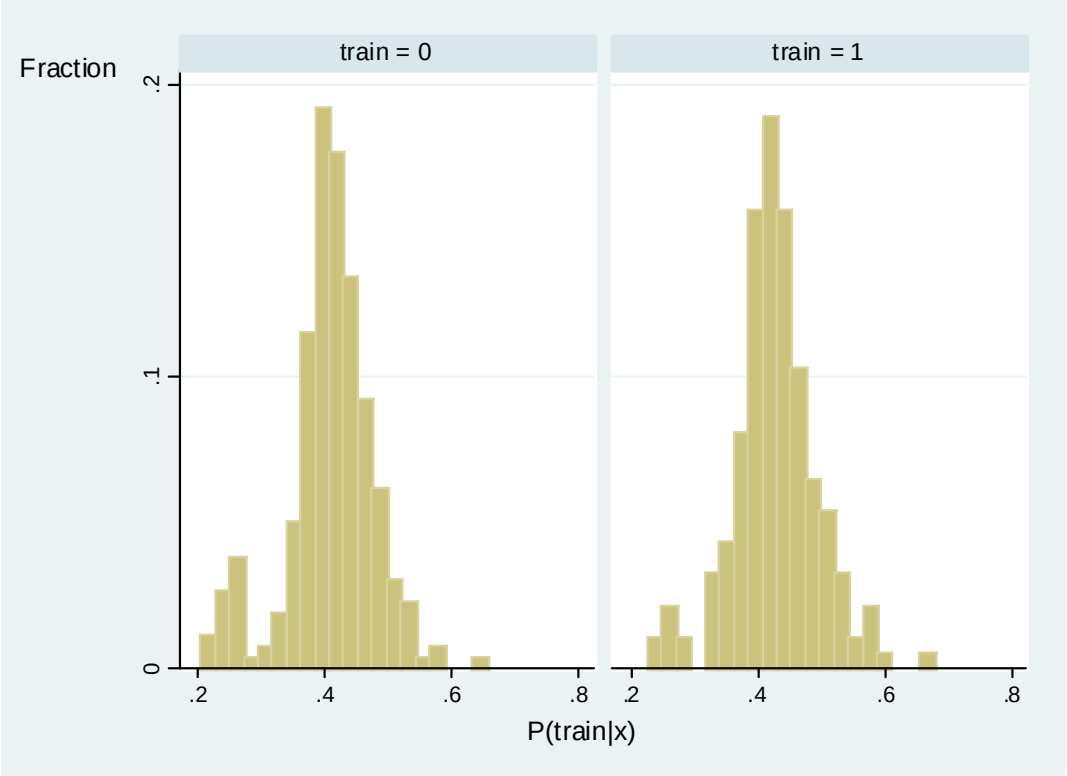
EXAMPLE: **Effects of Job Training**

• The file JTRAIN3 contains nonexperimental data. For comparison, the (much smaller) experimental data set, JTRAIN2 is also used. Finally, go back to JTRAIN3 and drop all people with estimated propensity score less than .05. Then redo the analysis.

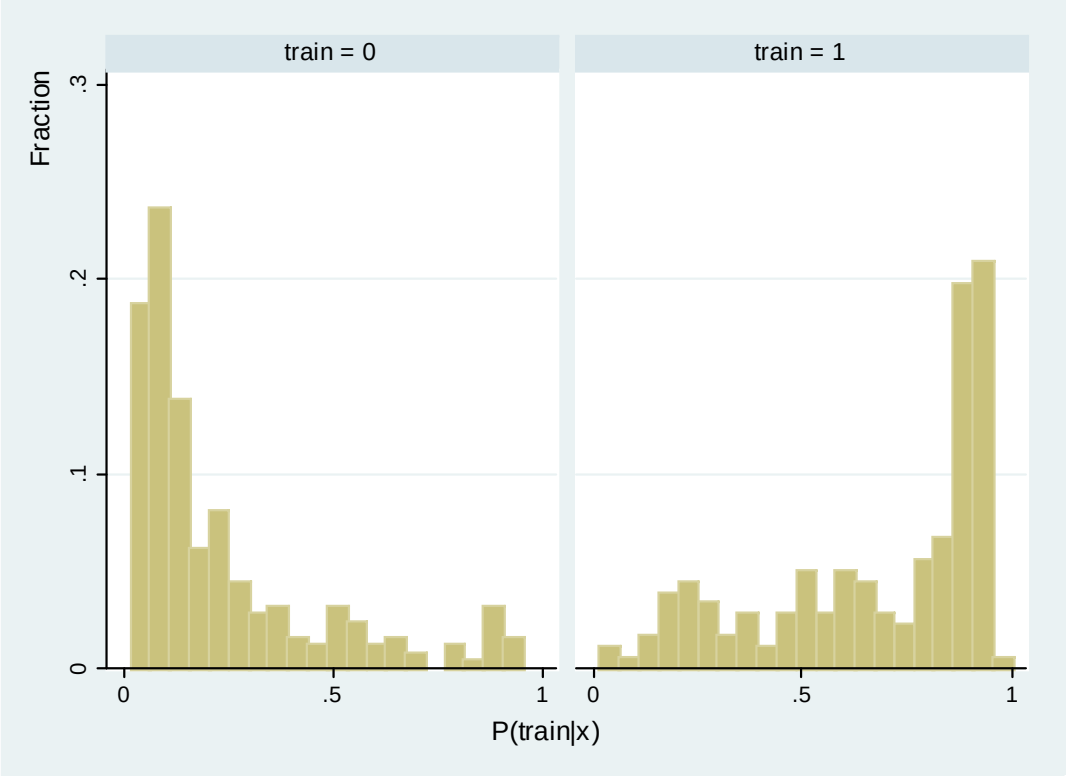
```
. use jtrain3

. logit train age educ black hisp married re74 re75

Logistic regression                                     Number of obs   =       2675
                                                        LR chi2(7)      =     872.82
                                                        Prob > chi2     =     0.0000
Log likelihood = -236.23799                             Pseudo R2       =     0.6488

------------------------------------------------------------------------------
       train |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0840291    .014761    -5.69   0.000    -.1129601    -.055098
        educ |  -.0624764   .0513973    -1.22   0.224    -.1632134    .0382605
       black |   2.242955   .3176941     7.06   0.000     1.620286    2.865624
        hisp |   2.094338   .5584561     3.75   0.000     .9997841    3.188892
     married |  -1.588358   .2602448    -6.10   0.000    -2.098428   -1.078287
        re74 |   -.117043   .0293604    -3.99   0.000    -.1745882   -.0594977
        re75 |  -.2577589   .0394991    -6.53   0.000    -.3351758   -.1803421
       _cons |   2.302714   .9112559     2.53   0.012     .5166853    4.088743
------------------------------------------------------------------------------
Note: 158 failures and 0 successes completely determined.

. predict phat
(option pr assumed; Pr(train))

. histogram phat, fraction by(train)
```


Fraction
train = 0
train = 1
0
.2
.4
.6
.8
1
P(train|x)

```
. use jtrain2

. logit train age educ black hisp married re74 re75

Logistic regression                                     Number of obs   =        445
                                                        LR chi2(7)      =       8.58
                                                        Prob > chi2     =     0.2840
Log likelihood = -297.80826                             Pseudo R2       =     0.0142

------------------------------------------------------------------------------
       train |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0107155    .014017     0.76   0.445    -.0167572    .0381882
        educ |   .0628366   .0558026     1.13   0.260    -.0465346    .1722077
       black |  -.3553063   .3577202    -0.99   0.321    -1.056425    .3458123
        hisp |  -.9322569   .5001292    -1.86   0.062    -1.912492    .0479784
     married |   .1440193   .2734583     0.53   0.598    -.3919492    .6799878
        re74 |  -.0221324   .0252097    -0.88   0.380    -.0715425    .0272777
        re75 |   .0459029   .0429705     1.07   0.285    -.0383177    .1301235
       _cons |  -.9237055   .7693924    -1.20   0.230    -2.431687    .5842759
------------------------------------------------------------------------------

. predict phat
(option pr assumed; Pr(train))

. histogram phat, fraction by(train)
```


Fraction
train = 0
train = 1
0
.1
.2
.2
.4
.6
.8
P(train|x)

```
. use jtrain3

. qui logit train age educ black hisp married re74 re75

. predict phat
(option pr assumed; Pr(train))

. drop if phat < .05
(2253 observations deleted)

. drop phat

. qui logit train age educ black hisp married re74 re75

. predict phat
(option pr assumed; Pr(train))

. histogram phat, fraction by(train)
```


train = 0
train = 1
Fraction
0
.1
.2
.3
0
.5
1
P(train|x)