PROGRAM EVALUATION: UNCONFOUNDED ASSIGNMENT

Jeff Wooldridge
Michigan State University
LABOUR Lectures, EIEF
October 18-19, 2011

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 nature of the assignment, intervention, or “treatment.”

- For example, is the “treatment” randomly assigned? Hardly ever in economics, and problematical even in clinical trials because those chosen to be eligible can and do opt out. But there is a push in some fields, for example, development economics, to use more randomized trials.
- With retrospective or observational data, a reasonable possibility is to assume that treatment is effectively randomly assigned conditional on observable covariates. (“Unconfoundedness” or “ignorability” of treatment 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? ("Confounded" assignment or "selection on unobservables" or "endogenous treatment")
- Often there is a component of self-selection in program evaluation.

- Broadly speaking, approaches to treament effect estimation fall into one of three situations: (1) Assume unconfoundedness of treatment, and then worry about how to exploit it in estimation; (2) Allow self-selection on unobservables but exploit an exogenous instrumental variable; (3) Exploit a “regression discontinuity” design, where the treatment is determined (or its probability) as a discontinuous function of observed variable.
- Here we consider estimation under unconfoundedness.

- Unconfoundedness leads to many possible estimation methods. We can usually put these into one of three categories: (1) regression adjustment; (2) propensity score weighting; (3) matching.
- Combinations of these methods can be very effective, but all maintain unconfoundedness.
- Unconfoundedness is fundamentally untestable, although in some cases there are ways to assess its plausibility or study sensitivity of estimates.

- A second key assumption is “overlap,” which concerns the similarity of the covariate distributions for the treated and untreated subpopulations. It plays a key role in any of the estimation methods based on unconfoundedness. In cases where parametric models are used, it can be too easily overlooked.
- If overlap is weak, may have to redefine the population of interest in order to precisely estimate a treatment effect on some subpopulation.

2. Basic Concepts

Counterfactual Outcomes and Parameters of Interest

- First assume a binary treatment. 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

$$Y(1) - Y(0).$$

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

$$Y_i(1) - Y_i(0).$$

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 $Y_i(0)$ and $Y_i(1)$ 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

$$\tau_{ate} = E[Y(1) - Y(0)].$$

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:

$$\tau_{att} = E[Y(1) - Y(0)|W = 1]$$

- With heterogeneous treatment effects, τ_{ate} and τ_{att} can be very different. ATE averages across gain from units that might never be subject to treatment.
- 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.
- We can also define ATEs and ATTs conditional on a set of observed covariates; in fact, some approaches to estimating τ_{ate} and τ_{att} rely on first estimating conditional average treatment effects.

Sampling Assumptions

- Assume independent, identically distributed observations from the underlying population. The data we would like to have is $\{(Y_i(0), Y_i(1)) \ : i = 1, \ldots, N\}$, but we only observe W_i and

$$Y_i = (1 - W_i)Y_i(0) + W_i Y_i(1) = Y_i(0) + W_i[Y_i(1) - Y_i(0)]. \qquad (4)$$

- Random sampling rules out treatment of one unit having an effect on other units. (So the "stable unit treatment value assumption," or SUTVA, is in force.)

Estimation under Random Assignment

• Strongest form of random assignment: $[Y(0), Y(1)]$ is independent of W. Then

$$E(Y|W = 1) - E(Y|W = 0) = E[Y(1)) - E(Y(0)] = \tau_{ate} = \tau_{att} \qquad (5)$$

under mean independence and the means on the left hand side can be estimated by using sample averages on the two subsamples.

• The randomization of treatment needed for the simple comparison-of-means estimator to consistently estimate the ATE is rare in practice.

Multiple Treatments

• If the treatment W_i takes on $G + 1$ levels, say $\{0, 1, \ldots, G\}$, it is straightforward to extend the counterfactual framework. Simply let $Y(0), \ldots, Y(G)$ denote the counterfactual outcomes associated with each level of treatment. If

$$\mu_g = E[Y(g)]$$

then we can define the expected gain in going from treatment level $g - 1$ to g as $\mu_g - \mu_{g-1}$. In some cases, $Y(0)$ might denote the response under no treatment, but generally the different values of W simply denote different treatment arms.

3. The Key Assumptions: Unconfoundedness 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. 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

$$(Y(0), Y(1)) \perp W \mid \mathbf{X}, \tag{1}$$

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

- We can also write unconfoundedness, or ignorability, as $D(W|Y(0), Y(1), \mathbf{X}) = D(W|\mathbf{X})$, 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).

- Essentially, unconfoundedness leads to a comparison-of-means after adjusting for observed covariates; even if one doubts we have “enough” of the “right” covariates, it is hard to envision not attempting such a comparison.

• 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 to the job training program.

• 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 assignment being randomized but then $\mathbf{X}$ includes a post-assignment variable that can be affected by assignment.)

• Can show that unconfoundedness generally fails unless $E(Y(g)|\mathbf{X}) = E(Y(g)), g = 0, 1.$

- 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 (1) is supposed to hold.

• Write $\mu_g(\mathbf{X}) \equiv E(Y(g)|\mathbf{X})$. 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.

• In convincing applications, $\mathbf{X}$ 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 unconfoundedness:

A.1′. Unconfoundedness in Conditional Mean:

$$E(Y(g)|W,\mathbf{X}) = E(Y(g)|\mathbf{X}),\ g = 0,1. \tag{2}$$

• Seems unlikely that this weaker version of the assumption holds without the stronger version. With weaker version, mean effects on different transformations of $Y(g)$ not identified.

- An argument in favor of an analyisis based on unconfoundedness is that the quantities we need to estimate are *nonparametrically identified*. Thus, if we used unconfoundedness we need impose few additional assumptions (other than overlap). By contrast, instrumental variables methods are either limited in what parameter they estimate or impose functional form and distributional restrictions.
- Can write down simple economic models where unconfoundedness holds, but the models limit the information available to agents when choosing "participation."

• To identify $\tau_{att} = E(Y(1) - Y(0)|W = 1)$, can get away with the weaker unconfoundedness assumption,

$$Y(0) \perp W \mid \mathbf{X}$$

or the mean version, $E(Y(0)|W, \mathbf{X}) = E(Y(0)|\mathbf{X})$. For example, the unit-specific gain, $Y_i(1) - Y_i(0)$, can depend on treatment status W_i in an arbitrary way.

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

$$0 < P(W = 1|\mathbf{X} = \mathbf{x}) < 1. \tag{3}$$

In other words, each unit in the defined population has some chance of being treated and some chance of not being treated. The probability of treatment as a function of $\mathbf{x}$ is known as the *propensity score*, which we denote

$$p(\mathbf{x}) = P(W = 1|\mathbf{X} = \mathbf{x}). \tag{4}$$

- *Strong Ignorability* [Rosenbaum and Rubin (1983)] = Unconfoundedness plus Overlap.
- For ATT, (3) can be relaxed to $p(\mathbf{x}) < 1$ for all $\mathbf{x} \in \mathcal{X}$; $p(\mathbf{x}) = 0$ is allowed (because we only average over the treated subpopulation).

4. Identification of Average Treatment Effects

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

$$\tau(\mathbf{x}) = E(Y(1) - Y(0)|\mathbf{X} = \mathbf{x}) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) \tag{5}$$

where $\mu_g(\mathbf{x}) \equiv E(Y(g)|\mathbf{X} = \mathbf{x}), g = 0, 1.$

- 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, it is always true (without any assumptions) that

$$\tau_{ate} = E(Y(1) - Y(0)) = E[\tau(\mathbf{X})] = E[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})] \quad (6)$$

It follows that τ_{ate} is identified if $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are identified over the support of $\mathbf{X}$, 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 unconfoundedness (and overlap), note that

$$\begin{aligned} E(Y|\mathbf{X}, W) &= (1 - W)E(Y(0)|\mathbf{X}, W) + WE(Y(1)|\mathbf{X}, W) \\ &= (1 - W)E(Y(0)|\mathbf{X}) + WE(Y(1)|\mathbf{X}) \\ &\equiv (1 - W)\mu_0(\mathbf{X}) + W\mu_1(\mathbf{X}), \end{aligned} \tag{7}$$

where the second equality holds by unconfoundedness. Define the always identified functions

$$m_0(\mathbf{X}) = E(Y|\mathbf{X}, W = 0), m_1(\mathbf{X}) = E(Y|\mathbf{X}, W = 1) \tag{8}$$

- Under overlap, $m_0(\cdot)$ and $m_1(\cdot)$ are nonparametrically identified on $\mathcal{X}$ because we assume the availability of a random sample on $(Y, \mathbf{X}, W)$.
- When we add unconfoundedness we identify $\mu_0(\cdot)$ and $\mu_1(\cdot)$ because

$$E(Y|\mathbf{X}, W = 0) = \mu_0(\mathbf{X}), E(Y|\mathbf{X}, W = 1) = \mu_1(\mathbf{X}) \tag{9}$$

- For ATT,

$$\begin{aligned} E(Y(1) - Y(0)|W) &= E[E(Y(1) - Y(0)|\mathbf{X}, W)|W] \\ &= E[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})|W]. \end{aligned} \tag{10}$$

- Therefore,

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

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

- In terms of the always identified mean functions,

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

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

By definition we can always estimate $E[m_1(\mathbf{X})|W = 1]$, and so, for τ_{att}, we can get by with "partial" overlap. Namely, we need to be able to estimate $m_0(\mathbf{x})$ for values of $\mathbf{x}$ taken on by the treatment group, which translates into $p(\mathbf{x}) < 1$ for all $\mathbf{x} \in \mathcal{X}$.

• We can also establish identification of τ_{ate} and τ_{att} using the propensity score. Also assuming unconfoundedness,

$$E\left[\frac{WY}{p(\mathbf{X})}\right] = E\left[\frac{WY(1)}{p(\mathbf{X})}\right] = E\left[\frac{E(W|\mathbf{X})E(Y(1)|\mathbf{X})}{p(\mathbf{X})}\right] = E(Y(1)), \quad (14)$$

$$E\left[\frac{(1-W)Y}{1-p(\mathbf{X})}\right] = E(Y(0)). \quad (15)$$

• In (14) we need $p(\mathbf{x}) > 0$ and in (15) we need $p(\mathbf{x}) < 1$ (both for all $\mathbf{x} \in \mathcal{X}$).

• Putting the two expressions together gives

$$\tau_{ate} = E\left[\frac{WY}{p(\mathbf{X})} - \frac{(1-W)Y}{1-p(\mathbf{X})}\right] = E\left\{\frac{[W-p(\mathbf{X})]Y}{p(\mathbf{X})[1-p(\mathbf{X})]}\right\}. \quad (16)$$

• 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.

• Makes intuitive sense that we only need $p(\mathbf{x}) < 1$ because τ_{att} is an average effect for those eventually treated. Therefore, for this parameter, it does not matter if some units have no chance of being treated. (In effect, this is one way to define the quantity of interest in a way that the necessary overlap assumption has a better chance of holding. But there are other ways based on **X**.)

Efficiency Bounds

• How well can we hope to do in estimate τ_{ate} or τ_{att}? Let $\sigma_0^2(\mathbf{X}) = Var(Y(0)|\mathbf{X})$ and $\sigma_1^2(\mathbf{X}) = Var(Y(1)|\mathbf{X})$. From Hahn (1998), the lower bounds for asymptotic variances of $\sqrt{N}$-consistent estimators are

$$E\left[\frac{\sigma_1^2(\mathbf{X})}{p(\mathbf{X})} + \frac{\sigma_0^2(\mathbf{X})}{[1-p(\mathbf{X})]} + [\tau(\mathbf{X}) - \tau_{ate}]^2 \right]$$

and

$$E\left[\frac{p(\mathbf{X})\sigma_1^2(\mathbf{X})}{\rho} + \frac{p(\mathbf{X})^2\sigma_0^2(\mathbf{X})}{\rho^2[1-p(\mathbf{X})]} + \frac{[\tau(\mathbf{X}) - \tau_{att}]^2 p(\mathbf{X})}{\rho^2} \right]$$

for τ_{ate} and τ_{att}, respectively, where $\rho = E[p(\mathbf{X})]$.

• These expressions assume the propensity score, $p(\cdot)$, is unknown. As shown by Hahn (1998), knowing the propensity score does not affect the variance lower bound for estimating τ, but it does change the lower bound for estimating τ_{att}.

• Estimators exist that achieve these bounds. The more mass on $p(\mathbf{x})$ closer to zero and one, the harder it is to estimate τ_{ate}. τ_{att} only cares about $p(\mathbf{x})$ close to unity.

5. Estimating ATEs

- When we assume unconfounded 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.
- Can mix the various approaches, and often this helps.
- Sometimes regression or matching are done on the propensity score. PS matching is especially popular.
- Need to keep in mind that all methods work under unconfoundedness and overlap. But they may behave quite differently when overlap is weak.

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 as complicated as full nonparametric regression.
- Key is that we compute a fitted values for each outcome for *all* units in sample. So even though we only use the treated units to obtain $\hat{m}_1(\mathbf{x})$, we need $\hat{m}_1(\mathbf{X}_i)$ for all $i = 1, \ldots, N$.

- The regression-adjustment estimators are

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

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

- Because the ATE as a function of **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 **x**. For example, let $\mathcal{R} \subset \mathcal{X}$ be a subset of the possible values of **x**. Then we can estimate

$$\tau_{ate,\mathcal{R}} = E(Y(1) - Y(0)|\mathbf{X} \in \mathcal{R})$$

as

$$\hat{\tau}_{ate,\mathcal{R}} = N_{\mathcal{R}}^{-1} \sum_{\mathbf{X}_i \in \mathcal{R}} [\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)] \tag{20}$$

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

- The restriction $\mathbf{X}_i \in \mathcal{R}$ can help with problems of overlap. If we have sufficient numbers of treated and control units with $\mathbf{X}_i \in \mathcal{R}$, $\tau_{ate,\mathcal{R}}$ can be identified when τ_{ate} is not.
- Of course, in problems with overlap, we might just redefine the population to begin with as $\mathbf{X} \in \mathcal{R}$. For example, only consider people with somewhat poor labor market histories to be eligible for job training.

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

$$\hat{\tau}_{ate,reg} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \bar{\mathbf{X}}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0) \tag{21}$$

where $\bar{\mathbf{X}}$ is the row vector of sample averages. (The definition of τ_{ate} means that we 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

$$Y_i \text{ on } 1, W_i, \mathbf{X}_i, W_i \cdot (\mathbf{X}_i - \bar{\mathbf{X}}), \quad i = 1, \ldots, N.$$

$\hat{\tau}_{ate,reg}$ 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, but unlikely to matter much.

- Note how $\mathbf{X}_i$ is demeaned before forming interaction. This is critical because we do not want to estimate $\alpha_1 - \alpha_0$ unless $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0$ is imposed. We want to estimate τ_{ate}.
- Demeaning the covariates before constructing the interactions is known to often "solve" the multicollinearity problem in regression. But it "solves" the problem because it redefines the parameter we are trying to estimate to be the ATE. Usually we can much more easily estimate an ATE than the treatment effect at $\mathbf{x} = \mathbf{0}$ which, except by fluke, is unlikely to be of much interest.

• 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.

• 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

$$Y_i \text{ on } 1, W_i, \mathbf{X}_i, W_i \cdot (\mathbf{X}_i - \bar{\mathbf{X}}_{\mathcal{R}}), \quad i = 1, \ldots, N$$

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 $\bar{\mathbf{X}}$. Might instead just restrict the analysis to $\mathbf{X}_i \in \mathcal{R}$ so that the parameters in the linear regression are estimated only using observations with $\mathbf{X}_i \in \mathcal{R}$

• 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:

$$Y_i \text{ on } 1, W_i, \mathbf{X}_i, \quad i = 1, \ldots, N. \tag{22}$$

• If linear models do not seem appropriate for $E(Y(0)|\mathbf{X})$ and $E(Y(1)|\mathbf{X})$, the specific nature of the $Y(g)$ can be exploited.

• 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. Still use this expression if $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_0$ is imposed.

• Or, for general $Y \geq 0$, Poisson 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 for standard error of $\hat{\tau}_{ate,reg}$.

• 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.

- From 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:

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

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

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

• Can use a formal nonparametric analysis. Imbens, Newey, and Ridder (2005) and Chen, Hong, and Tarozzi (2005) consider series estimation: essentially polynomial linear regression with an increasing number of terms. Estimator achieves the asymptotic efficiency bound for τ_{ate}.

• Heckman, Ichimura, and Todd (1997) and Heckman, Ichimura, Smith, and Todd (1998) use local linear regression. For kernel function $K(\cdot)$ and bandwidth $h_N > 0$, obtain, say, $\hat{m}_1(\mathbf{x})$ as $\hat{\alpha}_{1,\mathbf{x}}$ from

$$\min_{\alpha_{1,\mathbf{x}},\boldsymbol{\beta}_{1,\mathbf{x}}} \sum_{i=1}^{N} W_i K\left(\frac{\mathbf{X}_i - \mathbf{x}}{h_N}\right)(Y_i - \alpha_{1,\mathbf{x}} - (\mathbf{X}_i - \mathbf{x})\boldsymbol{\beta}_{1,\mathbf{x}})^2$$

and similarly for $\hat{m}_0(\mathbf{x})$.

• 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 $\mathbf{X}$ includes variables measuring labor market history – and then extrapolate this function to those who did not participate. This can lead to sensitive estimates if nonparticipants have very different values of $\mathbf{X}$.

• In the linear case with unrestricted regression functions, can see how lack of overlap can make $\hat{\tau}_{ate,reg}$ sensitive to changes in the specification. Can write $\hat{\tau}_{ate,reg}$ as

$$\hat{\tau}_{ate,reg} = (\bar{Y}_1 - \bar{Y}_0)) - (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_0)(f_0\hat{\boldsymbol{\beta}}_1 - f_1\hat{\boldsymbol{\beta}}_0)$$

where $f_0 = N_0/(N_0 + N_1)$ and $f_1 = N_1/(N_0 + N_1)$ are the relative fractions. If $\bar{\mathbf{X}}_1$ and $\bar{\mathbf{X}}_0$ are very different, minor changes in slope coefficients across the regimes can have large effects on $\hat{\tau}_{ate,reg}$.

• Nonparametric methods are not helpful in overcoming poor overlap. If they are global "series" estimators based on flexible parametric models, they require extrapolation. With local estimation methods we cannot easily estimate, say, $m_1(\mathbf{x})$ for $\mathbf{x}$ values far away from those in the treated subsample.

• At least using local methods the problem of overlap is more obvious: we have little or even no data to estimate the regression functions for values of $\mathbf{x}$ with poor overlap.

• Using τ_{att} has advantages because it requires only one extrapolation. From

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

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 $\hat{m}_0(\mathbf{x})$ for the treated group, and this can be difficult if we have units in the treated group with covariate values very different from all units in the control group.

• Classic study by Lalonde (1986): the nonexperimental data combined the treated group from the experiment with a random sample from a different source. The result was a much more heterogeneous control group than treatment group. Regression on the treatment group, where covariates had restricted range (particularly pre-training earnings), and using this to predict subsequent earnings for the control group (with some very high values of pre-training earnings), led to very poor imputations for estimating τ_{ate}.

• Things are better with τ_{att} because do have untreated observations similar to the control group. But should we use all control observations to estimate $m_0(\cdot)$? Local regression methods help so that the many controls in Lalonde's sample with, say, large pre-training earnings, do not affect estimation of $m_0(\cdot)$ for the low earners.

• Get better results by redefining the population, either based on propensity scores or a variable such as average pre-training earnings.

- 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.
- Note that it is easy to use sampling weights with regression adjustment. If stratification is based on the covariates, may not need to use weights for estimating mean functions. But would have to use weights in estimate τ_{ate} and τ_{att}.

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*).

$$\begin{aligned}\hat{\tau}_{ate,psw} &= N^{-1}\sum_{i=1}^{N}\left[\frac{W_i Y_i}{\hat{p}(\mathbf{X}_i)} - \frac{(1-W_i)Y_i}{1-\hat{p}(\mathbf{X}_i)}\right] \\ &= N^{-1}\sum_{i=1}^{N}\frac{[W_i - \hat{p}(\mathbf{X}_i)]Y_i}{\hat{p}(\mathbf{X}_i)[1-\hat{p}(\mathbf{X}_i)]}. \qquad (26)\end{aligned}$$

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

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

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

• Clear that $\hat{\tau}_{ate,psw}$ can be sensitive to the choice of model for $p(\cdot)$ because now tail behavior can matter when $p(\mathbf{x})$ is close to zero or one. (For τ_{att}, only close to one matters.)

• Can use (26) and (27) as motivation for trimming based on the propensity score.

• To exploit estimation error, write

$$\hat{\tau}_{ate,psw} = N^{-1} \sum_{i=1}^{N} \frac{[W_i - \hat{p}(\mathbf{X}_i)]Y_i}{\hat{p}(\mathbf{X}_i)[1 - \hat{p}(\mathbf{X}_i)]} \equiv N^{-1} \sum_{i=1}^{N} \hat{k}_i. \tag{28}$$

The adjustment for estimating $\boldsymbol{\gamma}$ by MLE turns out to be 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

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

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! They can be correlated with

$$k_i = \frac{[W_i - p(\mathbf{X}_i)]Y_i}{p(\mathbf{X}_i)[1 - p(\mathbf{X}_i)]},$$

which 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.

• An alternative is to use bootstrapping, where the binary response estimation and averaging (to get $\hat{\tau}_{ate,psw}$) are included in each bootstrap iteration. Unfortunately, lots of bootstrap samples may give fitted probabilities that are zero or one.

• It is conservative to ignore the estimation error in the $\hat{k}_i$ and simply treat it as data. That corresponds to just computing the standard error for a sample average: $se(\hat{\tau}_{ate,psw}) = \left[N^{-1}\sum_{i=1}^{N}(\hat{k}_i - \hat{\tau}_{ate,psw})^2\right]^{1/2}/\sqrt{N}$. This is always larger than (31) and is gotten by the regression $\hat{k}_i$ on 1.

• For $\hat{\tau}_{att,psw}$, adjustment to standard error somewhat different (Wooldridge, 2010, Chapter 21).

- 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 $.1 \leq \hat{p}(\mathbf{X}_i) \leq .9$ (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.

Regression on the Propensity Score

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

$$(Y(0), Y(1)) \perp W \mid p(\mathbf{X})$$

This is a key finding of Rosenbaum and Rubin (1983).

• Conditional independence 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 to break the dependence between W and $(Y(0), Y(1))$. We need not condition on $\mathbf{X}$.
- By iterated expectations,

$$\tau_{ate} = E(Y(1) - Y(0)) = E\{E[Y(1)|p(\mathbf{X})] - E[Y(0)|p(\mathbf{X})]\}.$$

- By unconfoundedness,

$$\begin{aligned} E[Y|p(\mathbf{X}), W] &= (1-W)E[Y(0)|p(\mathbf{X}), W] + WE[Y(1)|p(\mathbf{X}), W] \\ &= (1-W)E[Y(0)|p(\mathbf{X})] + WE[Y(1)|p(\mathbf{X})] \end{aligned}$$

and so

$$\begin{aligned} E[Y|p(\mathbf{X}), W=0] &= E[Y(0)|p(\mathbf{X})] \\ E[Y|p(\mathbf{X}), W=1] &= E[Y(1)|p(\mathbf{X})] \end{aligned}$$

• So, after estimating $p(\mathbf{x})$, we estimate $E[Y|p(\mathbf{X}), W = 0]$ and $E[Y|p(\mathbf{X}), W = 1]$ using each subsample.

• In the linear case,

$$Y_i \text{ on } 1, \hat{p}(\mathbf{X}_i) \text{ for } W_i = 0 \text{ and } Y_i \text{ on } 1, \hat{p}(\mathbf{X}_i) \text{ for } W_i = 1, \tag{33}$$

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

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

• Linearity might be a poor assumption because the fitted values are necessarily bounded.

• 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.

• Actually, somewhat more common is to drop the interaction term.

$$Y_i \text{ on } 1, W_i, \hat{p}(\mathbf{X}_i), i = 1, \ldots, N. \tag{36}$$

• Theoretically, regression on the propensity score in regression has little to offer compared with other methods.

- Linear regression estimates such as (36) should not be too sensitive to $\hat{p}_i$ close to zero or one, but that might only mask the problem of poor covariate balance.
- 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 $\hat{\tau}_{ate,regprop}$.

Combining Regression Adjustment and PS Weighting

- Question: Why use regression adjustment combined with PS weighting?
- Answer: With $\mathbf{X}$ 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 WLS 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

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

• 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 but analytical formulas not difficult.

Matching on Covariates

- 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 $Y_i(0)$, but we need to impute $Y_i(1)$. For each unit i in the treatment group, we observe $Y_i(1)$ but need to impute $Y_i(0)$.
- For τ_{ate}, matching estimators take the general form

$$\hat{\tau}_{ate,match} = N^{-1} \sum_{i=1}^{N} (\hat{Y}_i(1) - \hat{Y}_i(0))$$

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

- For τ_{att},

$$\hat{\tau}_{att,match} = N_1^{-1} \sum_{i=1}^{N} W_i (Y_i - \hat{Y}_i(0))$$

where this form uses the fact that $W_i Y_i = W_i Y_i(1)$ (we never need to impute $Y_i(1)$ 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 $\hat{Y}_i(1) = Y_i, \hat{Y}_i(0) = Y_h$ 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 $Y_i(0)$.

• Similarly, if $W_i = 0, \hat{Y}_i(0) = Y_i, \hat{Y}_i(1) = Y_h$ 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

$$\hat{Y}_i(1) = Y_i$$
$$\hat{Y}_i(0) = M^{-1} \sum_{h \in \aleph_M(i)} Y_h$$

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

- Similarly, if $W_i = 0$,

$$\hat{Y}_i(0) = Y_i$$
$$\hat{Y}_i(1) = M^{-1} \sum_{h \in \mathfrak{I}_M(i)} Y_h$$

where $\mathfrak{I}_M(i)$ contains the M treated nearest matches to observation i.

- Remarkably, 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 unconfoundedness, $(W, \mathbf{X})$ are effectively "exogenous."
- How can we obtain a confidence interval? Bootstrapping does not work with matching.

- Instead, the conditional variance of the matching estimator is

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

- The unconditional variance is more complicated because of a conditional bias (see Abadie and Imbens), but estimators are programmed in `nnmatch`. Need to "estimate" $Var(Y_i|, W_i, \mathbf{x}_i)$, but they do not have to be good pointwise estimates.

- AI suggest

$$\widehat{Var(Y_i|, W_i, \mathbf{X}_i)} = (Y_i - Y_{h(i)})^2/2$$

where $h(i)$ is the closest match to observation i with $W_{h(i)} = W_i$. [The idea is that $Y_i - Y_{h(i)} \approx U_i - U_{h(i)}$.]

- Could instead use flexible parametric models for first two moments of $D(Y_i|W_i, \mathbf{X}_i)$, exploiting the nature of Y. For example, if Y is binary, use flexible logits for $W_i = 0$, $W_i = 1$, which is what we would do for regression adjustment.

• There is another way to think of the variance estimator. Define the **sample average treatment effect**, τ_{sate}, as

$$\tau_{sate} = N^{-1} \sum_{i=1}^{N} [Y_i(1) - Y_i(0)]$$

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

• A similar definition holds for τ_{satt}.

Matching with Regression Adjustment

- The matching estimators have a large-sample bias if $\mathbf{X}_i$ has dimension greater than one, on the order of $N^{-1/K}$ where K is the number of covariates. Dominates the variance asymptotically when $K \geq 3$.
- The bias of the matching estimator comes from terms of the form $\mu_w(\mathbf{X}_{h(i)}) - \mu_w(\mathbf{X}_i)$ where $\mu_w(\mathbf{x}) = E(Y|\mathbf{X}, W = w)$.

• Let $\hat{\mu}_w$ be estimators – probably nonparametric – of the conditional means. Then define new imputations as

$$\tilde{Y}_i(1) = Y_i \text{ if } W_i = 1$$
$$\tilde{Y}_i(1) = Y_{h(i)} + \hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_1(\mathbf{X}_{h(i)}) \text{ if } W_i = 0$$
$$\tilde{Y}_i(0) = Y_i \text{ if } W_i = 0$$
$$\tilde{Y}_i(0) = Y_{h(i)} + \hat{\mu}_0(\mathbf{X}_i) - \hat{\mu}_0(\mathbf{X}_{h(i)}) \text{ if } W_i = 1$$

- The bias-corrected matching estimator is

$$\hat{\tau}_{ate,bcme} = N^{-1} \sum_{i=1}^{N} [\tilde{Y}_i(1) - \tilde{Y}_i(0)]$$

- The BCME has the same sampling variance as the matching estimator, but the bias has been removed from the asymptotic distribution [provided $\mu_w(\cdot)$ are sufficiently smooth].
- The `nnmatch` command in Stata allows for bias adjustment.

Matching on the Propensity Score

- It is also possible to match on the estimated propensity score. This is computationally easier because it is a single variable with range in $(0, 1)$.
- The Stata command is `psmatch2`, and it allows a variety of options. (For example, whether to estimate ATT or ATE, how many matches to use, whether to use smoothing.)
- Until recently, valid inference not available for (unsmoothed) PS matching unless we know propensity score. Bootstrapping not justified, but this is how Stata computes the standard errors.

- The technical problem is that matching is not smooth in $\hat{p}(\mathbf{X}_i)$. A small change in $\hat{p}(\mathbf{X}_i)$ can change the match (matching is discontinuous in the PS).
- Abadie and Imbens (2011, unpublished, "Matching on the Estimated Propensity Score"): Using matching with replacement, it is possible to estimate the sampling variance of the PS matching estimator.
- The estimator that ignores estimation of the PS turns out to be conservative. So can apply `nnmatch` with an estimated propensity score to obain conservative inference.

6. Assessing Unconfoundedness

- As mentioned, unconfoundedness is not directly testable. So any assessment is indirect.
- There are several possibilities. With multiple control groups, can establish that a “treatment effect” for, say, comparing two control groups is not statistically different from zero. For example, as in Heckman, Ichimura, and Todd (1997), can have ineligibles and eligible nonparticipants. If there is no treatment effect using, say, ineligibles as the control and eligibility as the treatment, have more faith in unconfoundedness for the actual treatment. But, of course, unconfoundeness of treatment and of eligibility are different.

• Can formalize by having three treatment values, $D_i \in \{-1, 0, 1\}$, with $D_i = -1$, $D_i = 0$ representing two different controls. If unconfoundedness holds with respect to D_i, then it follows that

$$Y_i \perp D_i \mid \mathbf{X}_i, D_i \in \{-1, 0\}$$

which is testable by using $D_i = -1$ as the "control" and $D_i = 0$ as the "treated" and estimating an ATE using the previous methods.

• Problem is that the implication only goes one way: passing this test does not mean unconfounded holds; it is suggestive.

• 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 consiste 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 $Y_{i0}(1)$. 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. 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 unconfoundedness by using a specific alternative mechanism. For example, suppose unconfoundedness holds conditional on an unobservable, V, in addition to $\mathbf{X}$:

$$Y_i(0), Y_i(1) \perp W_i \mid \mathbf{X}_i, V_i$$

If we parametrically specify $E[Y_i(g)|\mathbf{X}_i, V_i]$, $g = 0, 1$, specify $P(W_i = 1|\mathbf{X}_i, V_i)$, 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").
- 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) or Imbens and Wooldridge (2009) for details.

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

$$Y_i = \alpha + \tau W_i + \mathbf{X}_i\boldsymbol{\gamma} + u_i$$
$$E(\mathbf{X}_i' u_i) = 0$$

and then project a latent variable determining W_i onto the observables and unobservables,

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

- AET define "selection on unobservables is the same as selection on observables" as the restriction $\omega = \eta$. The idea is, other than the treatment W_i, the factors affecting Y_i, the observable part $\mathbf{X}_i\boldsymbol{\gamma}$ and the unobservable u_i, have the same regression effect on W_i^*. In counterfactual setting, $Y_i(0) = \alpha + \mathbf{X}_i\boldsymbol{\gamma} + u_i$. AET argue that in fact $\omega \leq \eta$ is reasonable, and so view estimates with $\omega = \eta$ as a lower bound (assuming positive selection and $\tau > 0$) and estimates with $\omega = 0$ (OLS in this case) as an upper bound.
- Can apply to other kinds of Y_i, such as binary.

- In case where Y_i follows linear model, estimation imposes

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

(OLS sets $\sigma_{uv} = 0$.) Cannot really estimate σ_{uv} even though it is technically identified.

- If we replace model for Y_i with probit, $\sigma_u^2 = 1$ and $\sigma_{uv} = \rho = Corr(u_i, v_i)$.

7. Assessing and Improving Overlap

- A simple 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.
- Look directly at the propensity scores or the log-odds of the propensity score. In other words, compute the normalized difference of a single function of $\mathbf{X}_i$.

- Also look directly at the histograms of estimated propensity scores for the treated and control groups. The command `psgraph` does this after using `psmatch2`.
- If there are problems with overlap in the original sample, may have to redefine the population. [Focusing on τ_{att} rather than τ_{ate} can solve part of the overlap problem because $P(W = 1|\mathbf{X}) = 0$ is allowed.]
- Earlier mentioned the rule of dropping i if $\hat{p}(\mathbf{X}_i) \notin [.1, .9]$
- Can lose a lot of data – including treated observations – and resulting population might not be what we want.

• An alternative approach is to use the estimated PS to match each treated unit with a single control unit, to obtain a new sample with the same number of treated and controls.

• After using all of the data to estimate the PS, for treated units order from largest to smallest PS. Starting at top, match the first treated unit to the closest control. Then do the same for the next treated unit (not replacing the control units). If there are N_1 treated units, we wind up with N_1 controls, too.

• The new (smaller – in some cases, much smaller) sample is better balanced. Can apply all the usual methods for τ_{att}.

- Has the advantage of keeping all treated observations. But the population is hard to interpret.
- Might be better to think about a sensible population ahead of time. For example, would people with above median incomes be eligible for a job training program?

Applications

Lalonde (1986) Job Training Data

- Focus on a nonexperimental data set constructed by Lalonde.

Everything works with the experimental data.

- Available as JTRAIN3.DTA at MIT Press web site.
- Response *re*78 is a corner solution. For regression adjustment, could use exponential, Tobit, or Cragg instead of linear regression.

```
. tab train

   =1 if in |
        job |
   training |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,490       93.08       93.08
          1 |        185        6.92      100.00
------------+-----------------------------------
      Total |      2,675      100.00

. reg re78 train, robust

Linear regression                               Number of obs =    2675
                                                F(  1,  2673) =  537.36
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0609
                                                Root MSE      =  15.152

------------------------------------------------------------------------------
             |               Robust
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       train |  -15.20478   .6559143   -23.18   0.000    -16.49093   -13.91863
       _cons |   21.55392    .311785    69.13   0.000     20.94256    22.16529
------------------------------------------------------------------------------
```

```
. * Regression adjustment:

. reg re78 train age educ black hisp married unem74 unem75 re74 re75, robust

Linear regression                                      Number of obs =    2675
                                                       F( 10,  2664) =  244.41
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.5871
                                                       Root MSE      =  10.064

------------------------------------------------------------------------------
             |               Robust
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       train |   .1153847    .831752     0.14   0.890     -1.51556     1.74633
         age |  -.0897655   .0232735    -3.86   0.000    -.1354014   -.0441296
        educ |   .5141238   .0924729     5.56   0.000     .3327978    .6954498
       black |  -.4542188   .4461159    -1.02   0.309    -1.328987    .4205498
        hisp |   2.197368    1.22836     1.79   0.074    -.2112669    4.606004
     married |   1.204787   .4963532     2.43   0.015     .2315101    2.178063
      unem74 |   2.389527   1.360835     1.76   0.079    -.2788731    5.057927
      unem75 |  -1.461964   1.412258    -1.04   0.301    -4.231196    1.307269
        re74 |     .31262   .0616021     5.07   0.000     .1918272    .4334129
        re75 |   .5436543   .0682471     7.97   0.000     .4098318    .6774769
       _cons |   .9536064   1.500485     0.64   0.525    -1.988628    3.895841
------------------------------------------------------------------------------
```

```
. qui reg re78 age educ black hisp married unem74 unem75 re74 re75 if ~train

. predict re78_0
(option xb assumed; fitted values)

. qui reg re78 age educ black hisp married unem74 unem75 re74 re75 if train

. predict re78_1
(option xb assumed; fitted values)

. gen ate_i = re78_1 - re78_0

. sum ate_i

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       ate_i |      2675   -8.819958     7.75979  -83.80467   7.277479

. sum ate_i if train

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       ate_i |       185    .8431944    3.262634  -13.71843   5.917717
```

```
. do ateregjtrain3.do

. clear all
. capture program drop ateboot

. program ateboot, eclass
  1.
.         * Estimate linear model on each treatment group
.         tempvar touse
  2.         gen byte ‘touse’ = 1
  3.         reg re78 age educ black hisp married unem74 unem75 re74 re75 if train
  4.         predict re78h_1
  5.         reg re78 age educ black hisp married unem74 unem75 re74 re75 if ~train
  6.         predict re78h_0
  7.
.         gen ate_i = re78h_1 - re78h_0
  8.         sum ate_i
  9.         scalar ate = r(mean)
 10.         sum ate_i if train
 11.         scalar att = r(mean)
 12.
.         matrix b = (ate, att)
 13.         matrix colnames b = ate att
 14.
.         ereturn post b , esample(‘touse’)
 15.         ereturn display
 16.
.         drop re78h_1 re78h_0 ate_i
 17.
. end

. use jtrain3
.
. bootstrap _b[ate] _b[att], reps(1000) seed(123): ateboot
(running ateboot on estimation sample)
```

```
Bootstrap replications (1000)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................    50
..................................................
..................................................  1000

Bootstrap results                                 Number of obs      =      2675
                                                  Replications       =      1000

      command:  ateboot
        _bs_1:  _b[ate]
        _bs_2:  _b[att]

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |  -8.819958   4.525198    -1.95   0.051    -17.68918    .0492667
       _bs_2 |   .8431944    .925601     0.91   0.362    -.9709502    2.657339
------------------------------------------------------------------------------

.
. program drop ateboot

end of do-file
```

```
. * Nearest-neighbor matching, no bias (regression) adjustment:

. nnmatch re78 train age educ black hisp married unem74 unem75 re74 re75,
          tc(ate)

Matching estimator:  Average Treatment Effect ate

Weighting matrix: inverse variance          Number of obs          =      2675
                                            Number of matches  (m) =         1

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATE |  -13.25807   4.125526    -3.21   0.001    -21.34396    -5.17219
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married unem74 unem75 re74 re75

. nnmatch re78 train age educ black hisp married unem74 unem75 re74 re75,
          tc(att)

Matching estimator:  Average Treatment Effect for the Treated

Weighting matrix: inverse variance          Number of obs          =      2675
                                            Number of matches  (m) =         1

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATT |    2.07348   1.673616     1.24   0.215    -1.206747    5.353707
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married unem74 unem75 re74 re75
```

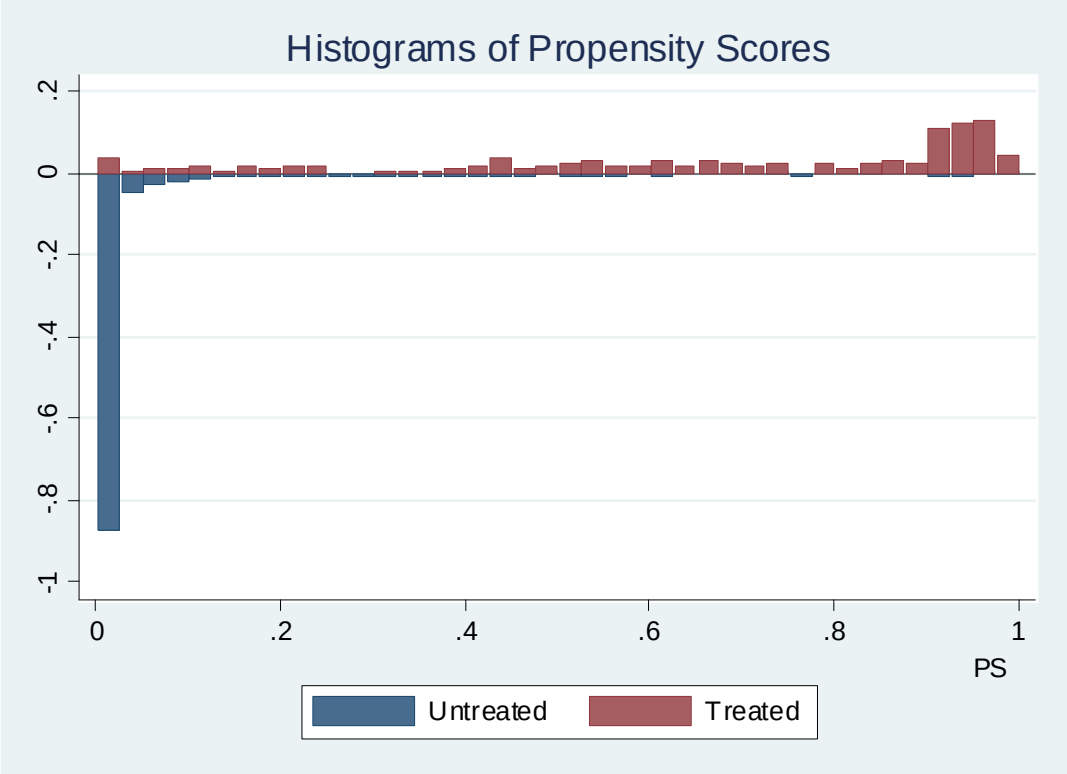
```
* With bias (regression) adjustment:

. nnmatch re78 train age educ black hisp married unem74 unem75 re74 re75,
          tc(ate) biasadj(bias)

Matching estimator:  Average Treatment Effect ate

Weighting matrix: inverse variance          Number of obs          =      2675
                                            Number of matches  (m) =         1

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATE |  -6.599812   3.413632    -1.93   0.053    -13.29041    .0907829
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married unem74 unem75 re74 re75
Bias-adj variables:  age educ black hisp married unem74 unem75 re74 re75

. nnmatch re78 train age educ black hisp married unem74 unem75 re74 re75,
          tc(att) biasadj(bias)

Matching estimator:  Average Treatment Effect for the Treated

Weighting matrix: inverse variance          Number of obs          =      2675
                                            Number of matches  (m) =         1

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATT |   2.415483   1.679509     1.44   0.150    -.8762945     5.70726
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married unem74 unem75 re74 re75
Bias-adj variables:  age educ black hisp married unem74 unem75 re74 re75
```

```
* Propensity score matching:

. logit train age educ black hisp married unem74 unem75 re74 re75

Logistic regression                                     Number of obs   =       2675
                                                        LR chi2(9)      =     926.52
                                                        Prob > chi2     =     0.0000
Log likelihood = -209.38931                             Pseudo R2       =     0.6887

------------------------------------------------------------------------------
       train |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1109206   .0177106    -6.26   0.000    -.1456327   -.0762084
        educ |  -.1008807   .0561133    -1.80   0.072    -.2108608    .0090994
       black |   2.650097   .3605668     7.35   0.000     1.943399    3.356795
        hisp |   2.247747   .5908963     3.80   0.000     1.089611    3.405882
     married |  -1.560628   .2817913    -5.54   0.000    -2.112928   -1.008327
      unem74 |   3.272456   .4887585     6.70   0.000     2.314507    4.230405
      unem75 |  -1.371405   .4545789    -3.02   0.003    -2.262363   -.4804465
        re74 |   .0201797   .0313149     0.64   0.519    -.0411963    .0815557
        re75 |  -.2743162   .0477066    -5.75   0.000    -.3678194   -.1808129
       _cons |   1.794543    .979261     1.83   0.067    -.1247735    3.713859
------------------------------------------------------------------------------
Note: 78 failures and 0 successes completely determined.
```

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

. sum phat

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        phat |      2675    .0691589    .2041418   8.86e-21   .9889598

. count if phat < .1
 2366

. count if phat > .9
   83

. psmatch2 train age educ black hisp married unem74 unem75 re74 re75,
           outcome(re78) logit ate

Logistic regression                               Number of obs   =       2675
                                                  LR chi2(9)      =     926.52
                                                  Prob > chi2     =     0.0000
Log likelihood = -209.38931                       Pseudo R2       =     0.6887

------------------------------------------------------------------------------
       train |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.1109206   .0177106    -6.26   0.000    -.1456326   -.0762085
        educ |  -.1008807   .0561131    -1.80   0.072    -.2108604    .0090991
       black |   2.650097   .3605654     7.35   0.000     1.943402    3.356792
        hisp |   2.247747   .5908943     3.80   0.000     1.089615    3.405878
     married |  -1.560628     .28179    -5.54   0.000    -2.112926   -1.008329
      unem74 |   3.272456   .4887569     6.70   0.000      2.31451    4.230402
      unem75 |  -1.371405   .4545768    -3.02   0.003    -2.262359   -.4804507
        re74 |   .0201797   .0313146     0.64   0.519    -.0411959    .0815552
        re75 |  -.2743162   .0477056    -5.75   0.000    -.3678175   -.1808149
```

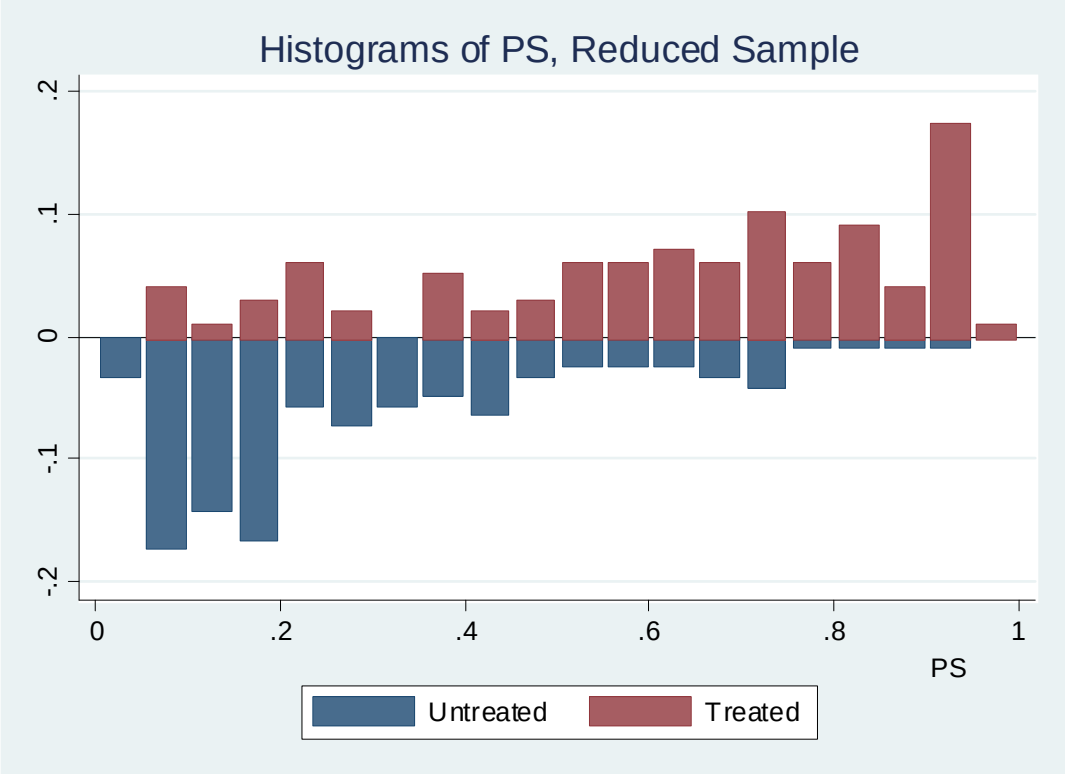
```
       _cons |   1.794543    .979257     1.83   0.067    -.1247656    3.713851
------------------------------------------------------------------------------
Note: 78 failures and 0 successes completely determined.
There are observations with identical propensity score values.
The sort order of the data could affect your results.
Make sure that the sort order is random before calling psmatch2.
------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+------------------------------------------------------------
            re78  Unmatched | 6.34914538   21.5539213  -15.2047759   1.15461436   -13.17
                        ATT | 6.34914538   4.10456445   2.24458093   1.58543438     1.42
                        ATU | 21.5539213   6.77791717  -14.7760042            .        .
                        ATE |                          -83.8046722            .        .
-----------------------------+------------------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.

           | psmatch2:
 psmatch2: |   Common
 Treatment |  support
assignment | On suppor |     Total
-----------+-----------+----------
 Untreated |     2,490 |     2,490
   Treated |       185 |       185
-----------+-----------+----------
     Total |     2,675 |     2,675
```

```
. psgraph, bin(40)
```



```
. keep if phat >= .1 & phat <= .9
(2449 observations deleted)

. psmatch2 train age educ black hisp married unem74 unem75 re74 re75,
           outcome(re78) logit ate

Logistic regression                                     Number of obs   =        226
                                                        LR chi2(9)      =      86.34
                                                        Prob > chi2     =     0.0000
Log likelihood = -111.48296                             Pseudo R2       =     0.2791

------------------------------------------------------------------------------
       train |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.138872   .0290637    -4.78   0.000    -.1958358   -.0819082
        educ |  -.1239302   .0739195    -1.68   0.094    -.2688096    .0209493
       black |   3.062403   .6672729     4.59   0.000     1.754572    4.370234
        hisp |   2.722197   .8657224     3.14   0.002     1.025412    4.418981
     married |  -1.942727   .4091421    -4.75   0.000    -2.744631   -1.140823
      unem74 |   4.273888   .8025175     5.33   0.000     2.700983    5.846794
      unem75 |  -2.345185   .6734993    -3.48   0.000     -3.66522   -1.025151
        re74 |   .0270945   .0445556     0.61   0.543    -.0602329    .1144218
        re75 |  -.4978745   .1039551    -4.79   0.000    -.7016227   -.2941263
       _cons |   3.203055   1.302987     2.46   0.014     .6492481    5.756862
------------------------------------------------------------------------------
There are observations with identical propensity score values.
The sort order of the data could affect your results.
Make sure that the sort order is random before calling psmatch2.
------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+------------------------------------------------------------
            re78  Unmatched | 6.55048374   6.73236973  -.181885992   .965331308    -0.19
                        ATT | 6.55048374   7.39334926  -.842865517   1.94904014    -0.43
                        ATU | 6.73236973   5.20654872  -1.52582101            .        .
                        ATE |                          -7.42042017            .        .
```

```
------------------------------+-----------------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.

           | psmatch2:
 psmatch2: |   Common
 Treatment |  support
assignment | On suppor |     Total
-----------+-----------+----------
 Untreated |       128 |       128
   Treated |        98 |        98
-----------+-----------+----------
     Total |       226 |       226
```


Histograms of PS, Reduced Sample
.2
.1
0
-.1
-.2
0
.2
.4
.6
.8
1
PS
Untreated
Treated

```
. nnmatch re78 train age educ black hisp married unem74 unem75 re74 re75,
          tc(att)

Matching estimator:  Average Treatment Effect for the Treated

Weighting matrix: inverse variance          Number of obs          =       226
                                            Number of matches  (m) =         1

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATT |   1.586371   1.763594     0.90   0.368    -1.870209    5.042951
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married unem74 unem75 re74 re75

. nnmatch re78 train age educ black hisp married unem74 unem75 re74 re75,
          tc(ate)

Matching estimator:  Average Treatment Effect ate

Weighting matrix: inverse variance          Number of obs          =       226
                                            Number of matches  (m) =         1

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATE |   .2957925   1.364623     0.22   0.828    -2.378819    2.970404
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married unem74 unem75 re74 re75
```

```
. * Can use simpler rules for selecting the sample based on covariates that
. * should matter for treatment assignment.

. use jtrain3

. des avgre

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
avgre           float  %9.0g                  (re74 + re75)/2

. sum avgre if train

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       avgre |       185    1.813815    3.679893          0   23.28835

. count if avgre > 10 & train
    6

. keep if avgre <= 10
(1910 observations deleted)
```

```
. tab train

   =1 if in |
        job |
   training |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        586       76.60       76.60
          1 |        179       23.40      100.00
------------+-----------------------------------
      Total |        765      100.00

. reg re78 train, robust

Linear regression                               Number of obs =     765
                                                F(  1,   763) =    8.77
                                                Prob > F      =  0.0032
                                                R-squared     =  0.0091
                                                Root MSE      =  9.0224

------------------------------------------------------------------------------
             |               Robust
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       train |   -2.03739     .68789    -2.96   0.003    -3.387772   -.6870082
       _cons |   8.185696   .3889923    21.04   0.000     7.422074    8.949318
------------------------------------------------------------------------------
```

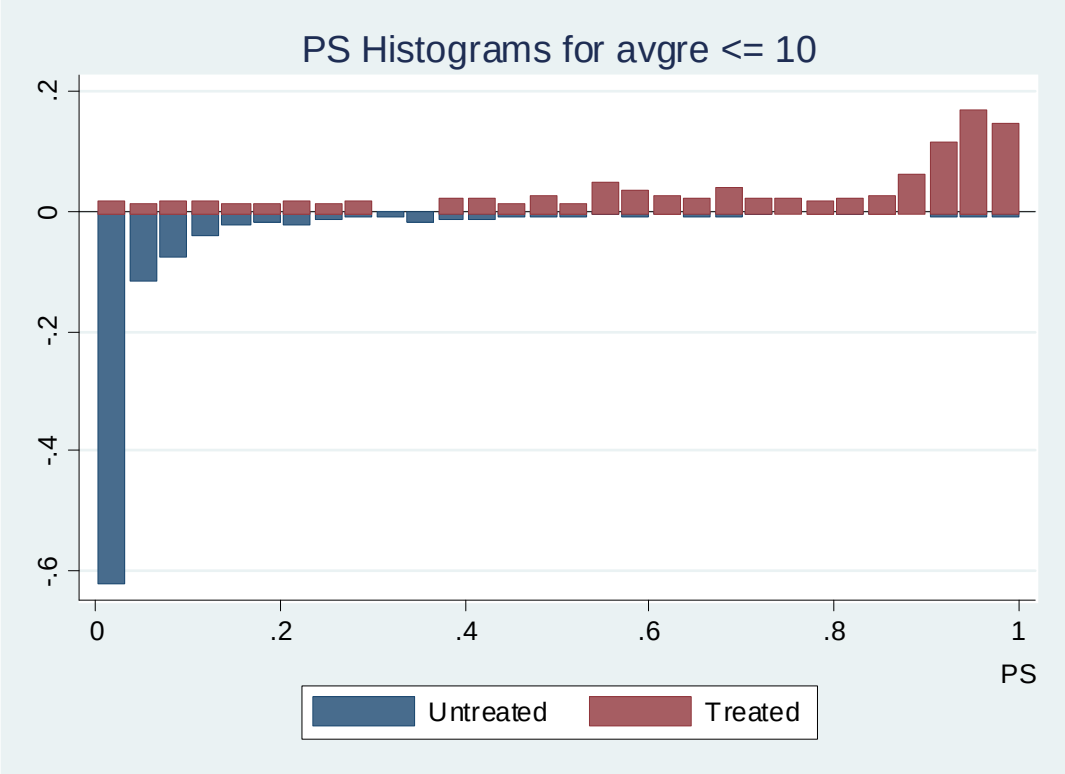
```
. reg re78 train age educ black hisp married unem74 unem75 re74 re75, robust

Linear regression                                      Number of obs =     765
                                                       F( 10,   754) =   27.48
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2472
                                                       Root MSE      =  7.9108

------------------------------------------------------------------------------
             |               Robust
        re78 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       train |   2.229707   .8322748     2.68   0.008     .5958553    3.863558
         age |  -.1258557   .0265928    -4.73   0.000    -.1780604   -.0736509
        educ |   .3048632   .1188998     2.56   0.011     .0714492    .5382773
       black |   -1.51669    .619149    -2.45   0.015    -2.732151   -.3012295
        hisp |  -.9766242   1.084878    -0.90   0.368    -3.106366    1.153117
     married |   2.012687   .6918517     2.91   0.004      .654502    3.370871
      unem74 |    1.69765    1.24419     1.36   0.173    -.7448382    4.140138
      unem75 |   .6463203   1.247726     0.52   0.605     -1.80311     3.09575
        re74 |   .4240549   .1024877     4.14   0.000     .2228597    .6252501
        re75 |   .8437341   .1606773     5.25   0.000     .5283061    1.159162
       _cons |   1.788739   1.998809     0.89   0.371    -2.135155    5.712632
------------------------------------------------------------------------------
```

```
. * Separate regressions:

. bootstrap _b[ate] _b[att], reps(1000) seed(123): ateboot
(running ateboot on estimation sample)

Bootstrap replications (1000)

Bootstrap results                                     Number of obs      =       765
                                                      Replications       =      1000

      command:  ateboot
        _bs_1:  _b[ate]
        _bs_2:  _b[att]

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |  -1.120044   1.429752    -0.78   0.433    -3.922306    1.682218
       _bs_2 |   3.170748   .9101968     3.48   0.000     1.386795    4.954701
------------------------------------------------------------------------------
.
. program drop ateboot

. * ATE estimate is negative but ATT is positive, large, and statistically
. * significant.
```

```
* PS weighting: still does not work well:

. do atepswjtrain3

. capture program drop ateboot

. program ateboot, rclass
  1.
.         * Estimate propensity score
.
.         logit train age educ black hisp married unem74 unem75 re74 re75
  2.         predict phat
  3.         gen kiate = (train - phat)*re78/(phat*(1 - phat))
  4.         sum kiate
  5.         return scalar atew = r(mean)
  6.         sum train
  7.         scalar rho = r(mean)
  8.         gen kiatt = (train - phat)*re78/(1 - phat)
  9.         sum kiatt
 10.         return scalar attw = r(mean)/rho
 11.
.         drop phat kiate kiatt
 12.
. end
```

```
. bootstrap r(atew) r(attw), reps(1000) seed(123): ateboot
(running ateboot on estimation sample)

Bootstrap results                                     Number of obs      =       765
                                                      Replications       =      1000

      command:  ateboot
        _bs_1:  r(atew)
        _bs_2:  r(attw)

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |  -3.021999   2.361112    -1.28   0.201    -7.649694    1.605696
       _bs_2 |  -.4351315   3.279641    -0.13   0.894    -6.863109    5.992846
------------------------------------------------------------------------------

. program drop ateboot
```

```
. psmatch2 train age educ black hisp married unem74 unem75 re74 re75,
           outcome(re78) logit ate

Logistic regression                               Number of obs   =        765
                                                  LR chi2(9)      =     492.18
                                                  Prob > chi2     =     0.0000
Log likelihood = -170.10568                       Pseudo R2       =     0.5913

------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+------------------------------------------------------------
            re78  Unmatched | 6.14830643   8.18569633  -2.03738991   .770511114    -2.64
                        ATT | 6.14830643   3.98120016   2.16710627   1.50085659     1.44
                        ATU | 8.18569633   9.13622201   .950525678            .        .
                        ATE |                           1.23518963            .        .
-----------------------------+------------------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.

           | psmatch2:
 psmatch2: |   Common
 Treatment |  support
assignment | On suppor |     Total
-----------+-----------+----------
 Untreated |       586 |       586
   Treated |       179 |       179
-----------+-----------+----------
     Total |       765 |       765
```


PS Histograms for avgre <= 10
.2
0
-.2
-.4
-.6
0
.2
.4
.6
.8
1
PS
Untreated
Treated

```
. nnmatch re78 train age educ black hisp married unem74 unem75 re74 re75,
          tc(att)

Matching estimator:  Average Treatment Effect for the Treated

Weighting matrix: inverse variance          Number of obs          =       765
                                            Number of matches  (m) =         1

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATT |    2.48946   1.753482     1.42   0.156    -.9473009    5.926221
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married unem74 unem75 re74 re75

. nnmatch re78 train age educ black hisp married unem74 unem75 re74 re75,
          tc(att) biasadj(bias)

Matching estimator:  Average Treatment Effect for the Treated

Weighting matrix: inverse variance          Number of obs          =       765
                                            Number of matches  (m) =         1

------------------------------------------------------------------------------
        re78 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATT |   2.676588   1.796669     1.49   0.136    -.8448188    6.197995
------------------------------------------------------------------------------
Matching variables:  age educ black hisp married unem74 unem75 re74 re75
Bias-adj variables:  age educ black hisp married unem74 unem75 re74 re75
```

Effects of Right Heart Catheterization on Death Rates

- The response variable, $Y_i = death_i$, is binary. Exploit that in regression adjustment. Matching and PS methods do not need to be altered (although one might want to if the methods are combined with regression adjustment).
- Treatment is $W_i = rhc_i$.

```
. tab rhc

      =1 if |
   received |
right heart |
catheteriza |
       tion |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      3,551       61.92       61.92
          1 |      2,184       38.08      100.00
------------+-----------------------------------
      Total |      5,735      100.00

. tab death

      =1 if |
    patient |
died within |
   180 days |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,013       35.10       35.10
          1 |      3,722       64.90      100.00
------------+-----------------------------------
      Total |      5,735      100.00
```

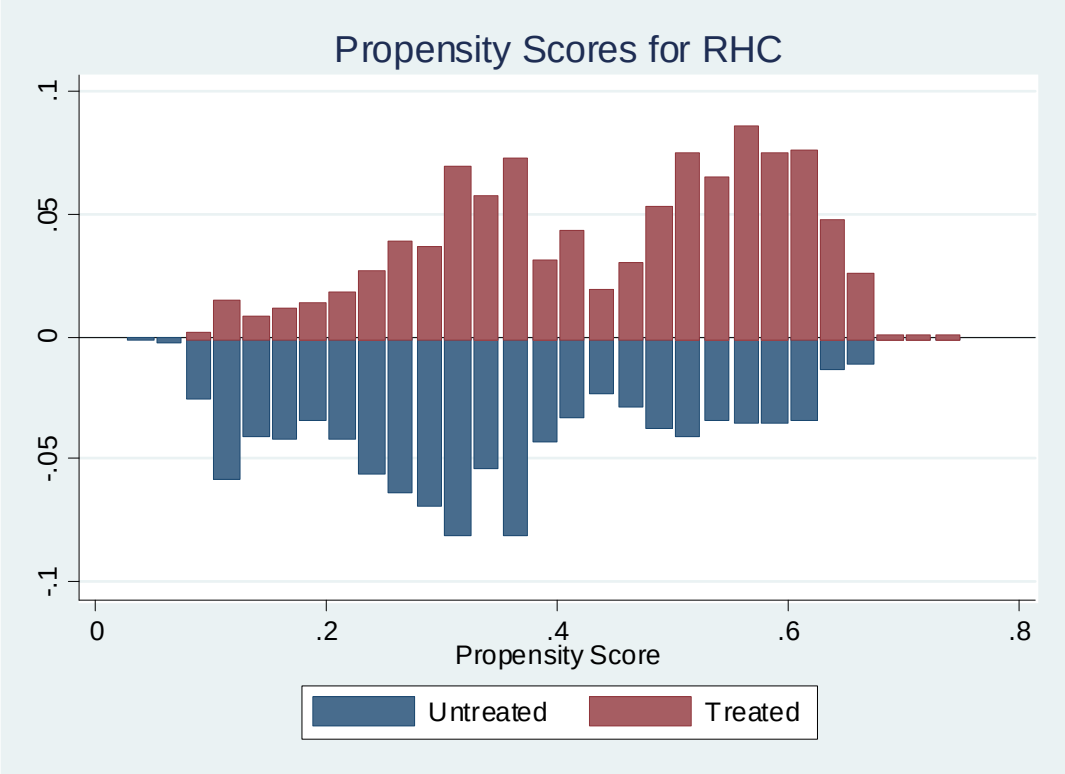
```
. reg death rhc, robust

Linear regression                                      Number of obs =    5735
                                                       F(  1,  5733) =   15.56
                                                       Prob > F      =  0.0001
                                                       R-squared     =  0.0027
                                                       Root MSE      =  .47673

------------------------------------------------------------------------------
             |               Robust
       death |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         rhc |   .0507212   .0128566     3.95   0.000     .0255174    .0759249
       _cons |   .6296818   .0081049    77.69   0.000     .6137931    .6455705
------------------------------------------------------------------------------
```

```
. * Regression adjustment using logit models for death, unrestricted
. * coefficients:

. do atereg_rhc

. capture program drop ateboot
.
. program ateboot, eclass
  1.
.         * Estimate logit on each treatment group
.
.         tempvar touse
  2.         gen byte ‘touse’ = 1
  3.         xi: logit death i.sex i.race i.income i.cat1 i.cat2 i.ninsclas age if rhc
  4.         predict d1hat
  5.         xi: logit death i.sex i.race i.income i.cat1 i.cat2 i.ninsclas age if ~rhc
  6.         predict d0hat
  7.         gen diff = d1hat - d0hat
  8.         sum diff
  9.         scalar ate = r(mean)
 10.         sum diff if rhc
 11.         scalar att = r(mean)
 12.         matrix b = (ate, att)
 13.         matrix colnames b = ate att
 14.         ereturn post b , esample(‘touse’)
 15.         ereturn display
 16.         drop d1hat d0hat diff _I*
 17.
. end
.
```

```
. bootstrap _b[ate] _b[att], reps(1000) seed(123): ateboot
(running ateboot on estimation sample)

Bootstrap results                                     Number of obs      =      5735
                                                      Replications       =      1000

      command:  ateboot
        _bs_1:  _b[ate]
        _bs_2:  _b[att]

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _bs_1 |   .0776176   .0129611     5.99   0.000     .0522143    .1030208
       _bs_2 |   .0656444     .01366     4.81   0.000     .0388713    .0924175
------------------------------------------------------------------------------

. program drop ateboot

. * Controlling for factors only makes the effect larger, not smaller.
```

```
. xi: logit rhc i.female i.race i.income i.cat1 i.cat2 i.ninsclas age
i.female          _Ifemale_0-1        (naturally coded; _Ifemale_0 omitted)
i.race            _Irace_0-2          (naturally coded; _Irace_0 omitted)
i.income          _Iincome_0-3        (naturally coded; _Iincome_0 omitted)
i.cat1            _Icat1_1-9          (naturally coded; _Icat1_1 omitted)
i.cat2            _Icat2_1-7          (naturally coded; _Icat2_1 omitted)
i.ninsclas        _Ininsclas_1-6      (naturally coded; _Ininsclas_1 omitted)

Iteration 0:   log likelihood = -3810.7005
Iteration 1:   log likelihood = -3503.6889
Iteration 2:   log likelihood = -3497.9775
Iteration 3:   log likelihood = -3497.9617
Iteration 4:   log likelihood = -3497.9617

Logistic regression                               Number of obs   =       5735
                                                  LR chi2(26)     =     625.48
                                                  Prob > chi2     =     0.0000
Log likelihood = -3497.9617                       Pseudo R2       =     0.0821

------------------------------------------------------------------------------
         rhc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _Ifemale_1 |   .1630495   .0587579     2.77   0.006     .0478861    .2782129
    _Irace_1 |   .0424279   .0827591     0.51   0.608    -.1197771    .2046328
    _Irace_2 |   .0393684   .1361998     0.29   0.773    -.2275784    .3063151
  _Iincome_1 |   .0443619   .0762726     0.58   0.561    -.1051296    .1938535
  _Iincome_2 |    .151793   .0892757     1.70   0.089    -.0231841    .3267701
  _Iincome_3 |   .1579471   .1140752     1.38   0.166    -.0656361    .3815303
    _Icat1_2 |    .498032    .107388     4.64   0.000     .2875553    .7085086
    _Icat1_3 |  -1.226306   .1495545    -8.20   0.000    -1.519428   -.9331849
    _Icat1_4 |  -.7173791   .1714465    -4.18   0.000    -1.053408   -.3813501
    _Icat1_5 |  -1.002513   1.085305    -0.92   0.356    -3.129671    1.124645
    _Icat1_6 |  -.6941957   .1260198    -5.51   0.000      -.94119   -.4472013
    _Icat1_7 |  -1.258815   .4833701    -2.60   0.009    -2.206203   -.3114273
    _Icat1_8 |  -.2076635   .1177652    -1.76   0.078     -.438479    .0231519
```

```
    _Icat1_9 |   1.003787   .0768436    13.06   0.000     .8531766    1.154398
    _Icat2_2 |   .9804654   1.465085     0.67   0.503    -1.891048    3.851979
    _Icat2_3 |  -.4141065   .4428411    -0.94   0.350    -1.282059    .4538461
    _Icat2_4 |  -.8864827   .8454718    -1.05   0.294    -2.543577    .7706116
    _Icat2_5 |   -.195389   .3933026    -0.50   0.619     -.966248      .57547
    _Icat2_6 |   1.034498    .369503     2.80   0.005     .3102859    1.758711
    _Icat2_7 |   .1415088   .3649828     0.39   0.698    -.5738443    .8568619
_Ininsclas_2 |   .1849583   .1216214     1.52   0.128    -.0534153    .4233318
_Ininsclas_3 |   .1082916    .152243     0.71   0.477    -.1900992    .4066824
_Ininsclas_4 |   .5216726   .1495659     3.49   0.000     .2285288    .8148164
_Ininsclas_5 |    .468176   .1122184     4.17   0.000      .248232    .6881199
_Ininsclas_6 |   .3742273   .1249122     3.00   0.003     .1294038    .6190508
         age |   .0006419    .002252     0.29   0.776    -.0037719    .0050557
       _cons |   -1.36677   .3834979    -3.56   0.000    -2.118412   -.6151284
------------------------------------------------------------------------------

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

. sum phat

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        phat |      5735    .3808195    .1564252   .0435625   .7379614

. qui xi: psmatch2 rhc i.female i.race i.income i.cat1 i.cat2 i.ninsclas age,
      outcome(death)
Note: S.E. does not take into account that the propensity score is estimated.

. psgraph
```


Propensity Scores for RHC
.1
.05
0
-.05
-.1
0
.2
.4
.6
.8
Propensity Score
Untreated
Treated

```
. xi: psmatch2 rhc i.female i.race i.income i.cat1 i.cat2 i.ninsclas age,
      outcome(death) logit ate


There are observations with identical propensity score values.
The sort order of the data could affect your results.
Make sure that the sort order is random before calling psmatch2.
------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+------------------------------------------------------------
           death  Unmatched |  .68040293    .62968178   .050721151   .012963982     3.91
                        ATT |  .68040293   .619505495   .060897436   .019021326     3.20
                        ATU |  .62968178   .705153478   .075471698            .        .
                        ATE |                           .069921534            .        .
-----------------------------+------------------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.

           | psmatch2:
 psmatch2: |   Common
 Treatment |  support
assignment | On suppor |     Total
-----------+-----------+----------
 Untreated |     3,551 |     3,551
   Treated |     2,184 |     2,184
-----------+-----------+----------
     Total |     5,735 |     5,735
```