TREATMENT EFFECT ESTIMATION WITH UNCONFOUNDED ASSIGNMENT

Jeff Wooldridge
Michigan State University
FARS Workshop, Chicago
January 6, 2012

1. Introduction

- What kinds of questions can we answer using a “modern” approach to treatment effect estimation?

1. Does using a Big 4 auditing firm affect the quality of audits?

2. Does using a talent agent affect CEO compensation?

3. Do CEO equity incentives affect the incidence of accounting irregularities?

- The main issue in treatment effect estimation concerns the nature of the assignment, intervention, or “treatment.”
- Is the “treatment” randomly assigned? Rarely in accounting applications. Can we assume the decision of whether to use a Big 4 firm is independent of all other factors (observed and unobserved)? Is using equity incentives unrelated to other factors that affect accounting irregularities?

- Without experimental data cannot assume random assignment. With observational (or retrospective) data, it may be reasonable to assume that treatment is effectively randomly assigned conditional on observable covariates.
- Called “unconfoundedness” or “ignorability” of treatment or “selection on observables.” Sometimes called “exogenous treatment.”

- Does assignment depend fundamentally on unobservables, where the dependence cannot be broken by controlling for observables?
- Called “confounded” assignment or “selection on unobservables” or “endogenous treatment.”
- Often there is a self-selection component to treatment assignment based on factors researchers cannot observe. Firms decide which auditing firm to use or how to structure CEO compensation.

- Three situations:

(1) Assume unconfoundedness of treatment, and then explore how to exploit it in estimation. (Leads to regression, propensity score, and matching methods.)

(2) Allow self-selection on unobservables and exploit exogenous instrumental variables. (Switching regression.)

(3) Exploit a “regression discontinuity” design, where the treatment (or its probability) is determined as a discontinuous function of an observed “forcing” variable.

- This workshop focuses on (1).

- The canonical setup is where there is a pre-treatment period and then an intervention, where some units are subjected to the treatment. The pre-treatment observables are used as controls to predict treatment assignment.
- Other scenarios, including panel data, can be handled, but one must be clear in defining the treatement effects and the nature of unconfoundedness assumptions.

- KEY POINT: Under unconfoundedness, regression methods identify treatment effect parameters, just as do propensity score weighting and matching approaches.
- Propensity score methods are not a panacea for the self-selection problem. They suffer systematic bias in cases where standard regression methods do.
- Practically, matching methods seem to work better than regression and weighting methods in some situations.

- Unconfoundedness is fundamentally untestable without extra information – just as with standard regression analysis. (After OLS estimation, we generally have no way of deciding whether the explanatory variables are “exogenous” without outside information.)
- In some cases there are ways to assess the plausibility of unconfoundedness, or study sensitivity of estimates.

- A second key assumption is “overlap,” which concerns the similarity of the covariate distributions for the treated and control 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.

Caution Concerning Data Structures

- Standard methods for estimating ATEs are derived for a cross-sectional setting (possibly with controls that are lagged outcomes).
- Panel data must be treated carefully. At a minimum, time series dependence must be accounted for in inference (not easily done with matching). There are conceptual issues, such as matching firms across time.
- With panel data, a standard “fixed effects” analysis may be more convincing, depending on how unconfoundedness is used.

2. Basic Concepts

Counterfactual Outcomes and Parameters of Interest

- Assume a binary treatment W – so two possible states of the world, $W = 0$ and $W = 1$.
- For each population unit, two potential outcomes: $Y(0)$ (the outcome without treatment) and $Y(1)$ (the outcome with treatment).
- $Y(0)$ and $Y(1)$ can be discrete, continuous, or some mix.

- $Y(0)$ is CEO compensation without using a talent agent, $Y(1)$ is compensation using an agent. (Continuous response.)
- $Y(0)$ is binary indicator of whether there is an accounting irregularity using a non-Big 4 firm, $Y(1)$ is the indicator using a Big 4 firm.

- The gain from treatment is

$$Y(1) - Y(0). \tag{1}$$

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

$$Y_i(1) - Y_i(0). \tag{2}$$

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

- Key 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. We will assume a random sample of units from the population, but we do not observe both outcomes.

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

$$\tau_{ate} = E[Y(1) - Y(0)]. \tag{3}$$

The expected gain for a randomly selected unit from the population.

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

$$\tau_{att} = E[Y(1) - Y(0)|W = 1] \tag{4}$$

• With heterogeneous treatment effects – that is, $Y_i(1) - Y_i(0)$ not constant – 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 discussion of the nature of the treatment. These definitions hold when whether assignment is randomized, unconfounded, or endogenous.

- How we *estimate* τ_{ate} and τ_{att} depends on what we assume about treatment assignment.
- We will also define ATEs and ATTs conditional on a set of observed covariates. Some approaches to estimating τ_{ate} and τ_{att} rely on first estimating *conditional* average treatment effects.
- Occasionally it is helpful to define average treatment effects in a sample, for example,

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

Sampling Assumptions

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

$$\begin{aligned} Y_i &= (1 - W_i)Y_i(0) + W_i Y_i(1) = Y_i(0) + W_i[Y_i(1) - Y_i(0)] \\ &= Y_i(0) + W_i \cdot Gain_i \end{aligned} \tag{5}$$

- 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: one unit’s treatment status has no effect on another unit’s outcome.
- Random sampling across all observations rules out panel data.

Estimation under Random Assignment

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

$$E(Y|W = 1) - E(Y|W = 0) = E[Y(1)] - E[Y(0)] = \tau_{ate} = \tau_{att} \qquad (6)$$

• $E(Y|W = 1)$ and $E(Y|W = 0)$ can be estimated by using sample averages on the two subsamples.

$$\hat{\tau}_{ate} = \bar{Y}_1 - \bar{Y}_0.$$

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

Many Treatment Levels

- If the treatment W_i takes on $G + 1$ levels numbered $\{0, 1, \ldots, G\}$, 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.
- Many estimation methods extend directly.

- Define the counterfactual means as

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

- The expected gain in going from treatment level $g - 1$ to g is $\mu_g - \mu_{g-1}$.
- If $Y(0)$ is the response under no treatment then $\mu_g - \mu_0$ is the average gain of treatment level g relative to no treatment.

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 representing the population distribution.
- The strongest form of unconfoundedness: Conditional on $\mathbf{X}$, the counterfactual outcomes are independent of W.
- In other words, for units in the subpopulation defined by $\mathbf{X} = \mathbf{x}$, assignment is randomized.

A.1. **Unconfoundedness**: Conditional on a set of covariates **X**, the pair of counterfactual outcomes, $[Y(0), Y(1)]$, is independent of W:

$$[Y(0), Y(1)] \perp W \mid \mathbf{X}, \tag{7}$$

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

- Can also write unconfoundedness, or ignorability, as

$$D[W|Y(0), Y(1), \mathbf{X}] = D(W|\mathbf{X}), \tag{8}$$

where $D(\bullet|\bullet)$ denotes conditional distribution.

- Unconfoundedness is controversial. It underlies standard regression methods to estimating treatment effects (via a “kitchen sink” regression that includes the treatment indicator along with controls).
- Essentially, unconfoundedness leads to a difference-in-means after adjusting for observed covariates.
- Even if we doubt we have “enough” of the “right” covariates, we probably would attempt such a comparison.

• A weaker version of unconfoundedness:

A.1′. Unconfoundedness in Conditional Mean:

$$E[Y(g)|W, \mathbf{X}] = E[Y(g)|\mathbf{X}], \; g = 0, 1. \tag{9}$$

• 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.

- IMPORTANT: Unconfoundedness is generally violated if **X** includes variables that are themselves affected by the treatment.
- For example, in studying whether hiring a talent agent affects CEO compensation, should not include in **X** variables possibly affected by "treatment," such as number of job interviews.

• An extreme case is where $[Y(0), Y(1)]$ is independent of W but $\mathbf{X}$ is not independent of W. (Think of assignment being randomized but then $\mathbf{X}$ includes a post-assignment variable that can be affected by the treatment.)

• Can show that unconfoundedness generally fails unless

$$E[Y(g)|\mathbf{X}] = E[Y(g)], \; g = 0, 1. \tag{10}$$

• Bottom line: It is not always better to put everything possible in $\mathbf{X}$.

- A compelling argument in favor of an analyisis based on unconfoundedness is that the quantities we need to estimate are *nonparametrically identified*. 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,

$$Y(0) \perp W \mid \mathbf{X} \tag{11}$$

or even the mean version,

$$E[Y(0)|W, \mathbf{X}] = E[Y(0)|\mathbf{X}] \tag{12}$$

• This allows the unit-specific gain, $Y_i(1) - Y_i(0)$, to depend on treatment status W_i in an arbitrary way.

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

$$0 < P(W = 1|\mathbf{X} = \mathbf{x}) < 1. \tag{13}$$

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

$$p(\mathbf{x}) = P(W = 1|\mathbf{X} = \mathbf{x}). \tag{14}$$

- **Strong Ignorability** [Rosenbaum and Rubin (1983)] = Unconfoundedness plus Overlap.
- For ATT, overlap can be relaxed to

$$p(\mathbf{x}) < 1 \text{ for all } \mathbf{x} \in \mathcal{X} \tag{15}$$

We can have $p(\mathbf{x}) = 0$ because we average only over the treated subpopulation.

4. Identification of Average Treatment Effects

- 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

$$\tau(\mathbf{x}) = E[Y(1) - Y(0)|\mathbf{X} = \mathbf{x}] = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) \tag{16}$$

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: it provides the average treatment effect for different segments of the population described by the observables, $\mathbf{X}$.

- By iterated expectations, it is always true that

$$\tau_{ate} = E[Y(1) - Y(0)] = E[\tau(\mathbf{X})] = E[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})] \quad (17)$$

- It follows that τ_{ate} is identified if $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are identified over the support (possibly values) of $\mathbf{X}$, because we observe a random sample on $\mathbf{X}$ and can average across its distribution.

- $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are identified under unconfoundedness and overlap:

$$\begin{aligned} E(Y|\mathbf{X}, W) &= (1-W)E[Y(0)|\mathbf{X}, W] + WE[Y(1)|\mathbf{X}, W] \\ &= (1-W)E[Y(0)|\mathbf{X}] + WE[Y(1)|\mathbf{X}] \\ &\equiv (1-W)\mu_0(\mathbf{X}) + W\mu_1(\mathbf{X}), \end{aligned} \tag{18}$$

where the second equality holds by unconfoundedness.

- Define the conditional means of the observed outcome as

$$m_0(\mathbf{X}) = E(Y|\mathbf{X}, W = 0), m_1(\mathbf{X}) = E(Y|\mathbf{X}, W = 1) \tag{19}$$

- 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

$$E(Y|\mathbf{X}, W = 0) = \mu_0(\mathbf{X}), E(Y|\mathbf{X}, W = 1) = \mu_1(\mathbf{X}) \tag{20}$$

- For ATT,

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

- Therefore,

$$\tau_{att} = E[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})|W = 1], \tag{21}$$

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

- In terms of the identified mean functions,

$$\tau_{ate} = E[m_1(\mathbf{X}) - m_0(\mathbf{X})] \tag{22}$$

$$\tau_{att} = E[m_1(\mathbf{X}) - m_0(\mathbf{X})|W = 1] \tag{23}$$

- 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. 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. Assuming unconfoundedness, can show [Wooldridge (2010, Chapter 21)]

$$E\left[\frac{WY}{p(\mathbf{X})}\right] = E[Y(1)] \tag{24}$$

$$E\left[\frac{(1-W)Y}{1-p(\mathbf{X})}\right] = E[Y(0)] \tag{25}$$

- Note where $p(\mathbf{x}) > 0$ and $p(\mathbf{x}) < 1$ come into play.
- Combining,

$$\tau_{ate} = E\left[\frac{WY}{p(\mathbf{X})} - \frac{(1-W)Y}{1-p(\mathbf{X})}\right] = E\left\{\frac{[W-p(\mathbf{X})]Y}{p(\mathbf{X})[1-p(\mathbf{X})]}\right\}. \tag{26}$$

• Can also show

$$\tau_{att} = E\left\{ \frac{[W - p(\mathbf{X})]Y}{\rho[1 - p(\mathbf{X})]} \right\}, \tag{27}$$

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

• Only need $p(\mathbf{x}) < 1$ because τ_{att} is an average effect for those eventually treated.

• The unconfoundedness and overlap assumptions to identify τ_{att} are weaker than for τ_{ate}.

5. Estimating ATEs

- When we assume unconfounded treatment and overlap, there are three general approaches to estimating the treatment effects:

1. Regression-based methods (on covariates or propensity score).
2. Propensity score weighting methods.
3. Matching methods (on covariates or propensity score).

- Can mix the various approaches, and often this helps.

- Need to remember that all methods assume under unconfoundedness and overlap. Even if unconfoundedness holds, they may behave quite differently when overlap is weak.
- Under regularity conditions, asymptotically efficient estimators exist under unconfoundedness and overlap. Using the propensity score in regression or matching on are not efficient.

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: We compute a fitted value for each outcome for *all* units in sample. For example, we only use the control units to obtain $\hat{m}_0(\cdot)$ but we need $\hat{m}_0(\mathbf{X}_i)$ for the treated units, too.

- The regression-adjustment estimators are

$$\hat{\tau}_{ate,reg} = N^{-1} \sum_{i=1}^{N} [\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)] \tag{28}$$

$$\hat{\tau}_{att,reg} = N_1^{-1} \sum_{i=1}^{N} W_i [\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)] \tag{29}$$

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

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

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

$$Y_i \text{ on } 1,\ W_i,\ \mathbf{X}_i,\ W_i \cdot (\mathbf{X}_i - \bar{\mathbf{X}}),\ \ i = 1, \ldots, N. \tag{30}$$

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

- Important to demean $\mathbf{X}_i$ before forming interaction, so that the coefficient on W_i is the estimate of τ_{ate}.
- Centering the covariates before constructing the interactions is known to often "solve" the multicollinearity problem in regression. It "solves" the problem because it redefines the parameter we are trying to estimate to be the ATE.

- With lots of covariates, might compute differences in fitted values, and average.
- In Stata:

```
reg y x1 x2 ... xK if ~treat
predict y0hat
reg y x1 x2 ... xK if treat
predict y1hat
gen ate_i = y1hat - y0hat
sum ate_i
```

• 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. It can be obtained from

$$Y_i \text{ on } 1,\ W_i,\ \mathbf{X}_i,\ W_i \cdot (\mathbf{X}_i - \bar{\mathbf{X}}_1),\ i = 1, \ldots, N. \tag{31}$$

• Or, just average the difference in fitted values over the treated subsample:

```
sum ate_i if treat
```

• 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:

$$\hat{\tau}_{ate,reg} = N^{-1} \sum_{i=1}^{N} [G(\hat{\alpha}_1 + \mathbf{X}_i\hat{\boldsymbol{\beta}}_1) - G(\hat{\alpha}_0 + \mathbf{X}_i\hat{\boldsymbol{\beta}}_0)]. \tag{32}$$

• Each summand is the difference in estimate probabilities under treatment and nontreatment for unit i, and the ATE just averages those differences.

- In Stata (for binary or fractional response):

```
glm y x1 x2 ... xK if ~treat, fam(bin)
link(logit)
predict y0hat
glm y x1 x2 ... xK if treat, fam(bin)
link(logit)
predict y1hat
gen atei = y1hat - y0hat
sum atei
```

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

$$\hat{\tau}_{ate,reg} = N^{-1} \sum_{i=1}^{N} [\exp(\hat{\alpha}_1 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_1) - \exp(\hat{\alpha}_0 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_0)]. \quad (33)$$

• In previous `glm` commands, substitute

`fam(poisson) link(log)`

• In nonlinear cases, can use delta method or bootstrap for standard error of $\hat{\tau}_{ate,reg}$. The delta method formula can be found in Wooldridge (2010, Chapter 21).

- 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 only large firms offer equity incentives. Estimating models of accounting irregularities will run into trouble if firm size is in **X**. We have to estimate $E(Y|\mathbf{X}, W = 1)$ using only firms with equity incentives and then extrapolate the mean function to firms not offering incentives.

• Nonparametric methods – see Imbens and Wooldridge (2009, *Journal of Economic Literature*) 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 ("kernel smoothing"), we cannot easily estimate $m_1(\mathbf{x})$ for $\mathbf{x}$ values far away from those in the treated subsample; similarly for $m_0(\mathbf{x})$.

• 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:

$$\hat{\tau}_{att,reg} = N_1^{-1} \sum_{i=1}^{N} W_i[\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)]. \quad (34)$$

- 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. We do not need to estimate $m_1(\mathbf{x})$ for values of $\mathbf{x}$ in the control group.
- We still need to estimate $\hat{m}_0(\mathbf{x})$ for the treated group.

Propensity Score Weighting

- The propensity score formula that establishes identification of τ_{ate} suggests an estimator of τ_{ate}:

$$\tilde{\tau}_{ate,psw} = N^{-1} \sum_{i=1}^{N} \left[\frac{W_i Y_i}{p(\mathbf{X}_i)} - \frac{(1 - W_i) Y_i}{1 - p(\mathbf{X}_i)} \right].$$

- $\tilde{\tau}_{ate,psw}$ is not feasible because it depends on the unknown 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 (probit or logit). Can show estimating the propensity score leads to a *smaller* asymptotic variance when the parametric model is correctly specified. (2) Use an explicit nonparametric approach, as in Hirano, Imbens, and Ridder (2003, *Econometrica*) or Li, Racine, and Wooldridge (2009, *JBES*).

$$\hat{\tau}_{ate,psw} = N^{-1} \sum_{i=1}^{N} \frac{[W_i - \hat{p}(\mathbf{X}_i)]Y_i}{\hat{p}(\mathbf{X}_i)[1 - \hat{p}(\mathbf{X}_i)]} \tag{35}$$

$$\hat{\tau}_{att,psw} = N^{-1} \sum_{i=1}^{N} \frac{[W_i - \hat{p}(\mathbf{X}_i)]Y_i}{\hat{\rho}[1 - \hat{p}(\mathbf{X}_i)]} \tag{36}$$

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 $p(\mathbf{x})$ close to unity matters.)

• These formulas show why trimming observations based on the PS is often used.

• Wooldridge (2010, Chapter 21) shows how to adjust the standard errors to account for estimation of the PS. It is easiest for a logit model.

• Write

$$\hat{k}_i \equiv \frac{[W_i - \hat{p}(\mathbf{X}_i)]Y_i}{\hat{p}(\mathbf{X}_i)[1 - \hat{p}(\mathbf{X}_i)]} \tag{37}$$

• The estimate $\hat{\tau}_{ate}$, along with a proper standard error, are obtained from the *intercept* in the regression

$$\hat{k}_i \text{ on } 1, \ \mathbf{X}_i(W_i - \hat{p}_i), i = 1, \ldots, N. \tag{38}$$

• Because the constant and $\mathbf{X}_i(W_i - \hat{p}_i)$ are orthogonal (FOC for logit), adding $\mathbf{X}_i(W_i - \hat{p}_i)$ does not change the intercept. But it reduces the sum of squared residuals.

- Also shows a result known in biostatistics: Asymptotically, we do no worse – and typically better – by adding covariates to the logit estimation even if they do not predict treatment!

- In Stata:

```
logit treat x1 x2 ... xK
predict phat
gen ehat = treat - phat
gen khat = ehat*y/(phat*(1 - phat))
```

```
gen x1eh = x1*eh
gen x2eh = x2*eh
⋮
gen xKeh = xK*eh
reg khat eh x1eh x2eh ... xKeh
```

- Obtain estimate and standard error as coefficient and standard error on intercept.

• An alternative is to use bootstrapping, where the propensity score estimation and averaging (to get $\hat{\tau}_{ate,psw}$) are included in each bootstrap iteration. Practical problem: Some bootstrap samples may give fitted probabilities that are zero or one.

- Conservative to ignore the estimation error in the $\hat{k}_i$ and simply treat it as a random sample. That corresponds to just running the regression

$$\hat{k}_i \text{ on } 1,\ i = 1,\ldots,N.$$

- For $\hat{\tau}_{att,psw}$, adjustment to standard error somewhat different (Wooldridge, 2010, Chapter 21). Bootstrapping still can be used.

- Can see directly from $\hat{\tau}_{ate,psw}$ and $\hat{\tau}_{att,psw}$ that the inverse probability weighted (IPW) estimators can be 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). Drop observations and start again on new, smaller sample.
- 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

• Key result of Rosenbaum and Rubin (1983): Given unconfoundedness conditional on $\mathbf{X}$, unconfoundedness holds if we condition only on $p(\mathbf{X})$:

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

and so

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

• After estimating $p(\cdot)$ by logit or probit, we estimate $E[Y|p(\mathbf{X}), W = 0]$ and $E[Y|p(\mathbf{X}), W = 1]$ using each subsample. For $\hat{\tau}_{ate}$, use the average difference in fitted values, as before.

• In the linear case, can obtain $\hat{\tau}_{ate}$ as the coefficient on W_i from the pooled regression

$$Y_i \text{ on } 1,\ W_i,\ \hat{p}(\mathbf{X}_i),\ W_i \cdot [\hat{p}(\mathbf{X}_i) - \hat{\mu}_{\hat{p}}],\ i = 1, \ldots, N \tag{39}$$

where $\hat{\mu}_{\hat{p}} = N^{-1} \sum_{i=1}^{N} \hat{p}(\mathbf{X}_i)$.

• The inference ignoring estimation of $p(\cdot)$ is conservative. Can use the bootstrap to obtain the smaller (valid) standard errors.

- Linear regression estimates such as (35) 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,

$$\hat{r}_i \equiv \log\left[\frac{\hat{p}(\mathbf{X}_i)}{1-\hat{p}(\mathbf{X}_i)} \right], \tag{40}$$

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 $\hat{\tau}_{ate,regprop}$.

- Theoretically, regression on the propensity score in regression has little to offer compared with other methods. It is not asymptotically efficient.

Combining Regression Adjustment and PS Weighting

- Question: Why use regression adjustment combined with PS weighting?
- Answer: With **X** having large dimension, still common to rely on parametric methods for regression and PS estimation. 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 logit or probit and obtain the estimated propensity scores as $p(\mathbf{X}_i, \hat{\boldsymbol{\gamma}})$.

- In the second step, we use regression or a quasi-likelihood method, where we weight by the inverse probability.
- If the conditional means are linear, solve the WLS problem

$$\min_{\alpha_1,\beta_1} \sum_{i=1}^{N} W_i(Y_i - \alpha_1 - \mathbf{X}_i\boldsymbol{\beta}_1)^2/p(\mathbf{X}_i,\hat{\boldsymbol{\gamma}}); \tag{41}$$

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

• ATE is estimated exactly as in the regression adjustment case, but with different estimates of $\alpha_g, \boldsymbol{\beta}_g$:

$$\hat{\tau}_{ate,pswreg} = N^{-1} \sum_{i=1}^{N} [(\hat{\alpha}_1 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_1) - (\hat{\alpha}_0 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_0)]. \quad (42)$$

• 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 also Wooldridge (2007, *Journal of Econometrics*)].

- Y continuous, negative and positive values: linear mean, least squares objective function, as above.
- Y binary or fractional: logit mean (not probit!), Bernoulli quasi-log likelihood:

$$\min_{\alpha_1,\boldsymbol{\beta}_1} \sum_{i=1}^{N} W_i\{(1-Y_i)\log[1-\Lambda(\alpha_1+\mathbf{X}_i\boldsymbol{\beta}_1)] + Y_i\log[\Lambda(\alpha_1+\mathbf{X}_i\boldsymbol{\beta}_1)]\}/p(\mathbf{X}_i,\hat{\boldsymbol{\gamma}}). \tag{43}$$

- 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:

$$\hat{\tau}_{ate,pswreg} = N^{-1} \sum_{i=1}^{N} [\Lambda(\hat{\alpha}_1 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_1) - \Lambda(\hat{\alpha}_0 + \mathbf{X}_i \hat{\boldsymbol{\beta}}_0)]. \quad (44)$$

- Sample Stata commands:

```
logit treat x1 x2 ... xK

predict phat

gen wght0 = 1/(1 - phat) if ~treat

gen wght1 = 1/phat if treat
```

- Linear Model:

```
reg y x1 x2 ... xK [pweight = wght0] if ~treat
predict yh0
reg y x1 x2 ... xK [pweight = wght1] if treat
predict yh1
gen ate_i = yh1 - yh0
sum ate_i
```

- For standard error, can bootstrap.

- Logistic model (binary or fractional Y):

```
glm y x1 x2 ... xK [pweight = wght0] if ~treat,
fam(bin) link(logit)
predict yh0
glm y x1 x2 ... xK [pweight = wght1] if treat,
fam(bin) link(logit)
predict yh1
gen ate_i = yh1 - yh0
sum ate_i
```

- Y nonnegative, including count, continuous, or corners at zero:

exponential mean, Poisson quasi-MLE:

```
fam(poisson) link(log)
```

- In each case, must include a constant in the index models for $E(Y|W, \mathbf{X})$ – the default.
- Rather than bootstrap, Stata module called `dr` can be used.

```
dr y treat, ovars(x1 ... xK) pvars(x1 ... xK)
dr y treat, ovars(x1 ... xK) pvars(x1 ... xK)
fam(bin) link(logit)
```

Matching

- Can match on a set of covariates or estimated propensity scores. Matching on a large set of covariates can be very computationally intensive.
- 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

$$\hat{\tau}_{ate,match} = N^{-1} \sum_{i=1}^{N} [\hat{Y}_i(1) - \hat{Y}_i(0)] \tag{45}$$

- Looks like regression adjustment except the imputed values are not fitted values from regression. If we observe $Y_i(1)$ then we use it, and same for $Y_i(0)$.

- For τ_{att},

$$\hat{\tau}_{att,match} = N_1^{-1} \sum_{i=1}^{N} W_i[Y_i - \hat{Y}_i(0)], \tag{46}$$

where this expression uses the fact that $Y_i = Y_i(1)$ for the treated subsample, and so we never have to impute $Y_i(1)$.

• For covariate matching, Abadie and Imbens (2006, *Econometrica*) consider several approaches. Simplest is to find a single match for each observation. Suppose i is a treated observation ($W_i = 1$). Then

$$\hat{Y}_i(1) = Y_i$$
$$\hat{Y}_i(0) = Y_h \text{ for } h \text{ such that } W_h = 0$$

and unit h is "closest" to unit i based on some metric (distance) in the covariates.

• For a treated unit i we find the "most similar" control observation, and use its response as $Y_i(0)$.

• If $W_i = 0$, $\hat{Y}_i(0) = Y_i$ and $\hat{Y}_i(1) = Y_h$ 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. (Diagonal Mahalanobis.)
- With a single variable to match on, the distance is just absolute value.

```
nnmatch y treat x1 x2 ... xK, tc(ate)
nnmatch y treat x1 x2 ... xK, tc(att)
```

- We can impute the missing values using an average of M nearest neighbors, or using all units h within a certain distance i – so-called "caliper matching."
- By choosing a distance and at most one match, can get no matches for some units.
- Variance estimation of matching estimators is tricky. It is easiest to use a conditional variance formula – programmed in `nnmatch` – but this changes the "parameter" of interest to τ_{sate}.
- Estimating the asymptotic variance when the parameter is τ_{ate} is hard due to a conditional bias.

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. Bias dominates the variance asymptotically when $K \geq 3$.

• The bias of the matching estimator comes from terms of the form $m_w(\mathbf{X}_{h(i)}) - m_w(\mathbf{X}_i)$ where $m_w(\mathbf{x}) = E(Y|\mathbf{X}, W = w)$.

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

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

- The bias-corrected matching estimator is

$$\hat{\tau}_{ate,bcme} = N^{-1} \sum_{i=1}^{N} [\tilde{Y}_i(1) - \tilde{Y}_i(0)] \tag{47}$$

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

`nnmatch y treat x1 x2 ... xK, tc(ate) biasadj(bias)`

- Unless one has strong priors, include the same covariates in propensity score and regression models. (This is imposed by `nnmatch`.) Exclusion restrictions do not help with unconfoundedness, and they can hurt.

Matching on the Propensity Score

- Matching on a full set of covariates can be computationally demanding. Instead, can match on the estimated propensity score – 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 the propensity score. Bootstrapping not justified, but this is how Stata currently computes the standard errors.

- 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.
- Useful conclusion: 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. Panel Data

- Lots of accounting applications with panel data. Several issues to address.

1. How are average treatment effects defined with panel data and an arbitrary pattern of treatments? Might there be persistent effects? Simpler to assume there are only static effects or that only those effects are of interest.

2. What covariates should be conditioned on? Many applications appear to ignore lagged outcomes in covariates. Can use all information in a sequential approach.

3. In matching approaches, should the matches be restricted? Without restrictions, can match two firms in the same time period, two firms in different time periods, or the same firm across different time periods.

- If matching is unrestricted, need the covariates to satisfy a “strict exogeneity” assumption. Then matching should be based on entire history of covariates or functions of them.
- In a pooled matching strategy, using lags of covariates does not solve lack of strict exogeneity. Covariates that are affect by past shocks cannot be conditioned on.

- A fixed effects approach effectively restricts matches to within a firm. Think of $T = 2$. FE is the same as differencing, and units without a change in treatment status do not affect estimation.
- Fixed effects estimation can be derived from a counterfactual framework where unconfoundedness holds conditional on unobserved heterogeneity and the history of covariates.
- Allowing within-firm matching similar to fixed effects for firms that switch: match is very likely to be within a firm over time.

• For firms with the same treatment status over time, how believable is matching (and the overlap assumption)? We cannot allow matching on an unobserved firm effect.

4. Statistical inference: How do we adjust usual statistics for time series dependence in the panel data? Difficult with matching. Use “cluster robust” options with regression or PS weighting.

• When a panel is treated as a long cross section with unrestricted matching, the usual bootstrap is incorrect: it ignores dependence across time. Need panel bootstrap (resampling all time periods for each *i*).

- Sometimes see PS matching done based only on first-period variables. Avoid this. Much less convincing than a fixed effects analysis – which allows unobserved time-constant covariates – or a sequential analysis that allows the propensity score to depend on the recent past.

- Defining and estimating general “dynamic treatment effects” is notationally hard; Wooldridge (2010, Chapter) contains a brief treatment with references.
- A sequential setup, where at time t the counterfactual outcomes are $[Y_t(0), Y_t(1)]$, is easier.

• Two approaches. First, assume unconfoundedness of the entire sequence of treatments, $\{W_{it} : t = 1, \ldots, \}$, conditional on unobserved heterogeneity, say $\mathbf{c}_i$, and a history of covariates, $\{\mathbf{X}_{it} : t = 1, \ldots, T\}$:

$$E[Y_{it}(0)|\mathbf{W}_i, \mathbf{X}_i, \mathbf{c}_i] = E[Y_{it}(0)|\mathbf{X}_i, \mathbf{c}_i]$$
$$E[Y_{it}(1)|\mathbf{W}_i, \mathbf{X}_i, \mathbf{c}_i] = E[Y_{it}(1)|\mathbf{X}_i, \mathbf{c}_i],$$

where $\mathbf{W}_i = (W_{i1}, \ldots, W_{iT})$ is the time series of all treatments.

• $\mathbf{X}_{it}$ cannot include factors that react to shocks to previous outcomes (so no lagged outcomes).

- Suppose

$$E[Y_{it}(0)|\mathbf{X}_i, \mathbf{c}_i] = \alpha_{t0} + \mathbf{X}_{it}\boldsymbol{\gamma}_0 + c_{i0}, \tag{48}$$

and the expected gain from treatment is a function of $\mathbf{X}_{it}$:

$$E[Y_{it}(1)|\mathbf{X}_i, \mathbf{c}_i] = E[Y_{it}(0)|\mathbf{X}_i, \mathbf{c}_i] + \eta_t + \mathbf{X}_{it}\boldsymbol{\gamma}_1. \tag{49}$$

• Since $Y_{it} = (1 - W_{it})Y_{it}(0) + W_{it}Y_{it}(1)$,

$$E(Y_{it}|\mathbf{W}_i, \mathbf{X}_i, \mathbf{c}_i) = \alpha_{t0} + \tau_t W_{it} + \mathbf{X}_{it}\boldsymbol{\gamma}_0 + W_{it} \cdot (\mathbf{X}_{it} - \boldsymbol{\xi}_t)\boldsymbol{\delta} + c_{i0} \qquad (50)$$

where $\boldsymbol{\xi}_t = E(\mathbf{X}_{it})$.

• Estimation is by fixed effects (to remove c_{i0}) including also time dummies, and interacting the treatment with $\mathbf{X}_{it} - \bar{\mathbf{X}}_t$.

• Setting $\boldsymbol{\delta} = \mathbf{0}$ and $\tau_t = \tau$ gives a standard fixed effects estimator of $\tau = \tau_{ate}$.

• A second approach uses unconfoundedness conditional on past observables.

• Let $\mathbf{X}_i^t$ denote a set of covariates observed prior to time t, including past outcomes $\{Y_{i,t-1}, \ldots, Y_{i1}\}$ and and past treatments $\{W_{i,t-1}, \ldots, W_{i1}\}$. Assume unconfoundedness conditional on $\mathbf{X}_i^t$:

$$D[W_{it}|Y_{it}(0), Y_{it}(1), \mathbf{X}_i^t] = D(W_{it}|\mathbf{X}_i^t) \tag{51}$$

• Under this assumption, we can, at each time period t, apply regression adjustment, propensity score weighting, or matching to obtain $\hat{\tau}_{ate,t}$.

• Might feel more comfortable using larger t that there is more to put in $\mathbf{X}_i^t$.

- There is a tradeoff between unconfoundedness and overlap. More in $\mathbf{X}_i^t$ is better for unconfoundedness, worse for overlap. The effective population may have to be reduced.
- For example, if we observe $W_{i1} = W_{i2} = \ldots = W_{i,t-1} = 1$, $P(W_{it} = 1|\mathbf{X}_i^t)$ may be very close to one.

- Using either approach, can get an overall average effect:

$$\hat{\tau}_{ate} = T^{-1} \sum_{t=1}^{T} \hat{\tau}_{ate,t} \tag{52}$$

- Getting an appropriate standard error requires accounting for the correlations in the estimators across time. Can use a "panel bootstrap," where cross-sectional units are resampled.

- Neither of the previous two approaches is more general than the other: one is based on unobserved heterogeneity (a “firm fixed effect”), the other on past observables.
- Both are preferred to the common practice of excluding past responses and treatments in $\mathbf{X}_i^t$ and also ignoring unobserved firm heterogeneity.

7. Application to Effects of Using a Big 4 Auditing Firm

- Subset of the data from Lawrence, Minutti-Meza, and Zhang (2011, The Accounting Review).
- LMZ_cross.dta is a small cross-sectional data set for 2006. LMZ_panel is a panel from 2001-2006, with values from 2000 as controls.
- Use same SICs as LMZ.
- Treatment variable is *big*4, an indicator for using a Big 4 auditing firm.
- Main outcome is *ada*, absolute discretionary accurals.

- Covariates include total assets, total market value, return on assets, measures of debts and liabilities. Also, industry and year dummies.
- LMZ pool data and treat a firm/year as the unit of observation. Inference probably wrong. They do not use firm fixed effects or fully dynamic matching.

```
. use LMZ_cross

. tab big4

       big4 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,015       26.77       26.77
          1 |      2,777       73.23      100.00
------------+-----------------------------------
      Total |      3,792      100.00

. sum ada

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         ada |      3792    .1005575    .1452539          0   2.229223
```

```
. reg ada big4, robust

Linear regression                                      Number of obs =    3792
                                                       F(  1,  3790) =   65.43
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0274
                                                       Root MSE      =  .14327

------------------------------------------------------------------------------
             |               Robust
         ada |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        big4 |   -.054318   .0067149    -8.09   0.000    -.0674831   -.0411528
       _cons |   .1403362   .0063556    22.08   0.000     .1278755    .1527969
------------------------------------------------------------------------------

. * Add lags of ada, big4:

. reg ada big4 ada_1 big4_1, robust

Linear regression                                      Number of obs =    3535
                                                       F(  3,  3531) =   29.79
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1403
                                                       Root MSE      =  .13359

------------------------------------------------------------------------------
             |               Robust
         ada |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        big4 |  -.0281794   .0112102    -2.51   0.012    -.0501584   -.0062004
       ada_1 |   .3914005   .0554027     7.06   0.000      .282776    .5000251
      big4_1 |  -.0113145   .0117728    -0.96   0.337    -.0343966    .0117676
       _cons |   .0931548   .0066316    14.05   0.000     .0801527    .1061569
------------------------------------------------------------------------------
```

```
. * Add LMZ controls:

. reg ada big4 ada_1 big4_1 log_assets aturn curr lev roa log_mkt lag_roa
      lag_lev lag_curr ind2-ind99, robust

Linear regression                                       Number of obs =    3535
                                                        F( 61,  3472) =       .
                                                        Prob > F      =       .
                                                        R-squared     =  0.3980
                                                        Root MSE      =  .11273

------------------------------------------------------------------------------
             |               Robust
         ada |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        big4 |  -.0037774   .0103369    -0.37   0.715    -.0240443    .0164896
      ada_1 |    .093561   .0436434     2.14   0.032     .0079916    .1791305
     big4_1 |  -.0050892   .0105566    -0.48   0.630    -.0257871    .0156086
  log_assets |  -.0150057   .0034664    -4.33   0.000    -.0218021   -.0082093
       aturn |   .0105236   .0035795     2.94   0.003     .0035054    .0175419
        curr |   .0005621   .0011111     0.51   0.613    -.0016163    .0027406
         lev |   .0619562   .0154145     4.02   0.000     .0317339    .0921786
         roa |  -.1779534   .0395166    -4.50   0.000    -.2554315   -.1004753
     log_mkt |   .0114269    .003407     3.35   0.001      .004747    .0181069
     lag_roa |  -.0540013   .0355773    -1.52   0.129    -.1237559    .0157534
     lag_lev |  -.0614061   .0179375    -3.42   0.001    -.0965753   -.0262369
    lag_curr |  -.0012331   .0009607    -1.28   0.199    -.0031167    .0006506
.
.
.
------------------------------------------------------------------------------

. * Effect is much smaller and statistically significant.
```

```
. * Now use separate linear regressions:

. * 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 ada ada_1 big4_1 log_assets aturn curr lev roa log_mkt l
                ag_roa lag_lev lag_curr ind2-ind99 if big4
  4.         predict adah1
  5.         reg ada ada_1 big4_1 log_assets aturn curr lev roa log_mkt
                lag_roa lag_lev lag_curr ind2-ind99 if ~big4
  6.         predict adah0
  7.
.         gen ate_i = adah1 - adah0
  8.         sum ate_i
  9.         scalar ate = r(mean)
 10.         sum ate_i if big4
 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 adah1 adah0 ate_i
 17.
. end
```

```
.
. 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      =      3792
                                                  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 |   -.002738   .0114702    -0.24   0.811    -.0252192    .0197431
       _bs_2 |   .0003525   .0146904     0.02   0.981    -.0284401     .029145
------------------------------------------------------------------------------

.
. program drop ateboot

.
end of do-file

. * To get regression adjustment, can use the double-robust approach
. * without any pvars:

. dr ada big4, ovars(ada_1 big4_1 log_assets aturn curr lev roa
```

```
    log_mkt lag_roa lag_lev lag_curr ind2-ind99)

Doubly Robust Estimate of the effect of big4 on ada
Using sandwich estimator of SE

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        big4 |   -.002738   .0052725    -0.52   0.604    -.0130718    .0075958
------------------------------------------------------------------------------
```

```
. * Now use PS matching. Use nnmatch to obtain conservative standard errors.

. * qui logit big4 ada_1 big4_1 log_assets aturn curr lev roa log_mkt
     lag_roa lag_lev lag_curr ind2-ind99

. predict phat
(option pr assumed; Pr(big4))
(307 missing values generated)

. sum phat

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        phat |      3485    .7268293    .3969844   .0005813   .9997486

. * Overlap looks like it may be a problem.

. count if phat < .05 | phat > .95
 2828

. sum phat if big4

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        phat |      2533    .9437114     .124395   .0036205   .9997486

. sum phat if ~big4

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        phat |       952    .1497679    .2784276   .0005813    .992681
```

```
. nnmatch ada big4 phat, tc(ate)
307 observations dropped due to treatment variable missing

Matching estimator:  Average Treatment Effect ate

                                        Number of obs          =      3485
                                        Number of matches  (m) =         1

------------------------------------------------------------------------------
         ada |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATE |   .0109919    .029483     0.37   0.709    -.0467938    .0687776
------------------------------------------------------------------------------
Matching variables:  phat

. nnmatch ada big4 phat, tc(att)
307 observations dropped due to treatment variable missing

Matching estimator:  Average Treatment Effect for the Treated

                                        Number of obs          =      3485
                                        Number of matches  (m) =         1

------------------------------------------------------------------------------
         ada |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        SATT |   .0237566   .0329039     0.72   0.470    -.0407339     .088247
------------------------------------------------------------------------------
Matching variables:  phat
```

```
. * Caliper matching does not help change sign of ate:

. psmatch2  big4 ada_1 big4_1 log_assets aturn curr lev roa log_mkt lag_roa
           lag_lev lag_curr ind2-ind99, out(ada) n(1) logit ate caliper(.03)

------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+------------------------------------------------------------
             ada  Unmatched | .084883983   .139056525  -.054172542   .005429904    -9.98
                         ATT | .084883983   .061127431   .023756552   .067533255     0.35
                         ATU | .140070144   .116690706  -.023379438            .         .
                         ATE |                            .011018563            .         .
-----------------------------+------------------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.

 psmatch2: |   psmatch2: Common
 Treatment |        support
assignment | Off suppo  On suppor |     Total
-----------+----------------------+----------
 Untreated |        14        938 |       952
   Treated |         0      2,533 |     2,533
-----------+----------------------+----------
     Total |        14      3,471 |     3,485
```

```
. * Inverse PS weighting:

. gen kate_i = (big4 - phat)*ada/(phat*(1 - phat))
(307 missing values generated)

. reg kate_i

      Source |       SS       df       MS              Number of obs =    3485
-------------+------------------------------           F(  0,  3484) =    0.00
       Model |           0     0           .           Prob > F      =       .
    Residual |  6634.63194  3484  1.90431456           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared =  0.0000
       Total |  6634.63194  3484  1.90431456           Root MSE      =    1.38

------------------------------------------------------------------------------
      kate_i |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .0411547   .0233759     1.76   0.078    -.0046771    .0869865
------------------------------------------------------------------------------

. * Probably too dangerous because of very small and large probabilities.
. * Bootstrap gives many samples with perfect fits of the PS.
```

```
. * Now double robustness, first using linear model, then exponential:

. dr ada big4, ovars(ada_1 big4_1 log_assets aturn curr lev roa log_mkt
      lag_roa lag_lev lag_curr ind2-ind99)
      pvars(ada_1 big4_1 log_assets aturn curr lev roa log_mkt lag_roa
      lag_lev lag_curr ind2-ind99)


Doubly Robust Estimate of the effect of big4 on ada
Using sandwich estimator of SE

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        big4 |   .0011662   .0126976     0.09   0.927    -.0237207    .0260531
------------------------------------------------------------------------------

. dr ada big4, ovars(ada_1 big4_1 log_assets aturn curr lev roa log_mkt
     lag_roa lag_lev lag_curr ind2-ind99)
      pvars(ada_1 big4_1 log_assets aturn curr lev roa log_mkt lag_roa
      lag_lev lag_curr ind2-ind99) fam(poiss) link(log)


Doubly Robust Estimate of the effect of big4 on ada
Using sandwich estimator of SE

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        big4 |    .011741   .0123333     0.95   0.341    -.0124319    .0359138
------------------------------------------------------------------------------
```

```
. * Now use firms with big4 == 0 in 2005:

. use LMZ_small

. do linreg_LMZ_small

. * 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 ada ada_1 log_assets aturn curr lev roa log_mkt
               lag_roa lag_lev lag_curr ind2-ind99 if big4
  4.         predict adah1
  5.         reg ada ada_1 log_assets aturn curr lev roa log_mkt
                lag_roa lag_lev lag_curr ind2-ind99 if ~big4
  6.         predict adah0
  7.
.         gen ate_i = adah1 - adah0
  8.         sum ate_i
  9.         scalar ate = r(mean)
 10.         sum ate_i if big4
 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 adah1 adah0 ate_i
```

```
 17.
. end

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

Bootstrap results                                     Number of obs      =       862
                                                      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 |  -.0607535   .1903262    -0.32   0.750    -.4337859     .312279
       _bs_2 |   .0005891   .0193423     0.03   0.976    -.0373212    .0384994
------------------------------------------------------------------------------
.
. program drop ateboot

. dr ada big4, ovars(ada_1 log_assets aturn curr lev roa log_mkt lag_roa
      lag_lev lag_curr ind2-ind99) pvars(ada_1 big4_1 log_assets
      aturn curr lev roa log_mkt lag_roa lag_lev lag_curr ind2-ind99)

Doubly Robust Estimate of the effect of big4 on ada
Using sandwich estimator of SE

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        big4 |  -.0242166   .0249231    -0.97   0.331     -.073065    .0246318
------------------------------------------------------------------------------
```

```
. * Now panel data:

. use LMZ_panel

. tab fyear

data year - |
     fiscal |      Freq.     Percent        Cum.
------------+-----------------------------------
       2000 |      4,639       15.66       15.66
       2001 |      4,483       15.13       30.79
       2002 |      4,478       15.12       45.91
       2003 |      4,259       14.38       60.28
       2004 |      4,100       13.84       74.12
       2005 |      3,874       13.08       87.20
       2006 |      3,792       12.80      100.00
------------+-----------------------------------
      Total |     29,625      100.00

. tab big4

       big4 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      5,568       18.79       18.79
          1 |     24,057       81.21      100.00
------------+-----------------------------------
      Total |     29,625      100.00
```

```
. * Fixed effects estimation:

. xtreg ada big4 log_assets aturn curr lev roa log_mkt lag_roa lag_lev
      lag_curr yr01-yr06, fe cluster(gvkey_n)

Fixed-effects (within) regression               Number of obs      =     29625
Group variable: gvkey_n                         Number of groups   =      6159

R-sq:  within  = 0.1307                         Obs per group: min =         1
       between = 0.4850                                        avg =       4.8
       overall = 0.3055                                        max =         7

                                                F(16,6158)         =     14.27
corr(u_i, Xb)  = 0.0855                         Prob > F           =    0.0000

                              (Std. Err. adjusted for 6159 clusters in gvkey_n)
------------------------------------------------------------------------------
             |               Robust
         ada |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         big4 |    .003193   .0047663     0.67   0.503    -.0061506    .0125365
   log_assets |  -.0132004   .0048506    -2.72   0.007    -.0227093   -.0036916
        aturn |   .0025449   .0016207     1.57   0.116    -.0006322    .0057221
         curr |   .0008496   .0005508     1.54   0.123    -.0002302    .0019293
          lev |  -.0528359   .0232103    -2.28   0.023    -.0983361   -.0073356
          roa |  -.2417109   .0662763    -3.65   0.000    -.3716355   -.1117863
      log_mkt |   .0138002   .0032338     4.27   0.000     .0074608    .0201395
      lag_roa |  -.0246634   .0198011    -1.25   0.213    -.0634804    .0141536
      lag_lev |   .0697862   .0267426     2.61   0.009     .0173615     .122211
    lag_curr |  -.0001963   .0005667    -0.35   0.729    -.0013073    .0009146
        yr01 |  -.0084786   .0034354    -2.47   0.014    -.0152133    -.001744
        yr02 |  -.0163516   .0028971    -5.64   0.000     -.022031   -.0106722
        yr03 |  -.0211816   .0028993    -7.31   0.000    -.0268653    -.015498
        yr04 |  -.0106873   .0032366    -3.30   0.001    -.0170322   -.0043425
        yr05 |  -.0246119   .0033542    -7.34   0.000    -.0311874   -.0180365
```

```
        yr06 |  -.0157228   .0036467    -4.31   0.000    -.0228716   -.0085739
       _cons |   .0927147   .0408614     2.27   0.023     .0126122    .1728173
-------------+----------------------------------------------------------------
     sigma_u |  .13570681
     sigma_e |  .12902638
         rho |   .5252185   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. * Pooled matching, as in LMZ. Standard errors incorrect because time series
. * correlation ignored. Also, estimated propensity score ignored:

. psmatch2 big4 log_assets aturn curr lev roa log_mkt lag_roa lag_lev
      lag_curr yr01-yr06 ind2-ind99, out(ada) n(1) ate logit

------------------------------------------------------------------------------
    Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+------------------------------------------------
         ada  Unmatched | .099083692    .15591834  -.056834648   .002551798   -22.27
                     ATT | .099083692   .107141449  -.008057757    .01024832    -0.79
                     ATU |  .15591834    .14861206   -.00730628            .        .
                     ATE |                          -.007916044            .        .
-----------------------------+------------------------------------------------
```

```
. * Sequential matching:

. psmatch2 big4 ada_1 big4_1 log_assets aturn curr lev roa log_mkt lag_roa
        lag_lev lag_curr ind2-ind99 if yr01, out(ada) n(1) ate
------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+------------------------------------------------------------
             ada  Unmatched | .111706441   .162546089  -.050839648   .006324887    -8.04
                        ATT | .111706441   .113419543  -.001713102   .066729983    -0.03
                        ATU | .162546089   .171438524   .008892435            .        .
                        ATE |                          -.000166068            .        .
-----------------------------+------------------------------------------------------------

. psmatch2 big4 ada_1 big4_1 log_assets aturn curr lev roa log_mkt lag_roa
        lag_lev lag_curr ind2-ind99 if yr02, out(ada) n(1) ate

------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+------------------------------------------------------------
             ada  Unmatched | .098112904   .153769922  -.055657018   .006005934    -9.27
                        ATT | .098112904   .103298256  -.005185352   .044479146    -0.12
                        ATU | .153769922   .149525088  -.004244833            .        .
                        ATE |                          -.005030568            .        .
-----------------------------+------------------------------------------------------------

. psmatch2 big4 ada_1 big4_1 log_assets aturn curr lev roa log_mkt lag_roa
        lag_lev lag_curr ind2-ind99 if yr03, out(ada) n(1) ate

------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+------------------------------------------------------------
             ada  Unmatched | .089687347   .150904827   -.06121748   .006238025    -9.81
                        ATT | .089687347   .112017109  -.022329762   .032756893    -0.68
                        ATU | .150904827   .166038042   .015133216            .        .
                        ATE |                          -.015891694            .        .
```

```
-----------------------------+-----------------------------------------------------------

. psmatch2 big4 ada_1 big4_1 log_assets aturn curr lev roa log_mkt lag_roa
        lag_lev lag_curr ind2-ind99 if yr04, out(ada) n(1) ate

------------------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+------------------------------------------------------------
             ada  Unmatched | .091364716   .172477806  -.081113091   .009079242    -8.93
                        ATT | .091364716   .099152418  -.007787702   .022627015    -0.34
                        ATU | .172477806   .101160802  -.071317004            .        .
                        ATE |                          -.021035827            .        .
-----------------------------+------------------------------------------------------------

. * Regression adjustment:

. dr ada big4  if yr04, ovars(ada_1 big4_1 log_assets aturn curr lev roa log_mkt
     lag_roa lag_lev lag_curr ind2-ind99)


Doubly Robust Estimate of the effect of big4 on ada
Using sandwich estimator of SE

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        big4 |  -.0100175   .0086784    -1.15   0.248    -.0270269    .0069919
------------------------------------------------------------------------------

. * PS weighting:

. dr ada big4  if yr04, pvars(ada_1 big4_1 log_assets aturn curr lev
     roa log_mkt lag_roa lag_lev lag_curr ind2-ind99)

Doubly Robust Estimate of the effect of big4 on ada
```

```
Using sandwich estimator of SE

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        big4 |   .0145192   .0566561     0.26   0.798    -.0965248    .1255631
------------------------------------------------------------------------------

. * Doubly-robust estimator:

. dr ada big4  if yr04, ovars(ada_1 big4_1 log_assets aturn curr lev roa
    log_mkt lag_roa lag_lev lag_curr ind2-ind99) pvars(ada_1 big4_1
    log_assets aturn curr lev roa log_mkt lag_roa lag_lev lag_curr ind2-ind99)


Doubly Robust Estimate of the effect of big4 on ada
Using sandwich estimator of SE

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        big4 |   .0077436   .0327389     0.24   0.813    -.0564234    .0719105
------------------------------------------------------------------------------

. psmatch2 big4 ada_1 big4_1 log_assets aturn curr lev roa log_mkt lag_roa
        lag_lev lag_curr ind2-ind99 if yr05, out(ada) n(1) ate

------------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+------------------------------------------------------
             ada  Unmatched | .078365145   .127552232  -.049187087   .004727829   -10.40
                        ATT | .078365145   .069721284   .008643861   .028699414     0.30
                        ATU | .127552232   .110698894  -.016853337            .        .
                        ATE |                            .002196133            .        .
-----------------------------+------------------------------------------------------
```

```
. psmatch2 big4 ada_1 big4_1 log_assets aturn curr lev roa log_mkt lag_roa
        lag_lev lag_curr ind2-ind99 if yr06, out(ada) n(1) ate

------------------------------------------------------------------------------
        Variable     Sample |    Treated     Controls   Difference         S.E.   T-stat
-----------------------------+------------------------------------------------------------
             ada  Unmatched | .084883983   .139056525  -.054172542   .005429904    -9.98
                        ATT | .084883983   .068711041   .016172942   .051404095     0.31
                        ATU | .139056525   .114961423  -.024095102            .         .
                        ATE |                           .005172891            .         .
-----------------------------+------------------------------------------------------------
Note: S.E. does not take into account that the propensity score is estimated.
```

8. Assessing Unconfoundedness

- Unconfoundedness is not directly testable. Any assessment is indirect.
- Several possibilities. Given several pre-treatment outcomes, can construct a treatment effect on a pseudo outcome and establish that it is not statistically different from zero.
- Suppose controls consist of time-constant characteristics, $\mathbf{Z}_i$, and three pre-assignment outcomes on the response, $Y_{i,-1}$, $Y_{i,-2}$, and $Y_{i,-3}$. Let the counterfactual outcomes be at $t = 0$, $Y_{i0}(0)$ and $Y_{i0}(1)$. Suppose we are willing to assume unconfoundedness given two lags:

$$Y_{i0}(0), Y_{i0}(1) \perp W_i \mid Y_{i,-1}, Y_{i,-2}, \mathbf{Z}_i \tag{53}$$

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

$$Y_{i,-1} \perp W_i, \mid Y_{i,-2}, Y_{i,-3}, \mathbf{Z}_i, \tag{54}$$

and this 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.

• Can regress $Y_{i,-1}$ on $Y_{i,-2}$, $Y_{i,-3}$, $\mathbf{Z}_i$ for $W_i = 0$ and $W_i = 1$ and peform Chow test.

• Or, estimate logit or probit for $P(W_i = 1|Y_{i,-1}, Y_{i,-2}, Y_{i,-3}, \mathbf{Z}_i)$ and test $Y_{i,-1}$ for significance.

• 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 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.

9. 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

$$normdiff_j = \frac{(\bar{X}_{1j} - \bar{X}_{0j})}{\sqrt{S_{1j}^2 + S_{0j}^2}} \tag{55}$$

- Imbens and Rubin discuss a rule-of-thumb: Normalized differences above about $.25$ are cause for concern.

- $normdiff_j$ is not the t statistic for comparing the means of the distribution. The t statistic depends explicitly on the sample sizes. 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.

- Easier to look directly at the propensity scores, or the log-odds of the propensity score. In other words, compute the normalized difference of a single "balancing" 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 for τ_{att}.]

• 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. Resulting population might not be what we want. Does not always solve the overlap problem.

• Or, 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.

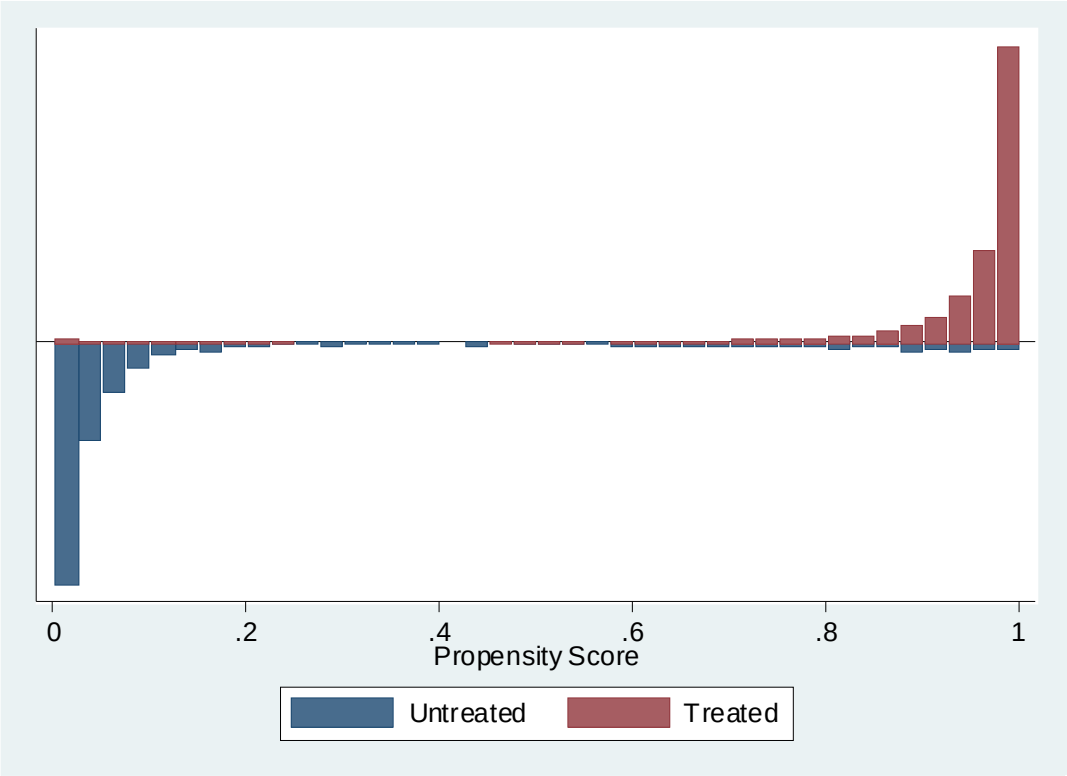
• The new smaller – in some cases, much smaller – sample is better balanced. Can apply all the usual methods for τ_{ate}.

- 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. If a firm has used a Big Four accounting firm over all twelve years of a sample, should it be included in a study of the effects of using one of the Big Four? These firms would fall out of a fixed effects analysis.

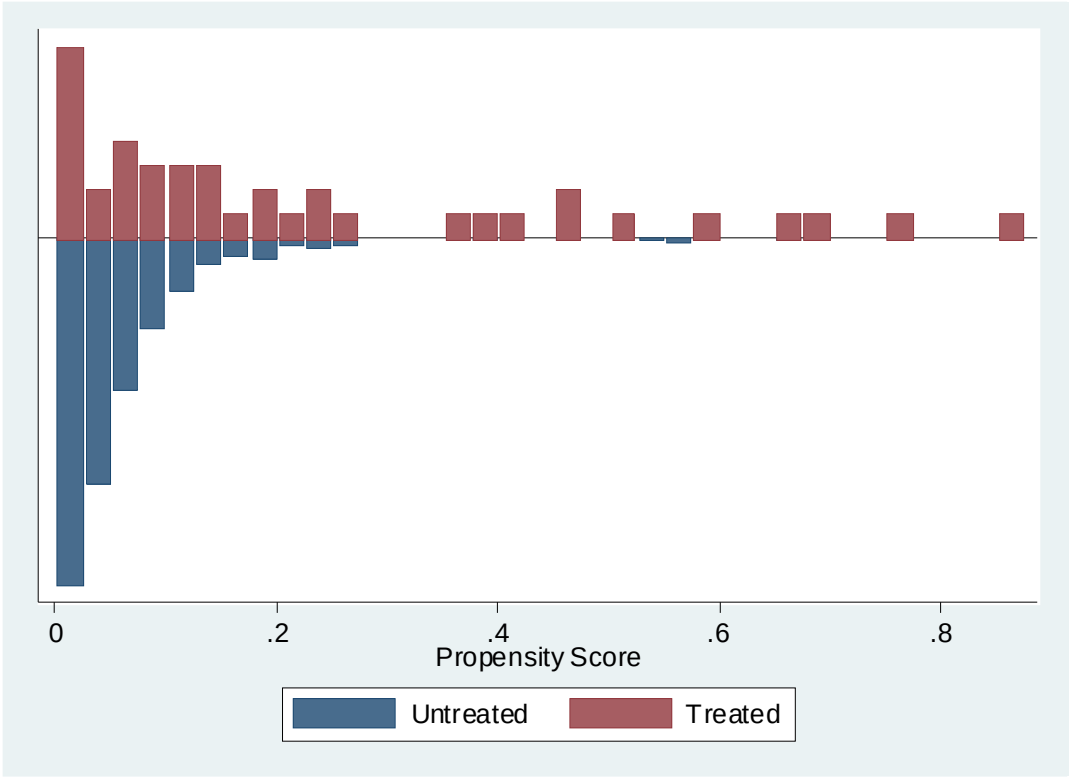
```
. * With full cross section in 2006:

. pstest ada_1 big4_1 log_assets aturn curr lev roa log_mkt lag_roa lag_lev lag_curr

-------------------------------------------------------------------------------
                         |       Mean                %%reduct |     t-test
    Variable     Sample | Treated Control    %%bias  |bias| |    t    p>|t|
-------------------------+----------------------------------+----------------
       ada_1  Unmatched | .07799   .12128    -31.1          |  -9.23  0.000
                Matched | .07799   .05381     17.4    44.1 |   9.21  0.000
                        |                                  |
      big4_1  Unmatched | .98381   .13866    324.7          | 105.72  0.000
                Matched | .98381    .9846     -0.3    99.9 |  -0.23  0.822
                        |                                  |
  log_assets  Unmatched | 6.8268    4.116    160.3          |  39.51  0.000
                Matched | 6.8268   6.4343     23.2    85.5 |   8.06  0.000
                        |                                  |
       aturn  Unmatched |  1.105   1.2815    -17.9          |  -5.04  0.000
                Matched |  1.105   .99216     11.5    36.0 |   5.26  0.000
                        |                                  |
        curr  Unmatched | 2.6542   3.3653    -21.2          |  -5.89  0.000
                Matched | 2.6542   2.5049      4.4    79.0 |   2.16  0.031
                        |                                  |
         lev  Unmatched | .21688   .19255      5.4          |   1.76  0.079
                Matched | .21688   .26278    -10.1   -88.6 |  -6.76  0.000
                        |                                  |
         roa  Unmatched | .00105  -.11701     31.0          |   9.71  0.000
                Matched | .00105   .01105     -2.6    91.5 |  -1.55  0.122
                        |                                  |
     log_mkt  Unmatched | 6.9848   4.4475    155.8          |  38.83  0.000
                Matched | 6.9848   6.7165     16.5    89.4 |   5.67  0.000
                        |                                  |
     lag_roa  Unmatched | .00577  -.09574     33.5          |  10.37  0.000
                Matched | .00577   .03216     -8.7    74.0 |  -5.38  0.000
                        |                                  |
```

```
     lag_lev  Unmatched | .20904   .17305     14.6          |   4.12  0.000
                Matched | .20904   .23189     -9.2    36.5  |  -3.80  0.000
                        |                                   |
    lag_curr  Unmatched | 2.7321    3.514    -17.4          |  -5.43  0.000
                Matched | 2.7321   2.7833     -1.1    93.5  |  -0.63  0.530
                        |                                   |
-------------------------------------------------------------------------------
```

```
. qui psmatch2  big4 ada_1 big4_1 log_assets aturn curr lev roa log_mkt lag_roa
      lag_lev lag_curr ind2-ind99, out(ada) n(1) logit ate
Note: S.E. does not take into account that the propensity score is estimated.

. psgraph, bin(40)
```



```
. * Firms with big4 == 0 in 2005:

. pstest phat

-------------------------------------------------------------------------------
                         |       Mean                %%reduct |     t-test
    Variable     Sample | Treated Control    %%bias  |bias| |    t    p>|t|
-------------------------+----------------------------------+-----------------
        phat  Unmatched | .22329   .05769     97.5          |  11.31  0.000
                Matched | .22329   .21494      4.9    95.0 |   0.17  0.864
                         |                                  |
-------------------------------------------------------------------------------

. psgraph, bin(40)
```


0
.2
.4
.6
.8
Propensity Score
Untreated
Treated