Lecture 3 –OLS and matching

Economics 8379
George Washington University

Instructor: Prof. Ben Williams

Review

Identification under random assignment:

- If D_i is randomly assigned then

$$E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0) = E(Y_{1i} - Y_{0i})$$

- this is because, more generally randomization implies that $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i$
- note that randomization also implies that $ATE = TT = TUT$ but *not* that $Y_{1i} - Y_{0i} = ATE$

Review

Estimation under random assignment:

- analogy principle
- other methods to improve efficiency
- regression adjustment can introduce bias in finite samples

Preview

Today's lecture:

- identification based on *conditional independence*:

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i \mid X_i$$

- what does OLS estimate under this assumption and other useful results about OLS
- matching estimators
- Campolieti, Gunderson, and Smith (2014)
- next week:
  - what to include in X_i
  - sensitivity analysis

Conditional independence

- The conditional independence assumption:

(M-1) $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i \mid X_i$

- *Selection on observables*
- We will explore what variables should and shouldn't be included in X_i later.
- weaker sufficient condition:

(M-1)' $E(Y_1 \mid D, X) = E(Y_1 \mid X)$ and $E(Y_0 \mid D, X) = E(Y_0 \mid X)$ (conditional mean independence)

Common support

- Identification also requires an auxiliary assumption.
- The common support assumption:

(M-2) $0 < Pr(D = 1 \mid X = x) < 1$ for each $x \in support(X)$

Common support

- Identification also requires an auxiliary assumption.
- The common support assumption:

(M-2) $0 < Pr(D = 1 \mid X = x) < 1$ for each $x \in support(X)$

- Notation:
  - $P(x) := Pr(D = 1 \mid X = x)$ is the *propensity score*
  - The support of a random variable is the set of values where its density (or pmf) is positive.
  - Let $S_d = Supp(X \mid D = d)$ and let $S_{10} = S_1 \cap S_0$.
- Under (M-2), $S_1 = S_0 = S_{10}$.

Identification of ATE

- First, (M-1)' implies that
  $E(Y \mid D = d, X) = E(Y_d \mid D = d, X) = E(Y_d \mid X)$.
- Therefore,

$$
\begin{aligned}
& E\left(E(Y \mid D = 1, X)\right) - E\left(E(Y \mid D = 0, X)\right) \\
&= E\left(E(Y_1 - Y_0 \mid X)\right) \\
&= E(Y_1 - Y_0)
\end{aligned}
$$

Identification of ATE

- Where does (M-2) come into play?

Identification of ATE

- Where does (M-2) come into play?
  - $E(Y_d \mid D = d, X = x)$ is only defined for $x \in S_d$
  - Therefore, $E(Y_d) = E\left(E(Y \mid D = d, X)\right)$ only holds if $S_d = Supp(X)$.
  - We need both $S_1 = Supp(X)$ and $S_0 = Supp(X)$

Identification of ATE

- Where does (M-2) come into play?
  - $E(Y_d \mid D = d, X = x)$ is only defined for $x \in S_d$
  - Therefore, $E(Y_d) = E\left(E(Y \mid D = d, X)\right)$ only holds if $S_d = Supp(X)$.
  - We need both $S_1 = Supp(X)$ and $S_0 = Supp(X)$– equivalent to (M-2)

Identification of ATT

- The treatment on the treated is identified under weaker assumptions.
- The derivation:

Identification of ATT

- The treatment on the treated is identified under weaker assumptions.
- The derivation:

$$
\begin{aligned}
& E(Y \mid D=1) - E\left(E(Y \mid D=0, X) \mid D=1\right) \\
&= E(Y_1 \mid D=1) - E\left(E(Y_0 \mid D=1, X) \mid D=1\right) \\
&= E(Y_1 - Y_0 \mid D=1)
\end{aligned}
$$

Identification of ATT

- The treatment on the treated is identified under weaker assumptions.
- The derivation:

$$
\begin{aligned}
& E(Y \mid D=1) - E\left(E(Y \mid D=0, X) \mid D=1\right) \\
&= E(Y_1 \mid D=1) - E\left(E(Y_0 \mid D=1, X) \mid D=1\right) \\
&= E(Y_1 - Y_0 \mid D=1)
\end{aligned}
$$

- So S_0 can be “bigger” but not “smaller”.

Common support

- pictures...

Identification without common support

- Let $\mathcal{C} = S_1 \cap S_0$ be the *common support*.
- If support conditions don't hold we can always identify

$$E(Y_1 - Y_0 \mid X \in \mathcal{C})$$

  under (M-1)' alone.
- Bounds on the treatment effects outside of $\mathcal{C}$ can be used to get bounds on the ATE or ATT.

Consider the linear regression model:

$$y_i = \beta' X_i + e_i$$

The OLS estimator of β minimizes the sum of squared residuals,

$$\sum_{i=1}^{n} (y_i - \beta' X_i)^2$$

The solution is

$$\hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i' \right)^{-1} \sum_{i=1}^{n} X_i y_i$$

If the data constitute a random sample and $E(e_i X_i) = 0$ then $\hat{\beta} \rightarrow_p \beta$ and has an asymptotic normal distribution.

If the data constitute a random sample and $E(e_i X_i) = 0$ then $\hat{\beta} \to_p \beta$ and has an asymptotic normal distribution.

- If we define $\beta = E(X_i X_i')^{-1} E(X_i y_i)$ and $e_i = y_i - \beta' X_i$ then $E(e_i X_i) = 0$ is automatically satisfied.
- In this case, $\beta' X_i$ provides minimum MSE linear approximation to the CEF, $E(Y_i \mid X_i)$.

If the data constitute a random sample and $E(e_i X_i) = 0$ then $\hat{\beta} \to_p \beta$ and has an asymptotic normal distribution.

- If we define $\beta = E(X_i X_i')^{-1} E(X_i y_i)$ and $e_i = y_i - \beta' X_i$ then $E(e_i X_i) = 0$ is automatically satisfied.
- In this case, $\beta' X_i$ provides minimum MSE linear approximation to the CEF, $E(Y_i \mid X_i)$.

If $E(e_i \mid X_i) = 0$ then $\hat{\beta}$ is unbiased.

If the data constitute a random sample and $E(e_i X_i) = 0$ then $\hat{\beta} \rightarrow_p \beta$ and has an asymptotic normal distribution.

If the data constitute a random sample and $E(e_i X_i) = 0$ then $\hat{\beta} \to_p \beta$ and has an asymptotic normal distribution.

- If we define $\beta = E(X_i X_i')^{-1} E(X_i y_i)$ and $e_i = y_i - \beta' X_i$ then $E(e_i X_i) = 0$ is automatically satisfied.
- In this case, $\beta' X_i$ provides minimum MSE linear approximation to the CEF, $E(Y_i \mid X_i)$.

If the data constitute a random sample and $E(e_i X_i) = 0$ then $\hat{\beta} \to_p \beta$ and has an asymptotic normal distribution.

- If we define $\beta = E(X_i X_i')^{-1} E(X_i y_i)$ and $e_i = y_i - \beta' X_i$ then $E(e_i X_i) = 0$ is automatically satisfied.
- In this case, $\beta' X_i$ provides minimum MSE linear approximation to the CEF, $E(Y_i \mid X_i)$.

If $E(e_i \mid X_i) = 0$ then $\hat{\beta}$ is unbiased.

Under the interpretation of the regression equation as the minimum MSE linear approximation to the CEF,

- OLS will typically be biased relative to β in finite samples.
- Heteroskedasticity is natural:
  - $\Longrightarrow$ always use robust standard errors
  - but weighting changes the estimand

Frisch-Waugh-Lovell (FWL) theorem

- For each k,

$$\beta_k = \frac{Cov(Y_i, \varepsilon_i^{(k)})}{Var(\varepsilon_i^{(k)})}$$

where $\varepsilon_i^{(k)}$ is the residual from regression X_{ik} on the rest of X_i.

Frisch-Waugh-Lovell (FWL) theorem

- For each k,

$$\beta_k = \frac{Cov(Y_i, \varepsilon_i^{(k)})}{Var(\varepsilon_i^{(k)})}$$

  where $\varepsilon_i^{(k)}$ is the residual from regression X_{ik} on the rest of X_i.
- What does this imply about a regression where $X_{i2}, \ldots, X_{iK}$ are, for example, regional dummy variables?
- Can replace Y_i with $\tilde{Y}_i^k$ (residual from regression Y_i on $X_{i2}, \ldots, X_{iK}$).
- This sample analogue also holds.

Omitted variable bias

- Suppose W_i denote some other variables excluded from the vector X_i.
- Let $\hat{beta}_1^s$ denote the coefficients on X_i in a regression that excludes.
- Let $\hat{\beta}_1^\ell$ and $\hat{\beta}_2^\ell$ denote the coefficients on X_i and W_i, respectively, when W_i is included.
- Then

$$\hat{\beta}_1^s = \hat{\beta}_1^\ell + \left((X'X)^{-1} X'W \right) \hat{\beta}_2^\ell$$

Omitted variable bias

- Suppose W_i denote some other variables excluded from the vector X_i.
- Let $\hat{beta}_1^s$ denote the coefficients on X_i in a regression that excludes.
- Let $\hat{\beta}_1^\ell$ and $\hat{\beta}_2^\ell$ denote the coefficients on X_i and W_i, respectively, when W_i is included.
- Then

$$\hat{\beta}_1^s = \hat{\beta}_1^\ell + \left((X'X)^{-1} X'W \right) \hat{\beta}_2^\ell$$

- "short equals long plus the effect of the omitted times the regression of omitted on included"

Fully saturated regression model

- boop

Regression and causality

- Consider a regression of Y_i on D_i (binary) and covariates X_i.
  - If $Y_{1i} - Y_{0i} = \delta$ and $E(Y_{0i} \mid X_i, D_i) = \gamma' X_i$ then δ is estimated consistently by OLS.

Regression and causality

- Consider a regression of Y_i on D_i (binary) and covariates X_i.
  - If $Y_{1i} - Y_{0i} = \delta$ and $E(Y_{0i} \mid X_i, D_i) = \gamma' X_i$ then δ is estimated consistently by OLS.
  - More generally, if D_i is continuous, $Y_{di} = \alpha + \delta D_i + \eta_i$, and $E(\eta_i \mid X_i, D_i) = \gamma' X_i$, then the OLS coefficient estimate on D_i is consistent for δ.
  - These results are most interesting when X_i is discrete but still assumes no heterogeneity.

Regression and causality

- Suppose that x_i is a vector if discrete variables and X_i is a vector of indicators that fully saturates the model in x_i – but not interactions with D_i.
  - example: x_i is years of schooling and gender and X_i is a dummy variable for each year of schooling, a dummy variable for gender, and an interaction between gender and each schooling level
- In this case, if Y_1, Y_0 are independent of D_i conditional on x_i then $E(Y_{0i} \mid X_i, D_i) = \gamma' X_i$ (or $E(\eta_i \mid X_i, D_i) = \gamma' X_i$) is trivially satisfied so the only other assumption is that of no heterogeneity in treatment effects (and linearity in D_i).

Regression and causality

- If D_i is randomly assigned then (M-1) is satisfied for any X that is measured at baseline.
- So what can we say about OLS in this case?
  - The no heterogeneity and linearity assumptions don't cause bias asymptotically because D_i is independent of X_i.
  - But controlling for X_i can introduce finite sample bias.
  - See Freedman (2008).

Regression and causality

- We will stick with the binary D_i case now.
- Define $\delta_x = E(Y_i \mid D_i = 1, X_i = x) - E(Y_i \mid D_i = 0, X_i = x)$.
  - Under (M-1)',

$$\delta_x = E(Y_{1i} - Y_{0i} \mid X_i = x)$$

Regression and causality

- Suppose the model is saturated in X_i but we now relax the no heterogeneity assumption.

Regression and causality

- Suppose the model is saturated in X_i but we now relax the no heterogeneity assumption.
- Then

$$\delta_R = \sum_x \delta_x \left(\frac{p(X_i)(1-p(X_i))Pr(X_i = x)}{\sum_x p(X_i)(1-p(X_i))Pr(X_i = x)} \right)$$

where $p(X_i) := Pr(D_i = 1 \mid X_i = x)$

Regression and causality

- Suppose the model is saturated in X_i but we now relax the no heterogeneity assumption.
- Then

$$\delta_R = \sum_x \delta_x \left(\frac{p(X_i)(1-p(X_i))Pr(X_i = x)}{\sum_x p(X_i)(1-p(X_i))Pr(X_i = x)} \right)$$

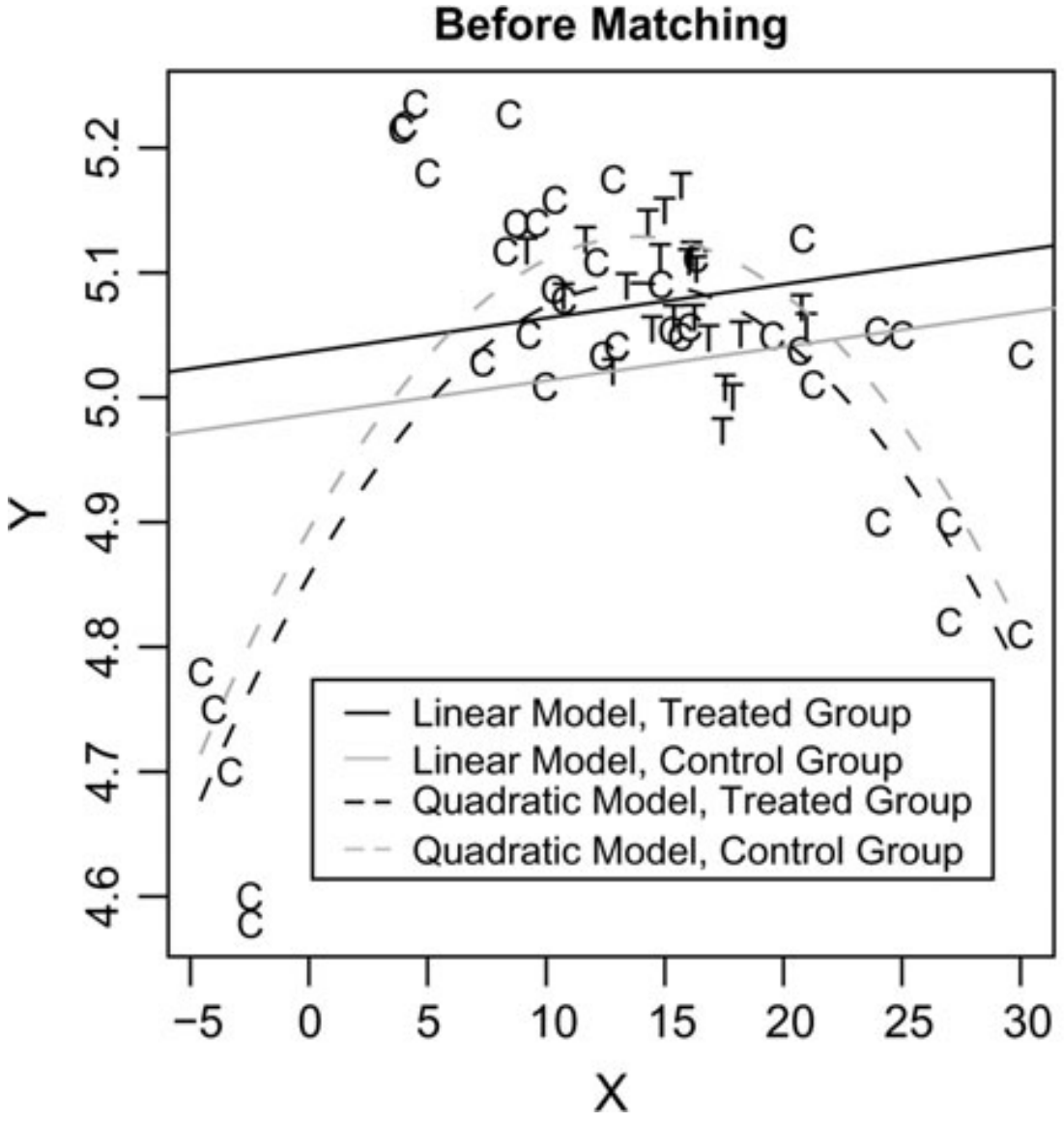
  where $p(X_i) := Pr(D_i = 1 \mid X_i = x)$
- When $Pr(D_i = 1 \mid X_i = x)$ equals 0 or 1, the weight is 0.
- This weighted average is different from the ATE, TT, and TUT, which can all also be seen as weighted averages of δ_x with different weights.

Why matching?

- Two reasons:
  1. The OLS weights are different from the ATE or ATT weights. This leads to different results if
     (a) $P(x) := Pr(D = 1 \mid X = x)$ varies in x
     (b) *and* δ_x varies in x
  2. Extrapolation:
     - If the model is not fully saturated in X then OLS extrapolates across observations (gives weight outside the common support).
     - Matching methods in a more controlled way.

Why matching?

- Consider the following example from Ho et al. (2007):



Exact matching

- Suppose X_i is discrete.
  - We can estimate δ_x for each value of x.
  - The ATE, for example, can be estimated simply as $n^{-1} \sum_{i=1}^{n} \hat{\delta}_{X_i}$
  - The ATT or other treatment effects can be estimates by average the δ_x's over subsamples.
    - This is what Angrist (1998) does (Table 3.3.1 in MHE)
  - Not feasible if there are too many "cells".

- If X_i is not discrete.
  - Discretize X_i.
  - Then do exact matching.
  - Todd (2006) calls this stratified or interval matching.
  - One version of this, called *coarsened exact matching* (CEM), has gained popularity lately.

Nearest neighbor

- One-to-one matching:
  - For each treated observation i choose the control observation j such that $d(X_i, X_j)$ is minimized.
  - Various different metrics can be used for d (Euclidean, Mahalanobis, etc.)
  - tie breaker necessary if X_i is discrete
- k-nearest neighbor matching
  - Choose the control observations that have the k smallest values of $d(X_i, X_j)$.
  - This does not avoid the possibility that a tie-breaker is needed.
  - Increases bias but reduces variance.

A class of matching estimators

- Let $\mathcal{I}_d = \{i : D_i = d\}$ and let $n_d = |\mathcal{I}_d|$ for each $d = 0, 1$.
- A class of matching estimators for the TT:

$$\hat{\Delta}_{TT} = \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} \left(Y_i - \sum_{j \in \mathcal{I}_0} w_{i,j} Y_j \right)$$

  - The weights should be calculated so that $\sum_{j \in \mathcal{I}_0} w_{i,j} Y_j$ is a good estimate of $E(Y_i \mid D_i = 0, X_i)$.
  - in other words, higher weights should be assigned to j with X_j close to X_i

A class of matching estimators

- The estimator can be viewed as follows:
  - First a new control sample is created by finding matches for each treatment observation.
  - The treatment and new control are “balanced”.
  - Then you take a difference in means – do what you would normally do when treatment is randomized.

Nonparametric regression-based matching

- Recall that

$$TT = E(Y \mid D = 1) - \int E(Y \mid D = 0, X = x) f_X(x \mid D = 1) dx$$

Nonparametric regression-based matching

- Recall that

$$TT = E(Y \mid D = 1) - \int E(Y \mid D = 0, X = x) f_X(x \mid D = 1) dx$$

- Therefore, we can use a nonparametric estimate of $g_0(x) := E(Y \mid D = 0, X = x)$ to calculate

$$\hat{\Delta}_{TT} = \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} (Y_i - \hat{g}_0(X_i))$$

Nonparametric regression-based matching

- Recall that

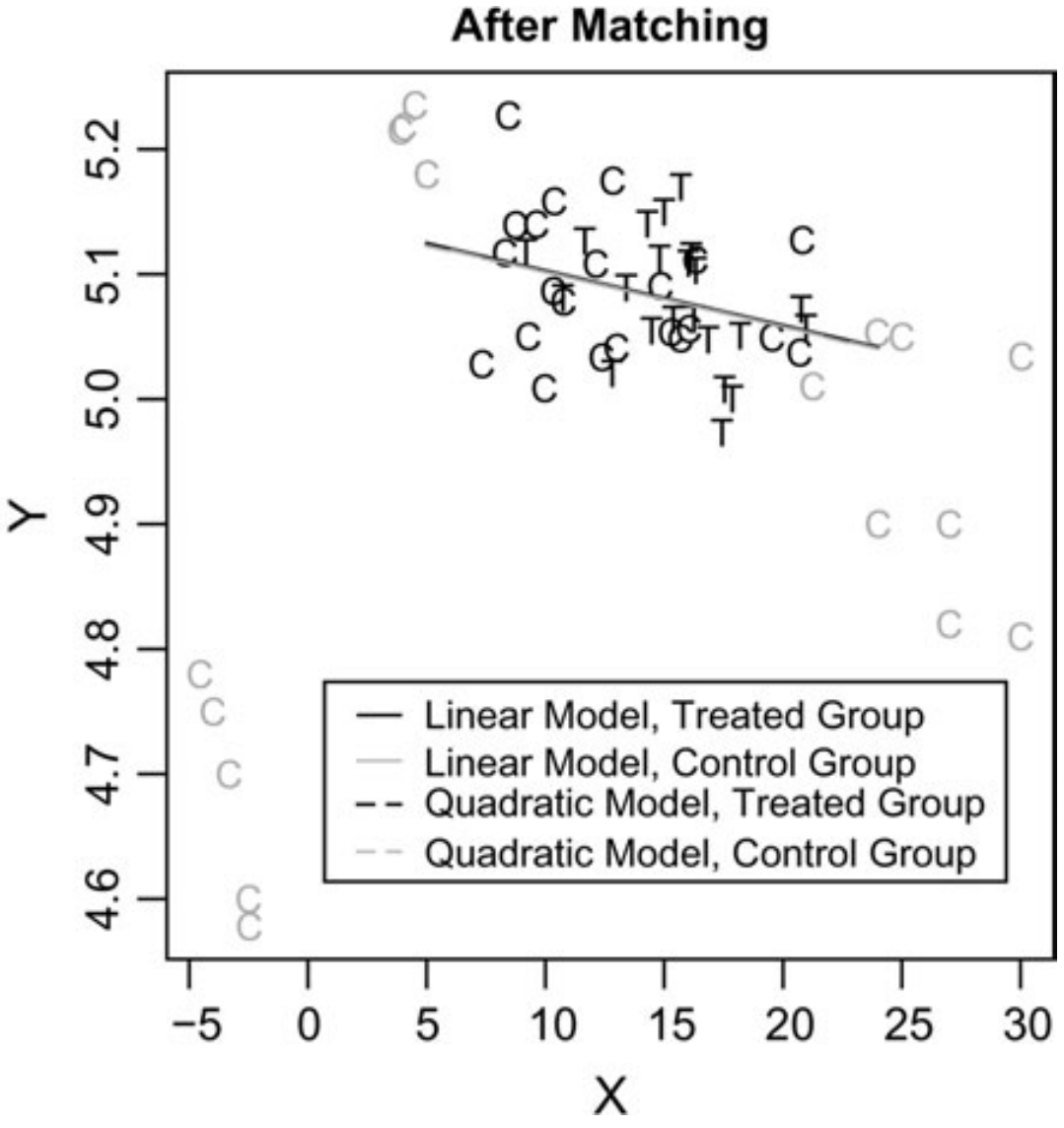
$$TT = E(Y \mid D = 1) - \int E(Y \mid D = 0, X = x) f_X(x \mid D = 1) dx$$

- Therefore, we can use a nonparametric estimate of $g_0(x) := E(Y \mid D = 0, X = x)$ to calculate

$$\hat{\Delta}_{TT} = \frac{1}{n_1} \sum_{i \in \mathcal{I}_1} (Y_i - \hat{g}_0(X_i))$$

- $\hat{g}_0(x)$ can be a kernel regression or local polynomial regression estimator, for example.

Example from Ho et al. (2007)



Other treatment effects

- Also, by (M-1) and (M-2),

$$TUT = \int E(Y \mid D = 1, X = x) f_X(x \mid D = 0) dx - E(Y \mid D = 0)$$

$$ATE = \int E(Y \mid D = 1, X = x) f_X(x) dx - \int E(Y \mid D = 0, X = x) f_X(x) dx$$

Other treatment effects

- Also, by (M-1) and (M-2),

$$TUT = \int E(Y \mid D = 1, X = x) f_X(x \mid D = 0) dx - E(Y \mid D = 0)$$

$$ATE = \int E(Y \mid D = 1, X = x) f_X(x) dx - \int E(Y \mid D = 0, X = x) f_X(x) dx$$

- We can, for example, use nonparametric estimates of g_0 and $g_1(x) := E(Y \mid D = 1, X = x)$,

$$\hat{\Delta}_{TUT} = \frac{1}{n_0} \sum_{i \in \mathcal{I}_0} (\hat{g}_1(X_i) - Y_i)$$

$$\hat{\Delta}_{ATE} = \frac{1}{n} \sum_{i \in \mathcal{I}_1 \cup \mathcal{I}_0} (\hat{g}_1(X_i) - \hat{g}_0(X_i))$$

Other treatment effects

- Also, by (M-1) and (M-2),

$$TUT = \int E(Y \mid D = 1, X = x) f_X(x \mid D = 0) dx - E(Y \mid D = 0)$$

$$ATE = \int E(Y \mid D = 1, X = x) f_X(x) dx - \int E(Y \mid D = 0, X = x) f_X(x) dx$$

- We can, for example, use nonparametric estimates of g_0 and $g_1(x) := E(Y \mid D = 1, X = x)$,

$$\hat{\Delta}_{TUT} = \frac{1}{n_0} \sum_{i \in \mathcal{I}_0} (\hat{g}_1(X_i) - Y_i)$$

$$\hat{\Delta}_{ATE} = \frac{1}{n} \sum_{i \in \mathcal{I}_1 \cup \mathcal{I}_0} (\hat{g}_1(X_i) - \hat{g}_0(X_i))$$

- It's straightforward to estimate things like $E(Y_1 - Y_0 \mid X_1 = x)$ as well.

Propensity score

- The propensity score –

$$P(x) = Pr(D = 1 \mid X = x)$$

  - under the conditional independence assumption $(Y_1, Y_0) \perp\!\!\!\perp D \mid X$,

$$E(Y_d \mid D, P(X)) = E(Y_d \mid P(X))$$

  - first shown by Rosenbaum and Rubin (1983)
  - match on $P(X)$!
    - this reduces the complexity of the estimation problem ...
    - if a functional form for $P(X)$ is known

Propensity score matching

1. Estimate the propensity score $P(x)$ through a logit, probit, semiparametric or nonparametric method.
2. Create $P_i = P(X_i)$ for each observation i
3. Use any of the matching estimators above, replacing X_i with P_i

Some theoretical results

- Kernel-based matching on X
  - if the bias in the first stage is small enough, we get $\sqrt{n}$ convergence and asymptotic normality with variance

$$V_{eff} = E\left(\frac{Var(Y_1 \mid X)}{P(X)} + \frac{Var(Y_0 \mid X)}{1 - P(X)}\right) + Var\left(g_1(X) - g_0(X)\right)$$

    - this is the semiparametric efficiency bound
    - $\sqrt{n}$ convergence and asymptotic normality – this means that for large samples it performs as well as a parametric method

Some theoretical results

- Kernel-based matching on X
  - if the bias in the first stage is small enough, we get $\sqrt{n}$ convergence and asymptotic normality with variance

$$V_{eff} = E\left(\frac{Var(Y_1 \mid X)}{P(X)} + \frac{Var(Y_0 \mid X)}{1 - P(X)}\right) + Var\left(g_1(X) - g_0(X)\right)$$

    - this is the semiparametric efficiency bound
    - $\sqrt{n}$ convergence and asymptotic normality – this means that for large samples it performs as well as a parametric method
  - however, the larger the dimension of X, the harder it is to reduce the first stage bias – curse of dimensionality

Some theoretical results

- Is use of the propensity score a solution to the curse of dimensionality?
  - case 1: the propensity score is known
    - the first stage bias will be easier to manage because of the dimension reduction
    - but it turns out that the asymptotic variance is larger!
    - Rothe (2016) partially resolves this "propensity score paradox": a modified estimator will obtain the efficiency bound *and* require weaker regularity conditions

Some theoretical results

- Is use of the propensity score a solution to the curse of dimensionality?
  - case 2: the propensity score is estimated parametrically
    - the specification can be logit or probit, for example
    - if the specification is the right one then the efficiency bound is attained
    - but generally the propensity score is misspecified
    - can have better finite sample performance
    - no curse of dimensionality

Some theoretical results

- Is use of the propensity score a solution to the curse of dimensionality?
  - case 3: the propensity score is estimated nonparametrically
    - attains the efficiency bound
    - has two *nuisance* parameters – asymptotics require some strong regularity conditions
    - curse of dimensionality returns!

Common support

- If (M-2) does not hold, we need to restrict the sample to the common support S_{10}:

$$E(E(Y_1 - Y_0 \mid X) \mid D = 1, X \in S_{10})$$

- Checking/enforcing
  - In the matching on X context, it is difficult to check for common support because X is high dimensional
    - King and Zeng (2007) – convex hull condition
    - a conservative approach
    - in some cases the convex hull is empty!

Common support

- Alternatively, trim observations where $P(X_i)$ is near 0 or 1.
- Smith and Todd (2005) suggest following the procedure:
  1. estimate $\hat{f}_{P(X)}(p \mid D = 1)$ and $\hat{f}_{P(X)}(p \mid D = 0)$
  2. remove observations where $\hat{f}_{P(X)}(P(X_i) \mid D = d) = 0$ for $d = 0$ or $d = 1$
  3. find a cutoff c_q so that removing those with $\hat{f}_{P(X)}(P(X_i) \mid D = 0) \leq c_q$ or $\hat{f}_{P(X)}(P(X_i) \mid D = 1) \leq c_q$ removes q percent of the remaining sample (1%, 2%, or 5%, in practice)
- If you don't do this, you will extrapolate out of sample!

Balance tests

- The covariates should have the same distribution in the treatment and the matched samples.
- There are various statistics – difference in means is the simplest.
- The standardized difference is also common – for each k,

$$SDIFF(X_k) = 100 \frac{\frac{1}{n_1}\sum_{i\in\mathcal{I}_1} X_{ki} - \sum_{j\in\mathcal{I}_0} w_{ij} X_{kj}}{\sqrt{\frac{1}{2} Var_{i\in\mathcal{I}_1}(X_{ki}) + \frac{1}{2} Var_{i\in\mathcal{I}_0}(X_{ki})}}$$

- No real guidance for how big is too big.
- Rosenbaum and Rubin (1985) say that 20 is a “large” value

Balance tests

- Ho et al. warn that smaller matched samples can be more likely to pass a balance test because of lower power.
- Propensity score based tests: $X \perp\!\!\!\perp D \mid P(X)$
  - regress X on flexible polynomial in $P(X), D$ and test significance of terms involving D
  - Shaikh et al. (2006) propose a test based on $f_{P(X)}(p \mid D = 1)$ and $f_{P(X)}(p \mid D = 0)$

LaLonde (1986)

- LaLonde's critique: observational methods cannot always reproduce the results from experimental study
- NSW experiment – job training targeted to those who have highest barriers to employment (10,000 participants across ten cities in the US in 1974-1975)
- comparison groups drawn from CPS and PSID – nationally representative longitudinal surveys
- outcome is earnings in 1978, Y_{1978}
- using this type of data to construct an observational estimate:
  - compare the NSW treatment to CPS or PSID comparison sample
  - use observational methods to control for differences
- the bias can be estimated:

$$bias = \hat{\Delta}_{TT}^{obs.} - \hat{E}^{NSW}(Y_{1978} \mid D = 1) - \hat{E}^{NSW}(Y_{1978} \mid D = 0)$$

Dehejia and Wahba

- LaLonde found that various regression, panel data, selection methods could not produce the same results as the NSW treatment-control comparison
- Dehejia and Wahba (1999,2002) argue that propensity score matching solves LaLonde's critique
  - match NSW treatment to comparison group, compare to experimental estimate
  - or directly match NSW control to comparison group
- Smith and Todd (2005) was written in response

Dehejia and Wahba

- note that selection here is different than if we had observational data from a context where individuals could choose to sign up for job training
  - we are controlling for differences in two study populations as well
  - the bias is the bias in estimating the treatment effect for the NSW experimental population (a select population)
  - *not* the average effect of job training
- The debate is very useful to read for understanding prevalent issues with propensity score matching.

Smith and Todd (2005)

Smith and Todd (2005)

- not robust
- reiterate concerns of Heckman, Ichimura, Smith and Todd (1998) about the importance of
  - rich set of conditioning variables available
  - dependent variable is measured in same way for both groups
  - comparison sample being from same local labor markets

Smith and Todd (2005)

- Inherent tradeoff between (M-1) and (M-2)
  - adding more controls to X (pre-program earnings for one year, two years) reduces the available sample
  - using the propensity score can sometimes mask this
  - note how Smith and Todd (2005) (and DW before them) use sample restrictions *before* matching on the propensity score
- issues with combining separate samples:
  - calendar time vs. program time
  - self-reported vs. administrative earnings records

Diff-in-diff matching

- Suppose there is an individual time-invariant fixed effect in earnings.
  - Using preprogram earnings as a control is not the right approach.
  - Instead we would difference out the fixed effect.
- We can combine this familiar approach with matching by redefining the dependent variable as $Y_{after} - Y_{before}$.
  - Then match and average.
  - Smith and Todd (2005) find that this approach is more robust in the NSW data than matching on pre-program earnings.

Setting

- Analysis of vocational rehab (VR) program of the Canada Pension Plan Disability Program (CPPD).
- Does the VR program work?

Comparison group

- Treatment group – cohort of individuals who started VR program in 1998
- Potential control groups:
  - all CPPD beneficiaries
  - VR dropouts
  - CPPD beneficiaries who are “reassessed”

Distribution of *P*

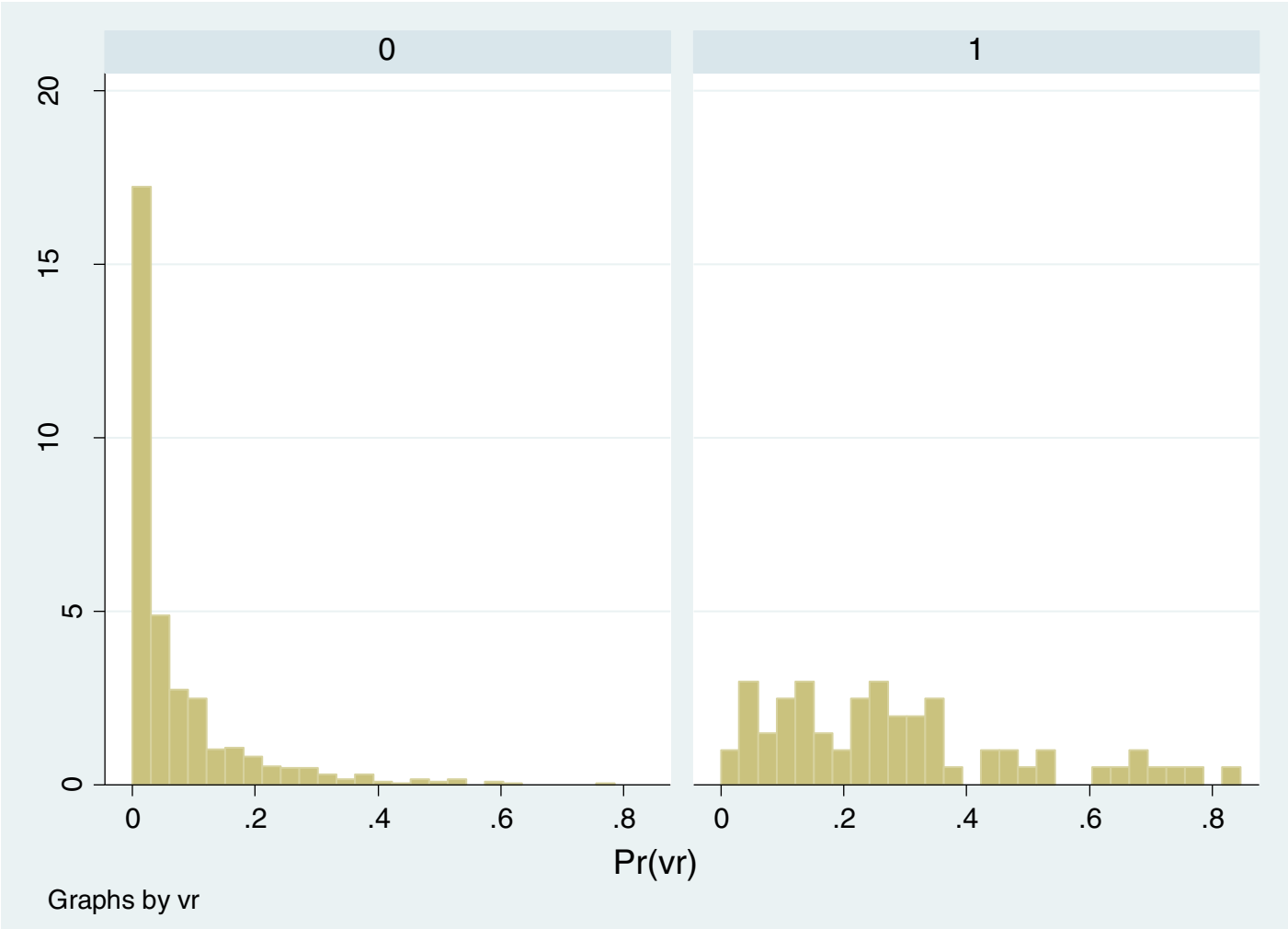


Figure 2 Histogram of propensity score for comparison and treatment groups, women.

Pre-matching balance

Table 3 Descriptive statistics, women

	Treatment		Comparison	
	Mean	**Standard deviation**	**Mean**	**Standard deviation**
Age (at onset of disability)	36.552	7.361	43.423	8.668
Married	0.507	0.504	0.618	0.486
Have children	0.522	0.503	0.410	0.492
[Less than high school]				
High school	0.552	0.501	0.325	0.469
Post-secondary	0.164	0.373	0.143	0.350
University degree	0.149	0.359	0.097	0.296
[Other]				
Infectious and parasitic diseases	–	–	–	–
Cancer	0.030	0.171	0.082	0.275
Blood diseases	0.000	0.000	0.000	0.000
Mental disorders	0.254	0.438	0.390	0.488
Diseases of the nervous system	0.045	0.208	0.107	0.309
Circulatory diseases	0.090	0.288	0.052	0.222
Respiratory diseases	–	–	–	–
Diseases of digestive system	0.090	0.288	0.017	0.131
Genitourinary system diseases	0.030	0.171	0.010	0.100
Musculoskeletal and soft-tissue disorders	0.299	0.461	0.263	0.441

Why matching?

- conditional independence is plausible

Why matching?

- conditional independence is plausible
- why not OLS?

estimators

- four different estimators:
  - kernel regression matching
  - llr matching
  - genetic algorithm matching
  - inverse probability weighting

Post-matching balance

Table 4 Standardized differences in treatment and comparison group, inverse probability weighting

	Men		Women	
	Before matching	After matching	Before matching	After matching
Age	−67.9	12.7	−61.2	1.8
High school	66.3	0.7	35.7	4.1
College	13.6	3.8	5.5	−4.8
University	3.3	−6.9	13.7	−1.1
Married	−4.6	6.1	−8.1	−7.8
Have children	6.7	−4.6	20.8	3.5
Infectious and parasitic diseases	8.1	−1.6	–	–
Cancer	8.2	−2.5	−14.0	−1.8
Mental disorders	−14.6	−0.7	−17.8	−6.8
Diseases of the nervous system	16.9	−2.5	−20.8	−2.0
Circulatory diseases	−18.8	8.8	−1.9	−4.8
Respiratory diseases	0.0	−7.6	–	–
Diseases of digestive system	−5.3	−13.4	20.3	23.5
Genitourinary system diseases	−9.3	−1.3	5.9	−6.0
Musculoskeletal and soft-tissue disorders	−4.4	−1.5	14.2	6.2
Congenital diseases	9.5	14.1	4.7	2.9
Zero earnings 1-year prior to application	9.5	6.2	21.8	10.8
Zero earnings 2-years prior to application	11.6	21.6	14.5	2.4
Earnings 1-year prior to application	−18.7	1.0	−9.5	−0.4
Earnings 2-years prior to application	−22.7	−13.9	−3.6	5.8
Unemployment rate	0.2	−0.2	−19.0	−8.2
Duration on CPPD program	−9.7	1.7	18.2	−2.2

Notes: Entries in the table are standardized differences between the treatment and comparison groups.

Results

Table 8 Estimates of the ATET of VR on individual outcomes, women

	Matching estimator		Inverse probability weighting	Genetic matching
	Kernel	Local linear		
	Sample restricted to propensity score values [0.001, 0.35]			
Leaving disability rolls	0.104	0.104	0.095	0.122
	(0.086)	(0.084)	(0.080)	(0.111)
Gainful employment	0.110	0.109	0.108	0.143
	(0.082)	(0.081)	(0.078)	(0.104)
Substantial gainful employment	0.169**	0.158**	0.144*	0.184*
	(0.079)	(0.079)	(0.078)	(0.103)
Sample size	631	631	631	631
	Sample restricted to propensity score values [0.001, 0.30]			
Leaving disability rolls	0.105	0.085	0.096	0.071
	(0.088)	(0.088)	(0.087)	(0.120)
Gainful employment	0.133	0.120	0.130	0.143
	(0.087)	(0.087)	(0.084)	(0.122)
Substantial gainful employment	0.180**	0.156*	0.159*	0.167
	(0.083)	(0.085)	(0.084)	(0.118)
Sample size	615	615	615	615
	Sample restricted to propensity score values [0.001,0.40]			
Leaving disability rolls	0.109	0.093	0.093	0.077
	(0.086)	(0.081)	(0.078)	(0.105)
Gainful employment	0.147	0.137	0.138	0.154
	(0.085)	(0.076)	(0.074)	(0.099)
Substantial gainful employment	0.195**	0.175**	0.171**	0.192**
	(0.081)	(0.075)	(0.074)	(0.098)
Sample size	642	642	642	642

Notes: * denotes statistical significance at the 10 percent level; ** denotes statistical significance at the 5 percent level. See notes for Table 7.

Conditional independence

When is $(Y_0, Y_1) \perp\!\!\!\perp D \mid X$?

- i.e., what should go in X?
- everything correlated with D_i?
- everything that has a causal effect on D_i?
- everything correlated with D_i that also has a causal effect on Y_{di}?
- the kitchen sink?

Modeling with equations

One way to think about it is by specifying a model, in equations...

Modeling with equations

One way to think about it is by specifying a model, in equations...

$$Y_{di} = \beta_d' X_i + U_{di}$$
$$D_i = \mathbf{1}(\gamma' Z_i + V_i \geq 0)$$

Modeling with equations

Example 1:

- If $Z_i = X_i$ ...

Modeling with equations

Example 1:

- If $Z_i = X_i$ ...then (M-1) is implied by $(U_1, U_0) \perp\!\!\!\perp V|X$.

Modeling with equations

Example 1:

- If $Z_i = X_i$ ...then (M-1) is implied by $(U_1, U_0) \perp\!\!\!\perp V|X$.
- The latter holds if:
  - (a) $(U_1, U_0) \perp\!\!\!\perp V$ and $(U_1, U_0, V) \perp\!\!\!\perp X$
  - or (b) $(U_1, U_0) \perp\!\!\!\perp V$ and $X \perp\!\!\!\perp V \mid (U_1, U_0)$
  - or (c) $(U_1, U_0) \perp\!\!\!\perp V$ and $X \perp\!\!\!\perp (U_1, U_0) \mid V$
- These conditions, and others, can be derived using

$$f_{U_1,U_0|V,X} = \frac{f_{U_1,U_0,V,X}}{\int f_{U_1,U_0,V,X} d(u_1, u_0)}$$

Modeling with equations

Example 2:

- Let $X_i = (X_{1i}, X_{2i})$ and $Z_i = (X_{1i}, \tilde{Z}_i)$ and $W_i = (X_1, X_2, \tilde{Z})$

Modeling with equations

Example 2:

- Let $X_i = (X_{1i}, X_{2i})$ and $Z_i = (X_{1i}, \tilde{Z}_i)$ and $W_i = (X_1, X_2, \tilde{Z})$
- Then
  - $(Y_1, Y_0) \perp\!\!\!\perp D \mid W$ if $(U_1, U_0) \perp\!\!\!\perp V \mid W$
  - $(Y_1, Y_0) \perp\!\!\!\perp D \mid X_1$ if $(U_1, U_0, X_2) \perp\!\!\!\perp (V, \tilde{Z}) \mid X_1$
  - $(Y_1, Y_0) \perp\!\!\!\perp D \mid X$ if $(U_1, U_0) \perp\!\!\!\perp (V, \tilde{Z}) \mid X$
  - $(Y_1, Y_0) \perp\!\!\!\perp D \mid Z$ if $(U_1, U_0, X_2) \perp\!\!\!\perp V \mid Z$

Modeling with equations

Example 3 – proxy control

- Suppose that $Z_i = \psi' X_i + \eta_i$.
- Then $(Y_1, Y_0) \perp\!\!\!\perp D \mid Z$ if
  - $(U_1, U_0) \perp\!\!\!\perp V \mid X, \eta$
  - *and* $X \perp\!\!\!\perp V \mid Z$
- How did I derive this?
  - First, using the outcome and selection equations, $(Y_1, Y_0) \perp\!\!\!\perp D \mid Z$ if $(X, U_1, U_0) \perp\!\!\!\perp V \mid Z$.
  - Second,

$$f_{X,U_1,U_0|V,Z} = \frac{f_{U_1,U_0|\eta,V,X,Z} f_{X|V,Z} f_{V,Z}}{\int f_{U_1,U_0|\eta,V,X,Z} f_{X|V,Z} f_{V,Z} d(x, u_1, u_0)}$$

  - Third, use $f_{U_1,U_0|\eta,V,X,Z} = f_{U_1,U_0|\eta,V,X}$.

Modeling with equations

Example 3 – proxy control

- Suppose that $Z_i = \psi' X_i + \eta_i$.
- Then $(Y_1, Y_0) \perp\!\!\!\perp D \mid Z$ if
  - $(U_1, U_0) \perp\!\!\!\perp V \mid X, \eta$
  - *and* $X \perp\!\!\!\perp V$
  - *and* $\eta \perp\!\!\!\perp V \mid X$
- The intuition is pretty clear if $\eta \equiv 0$ – in addition to the same condition for controlling for X we also need that X and V (and hence X and D) are only related through the scalar index $\psi' X$.

Modeling with equations

Example 3 – proxy control

- Suppose that $Z_i = \psi' X_i + \eta_i$.
- When is $(Y_1, Y_0) \perp\!\!\!\perp D \mid X$ (in the case where η is not 0)?

Modeling with equations

Example 3 – proxy control

- Suppose that $Z_i = \psi' X_i + \eta_i$.
- When is $(Y_1, Y_0) \perp\!\!\!\perp D \mid X$ (in the case where η is not 0)?
  - An application of example 2 shows that $(U_1, U_0) \perp\!\!\!\perp (\eta, V) \mid X$ is sufficient.
  - note: conditions for controlling for X vs controlling for Z are not nested.

MHE example

Mediator

- Suppose that $X_i = Z_i$ is an outcome of treatment – $X_i = \psi_0 + \psi_1 D_i + \eta_i$.
- It is unclear how the treatment effect should be defined.
- We can write $Y_{di} = \beta_{0d} + \beta_{1d} X_{di} + U_{di}$ where $X_{di} = \psi_0 + \psi_1 d + \eta_i$ ...
- ... so

$$Y_{1i} - Y_{0i} = \beta_{01} - \beta_{00} + (\beta_{11} - \beta_{10})(\psi_0 + \eta_i) + \beta_{11}\psi_1 + U_{1i} - U_{0i}$$

and

$$E(Y_{1i} - Y_{0i}) = \beta_{01} - \beta_{00} + (\beta_{11} - \beta_{10})\psi_0 + \beta_{11}\psi_1$$

MHE example

Mediator

- Suppose that $X_i = Z_i$ is an outcome of treatment – $X_i = \psi_0 + \psi_1 D_i + \eta_i$.
- It is unclear how the treatment effect should be defined.
- Alternatively,

$$Y_{1i} - Y_{0i} = \beta_{01} - \beta_{00} + (\beta_{11} - \beta_{10})X_i + U_{1i} - U_{0i}$$

and

$$E(Y_{1i} - Y_{0i}) = \beta_{01} - \beta_{00} + (\beta_{11} - \beta_{10})E(X_i)$$

MHE example

Mediator

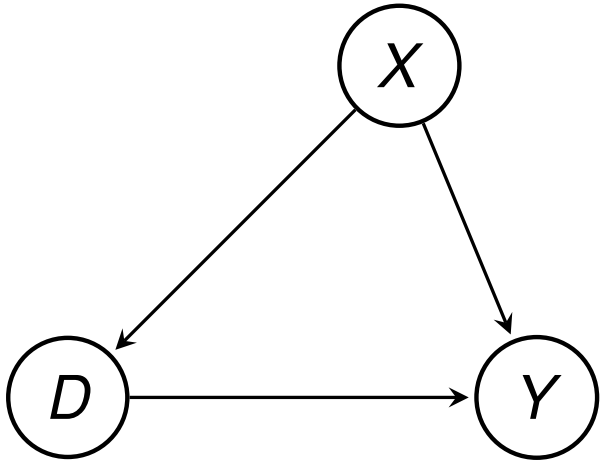
- What is identified?
  - If $E(\eta_i \mid D_i) = 0$ and $E(U_{di} \mid D_i) = 0$ then the first version of the ATE is identified from $E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0)$.
  - If $(U_0, U_1) \perp\!\!\!\perp (D, \eta)$ then $E(Y_i \mid D_i = 1, X_i = x) - E(Y_i \mid D_i = 0, X_i = x) = \beta_{01} - \beta_{00} + (\beta_{11} - \beta_{10})x$.
  - If η is correlated with U_1, U_0 – *confounded mediator* – version 2 is not identified but version 1 is – "bad control"
- More on mediation analysis in Pearl (2014).

Using DAGs

- A DAG is a *directed acyclic graph*.
- The graph is meant to encode causal relationships, in much the same way that our equations do.
- The *backdoor criterion* is a useful way to determine whether we should control for X_j.

Using DAGs

- An example DAG



Using DAGs

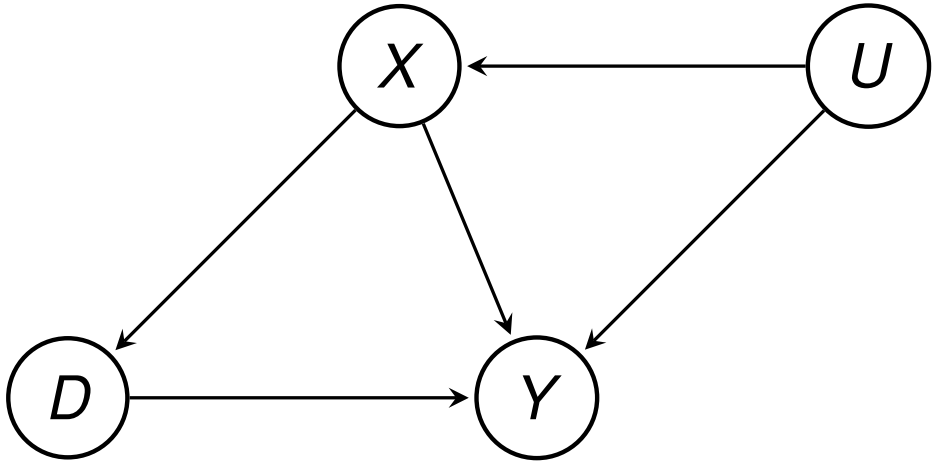
- Arrows are directional, indicating the direction of causality.
- Lack of an arrow between two variables means no causal effect.
- Simultaneity/reverse causality not allowed.
- A collider is a variable that has two arrows entering it
- Any other variable is a non-collider.

Using DAGs

- Backdoor criterion
  - X does not include any variables that are downstream from D
  - every "backdoor path" from D to Y – a path between D and Y including an arrow into D – either (a) includes a collider that is not part of X or (b) includes no colliders but includes a variable in X
- If this is satisfied then conditioning on X identifies the causal effect of D on Y.

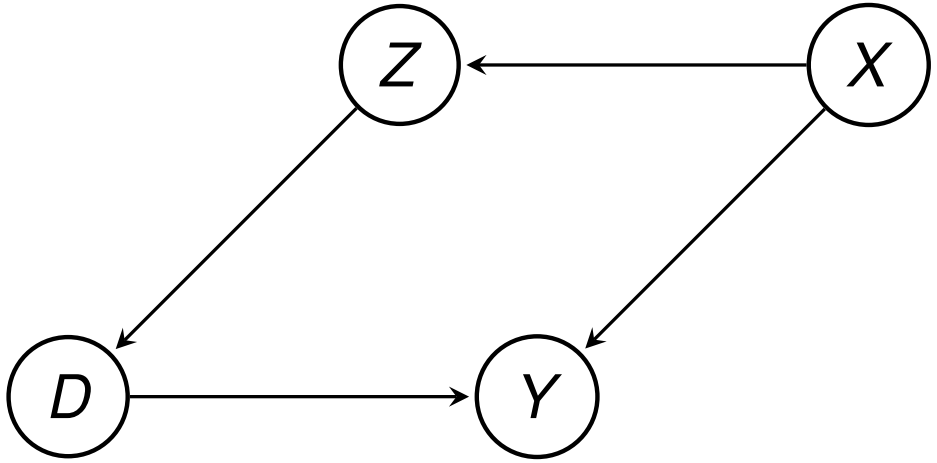
Using DAGs

- Example:



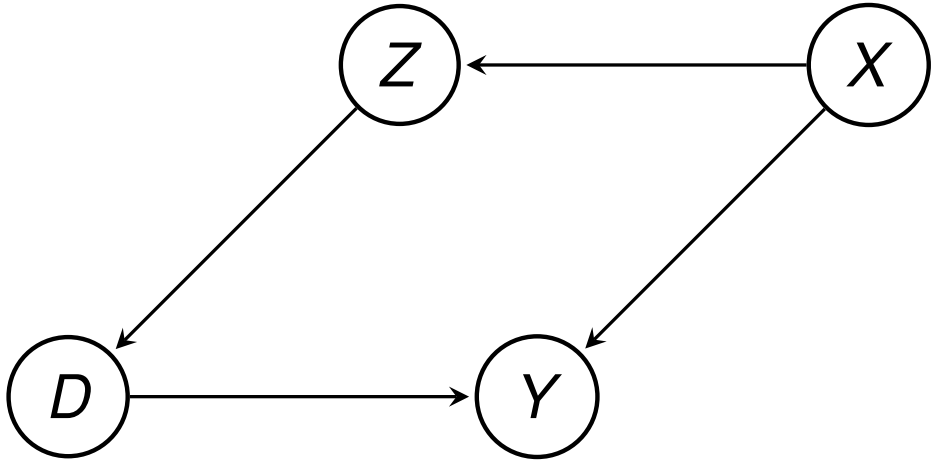
Using DAGs

- Example:



Using DAGs

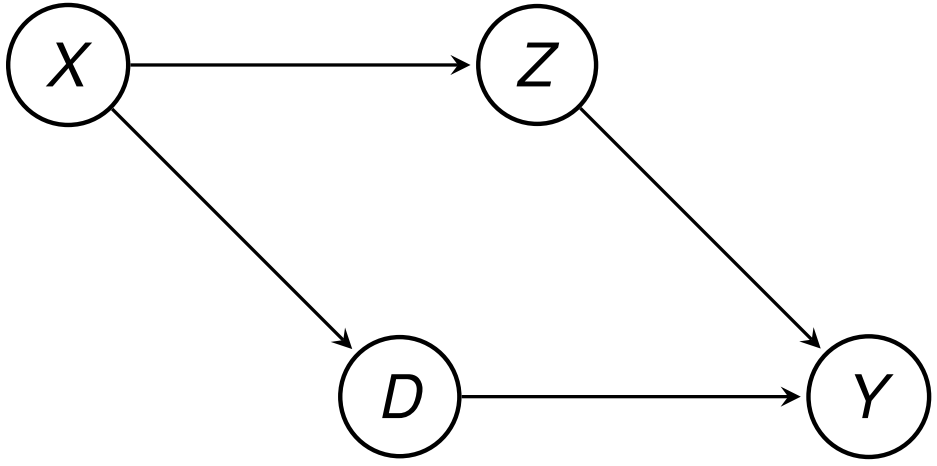
- Example:



- It's sufficient to control for Z.

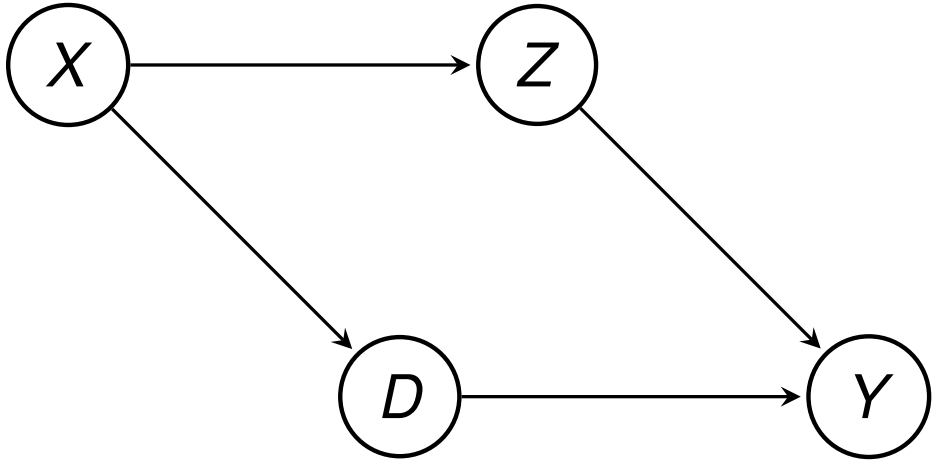
Using DAGs

- Example:



Using DAGs

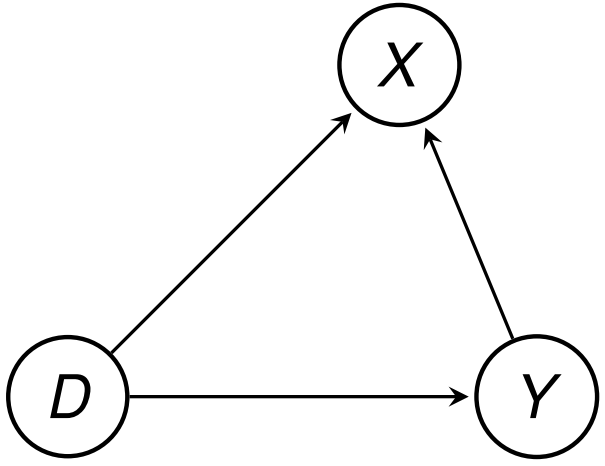
- Example:



- It's sufficient to control for Z.

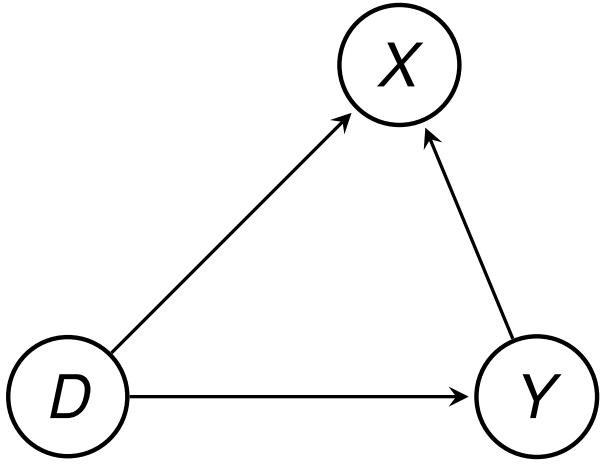
Using DAGs

- Collider bias example:



Using DAGs

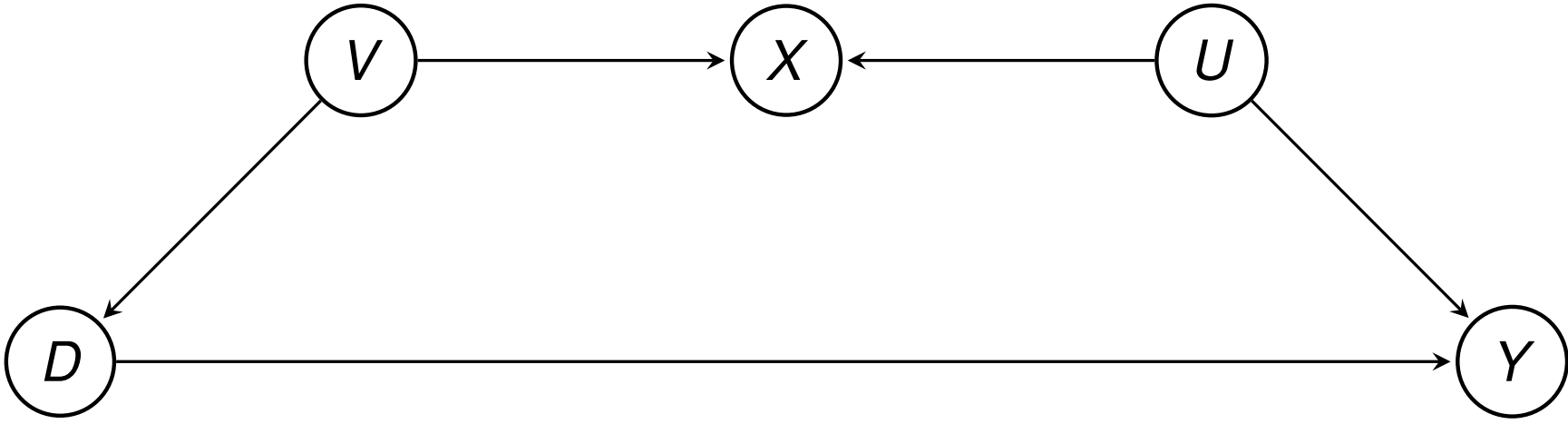
- Collider bias example:



- Don't control for X!

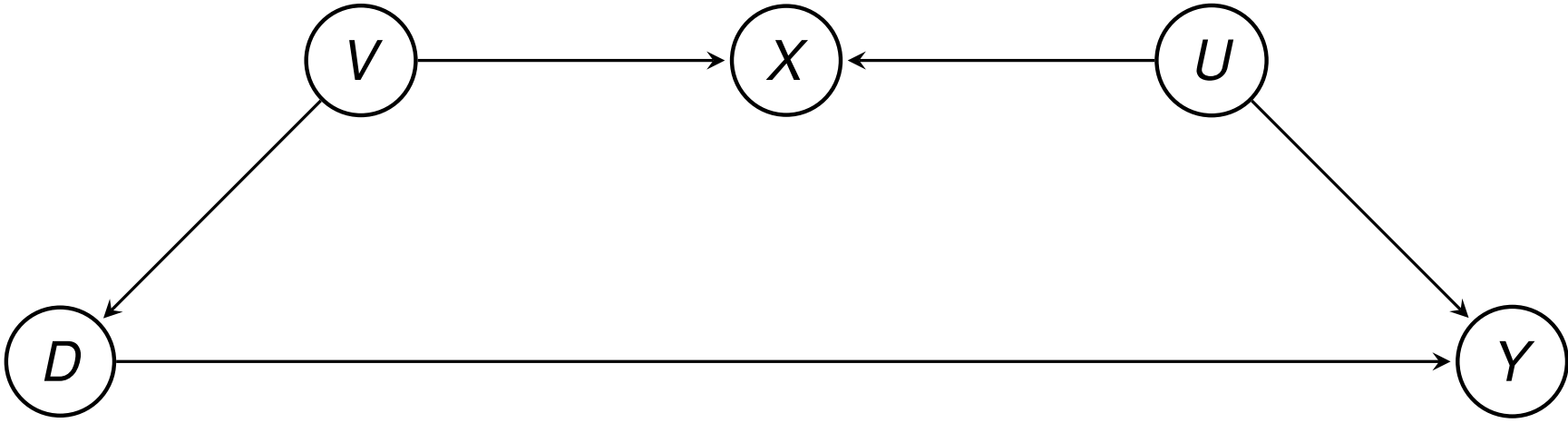
Using DAGs

- Collider bias example:



Using DAGs

- Collider bias example:



- Don't control for X!

Using DAGs

- Collider bias is similar but distinct from the confounded mediator problem.
- Both are reasons to not necessarily include everything related to both D and Y.
- See here for more examples.

Exacerbating bias

- Another reason to not include everything related to both D and Y:
- In reality, (M-1) is likely not satisfied exactly and including an additional control can make the bias worse.
  - Suppose that $Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + u_i$ where $Cov(D, u) \neq 0$
    - bias if X is included: $\frac{Cov(\tilde{D},u)}{Var(\tilde{D})}$
    - bias if X is omitted: $\frac{Cov(D,\beta_2 X+u)}{Var(D)}$
    - If X is included, the numerator is often smaller but the denominator is necessarily bigger!
  - "Throwing out the baby with the bathwater."

Robustness

- What if there are variables that we fail to control for...how bad can the bias be?
  - Altonji, Elder, Taber
  - Oster (2019)
  - Cinelli and Hazlett (2019)
  - Rosenbaum (1987) and Ichino et al. (2008)

Oster (2019)

- Suppose $Y = \beta X + \Psi\omega^0 + W_2 + \epsilon$ where X is scalar treatment, ω^0 are included confounders, W_2 is an index of unobserved confounders.
- result 1:
  - Suppose that $Cov(X, \Psi\omega^0)/Var(\Psi\omega^0) = Cov(X, W_2)/Var(W_2)$ and another more technical condition (Assumption 3).
  - Let $\tilde{\beta}$ and $\tilde{R}$ denote the coefficient on X and the R-squared from a regression of Y on X and ω^0.
  - Let β^0 and R^0 denote the coefficient on X and the R-squared from a regression of Y on X alone.
  - Then there is a unique, estimable value of ν such that $\tilde{\beta} - \nu\beta^0$ is a consistent estimator for β.

Oster (2019)

- Suppose $Y = \beta X + \Psi\omega^0 + W_2 + \epsilon$ where X is scalar treatment, ω^0 are included confounders, W_2 is an index of unobserved confounders.
- result 1:
  - Suppose that $Cov(X, \Psi\omega^0)/Var(\Psi\omega^0) = Cov(X, W_2)/Var(W_2)$ and another more technical condition (Assumption 3).
  - Let $\tilde{\beta}$ and $\tilde{R}$ denote the coefficient on X and the R-squared from a regression of Y on X and ω^0.
  - Let β^0 and R^0 denote the coefficient on X and the R-squared from a regression of Y on X alone.
  - Then there is a unique, estimable value of ν such that $\tilde{\beta} - \nu\beta^0$ is a consistent estimator for β.
  - Let R^{max} be the hypothetical R-squared from a regression of Y on X, ω^0, and W_2.
  - ν is a function of R^{max}.

Oster (2019)

- Suppose $Y = \beta X + \Psi\omega^0 + W_2 + \epsilon$ where X is scalar treatment, ω^0 are included confounders, W_2 is an index of unobserved confounders.
- result 2:
  - Define δ such that $\delta Cov(X, \Psi\omega^0)/Var(\Psi\omega^0) = Cov(X, W_2)/Var(W_2)$.
  - There is a unique value of δ (also a function of R^{max}) for which $\beta = 0$.
  - We can assess sensitivity by considering whether this is a plausible value for the proportionality parameter.
  - `psacalc` in Stata.

Cinelli and Hazlett (2019)

- Oster's proportionality constant maybe isn't easy to interpret.
- Cinelli and Hazlett (2019) derive a bias formula in terms of partial R^2.
- D treatment, X included, Z excluded.
- Suppose that $R^2_{Y\sim Z|X,D} = R^2_{D\sim Z|X}$
  - If this common partial R^2 is equal to

$$\frac{1}{2}\left(\sqrt{f_q^4 + 4f_q^2} - f_q^2\right)$$

    where $f_q^2 = q^2 R^2_{Y\sim D|X}/(1 - R^2_{Y\sim D|X})$ then including the unobserved confounder reduces the coefficient on D by $100q\%$.
  - They call $RV = RV_1$ the *robustness value*.

Cinelli and Hazlett (2019)

- They also have a robustness value at which statistical significance is lost.
- Also: “if Z explained all residual variance in the outcome how strongly associated with treatment would it need to be to eliminate the estimated effect?”

Cinelli and Hazlett (2019)

- They also have a robustness value at which statistical significance is lost.
- Also: "if Z explained all residual variance in the outcome how strongly associated with treatment would it need to be to eliminate the estimated effect?"
- The answer is $R^2_{Y \sim D|X}$.

Cinelli and Hazlett (2019)

- They also have a robustness value at which statistical significance is lost.
- Also: "if Z explained all residual variance in the outcome how strongly associated with treatment would it need to be to eliminate the estimated effect?"
- The answer is $R^2_{Y\sim D|X}$.
- They also provide tools for bounding the strength of unobserved confounders using observed covariates. (Section 4.4)
- This paper is easy to read and full of useful information for sensitivity analysis in a regression framework; `sensemakr` in R

Rosenbaum bounds

- Suppose that U_i is an unobserved confounder and let $Pr(D_i = 1 \mid X_i = x, U_i) = \exp(\beta' X_i + \gamma U_i)$. Then the relative odds of treatment for two observations with $X_i = X_j$, $u_i \neq u_j$ is

$$\Gamma := \frac{Pr(D_i = 1 \mid X_i, u_i)/Pr(D_i = 0 \mid X_i, u_i)}{Pr(D_j = 1 \mid X_j, u_j)/Pr(D_j = 0 \mid X_j, u_j)} = \exp(\gamma(u_i - u_j))$$

- This is bounded between $e^{-\gamma}$ and e^{γ} (scaling properly so that $\gamma > 0$).
- Based on Rosenbaum (1987), Stata codes `mhbounds` and `rbounds` provide bounds on the significance of estimated treatment effects for different specified values of Γ.
- This relies on some strong assumptions, only works for specific cases.
- See homework for an example of the use of `rbounds`.

Ichino et al. (2008)

- They simulate an unobserved confounder U with a given probability that depends on treatment (D) and outcomes (Y).
- The probabilities can be taken to match a particular observed covariate in X.
- The matching estimator is calculated in the simulated data using (U, X) instead of only X.
- This also relies on some strong assumptions, only works for specific cases (binary Y, e.g.)
- When U is simulated there is a corresponding odds ratio for outcomes ("outcome effect") and for treatment ("selection effect"). These vary with X_i but can be averaged.