TREATMENT EFFECT ESTIMATION WITH UNCONFOUNDED ASSIGNMENT

American Society of Health Economists Workshop
Philadelphia, PA
June 12, 2016

Jeff Wooldridge
Department of Economics
Michigan State University

1. Introduction
2. Basic Concepts
3. Unconfoundedness and Overlap
4. Identification of the Average Treatment Effects
5. A Strategy for Program Evaluation
6. Estimating the Propensity Score
7. Assessing and Improving Overlap
8. Estimating the Treatment Effects
9. Introduction to Regression Discontinuity
10. The Sharp RD Design
11. The Fuzzy RD Design
12. Unconfoundedness versus FRD
13. Additional Analyses

1. Introduction

- What kinds of questions can we answer using a potential outcomes approach to treatment effect estimation?

(i) What are the effects of a job training program on labor market outcomes?

(ii) What are the health spending effects of a newly available health insurance?

- The main issue in program evaluation concerns the nature of the assignment, intervention, or “treatment.”
- Is the “treatment” randomly assigned? (Still rare.)
- Is treatment randomly assigned conditional on observable covariates? (“Unconfoundedness” or “ignorability” or “selection on observables” or “exogenous” assignment.)

- Or, does assignment depend fundamentally on unobservables, where the dependence cannot be broken by controlling for observables? (“Confounded” assignment or “nonignorable” assignment or “selection on unobservables” or “endogenous” assignment.)
- Often there is a component of self-selection in program evaluation.
- Broadly speaking, approaches to treament effect estimation fall into one of three situations.

1. Assume unconfoundedness of treatment, and then exploit it in estimation.

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

3. Allow self-selection on unobservables; exploit an exogenous instrumental variable.

- These slides cover (1) and (2).

• Unconfoundedness leads to many possible estimation methods, including:

1. Regression adjustment
2. Propensity score weighting
3. Matching
4. Combinations

- All approaches maintain unconfoundedness.
- Unconfoundedness is fundamentally untestable.
- In some cases there are ways to assess the plausibility of unconfoundedness or study the sensitivity of estimates.

- A second key assumption is “overlap,” which concerns the similarity of the covariate distributions for the treated and untreated subpopulations.
- If overlap is weak, probably have to redefine the population of interest in order to precisely estimate a treatment effect on some subpopulation.

2. Basic Concepts

Counterfactual Outcomes and Parameters

- For each population unit, two possible outcomes:

  $Y(0)$ is the outcome without treatment

  $Y(1)$ is the outcome with treatment

- The binary indicator is W, where $W = 1$ denotes "treatment."

- The nature of $Y(0)$ and $Y(1)$ – discrete, continuous, some mix – is unspecified.

- The gain from treatment is

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

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

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

- If we could observe the gains for a random sample we could just average the gains across the random sample.

- Problem: For each unit i, only one of $Y_i(0)$ and $Y_i(1)$ is observed.
- In effect, we have a missing data problem.
- We obtain a random sample of units by we only observe W_i and

$$Y_i = (1 - W_i)Y_i(0) + W_i Y_i(1)$$

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

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

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

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

- With heterogeneous treatment effects, τ_{ate} and τ_{att} can be very different.
- Important: τ_{ate} and τ_{att} are defined without reference to a model.
- Definitions hold whether assignment is randomized, unconfounded, or endogenous.
- Identification of τ_{ate} and τ_{att} depends on what we assume about treatment assignment.

Sampling Assumptions

- Assume independent, identically distributed observations from the underlying population.
- The data we would like to have is

$$\{[Y_i(0), Y_i(1)] : i = 1, \ldots, N\}.$$

but we only observe W_i and Y_i:

$$\begin{aligned} Y_i &= (1 - W_i)Y_i(0) + W_i Y_i(1) \\ &= Y_i(0) + W_i[Y_i(1) - Y_i(0)]. \end{aligned}$$

- 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.)
- Spillovers and general equilibrium effects are tricky to handle.

3. The Key Assumptions: Unconfoundedness and Overlap

- For each i we draw, along with (W_i, Y_i), a vector of covariates, say $\mathbf{X}_i$.
- Let $\mathbf{X}$ be the random vector with a distribution in the population.

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

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

where the symbol "⊥" means "independent of" and "|" means "conditional on." □

• Within each segment of the population determined by different values of **X**, treatment assignment is random.

- For example, the probability of being chosen for a job training program depends on average labor market earnings over the past two years, but the assignment is random given average earnings.
- Unconfoundedness is controversial. In effect, it underlies standard methods to estimating treatment effects via "kitchen sink" regressions.

- Essentially, unconfoundedness leads to a comparison-of-means after adjusting for observed covariates.
- Even if we doubt we have enough of the “right” covariates, such a comparison makes sense.
- Can show unconfoundedness is generally violated if **X** includes variables that are themselves affected by the treatment.

- In evaluating a job training program, **X** should *not* include post-training schooling; it might be affected by being assigned or not to the job training program.
- In evaluating a change in payment plans on health costs, one should not include post-intervention measures of health in **X**.
- **X** should include variables that are measured prior to treatment assignment.
- **X** should not include instrumental variables.

• A weaker version of unconfoundedness is often sufficient:

A.1′. Unconfoundedness in Conditional Mean:

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

• With unconfoundedness the quantities we need to estimate are *nonparametrically identified* under weak assumptions.

• Instrumental variables methods are either limited in what parameter they estimate or impose functional form and distributional restrictions.

- The probability of treatment as a function of **x** is known as the **propensity score**:

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

- The propensity score plays a key role in studying overlap and in obtaining certain estimates.

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

$$0 < p(\mathbf{x}) < 1. \square$$

- Each unit in the defined population has some chance of being treated and some chance of not being treated.
- Called "overlap" because it ensures common support for

$$D(\mathbf{X}|W = 1) \text{ and } D(\mathbf{X}|W = 0).$$

- *Strong Ignorability* = Unconfoundedness plus Overlap (Rosenbaum and Rubin, 1983).

- Tension between ignorability and overlap:
    - ▶ To satisfy ignorability we want **X** to contain as many (pre-treatment) covariates as possible.
    - ▶ But the richer is **X** the more likely we are to find some partitions where units are never treated or always treated.

- Later we will study the distribution of $p(\mathbf{X})$ (actually, an estimate of it) to determine overlap.
- Why? Let $R = p(\mathbf{X})$, which generally takes values in $[0, 1]$.
- We can write the conditional density of R given W as

$$g(r|w) = \frac{r^w(1-r)^{1-w}h(r)}{\rho^w(1-\rho)^{1-w}}, w \in \{0,1\}, r \in [0,1]$$

- With overlap, $0 < R < 1$, and so

$$g(r|w) = 0 \text{ if and only if } h(r) = 0.$$

- In particular, overlap implies the supports of $g(r|w = 0)$ and $g(r|w = 1)$ are the same.
- We can study the distributions of the propensity scores for the control and treated groups to ensure "sufficient" overlap.

• For τ_{att} we can get by with weaker versions of ignorability and overlap:

ATT.1. **Unconfoundedness**:

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

• Allows W_i to be correlated with the (unobserved) gain from treatment, $Y_i(1) - Y_i(0)$.

ATT.2. **Overlap**:

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

• $p(\mathbf{x}) = 0$ is allowed because τ_{att} is conditional on $W = 1$.

4. Identification of Average Treatment Effects

- We can use three ways to show the treatment effects are identified under ignorability and overlap:

1. Conditional mean functions
2. Propensity score weighting
3. Ignorability conditional on the propensity score

Identification via Conditional Means

- Define the **average treatment effect conditional on x** as

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

where

$$\mu_0(\mathbf{x}) \equiv E[Y(0)|\mathbf{X} = \mathbf{x}]$$
$$\mu_1(\mathbf{x}) \equiv E[Y(1)|\mathbf{X} = \mathbf{x}]$$

are counterfactual conditional means.

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

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

- τ_{ate} is identified if $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are identified over the support of $\mathbf{X}$.

• To see $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are identified under unconfoundedness (and overlap), note that

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

where the second equality holds by unconfoundedness.

- Define the conditional means for *observable* Y as

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

- Under overlap, $m_0(\cdot)$ and $m_1(\cdot)$ are nonparametrically identified on $\mathcal{X}$ because we assume have a random sample on $(Y, \mathbf{X}, W)$.
- In general, $m_g(\cdot) \neq \mu_g(\cdot)$.

- However, with unconfoundedness

$$m_0(\mathbf{x}) = \mu_0(\mathbf{x}), m_1(\mathbf{x}) = \mu_1(\mathbf{x})$$

- For ATT,

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

- Therefore,

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

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

Summary

- Under ATE.1 and ATE.2, τ_{ate} is identified by

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

- Under ATT.1 and ATT.2, τ_{att} is identified by

$$\begin{aligned}\tau_{att} &= E[m_1(\mathbf{X}) - m_0(\mathbf{X})|W = 1] \\ &= E(Y|W = 1) - E[m_0(\mathbf{X})|W = 1]\end{aligned}$$

Identification via Propensity Score Weighting

- Ignorability in mean, ATE.1′, implies that W and $Y(g)$ are uncorrelated conditional on $\mathbf{X}$.
- Note that $WY = WY(1)$.
- By iterated expectations,

$$E\left[\frac{W \cdot Y}{p(\mathbf{X})}\right] = E\left[\frac{E(W|\mathbf{X}) \cdot E(Y(1)|\mathbf{X})}{p(\mathbf{X})}\right] = E[Y(1)].$$

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

- Putting the two expressions together gives

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

- Can also show

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

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

- For τ_{att}, does not matter if some units have no chance of being treated: $p(\mathbf{x}) = 0$.
- In effect, this is one way to define the quantity of interest in a way that the necessary overlap assumption has a better chance of holding.

Ignorability Conditional on the Propensity Score

- A key finding of Rosenbaum and Rubin (1983) is that if ignorability holds conditional on $\mathbf{X}$ then it holds conditional only on $p(\mathbf{X})$:

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

- $p(\mathbf{X})$ acts as a sufficient statistic for randomizing treatment.

5. A Strategy for Estimating Treatment Effects

1. Carefully define the population of interest.

2. Obtain data that coheres, as closely as possible, to the population.

3. Check overlap.

4. Adjust the sample to improve overlap, if necessary.

5. Use the adjusted sample and apply a suitable average treatment effect estimator.

6. If possible, evaluate unconfoundedness.

6. Estimating the Propensity Score

- The PS plays an important role in evaluating overlap and constructing estimates of ATEs.
- Typically, $P(W = 1|\mathbf{X} = \mathbf{x})$ is modeled using a flexible logit or probit.
- Specification searches cause little harm at the initial stage: the estimated propensity score depends on the data on the covariates and treatment assignment and not on the outcome variable Y.

- Imbens and Rubin (2015) describe a stepwise procedure for adding levels, squares, and cross products of elements of $\mathbf{X}$ to logit or probit.
- Model selection methods such as LASSO can be used.
- `teffects` in Stata uses `logit`, `probit`, or `hetprobit`; you supply the predictors.
- If certain outcomes on $\mathbf{X}_i$ perfectly predict $W_i = 0$ or $W_i = 1$, $\hat{p}(\mathbf{X}_i) = 0$ or $\hat{p}(\mathbf{X}_i) = 1$, and overlap fails.

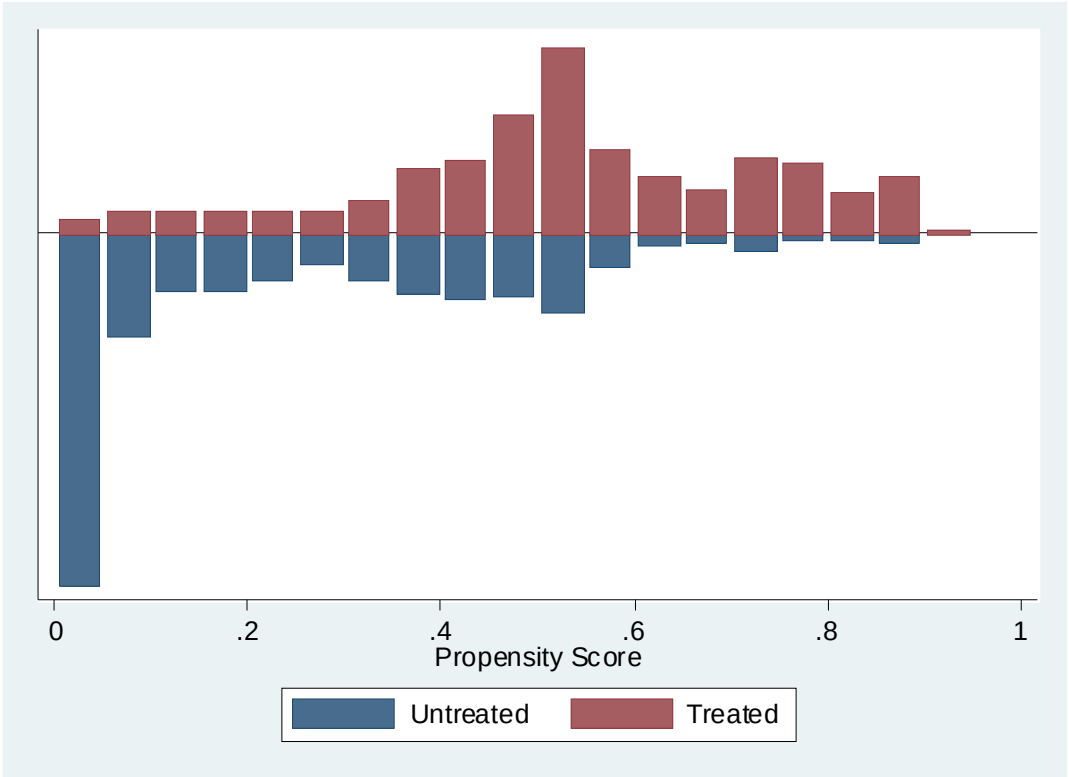
7. Assessing and Improving Overlap

- Most common approach to assessing overlap is to look at the distributions of the estimated propensity scores.
- The Stata command `psgraph` does this after using `psmatch2` (user written).
- The Stata command `teffects` uses smoothed density estimation.

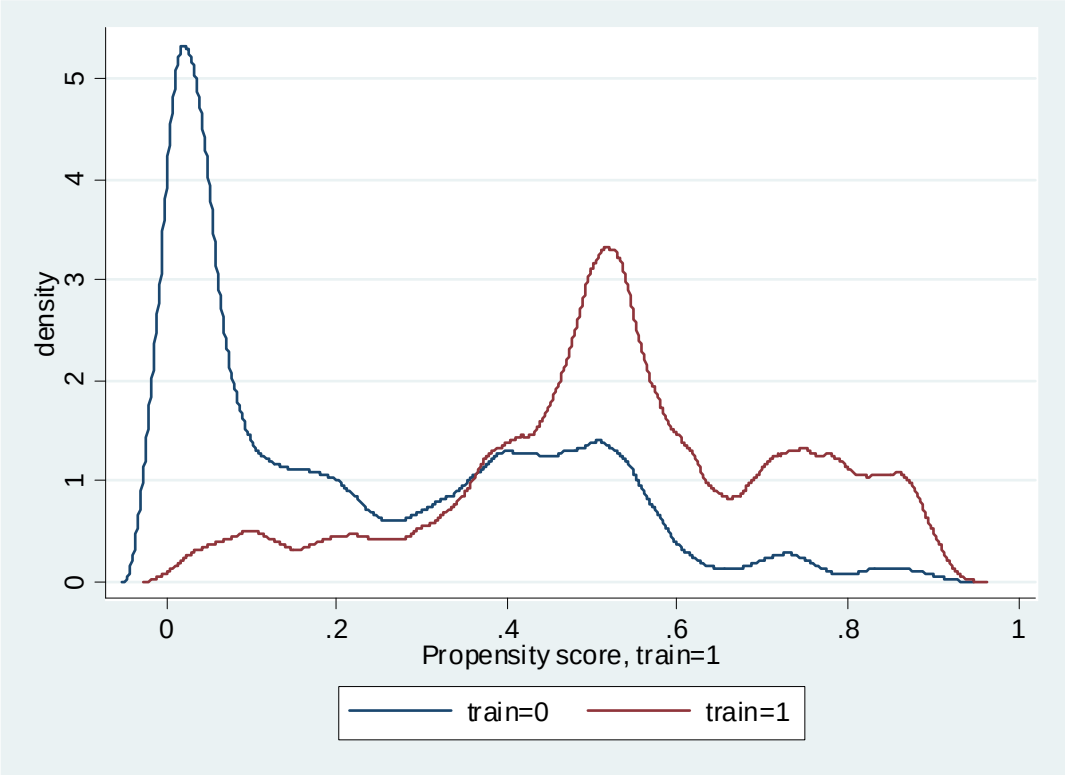
- If there are problems with overlap in the original sample, may have to redefine the population.
- Focusing on τ_{att} rather than τ_{ate} can solve much of the overlap problem because $P(W = 1|\mathbf{X}) = 0$ is allowed.

- JTRAIN98.DTA: Job training program implemented in 1997.
- Outcome is 1998 labor market earnings, *earn*98.
- The treatment variable is *train*.
- Controls are *earn*96, *unem*96, *age*, and *educ*. Include quadratics and interactions in the propensity score and regression adjustment.

```
. use jtrain98

. tab train

   =1 if in |
        job |
   training |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        754       66.73       66.73
          1 |        376       33.27      100.00
------------+-----------------------------------
      Total |      1,130      100.00

. qui logit train age c.age#c.age educ c.educ#c.educ c.age#c.age unem96 ///
      earn96 c.earn96#c.earn96 c.age#c.earn96 c.educ#c.earn96

. predict ph

. sum ph if train

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          ph |        376    .532687    .2010638   .0264422   .9134486

. sum ph if ~train

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          ph |        754   .2330367    .2222613   .0011963   .8939954
```

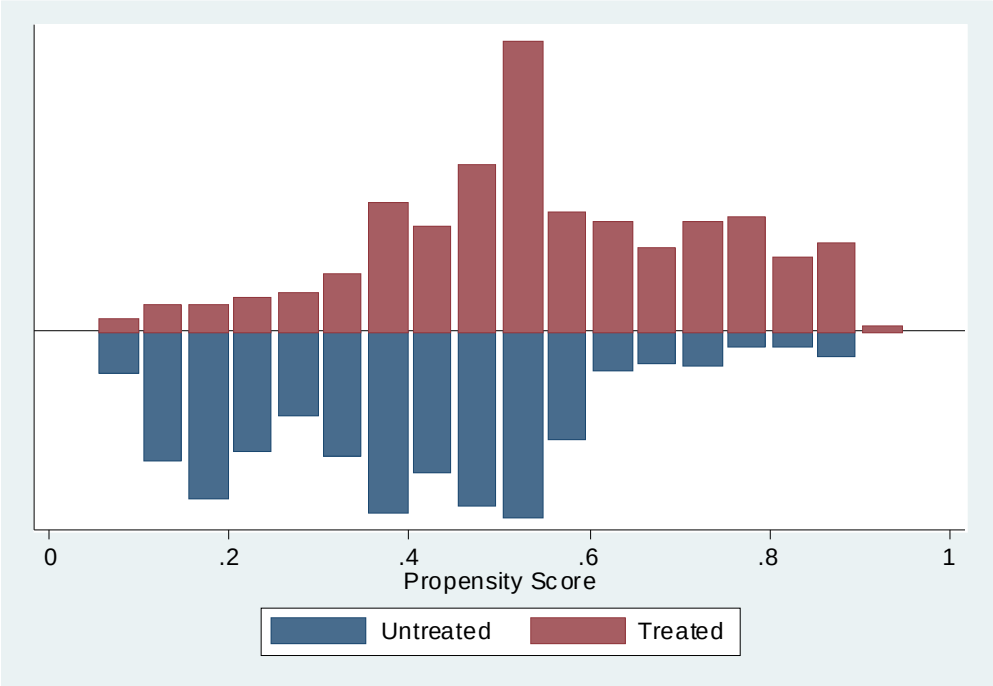

0
.2
.4
.6
.8
1
Propensity Score
Untreated
Treated

```
. qui teffects ipw (earn98) (train age c.age#c.age educ c.educ#c.educ ///
     c.age#c.educ unem96 earn96 c.earn96#c.earn96 c.age#c.earn96 c.educ#c.earn96)

. teffects overlap, ptl(1) name(ps_jtrain98_full_smooth, replace)
```



- For estimating ATE, Crump, Hotz, Imbens, and Mitnik (2009, Biometrika) propose a trimming method based on the propensity score.
- They "move the goalposts" and define a treatment effect that can be precisely estimated.
- Simple rule of thumb: drop observation i if $\hat{p}(\mathbf{X}_i) \notin [.1, .9]$.
- Generally, can lose a lot of data – including treated observations – and resulting population might not be what we want.

```
. count if ph < .1
  343

. count if ph > .9
  1

. keep if ph >= .1 & ph <= .9
(344 observations deleted)
```



• When interest centers on $\tau_{att} = E[Y(1) - Y(0)|W = 1]$, an alternative approach is to use the estimated PS to match each treated unit with a single control unit.

1. Use all data to estimate the PS.

2. Order treated units from largest to smallest PS.

3. Starting at top, match the first treated unit to the closest control. Then do the same for the next treated unit (without replacement). And so on.

• With N_1 treated units, we wind up with N_1 controls, too.

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

8. Estimating ATEs

- When we assume unconfounded treatment and overlap, there are three general approaches to estimating the treatment effects (although they can be combined):

(i) regression-based methods

(ii) propensity score methods

(iii) matching methods

- Can mix the various approaches, and often this helps.
- Matching on the propensity score is especially popular.
- Need to keep in mind that all methods produce consistent estimators under unconfoundedness and overlap.
- But they may behave quite differently when overlap is weak.

Regression Adjustment

1. 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 linear regression or as complicated as full nonparametric regression.

2. Compute fitted values for each outcome for *all* units in sample. Even though we only use the treated units to obtain $\hat{m}_1(\mathbf{x})$, we need $\hat{m}_1(\mathbf{X}_i)$ for all i.

- The regression-adjustment estimators are

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

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

- $\hat{m}_1(\mathbf{X}_i) - \hat{m}_0(\mathbf{X}_i)$ is the estimated gain from treatment for unit i.

- Alternatively, $\hat{\tau}_{att,reg}$ can be estimated as

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

- In the leading cases (linear regression, logit, Poisson) this is the same as using Y_i in place of $\hat{m}_1(\mathbf{X}_i)$.
- Linear case:

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

- $\hat{\tau}_{ate,reg}$ is simply the coefficient on W_i in the regression

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

- Use `teffects` in Stata to properly account for the sampling error in $\bar{\mathbf{X}}$.
- Demeaning the covariates before constructing the interactions "solves" the multicollinearity problem because it redefines the parameter we are trying to estimate to be the ATE.

- The 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.

- Linear regression adjustment in Stata 13+:

```
teffects ra (y x1 ... xK) (w)
teffects ra (y x1 ... xK) (w) atet
```

- The nature of Y can be exploited in nonlinear models.
- If Y is binary, use `logit` (default), `probit`, or `hetprobit`:

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

```
teffects ra (y x1 ... xK, logit) (w)
```

- For fractional Y, can use `glm`:

```
glm y i.w x1 ... xK
    i.w#c.x1 ... i.w#c.xK
    fam(bin) link(logit), robust
margins, dydx(i.w) vce(uncond)
```

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

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

```
teffects ra (y x1 ... xK, poisson) (w)
```

• 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.

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

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

• Have to estimate $\hat{m}_0(\mathbf{x})$ for values of $\mathbf{x}$ in the treated group.

- Classic study by Lalonde (1986, AER): The nonexperimental data combined the treated group from the experiment with a random sample from a different source.
- The control group was much more heterogeneous than the treatment group. Many controls had no similar treatment units.
- Very difficult to estimate τ_{ate} because $m_1(\mathbf{X}_i)$ poorly estimated for many i with $W_i = 0$.

- Better with τ_{att} because do have untreated observations similar to the control group.
- Get better results by redefining the population, either based on propensity scores or a variable such as average pre-training earnings.
- Think carefully about the population ahead of time. If high earners are not going to be eligible for job training, why include them in the analysis at all?
- The population is not immutable.

Propensity Score Weighting

- Sample analog of the earlier population moment:

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

- $\tilde{\tau}_{ate,psw}$ is not feasible because it depends on the propensity score $p(\cdot)$.
- We would not use it if we could! It is *better* to estimate the propensity score!

- Given $\hat{p}_i = \hat{p}(\mathbf{X}_i)$, a better strategy – Busso, Dinardo, and McCrary (2009, *JBES*) – is to use weights that sum to one.
- Combine into a single WLS problem:

$$\min_{\mu_0,\tau} \sum_{i=1}^{N} \left[\frac{(1-W_i)}{(1-\hat{p}_i)} + \frac{W_i}{\hat{p}_i} \right] (Y_i - \mu_0 - \tau W_i)^2.$$

- Default in Stata is logit model:

```
teffects ipw (y) (w x1 ... xK)
```

- For τ_{att}, the WLS problem is

$$\min_{\mu_0,\tau} \sum_{i=1}^{N} [(1 - W_i)\hat{p}_i/(1 - \hat{p}_i) + W_i](y_i - \mu_0 - \tau W_i)^2.$$

```
teffects ipw (y) (w x1 ... xK),
    atet
```

Combining Regression Adjustment and PS Weighting

- Question: Why use regression adjustment combined with PS weighting?
- Answer: With $\mathbf{X}$ having large dimension, still common to rely on parametric methods for regression and PS estimation. Might worry about misspecification.
- Idea: Let $m_0(\cdot, \boldsymbol{\delta}_0)$ and $m_1(\cdot, \boldsymbol{\delta}_1)$ be parametric functions for $E[Y|\mathbf{X}, W = g],\ g = 0, 1.$
- Let $p(\cdot, \boldsymbol{\gamma})$ be a parametric model for the propensity score.

1. Estimate $\boldsymbol{\gamma}$ by Bernoulli maximum likelihood and obtain the estimated propensity scores as $p(\mathbf{X}_i, \hat{\boldsymbol{\gamma}})$.

2. Use regression or a quasi-likelihood method, weighting by the inverse of the treatment probability.

- With linear means, $(\alpha_1, \boldsymbol{\beta}_1')'$ using the weighted problem

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

- For $\boldsymbol{\delta}_0$, we weight by $1/[1 - \hat{p}(\mathbf{X}_i)]$.
- ATE is estimated as before:

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

- 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.
- However, the mean and objective function must be properly chosen [Wooldridge (2007, Journal of Econometrics)].

- *Y* continuous, negative and positive values: linear mean, least squares objective function.

```
teffects ipwra (y x1 ... xK)
    (w x1 ... xK)
```

- *Y* binary or fractional: logit mean (not probit!), Bernoulli quasi-log likelihood.

```
teffects ipwra (y x1 ... xK, logit)
    (w x1 ... xK)
```

- Currently only works with *y* binary.

- $Y \geq 0$ (count, continuous, or corners at zero): exponential mean, Poisson QLL.

```
teffects ipwra (y x1 ... xK,
    poisson) (w x1 ... xK)
```

- Always include a constant in the models for $E(Y|W, \mathbf{X})$!

Matching on Covariates

- Matching estimators impute a value on the counterfactual outcome for each unit.
- For a unit i in the control group, we observe $Y_i(0)$, but we need to impute $Y_i(1)$.
- For each unit i in the treatment group, we observe $Y_i(1)$ but need to impute $Y_i(0)$.

- For τ_{ate}, matching estimators take the general form

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

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

- For τ_{att},

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

where this form uses the fact that $W_i Y_i = W_i Y_i(1)$ [we never need to impute $Y_i(1)$ for the treated subsample.]

- Abadie and Imbens (2006, Econometrica) consider several approaches.
- Simplest is to find a single match for each observation.
- If i is a treated observation ($W_i = 1$) then $\hat{Y}_i(1) = Y_i$ and $\hat{Y}_i(0) = Y_h$ for h such that $W_h = 0$ and unit h is "closest" to unit i based on some distance in the covariates.

- Similarly if $W_i = 0$: $\hat{Y}_i(0) = Y_i$ and $\hat{Y}_i(1) = Y_h$ for h such that $W_h = 1$ and unit h is "closest" to unit i.
- Valid inference with `teffects`.

```
teffects nnmatch (y x1 ... xK) (w)
```

- The default is a 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.
- Can use an average of the *M* nearest neighbors.
- Can combine covariate matching with regression adjustment (bias reduction).

```
teffects nnmatch (y x1 ... xK) (w),
    bias(x1 ... xK)
```

Matching on the Propensity Score

- Matching on the estimated PS is computationally much easier because $\hat{p}_i \in (0, 1)$.
- Works because unconfoundedness holds conditional on $p(\mathbf{X}_i)$.
- `teffects` computes proper standard errors.
- Default is a logit PS model.

```
teffects psmatch (y) (w x1 ... xK)
teffects psmatch (y) (w x1 ... xK) atet
```

Job Training Application

Full Sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\bar{Y}_1 - \bar{Y}_0$	RA	IPW	IPWLRA	IPWERA	CM	BACM	PSM
ATE	−2.05 (0.47)	2.23 (0.79)	1.31 (0.54)	2.07 (0.50)	2.01 (0.59)	1.34 (0.64)	1.71 (0.64)	1.65 (0.50)
ATT	−2.05 (0.47)	1.72 (.49)	1.68 (0.53)	1.69 (0.49)	1.69 (0.49)	1.72 (0.72)	1.87 (0.62)	2.39 (0.63)
N	1,130	1,130	1,130	1,130	1,130	1,130	1,130	1,130

RA = regression adjustment; IPW = inverse probability weighting; IPWLRA = inverse probability weighting, linear RA; IPWERA = inverse probability weighting, exponential RA; CM = covariate matching; BACM = bias adjusted covariate matching; PSM = propensity score matching

Trimmed Sample

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\bar{Y}_1 - \bar{Y}_0$	RA	IPW	IPWLRA	IPWERA	CM	BACM	PSM
ATE	0.74	1.69	1.79	1.70	1.68	1.69	1.76	1.89
	(0.50)	(0.51)	(0.56)	(0.51)	(0.51)	(0.61)	(0.61)	(0.54)
ATT	0.74	1.66	1.57	1.58	1.62	1.79	1.85	2.02
	(0.50)	(0.51)	(0.55)	(0.50)	(0.50)	(0.68)	(0.68)	(0.60)
N	786	786	786	786	786	786	1,130	786

9. Introduction to Regression Discontinuity

- Long history dating back to 1960s in Psychology (Thistlewait and Cook, 1960). Key work in econometrics by Van der Klaauw (2001), Hahn, Todd, and Van der Klaauw (2001). Application by Lee (2008) to election outcomes.
- Regression discontinuity (RD) designs exploit discontinuities in policy assignment.
- For example, there might be an age threshold at which one becomes eligible for a buyout plan, or an income threshold at which one becomes eligible for financial aid.

- General idea: Assume that units just on different sides of the discontinuity are similar. Their treatment status differs because of the institutional setup, and differences in outcomes can be attributed to the different treatment status.
- We consider the sharp design – where assignment follows a deterministic rule – and the fuzzy design, where the probability of being treated is discontinuous at a known point.

10. The Sharp RD Design

- Y_{i0} and Y_{i1} are the potential outcomes.
- The treatment status is W_i.
- A single covariate, X_i, determines treatment status – the "running variable" or "forcing variable."
- In the sharp regression discontinuity (SRD) design case,

$$W_i = 1[X_i \geq c].$$

- Along with X_i and W_i, we observe

$$Y_i = (1 - W_i)Y_{i0} + W_i Y_{i1}.$$

- Define $\mu_g(x) = E(Y_g|X = x), g = 0, 1.$
- Maintain the assumption that $\mu_g(\cdot),\ g = 0, 1$, are both continuous at c.
- The value of c is usually pretty arbitrary, so we are essentially assuming $\mu_g(\cdot)$ is continuous on its domain $\mathcal{X}$.
- Ideally, we might hope to estimate the ATE over the entire population:

$$\tau_{ate} = E[\mu_1(X) - \mu_0(X)]$$

• Because W is a deterministic function of X, unconfoundedness of treatment necessarily holds:

$$E(Y_g|X,W) = E(Y_g|X), g = 0,1.$$

• On the other hand, the overlap assumption necessarily fails. With $p(x) = P(W = 1|X = x)$,

$$p(x) = 0, x < c$$
$$p(x) = 1, x \geq c.$$

- Clearly we cannot use propensity score weighting or matching.
- Technically, we can use regression adjustment with parametric regression functions, but we must rely on extreme forms of extrapolation.
- Current theory/practice recognizes that without parametric models we cannot estimate $\mu_0(x)$ for $x \geq c$ and $\mu_1(x)$ for $x < c$.
- Therefore, τ_{ate} is *nonparametrically unidentified*.

• Consequently, the focus in the SRD setting is usually on a very specific parameter that is generally identified, the ATE at $x = c$:

$$\tau_c \equiv E(Y_1 - Y_0|X = c) = \mu_1(c) - \mu_0(c).$$

• For the constant treatment effect case, $\tau_c = \tau$. But in general we can only estimate the ATE for those at the margin of receiving the treatment.

• Even in the SRD case there are issues of external validity.

- τ_c is of interest primarily because it is *nonparametrically identified* by the SRD design.
- How? Write

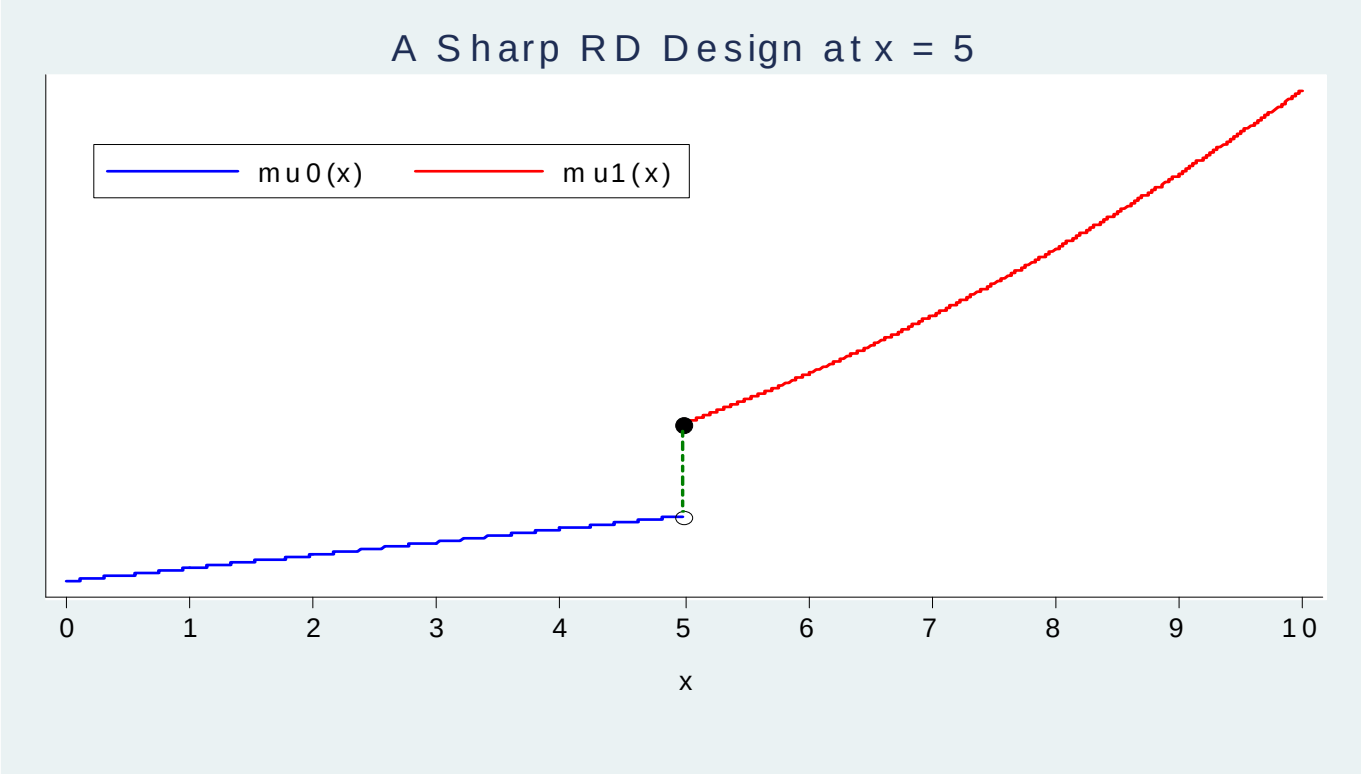
$$Y = 1[X < c]Y_0 + 1[X \geq c]Y_1,$$

and so

$$E(Y|X) = 1[X < c]\mu_0(X) + 1[X \geq c]\mu_1(X)$$

- Define $m(x) = E(Y|X = x)$ as the function we can estimate.

$$m(x) = 1[x < c]\mu_0(x) + 1[x \geq c]\mu_1(x)$$



- Using continuity of $\mu_0(\cdot)$ and $\mu_1(\cdot)$ at c,

$$m^-(c) \equiv \lim_{x \uparrow c} m(x) = \lim_{x \uparrow c} \mu_0(x) = \mu_0(c)$$

$$m^+(c) \equiv \lim_{x \downarrow c} m(x) = \lim_{x \downarrow c} \mu_1(x) = \mu_1(c)$$

and so

$$\tau_c = m^+(c) - m^-(c)$$

- Several approaches have been proposed for estimating $m^-(c)$ and $m^+(c)$.
- $\hat{m}^-(c)$ uses only data with $X_i < c$; $\hat{m}^+(c)$ uses data with $X_i \geq c$.
- First suppose $E(Y_g|X = x)$ are linear everywhere.
- Define $\mu_{0c} = \mu_0(c)$ and $\mu_{1c} = \mu_1(c)$, so that $\tau_c = \mu_{1c} - \mu_{0c}$.

- Write

$$Y_0 = \mu_{0c} + \beta_0(X - c) + U_0$$
$$Y_1 = \mu_{1c} + \beta_1(X - c) + U_1$$

- Plugging in and rearranging gives

$$Y = \mu_{0c} + \tau_c W + \beta_0(X - c) + \delta W \cdot (X - c) + U,$$

where $U = U_0 + W(U_1 - U_0)$.

- The estimate of τ_c is just the jump in the linear function at $x = c$.
- Different slopes are allowed on the different sides of the jumps.
- We could use the entire data set to run the regression

$$Y_i \text{ on } 1,\ W_i,\ (X_i - c),\ W_i \cdot (X_i - c)$$

and obtain $\hat{\tau}_c$ as the coefficient on W_i.

- Unlike when trying to estimate τ_{ate}, we center X_i about c rather than $\bar{X}$.

• Rather rely on linearity over the entire range of x, instead use **local linear regression**.

• Choose a "small" value $h > 0$ – the bandwidth – and use only the data satisfying $c - h < X_i < c + h$.

• This gives us a "local" method because it ignores data where x_i is sufficiently far from c.

- Equivalently, estimate two separate regressions:

$$Y_i \text{ on } 1, (X_i - c) \text{ for } c - h < X_i < c$$

$$Y_i \text{ on } 1, (X_i - c) \text{ for } c \leq X_i < c + h$$

- Then

$$\hat{\tau}_c = \hat{\mu}_{1c} - \hat{\mu}_{0c},$$

the difference in the two intercepts.

• Can see how sensitive estimates are to h. Tradeoff between bias and variance:

(i) As h decreases (we use less data), the bias shrinks but the variance increases.

(ii) As h increases (we use more data), the bias (generally) grows but the variance decreases.

• One can use higher order polynomials, but these seem to not work as well.

• Inference is standard when h is viewed as fixed: use a heteroskedasticity-robust t statistic.

- Imbens and Lemieux (2008, *Journal of Econometrics*) show that if h satisfies

$$h = o(N^{-a}),\ 1/5 < a < 2/5$$

then usual inference is still valid. (This choice ensures the estimator has no asymptotic bias.)

- In practice, this rate result is not very helpful.

- Can add regressors even though, under the pure sharp design, there is no need to for consistency.
- If the other regressors are $\mathbf{R}_i$, just run

$$Y_i \text{ on } 1,\ W_i,\ (X_i - c),\ W_i \cdot (X_i - c),\ \mathbf{R}_i$$

again only using data $c - h < X_i < c + h$.

- Using extra regressors is likely to have more of an impact when h is large; it might help reduce the bias from arising from the deterioration of the linear approximation.
- Adding $\mathbf{R}_i$ can improve the precision of $\hat{\tau}_c$.

- For response variables with special characteristics, can use local versions of other models, such as logistic and exponential functions.
- Does not seem to be done much in practice.
- In the linear regression case, Imbens and Lemieux summarize data-dependent methods for choosing the bandwidth, h.
- Should try different rules to check sensitivity of results for τ_c.

- Imbens and Kalyanaraman (2012, *Review of Economic Studies*) explicitly look at minimizing

$$E\{[\hat{\mu}_0(c) - \mu_0(c)]^2 + [\hat{\mu}_1(c) - \mu_1(c)]^2\},$$

a mean squared error for the two regression functions at the jump point.

- Not the same as the MSE for the actual estimand, which would be

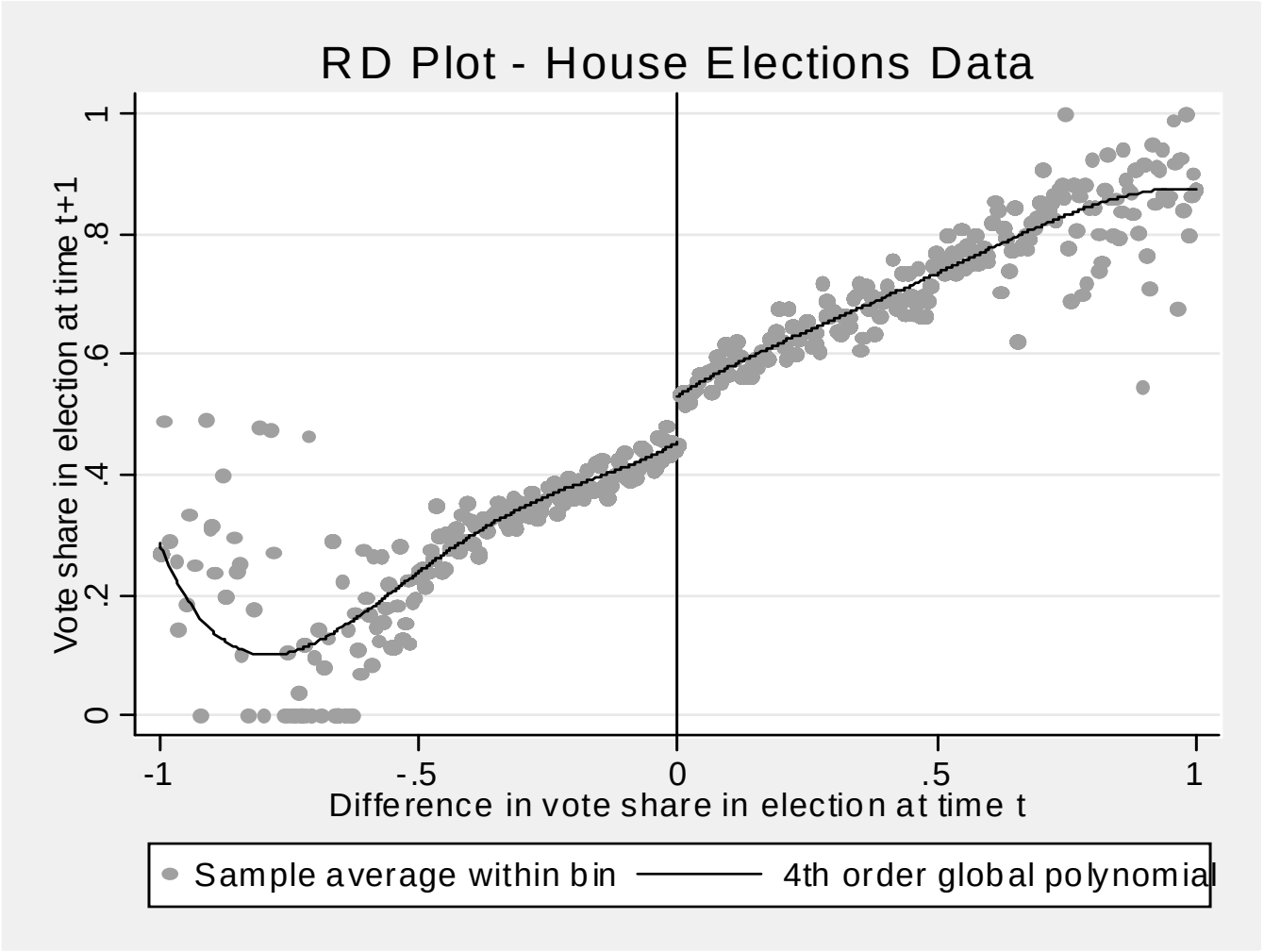
$$E[(\hat{\tau}_c - \tau_c)^2] = E\{([\hat{\mu}_1(c) - \hat{\mu}_0(c)] - [\mu_1(c) - \mu_0(c)])^2\}.$$

• Optimal bandwidth choice depends on second derivatives of the regression functions at $x = c$, the density of X_i at $x = c$, the conditional variances, and the kernel (often the uniform) used in local linear regression.

• But IK have shown how to make the choice of h purely data-dependent.

• Other authors have suggested other data-dependent bandwidth choices. Calonico, Cattaneo, and Titiunik (2014a, Econometrica) use a different choice.

- CCT (2014b, Stata Journal) allow three different possibilities, including IK.
- CCT also propose better behaved confidence intervals for bias-adjusted estimates of τ_c.
- Stata command is `rdrobust` (and also `rdplot`, `rdbwselect`).
- Data from Lee (2008), U.S. House of Representatives. Effects of incumbency on election outcomes.


RD Plot - House Elections Data
Vote share in election at time t+1
0
.2
.4
.6
.8
1
-1
-.5
0
.5
1
Difference in vote share in election at time t
Sample average within bin
4th order global polynomial

```
. rdrobust vote margin, c(0) kernel(uniform) bwselect(ik)
Preparing data.
Computing bandwidth selectors.
Computing variance-covariance matrix.
Computing RD estimates.
Estimation completed.

Sharp RD estimates using local polynomial regression.

         Cutoff c = 0 | Left of c  Right of c             Number of obs =
----------------------+----------------------             NN matches    =
        Number of obs |      1013        1030             BW type       =
 Order loc. poly. (p) |         1           1             Kernel type   =    Uniform
       Order bias (q) |         2           2
    BW loc. poly. (h) |     0.178       0.178
          BW bias (b) |     0.222       0.222
            rho (h/b) |     0.802       0.802

Outcome: vote. Running variable: margin.
------------------------------------------------------------------------------
               Method |    Coef.   Std. Err.      z      P>|z|     [95% Conf. Interval
----------------------+-------------------------------------------------------
         Conventional |   .08096    .00931    8.6918    0.000       .062701    .099212
               Robust |      -         -      5.6249    0.000       .047482    .098267
------------------------------------------------------------------------------
```

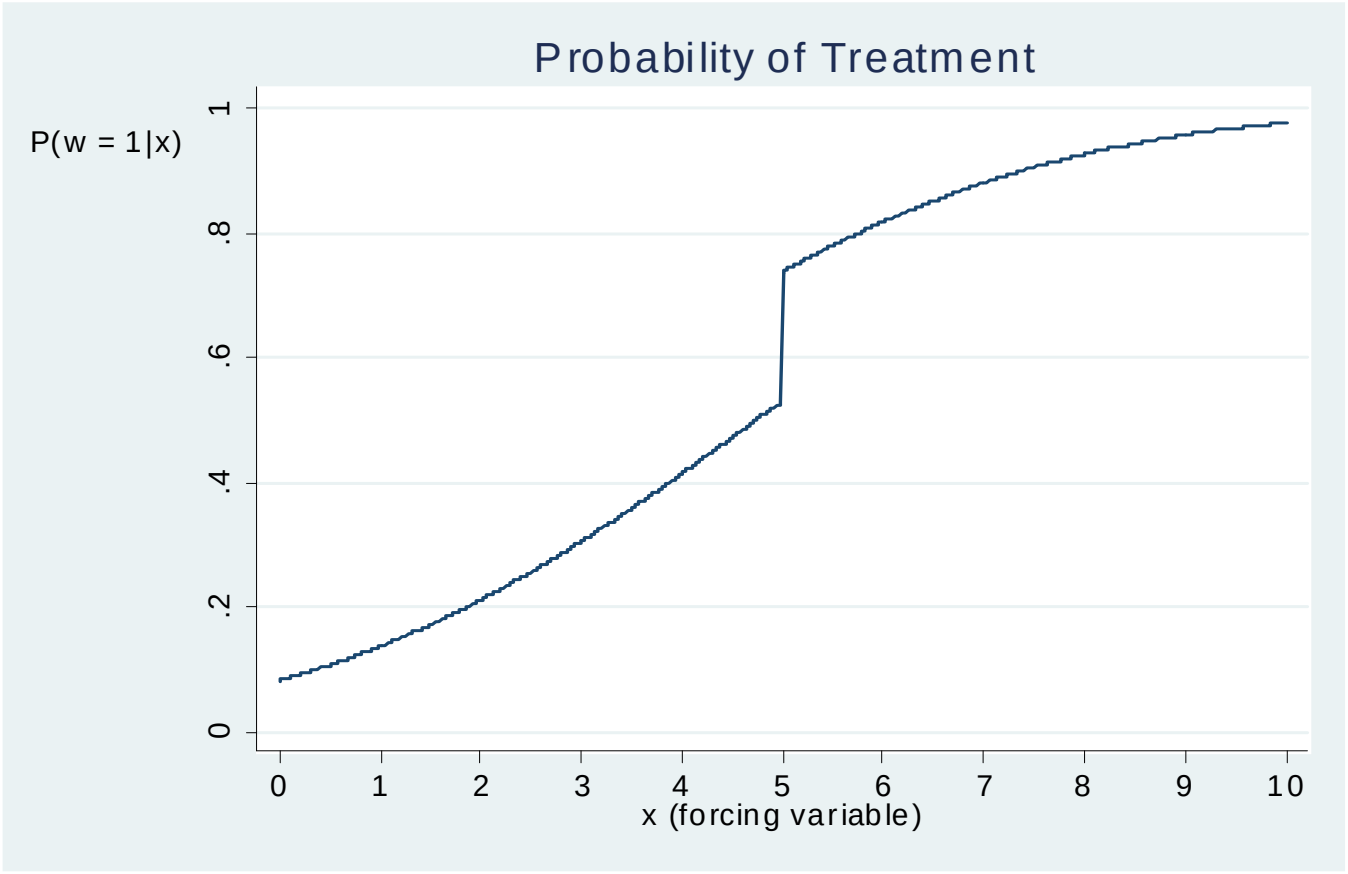
```
. rdrobust vote margin, c(0) kernel(uniform) bwselect(cct)
Preparing data.
Computing bandwidth selectors.
Computing variance-covariance matrix.
Computing RD estimates.
Estimation completed.

Sharp RD estimates using local polynomial regression.

         Cutoff c = 0 | Left of c  Right of c             Number of obs =
----------------------+-----------------------           NN matches    =
        Number of obs |       698         729            BW type       =
 Order loc. poly. (p) |         1           1            Kernel type   =    Uniform
       Order bias (q) |         2           2
    BW loc. poly. (h) |     0.120       0.120
          BW bias (b) |     0.228       0.228
            rho (h/b) |     0.525       0.525

Outcome: vote. Running variable: margin.
------------------------------------------------------------------------------
               Method |    Coef.   Std. Err.      z      P>|z|     [95% Conf. Interval
----------------------+-------------------------------------------------------
         Conventional |   .06622    .01126    5.8822    0.000       .044155    .088285
               Robust |      -         -      4.8827    0.000       .037744    .088366
------------------------------------------------------------------------------
```

11. The Fuzzy RD Design

- In the FRD case, the *probability* of treatment changes discontinuously at $x = c$.
- Define the propensity score as

$$P(W = 1|X = x) \equiv F(x).$$

- In addition to assuming $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are continuous at c, the key assumption for the FRD design is that $F(\cdot)$ is *discontinuous* at c: there is a discrete jump in the probability of treatment at the cutoff.


Probability of Treatment
P(w = 1|x)
0
.2
.4
.6
.8
1
0
1
2
3
4
5
6
7
8
9
10
x (forcing variable)

- To identify τ_c, we assume that $Y_1 - Y_0$ is independent of W, conditional on X.
- Allows treatment, W, to be correlated with Y_0 (after conditioning on X) but not with the unobserved gain from treatment.
- The unconfoundedness assumption for estimating ATT is that W is unconfounded with respect to Y_0, not $Y_1 - Y_0$.
- The unconfoundedness assumption for estimating ATE is that W is unconfounded with respect to Y_0 and $Y_1 - Y_0$.

• Again write $Y = Y_0 + W(Y_1 - Y_0)$ and use conditional independence of W and $Y_1 - Y_0$:

$$\begin{aligned} E(Y|X = x) &= E(Y_0|X = x) + E[W(Y_1 - Y_0)|X = x] \\ &= E(Y_0|X = x) + E(W|X = x)E(Y_1 - Y_0|X = x) \\ &\equiv \mu_0(x) + F(x)\tau(x). \end{aligned}$$

• As before, take limits from the right and left and use continuity of $\mu_0(\cdot)$ and $\tau(\cdot)$ at c:

$$\begin{aligned} m^+(c) &= \mu_0(c) + F^+(c)\tau_c \\ m^-(c) &= \mu_0(c) + F^-(c)\tau_c \end{aligned}$$

• It follows that, if $F^{+}(c) \neq F^{-}(c)$ – there is a jump in $F(\cdot)$ at $x = c$ – then

$$\tau_c = \frac{[m^{+}(c) - m^{-}(c)]}{[F^{+}(c) - F^{-}(c)]}.$$

• So, to identify τ_c in the FRD case, we assume

(i) $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are continuous at c.

(ii) $F^{+}(c) \neq F^{-}(c)$.

(iii) $(Y_1 - Y_0)$ and W are independent conditional on X.

- Imbens and Lemieux suggest estimating $m^-(c)$, $m^+(c)$, $F^-(c)$, and $F^+(c)$ by local linear regression.
- In other words, use

$$\hat{\tau}_c = \frac{[\hat{m}^+(c) - \hat{m}^-(c)]}{[\hat{F}^+(c) - \hat{F}^-(c)]},$$

where $\hat{m}^+(c) = \hat{\alpha}_{1c}$, $\hat{m}^-(c) = \hat{\alpha}_{0c}$, $\hat{F}^+(c) = \hat{\theta}_{1c}$, and $\hat{F}^-(c) = \hat{\theta}_{0c}$ are the intercepts from four local linear regressions.

• For example, $\hat{\alpha}_{1c}$ is from Y_i on 1, $(X_i - c)$, $c \leq X_i < c + h$ and $\hat{\theta}_{1c}$ is from W_i on 1, $(X_i - c)$, $c \leq X_i < c + h$.

• Conveniently, Hahn, Todd, and Van der Klaauw (2001) show that the IV estimator of

$$Y_i = \alpha_{0c} + \tau_c W_i + \beta_0 (X_i - c) + \gamma 1[X_i \geq c] \cdot (X_i - c) + R_i$$

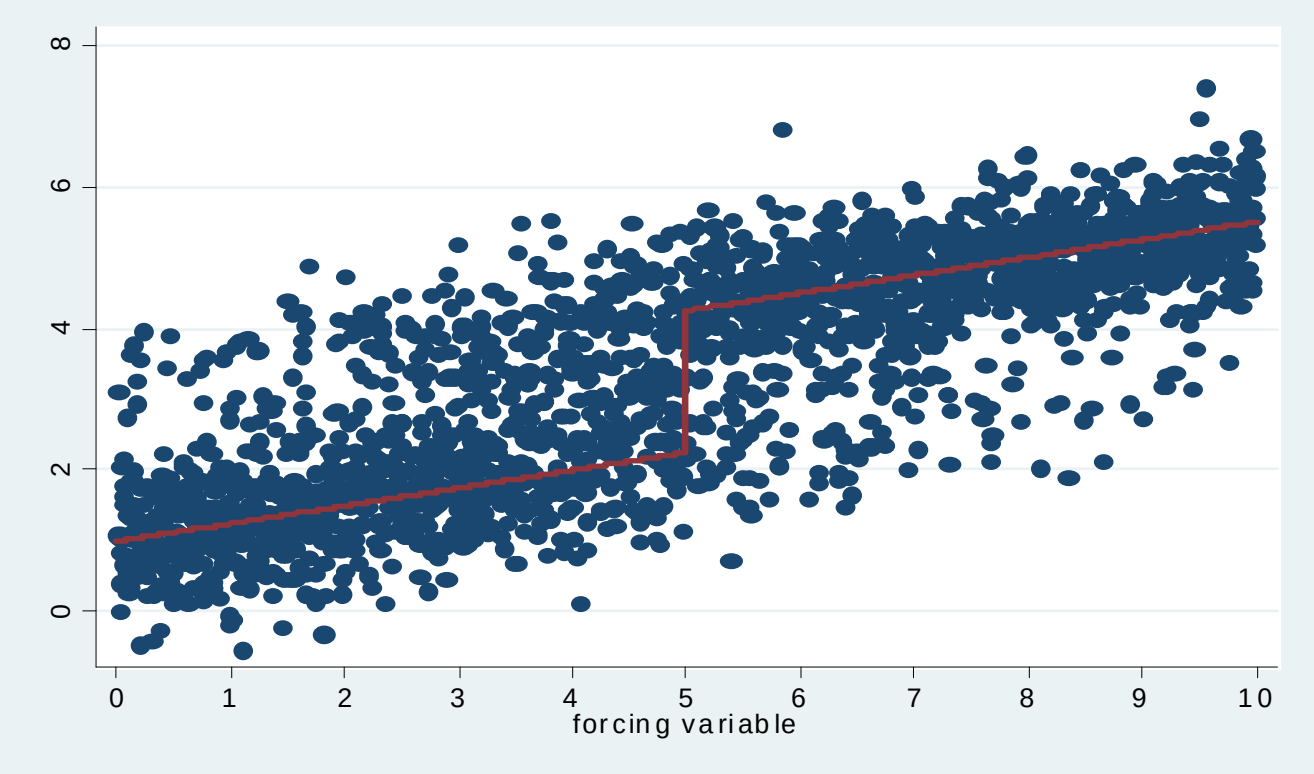
using $Z_i \equiv 1[X_i \geq c]$ as the IV for W_i produces $\hat{\tau}_c$ as the coefficient on W_i.

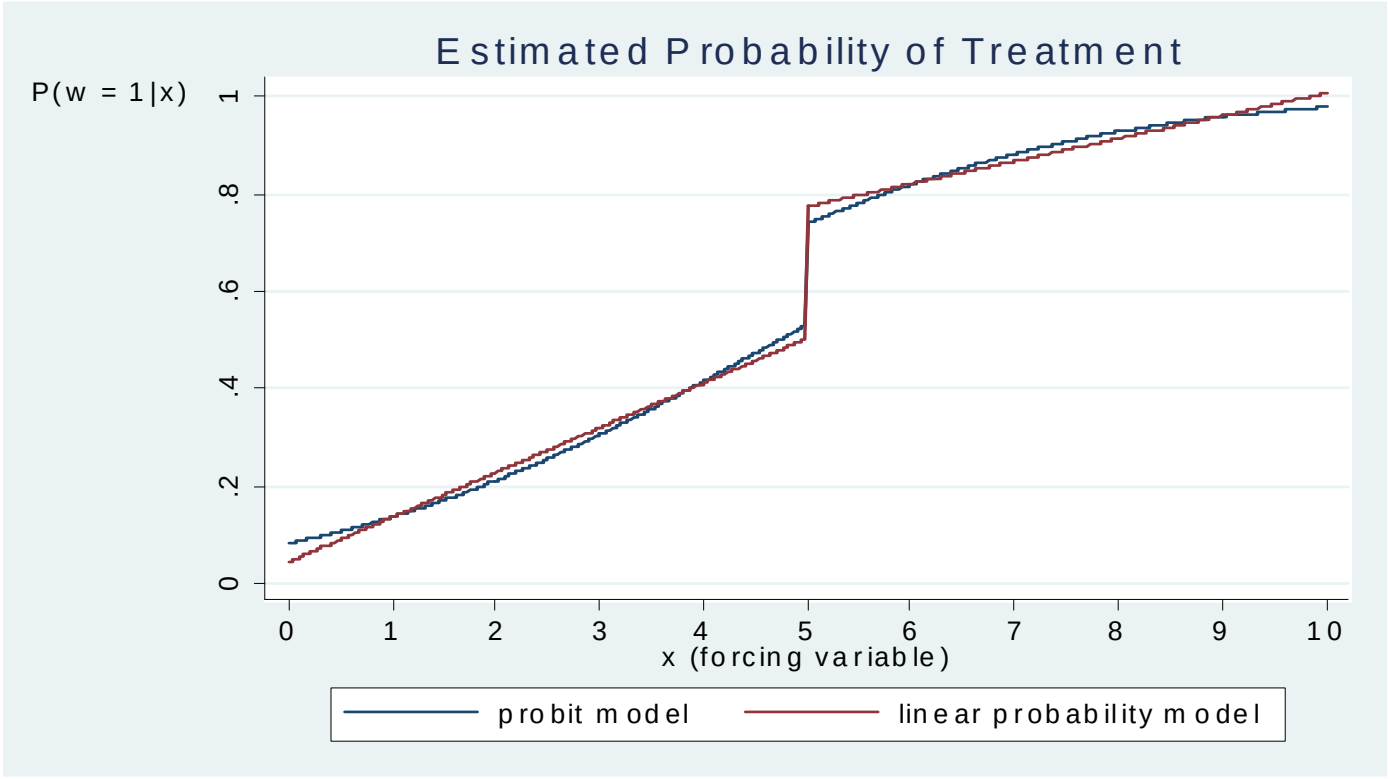
• This is an algebraic equivalence.

- If h is fixed, or is decreasing "fast enough," can use the usual heteroskedasticity-robust IV standard error.
- Now have two choices of bandwidths because need to estimate $E(W|X = x)$ for $x < c$ and $x \geq c$ in addition to $E(Y|X = x)$ for $x < c$ and $x \geq c$.
- Could choose one based on $E(Y|X = x)$ and use it for both, or choose them separately using, say, Imbens and Kalraynaram (2012).

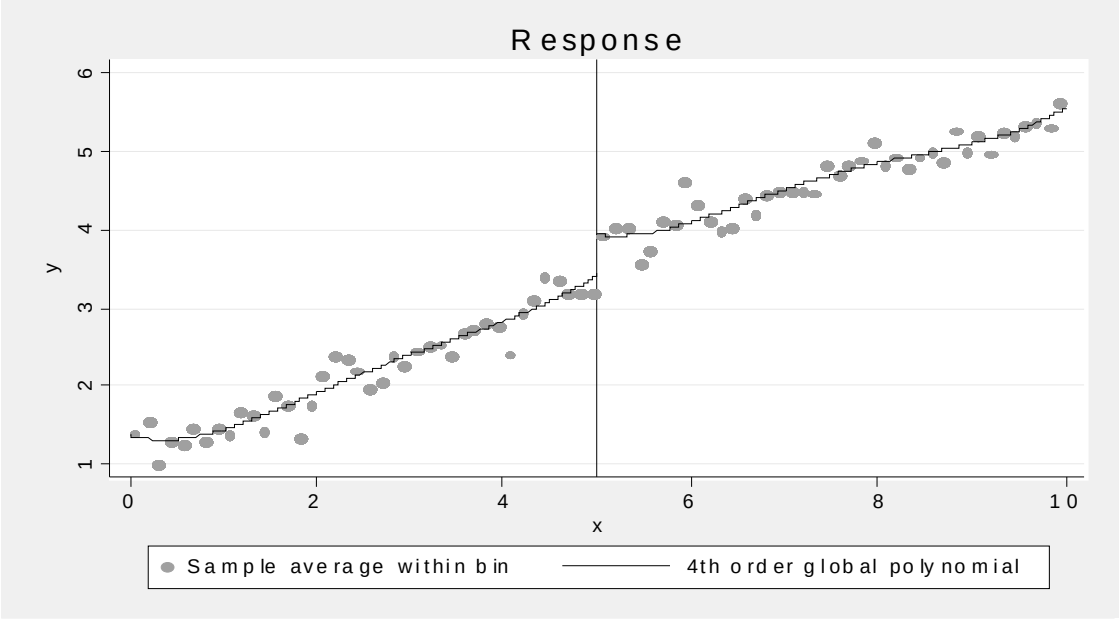
EXAMPLE: **Generated Data**. Forcing variable is $X \sim Uniform[0, 10]$, the rule that predicts treatment is $Z = 1[X \geq 5]$, and W is the actual treatment indicator. The outcome variable is Y.

```
. tab w z

     =1 if |     =1 if x >= 5
   treated |         0          1 |     Total
-----------+----------------------+----------
         0 |       727        111 |       838
         1 |       273        889 |     1,162
-----------+----------------------+----------
     Total |     1,000      1,000 |     2,000
```


0
2
4
6
8
0
1
2
3
4
5
6
7
8
9
10
forcing variable


Estimated Probability of Treatment
P(w = 1|x)
0
.2
.4
.6
.8
1
0
1
2
3
4
5
6
7
8
9
10
x (forcing variable)
probit model
linear probability model


Treatment
w
0
.2
.4
.6
.8
1
0
2
4
6
8
10
x
Sample average within bin
4th order global polynomial


Response
y
1
2
3
4
5
6
0
2
4
6
8
10
x
Sample average within bin
4th order global polynomial

```
. rdrobust y x, c(5) fuzzy(w) kernel(uniform) bwselect(ik)
Preparing data.
Computing bandwidth selectors.
Computing variance-covariance matrix.
Computing RD estimates.
Estimation completed.

Sharp RD estimates using local polynomial regression.

         Cutoff c = 5 | Left of c  Right of c            Number of obs =
----------------------+----------------------            NN matches    =
        Number of obs |       457         457            BW type       =
 Order loc. poly. (p) |         1           1            Kernel type   =    Uniform
       Order bias (q) |         2           2
    BW loc. poly. (h) |     2.284       2.284
          BW bias (b) |     2.453       2.453
            rho (h/b) |     0.931       0.931
```

Structural Estimates. Outcome: y. Running variable: x. Instrument: w.

Method	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conventional	1.902	.33225	5.7248	0.000	1.25084	2.55322
Robust	-	-	3.1965	0.001	.592048	2.46885

First-Stage Estimates. Outcome: w. Running variable: x.

Method	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conventional	.25409	.06018	4.2224	0.000	.136144	.37203
Robust	-	-	3.8253	0.000	.162191	.503032

```
. rdrobust y x, c(5) fuzzy(w) kernel(uniform) bwselect(cct)
Preparing data.
Computing bandwidth selectors.
Computing variance-covariance matrix.
Computing RD estimates.
Estimation completed.

Sharp RD estimates using local polynomial regression.

         Cutoff c = 5 | Left of c  Right of c            Number of obs =
----------------------+----------------------            NN matches    =
        Number of obs |       349         349            BW type       =
 Order loc. poly. (p) |         1           1            Kernel type   =    Uniform
       Order bias (q) |         2           2
    BW loc. poly. (h) |     1.747       1.747
          BW bias (b) |     2.975       2.975
            rho (h/b) |     0.587       0.587
```

Structural Estimates. Outcome: y. Running variable: x. Instrument: w.

Method	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conventional	1.6887	.33542	5.0345	0.000	1.03126	2.34607
Robust	-	-	3.9948	0.000	.809434	2.36874

First-Stage Estimates. Outcome: w. Running variable: x.

Method	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Conventional	.29458	.06989	4.2152	0.000	.157606	.431551
Robust	-	-	3.6415	0.000	.139123	.463432

```
. * IV estimation "by hand" with bw = 1:

. ivregress 2sls y x zx_5 (w = z) if x > 4 & x < 6, robust

Instrumental variables (2SLS) regression          Number of obs =     400
                                                  F(  3,   396) =   62.45
                                                  Prob > F      =  0.0000
                                                  R-squared     =  0.6377
                                                  Root MSE      =  .73069

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           w |   1.241897   .5772794     2.15   0.032     .1069813    2.376812
           x |   .5988192   .1930259     3.10   0.002     .2193356    .9783028
        zx_5 |  -.2123672   .2431103    -0.87   0.383    -.6903155    .2655811
       _cons |  -.1820881   .7057876    -0.26   0.797    -1.569647    1.205471
------------------------------------------------------------------------------
Instrumented:  w
Instruments:   x zx_5 z
------------------------------------------------------------------------------
```

```
. * IV estimation "by hand" with bw = 2:

. ivreg y x zx_5 (w = z) if x > 3 & x < 7, robust

Instrumental variables (2SLS) regression             Number of obs =     800
                                                     F(  3,   796) =  351.50
                                                     Prob > F      =  0.0000
                                                     R-squared     =  0.7662
                                                     Root MSE      =  .61919

------------------------------------------------------------------------------
             |               Robust
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           w |   1.775465   .3267695     5.43   0.000     1.134033    2.416897
           x |   .3471895   .0726118     4.78   0.000     .2046563    .4897226
        zx_5 |  -.0991082   .0772654    -1.28   0.200    -.2507762    .0525599
       _cons |   .7060606   .1912344     3.69   0.000     .3306773    1.081444
------------------------------------------------------------------------------
Instrumented:  w
Instruments:   x zx_5 z
------------------------------------------------------------------------------
```

12. Unconfoundedness versus FRD

- In the FRD case, overlap can hold (although it might be weak in practice). If overlap holds, we can compare regression adjustment to the FRD methods of the previous section.
- Useful to return to the linear formulation:

$$Y = \eta_c + \tau_c W + \beta_0(X - c) + \delta W \cdot (X - c) + U_0 + W(U_1 - U_0).$$

- The unconfoundedness assumption for estimating ATE is

$$(U_0, U_1) \perp W \mid X,$$

and then

$$\begin{aligned} E[U_0 + W(U_1 - U_0)|W,X] &= E(U_0|W,X) + WE[(U_1 - U_0)|W,X] \\ &= 0. \end{aligned}$$

- OLS (or local regression) would consistently estimate τ_c.
- If we believe full unconfoundedness and the linear functional form, we can use all of the data and average across X_i to estimate τ_{ate}.
- Using regression adjustment, the estimator of τ_c is

$$\tilde{\tau}_c = \tilde{m}_1(c) - \tilde{m}_0(c),$$

where $\tilde{m}_1(x)$ is estimated using the $W_i = 1$ observations and $\tilde{m}_0(x)$ is estimated using the $W_i = 0$ observations.

• If we assume the less restrictive version of unconfoundedness, that is,

$$U_1 - U_0 \perp W \mid X,$$

but allow U_0 to be correlated with W – then the OLS estimator is inconsistent.

• But the estimator that exploits the fuzzy RD design,

$$\hat{\tau}_c = \frac{[\hat{m}^+(c) - \hat{m}^-(c)]}{[\hat{F}^+(c) - \hat{F}^-(c)]},$$

is consistent.

- The FRD estimator $\hat{\tau}_c$ exploits the jumps in the means and treatment probabilities at $x = c$.
- The "+" quantities use data only with $X_i \geq c$ and the "–" quantities use data only with $X_i < c$.

- Another benefit of $\hat{\tau}_c$ is that it is consistent for the ATE for compliers at $x = c$ *without* unconfoundedness, provided we add a monotonicity assumption.
- Let $W(a)$ denote treatment status if the cutoff point were a (rather than c), and think of this as a function of potential cutoff points at least over some interval that includes c.
- The monotonicity assumption is that $W(\cdot)$ is nonincreasing at $a = c$.

- Suppose that the cutoff is determined by age: those with $X \geq c$ are eligible.
- Now suppose the eligibility age is lowered to $c - 1$.
- The monotonicity assumption is (a local version of)

$$W(c-1) \geq W(c),$$

which rules out the possibility that a person would participate if eligible at age c, $W(c) = 1$, but would refuse to participate if the eligibility age were lowered , $W(c-1) = 0$.

13. Additional Analyses

- As a “falsification” check, can do use plots for other covariates that should not be affected by the threshold.
- Can also do RD estimation for other covariates and hope to find no effect at the threshold.

- A histogram of the forcing variable can be used to verify it is not being manipulated around the threshold. McCrary (2008, Journal of Econometrics) proposes a formal test.
- A family might be able to manipulate its income to stay below a threshold. Much less likely is manipulation of something like a test score or voting outcome.