Treatment Effects

ECON 413

Murat Genç

Overview I

1. Introduction
2. Main Scenario
3. Basics
   - The Counterfactual Framework
   - Types of Treatment Effects
4. Estimation under Random Assignment
   - Difference-in-Means (DIM) Estimator
     - Mean Independence
5. Methods Assuming Ignorability (or Unconfoundedness) of Treatment
   - Identification based on regression functions
   - Identification based on propensity score weighting
6. Estimation Methods Based on Selection on Observables
   - Regression Adjustment
     - Overlap Issues
     - Regression Adjustment with Randomized Assignment
     - Stata implementation of RA
   - Propensity Score Methods
     - Inverse Probability Weighted (IPW) Estimator
     - Regression on the Propensity Score
     - IPW: Stata Implementation
   - Doubly-Robust Estimation: Combining Regression Adjustment and PS Weighting
   - Matching Methods

- Choosing Among Estimators
- Evaluating the Quality of Matching

7. Estimation Methods Based on Selection on Unobservables (Hidden Bias)
   - Instrumental Variables
   - Selection Model
   - Difference-in-Differences (DID)

8. Regression Discontinuity Design (RD)

9. Final Remarks

- The causal effect of a binary treatment on outcomes.
- A central component of empirical research in economics and many other disciplines.
- The study of cause-and-effect relationships. (Programme evaluation)
- To what extent can the *net difference* observed in outcomes between treated and nontreated groups be attributed to the intervention, given that all other things are held constant (or ceteris paribus)?
- "Treatment effect could be the single most important topic for science."(Myoung-Jae Lee (2005) *Micro-econometrics for Policy, Program, and Treatment Effects* Oxford University Press.)

- The causal effect of a training programme on wages.
- The effect of attending university on wages.
- The word 'treatment' is used in a broad sense. It can refer to many things: receiving a benefit from WINZ, changes in rules and regulations, smoking, alcohol abuse, etc. Almost everything can be called treatment.
- Similar to analysing the effectiveness of a medical treatment. Except we work with observational data. (No randomised controlled experiments.)
- We try our best to approximate a real experiment with the data we have. (The manner in which this is done is known as the *identification strategy*.)
- Quasi-experimental methods.

- Suppose we want to know the effect of a drug (a treatment) on blood pressure (a response variable) by comparing two people, one treated and the other not.
- If the two people are exactly the same, other than the treatment status, then the difference between their blood pressures can be taken as the effect of the drug on blood pressure.
- If they differ in some other way than in the treatment status, however, the difference in blood pressures may be due to differences other than the treatment status difference.
- The main catchphrase in treatment effect is **compare comparable people**.
- Of course, it is impossible to have exactly the same people.
- The issue is what can be done to solve this problem.

- Suppose we want to know the effect of a childhood education program (treatment) at age 5 on a cognition test score (response) at age 10. How do we know if the treatment is effective?
- We need to compare two *potential* test scores at age 10, one (y_1) with the treatment and the other (y_0) without.
- If $y_1 - y_0 > 0$, then we can say that the program worked.
- However, we *never* observe *both* y_0 and y_1 for the same child as it is impossible to go back to the past and '(un)do' the treatment.
- The observed response is $y = dy_1 + (1 - d)y_0$ where $d = 1$ means treated and $d = 0$ means untreated.
- Instead of the individual effect $y_1 - y_0$, we may look at the *mean effect* $E(y_1 - y_0) = E(y_1) - E(y_0)$ to define the treatment effectiveness as $E(y_1 - y_0) > 0$.

- The ideal way to find the mean effect is a randomised experiment: get a number of children and divide them **randomly** into two groups, one treated (*treatment group*, or '$d = 1$ group', or 'T group') from whom y_1 is observed, and the other untreated (*control group*, '$d = 0$ group') from whom y_0 is observed.
- The role of randomisation is to choose (in a particular fashion) the 'path' 0 or 1 for each child. At the end of each path, there is the outcome y_0 or y_1 waiting, which is not affected by randomisation.
- The particular fashion is that two groups are homogenous on average in terms of the variables other than d and y.

- However, randomisation is hard to do. If the program seems harmful, it would be unacceptable to randomise any child to group T; if the program seems beneficial, the parents would be unlikely to let their child be randomised to group C.
- An alternative is to use observational data where the children (i.e., their parents) self-select the treatment.
- Suppose the program is perceived as good and requires a hefty fee. Then the T group could be markedly different from the C group: the T group's children could have lower (baseline) cognitive ability at age 5 and richer parents.
- Suppose x denotes the observed variables and ε denotes unobserved variables that would matter for y. For instance, x consists of the baseline cognitive ability at age 5 and parents' income, and ε consists of the child's genes and lifestyle.

- Suppose we ignore the differences across the two groups in x or ε just to compare the test scores at age 10.
- Since the T group are likely to consist of children of lower baseline cognitive ability, the T group's test score at age 10 may turn out to be smaller than the C group's.
- We may then falsely conclude no effect of treatment or even a negative effect.
- Clearly, this comparison is wrong: we will have compared incomparable subjects, in the sense that the two groups differ in the observable x or unobservable ε.
- The group mean difference $E(y|d=1) - E(y|d=0)$ may not be the same as $E(y_1 - y_0)$, because

$$E(y|d=1) - E(y|d=0) = E(y_1|d=1) - E(y_0|d=0) \neq E(y_1) - E(y_0).$$

- $E(y_1|d=1)$ is the mean treated response for the richer and less able T group, which is likely to be different from $E(y_1)$, the mean treated response for the C and T groups combined. Similarly, $E(y_0|d=0) \neq E(y_0)$.
- The difference in the observable x across the two groups may cause *overt bias* for $E(y_1 - y_0)$ and the difference in the unobservable ε may cause *hidden bias*. In either case, the treated and control groups differ prior to treatment in ways that matter for the outcomes.
- Dealing with the difference in x or ε is the main task in finding treatment effects with observational data.

- If there is no difference in ε, then only the difference in x should be taken care of.
- The basic way to remove the difference (or imbalance) in x is to select T and C group subjects that share the same x, which is called **matching**.
- That is, compare children whose baseline cognitive ability and parents' income are the same. This yields

$$\begin{aligned} E(y|x,d=1) - E(y|x,d=0) &= E(y_1|x,d=1) - E(y_0|x,d=0) \\ &= E(y_1|x) - E(y_0|x) = E(y_1 - y_0|x). \end{aligned}$$

- The variable d in $E(y_j|x,d)$ drops out once x is conditional on as if d is randomised given x.
- This assumption $E(y_j|x,d) = E(y_j|x)$ is **selection-on-observables** or **ignorable treatment**.

- There are two problems with matching.
  - *Dimension problem*: If x is high-dimensional, it is hard to find control and treat subjects that share exactly the same x.
  - *Support problem*: The T and C groups do not overlap in x.
- For example, suppose x is parental income per year and $d = 1[x \geq \tau]$ where $\tau = \$100,000$ (i.e., when parents who earn more than \$100,000 all put their child in the treatment group). Then the T group are all rich and the C group are all (relatively) poor and there is no overlap in x across the two groups.

- For x to cause an overt bias, it is necessary that x alters the probability of receiving the treatment.
- This provides a way to avoid the dimension problem in matching on x: match instead on the one-dimensional **propensity score** $\pi(x) \equiv P(d = 1) = E(d|x)$.
- That is, compute $\pi(x)$ for both groups and match only on $\pi(x)$.
- $\pi(x)$ can be estimated with logit or probit.

- The support problem is binding when both $d = 1[x \geq \tau]$ and x affect (y_0, y_1): x should be controlled for, which is, however, impossible due to no overlap in x.
- Due to $d = 1[x \geq \tau]$, $E(y_0|x)$ and $E(y_1|x)$ have a break (discontinuity) at $x = \tau$; this case is called **regression discontinuity** (or **before-after** if x is time).
- The support problem cannot be avoided, but subjects near the threshold τ are likely to be similar and thus comparable.
- This comparability leads to 'threshold (or borderline) randomisation', and this randomisation identifies $E(y_1 - y_0|x \approxeq \tau)$, the mean effect for the subpopulation $x \approxeq \tau$.

- Suppose there is no dimension nor support problem, and we want to find comparable control subjects (*controls*) for each treated subject (*treated*) with matching.
- The matched controls are called a 'comparison group', and there are decisions to make in finding a comparison group.
- First, how many controls there are for each treated. If one, we get **pair matching**, and if many, we get **multiple matching**.
- Second, in the case of multiple matching, exactly how many, and whether the number is the same for all treated or different to be determined.
- Third, whether a control is matched only once or multiple times.
- Fourth, whether to pass over (i.e., drop) a treated or not if no good matched control is found.
- Fifth, to determine a 'good' match, a distance should be chosen for $|x_0 - x_1|$ for treated x_1 and control x_0.

- Once these decisions are made and matching is implemented, the matching success is gauged by checking balance of x across the two groups.
- Although it seems easy to pick the variables to avoid overt bias, selecting x can be deceptively difficult. For example, if there is an observed variable w that is affected by d and affects y, should w be included in x?
- Dealing with hidden bias due to imbalance in unobservable ε is more difficult than dealing with overt bias, simply because ε is not observed.
- However, there are many ways to remove or determine the presence of hidden bias.

- There are several other issues as well.
- Firstly, the mean effect is not the only effect of interest. We may be more interested in lower quantiles of $y_1 - y_0$ than in $E(y_1 - y_0)$. Alternatively, instead of mean or quantiles, whether or not y_0 and y_1 have the same marginal distribution may also be interesting.
- Secondly, instead of matching, it is possible to *control for x by weighting the T and C group samples differently*.
- Thirdly, the T and C groups may be observed multiple times over time (before and after the treatment), which leads us to **difference in differences (DID)**, and related study designs.
- Fourthly, binary treatments are generalised into **multiple treatments** that include **dynamic treatments** where binary treatments are given repeatedly over time.

- The standard problem in treatment evaluation involves the inference of a causal connection between the treatment and the outcome.
- We observe $(y_i, \mathbf{x}_i, d_i)$, $i = 1, \ldots, N$, and the impact of a hypothetical change in d on y, holding $\mathbf{x}$ constant, is of interest.
- The outcome variable of interest needs to be compared in the treated and nontreated states.
- However, no individual is simultaneously observed in both states. (The fundamental problem of causal inference.)
- Hence, the situation is akin to one of missing data, and it can be tackled by methods of causal inference carried out in terms of **counterfactuals**.

- **counterfactual**: A **potential outcome**,i.e.,the state of affairs that would have happened in the absence of the cause.
- Note the use of the subjunctive mood (i.e., contingent on what "what would have happened". The counterfactual is not observed in real data.
- The task is to use known information to impute a missing value for a hypothetical and unobserved outcome.
- Individuals selected into either treatment or nontreatment groups have potential outcomes in both states: that is, the one in which they are observed and the one in which they are not observed. (Neyman-Rubin framework.)
- The observed outcome: $y_i = y_{1i}d_i + y_{0i}(1 - d_i)$ where states 0 and 1 correspond to nontreatment and treatment, and $d_i = 1$ denote the receipt of treatment.
- Causal effect: $\delta_i = y_{1i} - y_{0i}$.

- Causal relation is different from associative relation such as correlation or covariance: we need (d_i, y_{0i}, y_{1i}) in the former to get $y_{1i} - y_{0i}$, while we need only (d_i, y_i) in the latter; of course an associative relation suggests a causal relation.
- Correlation, $Corr(d_i, y_i)$, between d_i and y_i is an association; so is the least squares estimator (LSE) $Cov(d_i, y_i)/Var(d_i)$.
- Thus, OLS is used only for association although we tend to interpret LSE findings in practice as if they are causal findings.
- When an association between two variables d_i and y_i is found, it is useful to think of the following three cases.

1. d_i influences y_i unidirectionally: $(d_i \longrightarrow y_i)$
2. y_i influences d_i unidirectionally: $(d_i \longleftarrow y_i)$
3. There are third variables, w_i, that influence both d_i and y_i unidirectionally although there is no direct relationship between d_i and y_i: $(d_i \longleftarrow w_i \longrightarrow y_i)$.

- In treatment analysis, we fix the cause and try to find the effect; thus case 2 is ruled out.
- What is difficult is to tell case 1 from 3 which is a "common factor" case (w_i is common variables for d_i and y_i).

- The triple (y_0, y_1, d) represents a random vector from the underlying population.
- For a random draw i from the population, we write (y_{0i}, y_{1i}, d_i).
- Our interest is $\delta = y_1 - y_0$.
- Because this is a random variable (that is, individual specific), we must be clear about what feature of its distribution we want to estimate.
- Several possibilities have been suggested.
- **Average treatment effect (ATE)**: The expected effect of treatment on a randomly drawn person from the population. (Rosenbaum & Rubin (1983))

$$\tau_{ate} \equiv E(y_1 - y_0) \tag{1}$$

- This is the most popular treatment effect, due to the linearity of the $E(\cdot)$:

$$E(y_1 - y_0) = E(y_1) - E(y_0)$$

  so that the mean effect can be found from the two marginal means of the treatment and control groups.
- **Average treatment effect on the treated (ATT)**: The mean effect for those who actually received the treatment. (Heckman (1992))

$$\tau_{att} \equiv E(y_1 - y_0 | d = 1) \tag{2}$$

- With heterogeneous treatment effects, (1) and (2) can be very different. The ATE might average across the gain from units that would be very unlikely to be subject to treatment (but this depends how the population is defined).

- τ_{ate} has "external validity" in that it tells us something about a randomly drawn unit from the population. τ_{att} is specific to the particular program assignment mechanism.
- Important point: τ_{ate} and τ_{att} are defined without reference to a model or a discussion of the nature of the treatment. In particular, these definitions hold when whether assignment is randomized, unconfounded, or endogenous.
- As noted previously, the difficulty in estimating (1) or (2) is that we observe only y_0 or y_1, not both. The observed outcome is

$$y = y_0(1 - d) + dy_1. \tag{3}$$

- The question: How can we estimate τ_{ate} or τ_{att} with a random sample on y and d (and usually some observed covariates)?
- How we do it depends on what we assume about treatment assignment.
- **Random assignment**: Suppose that the treatment indicator d is statistically independent of (y_0, y_1), as would occur when treatment is *randomised* across agents.
- Behavioural implication of this assumption is that participation in the treatment programme does not depend on outcomes.
- One implication of independence between treatment status and the potential outcome is that τ_{ate} and τ_{att} are identical: $E(y_1 - y_0)|d = 1) = E(y_1 - y_0)$.

- Furthermore, estimation of τ_{ate} is simple. Using equation (3), we have

$$E(y|d=1) = E(y_1|d=1) = E(y_1),$$

where the last equality follows because y_1 and d are independent. Similarly,

$$E(y|d=0) = E(y_0|d=0) = E(y_0).$$

- It follows that

$$\tau_{ate} = \tau_{att} = E(y|d=1) - E(y|d=0).$$

- An unbiased and consistent estimator of $E(y|d=1)$ is the sample average on the treated subsample and similarly for $E(y|d=0)$.
- So, the treatment effect is easily estimated by a difference in sample means: the sample average of y for the treated units minus the sample average of y for the untreated units.

- Thus, randomised treatment guarantees that the difference-in-means (DIM) estimator from basic statistics is unbiased, consistent, and asymptotically normal.
- In fact, these properties will also hold under the weaker assumption of **mean independence**: $E(y_0|d) = E(y_0)$ and $E(y_1|d) = E(y_1)$.
- Randomisation of treatment is often infeasible. In most cases, individuals at least partly determine whether they receive treatment, and their decisions may be related to the benefits of or gain from treatment. In other words, there is **self-selection** into treatment.
- It turns out that τ_{att} can be consistently estimated as a difference in means under the weaker assumption that d is independent of y_0, without placing any restriction on the relationship between d and y_1.

- To see this, note that we can always write[1]

$$E(y|d=1)-E(y|d=0) = E(y_0|d=1)-E(y_0|d=0)+E(y_1-y_0|d=1)$$
$$= E(y_0|d=1) - E(y_0|d=0) + \tau_{att}.$$

- If y_0 is mean independent of d, that is, $E(y_0|d) = E(y_0)$, then the first term disappears, and so the difference in means estimator is an unbiased estimator of τ_{att}.
- Unfortunately, the assumption $E(y_0|d) = E(y_0)$ is still a pretty strong assumption. For example, suppose that people are randomly made eligible for a voluntary job training programme. This assumption implies that the participation decision is unrelated to what people would earn in the absence of the programme.

[1] The equation for the observed outcome implies that $E(y|d=1) = E(y_1|d=1)$ and $E(y|d=0) = E(y_0|d=0)$.

- A useful expression relating τ_{att} and τ_{ate} is obtained by writing $y_0 = \mu_0 + v_0$ and $y_1 = \mu_1 + v_1$, where $\mu_g = E(y_g)$, $g = 0, 1$. Then

$$y_1 - y_0 = (\mu_1 - \mu_0) + (v_1 - v_0) = \tau_{ate} + (v_1 - v_0).$$

  Taking the expectation of this equation conditional on $d = 1$ gives

$$\tau_{att} = \tau_{ate} + E(v_1 - v_0|d = 1).$$

- We can think of $v_1 - v_0$ as the person-specific gain from participation–that is, the deviation from the population mean–and so τ_{att} differs from τ_{ate} by the expected person-specific gain for those who participated.
- If $y_1 - y_0$ is not mean independent of d, τ_{att} and τ_{ate} generally differ.

- Fortunately, we can estimate τ_{ate} and τ_{att} under assumptions less restrictive than independence between (y_0, y_1) and d.
- In most cases, we can collect data on individual characteristics and relevant pretreatment outcomes.
- If, in an appropriate sense, treatment depends on the observables and not on the unobservables determining (y_0, y_1), then we can estimate average treatment effects quite generally, as we show next.

- Let $\mathbf{x}$ denote a vector of observed covariates, so that the population is described by $(y_0, y_1, d, \mathbf{x})$.
- We observe y, d, and $\mathbf{x}$, where y is given by equation (3).
- When d and (y_0, y_1) are correlated, we need an assumption in order to identify treatment effects.
- Rosenbaum and Rubin (1983) introduced the following assumption, which they called **ignorability of treatment** (given covariates $\mathbf{x}$):

ASSUMPTION

ATE.1 (Ignorability) *Conditional on* $\mathbf{x}$*,* d *and* (y_0, y_1) *are independent.*

- That is, conditional on the covariates, the assignment of participants to treated and untreated groups is independent of the outcome of nontreatment and the outcome of treatment.

Ignorability (or Unconfoundedness) of Treatment

- Also known as **unconfoundedness**, **selection on observables**, **conditional independence**, and **exogeneity**.
- d and (y_0, y_1) might be correlated but not once we control for characteristics $\mathbf{x}$. For example, the probability of being chosen for a job training program differs by education levels but is the same at a given level of education.
- A useful way to express ignorability (conditional on $\mathbf{x}$): $D(d|y_0, y_1, \mathbf{x}) = D(d|x)$, where $D(\cdot|\cdot)$ denotes conditional distribution.
- Unconfoundedness is controversial. In effect, it underlies standard regression methods to estimating treatment effects (via a "kitchen sink" regression that includes covariates, the treatment indicator, and possibly interactions).
- Ignorability of treatment certainly holds if d is a deterministic function of $\mathbf{x}$, which is why some authors called it *selection on observables*.

- Intuitively, the ignorability assumption has a better chance of holding when the set of control variables, $\mathbf{x}$, is richer. But one must be careful not to include variables in $\mathbf{x}$ that can themselves be affected by treatment. That would cause ignorability to fail. For example, in evaluating a job training program, $\mathbf{x}$ should not include post-training schooling because that might have been chosen in response to being assigned or not assigned to the program. We would not want to hold post-training schooling fixed.
- In fact, suppose (y_0, y_1) is independent of d but $D(\mathbf{x}|d) \neq D(\mathbf{x})$. In other words, assignment is randomized with respect to (y_0, y_1) but not with respect to $\mathbf{x}$. (Think of random assignment but then $\mathbf{x}$ is defined to include other outcomes affected by d.) Then ignorability generally fails unless $E(y_g|\mathbf{x}) = E(y_g)$, $j = 0, 1$.
- Good candidates for inclusion in $\mathbf{x}$ are variables measured prior to treatment assignment, including past outcomes on y. (Of course, gender, race, and other demographic variables can be included as well.)

- Variables that satisfy instrumental variables assumptions–they are independent of observables that affect (y_0, y_1) but help predict d–should be excluded because their inclusion increases bias in standard regression adjustment estimators unless ignorability holds without the instrument-like variables.
- Unfortunately, ignorability is fundamentally untestable because we only observe $(y, d, \mathbf{x})$. But a sensitivity analysis, similar to studying omitted variables, can be performed.

- For many purposes, it suffices to assume ignorability in a **conditional mean independence** sense:

ASSUMPTION

ATE.1′ (Ignorability in Mean): *(a)* $E(y_0|\mathbf{x}, d) = E(y_0|\mathbf{x})$; *and (b)* $E(y_1|\mathbf{x}, d) = E(y_1|\mathbf{x})$.

- Naturally, ATE.1 implies ATE.1′.
- ATE.1′ allows $Var(y_0|\mathbf{x}, d)$ and $Var(y_1|\mathbf{x}, d)$ to depend on d.
- The idea is this: if we can observe enough information (contained in $\mathbf{x}$) that determines treatment, then (y_0, y_1) might be mean independent of d, conditional on $\mathbf{x}$. Loosely, even though (y_0, y_1) and d might be correlated, they are uncorrelated once we partial out $\mathbf{x}$.
- Under conditional mean independence, the **ATE conditional on x** and the **ATT conditional on x** are identical.

- Assuming that ignorability holds, what is the additional assumption we need to identify the unconditional average treatment effect, τ_{ate}?
- From the law of iterated expectations,

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

  where the expectations are over the distribution of $\mathbf{x}$, and $\mu_g(\mathbf{x}) = E(y_g|\mathbf{x})$, $g = 0, 1$.
- Estimating τ_{ate} requires being able to observe both control and treated units for every outcome on $\mathbf{x}$.
- This assumption is called the **overlap** assumption:

ASSUMPTION

ATE.2 (Overlap): *For all* $\mathbf{x} \in \mathcal{X}$, *where* $\mathcal{X}$ *is the support of the covariates,* $0 < P(d = 1|\mathbf{x}) < 1$.

- The probability of treatment, as a function of $\mathbf{x}$ is called the **propensity score**.
- We denote it $p(\mathbf{x}) = P(d = 1|\mathbf{x})$.
- The overlap assumption rules out the possibility that the propensity score is ever zero or one.
- Ignorability plus overlap is called **strong ignorability**.
- Strong ignorability is critical in estimating τ_{ate}. But the following weaker versions suffice for identifying τ_{att}:

ASSUMPTION

ATT.1′ (Ignorability in Mean): $E(y_0|\mathbf{x}, d) = E(y_0|\mathbf{x})$.

ASSUMPTION

ATT.2 (Overlap): *For all* $\mathbf{x} \in \mathcal{X}$, $0 < P(d = 1|\mathbf{x}) < 1$.

- Given ignorability and overlap assumptions, there are two ways in which τ_{ate} and τ_{att} are identified: based on regression functions or using propensity score weighting.
- Define the **average treatment effect conditional on x** as

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

  where $\mu_g(x) \equiv E(y_g|\mathbf{x})$, $g = 0, 1$.
- The function $\tau(\mathbf{x})$ is of interest on its own right, as it provides the mean effect for different segments of the population described by the observables, $\mathbf{x}$.
- By iterated expectations,

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

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

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

$$\begin{aligned} E(y|\mathbf{x},d) &= (1-d)E(y_0|\mathbf{x},d) + dE(y_1|\mathbf{x},d) \\ &= (1-d)E(y_0|\mathbf{x}) + dE(y_1|\mathbf{x}) \\ &\equiv (1-d)\mu_0(\mathbf{x}) + d\mu_1(\mathbf{x}), \end{aligned} \tag{4}$$

  where the second equality holds by ignorability. So

$$E(y|\mathbf{x},d=0) = \mu_0(\mathbf{x})$$

$$E(y|\mathbf{x},d=1) = \mu_1(\mathbf{x})$$

- The functions $\mu_0(\mathbf{x})$, $\mu_1(\mathbf{x})$ are consistently estimable from the data because we have a random sample on $(y,\mathbf{x},d)$.

- We need to estimate $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. This is possible if the overlap assumption holds.
- Note that, by definition, $E(y|\mathbf{x}, d = 0)$ will be estimated only using the control group and $E(y|\mathbf{x}, d = 1)$ will be estimated only using the treatment group.
- Define the estimable regression functions for the control and treatment groups as

$$m_0(\mathbf{x}) \equiv E(y|\mathbf{x}, d = 0) \quad \text{and} \quad m_1(\mathbf{x}) \equiv E(y|\mathbf{x}, d = 1).$$

- Under ignorability, $m_1(\mathbf{x}) = \mu_1(\mathbf{x})$ and $m_0(\mathbf{x}) = \mu_0(\mathbf{x})$. (If ignorability fails, $m_g(\mathbf{x}) \neq \mu_g(\mathbf{x})$. This is why we use different notation, $m(\cdot)$ and $\mu(\cdot)$.)
- If the overlap assumption holds, we can identify $m_g(\cdot)$ for all $\mathbf{x} \in \mathcal{X}$
- Then, in terms of the estimable mean functions,

$$\tau_{ate} = E[m_1(\mathbf{x}) - m_0(\mathbf{x})]. \tag{5}$$

- Note that the expected value is over the distribution of $\mathbf{x}$. In practice, with a random sample, we use sample averaging.

- If we are able to identify $\tau(\mathbf{x})$ at all $\mathbf{x} \in \mathcal{X}$, which we can under overlap, then we can also identify the average treatment effect on any subset of the population defined by $\mathbf{x}$.
- For the average treatment effect on the treated, i.e., ATT, note that

$$\begin{aligned} E(y_1 - y_0|d) &= E[E(y_1 - y_0|\mathbf{x}, d)|d] = E[E(y_1 - y_0|\mathbf{x})|d] \\ &= E[\mu_1(\mathbf{x}) - \mu_0(\mathbf{x})|d], \end{aligned}$$

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

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

  and we know $\mu_0(\cdot)$ and $\mu_1(\cdot)$ are identified.
- In terms of estimable functions,

$$\tau_{att} = E[m_1(\mathbf{x}) - m_0(\mathbf{x})|d = 1]. \tag{6}$$

- We can also establish identification using propensity score weighting.
- Because $dy = dy_1$, and ignorability implies that d and y_g are uncorrelated conditional on $\mathbf{x}$, by iterated expectations,

$$\begin{aligned} E\left[\frac{dy}{p(\mathbf{x})}\middle|\mathbf{x}\right] &= E\left[\frac{dy_1}{p(\mathbf{x})}\middle|\mathbf{x}\right] = E\left\{E\left[\left(\frac{dy_1}{p(\mathbf{x})}\middle|\mathbf{x},d\right)\right]\middle|\mathbf{x}\right\} \\ &= E\left\{\frac{dE(y_1|\mathbf{x},d)}{p(\mathbf{x})}\middle|\mathbf{x}\right\} = E\left\{\frac{dE(y_1|\mathbf{x})}{p(\mathbf{x})}\middle|\mathbf{x}\right\} = E\left\{\frac{d}{p(\mathbf{x})}\middle|\mathbf{x}\right\}\mu_1(\mathbf{x}) \\ &= \mu_1(\mathbf{x}) \end{aligned}$$

  since $E(d|\mathbf{x}) = p(\mathbf{x})$.
- A similar argument shows that

$$E\left[\frac{(1-d)y}{(1-p(\mathbf{x}))}\middle|\mathbf{x}\right] = \mu_0(\mathbf{x}).$$

- Combining these two results and simple algebra gives

$$\tau_{ate} = E\left\{\frac{[d-p(\mathbf{x})]y}{p(\mathbf{x})[1-p(\mathbf{x})]}\right\} = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}). \tag{7}$$

- Clear from (7) that the overlap assumption is needed: $p(\mathbf{x})$ and $1 - p(\mathbf{x})$ must both be different from zero for all $\mathbf{x}$.
- Intuitively, if we want an average effect over the stated population, then at each $\mathbf{x}$ there must be units in the control and treatment groups.
- We can also show that

$$\tau_{att} = E\left\{\frac{[d - p(\mathbf{x})]y}{\rho[1 - p(\mathbf{x})]}\right\} \tag{8}$$

  where $\rho \equiv P(d = 1)$ is the unconditional probability of treatment.
- Now, we only need to keep $p(\mathbf{x})$ away from unity. Makes intuitive sense because τ_{att} is an average effect for those eventually treated. Therefore, it does not matter if some units have no chance of being treated; they are excluded from the averaging anyway.

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

- Why do many have a preference for PS methods over regression methods?
  1. Estimating the PS requires only a single parametric or nonparametric estimation. Regression methods require estimation of $E(y|d=0,\mathbf{x})$ and $E(y|d=1,\mathbf{x})$ as well as accounting for the nature of y (continuous, discrete, some mixture?)
  2. We have good binary response models for estimating $P(d=1|\mathbf{x})$. Do not need to worry about the nature of y.
  3. Simple propensity score methods have been developed that are asymptotically efficient (although the estimators may not be practically the best, or need some adjustment).
  4. PS methods seem more exotic compared with regression.

Table 1.2 A taxonomy of policy evaluation methods according to the identification assumption, type of specification, and data structure

	Identification assumption		Type of specification		Data structure	
	Selection on observables	Selection on unobservables	Structural	Reduced-form	Cross-section	Longitudinal or repeated cross-section
Regression-adjustment	x			x	x	
Matching	x			x	x	
Reweighting	x				x	
Instrumental-variables	x	x	x		x	
Selection-model	x	x	x		x	
Regression-discontinuity-design	x (sharp)	x (fuzzy)	x (fuzzy)	x (sharp)		
Difference-in-differences	x	x		x		x

Table 1.3 An assessment of the comparative advantages and drawbacks of econometric methods for program evaluation

Method	Advantages	Drawbacks
Regression-adjustment (Control-function regression)	*Suitable for observable selection* *Not based on distributional hypotheses*	*Not suitable for unobservable selection* *Based on a parametric estimation*
Matching	*Suitable for observable selection* *Not based on distributional hypotheses* *Based on a nonparametric estimation*	*Not suitable for unobservable selection* *Sensitive to sparseness (weak overlap)* *Sensitive to confounders' unbalancing*
Reweighting	*Suitable for observable selection* *Not based on distributional hypotheses* *Based on a semi-parametric estimation*	*Not suitable for unobservable selection* *Sensitive to propensity-score specification and/or weighting schemes*
Selection-model	*Suitable for both observable and unobservable selection*	*Based on distributional hypotheses* *Based on a parametric estimation*
Instrumental-variables	*Suitable for both observable and unobservable selection* *Not based on distributional hypotheses*	*Availability of instrumental variables* *Based on a parametric estimation*
Regression-discontinuity-design	*Reproducing locally a natural experiment (randomization)* *No distributional hypothesis* *Extendable to nonparametric*	*Availability of a "forcing" variable* *Choice of the cutoff and of an appropriate bandwidth*

- We have seen that under conditional mean independence,

$$\tau_{ate}(\mathbf{x}) = m_1(\mathbf{x}) - m_0(\mathbf{x}),$$

  where $m_0(\mathbf{x}) \equiv E(y|\mathbf{x}, d = 0)$ and $m_1(\mathbf{x}) \equiv E(y|\mathbf{x}, d = 1)$.
- $m_0(\mathbf{x})$ and $m_1(\mathbf{x})$ depend entirely on observables, and so they can be consistently estimated.
- As soon as consistent estimators of them, $\hat{m}_0(\mathbf{x})$ and $\hat{m}_1(\mathbf{x})$, are available, we can estimate the causal effects by using the sample equivalents:

$$\hat{\tau}_{ate,reg} = \frac{1}{N}\sum_{i=1}^{N}[\hat{m}_1(\mathbf{x}_i) - \hat{m}_0(\mathbf{x}_i)] \qquad \text{and} \tag{9}$$

$$\hat{\tau}_{att,reg} = \left(\sum_{i=1}^{N} d_i\right)^{-1} \sum_{i=1}^{N} d_i \cdot [\hat{m}_1(\mathbf{x}_i) - \hat{m}_0(\mathbf{x}_i)]. \tag{10}$$

- These are called **regression adjustment (RA)** estimators. (Essentially **control function estimators** where control variables are used.)
- Notice that $\hat{\tau}_{att,reg}$ simply averages the differences in predicted values over the subsample of treated units.
- Notice also that we must observe the same set of covariates for the treated and untreated groups. While we can think of the counterfactual setting as being a missing data problem on (y_{i0}, y_{i1}), we assume we do not have missing data on $(d_i, \mathbf{x}_i)$.
- The key implementation issue is how to obtain $\hat{m}_0(\mathbf{x})$ and $\hat{m}_1(\mathbf{x})$.
- This can be done parametrically, semi-parametrically, or nonparametrically.

- RA estimators basically use a two-step procedure.
  1. Fit separate regression models of the outcome on a set of covariates to obtain $\hat{m}_0(\mathbf{x})$ from the "control" subsample, $d_i = 0$, and $\hat{m}_1(\mathbf{x})$ from the "treated" subsample, $d_i = 1$.
  2. Compute fitted values in each case for *all* units in the sample and average them. These averages reflect the potential outcome means (POM). The contrasts of these averages provide estimates of the ATEs. By restricting the computations of the means to the subset of treated subjects, we obtain the ATTs.
- RA estimators are consistent as long as the treatment is independent of the potential outcomes after conditioning on the covariates.
- The model for the outcome variable can be linear or nonlinear.

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

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

  we can easily estimate the ATE for subpopulations described by functions of $\mathbf{x}$.
- If there is not sufficient overlap, $\hat{\tau}_{reg}(\mathbf{x})$ can be a poor estimator for certain values of $\mathbf{x}$.
- Let $\mathcal{R} \subset \mathcal{X}$ be some subset of the possible values of $\mathbf{x}$. We can estimate

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

  as

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

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

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

$$\hat{\tau}_{ate} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \bar{\mathbf{x}}(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0), \tag{11}$$

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

- The linear regression estimate of τ_{att} is

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

  where $\mathbf{x}_1$ is the average of the $\mathbf{x}_i$ over the treated subsample. $\hat{\tau}_{att,reg}$ can be close to $\hat{\tau}_{ate,reg}$ if (1) $\hat{\beta}_1 \approx \hat{\beta}_0$ or (2) $\mathbf{x} \approx \mathbf{x}_1$.
- More generally, if we want to use linear regression to estimate $\hat{\tau}_{ate,\mathcal{R}} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \mathbf{x}_{\mathcal{R}}(\hat{\beta}_1 - \hat{\beta}_0)$, where $\mathbf{x}_{\mathcal{R}}$ is the average over some subset of the sample, then the regression

$$y_i \text{ on } 1,\ d_i,\ \mathbf{x}_i,\ d_i \cdot (\mathbf{x}_i - \mathbf{x}_{\mathcal{R}}),\quad i = 1, ..., N$$

  can be used. Note that it uses all the data to estimate the parameters; it simply centers about $\bar{\mathbf{x}}_{\mathcal{R}}$ rather than $\bar{\mathbf{x}}$.
- If common slopes are imposed, $\hat{\beta}_1 = \hat{\beta}_0$, $\hat{\tau}_{ate,reg} = \hat{\tau}_{att,reg}$ is just the coefficient on d_i from the regression across all observations:

$$y_i \text{ on } 1,\ d_i,\ \mathbf{x}_i,\quad i = 1, ..., N. \tag{12}$$

- If linear models do not seem appropriate for $E(y_0|\mathbf{x})$ and $E(y_1|\mathbf{x})$, we can exploit the specific nature of y_g.
- If y is a binary response, or a fractional response, estimate logit or probit separately for the $d_i = 0$ and $d_i = 1$ subsamples and average differences in predicted values:

$$\hat{\tau}_{ate,reg} = N^{-1} \sum_{i=1}^{N} [G(\hat{\alpha}_1 + \mathbf{x}_i\hat{\boldsymbol{\beta}}_1) - G(\hat{\alpha}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}_0)]. \tag{13}$$

- Each summand in the equation above is the difference in estimated probabilities under treatment and nontreatment for unit i, and the ATE just averages those differences. Use the same approach even if $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_0$ is imposed.
- For general $y \geq 0$, Poisson or gamma regression with exponential mean is attractive:

$$\hat{\tau}_{ate,reg} = N^{-1} \sum_{i=1}^{N} [\exp(\hat{\alpha}_1 + \mathbf{x}_i\hat{\boldsymbol{\beta}}_1) - \exp(\hat{\alpha}_0 + \mathbf{x}_i\hat{\boldsymbol{\beta}}_0)]. \tag{14}$$

- Regardless of the mean function, without good overlap in the covariate distribution, we must extrapolate a parametric model – linear or nonlinear – into regions where we do not have much or any data. For example, suppose, after defining the population of interest for the effects of job training, those with better labor market histories are unlikely to be treated. Then, we have to estimate $E(y|\mathbf{x},d=1)$ only using those who participated – where $\mathbf{x}$ includes variables measuring labor market history – and then extrapolate this function to those who did not participate. This leads to sensitive estimates if nonparticipants have very different values of $\mathbf{x}$.
- Nonparametric methods are not helpful in overcoming poor overlap because they are either based on flexible parametric models (and so require extrapolation) or use local averaging (in which case we cannot estimate $m_1(\mathbf{x})$ for $\mathbf{x}$ values far away from those in the treated subsample).
- The most common local smoothing method, based on kernel estimation, would at least let you know there is very little data to estimate the regression function for values of $\mathbf{x}$ with poor overlap.

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

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

  Therefore, we only need to estimate $m_1(\mathbf{x})$ for values of $\mathbf{x}$ taken on by the treated group, which we can do well. Unlike with the ATE, we do not need to estimate $m_1(\mathbf{x})$ for values of $\mathbf{x}$ in the untreated group. But we need to estimate $\hat{m}_0(\mathbf{x}_i)$ for treated individuals i, and this can be difficult if we have units in the treated group very different from all units in the control group.
- A "solution" is to "balance" the sample by dropping observations that are either very unlikely or very likely to receive treatment, based on the values of $\mathbf{x}$. This is often done based on the propensity score, which we cover below. This effectively changes the population that we are studying.
- It also makes sense to think more carefully about the population ahead of time. If high earners are not going to be eligible for job training, why include them in the analysis at all? The notion of a population is not immutable.

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

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

  and d_i is independent of y_{i0} and therefore v_{i0}.
- Simple regression of y_i on $1, d_i$ is unbiased and consistent for τ.

- But writing the linear projection

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

  we have

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

  where, by randomized assignment, d_i is uncorrelated with $\mathbf{x}_i$ and u_{i0}. So multiple regression is consistent for τ. If $\boldsymbol{\beta}_0 \neq \mathbf{0}$, $Var(u_{i0}) < Var(v_{i0})$, and so adding $\mathbf{x}_i$ reduces the error variance.

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

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

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

- The only caveat is that if $E(y_{i0}|\mathbf{x}) \neq \alpha_0 + \mathbf{x}_i\boldsymbol{\beta}_0$, the OLS estimator of τ is only guaranteed to be consistent, not unbiased. This distinction can be relevant in small samples (as often occurs in true experiments).
- With nonconstant treament effect, add the linear projecton $y_{i1} = \alpha_1 + \mathbf{x}_i\boldsymbol{\beta}_1 + u_{i1}$, so that $\tau_{ate} = \tau = (\alpha_1 - \alpha_0) + \boldsymbol{\mu}_{\mathbf{x}}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)$.
- Now we can write

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

  with $\boldsymbol{\delta} \equiv \boldsymbol{\beta}_1 - \boldsymbol{\beta}_0$ and $e_i \equiv u_{i1} - u_{i0}$.
- Under random assignment of treatment, $(e_i, \mathbf{x}_i)$ is independent of d_i, so d_i is uncorrelated with all other terms in the equation. OLS is consistent for τ but it is generally biased unless the equation represents $E(y_i|d_i, \mathbf{x}_i)$.

- Further,
$$E(\mathbf{x}_i' d_i e_i) = E(d_i)E(\mathbf{x}_i' e_i) = \mathbf{0}$$
and so $\mathbf{x}_i$ and $d_i \cdot (\mathbf{x}_i - \mu_{\mathbf{x}})$ are uncorrelated with $u_{i0} + d_i e_i$ (and this term has zero mean). So OLS consistently estimates all parameters: α_0, τ, $\boldsymbol{\beta}_0$, and $\boldsymbol{\delta}$.
- As a bonus from including covariates interacted with the treatment, we can estimate ATEs as a function of $\mathbf{x}$:
$$\hat{\tau}(\mathbf{x}) = \hat{\tau} + (\mathbf{x} - \bar{\mathbf{x}})\hat{\boldsymbol{\delta}}.$$
- If the $E(y_g|\mathbf{x})$ are not linear, $\hat{\tau}(\mathbf{x})$ is not a consistent estimator of $\tau(\mathbf{x}) = E(y_1 - y_0|\mathbf{x})$, but it should be a reasonable approximation in many cases.

teffects ra (*ovar omvarlist* [, *omodel* **noconstant**]) (*tvar*) [*if*] [*in*] [*weight*]
[, *stat options*]

ovar is a binary, count, continuous, fractional, or nonnegative outcome of interest.

omvarlist specifies the covariates in the outcome model.

tvar must contain integer values representing the treatment levels.

omodel	Description
Model	
linear	linear outcome model; the default
logit	logistic outcome model
probit	probit outcome model
hetprobit(*varlist*)	heteroskedastic probit outcome model
poisson	exponential outcome model
flogit	fractional logistic outcome model
fprobit	fractional probit outcome model
fhetprobit(*varlist*)	fractional heteroskedastic probit outcome model

omodel specifies the model for the outcome variable.

Stat	
ate	estimate average treatment effect in population; the default
atet	estimate average treatment effect on the treated
pomeans	estimate potential-outcome means

The effect of a mother's smoking on infant birthweight using data from Cattaneo (2010).

```
. use http://www.stata-press.com/data/r14/cattaneo2, clear
(Excerpt from Cattaneo (2010) Journal of Econometrics 155: 138-154)
. codebook, compact
```

Variable	Obs	Unique	Mean	Min	Max	Label
bweight	4642	732	3361.68	340	5500	infant birthweight (grams)
mmarried	4642	2	.6996984	0	1	1 if mother married
mhisp	4642	2	.0340371	0	1	1 if mother hispanic
fhisp	4642	2	.037053	0	1	1 if father hispanic
foreign	4642	2	.0534252	0	1	1 if mother born abroad
alcohol	4642	2	.0323137	0	1	1 if alcohol consumed during pregnancy
deadkids	4642	2	.259371	0	1	previous births where newborn died
mage	4642	33	26.50452	13	45	mother´s age
medu	4642	18	12.68957	0	17	mother´s education attainment
fage	4642	46	27.26713	0	60	father´s age
fedu	4642	17	12.3072	0	17	father´s education attainment
nprenatal	4642	30	10.75808	0	40	number of prenatal care visits
monthslb	4642	173	23.07497	0	272	months since last birth
order	4642	12	1.892072	0	12	order of birth of the infant
msmoke	4642	4	.3996122	0	3	cigarettes smoked during pregnancy
mbsmoke	4642	2	.1861267	0	1	1 if mother smoked
mrace	4642	2	.840586	0	1	1 if mother is white
frace	4642	2	.8136579	0	1	1 if father is white
prenatal	4642	4	1.201853	0	3	trimester of first prenatal care visit
birthmonth	4642	12	6.540069	1	12	month of birth
lbweight	4642	2	.0603188	0	1	1 if low birthweight baby
fbaby	4642	2	.4379578	0	1	1 if first baby
prenatal1	4642	2	.8013787	0	1	1 if first prenatal visit in 1 trimester

```
. tabstat bweight prenatal mmarried mage fbaby, by( mbsmoke)
Summary statistics: mean
  by categories of: mbsmoke (1 if mother smoked)
```

mbsmoke	bweight	prenatal	mmarried	mage	fbaby
nonsmoker	3412.912	1.177607	.7514558	26.81048	.4531498
smoker	3137.66	1.30787	.4733796	25.16667	.3715278
Total	3361.68	1.201853	.6996984	26.50452	.4379578

```
. teffects ra (bweight prenatal1 mmarried mage fbaby) (mbsmoke)
Iteration 0:   EE criterion =  7.734e-24
Iteration 1:   EE criterion =  1.196e-25
Treatment-effects estimation                    Number of obs     =      4,642
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
```

bweight	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
mbsmoke (smoker vs nonsmoker)	-239.6392	23.82402	-10.06	0.000	-286.3334	-192.945
POmean						
mbsmoke nonsmoker	3403.242	9.525207	357.29	0.000	3384.573	3421.911

- The average birthweight if all mothers were to smoke would be 240 grams less than the average of 3,403 grams that would occur if none of the mothers had smoked.
- This shows us the average amount by which infants' weights are affected by their mothers' decision to smoke. We may instead be interested in knowing the average amount by which the weight of babies born to smoking mothers was decreased as a result of smoking. The ATT provides us the answer.

```
. teffects ra (bweight prenatal1 mmarried mage fbaby) (mbsmoke), atet
Iteration 0:   EE criterion =  7.629e-24
Iteration 1:   EE criterion =  2.697e-26
Treatment-effects estimation                    Number of obs     =      4,642
Estimator      : regression adjustment
Outcome model  : linear
Treatment model: none
```

bweight	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATET						
mbsmoke (smoker vs nonsmoker)	-223.3017	22.7422	-9.82	0.000	-267.8755	-178.7278
POmean						
mbsmoke nonsmoker	3360.961	12.75749	263.45	0.000	3335.957	3385.966

- The average birthweight is 223 grams less when all the mothers who smoke do so than the average of 3,361 grams that would have occurred if none of these mothers had smoked.
- The ATT differs from the ATE because the distribution of the covariates among mothers who smoke differs from the distribution for nonsmoking mothers.

- Birthweights cannot be negative, though it is possible for a linear regression model to make negative predictions. A common way to enforce nonnegative predictions is to use an exponential conditional mean model, which is commonly fitted using the Poisson quasi maximum-likelihood estimator.

```
. teffects ra (bweight prenatal1 mmarried mage fbaby,poisson) (mbsmoke)
Iteration 0:   EE criterion =  3.950e-17
Iteration 1:   EE criterion =  1.231e-23
Treatment-effects estimation                    Number of obs     =      4,642
Estimator      : regression adjustment
Outcome model  : Poisson
Treatment model: none
```

bweight	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
mbsmoke (smoker vs nonsmoker)	-239.6669	23.83757	-10.05	0.000	-286.3877	-192.9462
POmean						
mbsmoke nonsmoker	3403.178	9.526006	357.25	0.000	3384.508	3421.849

- The formula that establishes identification of τ_{ate} based on population moments, (7), suggests an immediate estimator of τ_{ate}:

$$\tilde{\tau}_{ate,psw} = N^{-1} \sum_{i=1}^{N} \left[\frac{d_i y_i}{p(\mathbf{x}_i)} - \frac{(1-d_i) y_i}{1-p(\mathbf{x}_i)} \right]. \tag{15}$$

- $\tilde{\tau}_{ate,psw}$ is not feasible because it depends on the propensity score $p(\cdot)$.
- Interestingly, we would not use it if we could! Even if we know $p(\cdot)$, $\tilde{\tau}_{ate,psw}$ is not asymptotically efficient. It is *better* to estimate the propensity score!
- Let $\hat{p}(\mathbf{x})$ denote such an estimator obtained using the random sample $\{(d_i, \mathbf{x}_i) : i = 1, \ldots, N\}$.

- Then (7) suggests the estimator

$$\hat{\tau}_{ate,psw} = N^{-1}\sum_{i=1}^{N}\left[\frac{d_i y_i}{\hat{p}(\mathbf{x}_i)} - \frac{(1-d_i)y_i}{1-\hat{p}(\mathbf{x}_i)}\right] = N^{-1}\sum_{i=1}^{N}\frac{[d_i - \hat{p}(\mathbf{x}_i)]y_i}{\hat{p}(\mathbf{x}_i)[1-\hat{p}(\mathbf{x}_i)]}. \tag{16}$$

  while (8) suggests

$$\hat{\tau}_{att,psw} = N^{-1}\sum_{i=1}^{N}\frac{[d_i - \hat{p}(\mathbf{x}_i)]y_i}{\hat{\rho}[1-\hat{p}(\mathbf{x}_i)]}, \tag{17}$$

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

- These are called **inverse probability weighted estimator**s. Weighted averages of the outcomes for each treatment level. Very simple to compute given $\hat{p}(\mathbf{x}_i)$. No model is required for the outcome variable.
- Not surprisingly, these estimators are consistent under Assumptions ATE' and $ATE.2$, or Assumptions $ATT.1'$ and $ATT.2$, respectively.

- As for estimating the propensity score, Rosenbaum and Rubin (1983) suggest using a flexible logit model, where various functions of $\mathbf{x}$–for example, levels, squares, and interactions–are included. As noted by Robins and Rotnitzky (1995, JASA), one never does worse by adding functions of $\mathbf{x}_i$ to the PS model, even if they do not predict treatment! If the functions are correlated with

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

  including them in the logit reduces the error variance in e_i.
- Clearly the sample size is important in deciding on how flexible the model can be.

- Can see directly from $\hat{\tau}_{ate,psw}$ and $\hat{\tau}_{att,psw}$ that the inverse probability weighted (IPW) estimators can be very sensitive to extreme values of $\hat{p}(\mathbf{x}_i)$. $\hat{\tau}_{att,psw}$ is sensitive only to $\hat{p}(\mathbf{x}_i) \approx 1$, but $\hat{\tau}_{ate,psw}$ is also sensitive to $\hat{p}(\mathbf{x}_i) \approx 0$.
- Imbens and coauthors have provided a rule-of-thumb: only use observations with $.10 \leq \hat{p}(\mathbf{x}_i) \leq .90$ (for ATE).
- Sometimes the problem is $\hat{p}(\mathbf{x}_i)$ "close" to zero for many units, which suggests the original population was not carefully chosen.
- After using the PS to choose a new "population," redo the analysis (regression, matching, or PS weighting) where all estimates are based on the new, smaller sample. Of course, because the PS has been estimated, our new "population" is depends on the sample from the original population.

- IPW estimator belongs to a class of estimators known as **reweighting estimators**.
- These estimators are based on a simple idea. When the treatment is not randomly assigned, we expect that the treated and untreated units present very different distributions of their observable characteristics. If this is the case, the distribution of the variables feeding into **x** could be strongly unbalanced.
- To reestablish some balance in the covariates' distributions, a suitable way could be that of weighting the observations by suitable weights and then using a Weighted least squares (WLS) framework to estimate the ATEs.

- The general formula for the Reweighting estimator of ATEs takes the following form:

$$\hat{\tau}_{ate} = \frac{1}{N_1}\sum_{i=1}^{N}\omega_1(i)\cdot d_i\cdot y_i - \frac{1}{N_0}\sum_{j=1}^{N}(1-d_i)\cdot\omega_0(j)\cdot y_j$$

$$\hat{\tau}_{att} = \frac{1}{N_1}\sum_{i=1}^{N} d_i\cdot y_i - \frac{1}{N_0}\sum_{j=1}^{N}(1-d_i)\cdot\omega_0(j)\cdot y_j$$

  where the weights $\omega_0(\cdot)$ and $\omega_1(\cdot)$ add to one in specific cases only.
- IPW estimator does this in a specific way, using the propensity score inverse probability.
- It is based on the intuitive idea of penalizing (advantaging) treated units with higher (lower) probability to be treated and advantaging (penalizing) untreated units with higher (lower) probability to be treated, thus rendering the two groups as similar as possible.

- A different use of the estimated propensity scores is in regression adjustment. A simple, still somewhat popular estimate is obtained from the OLS regression

$$y_i \quad \text{on} \quad 1, d_i, \hat{p}(\mathbf{x}_i), \quad i = 1, \ldots, N; \tag{18}$$

the coefficient on d_i, say, $\hat{\tau}_{ate,psreg}$ is the estimate of τ_{ate}.
- The idea is that the estimated propensity score should be sufficient in controlling for correlation between the treatment, d_i, and the covariates $\mathbf{x}_i$.
- It can actually be shown formally that ignorability holds conditional only on $p(\mathbf{x})$. In other words, it is sufficient to condition only on the propensity score so break the dependence between d and (y_0, y_1). We need not condition on $\mathbf{x}$.
- This implies that

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

- We can then show that

$$\begin{aligned} E[y|p(\mathbf{x}), d &= 0] = E[y_0|p(\mathbf{x})] \\ E[y|p(\mathbf{x}), d &= 1] = E[y_1|p(\mathbf{x})] \end{aligned}$$

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

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

$$y_i \text{ on } 1, \hat{p}(\mathbf{x}_i) \text{ for } d_i = 0 \text{ and } y_i \text{ on } 1, \hat{p}(\mathbf{x}_i) \text{ for } d_i = 1, \tag{19}$$

  which gives fitted values $\hat{\alpha}_0 + \hat{\gamma}_0 \hat{p}(\mathbf{x}_i)$ and $\hat{\alpha}_1 + \hat{\gamma}_1 \hat{p}(\mathbf{x}_i)$, respectively.
- A consistent estimator of τ_{ate} is

$$\hat{\tau}_{ate,regps} = N^{-1} \sum_{i=1}^{N} [(\hat{\alpha}_1 - \hat{\alpha}_0) + (\hat{\gamma}_1 - \hat{\gamma}_0)\hat{p}(\mathbf{x}_i)]. \tag{20}$$

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

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

  as regressors when y has a wide range. So, regress y_i on $1, \hat{r}_i, \hat{r}_i^2, ..., \hat{r}_i^Q$ for some Q using both the control and treated samples, and then average the difference in fitted values to obtain $\hat{\tau}_{ate,regprop}$.
- On balance, regression on the propensity score (or functions of it) has little to offer compared with weighting by the propensity score, provided the overlap issue is attended to. The PS weighted estimator does not require us to model $E[y_g|p(\mathbf{x})]$, and PS weighting can be asymptotically efficient.

IPW: Stata Implementation

- `teffects ipw` implements the inverse-probability weighting estimator.
- Continue with our example of the effect of a mother's smoking on birthweight.
- We can use a probit model to predict treatment status, using `prenatal1`, `mmarried`, `mage`, the square of `mage`, and `fbaby` as explanatory variables:

```
. use http://www.stata-press.com/data/r15/cattaneo2, clear
(Excerpt from Cattaneo (2010) Journal of Econometrics 155: 138-154)
. teffects ipw (bweight) (mbsmoke mmarried c.mage##c.mage fbaby medu, probit)
Iteration 0:   EE criterion =  4.621e-21
Iteration 1:   EE criterion =  7.367e-26
Treatment-effects estimation                    Number of obs     =      4,642
Estimator      : inverse-probability weights
Outcome model  : weighted mean
Treatment model: probit
```

bweight	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
mbsmoke (smoker vs nonsmoker)	-230.6886	25.81524	-8.94	0.000	-281.2856	-180.0917
POmean						
mbsmoke nonsmoker	3403.463	9.571369	355.59	0.000	3384.703	3422.222

The average birthweight if all mothers were to smoke would be 231 grams less than the average of 3,403 grams that would occur if none of the mothers had smoked. (It was 240 with `ra`.)

We use the atet option to estimate ATT:

```
. teffects ipw (bweight) (mbsmoke mmarried c.mage##c.mage fbaby medu, probit),atet
Iteration 0:   EE criterion =  4.621e-21
Iteration 1:   EE criterion =  9.204e-27
Treatment-effects estimation                    Number of obs     =      4,642
Estimator      : inverse-probability weights
Outcome model  : weighted mean
Treatment model: probit
```

bweight	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATET						
mbsmoke (smoker vs nonsmoker)	-225.1773	23.66458	-9.52	0.000	-271.559	-178.7955
POmean						
mbsmoke nonsmoker	3362.837	14.20149	236.79	0.000	3335.003	3390.671

The average birthweight is 225 grams less when all the mothers who smoke do so than the average of 3,363 grams that would have occurred if none of these mothers had smoked. (It was 223 with `ra`.)

Doubly-Robust Estimation: Combining Regression Adjustment and PS Weighting

- Combining different methods may sometimes lead to an estimation of the treatment effects having better properties in terms of robustness. Combining PS weighting and regression adjustment is one of these cases, resulting in what is called a **doubly-robust estimator**.
- More robust, hence the name.
- More robust, because, surprisingly, only one of the two models need to be correctly specified to consistently estimate the treatment effects.
- Intuitively, the doubly-robust estimator is an IPW that includes an augmentation term that corrects the estimator when the treatment model is misspecified. When the treatment is correctly specified, the augmentation term vanishes as the sample size becomes large.

Doubly-Robust Estimation: Combining Regression Adjustment and PS Weighting

The application of the Doubly-robust estimator is as follows:

- Define a parametric function for the conditional mean of the two potential outcomes as $m_0(\mathbf{x}, \boldsymbol{\delta}_0)$ and $m_1(\mathbf{x}, \boldsymbol{\delta}_1)$, respectively, and let $p(\mathbf{x}, \gamma)$ be a parametric model for the propensity-score.
- Estimate $\hat{p}(\mathbf{x}_i)$ by the maximum likelihood (logit or probit).
- Apply a WLS regression using as weights the inverse probabilities to obtain, by assuming a linear form of the conditional mean, the parameters' estimation as:

$$\min_{\alpha_1, \boldsymbol{\beta}_1} \sum_{i=1}^{N} d_i (y_i - \alpha_1 - \mathbf{x}_i \boldsymbol{\beta}_1)^2 / \hat{p}(\mathbf{x}_i)$$

$$\min_{\alpha_0, \boldsymbol{\beta}_0} \sum_{i=1}^{N} (1 - d_i)(y_i - \alpha_0 - \mathbf{x}_i \boldsymbol{\beta}_0)^2 / \hat{p}(\mathbf{x}_i)$$

- Finally, estimate ATEs by Regression-adjustment as:

$$\hat{\tau}_{ate,pswreg} = N^{-1} \sum_{i=1}^{N} [(\hat{\alpha}_1 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_1) - (\hat{\alpha}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_0)] \tag{21}$$

- This is the same formula as linear regression adjustment, but we are using different estimates of $\alpha_g, \boldsymbol{\beta}_g$, $g = 0, 1$.

Doubly-Robust Estimation: Combining Regression Adjustment and PS Weighting

- For ATT:

$$\hat{\tau}_{att,pswreg} = N_1^{-1} \sum_{i=1}^{N} [(\hat{\alpha}_1 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_1) - (\hat{\alpha}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}}_0)] \tag{22}$$

- Two different arguments are invoked to illustrate why the Doubly-robust estimator is consistent (see Wooldridge 2010, pp. 931–932):
  1. In the first case, the conditional mean is correctly specified but the propensity score function is freely misspecified. In this case, robustness is assured by the fact that WLS consistently estimate the parameters independently of the specific function of $\mathbf{x}$ used to build weights. Thus, even an incorrect propensity-score does not affect ATEs consistency.
  2. In the second case, the conditional mean is misspecified but the propensity-score function is correctly specified. We know that with the selection probabilities correctly specified that IPW consistently estimates the linear projection parameters in $L(y_g|1,\mathbf{x}) = \alpha_g + \mathbf{x}\boldsymbol{\beta}_g$ (because these parameters solve the population least squares problem).
     So the estimator of τ based on $\tau = (\alpha_1 + \mu_{\mathbf{x}}\boldsymbol{\beta}_1) - (\alpha_0 + \mu_{\mathbf{x}}\boldsymbol{\beta}_0)$, given in (21) is consistent. This also continues to hold when we consider functions of $\mathbf{x}$.
- The previous results can be seen to hold, with slight modifications, even in the case of binary, fractional and count response variables, provided that the corresponding conditional mean function is considered.

- `teffects aipw` implements the doubly-robust estimator.
- The following code uses a probit model to predict treatment status as a function of `mmarried`, `mage`, and `fbaby`; to maximize the predictive power of this model, we use factor-variable notation to incorporate quadratic effects of the mother's age, the only continuous covariate in our model. We use linear regression to model `birthweight`, using `prenatal1`, `mmarried`, `mage`, and `fbaby` as explanatory variables.

Stata implementation of Doubly-Robust Estimation

```
. teffects aipw (bweight prenatal1 mmarried mage fbaby)(mbsmoke mmarried c.mage##c.mage fbaby medu,
> probit)

Iteration 0:   EE criterion =  4.629e-21
Iteration 1:   EE criterion =  1.837e-25

Treatment-effects estimation                    Number of obs     =      4,642
Estimator      : augmented IPW
Outcome model  : linear by ML
Treatment model: probit
```

bweight	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
mbsmoke (smoker vs nonsmoker)	-230.9892	26.21056	-8.81	0.000	-282.361	-179.6174
POmean						
mbsmoke nonsmoker	3403.355	9.568472	355.68	0.000	3384.601	3422.109

- Matching estimators are based on imputing a value on the counterfactual outcome for each unit.
- The idea behind Matching is simple, intuitive, and attractive, and this can partly explain its popularity. It can be summarized in the following statement: *Recovering the unobservable potential outcome of one unit using the observable outcome of* **similar** *units in the opposite status*.
- In our birthweight and smoking example, we could select a mother who smokes and select a mother of the same age who does not smoke and compare the birthweights of the infants. The data for each member serve as the potential outcome for the other. ("Matching" (in observable characteristics) mothers.)
- Need matching criteria.
- For a single covariate such as age, identifying a pair of comparable mothers is not difficult. If we have a second covariate that is categorical, such as race, we might still be able to identify pairs of mothers who are the same age and of same race. However, if we have covariates that are measured on continuous scales or have more than a few discrete ones, then finding identical matches is a challenge.

- The solution is to use what is called a similarity measure, which is a statistic that measures how "close" two observations are.
- This "distance" can be measured in two ways:
  1. based on the observables $\mathbf{x}$, so that one can calculate, in a meaningful way, how far $\mathbf{x}_i$ is from $\mathbf{x}_j$, where unit j is assumed to be in the opposite treatment group (known as covariates Matching or C Matching)
  2. or on the basis of only one single index-variable, the propensity score $p(\mathbf{x}_i)$, synthesizing all covariates in a one-dimension variable (known as propensity-score Matching or PS Matching)
- Irrespective of the method chosen, the estimation of the average treatment effect for unit i, $ATE_i(\mathbf{x}_i)$ would be simply given by

$$\widehat{ATE}_i(\mathbf{x}_i) = y_{1i} - \hat{y}_{0i},$$

assuming that y_{0i} is perfectly estimated by using some average of the outcome of (matched) untreated individuals.

- The estimators of the causal effects are then obtained by averaging properly these quantities over i:

$$\hat{\tau}_{ate} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{y}_{1i} - \hat{y}_{0i} \right) \tag{23}$$

$$\hat{\tau}_{att} = \frac{1}{N_1} \sum_{i=1}^{N} d_i \left(y_{1i} - \hat{y}_{0i} \right) \tag{24}$$

  where N_1 is the number of treated units.

- Thus, Matching directly uses the observed outcome for treated instead of an estimation of the conditional predictions as in the Regression Adjustment approach.

- In principle, Matching identifies ATEs only under two assumptions: conditional mean independence (ATE.1′) and overlap (ATE.2).
- However, this is only so if the Matching is *exact*, i.e., only if it is possible to build a finite number of cells based on crossing the values taken by the various **x**. When this is not possible, as usually happens, when **x** contains at least one continuous variable, then we need a third assumption in order to identify ATEs:

ASSUMPTION

Balancing: *After matching, the covariates' distribution in the treated and control groups is the same.*

- In order for Matching to be a reliable procedure for estimating the actual ATEs, we have to rely on a "plausible degree" of balancing over the observables; this should be possible to test using some suitable test statistic after Matching is completed.
- Therefore, at least in principle, only when Matching passes the "balancing test," can we rely on the estimated treatment effects.

- Matching treated and untreated individuals on their observable characteristics **x**.
- Practicable when the covariates are discrete and the sample contains many observations at each distinct value of **x**.
- The data are stratified into *cells* defined by each particular value of **x**.
- Within each cell (i.e., conditioning on **x**), one computes the *difference* between the average outcomes of the treated and that of the controls.
- These differences are averaged with respect to the distribution of **x** in the population of treated to obtain ATE.
- So, the average treatment effects are a weighted average of the treatment effects with weights equal to the probability of **x** within the set of treated or untreated individuals.

- Exact Matching is feasible only when **x** has a very small dimensionality (taking, for instance, just three values). But if the sample is small, the set of covariates **x** is large and many of them take discrete multivalues or, even worse, they are continuous variables, then exact Matching is unfeasible.
- For example, if **x** is made of K binary variables, then the number of cells becomes 2^K, and this number increases further if some variables take more than two values.
- If the number of cells (or "blocks") is very large with respect to the size of the sample, it is possible that some cells contain only treated or only control subjects. Thus, the calculus of ATEs might become unfeasible and ATEs not identied.
- If variables are all continuous, as happens in many socioeconomic applications, it would be even impossible to build cells.

- To avoid this drawback, known as the *dimensionality problem*, Rosenbaum and Rubin (1983) have suggested that individuals are matched according to the propensity score.
- Using the propensity score permits to reduce the multidimensionality to a single scalar dimension, $p(\mathbf{x})$.
- In a parametric context, the estimation of the propensity score is usually obtained through a probit (or logit) regression of d on the variables contained in x. Once the scores are obtained, one may match treated and control units with the same propensity score and then averaging on the differences so obtained.
- The problem is that although the propensity score is a singleton index, it is still a "continuous" variable, and this prevents us from being able to perform an exact Matching.

- A *discretization* procedure of the propensity-score may be implemented to approximate the Exact-Matching approach.

1. Estimating the propensity score:
   - First, start with a parsimonious specication in order to estimate the propensity-score for each individual, using probit, logit, etc.
   - Second,order the individuals according to the estimated propensity score (from the lowest to the highest value).
2. Identify the number of strata by satisfying the balancing property:
   - Third, stratify all observations into blocks such that in each block, the estimated propensity-scores for the treated and the controls are not statistically different:
     - Start with five blocks of equal score range $\{0 - 0.2, ..., 0.8 - 1\}$
     - Test whether the means of the scores for the treated and the controls are statistically different in each block (balancing of the propensity-score)

- If they are, increase the number of blocks and test again
- If not, proceed to the next step

- Fourth, test whether the balancing property holds in all strata for all covariates:
  - For each covariate, test whether the means for the treated and for the controls are statistically different in all strata (balancing for covariates)
  - If one covariate is not balanced in one block, split the block and test again within each finer block
  - If one covariate is not balanced in all blocks, modify the logit/probit/linear estimation of the propensity score adding more interaction and higher order terms and then test the balancing property again.

3 Estimating ATEs:once the balancing property is satisfied and, thus, the optimal number of strata is found, then an (weighted) average of the DIM estimators calculated in the nal blocks provides an estimation of ATEs.

- This procedure is called **stratification Matching**.
- It is one of many types of Matching estimators that can be used.
- It may be rather demanding, as it may be difficult to assure balancing for all covariates within all strata.
- There are other matching methods that provide a less restrictive and, thus, easier way to obtain reliable estimates of ATEs, without requiring to build blocks.
- A typical procedure for estimating ATEs by these approaches takes the following form.

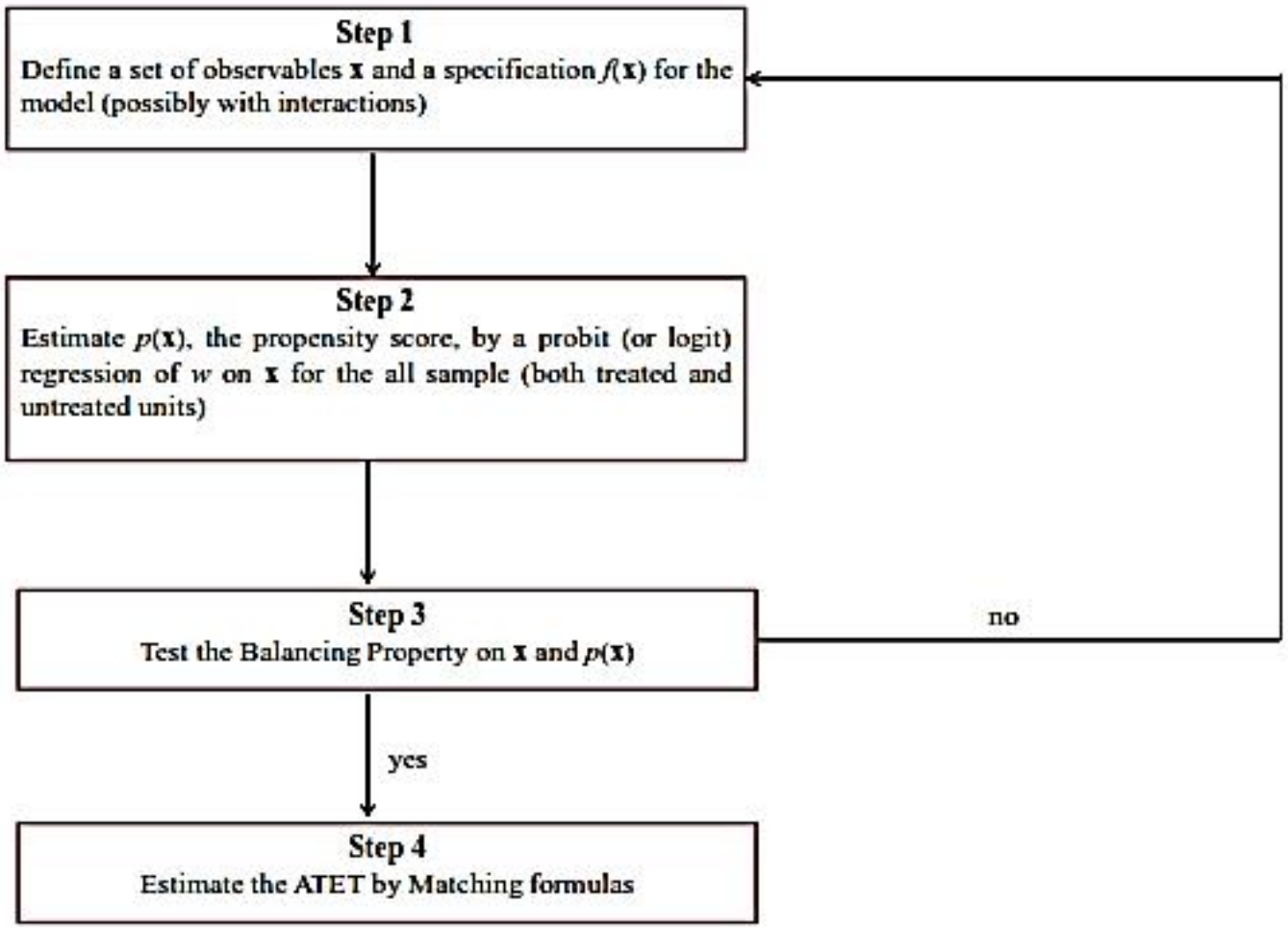


Fig. 2.6 Flow diagram of a Matching protocol

- In this procedure, one should apply Matching estimation just when for each $\mathbf{x}$ and for $p(\mathbf{x})$, no difference emerges in terms of the mean of treated and matched untreated units.
- The advantage of this approach is that it does not require balancing for each x in $\mathbf{x}$ and for $p(\mathbf{x})$ in each stratum since, comparatively, it is "as if" only one single block was built.

- It is clear that perfect balancing with covariates Matching is impossible.
- The other alternative is matching on propensity scores.
- There are at least three reasons to prefer propensity-score Matching over covariates Matching:
  - conditioning on $p(\mathbf{x})$ rather than $\mathbf{x}$ does not undermine consistency and does not increase the variance (precision) of estimation;
  - working with $p(\mathbf{x})$ is easier than working with $\mathbf{x}$, as $p(\mathbf{x})$ is a single variable indexing the overall $\mathbf{x}$. It is computationally preferable to work on only one dimension rather than on k dimensions;
  - knowing $p(\mathbf{x})$ may be interesting per se, having a meaningful theoretical interpretation as it derives from the behavioral selection rule adopted by the individuals within the program/experiment.

- Propensity score methods call for a good model to generate the scores. Our interest is in estimating consistently the participation probability rather than the estimates of parameters in the propensity score function. A better statistical fit for the propensity score is more likely to result from a flexible parametric or nonparametric model.
- In implementing matching based on $p(\mathbf{x})$ three issues are relevant: (1) whether to match with or without replacement, (2) the number of units to use in the comparison set, and (3) the choice of the matching method.
- Matching without replacement means that any observation in the comparison group is matched to no more than one treated observation, that which is the closest match, whereas with replacement means that there can be multiple matches.

- If matching without replacement, the smallness of the comparison set would mean that the matches may not be very close in terms of $p(\mathbf{x})$, which will increase the bias of the estimator.
- The issue of choosing the number of cases in the comparison set involves tradeoff between bias and variance. By using a single closest match to a treated case, one reduces the bias, but by including more matched controls, the variance is reduced whereas bias increases if the additional observations are inferior matches for the treated observations.
- A partial solution is to use a predefined neighborhood in terms of a radius around the $p(\mathbf{x})$ of the treated observation and to exclude matches that lie outside this neighborhood. In other words, one only uses the better matches. This is called **caliper matching**.

- Although there are different Matching methods, the imputation of the missing counterfactual follows the following general rule:

$$\hat{y}_{0i} = \begin{cases} y_i & \text{if } d_i = 0 \\ \sum_{j \in C(i)} h(i,j) y_j & \text{if } d_i = 1 \end{cases}$$

and

$$\hat{y}_{1i} = \begin{cases} \sum_{j \in C(i)} h(i,j) y_j & \text{if } d_i = 0 \\ y_i & \text{if } d_i = 1 \end{cases}$$

where $C(i)$, called the "neighborhood" of i, is the set of indices j for the units matched with unit i; $0 < h(i,j) < 1$ are weights to apply to the single j matched with i, and they generally increase as soon as j is closer to i. Observe that i may be treated or untreated.

- Different Matching methods are obtained by specifying different forms of weights $h(i,j)$ and of the set $C(i)$.

Table 2.3 Different Matching methods for estimating ATEs according to the specification of C(i) and $h(i,j)$

Matching method	C(i)	$h(i,j)$
One-nearest-neighbor	$\{\text{Singleton } j : \min_j \|p_i - p_j\|\}$	1
M-nearest-neighbors	$\{\text{First}M\ j : \min_j \|p_i - p_j\|\}$	$\frac{1}{M}$
Radius	$\{j : \|p_i - p_j\| < r\}$	$\frac{1}{N_{C(i)}}$
Kernel	All control units (C)	$\frac{K_{ij}}{\sum_{j\in C} K_{ij}}$
Local-linear	All control units (C)	$\frac{K_{ij}L_i^2 - K_{ij}\widehat{\Delta}_{ij}L_i^1}{\sum_{j\in C}\left(K_{ij}L_i^2 - K_{ij}\widehat{\Delta}_{ij}L_i^1 + r_L\right)}$
Ridge	All control units (C)	$\frac{K_{ij}}{\sum_{j\in C} K_{ij}} + \frac{\widetilde{\Delta}_{ij}}{\sum_{j\in C}\left(K_{ij}\widetilde{\Delta}^2{}_{ij} + r_R h\lvert\widetilde{\Delta}_{ij}\rvert\right)}$
Stratification	All control units (C)	$\frac{\sum_{b=1}^{B} \mathbf{1}[p(\mathbf{x}_i) \in I(b)]\cdot\mathbf{1}[p(\mathbf{x}_j)\in I(b)]}{\sum_{b=1}^{B}\mathbf{1}[p(\mathbf{x}_j)\in I(b)]}$

`teffects nnmatch` implements the nearest-neighbor matching (NNM) estimator.

`teffects nnmatch` (*ovar omvarlist*) (*tvar*) [*if*] [*in*] [*weight*]
[, *stat options*]

ovar is a binary, count, continuous, fractional, or nonnegative outcome of interest.

omvarlist specifies the covariates in the outcome model.

tvar must contain integer values representing the treatment levels. Only two treatment levels are allowed.

stat	Description
Stat	
`ate`	estimate average treatment effect in population; the default
`atet`	estimate average treatment effect on the treated

options	Description
Model	
`nneighbor(#)`	specify number of matches per observation; default is `nneighbor(1)`
`biasadj(varlist)`	correct for large-sample bias using specified variables
`ematch(varlist)`	match exactly on specified variables

- The option `biasadj(varlist)` specifies that a linear function of the specified covariates can be used to correct for a large sample bias that exists when matching on more than one continuous covariate.
- The option `ematch(varlist)` specifies that the variables in `varlist` match exactly.
- We continue with our birthweight example.

Using the Mahalanobis distance, the default. (The "nearest" is determined by using a weighted function of the covariates for each observation. The weights are based on the inverse of the covariates' variance-covariance matrix.)

```
. teffects nnmatch (bweight prenatal1 mmarried mage fbaby) (mbsmoke)
Treatment-effects estimation                   Number of obs      =      4,642
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Distance metric: Mahalanobis                               max =        139
```

bweight	Coef.	AI Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
mbsmoke (smoker vs nonsmoker)	-240.3306	28.43006	-8.45	0.000	-296.0525	-184.6087

The results obtained are almost identical to those found using RA. The average birthweight if all mothers were to smoke would be 240 grams less than the average that would occur if none of the mothers had smoked.

Using the `ematch()` option to require exact matches on the binary variables `prenatall`, `married`, and `fbaby`.

```
. teffects nnmatch (bweight mage)(mbsmoke),ematch(prenatal1 mmarried fbaby) metric(euclidean)

Treatment-effects estimation                   Number of obs      =      4,642
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Distance metric: Euclidean                                 max =        139
```

bweight	Coef.	AI Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
mbsmoke (smoker vs nonsmoker)	-240.3306	28.43006	-8.45	0.000	-296.0525	-184.6087

The results are identical in this case.

Nearest-neighbor matching estimators are not consistent when matching on two or more continuous covariates. A bias-adjusted estimator can be specified by using the option `biasadj()`.

```
. teffects nnmatch (bweight mage fage)(mbsmoke),ematch(prenatal1 mmarried fbaby) biasadj(mage fage)

Treatment-effects estimation                   Number of obs      =      4,642
Estimator      : nearest-neighbor matching     Matches: requested =          1
Outcome model  : matching                                  min =          1
Distance metric: Mahalanobis                               max =         25
```

	bweight	Coef.	AI Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE							
	mbsmoke						
(smoker vs nonsmoker)		-223.8389	26.19973	-8.54	0.000	-275.1894	-172.4883

ATE is now 224 grams, instead of 240.

- `teffects psmatch` implements propensity score matching estimator.
- The propensity score, the probability of treatment, can be predicted by a logit, probit, or heteroskedastic probit model. (The default is logit.)
- PS Matching does not require a bias correction, since it matches units on a single continuous index.
- A three-step procedure:
  - Estimate the propensity score.
  - Choose a matching algorithm that will use the estimated propensity scores to match untreated units to treated units.
  - Estimate the impact of the intervention with the matched sample and calculate standard errors.

Because the performance of PSM hinges upon how well we can predict the propensity scores, we will include both linear and quadratic terms for mage, the only continuous variable in our model.

```
. teffects psmatch (bweight) (mbsmoke mmarried c.mage##c.mage fbaby medu)
Treatment-effects estimation                   Number of obs      =      4,642
Estimator      : propensity-score matching     Matches: requested =          1
Outcome model  : matching                                   min =          1
Treatment model: logit                                      max =         74
```

bweight	Coef.	AI Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
mbsmoke (smoker vs nonsmoker)	-210.9683	32.021	-6.59	0.000	-273.7284	-148.2083

The average birthweight if all mothers were to smoke would be 211 grams less than the average that would occur if none of the mothers had smoked.

In the previous example, each subject was matched to at least one other subject, which is the default behavior for `teffects psmatch`. However, we can request that `teffects psmatch` match each subject to multiple subjects with the opposite treatment level by specifying the `nneighbor()` option. Matching on more distant neighbors can reduce the variance of the estimator at a cost of an increase in bias.

```
. teffects psmatch (bweight) (mbsmoke mmarried c.mage##c.mage fbaby medu),nneighbor(4)

Treatment-effects estimation                   Number of obs      =      4,642
Estimator      : propensity-score matching     Matches: requested =          4
Outcome model  : matching                                  min =          4
Treatment model: logit                                     max =         74
```

bweight	Coef.	AI Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
mbsmoke (smoker vs nonsmoker)	-224.006	29.88627	-7.50	0.000	-282.582	-165.43

- No definitive way to select.
- Three tradeoffs.
- First, if the outcome model is correctly specied, the RA estimator will break down more slowly than the IPW, AIPW, or PSM estimators as the overlap assumption begins to fail. This result depends critically on the ability of the RA estimator to predict into regions in which there are little data.
- Second, if the overlap assumption holds, the AIPW estimator has the double-robust property.
- Third, all the estimators require the same assumptions, so if each is correctly specified, they should all produce similar results. Of course, just because they produce similar results does not mean that they are correctly specified; it is possible that they are just behaving similarly in response to some underlying problem.

- If the model or matching method is well specified, it should balance the covariates.
- Stata provides a range of techniques and tests check the specification of the conditioning method used by the Matching estimators.
- These diagnostic statistics and plots are produced by the `tebalance` command.

`tebalance` *subcommand* ... [, *options*]

subcommand	Description
`summarize`	compare means and variances in raw and balanced data
`overid`	overidentification test
`density`	kernel density plots for raw and balanced data
`box`	box plots for each treatment level for balanced data

Balance after estimators that use weighting

- Using `teffects ipw` to estimate the effect of a mother's smoking behavior (`mbsmoke`) on the birthweight of her child (`bweight`), controlling for marital status (`mmarried`), the mother's age (`mage`), whether the mother had a prenatal doctor's visit in the baby's first trimester (`prenatal1`), and whether this baby is the mother's first child (`fbaby`).

```
. quietly teffects ipw (bweight) (mbsmoke mmarried mage prenatal1 fbaby)
. tebalance overid
Iteration 0:   criterion =  .02146858
Iteration 1:   criterion =  .02159149  (backed up)
Iteration 2:   criterion =  .02177784
Iteration 3:   criterion =  .02260111
Iteration 4:   criterion =  .02267958
Iteration 5:   criterion =   .0229244
Iteration 6:   criterion =  .02430837
Iteration 7:   criterion =  .02456728
Iteration 8:   criterion =  .02488445
Iteration 9:   criterion =  .02530121
Iteration 10:  criterion =  .02545952
Iteration 11:  criterion =  .02550143
Iteration 12:  criterion =  .02552767
Iteration 13:  criterion =  .02554415
Iteration 14:  criterion =  .02554513
Iteration 15:  criterion =  .02554514
Overidentification test for covariate balance
        H0: Covariates are balanced:
        chi2(5)      =  38.1464
        Prob > chi2  =   0.0000
```

We reject the null hypothesis that the specified treatment model balances the covariates

We reject the null hypothesis that the specied treatment model balances the covariates

Balance after estimators that use weighting

- We can use `tebalance summarize` to see where the problem lies. To get an idea of the extent to which the covariates are unbalanced, we begin by summarizing the covariates by group in the raw data by specifying the baseline option.

```
. tebalance summarize, baseline
  Covariate balance summary
                                              Raw        Weighted
                        Number of obs =       4,642       4,642.0
                        Treated obs   =         864       2,315.3
                        Control obs   =       3,778       2,326.7

------------------------------------------------------------------
                |               Means              Variances
                |     Control     Treated     Control     Treated
----------------+-------------------------------------------------
       mmarried |   .7514558   .4733796   .1868194   .2495802
           mage |   26.81048   25.16667   31.87141   28.10429
      prenatal1 |   .8268925   .6898148   .1431792   .2142183
          fbaby |   .4531498   .3715278   .2478707   .2337654
------------------------------------------------------------------
```

It seems that the covariates may not be balanced in the raw data. For example, the distribution of the mother's age may differ over the treatment groups. We can investigate the differences further with standardized differences and variance ratios. A perfectly balanced covariate has a standardized difference of zero and variance ratio of one. There are no standard errors on these statistics, so inference is informal.

Balance after estimators that use weighting

By omitting the baseline option, we obtain these diagnostic statistics for the raw data and the weighted data.

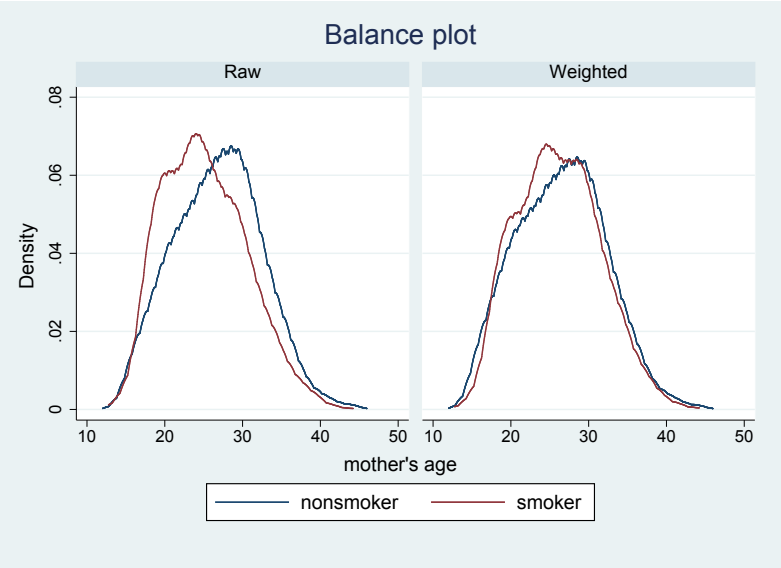
```
. tebalance summarize
  Covariate balance summary
                                          Raw      Weighted
                        Number of obs =         4,642       4,642.0
                        Treated obs   =           864       2,315.3
                        Control obs   =         3,778       2,326.7

                    Standardized differences          Variance ratio
                             Raw    Weighted                Raw   Weighted
-------------------------------------------------------------------------
        mmarried |   -.5953009   -.0105562        1.335944   1.009079
           mage |    -.300179   -.0672115        .8818025   .8536401
       prenatal1 |   -.3242695   -.0156339        1.496155   1.023424
           fbaby |   -.1663271    .0257705        .9430944   1.005698
-------------------------------------------------------------------------
```

We see that for `mmarried`, `prenatal1`, and `fbaby`, our model improved the level of balance. The weighted standardized differences are all close to zero and the variance ratios are all close to one. However, output indicates that `mage` may not be balanced by our model. The weighted standardized difference is close to zero, but the weighted variance ratio still appears to be considerably less than one.

Balance after estimators that use weighting

We can use `tebalance density` to look at how the densities of `mage` for treated and control groups differ.

```
. tebalance density mage
```



The plots also indicate a lack of balance in mage between the treatment groups.

Balance after estimators that use weighting

To try to achieve better balance, we specify a richer model with interactions between mage and the other covariates and look at the resulting standardized differences.

```
. quietly teffects ipw (bweight) (mbsmoke mmarried mage prenatal1 fbaby c.mage#(c.mage i.mmarried pr
> enatal1))
. tebalance summarize
  Covariate balance summary
                                                Raw      Weighted
                     Number of obs =          4,642       4,642.0
                     Treated obs   =            864       2,329.1
                     Control obs   =          3,778       2,312.9
```

	Standardized differences		Variance ratio	
	Raw	Weighted	Raw	Weighted
mmarried	-.5953009	.0053497	1.335944	.9953184
mage	-.300179	.0410889	.8818025	1.076571
prenatal1	-.3242695	.0009807	1.496155	.9985165
fbaby	-.1663271	-.0130638	.9430944	.9965406
mage#				
mage	-.3028275	.0477465	.8274389	1.109134
mmarried#				
mage				
married	-.6329701	.0197209	1.157026	1.034108
prenatal1#				
mage				
Yes	-.4053969	.0182109	1.226363	1.032561

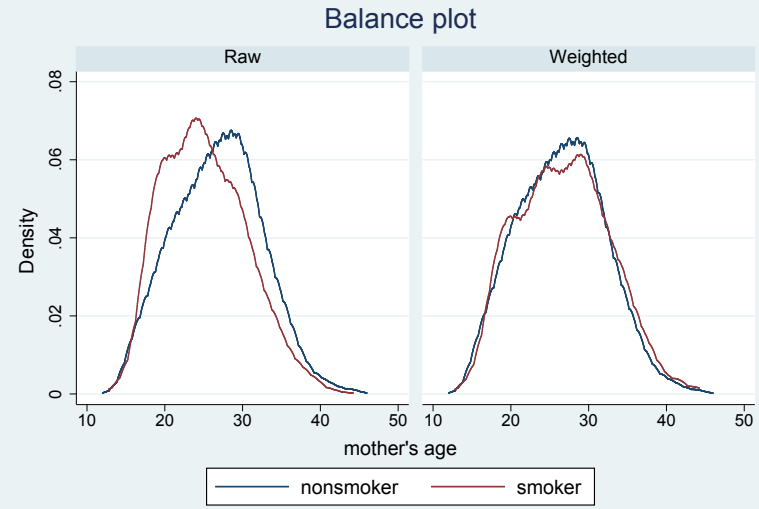
Balance after estimators that use weighting

The standardized difference and variance ratio results for mage look closer to the expected values of zero and one, so we proceed to the formal test.

```
. tebalance overid
Iteration 0:   criterion =   .0602349
Iteration 1:   criterion =  .06172749  (backed up)
Iteration 2:   criterion =  .06427888  (backed up)
Iteration 3:   criterion =  .06488802  (backed up)
Iteration 4:   criterion =  .06526377  (backed up)
Iteration 5:   criterion =   .0664351
Iteration 6:   criterion =   .0715371
Iteration 7:   criterion =  .07635392
Iteration 8:   criterion =  .07668793
Iteration 9:   criterion =  .07680736
Iteration 10:  criterion =  .07691632
Iteration 11:  criterion =  .07752589
Iteration 12:  criterion =  .07770959
Iteration 13:  criterion =  .07772689
Iteration 14:  criterion =  .07773172
Iteration 15:  criterion =  .07774686
Iteration 16:  criterion =  .07775314
Iteration 17:  criterion =  .07775324
Overidentification test for covariate balance
         H0: Covariates are balanced:
         chi2(8)      =  11.8612
         Prob > chi2  =   0.1575
```

We do not reject the null hypothesis that the specied treatment model balances the covariates.

Balance after estimators that use weighting



Not much difference now between the densities of `mage` for treated and control groups.

- When selection into a program is driven not only by observables but also by unobservable-to-the-analyst factors, then the conditional mean independence (CMI) hypothesis no longer holds and Regression-adjustment (including Control function regression), Matching, and Reweighting generally bring biased estimates of treatment effects.
- In the regression approach, the treatment binary variable d becomes endogenous, that is, correlated with the error term, thus preventing ordinary least squares (OLS) from producing consistent estimates of regression parameters, including ATE and ATET.
- In the case of Matching (and propensity-score based Reweighting, for instance), the bias depends on excluding relevant covariates from the variables generating the actual propensity-score and/or from the matching procedure applied on units (as, for instance, in the nearest-neighbor approach).

- There are three methods to cope with selection on unobservables: Instrumental-variables (IV), Selection-models (SM), and Difference-in-differences (DID).
- The application of IV requires the availability of at least one instrumental-variable, i.e., a variable in the dataset which is directly correlated with the selection process, but (directly) uncorrelated with the outcome.
- Selection-models restore consistency under the assumption of joint normality of the error terms of the potential outcomes and of the selection equation.
- The DID estimator requires to have observations before and after the policy event, either for different or for the same set of individuals.

- The IV method makes use of at least one exogenous variable z, the "instrumental variable," which is assumed to have the following two fundamental properties:
  - z is (directly) correlated with treatment d
  - z is (directly) uncorrelated with outcome y.
- These two requirements imply that the selection into program should possibly depend on the same factors affecting the outcome plus z, the instrument, assumed to not directly affect the outcome. The relation between the endogenous variable d and the outcome y can exist (so that empirical correlation might not be zero), but it can be only an "indirect link" produced by the "direct effect" of z on d.
- Algebraically, this represents the classical *exclusion restriction* assumption under which IV methods identify the casual parameters of interest.

- Difference-in-means (DIM) estimator is equal to the coefficient α obtained by an OLS regression of the simple univariate linear model:

$$y = \mu + \alpha d + u$$

so that

$$\alpha = E(y|d = 1) - E(y|d = 0) = DIM.$$

- Suppose now that the selection-into-treatment was driven by a factor x, that is unobservable-to-the-analyst. We want to characterize this situation and show that IV provides an unbiased estimate of α.
- Such a situation entails that the outcome is also function of x. In other words, the true process generating y is:

$$y = \mu + \alpha d + \beta x + u$$

$$y = \mu + \alpha d + \beta x + u$$

- Since x is unobservable, it is part of the error term; thus the model becomes:

$$y = \mu + \alpha d + u^* \tag{25}$$

  with $u^* = \beta x + u$ showing that the treatment d and the new error term u^* are related, for the selection-into-treatment is supposed to depend on x.
- A simple OLS of regression (25), therefore, leads to a biased estimation of α.

- In fact,

$$\alpha_{OLS} = \frac{Cov(y,d)}{Var(d)} = \frac{Cov(\mu + \alpha d + \beta x + u, y)}{Var(d)}$$
$$= \alpha \frac{Var(d)}{Var(d)} + \beta \frac{Cov(x,d)}{Var(d)},$$
$$\Longrightarrow \alpha_{OLS} = \alpha + \beta \frac{Cov(x,d)}{Var(d)}$$
$$\Longrightarrow \alpha_{OLS} = \alpha + \beta\{E(x|d=1) - E(x|d=0)\}.$$

proves that in the case of unobservable selection, a standard OLS is a biased estimator.

- An IV approach can restore consistency, provided that an instrumental-variable z, correlated with d but uncorrelated with u^*, is available.
- If we assume that u is a pure random component, thus uncorrelated by definition with z, we can show that:

$$Cov(z, u^*) = Cov(z, \beta x + u) = \beta Cov(z, x) + Cov(z, u)$$
$$= \beta Cov(z, x) = 0$$

- By starting from (25), and assuming that $Cov(z, u^*)$ is zero, with z as an instrument, we have that:

$$Cov(z, u^*) = Cov(z, y - \mu - \alpha d) = Cov(y, z) - \alpha Cov(d, z) = 0$$

$$\Longrightarrow \alpha_{IV} = \frac{Cov(y, z)}{Cov(d, z)}.$$

- It can now easily be shown that the IV estimator $\hat{\alpha}$ is consistent.

- The observed outcome y can be written as

$$y = \mu_0 + (\mu_1 - \mu_0)d + v_0 + d(v_1 - v_0), \tag{26}$$

  where $\mu_g = E(y_g)$ and $v_g = y_g - \mu_g$, $g = 0, 1$.
- This equation, assuming conditional mean independence does not hold, yields:

$$E(v_1|d, \mathbf{x}) \neq E(v_1|\mathbf{x}) \qquad \text{and} \quad E(v_0|d, \mathbf{x}) \neq E(v_0|\mathbf{x})$$

- If we assume that the stochastic parts are the same, that is, $v_1 = v_0$, then the interaction term disappears so that

$$y = \mu_0 + (\mu_1 - \mu_0)d + v_0. \tag{27}$$

- This equation implies that

$$\tau_{ate} = \tau_{att} = \mu_1 - \mu_0 \tag{28}$$

- The two properties that the instrumental variable, z, should have can be written as:

$$E(v_0|\mathbf{x}, z) = E(v_0|\mathbf{x}) \quad \Leftrightarrow \quad z \text{ is uncorrelated with } v_0 \tag{29}$$

$$E(d|\mathbf{x}, z) \neq E(d|\mathbf{x}) \quad \Leftrightarrow \quad z \text{ is correlated with } d \tag{30}$$

- If we assume that $E(v_0|\mathbf{x}) = \mathbf{x}'\beta$, equation (29) implies, after some algebra, that

$$y = \mu_0 + d\tau_{ate} + \mathbf{x}'\beta + u, \tag{31}$$

where $(\mathbf{x}, z)$ are uncorrelated with the error term u, i.e., $(\mathbf{x}, z)$ are exogenous, but the error term u is still correlated with d, the treatment. Thus, OLS estimation of (31) is inconsistent.

- All of this can be more compactly summarised in the following two-equation structural system:

$$\begin{cases} \text{(a) } y_i = \mu_0 + d_i\tau_{ate} + \mathbf{x}_i'\boldsymbol{\beta} + u_i \\ \text{(b) } d_i^* = \eta + \mathbf{q}_i'\boldsymbol{\delta} + \varepsilon_i \\ \text{(c) } d_i = \begin{cases} 1 & \text{if} \quad d_i^* \geq 0 \\ 0 & \text{if} \quad d_i^* < 0 \end{cases} \\ \text{(d) } \mathbf{q}_i = (\mathbf{x}_i, z_i) \end{cases} \tag{32}$$

where ATE cannot be consistently estimated by an OLS of (3.2a), since without the conditional mean independence assumption, we have $Cov(u_i, \varepsilon_i) \neq 0$, thus d is endogenous in this equation.

- In this system, (3.21a) is known as the *outcome equation*, whereas (3.21b)–or, equivalently, (3.21c)–is known as the *selection equation*, and (3.21d) as the identifying *exclusion restriction*.
- Because the only endogenous variable in the outcome equation is binary, equation (3.21a) is called a **dummy endogenous variable** equation.
- Consistent estimation of ATE in this system relies on three IV methods:
  - Direct 2SLS
  - Probit (or Logit)-OLS
  - Probit (or Logit)-2SLS

- The traditional IV procedure.
- Two sequential OLS regressions in order to calculate the predictions of the endogenous variable d in the first step, and using these predictions as a regressor in the outcome equation in place of the actual d in the second step.
- This approach assumes that the probability to be treated given $\mathbf{x}$ takes a linear form. That is, the first step uses a linear probability model to estimate the selection equation.
- The robustness of this approach hinges mainly on the quality of the chosen instrumental variable z, as a weak instrument (a z poorly correlated with the treatment d) can inflate parameters' standard errors, thus making estimates highly imprecise.

- Replaces the linear probability model in the first stage with nonlinear probability functions (probit or logit).
- So, the error term of the latent selection equation in (3.21b) is assumed to be standard normal or logistic.
- Consistency depends on relying on a correctly specified propensity-score model.

- Caters for the possibility that the propensity-score model in Probit-OLS is not correctly specified.
- Uses a 2SLS (instead of OLS) after running the probit model.
- It works as follows: first, apply a probit (logit) of d on $\mathbf{x}$ and z, thus obtaining the "predicted probability of d"; then, use these probabilities to apply a (direct) 2SLS with the predicted probabilities obtained from the probit (logit) estimation being used as an instrument for d.
- This procedure leads to higher efficiency than that of Direct-2SLS.
- The very advantage of using Probit-2SLS is that, unlike Probit-OLS, it returns consistent estimations even when the first-step probit is incorrectly specified (although, it is no more efcient in this case).

- From a technical point of view, when using Probit-OLS or Probit-2SLS, identifying $(\mu_0, \tau_{ate}, \boldsymbol{\beta})$ in the outcome equation (3.21a) does not require one to introduce z as additional regressor in the selection equation (3.21b).
- Since $F(\mathbf{x}'\boldsymbol{\beta})$ is a nonlinear function of $\mathbf{x}$, then it is not perfectly collinear with $\mathbf{x}$. $F(\mathbf{x}'\boldsymbol{\beta})$ can, therefore, be used as an instrument along with $\mathbf{x}$, since it does not produce problems of collinearity.
- Problems due to collinearity can, however, emerge when $F(\cdot)$ is assumed to be linear (as in the case of the linear probability model).
- When using IV methods such as Probit-OLS and Probit-2SLS, it is, therefore, recommended to have access to at least one instrument z, which can be exploited in the estimation of the selection equation.

- Our discussion so far did not take into account either observable or unobservable heterogeneity. (We assumed that $v_1 = v_0$ in (26).)
- When we eliminate this assumption, minor changes need to be incorporated into these IV procedures.
- If $v_1 \neq v_0$, then $\tau_{ate} \neq \tau_{att}$.
- This assumption states that the same unit has a different reaction to variations in the vector of observables $\mathbf{x}$ when it is treated and untreated.
- For many empirical applications, this seems a more general and reasonable assumption.

- Suppose that v_1 and v_0 are independent of z, and

$$v_0 = g_0(\mathbf{x}) + e_0 \quad \text{with} \quad E(e_0|\mathbf{x}, z) = 0$$

$$v_1 = g_1(\mathbf{x}) + e_1 \quad \text{with} \quad E(e_1|\mathbf{x}, z) = 0$$

  so that $g_0(\mathbf{x}) = E(v_0|\mathbf{x})$ and $g_1(\mathbf{x}) = E(v_1|\mathbf{x})$. Then the observed outcome can be expressed as:

$$y = \mu_0 + \alpha d + g_0(\mathbf{x}) + d[g_1(\mathbf{x}) - g_0(\mathbf{x})] + e_0 + d(e_1 - e_0) \tag{33}$$

- By assuming that $g_0(\mathbf{x}) = \mathbf{x}'\beta_0$, $g_1(\mathbf{x}) = \mathbf{x}'\beta_1$, and $\varepsilon = e_0 + d(e_1 - e_0)$, we can obtain the following regression model

$$y = \mu_0 + d\tau_{ate} + \mathbf{x}'\beta_0 + d(\mathbf{x} - \mu_{\mathbf{x}})\beta + \varepsilon \tag{34}$$

- This model contains two endogenous variables, d and $d(\mathbf{x} - \mu_{\mathbf{x}})$.
- How can we deal with this additional endogenous variable?
- Suppose that $e_1 = e_0$. (That is, *only observable heterogeneity.* This is a quite strong assumption, but one that holds in many applications, especially when we have access to a large set of observable variables and are sure that diversity in units' outcome response is driven by these (available) observable factors.
- Then $\varepsilon = e_0$, and we can conclude that

$$y = \mu_0 + \alpha d + \mathbf{x}'\beta_0 + d(\mathbf{x} - \mu_{\mathbf{x}}) + e_0 \tag{35}$$

- Thus what remains in the model is simply the endogeneity due to d and $d(\mathbf{x} - \mu_{\mathbf{x}})$.

- The following procedure is therefore suitable in order to obtain a consistent estimation of the parameters:
  - Apply a probit of d on $\mathbf{x}$ and z, obtaining p_d, i.e., the "predicted probability of d."
  - Estimate the following equation: $y_i = \mu_0 + \alpha d + \mathbf{x}'\beta_0 + d(\mathbf{x} - \mu_{\mathbf{x}}) + error_i$ using as instruments: $1, p_d, \mathbf{x}_i, p_d(\mathbf{x}_i - \mu_{\mathbf{x}})$.
- This is equivalent to the Probit-2SLS we have seen. Of course, either Direct-2SLS or Probit-OLS procedure can also be applied here with minimal changes.
- A particularly attractive property of a model with heterogeneity is that various functions and interactions of $(\mathbf{x}, z)$ can be used to generate additional instruments, in order to obtain an overidentified setting and thus test the (joint) exogeneity of the instruments.

- Suppose now that $e_1 \neq e_0$. (*both observable and unobservable heterogeneities*) That is, the unobservable component affecting the outcome for a given unit is different when such a unit is treated or untreated.
- The error term, $d(e_1 - e_0)$ in equation (33) now contains the endogenous variable d so that the mean of $d(e_1 - e_0)$ conditional on $\mathbf{x}$ and z is not equal to zero. Thus, to restore consistent estimation, we need to assume some additional conditions.
- One possible solution is to assume that $E[d(e_1 - e_0)|\mathbf{x}, z] = E[d(e_1 - e_0)]$. Then, any function of $(\mathbf{x}, z)$ can be used as instrument in the outcome equation. One can, therefore, apply an IV procedure identical to we have just seen.

- This IV estimator is consistent but generally not efficient. In order to obtain an efficient estimation, one needs to introduce some additional hypotheses.
- It is typical to use the Heckman Selection model ("Heckit").

- Heckman's sample selection model.
- Consistently estimates the parameters in the system (32) without the necessity of including an instrument.
- The cost of not having an instrument to rely on is the necessity for additional assumptions, in particular the joint normality of the error terms.
- Recall that the outcome equation is written as the following, equation (33), for the most general case with both observable and unobservable heterogeneity:

$$y = \mu_0 + \alpha d + g_0(\mathbf{x}) + d[g_1(\mathbf{x}) - g_0(\mathbf{x})] + e_0 + d(e_1 - e_0),$$

which, after some manipulation, leads to

$$y = \mu_0 + \alpha d + \mathbf{x}'\beta_0 + d(\mathbf{x} - \mu_{\mathbf{x}})'\beta + e_0 + d(e_1 - e_0).$$

- A generalized Heckit model can be implemented to obtain consistent and efficient estimates of the parameters if the following assumptions are made:
  1. $y = \mu_0 + \alpha d + \mathbf{x}'\boldsymbol{\beta}_0 + d(\mathbf{x} - \mu_{\mathbf{x}})'\boldsymbol{\beta} + u$
  2. $E(e_1|\mathbf{x}, z) = E(e_0|\mathbf{x}, z) = 0$
  3. $d = 1[\theta_0 + \mathbf{x}'\boldsymbol{\theta}_1 + \theta_2 z + a \geq 0]$
  4. $E(a|\mathbf{x}, z) = 0$
  5. $(a, e_0, e_1) \sim \mathbf{N}$
  6. $a \sim N(0, 1) \Longrightarrow \sigma_a = 1$
  7. $u = e_0 + d(e_1 - e_0)$

  where the most crucial hypothesis here is that of assuming a trivariate normal distribution of the error terms of the potential outcomes (e_1, e_0) and of the selection equation (a).

- Observe that, although z is included as regressor, the identification of such a model does not require an instrumental variable to be specied. The normality assumption is sufficient to obtain consistent results.
- A two-step procedure:
  1. Run a probit of d_i on $(1, \mathbf{x}_i, z_i)$ and get $(\hat{\phi}_i, \hat{\Phi}_i)$
  2. Run OLS of y_i on: $\left[1, d_i, d_i(\mathbf{x}_i - \mu_{\mathbf{x}})_i, d_i \frac{\hat{\phi}_i}{\hat{\Phi}_i}, (1-d_i)\frac{\hat{\phi}_i}{1-\hat{\Phi}_i}\right]$
- This regression can be written as

$$\begin{aligned} E(y|\mathbf{x}, z, d) &= \mu_0 + \alpha d + \mathbf{x}'\boldsymbol{\beta}_0 + d(\mathbf{x} - \mu_{\mathbf{x}}{}'\boldsymbol{\beta} + \rho_1 d\lambda_1(\mathbf{q}'\boldsymbol{\theta}) \\ &+ \rho_0(1-d)\lambda_0(\mathbf{q}'\boldsymbol{\theta}) \end{aligned} \tag{36}$$

where $\lambda_1(\mathbf{q}'\boldsymbol{\theta}) = \frac{\phi(\mathbf{q}'\boldsymbol{\theta})}{\phi(\mathbf{q}'\boldsymbol{\theta})}$ and $\lambda_0(\mathbf{q}'\boldsymbol{\theta}) = \frac{\phi(\mathbf{q}'\boldsymbol{\theta})}{1-\phi(\mathbf{q}'\boldsymbol{\theta})}$ are the inverse Mills ratios.

- The null hypothesis $H_0 : \rho_1 = \rho_0 =$, if accepted, allows one to conclude that there is no selection on unobservables.
- Since under the joint normality assumption, the model is fully parametric, a MLE can be employed, thus not only yielding consistent but also efficient estimations of the causal parameters. Generally, however, maximum likelihood estimation can result in convergence problems, especially when many discrete control variables are used. In such cases, the two-step procedure is a valuable (although less efficient) alternative.

- User-written command `ivtreatreg` (Cerulli 2014).
- `ivtreatreg` fits four binary treatment models with and without idiosyncratic or heterogeneous ATEs. Depending on the model specied, it provides consistent estimation of ATEs under the hypothesis of selection-on-unobservables by using IV and a generalized Heckman-style selection model.
- The four models fit by `ivtreatreg` are direct-2sls (IV regression by direct two-stage least squares), probit-ols (IV two-step regression by probit and OLS), probit-2sls (IV regression by probit and two-stage least squares), and heckit (Heckman two-step selection model).

`ivtreatreg` *outcome treatment* [*varlist*] [*if*] [*in*] [*weight*], `model(`*modeltype*`)`
[`hetero(`*varlist_h*`) iv(`*varlist_iv*`) conf(`#`) graphic vce(`*vcetype*`) beta`
`const(noconstant) head(noheader)`]

where *outcome* specifies the target variable that is the object of the evaluation, *treatment* specifies the binary treatment variable (that is, 1 = treated or 0 = untreated), and *varlist* defines the list of exogenous variables that are considered as observable confounders.

modeltype	Description
`direct-2sls`	IV regression fit by direct two-stage least squares
`probit-2sls`	IV regression fit by probit and two-stage least squares
`probit-ols`	IV two-step regression fit by probit and OLS
`heckit`	Heckman two-step selection model

`iv`(varlist_iv) specifies the variables to be used as instruments. This option is required with `model(direct-2sls)`; it is optional with other *modeltypes*.

- The dataset `fertil2.dta` contains cross-sectional data on 4,361 women of childbearing age in Botswana.
- It contains 28 variables on various female and family characteristics.
- We are interested in evaluating the impact of the variable `educ7` (taking value 1 if a woman has seven years of education or more and 0 otherwise) on the number of family children (`children`). Several conditioning (or confounding) observable factors are included in the dataset, such as the age of the woman (`age`), whether or not the family owns a TV (`tv`), and whether or not the woman lives in a city (`urban`), and so forth.

To inquire about the relationship between education and fertility, we first estimate ATE by using the simple DIM estimator:

```
. use "\\commerce\obsUser01\mgenc\Documents\COURSES 2017\ECON413\Wooldridge Econometric Analysi
> ta Files\fertil2.dta", clear
. **First simple Difference-in-mean (DIM) estimator
. reg children educ7
```

Source	SS	df	MS	Number of obs	=	4,361
				F(1, 4359)	=	810.08
Model	3373.65898	1	3373.65898	Prob > F	=	0.0000
Residual	18153.5174	4,359	4.16460596	R-squared	=	0.1567
				Adj R-squared	=	0.1565
Total	21527.1763	4,360	4.93742577	Root MSE	=	2.0407

children	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ7	-1.770068	.0621908	-28.46	0.000	-1.891994	-1.648143
_cons	3.25129	.0463564	70.14	0.000	3.160408	3.342172

The mean of the treated ones, which are the children in the group of more educated women, minus the mean of the untreated ones, which are the children in the group of less educated women is 1.77 with a tvalue of 28.46. This means that women with more education show about two children fewer than women with less education, without ceteris paribus conditions.

Adding confounding factors.

```
. use "\\commerce\obsUser01\mgenc\Documents\COURSES 2017\ECON413\Wooldridge Econometric Analysis\Sta
> ta Files\fertil2.dta", clear
. **Adding confounding factors
. reg children educ7 age agesq evermarr urban electric tv
```

Source	SS	df	MS	Number of obs	=	4,358
				F(7, 4350)	=	880.03
Model	12607.4006	7	1801.05723	Prob > F	=	0.0000
Residual	8902.63153	4,350	2.04658196	R-squared	=	0.5861
				Adj R-squared	=	0.5855
Total	21510.0321	4,357	4.93689055	Root MSE	=	1.4306

children	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ7	-.3935524	.0495534	-7.94	0.000	-.4907024	-.2964025
age	.2719307	.0171033	15.90	0.000	.2383996	.3054618
agesq	-.001896	.0002752	-6.89	0.000	-.0024356	-.0013564
evermarr	.6947417	.0523984	13.26	0.000	.5920142	.7974691
urban	-.2437082	.0460252	-5.30	0.000	-.333941	-.1534753
electric	-.336644	.0754557	-4.46	0.000	-.4845756	-.1887124
tv	-.3259749	.0897716	-3.63	0.000	-.501973	-.1499767
_cons	-3.526605	.2451026	-14.39	0.000	-4.007131	-3.046079

OLS estimate of ATE is 0.394 with a t-value of 7.94, still in absence of heterogeneous treatment. This is still significant, but the magnitude, as expected, dropped considerably compared with the difference-in-mean estimation, thus showing that confounders are relevant.

- Results change dramatically we do IV estimation.
- The specification we use adopts the covariate `frsthalf` as the IV and takes value 1 if the woman was born in the first six months of the year and 0 otherwise. This variable is partially correlated with `educ7`, but it should not have any direct relationship with the number of family children.
- We estimate the specification with the `probit-2sls` option.
- Results on the probit show that `frsthalf` is partially correlated with `educ7`, thus it can be reliably used as an instrument for this variable. Step 2 shows that the ATE (again, the coefficient of `educ7`) is no more significant and that it changes sign, becoming positive and equal to 0.30.

We can now also calculate the ATET and ATENT, and inspect the cross-unit distribution of these effects. `ivtreatreg` returns these parameters as scalars (along with treated and untreated sample size).

```
. display e(ate)
.30040074
. display e(atet)
.89829002
. display e(atent)
-.44688343
. display e(N_tot)
4358
. display e(N_treat)
2421
. display e(N_untreat)
1937
```

```
. **To get the standard errors for testing ATET and ATENT signicance, we can easily
. *implement a bootstrap procedure as follows:
. bootstrap atet=e(atet) atent=e(atent), rep(100): ivtreatreg children educ7 age agesq evermarr urba
> n electric tv, hetero(age agesq evermarr urban) iv(frsthalf) model(probit-2sls)
(running ivtreatreg on estimation sample)

Bootstrap replications (100)
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
..................................................   50
..................................................   100

Bootstrap results                               Number of obs     =      4,358
                                                Replications      =        100

      command:  ivtreatreg children educ7 age agesq evermarr urban electric tv, hetero(age agesq
                    evermarr urban) iv(frsthalf) model(probit-2sls)
         atet:  e(atet)
        atent:  e(atent)
```

	Observed Coef.	Bootstrap Std. Err.	z	P>|z|	Normal-based [95% Conf.	Interval]
atet	.89829	.6267185	1.43	0.152	-.3300556	2.126636
atent	-.4468834	.458569	-0.97	0.330	-1.345662	.4518952

```
. *The results show that both ATET and ATENT are not signicant and show quite
. *different values, but the values are not far from that of ATE.
```

ATE results for all four models

```
. estimates table probit_ols direct_2sls probit_2sls heckit,b(%9.2f) keep(educ7 G_fv) star
```

Variable	probit_ols	direct_2sls	probit_2sls	heckit
educ7		-1.04	0.30	-1.92***
G_fv	-0.11			

legend: * p<0.05; ** p<0.01; *** p<0.001

The ATE obtained by IV methods is consistently not significant, but it has a positive value for only `probit-2sls`. The rest of the ATEs consistently show negative values–meaning that more-educated women would have been more fertile if they had been less educated. `heckit` is a little more puzzling because the result is significant and very close to the difference-in-mean estimation that is highly suspected as biased. This could be because the identification conditions of heckit are not met in this dataset.

Commands used in this example

```
regress children educ7
estimates store ttest
reg children educ7 age agesq evermarr urban electric tv
ivtreatreg children educ7 age agesq evermarr urban electric tv,hetero(age agesq evermarr urban) iv(f
> rsthalf) model(heckit) graphic
estimates store heckit
display e(ate)
display e(atet)
display e(atent)
display e(N_tot)
display e(N_treat)
display e(N_untreat)
bootstrap atet=e(atet) atent=e(atent), rep(100): ivtreatreg children educ7 age agesq evermarr urban
> electric tv, hetero(age agesq evermarr urban) iv(frsthalf) model(probit-2sls)
ivtreatreg children educ7 age agesq evermarr urban electric tv,hetero(age agesq evermarr urban) iv(f
> rsthalf) model(probit-ols) graphic
estimates store probit_ols
ivtreatreg children educ7 age agesq evermarr urban electric tv,hetero(age agesq evermarr urban) iv(f
> rsthalf) model(direct-2sls) graphic
estimates store direct_2sls
ivtreatreg children educ7 age agesq evermarr urban electric tv,hetero(age agesq evermarr urban) iv(f
> rsthalf) model(probit-2sls) graphic
estimates store probit_2sls
estimates table probit_ols direct_2sls probit_2sls heckit,b(%9.2f) keep(educ7 G_fv) star
```

- Stata's `etregress` command also estimates the ATE in a linear regression model with an endogenous binary treatment variable.
- In contrast to `ivtreatreg`, the `etregress` module assumes a homogenous reaction of potential outcomes to confounders, but it offers the advantage of exploiting a full maximum likelihood approach besides the two-step consistent estimator.
- The basic syntax of this command is:

  `etregress` *depvar* [*indepvars*], `treat`($depvar_t = indepvars_t$) [`twostep`]

  where *depvar* is the outcome; *indepvars* the exogenous covariates explaining the outcome equation; $depvar_t$ is the endogenous treatment; and $indepvars_t$ the confounders explaining the selection equation.

- A powerful approach to deal with endogenous selection without the need for instrumental-variables or additional distributional assumptions.
- DID is suitable where observational data for treated and untreated units are available both before and after treatment.
- Two types of DID estimators depending upon whether the data are a pure longitudinal dataset (panel data) or a repeated cross section.
- In the first case (panel), the same unit (either treated or untreated) is observed before and after a treatment occurred; in the second case (repeated cross section), the units observed before and after treatment (either treated or not) may be different.
- Identication assumptions of both types of DID are, however, the same.

- Card & Krueger (1994) Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *AER*, **90**, 1397-1420.
- Suppose you are interested in the effect of minimum wages on employment (a classic and controversial question in labour economics).
- In a competitive labour market, increases in the minimum wage would move us up a downward-sloping labour demand curve.
  - $\rightarrow$ employment would fall.
- Card and Krueger analyse the effect of a minimum wage increase in New Jersey using a differences-in-differences methodology.

- In February 1992 NJ increased the state minimum wage from \$4.25 to \$5.05. Pennsylvania's minimum wage stayed at \$4.25.
- They surveyed about 400 fast food stores both in NJ and in PA both before and after the minimum wage increase in NJ.
- Define

  Y_{1ist}: employment at restaurant i, state s , time t with a high w^{min}
  Y_{0ist}: employment at restaurant i, state s , time t with a low w^{min}

- In practice of course we only see one or the other.

- Assume that

$$E[Y_{0ist}|s,t] = \gamma_s + \lambda_t \qquad \text{common trend assumption}$$

  where γ_s is a location-specific effect and λ_t is a time-specific effect.
- This assumption simply says that the nontreatment employment time trend in NJ (the treated location) and PA (the nontreated one) are the same. In the absence of a minimum wage change,employment is determined by the sum of a time-invariant state effect, γ_s and a year effect λ_t that is common across states.
- Let D_{st} be a dummy for high-minimum wage states and periods.

- Assuming constant average treatment effect over s and t, that is, the treatment effect is $E[Y_{1ist} - Y_{0ist}|s,t] = \delta$, observed employment can be written as:

$$Y_{ist} = \gamma_s + \lambda_t + \delta D_{st} + \varepsilon_{ist}$$

- The differences-in-differences strategy amounts to comparing the change in employment in NJ to the change in employment in PA.

	New Jersey	Pennsylvania
Employment in February	$E[Y_{ist}\|s = NJ, t = Feb] = \gamma_{NJ} + \lambda_{Feb}$	$E[Y_{ist}\|s = PA, t = Feb] = \gamma_{PA} + \lambda_{Feb}$
Employment in November	$E[Y_{ist}\|s = NJ, t = Nov] = \gamma_{NJ} + \lambda_{Nov} + \delta$	$E[Y_{ist}\|s = PA, t = Nov] = \gamma_{PA} + \lambda_{Nov}$
Nov - Feb	$\lambda_{Nov} - \lambda_{Feb} + \delta$	$\lambda_{Nov} - \lambda_{Feb}$
Diff in Diff	$(\delta + \lambda_{Nov} - \lambda_{Feb}) - (\lambda_{Nov} - \lambda_{Feb}) = \delta$	

- The population differences-in-differences are:

$$E[Y_{ist}|s = NJ, t = Nov] - E[Y_{ist}|s = NJ, t = Feb]$$

$$-(E[Y_{ist}|s = PA, t = Nov] - E[Y_{ist}|s = PA, t = Feb]) = \delta$$

- This is estimated using the sample analog of the population means.

Variable	Stores by state: PA (i)	NJ (ii)	Difference, NJ − PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

- Surprisingly, employment rose in NJ relative to PA after the minimum wage change.

- We can estimate the differences-in-differences estimator in a regression framework.
- Advantages:
  - It is easy to calculate standard errors.
  - We can control for other variables which may reduce the residual variance (lead to smaller standard errors).
  - It is easy to include multiple periods.
  - We can study treatments with different treatment intensity. (e.g. varying increases in the minimum wage for different states).
- The typical regression model that we estimate is:

$$Outcome_{it} = \beta_0 + \beta_1 post_t + \beta_2 treat_i + \beta_3 (treat * post)_{it} + \varepsilon$$

$treat_i$ is a dummy equal to 1 if the observation is in the treatment group, and $post_t$ is a post treatment dummy.

- The estimated coefficients have the following interpretations:
  - $\hat{\beta}_0$: the mean outcome of the control group at the baseline (before treatment).
  - $\hat{\beta}_0 + \hat{\beta}_1$: the mean outcome of the control group in the follow-up (after treatment).
  - $\hat{\beta}_2$: the difference between the treated and the control groups at the baseline.
  - $\hat{\beta}_0 + \hat{\beta}_2$: The mean outcome of the treated group at the baseline.
  - $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$: the mean outcome of the treated group in the follow-up.
  - $\hat{\beta}_3$: the DID estimand.

- In the Card & Krueger case the equivalent regression model would be:

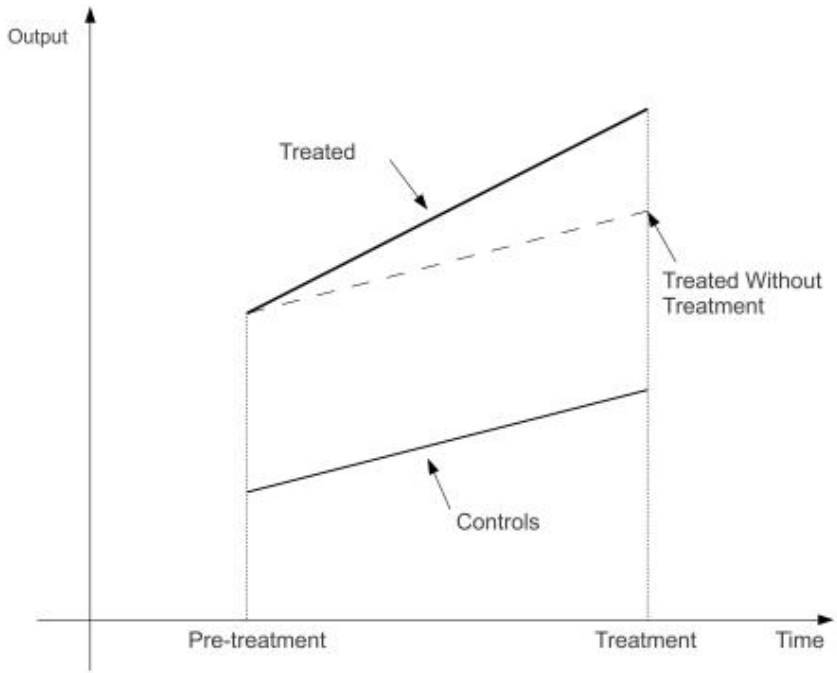
$$Y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \delta(NJ_S \times d_t) + \varepsilon_{ist}$$

  where NJ is a dummy which is equal to 1 if the observation is from NJ, and d is a dummy which is equal to 1 if the observation is from November ($post$).
- The equation takes the following values

  PA Pre: α
  PA Post: $\alpha + \lambda$
  NJ Pre: $\alpha + \gamma$
  NJ Post: $\alpha + \gamma + \lambda + \delta$

- Difference-in-Differences estimate: (NJ Post-NJ Pre)-(PA Post- PA Pre)=δ

- The crucial assumption is the "parallel paths" assumption: the outcome in treatment and control groups would follow the same time trend in the absence of treatment. (Without treatment, the average change for the treated would have been equal to the observed average change for the controls.)



- Common trend assumption is a strong assumption. It's difficult to verify it, but one often uses pre-treatment data to show that the trends are the same.
- Correcting for possible differences in time trends across the treatment and control groups is necessary in order for DID to remain unbiased.
- One way to relax the common trend assumption would be to allow the DID equation to contain a location-specic trend coefficient. But this would require at least three periods; and using just three periods to infer the difference in pre–and post-trend may be questionable.
- A second possibility may be to add covariates as a source of omitted location-specific trends.

- Adding further covariates is a significant advantage of DID compared with other methods.
- Even when the common-trend is not violated, including additional covariates (either t-invariant or s-invariant or unit specific) helps to increase the precision of the ATE's estimation (efficiency) provided, of course, that the model is correctly specied (i.e., the covariates are the correct predictors of outcome's DGP).
- Including leads into the model is another easy way to analyze pre-trends.

- Lags can also be included to analyse whether the treatment effect changes over time after treatment.

$$Y_{ist} = \gamma_s + \lambda_t + \sum_{r=-q}^{-1} \delta_r D_{sr} + \sum_{r=0}^{m} \delta_r D_{sr} + \mathbf{x}'\boldsymbol{\beta} + \varepsilon_{ist}$$

  - treatment occurs in year 0.
  - includes q leads or anticipatory effects.
  - includes m leads or post treatment effects.

- Same units are observed before and after the treatment.
- Suppose we have data for two points in time ($t = (0, 1)$ as in the crosssection case.
- DID is defined as the OLS estimator of α in the following regression:

$$\begin{cases} t = 1: & Y_{i1} = \mu_1 + \alpha D_{i1} + u_{i1} \\ t = 0: & Y_{i0} = \mu_0 + \alpha D_{i0} + u_{i0} \\ D_{i0} = 0 & \end{cases}$$

where estimation is only carried out for those units which are untreated in $t = 0$. By subtracting, we then obtain:

$$\begin{cases} \Delta Y_{it} & = \mu + \alpha \Delta D_{i1} + \Delta u_{i1} \\ D_{i0} = 0 \end{cases}$$

with $\mu = \mu_1 - \mu_0$, which is equivalent to

$$\begin{cases} \Delta Y_{it} & = \mu + \alpha D_{i1} + \Delta u_{i1} \\ D_{i0} = 0 \end{cases}$$

- As in the repeated cross-section case, one can control also for time and individual effects in order to preserve exogeneity.
- DID with panel data can also be easily extended to the case of dynamic treatment by introducing lags and leads as we did in the cross-section case.

- Although it can easily be done manually, it is best to use the user-written command `diff`.[2]

```
diff outcome_var [if] [in] [weight], period(varname) treated(varname)
    [cov(varlist) kernel id(varname) bw(#) ktype(kernel) rcs qdid(quantile)
    pscore(varname) logit support addcov(varlist) cluster(varname) robust
    bs reps(int) test report nostar export(filename)]
```

- Data from Card & Krueger (1994).

[2] Villa, J.M. (2016) diff: Simplifying the estimation of difference-in-differences treatment effects. *Stata Journal*, **16**, 52-71.

```
. use cardkrueger1994.dta, clear
(Sample dataset from Card and Krueger (1994))

. des

Contains data from cardkrueger1994.dta
  obs:           780                          Sample dataset from Card and Krueger (
 vars:             8                          12 Mar 2014 14:03
 size:        11,700
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
id                int     %8.0g                 Store ID
t                 byte    %8.0g                 Feb. 1992 = 0; Nov. 1992 = 1
treated           long    %8.0g      treated    New Jersey = 1; Pennsylvania = 0
fte               float   %9.0g                 Output: Full Time Employment
bk                byte    %8.0g                 Burger King == 1
kfc               byte    %8.0g                 Kentucky Fried Chicken == 1
roys              byte    %8.0g                 Roy Rogers == 1
wendys            byte    %8.0g                 Wendy´s == 1
-------------------------------------------------------------------------------
Sorted by: id  t  treated
```

```
. l in 1/15
```

	id	t	treated	fte	bk	kfc	roys	wendys
1.	1	0	NJ	31	1	0	0	0
2.	1	1	NJ	40	1	0	0	0
3.	2	0	NJ	13	1	0	0	0
4.	2	1	NJ	12.5	1	0	0	0
5.	3	0	NJ	12.5	0	1	0	0
6.	3	1	NJ	7.5	0	1	0	0
7.	4	0	NJ	16	0	0	1	0
8.	4	1	NJ	20	0	0	1	0
9.	5	0	NJ	20	0	0	1	0
10.	5	1	NJ	25	0	0	1	0
11.	6	0	NJ	3	0	0	1	0
12.	6	1	NJ	6	0	0	1	0
13.	9	0	NJ	32	1	0	0	0
14.	9	1	NJ	16	1	0	0	0
15.	10	0	NJ	25	1	0	0	0

```
. su id t treated fte bk kfc roys wendys
```

Variable	Obs	Mean	Std. Dev.	Min	Max
id	780	247.2641	148.644	1	522
t	780	.5	.5003208	0	1
treated	780	.8051282	.3963561	0	1
fte	780	17.58109	9.095066	0	80
bk	780	.4179487	.4935381	0	1
kfc	780	.2051282	.4040544	0	1
roys	780	.2435897	.4295233	0	1
wendys	780	.1333333	.3401528	0	1

```
. *single DID with no covariates
. diff fte, treated(treated) period(t)
```

DIFFERENCE-IN-DIFFERENCES ESTIMATION RESULTS

Number of observations in the DIFF-IN-DIFF: 780

	Baseline	Follow-up	
Control:	76	76	152
Treated:	314	314	628
	390	390	

Outcome var.	fte	S. Err.	t	P>|t|
Baseline				
Control	20.013			
Treated	17.069			
Diff (T-C)	-2.944	1.160	-2.54	0.011**
Follow-up				
Control	17.523			
Treated	17.518			
Diff (T-C)	-0.005	1.160	-0.00	0.997
Diff-in-Diff	2.939	1.641	1.79	0.074*

R-square: 0.01

- Means and Standard Errors are estimated by linear regression

**Inference: *** p<0.01; ** p<0.05; * p<0.1

```
. *single DID with covariates
. diff fte, treated(treated) period(t) cov(bk kfc roys)
DIFFERENCE-IN-DIFFERENCES WITH COVARIATES

DIFFERENCE-IN-DIFFERENCES ESTIMATION RESULTS
Number of observations in the DIFF-IN-DIFF: 780
            Baseline       Follow-up
   Control: 76             76           152
   Treated: 314            314          628
            390            390
```

Outcome var.	fte	S. Err.	t	P>|t|
Baseline				
Control	21.342			
Treated	19.003			
Diff (T-C)	-2.339	1.052	-2.22	0.026**
Follow-up				
Control	18.852			
Treated	19.452			
Diff (T-C)	0.600	1.052	0.57	0.569
Diff-in-Diff	2.939	1.485	1.98	0.048**

```
R-square:    0.19
- Means and Standard Errors are estimated by linear regression
**Inference: *** p<0.01; ** p<0.05; * p<0.1
```

- DID estimators require the parallel paths assumption when there is only one pretreatment period.
- When several pretreatment periods are available, the assumption equivalent to parallel paths is referred to as *common trends*.
- If there are pretreatment trend differentials, it is customary to adjust the econometric specification to try to accommodate for those differences by including linear trends.
- But a more general model would be more appropriate.
- `didq` is a user-written command that allows one to make different assumptions. (Mora, R. & I. Reggio (2015) didq: A command for treatment-effect estimation under alternative assumptions. *Stata Journal*, **15**, 796-808.)

- A special situation where the selection into treatment is highly determined by the level assumed by a specific variable s (called "forcing" variable), defining a threshold $\bar{s}$ separating treated and untreated units. (Usually due to institutional or logical structures.)
- Generally, RD designs exploit discontinuities in policy assignment.
- For example, there might be an age threshold at which one becomes eligible for pension plan vesting, or an income threshold at which one becomes eligible for financial aid.
- One assumes that units just on different sides of the discontinuity are essentially the same in unobservables that affect the relevant outcome.
- The treatment statuses of the two groups differ, because of the institutional setup, in which case differences in outcomes can be attributed to the different treatment statuses.

- Two types of RD design:
  - **Sharp RD**: when the relation between treatment and the forcing variable is *deterministic*, thus creating a strict "jump" in the probability of receiving treatment at the threshold
  - **Fuzzy RD**: when this relation is *stochastic*, thus producing a milder jump at the threshold

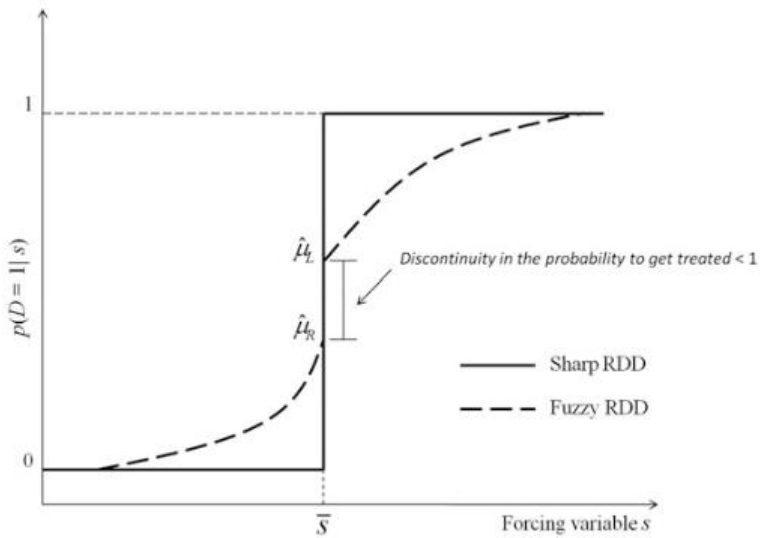


Fig. 4.2 Discontinuity in the probability to be treated in the sharp and fuzzy RDD

- The idea behind RDD is that, in a neighborhood of the threshold, conditions for a natural experiment (i.e., a random assignment to treatment) are restored. Therefore, as long as: (1) the threshold is well identified and (2) the treatment is dependent on s, the analyst can obtain the policy effect simply by comparing the mean outcome of individuals laying on the left and the mean outcome of individuals laying on the right of the threshold.
- The identified effect is a **local average treatment effect (LATE)**.

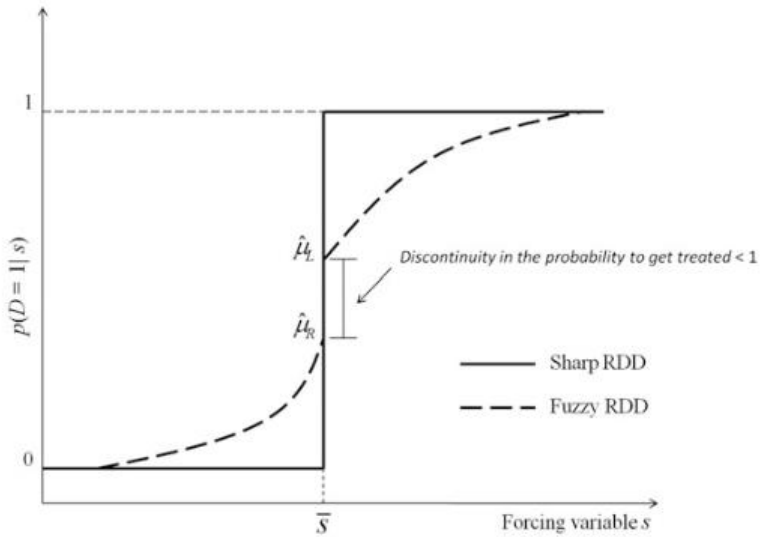
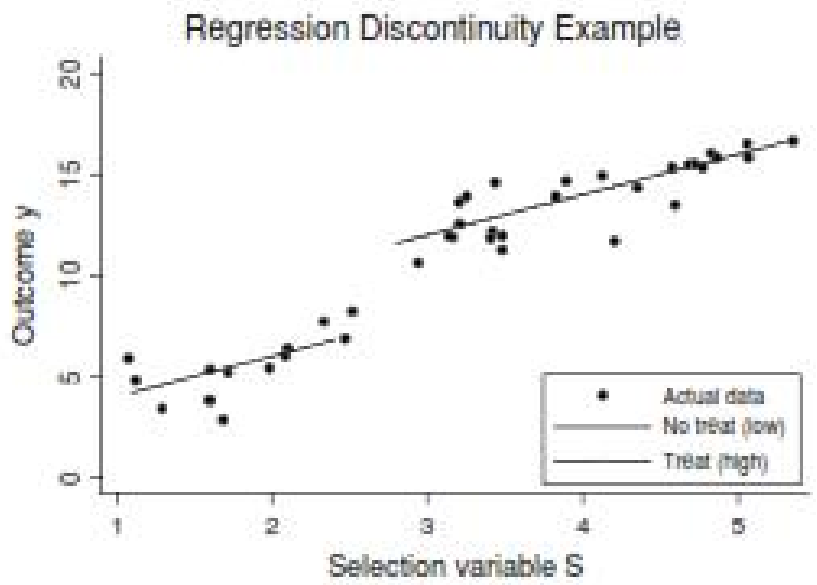


Fig. 4.2 Discontinuity in the probability to be treated in the sharp and fuzzy RDD

- Among quasi-experimental methods, RD techniques are considered to have the highest internal validity (the ability to identify causal relationships in the current research setting). Their external validity (ability to generalize findings to similar contexts) may be less impressive, as the estimated treatment effect is local to the discontinuity.
- There are four crucial elements to a RD design:
  - Treatment is not randomly assigned, but dependent at least in part on an observable assignment variable S.
  - There is a discontinuity at some cutoff value of the assignment variable in the level of treatment.
  - Individuals cannot manipulate their status to affect whether they fall on one side of the cutoff or the other. Those near the cutoff are assumed to be exchangeable or otherwise identical.
  - Other variables are smooth functions of the assignment variable, conditional on treatment. That is, the only reason the outcome variable should jump at the cutoff is due to the discontinuity in the level of treatment. (Note this differs from IV, in that the assignment variable can have a direct impact on the outcome, not just on the treatment, though not a discontinuous impact.)

- Individuals are assigned to treatment and control groups solely on the basis of an observed continuous measure S, the selection. Those falling below the distinct cutoff S do not receive treatment and constitute the control group whereas those that are above the cutoff receive treatment ($D = 1$).
- That is, the treatment assignment occurs through a known and measured deterministic decision rule: $D_i = 1[S_i \geq \bar{S}]$.



$$E[u|D,S] = E[u|S]$$

where u is the error in the outcome equation. Because S is the only systematic determinant of D, S will capture any correlation between D and u.

- With $D_i = D(S_i) = 1[S_i \geq \bar{S}]$, a dependence between D_i and u_i would make OLS an inconsistent estimator of α.
- So, we specify and include the conditional mean function $E[u|D,S]$ as a 'control function' in the outcome equation:
- Outcome equation:

$$y_i = \beta + \alpha D_i + k(S_i) + \varepsilon_i$$

where $\varepsilon_i = y_i - E[y_i|D_i,S_i]$.

- If $k(S)$ is correctly specified, the regression will consistently estimate α.
- If $k(S)$ is linear then α will be estimated by the distance between the two linear parallel regression lines at the cutoff point, which in this case equals the difference between the two intercepts. It is an unbiased estimate of the common treatment effect if the control function is linear.

- In the "sharp" or "deterministic assignment" version, the estimated treatment effect is just the jump in expected outcomes at the cutoff: in other words, the expected outcome for units just above the cutoff (who get treated), call this y^+, minus the expected outcome for units just below the cutoff (who don't get treated, but are supposed to be otherwise identical), call this y^-, or

$$\widehat{LATE} = (y^+ - y^-)$$

since the jump in the level of treatment is exactly one unit at the cutoff.

- In the " fuzzy" or " assignment" version, the jump in outcomes is "caused" by some jump in treatment that need not be one. Then we just form the ratio of the jump in outcomes to the jump in treatment. The Local Wald Estimator of LATE is thus $(y^+ - y^-)/(x^+ - x^-)$, where $(x^+ - x^-)$ is the estimated discontinuous jump in expected treatment.
- Note that this second estimator reduces to the first given "deterministic assignment" since $(x^+ - x^-) = 1$ in this case, so the distinction between "sharp" and " fuzzy" RD is not too sharp.

Polynomial or Local Polynomial in S

- There is a great deal of art involved in the choice of some continuous function of the assignment variable S for treatment and outcomes.
- A high-order polynomial of S is often used to estimate separately on both sides of the discontinuity. Better yet, a local polynomial, local linear model, or local mean smoother may be used, where one must choose a kernel or bandwidth parameter. The art is in the choice of these.

- The `votex` dataset contains information for 349 of the 435 Congressional districts in the 102nd US Congress. `lne` is the logarithm of Federal expenditures in the district (evidence of the member of Congress 'bringing home the bacon'.) Variable `d` is the Democratic vote share minus 0.5, so that it is positive for Democratic districts and negative for Republican districts.
- Having a Democratic representative in the US Congress may be considered a treatment applied to a Congressional district, and the assignment variable S is the vote share garnered by the Democratic candidate. At $S = 50\%$, the probability of treatment=1 jumps from zero to one.
- Suppose we are interested in the effect a Democratic representative has on the federal spending within a Congressional district.
- User-written package `rd`.

Brief Example in Stata

```
. use "\\commerce\obsUser01\mgenc\Documents\COURSES 2017\ECON413\Slides\votex.dta", clear
(102nd Congress)
. des
Contains data from \\commerce\obsUser01\mgenc\Documents\COURSES 2017\ECON413\Slides\votex.dta
  obs:           349                          102nd Congress
 vars:            19                          23 May 2017 00:47
 size:        37,692
```

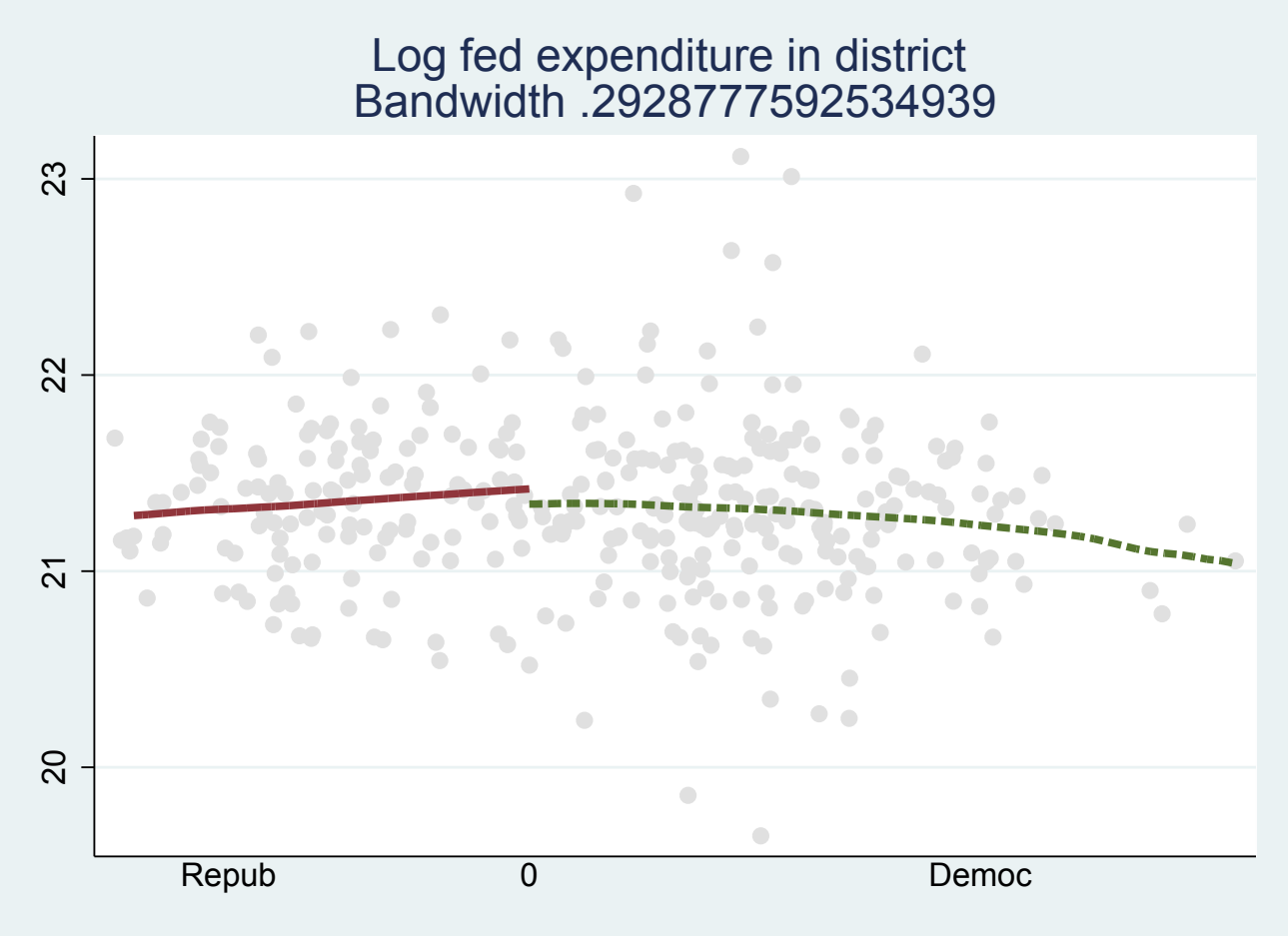
variable name	storage type	display format	value label	variable label
fips	byte	%8.0g	fips	State code
district	byte	%8.0g		Congr district
d	double	%10.0g		Dem vote share minus .5
win	byte	%9.0g		Dem Won Race
lne	float	%9.0g		Log fed expenditure in district
i	byte	%9.0g		Incumbent
votingpop	long	%12.0g		Voting Age Population
votpop	double	%10.0g		Voting Age Population Share
populatn	long	%12.0g		Population
black	double	%12.0g		Black Population Share
blucllr	double	%12.0g		Blue-collar Population Share
farmer	double	%12.0g		Farmer Population Share
fedwrkr	double	%12.0g		Fed Worker Population Share
forborn	double	%12.0g		Foreign Born Population Share
manuf	double	%12.0g		Manufactur Population Share
unemplyd	double	%12.0g		Unemp Population Share
union	float	%9.0g		Unionized Population Share
urban	double	%12.0g		Urban Population Share
veterans	double	%12.0g		Veteran Population Share

```
Sorted by: fips  district
```

```
. rd lne d, gr mbw(100) line(`"xla(-.2 "Repub" 0 .3 "Democ", noticks)"´)
Two variables specified; treatment is
assumed to jump from zero to one at Z=0.

 Assignment variable Z is d
 Treatment variable X_T unspecified
 Outcome variable y is lne

Command used for graph: lpoly; Kernel used: triangle (default)
Bandwidth: .29287776; loc Wald Estimate: -.07739553
Estimating for bandwidth .2928777592534939
```

lne	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwald	-.0773955	.1056062	-0.73	0.464	-.28438	.1295889

The estimate of the LATE is not signicantly different from zero in this case. Interestingly, as we move from 'safe' seats in either party toward the contested boundary, expenditures modestly increase.


Log fed expenditure in district
Bandwidth .2928777592534939
23
22
21
20
Repub
0
Democ

- None of these methods is perfect. The gold standard, an RCT, has the best internal validity but may have poor external validity. Of methods using observational data, the RD design is closest to an RCT, and also has high internal validity but low external validity. IV methods can eliminate bias from selection on unobservables in the limit, but may have very poor performance in finite samples. The hypothetical internal validity of IV is high, but the practical internal validity of IV is often low, and the external validity not much greater than RD.
- Matching and reweighting methods can eliminate bias due to selection on observables, and give efficient estimates of many types of treatment effects in many settings, but it is rarely the case that selection depends only on observables, in which case matching can actually exacerbate bias. Regression or matching methods applied to population data often have very high external validity, but internal validity that is often questionable.

- In practice, the data often dictate the method. If one has access to experimental data, one worries less about selection (though IV is often used to correct for selection of treatment status contrary to assignment). Given observational data, if one can find a discontinuity in expected treatment with respect to an observable assignment variable, one uses RD; if one can conceive of plausible excluded instruments, one uses IV. In the absence of these features of the data, repeated measures may used to control for invariant unobservables, or observations may be matched on observables.