Causal Analysis

Impact Evaluation and Causal Machine Learning with Applications in R

Chapter 4: Selection on Observables (1)

Table of Contents

Motivation

- A large share of empirical analyses is based on observational rather than experimental data, e.g., from:
  - Surveys (e.g., an online survey among customers).
  - Company data (e.g., product features and sales in stores).
  - Administrative data (e.g., information on labor market performance and public transfer payments).
- Observational data often contain (one or more) outcomes, covariates, and a treatment, which is not randomly assigned.
- In this case, the independence assumption $\{Y(1), Y(0)\} \perp D$ is generally implausible.
- Thus, a simple comparison of mean outcomes between treated and untreated groups is inadequate for assessing the ATE.

Selection on Observables (1)

Conditional independence/selection on observables

Observed covariates are rich enough to control for all confounders.
$\Rightarrow$ After conditioning on X, D is as good as randomly assigned.

- Assumption is satisfied if:
  - We directly observe all covariates with an effect on both the treatment and the outcome, or
  - Controlling for the covariates blocks the effects of unobserved confounders on the treatment, the outcome, or both.
- The plausibility of this assumption has to be scrutinized based on theory, domain knowledge, or prior empirical findings.

Selection on Observables (2)

Common support

For any combination of covariate values occurring in the population, there are both treated and nontreated subjects.

- Rules out that covariates deterministically predict treatment.
- Implies that the **propensity score** $p(X) = \Pr(D = 1|X)$ (i.e., the conditional treatment probability) is between zero and one.

No posttreatment covariates

Covariates are measured at or prior to treatment assignment and are thus not affected by the treatment.

- Otherwise, controlling for them may condition away part of the treatment effect or introduce collider bias.

Selection on Observables (3)

- Formal assumptions for selection on observables:

$$\{Y(1), Y(0)\} \perp D|X, \qquad 0 < p(X) < 1, \qquad X(1) = X(0) = X \tag{4.1}$$

Conditional independence | Common support | No posttreatment covariates

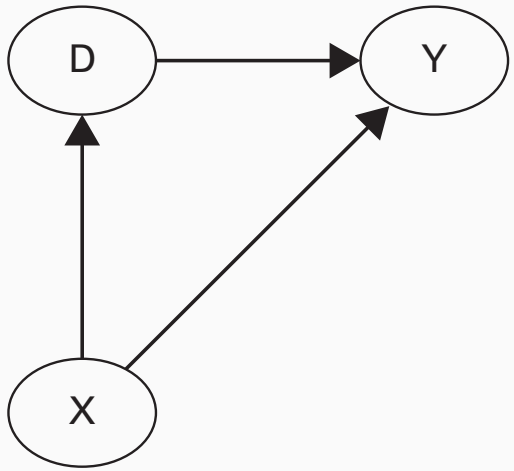


Figure 4.1: Selection on observables

- Conditional on covariates X, no unobserved variables influence both treatment D and outcome Y.

Selection on Observables for the ATET

When identifying the ATET (rather than the ATE), the assumptions on the previous slide may be relaxed:

- Conditional independence: $Y(0) \perp D|X$
  - Applies only to the potential outcomes under nontreatment.
- Common support: $p(X) < 1$
  - We only need to find nontreated individuals who are comparable to the treated.
  - Not necessary to find comparable treated observations for every nontreated observation.
  - Implies that $p(X)$ may be zero for some values of X (i.e., only nontreated observations exist with such values).

Conditional Average Treatment Effect (CATE)

- The conditional mean outcome given that treatment D is equal to $d \in \{0, 1\}$ and $X = x$ is denoted by:

$$\mu_d(x) = E[Y|D = d, X = x]$$

- Under the conditional independence assumption, $\mu_1(x) - \mu_0(x)$ identifies the causal effect among subjects with covariates $X = x$:

$$\Delta_x = E[Y(1)|X = x] - E[Y(0)|X = x] = \mu_1(x) - \mu_0(x) \tag{4.2}$$

- Δ_x is the conditional average treatment effect (CATE).
- The ATE is identified by averaging CATEs across all values of x, which the covariates X take in the population:

$$\Delta = E[\mu_1(X) - \mu_0(X)] \tag{4.3}$$

ATET and ATENT

- In the selection-on-observables framework, treated and nontreated groups may differ in X, and thus in causal effects.
- Effects for subpopulations, such as the treated, may be more relevant than the ATE if not everyone can/should be treated.
- To identify the ATET, the CATEs are averaged across the covariate values x appearing among the treated population:

$$\Delta_{D=1} = E[\mu_1(X)|D=1] - E[\mu_0(X)|D=1] = E[Y|D=1] - E[\mu_0(X)|D=1] \quad (4.4)$$

  - The second equality follows from the law of iterated expectations: $E[\mu_1(X)|D=1] = E[E[Y|D=1,X]|D=1] = E[Y|D=1]$

- The ATENT is identified analogously:

$$\Delta_{D=0} = E[\mu_1(X)|D=0] - E[\mu_0(X)|D=0] = E[\mu_1(X)|D=0] - E[Y|D=0] \quad (4.5)$$

Table of Contents

Linear Regression (1)

- To identify the CATE, we may be tempted to define the following linear regression model for $E[Y(D)|X] = E[Y|D,X] = \mu_D(X)$:

$$\mu_D(X) = \alpha + \beta_D D + \beta_{X_1} X_1 + \cdots + \beta_{X_K} X_K \tag{4.6}$$

- OLS estimate:

$$\hat{\beta}_D = \frac{\widehat{Cov}(Y_i, D_i|X_i)}{\widehat{Var}(D_i|X_i)} \tag{4.7}$$

- Problem: In a selection-on-observables framework, D and X are generally correlated.
- Implications of $Cov(D,X) \neq 0$:
  - Larger variance of $\hat{\beta}_D$.
  - If the relationship between Y and X is misspecified, this bias spills over to the estimation of β_D: $\hat{\beta}_D$ is biased and inconsistent.

Linear Regression (2)

- Sources of misspecification:
  - Omission of interaction terms between covariates (e.g., $X_1 \cdot X_2$).
  - Omission of higher-order terms (e.g., X_1^2) capturing nonlinear relationships.
- If treatment effects differ across X, omitting interactions between D and X leads to a biased and inconsistent estimate of the CATE.
- In the linear model, $E[\beta_D] = E[\Delta_X] = \Delta$.
- This implies that average effects are homogeneous across values of X: CATE = ATE = ATET.
- Such homogeneity is often implausible in empirical applications.
- Prefer methods that allow for heterogeneous effects across X, unless strong prior knowledge suggests homogeneous effects.

Series Regression (1)

- To make the model more flexible, add interaction terms and higher-order terms as additional regressors:

$$\mu_D(X) = \alpha + \beta_D D + \beta_{X_1} X_1 + \cdots + \beta_{X_K} X_K + \beta_{D,X_1} D X_1 + \ldots + \beta_{D,X_k} D X_K + \beta_{X_1^2} X_1^2 + \cdots + \beta_{X_1 X_2} X_1 X_2 + \ldots \tag{4.8}$$

- Estimate $\hat{\mu}_1(X)$ by setting $D = 1$ and $\hat{\mu}_0(X)$ by setting $D = 0$.
- Compute the ATE by averaging the CATEs in the sample:

$$\hat{\Delta} = \frac{1}{n} \sum_{i=1}^{n} [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)] \tag{4.9}$$

- Compute the ATET by averaging the CATEs in the subsample of treated observations:

$$\hat{\Delta}_{D=1} = \frac{1}{n_1} \sum_{i:D_i=1} [\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)] \tag{4.10}$$

Series Regression (2)

- Direct estimation of the ATE that avoids the two-step procedure of computing CATEs and averaging (Imbens and Wooldridge, 2009):

$$\hat{\mu}_D(X) = \hat{\alpha} + \underbrace{\hat{\beta}_D}_{\hat{\Delta}} D_i + \beta_{X_1} X_{i1} + \cdots + \beta_{X_K} X_{iK} + \beta_{D,X_1} D_i \cdot (X_{i1} - \bar{X}_1) + \cdots$$
$$+ \beta_{D,X_k} D_i \cdot (X_{iK} - \bar{X}_K) + \beta_{X_1^2} X_1^2 + \cdots + \beta_{X_1 X_2} X_1 \cdot X_2 + \cdots \quad (4.11)$$

- Or run two separate regressions for $D = 1$ and $D = 0$ without including interaction terms between D and X:

$$\mu_1(X) = \alpha_1 + \beta_{X_1,1} X_1 + \cdots + \beta_{X_K,1} X_K + \beta_{X_1^2,1} X_1^2 + \cdots + \beta_{X_1 X_2,1} X_1 \cdot X_2 + \cdots ,$$
$$\mu_0(X) = \alpha_0 + \beta_{X_1,0} X_1 + \cdots + \beta_{X_K,0} X_K + \beta_{X_1^2,0} X_1^2 + \cdots + \beta_{X_1 X_2,0} X_1 \cdot X_2 + \cdots \quad (4.12)$$

- Use these to obtain $\hat{\mu}_1(X)$ and $\hat{\mu}_0(X)$ and compute the ATE or ATET based on averaging appropriately.

Leave-One-Out Cross-Validation (1)

- Choosing the number of interaction and higher-order terms involves a trade-off:
  - Including too few terms may induce a bias in treatment effect estimation due to poor approximation of $\mu_D(X)$.
  - Including too many terms with little or no influence on $\mu_D(X)$ may increase the variance due to overfitting.
- The problem with overfitting is that $\hat{\mu}_D(X)$ also captures part of the sample-specific error terms, leading to poor generalization.
- The goal is to find a specification that optimally balances bias and variance by minimizing the overall estimation error (MSE).
- Leave-one-out cross-validation can be used for this purpose.

Leave-One-Out Cross-Validation (2)

- To avoid overfitting, leave observation i out when estimating $\hat{\mu}_{1,-i}(X_i)$, and then compute the squared residual for i.
- Use this leave-one-out estimation for all treated observations and sum up the squared residuals:

$$\sum_{i:D_i=1} [Y_i - \hat{\mu}_{1,-i}(X_i)]^2 \tag{4.13}$$

- Repeat procedure for different model specifications in terms of p (number of interaction and higher-order terms).
- Select the one that minimizes the sum of squared residuals:

$$p_{\text{opt}} = \arg\min_{p \in P} \sum_{i:D_i=1} [Y_i - \hat{\mu}_{1,-i,p}(X_i)]^2 \tag{4.14}$$

- Finally, use optimal p_{opt} to estimate $\mu_1(X_i)$ in the full sample:

$$\hat{\mu}_1(X_i) = \hat{\mu}_{1,p_{\text{opt}}}(X_i) \tag{4.15}$$

Leave-One-Out Cross-Validation (3)

- The leave-one-out cross-validation procedure can also be applied to the estimation of $\mu_0(X)$.
- The number of terms in p_{opt} depends on the number of observations and tends to grow as the sample size increases.
- Reason: Variance tends to decrease in larger samples, allowing for greater model flexibility to reduce bias.
- However, the optimal number of terms grows at a slower pace than the sample size.
- Otherwise, the bias would be reduced at a faster rate than the variance, which is not optimal for minimizing the MSE.

Global vs. Local Estimation Methods

- Global methods (e.g., linear or series regression) estimate $\mu_D(X)$ for any covariate value X, even where no data is observed.
- However, predictions may be quite poor for X values far beyond the observed data.
- In contrast, local methods do not permit such predictions.
- Idea: Estimate $\mu_1(x)$ as a local average of treated observations with values of X close to x (i.e., within a bandwidth h around x):

$$\hat{\mu}_{1,h}(x) = \frac{\sum_{i:D_i=1} I\{|X_i - x| \leq h\} \cdot Y_i}{\sum_{i:D_i=1} I\{|X_i - x| \leq h\}} \tag{4.16}$$

  - $\sum_{i:D_i=1} I\{|X_i - x| \leq h\}$: Number of treated observations with all covariate values within the bandwidth.

- $\hat{\mu}_{1,h}(x)$ depends on the specific choice of the bandwidth h.
- All observations within the bandwidth get the same weight.

Local Constant Kernel Regression

- A kernel function gives more weight to observations within the bandwidth whose covariate values are closer to x.
- Local constant kernel regression or Nadaraya Watson (1964) estimator:

$$\hat{\mu}_{1,h}(x) = \frac{\sum_{i:D_i=1}^{n} \mathcal{K}\left(\frac{X_i - x}{h}\right) \cdot Y_i}{\sum_{i:D_i=1}^{n} \mathcal{K}\left(\frac{X_i - x}{h}\right)} \tag{4.17}$$

- Kernel function $\mathcal{K}(a)$ (with a being a specific value for $\frac{X_i - x}{h}$) is assumed to satisfy the following conditions:

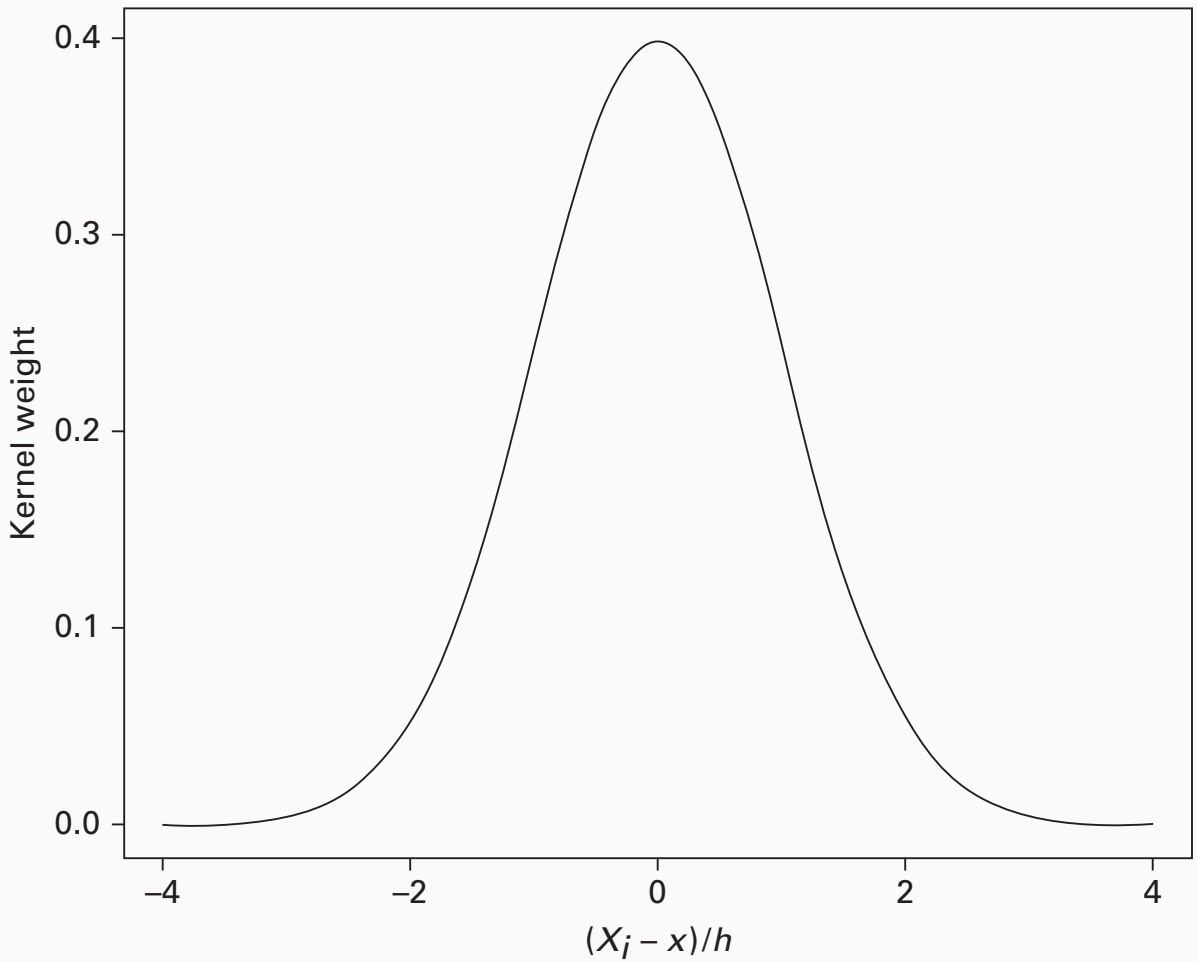
$$\int \mathcal{K}(a)da = 1 \qquad \int a\mathcal{K}(a)da = 0 \qquad \int a^2\mathcal{K}(a)da < \infty$$

Integrates to 1 | Symmetric around zero | Bounded second order

- Examples of kernel functions satisfying these properties: Standard normal kernel, Epanechnikov kernel, triangular kernel.

Illustration of Kernel Weights

Figure 4.2: Standard normal kernel function



Bandwidth Choice (1)

- Kernel functions assign greater weight to observations for which $\frac{X_i - x}{h}$ is close to zero.
- The weight is greater when X_i is close to x.
- The bandwidth h determines by how much the kernel weight depends on the absolute difference of $X_i - x$:
  - When h is large, weights are more uniform and less dependent on $X_i - x$.
  - When h is small, only observations with X_i close to x receive a nonnegligible weight.
- Choosing h involves a bias–variance trade-off:
  - A large h may entail substantial bias by giving large weights to observations whose value X_i is far from x.
  - A small h may entail high variance by giving weight to very few observations (overfitting if their errors ε_i do not average out).

Bandwidth Choice (2)

- Use leave-one-out cross-validation for finding the optimal bandwidth h_{opt} among a range of candidate values H:

$$h_{\text{opt}} = \arg\min_{h \in H} \sum_{i:D_i=1} [Y_i - \hat{\mu}_{1,-i,h}(X_i)]^2 \tag{4.18}$$

- Then estimate $\mu_1(X_i)$ using the full sample and the optimal bandwidth h_{opt}:

$$\hat{\mu}_1(X_i) = \hat{\mu}_{1,h_{\text{opt}}}(X_i) \tag{4.19}$$

- h_{opt} tends to decrease as the sample size grows to reduce the bias from relying on observations with X values too far from x.
- However, h should decrease at a slower pace than the growth of the sample size.
- Otherwise, the variance would be too large relative to the bias, which would not be optimal for minimizing the MSE of $\hat{\mu}_1(X_i)$.

Local Linear Regression

- Local linear regression combines kernel-based weighting and regression.
- Idea: Run a weighted linear regression of Y_i on X_i within treatment groups.
  - The weight of each observation corresponds to $\mathcal{K}\left(\frac{X_i - x}{h}\right)$.
  - Observations with X_i close to x receive more weight.
- Permits estimating regression coefficients that are specific to the covariate value x at which the CATE is to be computed.
- Compared to local constant regression, local linear regression generally has a smaller bias at the boundaries of the data.

Convergence Rates of Estimators

- Kernel-based estimates of $\mu_1(X)$, $\mu_0(X)$, and the CATE converge slower than the fastest possible convergence rate of $\frac{1}{\sqrt{n}}$.
- Reason: Kernel regression strongly depends on a subset of observations with covariate values close to x.
- In contrast, linear regression achieves $\frac{1}{\sqrt{n}}$ by exploiting the entire sample, but at the price of imposing linearity.
- Even though the CATE converges slower, the estimation of the ATE and ATET can be $\sqrt{n}$-consistent under specific conditions.
- Reason: Averaging over many CATEs with different values of x may average out the estimation errors at a specific x.
- Series regression estimates of $\mu_1(X)$, $\mu_0(X)$, and the CATE also converge slower than $\frac{1}{\sqrt{n}}$.

Table of Contents

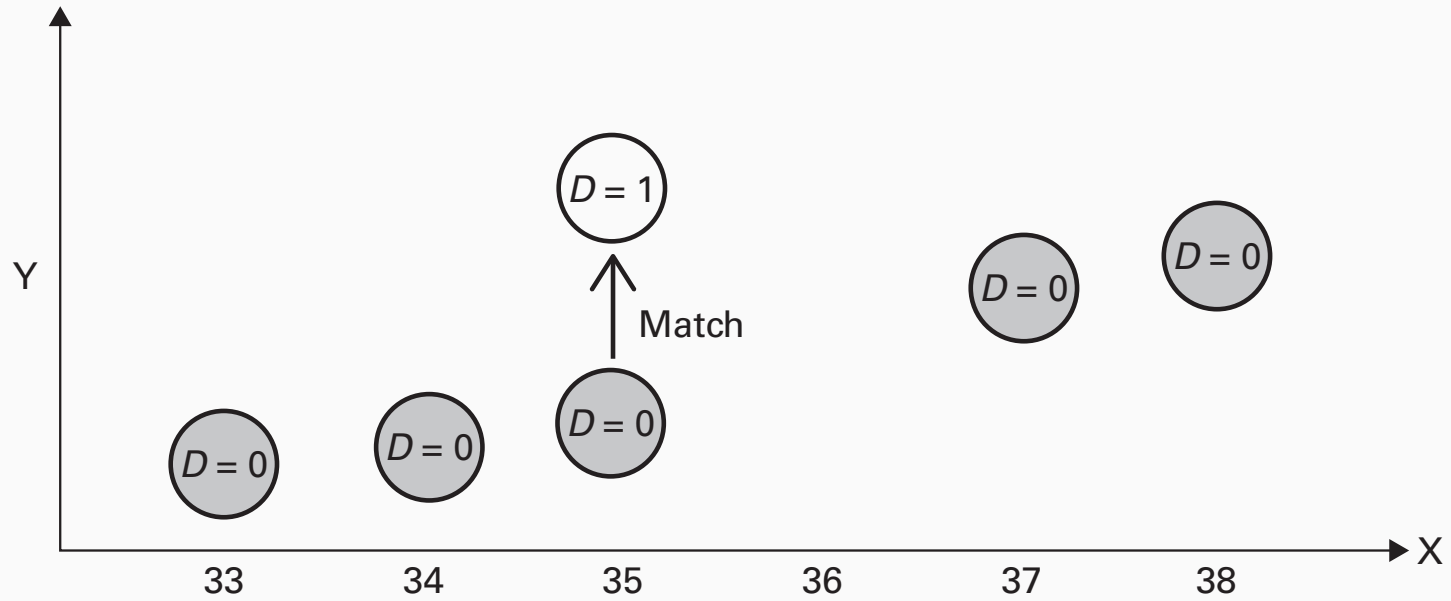
Motivation of Matching

- Idea of matching: Find and match treated and nontreated observations with similar (or, ideally, identical) covariate values.
- See e.g. Heckman, Ichimura, and Todd (1998), Heckman, Ichimura, Smith, and Todd (1998), Dehejia and Wahba (1999), and Lechner, Miquel, and Wunsch (2011).
- Goal: Create treated and nontreated groups that are comparable in their covariate distributions.
- Matching with or without replacement: Can an observation serve as a match only once or multiple times?
- Matching without replacement reduces variance by not reusing observations but may increase bias due to lower match quality.
- Matching with replacement reduces bias by finding the closest possible match; preferred when potential matches are limited.
- In what follows, we focus on matching with replacement.

Pair Matching (1)

- Pair matching: For each observation in one treatment group, find the best match in the other treatment group in terms of X.

Figure 4.3: Pair matching



- In this example, X consists of only one covariate, like age.

Pair Matching (2)

- Formal definition of ATET and ATENT estimates via pair matching:

$$\hat{\Delta}_{D=1} = \frac{1}{n_1} \sum_{i:D_i=1} \{Y_i - \sum_{j:D_j=0} I\{||X_j - X_i|| = \min_{l:D_l=0} ||X_l - X_i||\} Y_j\} \tag{4.20}$$

$$\hat{\Delta}_{D=0} = \frac{1}{n_0} \sum_{i:D_i=0} \{\sum_{j:D_j=1} I\{||X_j - X_i|| = \min_{l:D_l=1} ||X_l - X_i||\} Y_j - Y_i\} \tag{4.21}$$

  - $||X_j - X_i||$ measures the distance between vectors X_j and X_i.

- By the law of total probability, the ATE can be expressed as:

$$\Delta = \Pr(D = 1) \cdot \Delta_{D=1} + \Pr(D = 0) \cdot \Delta_{D=0} \tag{4.22}$$

- The ATE is thus estimated as a weighted average of ATET and ATENT, using the shares of both groups in the sample as weights:

$$\hat{\Delta} = \frac{n_1}{n} \cdot \hat{\Delta}_{D=1} + \frac{n_0}{n} \cdot \hat{\Delta}_{D=0} \tag{4.23}$$

Distance Metrics (1)

How should we define the distance metric $||X_j - X_i||$?

- **Euclidean distance:**

$$||X_j - X_i||_{\text{Euclidean}} = \sqrt{\sum_{k=1}^{K}(X_{jk} - X_{ik})^2} \tag{4.24}$$

  - Sum the squared differences across all covariates k and select the observation j with the smallest overall distance.
  - Problem: A specific difference (e.g., of 1) is considered equally important for each covariate X_k, independent of its distribution.

- **Standardized Euclidean distance:**

$$||X_j - X_i||_{\text{Variance}} = \sqrt{\sum_{k=1}^{K}\frac{(X_{jk} - X_{ik})^2}{\widehat{Var}(X_k)}} \tag{4.25}$$

  - Normalizes any covariate difference between j and i based on the inverse of the sample variance of the respective covariate.

Distance Metrics (2)

- Mahalanobis distance:

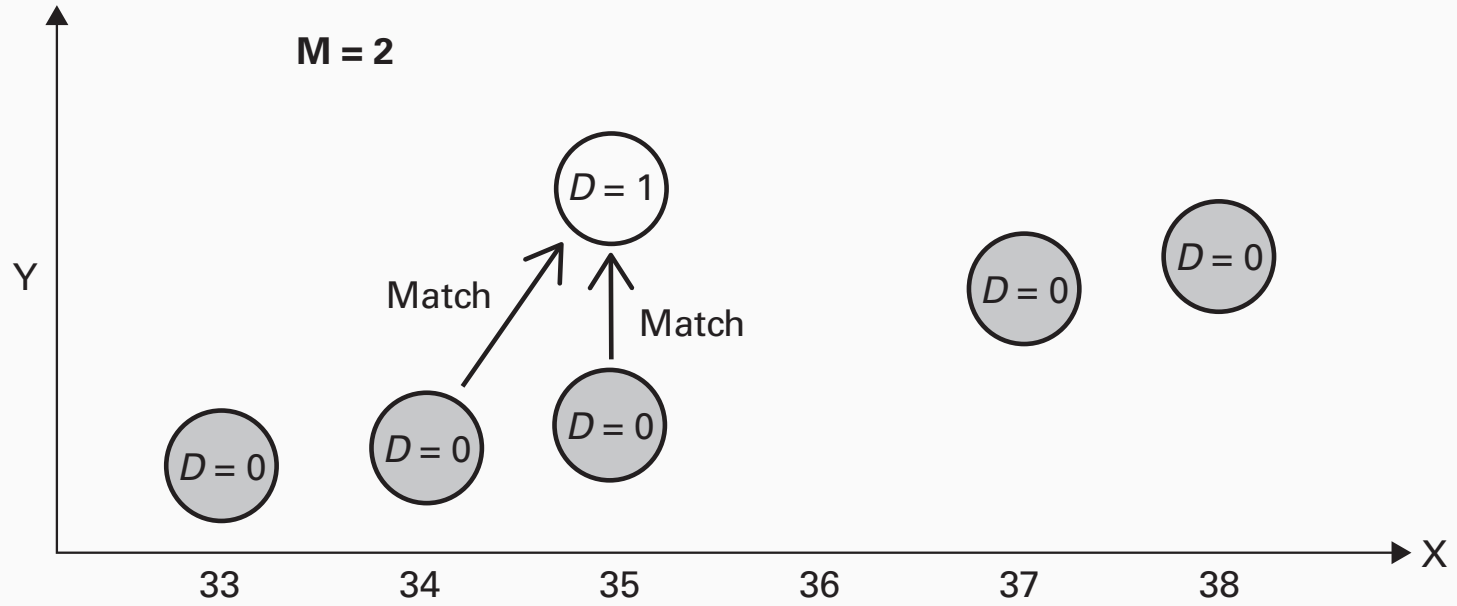
$$||X_j - X_i||_{\text{Mahalanobis}} = \sqrt{\sum_{k=1}^{K}\sum_{l=1}^{K} \frac{(X_{jk} - X_{ik})(X_{jl} - X_{il})}{\widehat{Cov}(X_k, X_l)}} \tag{4.26}$$

  - Incorporates inverse weighting by both the variance and covariance of the covariates.
  - Less weight when X_k strongly correlates with other covariates, as good matches for them likely imply a decent match for X_k.
  - Greater weight when X_k is independent of other covariates to ensure a satisfactory match quality for X_k.

1:*M* Matching (1)

- 1:*M* (one-to-many) matching: Match the *M* closest observations from the other treatment group to reference observation *i*.

Figure 4.4: 1:*M* matching



- In this example, we have one covariate X and $M = 2$.

1:*M* Matching (2)

- The 1:*M* matching estimator of the ATET corresponds to:

$$\hat{\Delta}_{D=1} = \frac{1}{n_1} \sum_{i:D_i=1} [Y_i - \hat{\mu}_0(X_i)] \tag{4.27}$$

- Let $J(i)$ denote the set of M nontreated observations matched to a treated reference observation i. Then, $\hat{\mu}_0(X_i)$ corresponds to:

$$\hat{\mu}_0(X_i) = \frac{1}{M} \sum_{j \in J(i)} Y_j \tag{4.28}$$

- Equivalent expression for the ATET estimator:

$$\hat{\Delta}_{D=1} = \frac{1}{n_1} \sum_{i=1}^{n} \left[D_i - \frac{W_i}{M} \right] \cdot Y_i \tag{4.29}$$

- W_i: Number of times a nontreated observation is matched to any treated observation.

Radius Matching (1)

- Radius matching defines a maximum admissible dissimilarity in X between matched treated and nontreated observations.
- Let $\mathcal{B}$ be the threshold for the distance metric (e.g., variance or Mahalanobis distance).
- For any treated reference observation i, estimate $\mu_0(X_i)$ as the average of all nontreated observations within $\mathcal{B}$:

$$\hat{\mu}_0(X_i) = \frac{\sum_{j:D_j=0} I\{||X_j - X_i|| \leq \mathcal{B}\} \cdot Y_j}{\sum_{j:D_j=0} I\{||X_j - X_i|| \leq \mathcal{B}\}} \tag{4.30}$$

- The more similar potential matches are available in the data, the more comparison observations are actually matched.

Radius Matching (2)

- In contrast to 1:M matching, radius matching does not fix the number of matches, but makes this choice data-dependent.
- This may decrease variance (without large costs in terms of bias) if many similar comparison observations exist.
- A kernel function can also be used to make weights dependent on the magnitude of the distance metric:

$$\hat{\mu}_0(X_i) = \frac{\sum_{j:D_j=0} \mathcal{K}\left(\frac{||X_j - X_i||}{\mathcal{B}}\right) \cdot Y_j}{\sum_{j:D_j=0} \mathcal{K}\left(\frac{||X_j - X_i||}{\mathcal{B}}\right)} \tag{4.31}$$

Asymptotic Properties of Matching Estimators

- Pair matching and 1:M matching estimators are not necessarily $\sqrt{n}$-consistent if X contains more than one continuous element (Abadie and Imbens, 2006).
- Reason: Using a fixed number of matches does not optimally trade off the bias and variance of the estimator.
- In contrast, kernel matching can attain $\sqrt{n}$-consistency if the bandwidth h is appropriately adapted to the sample size.
- Even under $\sqrt{n}$-consistency, pair or 1:M matching tends to have a higher asymptotic variance than the most precise estimators relying on the same assumptions.

Variance Estimation for Matching Estimators (1)

- Bootstrapping approaches are inconsistent for pair and 1:M matching due to the discontinuity of weights (Abadie and Imbens, 2008).
- Only the selected one or M matches receive positive weight; all other observations in the sample have zero weight.
- A consistent asymptotic approximation of the estimator's variance (for the ATET) is given by:

$$Var(\hat{\Delta}_{D=1}) = \frac{1}{n_1}\left\{E[(\Delta_{X_i} - \Delta_{D=1})^2|D_i = 1] + E\left[\frac{1}{n_1}\sum_{i=1}^{n}\left[D_i - (1-D_i)\cdot\frac{W_i}{M}\right]^2 \cdot \sigma^2(D_i, X_i)\right]\right\} \quad (4.32)$$

  - $\sigma^2(D_i, X_i) = Var(Y|D = D_i, X = X_i)$ is the conditional variance of the outcome, given the treatment and the covariates.

Variance Estimation for Matching Estimators (2)

- The variance on the previous slide can be estimated by:

$$\widehat{Var}(\hat{\Delta}_{D=1}) = \frac{1}{n_1}\left\{\frac{1}{n_1}\sum_{i=1}^{n} D_i \cdot [Y_i - \hat{\mu}_0(X_i) - \hat{\Delta}_{D=1}]^2 + \frac{1}{n_1}\sum_{i=1}^{n}(1-D_i) \cdot \left[\frac{W_i \cdot (W_i - 1)}{M^2}\right] \cdot \hat{\sigma}^2(D_i, X_i)\right\} \tag{4.33}$$

- $\hat{\sigma}^2(D_i, X_i)$ is estimated via matching within the nontreated group:

$$\hat{\sigma}^2(D_i, X_i) = \frac{M}{M+1} \cdot \left(Y_i - \frac{1}{M}\sum_{j \in \mathcal{J}(i)} Y_j\right)^2 \tag{4.34}$$

  - $\mathcal{J}_i$: Set of M nontreated observations that are closest to some nontreated reference observation i.

- $\hat{\sigma}^2(D_i, X_i)$ is inconsistent (fixed M), but averaging across treated observations in (4.33) yields a consistent variance estimator.

Bias Correction (1)

- Pair or 1:M matching can be combined with a regression-based bias correction to improve the properties of the estimators.
- Bias arises because matched observations $j \in J(i)$ typically do not have exactly the same X values as reference observation i.
- Reconsider the estimator of the ATET:

$$\hat{\Delta}_{D=1} = \frac{1}{n_1} \sum_{i:D_i=1} [Y_i - \hat{\mu}_0(X_i)] \quad \text{with} \quad \hat{\mu}_0(X_i) = \frac{1}{M} \sum_{j \in J(i)} Y_j$$

- Correct for the bias due to $X_j - X_i \neq 0$ by modifying $\hat{\mu}_0(X_i)$:

$$\hat{\mu}_0(X_i) = \frac{1}{M} \sum_{j \in J(i)} [Y_j - (\tilde{\mu}_0(X_j) - \tilde{\mu}_0(X_i))] \tag{4.35}$$

  - $\tilde{\mu}_0(X)$: Estimate from regressing Y on X among the nontreated.

Bias Correction (2)

- Bias correction removes the bias without affecting the asymptotic variance (Rubin, 1979, Abadie and Imbens, 2011).
- Under specific conditions, it entails a $\sqrt{n}$-consistent and asymptotically normal ATET estimator.
- This approach may reduce the bias even if the regression model for Y given X and $D = 0$ is somewhat misspecified.
- In small samples, the bias correction increases variance of ATET estimation due to estimating $\tilde{\mu}_0(X)$.
- Kernel or radius matching with nondiscontinuous, smooth kernel weights are less problematic for bootstrap-based inference.
- Bias correction may still be beneficial in terms of bias reduction in this case.

Table of Contents

Motivation

- Caveat of methods like covariate matching, kernel, or series regression that control for X nonparametrically:

Curse of dimensionality:

As the number of covariates in X and possible covariate values increases, it becomes harder to find good matches in finite samples.

- Alternative approach: Control for the *propensity score* instead of directly controlling for X.

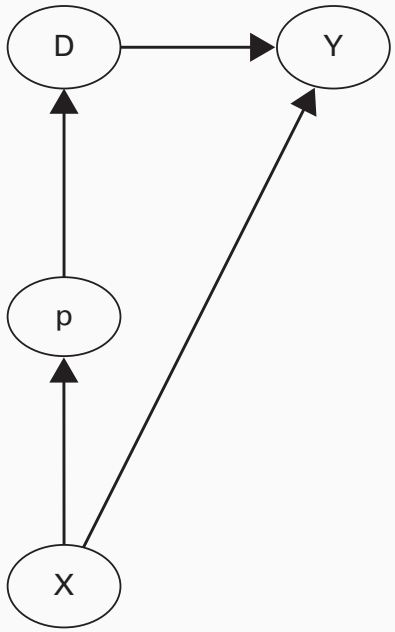
Propensity score

The conditional treatment probability given the covariates:
$p(X) = \Pr(D = 1|X)$

Balancing Property of the Propensity Score

- Conditioning on $p(X)$ balances covariate distribution across treatment groups $\Rightarrow X \perp D | p(X)$ (Rosenbaum and Rubin, 1983b).

Figure 4.5: A causal graph including the propensity score (denoted by p)



- Controlling for $p(X)$ blocks any impact of X on D and thus, X no longer jointly affects D and Y.

Propensity Score Matching (1)

- The ATE and ATET can also be identified by controlling for the propensity score $p(X)$ instead of X:

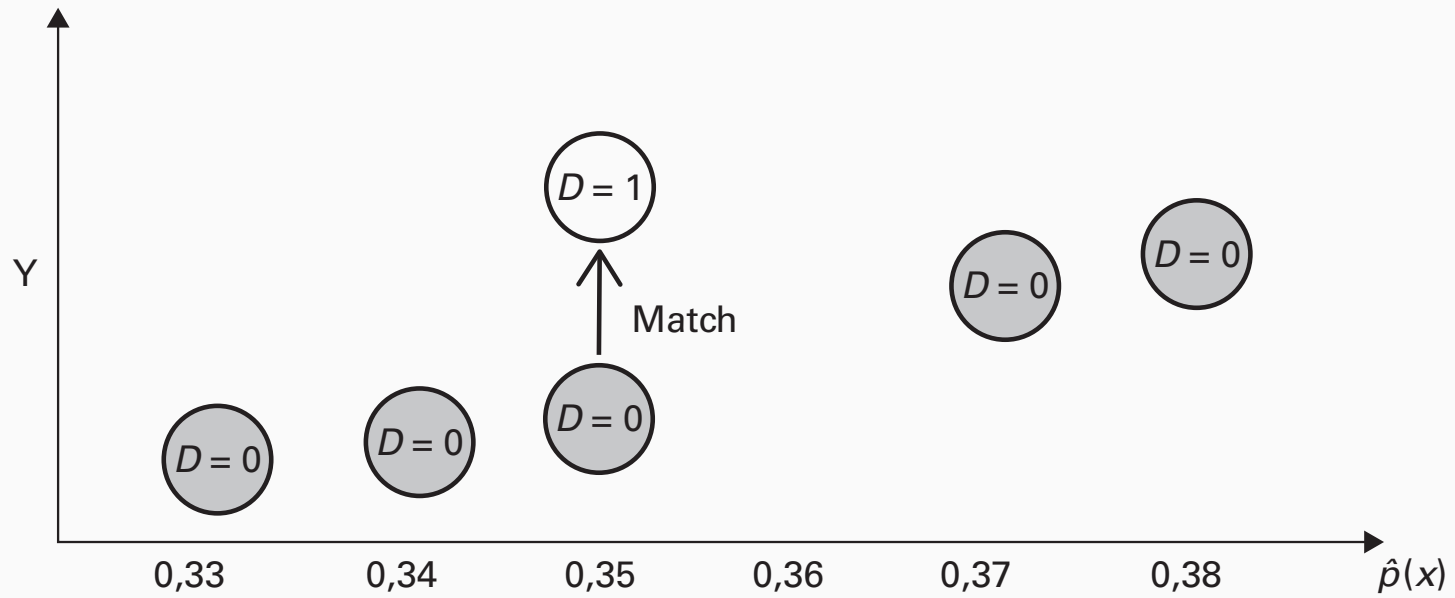
$$\begin{aligned} \Delta &= E[\mu_1(p(X)) - \mu_0(p(X))], \\ \Delta_{D=1} &= E[\mu_1(p(X)) - \mu_0(p(X))|D=1] \\ &= E[Y|D=1] - E[\mu_0(p(X))|D=1] \end{aligned} \tag{4.36}$$

- Therefore, we may estimate treatment effects by substituting X with an estimate of $p(X)$ in matching or regression approaches.
- Propensity score matching is often preferred over covariate matching:
  - Matches only on an estimate of $p(X)$ (a single variable) instead of high-dimensional X.
  - No need for a distance metric because we do not aggregate the distances in several covariates.

Propensity Score Matching (2)

- Match to a treated reference observation the nontreated subject with the most similar estimated propensity score $\hat{p}(X)$.

Figure 4.6: Propensity score matching



Propensity Score Estimation (1)

- Nonparametric estimation of $p(X)$ (e.g., by kernel or series regression) still suffers from the curse of dimensionality.
- To avoid this, $p(X)$ is often estimated using a parametric binary choice model:

$$\Pr(D = 1|X) = p(X) = \Lambda(\alpha_0 + \alpha_{X_1} X_1 + \cdots + \alpha_{X_K} X_K) \tag{4.37}$$

  - Combines a linear index $\alpha_0 + \alpha_{X_1} X_1 + \cdots + \alpha_{X_K} X_K$ and a nonlinear link function Λ of that index.
  - Λ is typically either a normal or logistic distribution function.
  - Implies that $p(X)$ is estimated using a probit or logit model and thus strictly between 0 and 1.
- Propensity score estimation may be inconsistent if:
  - Treatment decision cannot be modeled using a linear index of X.
  - Unobservables do not follow the assumed distribution (i.e., normal for probit, logistic for logit models).

Propensity Score Estimation (2)

- Parametric binary choice models (e.g., probit/logit models) are conventionally estimated by maximum likelihood estimation:

$$\hat{\alpha_0}, \ldots, \hat{\alpha}_{X_K} = \arg \max_{\alpha_0^*, \ldots, \alpha_{X_K}^*} \sum_{i=1}^{n} D_i \ln(\Lambda(\alpha_0^* + \cdots + \alpha_{X_K}^* X_{iK})) + (1 - D_i) \ln(1 - \Lambda(\alpha_0^* + \cdots + \alpha_{X_K}^* X_{iK})) \tag{4.38}$$

  - Intuition: Find coefficient values that maximize the joint likelihood to obtain the treatment states observed in the sample.

- Propensity score estimate:

$$\hat{p}(X) = \Lambda(\hat{\alpha}_0 + \hat{\alpha}_{X_1} X_1 + \cdots + \hat{\alpha}_{X_K} X_K) \tag{4.39}$$

Variance Estimation for Propensity Score Matching

- Matching on $\hat{p}(X)$ has a different variance than matching directly on X.
- Variance estimator for propensity score matching must include a correction term to account for uncertainty from estimating $p(X)$.
- Ignoring the correction term and implicitly assuming that the propensity score is known leads to biased variance estimates.
- For ATE estimation, ignoring the correction generally overestimates the true variance (bias is never negative).
- Bootstrapping accounts for uncertainty from estimating $p(X)$ by re-estimating both $\hat{p}(X)$ and the ATET in each bootstrap sample.

Table of Contents

Inverse Probability Weighting (1)

Inverse probability weighting (IPW) (Horvitz and Thompson, 1952)

Observations with propensity scores that are underrepresented (overrepresented) in their treatment groups relative to some target population are given more (less) weight.

- The ATE is identified by:

$$\Delta = E[\mu_1(X) - \mu_0(X)] = E\left[\frac{E[Y \cdot D|X] \cdot D}{p(X)} - \frac{E[Y \cdot (1-D)|X] \cdot (1-D)}{1-p(X)}\right]$$

$$= E\left[\frac{Y \cdot D}{p(X)} - \frac{Y \cdot (1-D)}{1-p(X)}\right] \tag{4.40}$$

- The ATET is identified by:

$$\Delta_{D=1} = E\left[\frac{Y \cdot D}{\Pr(D=1)} - \frac{Y \cdot (1-D) \cdot p(X)}{(1-p(X)) \cdot \Pr(D=1)}\right] \tag{4.41}$$

Inverse Probability Weighting (2)

- When estimating treatment effects like the ATE based on sample analogs of IPW equations, normalization is recommended:

$$\hat{\Delta} = \sum_{i=1}^{n} \frac{Y_i \cdot D_i}{\hat{p}(X_i)} \Big/ \sum_{i=1}^{n} \frac{D_i}{\hat{p}(X_i)} - \sum_{i=1}^{n} \frac{Y_i \cdot (1 - D_i)}{1 - \hat{p}(X_i)} \Big/ \sum_{i=1}^{n} \frac{1 - D_i}{1 - \hat{p}(X_i)} \tag{4.42}$$

  - $\sum_{i=1}^{n} D_i / \hat{p}(X_i)$ and $\sum_{i=1}^{n} (1 - D_i)/(1 - \hat{p}(X_i))$ normalize the weights such that they add up to 1 within the treatment groups.

- In smaller samples, normalized sample analogs typically entail better effect estimation, see Busso, DiNardo, and McCrary (2014).
- $\hat{p}(X_i)$ may be estimated parametrically or nonparametrically, e.g. based on series estimation (Hirano, Imbens, and Ridder, 2003) or kernel regression (Ichimura and Linton, 2005).

Inverse Probability Weighting (3)

- Advantages of IPW
  - Computationally inexpensive
  - No need to choose tuning parameters (e.g., number of matches).
  - If the propensity score is nonparametrically estimated, it can attain the semiparametric efficiency bound (lowest possible asymptotic variance).
- Disadvantages of IPW
  - Estimates may be more sensitive to errors in propensity scores close to 1 or 0 (higher variance; especially in small samples).
  - May be less robust (i.e., more prone to estimation errors) when using an incorrect model for the propensity score than matching.

Empirical Likelihood Methods (1)

- After weighting by the inverse of the true $p(X)$, the covariates X and any of their moments are balanced across treatment groups.
- For the ATET, this implies:

$$E\left[\frac{\tilde{X}\cdot D}{\Pr(D=1)}-\frac{\tilde{X}\cdot(1-D)\cdot p(X)}{(1-p(X))\cdot\Pr(D=1)}\right]=0$$

$$\Leftrightarrow E\left[\tilde{X}\cdot D-\frac{\tilde{X}\cdot(1-D)\cdot p(X)}{1-p(X)}\right]=0 \tag{4.43}$$

  - $\tilde{X}$ is a function of X (e.g., $\tilde{X}=X$ for balancing the mean and $\tilde{X}=(X-E[X])^2$ for balancing the variance across treatment groups).
- Problem: The initial propensity score estimate $\hat{p}(X_i)$ may not fully balance $\tilde{X}$ in the sample.
- **Empirical likelihood (EL) methods** aim to modify $\hat{p}(X_i)$ to ensure that $\tilde{X}$ is as similar as possible across treatment groups.

Empirical Likelihood Methods (2)

EL methods (Graham, Pinto, and Egel, 2012, Imai and Ratkovic, 2014)

Modify an initial propensity score estimate $\hat{p}(X_i)$ (through changing the coefficients) until predefined moments of X are maximally balanced across treatment groups.

- Aim: Enforcing the balance condition on the previous slide to hold in the sample:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\tilde{X}_i \cdot D_i - \frac{\tilde{X}_i \cdot (1 - D_i) \cdot \tilde{p}(X_i)}{1 - \tilde{p}(X_i)}\right] = 0 \tag{4.44}$$

  - $\tilde{p}(X_i)$ is an adjusted version of $\hat{p}(X_i)$ that fully balances $\tilde{X}$.
- Avoids manually searching for propensity score specifications that entail decent balancing.
- $\tilde{p}(X_i)$ can be used not only for IPW but also for other estimators like propensity score matching.

Entropy Balancing

Entropy balancing (EB) (Hainmueller, 2012)

Iteratively modifies initial (e.g., uniform) default weights until the predefined balance criterion with regard to X is maximized.

Constraint: Weights must sum to 1 (and be nonnegative) in either treatment group.

- Both EL and EB aim at perfect covariate balance to make treated and nontreated observations fully comparable.
- This avoids bias from dissimilarities in X, but may increase the variance of the estimator.
- In contrast to EL, EB does not require an initial estimate of the propensity score for computing the final weights.

Table of Contents

Doubly Robust Methods (1)

- We may combine models for conditional mean outcomes $\mu_1(X), \mu_0(X)$ and propensity scores $p(X)$ when evaluating treatment effects.
- Entails so-called **doubly robust (DR)** expressions of the ATE (Δ) and ATET ($\Delta_{D=1}$), see Robins, Rotnitzky, and Zhao (1994), Robins and Rotnitzky (1995), and Hahn (1998):

$$\Delta = E\left[\phi(X)\right],$$

$$\text{with } \phi(X) = \mu_1(X) - \mu_0(X) + \frac{(Y - \mu_1(X)) \cdot D}{p(X)} - \frac{(Y - \mu_0(X)) \cdot (1 - D)}{1 - p(X)},$$

$$\Delta_{D=1} = E\left[\frac{(Y - \mu_0(X)) \cdot D}{\Pr(D = 1)} - \frac{(Y - \mu_0(X)) \cdot (1 - D) \cdot p(X)}{(1 - p(X)) \cdot \Pr(D = 1)}\right] \tag{4.45}$$

- $\phi(X)$ is the efficient influence function.

Doubly Robust Methods (2)

- DR, IPW, and outcome regression are equivalent for identifying causal effects, when $\mu_1(X), \mu_0(X), p(X)$ are correctly specified.
- This follows from the fact that by the law of iterated expectations, some terms cancel out:

$$E\left[\frac{(Y-\mu_1(X))\cdot D}{p(X)} - \frac{(Y-\mu_0(X))\cdot(1-D)}{1-p(X)}\right] = E\left[\frac{\varepsilon\cdot D}{p(X)} - \frac{\varepsilon\cdot(1-D)}{1-p(X)}\right] = 0 \text{ and } (4.46)$$

$$E\left[\frac{-\mu_0(X)\cdot D}{\mathbf{Pr}(D=1)} - \frac{-\mu_0(X)\cdot(1-D)\cdot p(X)}{(1-p(X))\cdot\mathbf{Pr}(D=1)}\right] = E\left[\mu_0(X)\cdot\left(\frac{p(X)}{\mathbf{Pr}(D=1)} - \frac{p(X)}{\mathbf{Pr}(D=1)}\right)\right] = 0,$$

- with the error term $\varepsilon = Y - \mu_D(X)$ and $E[\varepsilon|D,X] = 0$.

Doubly Robust Methods (3)

- DR estimators are consistent if *either* the conditional mean outcomes *or* the propensity scores are correctly specified.

  $\Rightarrow$ Two chances for correct specification.
- This makes DR estimators more robust than outcome regression (relies on $\mu_1(X), \mu_0(X)$ only) and IPW (relies on $p(X)$ only).
- If both models are correct, DR estimation is semiparametrically efficient.
- This also holds when $\mu_1(X), \mu_0(X)$ and $p(X)$ are estimated nonparametrically (e.g., by kernel or series regression).
- In small samples, nonparametric DR has lower bias and variance than nonparametric versions of IPW and outcome regression.
- This better finite sample behavior makes DR estimation attractive even when IPW and outcome regression are consistent.

Doubly Robust Methods (4)

- Targeted Maximum Likelihood Estimation (TMLE) is another DR method (van der Laan and Rubin, 2006).
- Step 1: Obtain initial estimates of $\mu_1(X)$ and $\mu_0(X)$ by regression.
- Step 2: Update (robustify) these estimates by regressing them on a function of the estimated propensity score $p(X)$.
- If outcome regressions are misspecified but $p(X)$ is correct, TMLE corrects the bias. If outcome regressions are correct but $p(X)$ is misspecified, TMLE does no harm.
- Another DR approach is weighted outcome regression using IPW weights based on the propensity score.
- Common principle of DR methods is to combine outcome regression and propensity score estimation to exploit all information in the data to control for confounding.