Lecture 12: Average Treatment Effect

April 1, 2025

Introduction

- What kinds of questions can we answer using a "modern" approach to treatment effect estimation? Here are some examples:
  1. What are the effects of a job training program on employment or labor earnings?
  2. What are the effects of a school voucher program on student performance?
  3. Does a certain medical intervention increase the likelihood of survival?
- The main issue in program evaluation concerns the assignment of the binary intervention, or “treatment”.

Introduction

- For example, is the "treatment" randomly assigned? (Hardly ever in business and economics, and problematical even in clinical trials because those chosen to be eligible can and do opt out.)
- A more reasonable possibility is that the treatment is effectively randomly assigned conditional on observable covariates. ("Ignorability" of treatment, "unconfoundedness," or "selection on observables." Sometimes called "exogenous treatment.")

Introduction

- Or, does assignment depend fundamentally on unobservables, where the dependence cannot be broken by controlling for observables? ("Nonignorable" treatment, "confounded" assignment, "selection on unobservables," or "endogenous treatment.")
- Often there is a component of self-selection in program evaluation.

Introduction

- Nevertheless, start with unconfoundedness because it is often all we have (and is a good starting point in any case). A key point is that, under the ignorability or unconfoundedness assumption, regression methods with the covariates as controls have the same ability - at least in theory - of identifying the treatment effect parameters. Therefore, propensity score methods and/or matching methods are not a panacea for the self-selection problem.

Basic Concepts

Counterfactual Outcomes and Parameters of Interest

- For each population unit, two possible outcomes: y_0 (the outcome without treatment) and y_1 (the outcome with treatment). The binary "treatment" indicator is w, where $w = 1$ denotes "treatment." The nature of y_0 and y_1 - discrete, continuous, some mix - is, for now, unspecified. (The generality this affords is one of the attractions of the **Rubin Causal Model**.)
- The gain from treatment is

$$y_1 - y_0 \tag{1}$$

Basic Concepts

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

$$y_{i1} - y_{i0}$$

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

- Problem: For each unit i, only one of y_{i0} and y_{i1} is observed.
- In effect, we have a missing data problem (even though we will eventually assume a random sample of units).

Basic Concepts

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

$$\tau_{\text{ate}} = E\left(y_1 - y_0\right). \tag{2}$$

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

  This is sometimes called the average causal effect.

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

$$\tau_{\text{att}} = E\left(y_1 - y_0 \mid w = 1\right) \tag{3}$$

Basic Concepts

- 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.
- Not suprisingly, how we estimate τ_{ate} and τ_{att} depends on what we assume about treatment assignment.

Sampling Assumptions

- Assume independent, identically distributed observations from the underlying population. The data we would like to have is $\{(y_{i0}, y_{i1}) : i = 1, \ldots, N\}$, but we only observe w_i and

$$y_i = (1 - w_i)\, y_{i0} + w_i y_{i1} \tag{4}$$

- Random sampling rules out treatment status of one unit having an effect on other units.
- Also implies that the outcome for unit i does not affect the outcome for other members of the population.

Estimation under Random Assignment

- With $y = (1 - w)y_0 + wy_1$ we can always write

$$E(y \mid w) = (1 - w)E\left(y_0 \mid w\right) + wE\left(y_1 \mid w\right)$$

- Strongest form of random assignment: (y_0, y_1) is independent of w.
- It follows that $E(y \mid w = 1) = E\left(y_1\right)$ and $E(y \mid w = 0) = E\left(y_0\right)$, and so

$$E(y \mid w = 1) - E(y \mid w = 0) = E\left(y_1\right) - E\left(y_0\right) = \tau_{\text{ate}} \ = \tau_{\text{att}} \ . \qquad (5)$$

- An unbiased and consistent estimator of $E(y \mid w = 1)$ is the sample average on the treated subsample and similarly for $E(y \mid w = 0)$. The estimator $\hat{\tau}_{\text{ate}}$ is just the simple difference-in-means estimator.
- The randomization of treatment needed for the simple comparison-of-means estimator to consistently estimate the ATE is rare in practice but not unheard of. (Eligibility is sometimes randomly assigned, but actual participation need not be.)

Key Assumptions: Ignorability and Overlap

- Rather than assume random assignment, for each unit i we also draw a vector of covariates, $\mathbf{x}_i$. Let $\mathbf{x}$ be the random vector with a distribution in the population.

  A.1. Ignorability (Unconfoundedness): Conditional on a set of covariates $\mathbf{x}$, the pair of counterfactual outcomes, (y_0, y_1), is independent of w, which is often written as

  $$(y_0, y_1) \perp w \mid \mathbf{x}, \tag{6}$$

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

Key Assumptions: Ignorability and Overlap

- w 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 $\mathbf{x}$): $D(w \mid y_0, y_1, \mathbf{x}) = D(w \mid \mathbf{x})$, where $D(\cdot \mid \cdot)$ 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).

Key Assumptions: Ignorability and Overlap

- Most often $\mathbf{x}$ includes variables that are measured prior to treatment assignment, such as previous labor market history. Of course, gender, race, and other demographic variables can be included.
- Can show unconfoundedness is generally violated if $\mathbf{x}$ includes variables that are themselves affected by the treatment. 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.

Key Assumptions: Ignorability and Overlap

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

$$0 < P(w = 1 \mid \mathbf{x}) < 1. \tag{7}$$

In other words, each unit in the defined population has some chance of being treated and some chance of not being treated.

- We define the **propensity score** as

$$p(\mathbf{x}) = P(w = 1 \mid \mathbf{x}), \mathbf{x} \in \mathcal{X}. \tag{8}$$

- **Strong Ignorability** [Rosenbaum and Rubin (1983)] = Ignorability + Overlap.

Identification I

- Use two ways to show the treatment effects are identified under ignorability.
- First is based on regression functions. Define **the average treatment effect conditional on** x as

$$\tau(\mathbf{x}) = E\left(y_1 - y_0 \mid \mathbf{x}\right) = E\left(y_1 \mid \mathbf{x}\right) - E\left(y_0 \mid \mathbf{x}\right) = \mu_1(\mathbf{x}) - \mu_0(\mathbf{x}).$$

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

Identification I

- By iterated expectations,

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

  It follows that 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,

$$\begin{aligned} E(y \mid \mathbf{x}, w) &= (1-w)E\left(y_0 \mid \mathbf{x}, w\right) + wE\left(y_1 \mid \mathbf{x}, w\right) \\ &= (1-w)E\left(y_0 \mid \mathbf{x}\right) + wE\left(y_1 \mid \mathbf{x}\right) \\ &\equiv (1-w)\mu_0(\mathbf{x}) + w\mu_1(\mathbf{x}), \end{aligned} \tag{9}$$

  where the second equality holds by ignorability and $\mu_g(\mathbf{x}) \equiv E\left(y_g \mid \mathbf{x}\right)$, $g = 0, 1$. So

$$\begin{aligned} E(y \mid \mathbf{x}, w = 0) &= \mu_0(\mathbf{x}) \\ E(y \mid \mathbf{x}, w = 1) &= \mu_1(\mathbf{x}) \end{aligned}$$

Identification I

- The functions $E(y \mid \mathbf{x}, w = 0), E(y \mid \mathbf{x}, w = 1)$ are consistently estimable from the data because we have a random sample on $(y, \mathbf{x}, w)$. But overlap is critical. We need to estimate $\mu_0(\mathbf{x})$ and $\mu_1(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. But by defintion $E(y \mid \mathbf{x}, w = 0)$ will be estimated only using the control group and $E(y \mid \mathbf{x}, w = 1)$ will be estimated only using the treatment group. (More on the overlap issue when we consider estimation.)

Identification I

- For ATT note that

$$\begin{aligned} E\left(y_1 - y_0 \mid w\right) &= E\left[E\left(y_1 - y_0 \mid \mathbf{x}, w\right) \mid w\right] = E\left[E\left(y_1 - y_0 \mid \mathbf{x}\right) \mid w\right] \\ &= E\left[\mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) \mid w\right], \end{aligned}$$

  where the second equality holds by ignorability (in the mean), that is, $E\left(y_1 - y_0 \mid \mathbf{x}, w\right) = E\left(y_1 - y_0 \mid \mathbf{x}\right)$.

- So

$$\tau_{\text{att}} = E\left[\mu_1(\mathbf{x}) - \mu_0(\mathbf{x}) \mid w = 1\right],$$

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

Identification II

- We can also establish identification using propensity score weighting. Ignorability implies that w and y_g are uncorrelated conditional on $\mathbf{x}$, and so, by iterated expectations,

$$\begin{aligned} E\left[\frac{wy}{p(\mathbf{x})}\right] &= E\left[\frac{wy_1}{p(\mathbf{x})}\right] = E\left[\left(\frac{wy_1}{p(\mathbf{x})}\middle|\ \mathbf{x}\right)\right] \\ &= E\left[\frac{E(w \mid \mathbf{x})E\left(y_1 \mid \mathbf{x}\right)}{p(\mathbf{x})}\right] = E\left[E\left(y_1 \mid \mathbf{x}\right)\right] = E\left(y_1\right) \end{aligned} \tag{10}$$

  A similar argument shows

$$E\left[\frac{(1-w)y}{1-p(\mathbf{x})}\right] = E\left(y_0\right). \tag{11}$$

Identification II

- Putting the two expressions together gives

$$\tau_{\text{ate}} = E\left[\frac{wy}{p(\mathbf{x})} - \frac{(1-w)y}{1-p(\mathbf{x})}\right] = E\left\{\frac{[w-p(\mathbf{x})]y}{p(\mathbf{x})[1-p(\mathbf{x})]}\right\}. \tag{12}$$

- Clear from (12) 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.

Estimating ATEs

- 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.

Regression

Regression Adjustment

- First step is to obtain $\hat{m}_0(\mathbf{x})$ from the "control" subsample, $w_i = 0$, and $\hat{m}_1(\mathbf{x})$ from the "treated" subsample, $w_i = 1$. Can be as simple as (flexible) linear regression or full nonparametric regression.
- Compute fitted values in each case for all units in sample. (This is made easy in Stata using the "predict" command because a fitted value is computed for all units with nonmissing $\mathbf{x}_i$, even if a unit was not used in estimation.)

Regression

- The regression-adjustement estimates are

$$\hat{\tau}_{\text{ate,reg}} = N^{-1} \sum_{i=1}^{N} \left[\hat{m}_1 \left(\mathbf{x}_i \right) - \hat{m}_0 \left(\mathbf{x}_i \right) \right] = N^{-1} \sum_{i=1}^{N} \hat{\tau} \left(\mathbf{x}_i \right) \tag{13}$$

$$\hat{\tau}_{\text{att,reg}} = N_1^{-1} \sum_{i=1}^{N} w_i \left[\hat{m}_1 \left(\mathbf{x}_i \right) - \hat{m}_0 \left(\mathbf{x}_i \right) \right] = N_1^{-1} \sum_{i=1}^{N} w_i \hat{\tau} \left(\mathbf{x}_i \right) \tag{14}$$

where $N_1 = \sum_{i=1}^{N} w_i$ is the number of treated units.

Regression

- How does overlap affect estimation of τ_{ate} and τ_{att} ? Note that $\hat{\tau}_{\text{ate,reg}}$ requires two kinds of extrapolation: we must evaluate $\hat{m}_0(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}_i$ for treated i and we must evaluate $\hat{m}_1(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}_i$ for untreated i.
- When we use parametric models for $m_g(\cdot)$, extrapolation is easy. But it may be hiding a problem. The estimates of the mean functions where data are scarce may be very sensitive to functional form.
- Nonparametric methods that use local averaging will reveal overlap problems.

Regression

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

$$\hat{\tau}_{\text{ate,reg}} = (\hat{\alpha}_1 - \hat{\alpha}_0) + \overline{\mathbf{x}} \left(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0 \right), \tag{15}$$

  where $\overline{\mathbf{x}}$ 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.)

- Easiest way to obtain standard error for $\hat{\tau}_{\text{ate, reg}}$ is to ignore sampling error in $\overline{\mathbf{x}}$ and use the coefficient on w_i in the regression

$$y_i \text{ on } 1, w_i, \mathbf{x}_i, w_i \cdot (\mathbf{x}_i - \overline{\mathbf{x}}), i = 1, \ldots, N. \tag{16}$$

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

Regression

- 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 $E(y \mid \mathbf{x}, w = 1)$ 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 $\mathbf{x}$.

Propensity Score Weighting

- The formula that establishes identification of τ_{ate} base on population moments suggests an imediate estimator of τ_{ate} :

$$\tilde{\tau}_{\text{ate, psw}} = N^{-1} \sum_{i=1}^{N} \left[\frac{w_i y_i}{p(\mathbf{x}_i)} - \frac{(1 - w_i) y_i}{1 - p(\mathbf{x}_i)} \right]. \tag{17}$$

- $\tilde{\tau}_{\text{ate, psw}}$ 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 $p(\cdot)$, $\tilde{\tau}_{\text{ate, psw}}$ is not asymptotically efficient. It is better to estimate the propensity score!

Propensity Score Weighting

- Two approaches: (1) Model $p(\cdot)$ parametrically, in a flexible way.

  Can show estimating the propensity score leads to a smaller asymptotic variance when the parametric model is correctly specified. (2) Use an explicit nonparametric approach, as in Hirano, Imbens, and Ridder (2003, Econometrica) or Li, Racine, and Wooldridge (2009, JBES).

$$\hat{\tau}_{\text{ate, psw}} = N^{-1} \sum_{i=1}^{N} \left[\frac{w_i y_i}{\hat{p}(\mathbf{x}_i)} - \frac{(1 - w_i) y_i}{1 - \hat{p}(\mathbf{x}_i)} \right] = N^{-1} \sum_{i=1}^{N} \frac{[w_i - \hat{p}(\mathbf{x}_i)] y_i}{\hat{p}(\mathbf{x}_i)[1 - \hat{p}(\mathbf{x}_i)]}. \tag{18}$$

- Very simple to compute given $\hat{p}(\cdot)$.

Propensity Score Weighting

- τ_{att} is estimated using identical reasoning:

$$\hat{\tau}_{\text{att psw}} = N^{-1} \sum_{i=1}^{N} \frac{[w_i - \hat{p}(\mathbf{x}_i)] y_i}{\hat{\rho}[1 - \hat{p}(\mathbf{x}_i)]}, \tag{19}$$

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

Propensity Score Weighting

- Can see directly from $\hat{\tau}_{\text{ate, psw}}$ and $\hat{\tau}_{\text{att, psw}}$ that the inverse probability weighted (IPW) estimators can be very sensitive to extreme values of $\hat{p}(\mathbf{x}_i)$. $\hat{\tau}_{\text{att, psw}}$ is sensitive only to $\hat{p}(\mathbf{x}_i) \approx 1$, but $\hat{\tau}_{\text{ate, psw}}$ 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.

Combining Regression and Propensity Score: reduce dimensionality of regression

Regression on the Propensity Score

- The motivation is that one can show, given ignorability, that ignorability actually holds conditional only on $p(\mathbf{x})$:

$$(y_0, y_1) \perp w \mid p(\mathbf{x}),$$

which, of course, implies

$$E\left[y_g \mid p(\mathbf{x}), w\right] = E\left[y_g \mid p(\mathbf{x})\right], g = 0, 1.$$

- In other words, it is sufficient to condition only on the propensity score so break the dependence between w and (y_0, y_1). We need not condition on $\mathbf{x}$.

Combining Regression and Propensity Score

- By iterated expectations,

$$\tau_{\text{ate}} = E(y_1 - y_0) = E\{E[y_1 \mid p(\mathbf{x})] - E[y_0 \mid p(\mathbf{x})]\}.$$

- Now we can obtain a conditional expectation for the observable y :

$$\begin{aligned} E[y \mid p(\mathbf{x}), w] &= (1-w)E[y_0 \mid p(\mathbf{x}), w] + wE[y_1 \mid p(\mathbf{x}), w] \\ &= (1-w)E[y_0 \mid p(\mathbf{x})] + wE[y_1 \mid p(\mathbf{x})] \end{aligned}$$

  where the second equality follows by the previous result.

Combining Regression and Propensity Score

- We have shown that

$$E[y \mid p(\mathbf{x}), w = 0] = E\left[y_0 \mid p(\mathbf{x})\right]$$
$$E[y \mid p(\mathbf{x}), w = 1] = E\left[y_1 \mid p(\mathbf{x})\right]$$

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

Combining Regression and Propensity Score

- In the linear case, $E\left[y_g \mid p(\mathbf{x})\right] = \alpha_g + \gamma_1 p(\mathbf{x}), g = 0, 1$, and we use

$$y_i \text{ on } 1, \hat{p}\left(\mathbf{x}_i\right) \text{ for } w_i = 0 \text{ and } y_i \text{ on } 1, \hat{p}\left(\mathbf{x}_i\right) \text{ for } w_i = 1, \tag{20}$$

  which gives fitted values $\hat{\alpha}_0 + \hat{\gamma}_0 \hat{p}\left(\mathbf{x}_i\right)$ and $\hat{\alpha}_1 + \hat{\gamma}_1 \hat{p}\left(\mathbf{x}_i\right)$, respectively.

- A consistent estimator of τ_{ate} is

$$\hat{\tau}_{\text{ate,regps}} = N^{-1} \sum_{i=1}^{N} \left[\left(\hat{\alpha}_1 - \hat{\alpha}_0\right) + \left(\hat{\gamma}_1 - \hat{\gamma}_0\right) \hat{p}\left(\mathbf{x}_i\right)\right]. \tag{21}$$

Combining Regression and Propensity Score

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

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

  as regressors when y has a wide range. So, regress y_i on $1, \hat{r}_i, \hat{r}_i^2, \ldots, \hat{r}_i^Q$ for some Q using both the control and treated samples, and then average the difference in fitted values to obtain $\hat{\tau}_{\text{ate,regprop}}$.

Matching

- Matching estimators are based on imputing a value on the counterfactual outcome for each unit. That is, for a unit i in the control group, we observe y_{i0}, but we need to impute y_{i1}. For each unit i in the treatment group, we observe y_{i1} but need to impute y_{i0}.
- For τ_{ate} , matching estimators take the general form

$$\hat{\tau}_{\text{ate,match}} = N^{-1} \sum_{i=1}^{N} (\hat{y}_{i1} - \hat{y}_{i0})$$

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

Matching

- For τ_{att},

$$\hat{\tau}_{\text{att,match}} = N_1^{-1} \sum_{i=1}^{N} w_i \left(y_i - \hat{y}_{i0} \right)$$

  where this form uses the fact that y_{i1} is always observed for the treated subsample. (In other words, we never need to impute y_{i1} for the treated subsample.)

Matching

- Abadie and Imbens (2006, Econometrica) consider several approaches. The simplest is to find a single match for each observation. Suppose i is a treated observation ($w_i = 1$). Then $\hat{y}_{i1} = y_i, \hat{y}_{i0} = y_h$ for h such that $w_h = 0$ and unit h is "closest" to unit i based on some metric (distance) in the covariates. In other words, for the treated unit i we find the "most similar" untreated observation, and use its response as y_{i0}. Similarly, if $w_i = 0, \hat{y}_{i0} = y_i, \hat{y}_{i1} = y_h$ where now $w_h = 1$ and $\mathbf{x}_h$ is "closest" to $\mathbf{x}_i$.
- Abadie and Imbens matching has been programmed in Stata in the command "nnmatch." The default is to use the single nearest neighbor.

Matching

- The default matrix in defining distance is the inverse of the diagonal matrix with sample variances of the covariates on the diagonal. [That is, diagonal Mahalanobis.]
- More generally, we can impute the missing values using an average of M nearest neighbors. If $w_i = 1$ then

$$\hat{y}_{i1} = y_i$$
$$\hat{y}_{i0} = M^{-1} \sum_{h \in \mathbb{N}_M(i)} y_h$$

where $\aleph_M(i)$ contains the M untreated nearest matches to observation i, based on the covariates. So for all $h \in \aleph_M(i), w_h = 0$.

Matching

- With ties, there can be more than M elements in $\aleph_M(i)$, and then M is replaced with the number of elements in $\aleph_M(i)$.
- Similarly, if $w_i = 0$,

$$\hat{y}_{i0} = y_i$$
$$\hat{y}_{i1} = M^{-1} \sum_{h \in \Im_M(i)} y_h$$

  where $\Im_M(i)$ contains the M treated nearest matches to observation i.