Introduction to Econometrics II: Recitation 10

Seung-hun Lee*

April 11th, 2022

1 Semiparametric Regression Models

1.1 Why are they called semiparametric?

Semiparametrics can be thought of as a middle ground between nonparametric and parametric regression. In fact, they share some traits that exist in parametric and nonparametric estimations. This becomes clear when we understand semiparametric estimation in the robustness-efficiency tradeoff context.

Nonparametric estimation is the most robust estimators conceivable in the sense that the consistency and asymptotic normality of these estimation does not depend on getting the assumption about specification right. However, this is not the most efficient estimation since it converges very slowly to the true distribution and significantly larger number of observation to achieve similar degree of precision. Parametric estimation is on the opposite end. They always converge to the true distribution at a nice speed of $\sqrt{n}$. The consistency and asymptotic normality of parametric models depend on specification assumption. Thus, robustness fails if the model is misspecified.

Semiparametric estimation can be understood as a method that stands on the middle ground. They are less efficient than parametric estimators since semiparametric estimation contains part of a nonparametric estimation by design and has larger variance. However, because of this exact reason, it performs better than parametric estimation in robustness since semiparametric estimation is robust in larger class of distribution than in parametric estimation. In a similar approach, semiparametric estimation is not more robust compared to nonparametrics but performs better in efficiency compared to nonparametric estimation.

*Contact me at sl4436@columbia.edu if you spot any errors or have suggestions on improving this note.

1.2 Partially linear models

There are two ways to set this up. The first way to think about semiparametric estimation is to consider the **partially linear models**. Suppose that we partition the covariates into two unoverlapping spaces - X and Z. Also assume that $E[\epsilon|X, W] = 0$. We work with the DGP of the following form

$$y_i = X_i\beta + g(Z_i) + \epsilon_i$$

where $\beta \in \dim(X_i)$ represents a coefficient for the linearly regressed terms and $g(\cdot)$ is a nonparametric portion of the regression.

To estimate β, we use the fact that

$$\begin{aligned} E[y_i|Z_i] &= E[X_i|Z_i]\beta + E[g(Z_i)|Z_i] + E[\epsilon_i|Z_i] \\ &= E[X_i|Z_i]\beta + g(Z_i) + E[\epsilon_i|Z_i] \\ &= E[X_i|Z_i]\beta + g(Z_i) + E[E[\epsilon_i|X_i, Z_i]|Z_i] = E[X_i|Z_i]\beta + g(Z_i) \end{aligned}$$

Given this, we can write

$$y_i - E[y_i|Z_i] = \{X_i - E[X_i|Z_i]\}\beta + \epsilon_i$$

Then we follow this procedure

1. Nonparametrically estimate $E[X_i|Z_i]$ and $E[y_i|Z_i]$. Then define $\widehat{X}_i = X_i - \widehat{E}[X_i|Z_i]$ and $\widehat{Y}_i = y_i - \widehat{E}[y_i|Z_i]$, where $\widehat{E}$ are nonparametric estimators
2. Regress $\widehat{Y}$ onto $\widehat{X}$ to get an estimate of β
3. We can estimate $g(\cdot)$ by nonparametrically regressing $y_i - X_i\hat{\beta}$ onto Z_i

Before proceeding with the estimation, we need to set a condition for identification. The requirement to identifying β involves a rank condition on $\widehat{X}$. For the second-step estimator on the β to be defined, we need $\widehat{X}$ to be a full column rank matrix. This is broken down when any linear combination of $X\beta$ is a deterministic function of variables in Z, or if elements of X is perfectly predicted by Z. In such case, the $\widehat{X}$ will be comprised of some columns that are not linearly independent and the $(\widehat{X}'\widehat{X})$ will not have an inverse function.

As far as approximation is concerned, β follows the properties of parametric estimators. That is, it converges at rate $n^{-1/2}$ if $\dim(Z_i) < 4$. The reason behind this is kernel averaging. Note that

$$\hat{\beta} = (\widehat{X}'\widehat{X})^{-1}(\widehat{X}'\widehat{Y})$$

This involves sums such as $X'\widehat{E}[Y|Z]$, which is

$$X'\widehat{E}[Y|X] = \sum_i X_i \widehat{E}[Y|Z = Z_i] = \sum_i X_i \frac{\sum_j K\left(\frac{Z_j - Z_i}{h}\right) Y_j}{\sum_j K\left(\frac{Z_j - Z_i}{h}\right)}$$

As a whole, $\hat{\beta}$ involves averaging this over all possible values of X_i with

$$\frac{1}{n}\sum_i X_i \frac{\sum_j K\left(\frac{Z_j - Z_i}{h}\right) Y_j}{\sum_j K\left(\frac{Z_j - Z_i}{h}\right)}$$

which reduces the nonparametric estimation variance by factors of n. It is better to undersmoothe here and get a nonparametric root mean squared error (or standard error) of order lower than $n^{\frac{-4}{4+d}}$ for $\widehat{X}'\widehat{Y}$. For $d < 4$, this is faster than $n^{-\frac{1}{2}}$, which is the parametric rate of convergence. This is the part where semiparametric estimation achieves better efficiency than nonparametric ones.

Unfortunately, estimating $g(\cdot)$ follows the same properties as nonparametric estimators. It has a slower convergence rate, which becomes even slower with more dimensions of Z_i. This part is done by nonparametrically regressing $y_i - X_i\hat{\beta}$ onto Z_i

1.3 Single index models

In a single index model, the conditional mean function $E[Y|X]$ is written as

$$E[Y|X] = F_0(X\beta)$$

where the unknowns are the β parameter in the index $X\beta$ and F_0, which we will not set any restrictions on. These two are the parameters of interest in our regression. We proceed by estimating for β and then for F_0.

Identification of a β parameters in the single index models rests on the idea that the "marginal rate of substitution" is independent of X. Let X_j, X_k be the two continuously distributed variables in X and that $m(x) \equiv E[Y|X]$. Then what we can get is the following relation

$$\frac{m_j'(X)}{m_k'(X)} = \frac{F_0'(X\beta)\beta_j}{F_0'(X\beta)\beta_k} = \frac{\beta_j}{\beta_k}$$

This relation gets us several estimators for $\frac{\beta_j}{\beta_k}$: The ratio of the β estimates. To pin down

an estimate for a particular β, we normalize one of the estimates to 1, say β_1 - similar to designating one good as a numeraire good in microeconomics. Also, we choose some value of X, say an average value X_0. Then, we can obtain the average derivative estimator

$$\hat{\beta}_j = \frac{m_j'(X_0)}{m_1'(X_0)}$$

This is simple in concept but not necessarily the best thing. It depends on the kernel estimation of the conditional mean function and the parameters depend on the choice of which variable to normalize on.

Another way of doing this is to apply a density-weighted average derivative estimator which has the form

$$\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} y_i \hat{f}_n'(X_i)$$

where $\hat{f}_n$ is the density estimate of X and $\hat{f}_n'(X_i)$ is the estimate for the derivative of f. Thus, we can write

$$\hat{\beta} = \frac{1}{n(n-1)}\sum_{i=1}^{n} y_i \frac{1}{h^{\dim(X)+1}}\sum_{j\neq i} K'\left(\frac{X_i - X_j}{h_n}\right)$$

The justification for this is that since $\frac{m_j'(X)}{m_k'(X)} = \frac{\beta_j}{\beta_k}$ for all X, we can also claim that the equality is preserved when we integrate by multiplying any $g(x)$ function and then integrating over X, or $\frac{\int m_j'(x)g(x)dx}{\int m_k'(x)g(x)dx} = \frac{\int \beta_j g(x)dx}{\int \beta_k g(x)dx}$. Also, using integration by parts, and the fact that

$$(m(x)g(x))' = m'(x)g(x) + m(x)g'(x)$$

holds, we can write

$$\begin{aligned}\int m_j'(x)g(x)dx &= m(x)g(x)]_{-\infty}^{\infty} - \int m(x)g_j'(x)dx \\ &= -\int m(x)g_j'(x)dx \ (\because m(\cdot) \text{ continous for all components of } x.)\end{aligned}$$

Here, we take $g(x) = -\frac{f^2(x)}{2}$. This results in $g'(x) = -f(x)f'(x)$ and

$$-\int m(x)g_j'(x)dx = \int m(x)f'(x)f(x)dx = E[m(x)f'(X)] = E[E[Y|X]f'(X)] = E[Yf'(X)]$$

From here, the sample analogue of this value will become our density-weighted average derivative estimator. This would be our estimator of β.

To estimate F_0, calculate the index $\widehat{Z} = X\hat{\beta}$. Then, we nonparametrically estimate Y onto $\widehat{Z}$ to obtain $\widehat{F}$, which estimate F_0. This is a purely nonparametrical part, which results in slower convergence. However, the index allows us to effectively reduce the dimensions involved in this step to 1, avoiding much of the curse of dimensionality.

1.4 Specification testing

We can formulate a test comparing parametric a pair out of the three possible distributions - parametric, nonparametric, and semiparametric. So we can test parametric against semiparametric, semiparametric against nonparametric, and so on. To construct the test, we first estimate a candidate model $\tilde{m}(x, \hat{\theta})$ and then come up with an estimation for the test model $\hat{m}_n(x)$ (nonparametric one, for example). Then the test statistic would be based on an absolute of

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{m}(x_i, \hat{\theta}))(\hat{m}_n(x_i) - \tilde{m}(x_i, \hat{\theta}))$$

and test this against a limiting distribution of the test statistic under the null (different depending on the setup of the null).

2 Treatment Effects

2.1 Causality and Treatment Assignment

Now, our interest is in finding out whether some X variable **causes** Y, not just whether they are correlated. Generally, correlation is not enough to argue causation. Suppose that you find that $cov(X, Y) \neq 0$. This can happen because

- X do cause Y, which is good for us. But..
- Y could also cause X. So there is a reverse causality bias here
- Z mutually affects X and Y. This is an omitted variable bias and leads to nonzero correlation even if X and Y has no connection whatsoever.

In any study examining program evaluation or quasi-experimental setting, the key issue is the assignment of the treatment, which is considered binary for the moment. There are three possible cases, which are

- **Random Assignment**: This would be a case when we can guarantee that the assignment to the treatment arms (treatment and control) are determined by chance - like flipping a coin or random draw of numbers. This rarely happens in social science except for randomized control trials, as people can voluntarily opt in and out of treatment.
- **Selection on Observables**: The treatment assignment is effectively random once we condition on some observable covariates. This is also known as ignorability or unconfoundedness assumptions.
- **Selection on Unobservables**: The assignment depends fundamentally on unobservables, or in other words, we cannot break down the dependence structure of assignment using observed variables. For instance, even in a clinical trials that involve voluntary participation, it is very likely that (lack of) risk averseness determines participation. How do we control for risk averseness?

2.2 Framework for Treatment Effect and Various Parameters of Interest

We will consider a binary treatment variable - whether individual i received a treatment or not (in the treatment group or control group). Define a variable D_i s.t.

$$D_i = \begin{cases} 1 & \text{If treated} \\ 0 & \text{If not treated} \end{cases}$$

In the above notation, i indexes the unit of the treatment - an individual, household, county, firm, and etc. For each unit i, there are two possible outcomes. One is the outcome without treatment, $y_i(0)$. The other is the outcome with treatment, $y_i(1)$. This can be seen as a counterfactual framework in the sense that if we get $y_i(1)$ for unit i, we cannot get $y_i(0)$ and vice versa. We always have a missing data problem in this regard. A useful way of writing this is

$$y_i = D_i y_i(1) + (1 - D_i) y_i(0)$$

The left hand side is an outcome for unit i, treated or untreated. *This is observed for every i*. The right hand side is meant to capture that the observed outcome for individual i is either *one* of $y_i(1)$ or $y_i(0)$. This framework is sometimes called **potential outcome framework** and will come in handy when we analyze various treatment effect results. Also, our understanding of the treatment assignment process - whether this was random, selection on observables,

or selected based on unobservables - determine which parameter of the treatment we can estimate.

We are interested in how an outcome for unit i changes between treatment and control, or how does the outcome for each i change when it is treated (vs. in control)? In other words,

$$TE_i = y_i(1) - y_i(0)$$

Another parameter of interest could be the treatment effect averaged over those who share the common covariate value, or what happens if we shift all i's with $X_i = x$ from none treated to all treated. This is

$$TE(x) = E(TE_i|X_i = x) = E(y_i(1) - y_i(0)|X_i = x)$$

We could also be interested in the treatment effect averaged over the population. This is the average treatment effect, defined as

$$ATE = E(TE_i) = E(y_i(1) - y_i(0))$$

Of course there are more, most of which we will learn over the course of this sequence. There are average treatment effect on the treated, ATE for the untreated, intent to treat estimate (ITT, in case of imperfect treatment compliance) and so on. For the average treatment on the treated, we can write

$$ATT = E(TE_i|D_i = 1)$$

Now that we have defined some parameters of interest, we go back to the missing data problem and how that affects our estimation of the treatment effects. We are forced to make an assumption on the things that we cannot observe. Take the ATE for example. We write

$$\begin{aligned} E[y_i(1) - y_i(0)] &= E[y_i(1)] - E[y_i(0)] \\ &= \{\Pr(D_i = 1)E[y_i(1)|D_i = 1] + (1 - \Pr(D_i = 1))E[y_i(1)|D_i = 0]\} \\ &\quad - \{\Pr(D_i = 1)E[y_i(0)|D_i = 1] + (1 - \Pr(D_i = 1))E[y_i(0)|D_i = 0]\} \end{aligned}$$

Here is the problem: We can get what $E[y_i(1)|D_i = 1]$ and $E[y_i(0)|D_i = 0]$ are from the data.

This is because

$$E[y_i|D_i = 1] = E[1 \cdot y_i(1) - (1-1) \cdot y_i(0)|D_i = 1] = E[y_i(1)|D_i = 1]$$
$$E[y_i|D_i = 0] = E[0 \cdot y_i(1) - (1-0) \cdot y_i(0)|D_i = 0] = E[y_i(0)|D_i = 0] \quad \text{(TE)}$$

So we can get $E[y_i(1)|D_i = 1]$ and $E[y_i(0)|D_i = 0]$ from the data - take the expected value of observed y_i conditional on $D_i = 1$ and 0. We cannot do the same for $E[y_i(1)|D_i = 0]$ and $E[y_i(1)|D_i = 0]$. This forces us to make an assumption about what the two unobserved values could be.

Before proceeding any further, we also assume that the that an observation is only affected by its own treatment, or stable unit treatment value assumption (SUTVA). This implies that there is no un-modeled spillovers (or no equilibrium effects). Mathematically speaking, we write

$$y_i \perp\!\!\!\perp (D_{-i}, y_{-i})|D_i$$

where $-i$ indicates treatment unit without i. This can be broken when the experiment is scaled up, or when we extrapolate small-scale (e.g. village level experiment) to a larger-scale experiment (e.g. state level experiment). Think about vaccination, or any other health related treatment. Thus, external validity of the treatment outcome should be interpreted with caution.

2.3 Random Assignment

We are back to having to assume what $E[y_i(1)|D_i = 0]$ and $E[y_i(1)|D_i = 0]$ are. A **random assignment** assumes that the outcome is independent of the treatment status. More formally

$$(y_i(0), y_i(1)) \perp\!\!\!\perp D_i \quad \text{(RA)}$$

This implies that

$$E[y_i(1)] = E[y_i(1)|D_i = 1] = E[y_i(1)|D_i = 0]$$

and similarly for $E[y_i(0)]$. If we go back to the (TE) equation, the problem was that we were unable to make any statement about $E[y_i(1)|D_i = 0]$ and $E[y_i(1)|D_i = 0]$. Now what we are

doing is to equate (first equality from data, second from RA assumption)

$$E[y_i|D_i = 1] = E[y_i(1)|D_i = 1] = E[y_i(1)|D_i = 0]$$
$$E[y_i|D_i = 0] = E[y_i(0)|D_i = 0] = E[y_i(0)|D_i = 1]$$

and effectively make the statements about the previously unobserved values. In practice, researchers conduct a balance test to approach this. This allows us to rewrite the ATE as

$$\begin{aligned} E[y_i(1) - y_i(0)] &= E[y_i(1)] - E[y_i(0)] \\ &= E[y_i(1)|D_i = 1] - E[y_i(0)|D_i = 0] \ (\because \text{RA}) \\ &= E[y_i|D_i = 1] - E[y_i|D_i = 0] \ (\because \text{TE}) \end{aligned}$$

Therefore, we can estimate ATE by mapping $E[y_i(1)]$ to $E[y_i|D_i = 1]$, $E[y_i(0)]$ to $E[y_i|D_i = 0]$. We can obtain this using many methods, even nonparametric. Even an OLS would get us this estimate; Take the following setup.

$$y_i = \beta_0 + \beta_1 D_i + e_i$$

assuming $E[e_i|D_i] = 0$, which is a random assignment. In the above context

$$\begin{aligned} &E[y_i|D_i = 0] = \beta_0, \ E[y_i|D_i = 1] = \beta_0 + \beta_1 \\ \implies &E[y_i|D_i = 1] - E[y_i|D_i = 0] = \beta_1 \end{aligned}$$

This implies that the ATE can be derived by estimating β_1 with a simple OLS.

2.4 Conditional Independence Assumption

In this setup, we are assuming that conditional on X_i, the outcomes and D_i are independent. Formally, we can write

$$(y_i(0), y_i(1)) \perp\!\!\!\perp D_i|X_i \tag{CIA}$$

We can alternatively conceptualize this framework in the following way: For $y_i(d)$, where $d \in \{0, 1\}$, we write

$$y_i(d) = \mu(X_i, d) + \epsilon_i(d), \ \ E[e_i(d)|X_i] = 0 \ \forall d$$

As for D_i, write

$$D_i = \mathbb{1}[u_i < p(X_i)]$$

where $u_i \equiv U[0,1]$ determines selection and $p(X_i)$ can be interpreted as a propensity score - how much are you likely to be treated given that you have covariates specified as some X_i. You can understand $[u_i < p(X_i)]$ as determining the size of the treatment region - the larger the $p(X_i)$, the likelihood of individual i being treated increases. This framework, along with the (CIA), allows us to say the following:

$$\begin{aligned}
E[y_i(1)|X_i = x] &\implies E[y_i(1)|D_i = 1, X_i = x] = E[y_i(1)|D_i = 0, X_i = x] \ (\because \text{CIA}) \\
&\implies E[\mu(x,1) + \epsilon_i(1)|1, x] = E[\mu(x,1) + \epsilon_i(1)|0, x] \\
&\implies E[\mu(x,1)|1, x] + E[\epsilon_i(1)|1, x] = E[\mu(x,1)|0, x] + E[\epsilon_i(1)|0, x] \\
&\implies E[\epsilon_i(1)|1, x] = E[\epsilon_i(1)|0, x] \ (\because \mu(x,1) \text{ is same for both cases}) \\
&\implies E[\epsilon_i(1)|u_i < p(x), x] = E[\epsilon_i(1)|u_i > p(x), x] \\
&\implies (\epsilon_i(1), \epsilon_i(0)) \perp\!\!\!\perp u_i|X_i \qquad \text{(CIA2)}
\end{aligned}$$

So we can write conditional independence assumption as $(\epsilon_i(1), \epsilon_i(0)) \perp\!\!\!\perp u_i|X_i$, meaning that the unobserved assignment element u_i is independent of the unobserved element in outcome process $(\epsilon_i(1), \epsilon_i(0))$ once observables x_i's are controlled for.

There is one additional caveat to this condition. It must be that $p(X_i) \in (0,1)$ - it should not be 0 or 1. If $p(X_i) = 1$, then for every possible u_i, $D_i = 1$ - everyone gets treated. Therefore, there is no meaningful statement we can make about the $D_i = 0$ case. We can say similarly for $p(x_i) = 0$ case. In some textbooks, this is also known as **overlap** assumption - each unit in the defined population should have some chance of being treated and some chance of being in the control group. [1] Propensity scores matter in that it captures the idea that the dependence between treatment assignment and treatment effects only occur through the probability of the treatment, determined by the values of X_i's. It also reduces dimension reduction in that propensity score expresses information from many dimensions of X_i into a single number.

What we want to learn involves understanding the joint distribution of the variables $y_i(0)$ and $y_i(1)$. Take the treatment effect at $X_i = x$, written as

$$TE(x) = \mu(x,1) - \mu(x,0)$$

This is the effect of the treatment on the outcome y_i for $X_i = x$. The interpretation is that if

[1] If both (CIA) and overlap is satisfied, we say that have **strong ignorability** (Rosenbaum, Rubin 1983).

we shift everyone with $X_i = x$ from the control group to the treatment group, the average outcome increases by $TE(x)$. With the CIA framework, we can write

$$\begin{aligned}
E[y_i|1,x] - E[y_i|0,x] &= E[\mu(x,1) + \epsilon_i(1)|1,x] - E[\mu(x,0) + \epsilon_i(0)|0,x] \\
&= E[\mu(x,1)|1,x] + E[\epsilon_i(1)|1,x] - E[\mu(x,0)|0,x] - E[\epsilon_i(0)|0,x] \\
&= \mu(x,1) + E[\epsilon_i(1)|x] - \mu(x,0) - E[\epsilon_i(0)|x] \ (\because \text{CIA2}) \\
&= \mu(x,1) - \mu(x,0) \ (\because E[\epsilon_i(d)|X_i] = 0 \ \forall d)
\end{aligned}$$

It is also possible to write this as $TE(x) = E[TE_i|X_i = x]$, where $TE_i = y_i(1) - y_i(0)$.

$$\begin{aligned}
E[y_i(1) - y_i(0)|X_i = x] &= E[y_i(1)|X_i = x] - E[y_i(0)|X_i = x] \\
&= (\Pr(1|x) \cdot E[y_i(1)|1,x] + (1 - \Pr(1|x))E[y_i(1)|0,x]) \\
&\quad - (\Pr(1|x) \cdot E[y_i(0)|1,x] + (1 - \Pr(1|x))E[y_i(0)|0,x]) \\
&= E[y_i(1)|1,x] - E[y_i(0)|0,x] \ (\because \text{CIA}) \\
&= E[y_i|1,x] - E[y_i|0,x]
\end{aligned}$$

This implies that if CIA condition holds, $TE(x)$ can be easily estimated by

$$\hat{\mu}(1,x) - \hat{\mu}(0,x)$$

for some estimator. CIA assumption allows us to map $E[y_i(1)|X_i = x]$ to $E[y_i|D_i = 1, X_i = x]$ and $E[y_i(0)|X_i = x]$ to $E[y_i|D_i = 0, X_i = x]$, which is impossible without it. This can be done nonparametrically or even with some OLS with controls. Write

$$y_i = \beta_0 + \beta_1 D_i + X_i\gamma + e_i$$

where X_i is a set of controls. Assuming $E[e_i|D_i, X_i] = 0$, we get

$$\begin{gathered}
E[y_i|D_i = 0, X_i = x] = \beta_0 + x\gamma, \ E[y_i|D_i = 1, X_i = x] = \beta_0 + \beta_1 + x\gamma \\
\implies E[y_i|D_i = 1, X_i = x] - E[y_i|D_i = 0, X_i = x] = \beta_1
\end{gathered}$$

So β_1 captures our $TE(x)$.

More efficient and/or robust estimators of $TE(x)$ uses a first step estimator of the propensity score. Define $\hat{p}_n(x) = \Pr(D_i = 1|X_i = x)$ as our estimator of the propensity score. The most semiparametrically efficient estimator of $E(TE(x)|x \in A)$ is the **inverse probability weighting** estimator. The steps are as follows

1. Estimate the propensity score $p(X_i)$ by computing (nonparametrically, or parametrical methods such as LPM/logit/probit)

$$\hat{p}_n(X_i) = \Pr(D_i = 1 | X_i = x)$$

2. Use the following formula to estimate $E(TE(x)|x \in A)$:

$$\frac{\sum_{D_i=1, x_i \in A} a_i y_i}{\sum_{D_i=1, x_i \in A} a_i} - \frac{\sum_{D_i=0, x_i \in A} b_i y_i}{\sum_{D_i=0, x_i \in A} b_i}$$

where $a_i = \frac{1}{\hat{p}_n(x_i)}, b_i = \frac{1}{1-\hat{p}_n(x_i)}$. a_i puts more weight on the least likely treated (vice versa for b_i) and balances the distribution out, leading to lesser variance.

The above method involves summation over different domains. There is another way to work this out that is regression friendly. The above can also be written as

$$\frac{\sum_{i=1, x_i \in A}^{N} \frac{D_i y_i}{\hat{p}(X_i)}}{\sum_{i=1, x_i \in A}^{N} \frac{D_i}{\hat{p}(X_i)}} - \frac{\sum_{i=1, x_i \in A}^{N} \frac{(1-D_i) y_i}{1-\hat{p}(X_i)}}{\sum_{i=1, x_i \in A}^{N} \frac{1-D_i}{1-\hat{p}(X_i)}}$$

The reason this works is as follows: Notice that

$$\begin{aligned} D_i y_i &= D_i(D_i y_i(1) + (1 - D_i) y_i(0)) = D_i y_i(1) \\ (1 - D_i) y_i &= (1 - D_i)(D_i y_i(1) + (1 - D_i) y_i(0)) = (1 - D_i) y_i(0) \end{aligned}$$

Thus, we have

$$\begin{aligned} E\left[\frac{D_i y_i}{p(X_i)}\right] &= E\left[E\left[\frac{D_i y_i(1)}{p(X_i)} | X_i\right]\right] \\ &= E\left[\frac{E[D_i|X_i] E[y_i(1)|X_i]}{p(X_i)}\right] \\ &= E\left[\frac{p(X_i) E[y_i(1)|X_i]}{p(X_i)}\right] = E[E[y_i(1)|X_i]] = E[y_i(1)] \end{aligned}$$

so we have $E\left[\frac{D_i y_i}{p(X_i)}\right] = E[y_i(1)], E\left[\frac{D_i y_i}{p(X_i)} | X_i\right] = E[y_i(1)|X_i]$. We can make the similar arguments for $y_i(0)$ and get $E\left[\frac{(1-D_i) y_i}{1-p(X_i)}\right] = E[y_i(0)], E\left[\frac{(1-D_i) y_i}{1-p(X_i)} | X_i\right] = E[y_i(0)|X_i]$. We can also

show that the denominator terms average to 1. Thus, the ATE for $X_i \in A$ can be written as

$$\frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i y_i}{p(X_i)} - \frac{(1-D_i)y_i}{1-p(X_i)}\right) \ \forall X_i \in A$$

A slightly less efficient but 'safer' estimation method for $TE(x)$ is a **doubly robust estimator**. We define the following setup

$$\Psi = \mu(1,X) - \mu(0,X) + \frac{D}{p(X)}(y(1) - \mu(1,X)) - \frac{1-D}{1-p(X)}(y(0) - \mu(0,X))$$

The key is that $E[\Psi|X=x] = TE(x) = E[TE_i|X=x]$ if just one of the two conditions hold

1. $p(x) = \Pr(D=1|X=x)$
2. $\mu(d,x) = E[y(d)|X=x]$ for $d \in \{0,1\}$

We can see this from

$$E[\Psi|X=x] = \mu(1,x) - \mu(0,x) + \frac{E[D|x]}{p(x)}(E[y(1)|x] - \mu(1,x)) - \frac{1-E[D|x]}{1-p(x)}(E[y(0)|x] - \mu(0,x))$$

If condition 1 holds, we have $p(x) = \Pr(D=1|X=x) = E[D|x]$ so $E[\Psi|X=x]$ reduces to $E[y(1)|x] - E[y(0)|x] = E[y(1) - y(0)|x]$. If condition 2 holds, we have $E[y(1)|x] = \mu(x,1)$, and $E[y(0)|x] = \mu(x,0)$. Thus $E[\Psi|X=x] = \mu(1,x) - \mu(0,x)$. This estimator is safer in the sense that even if one of the two is misspecified, we can still back out a consistent estimator.

2.5 Matching Estimation

The idea behind matching estimation is to impute the values for something we cannot get from the data: Namely, $Y_i(0)$ for those in $D_i = 1$ and $Y_i(1)$ for those in $D_i = 0$. In other words, you try to find a 'close' match for a particular unit i in a different treatment arm. When we say we find k-closest neighbors for unit i in $D_i = 0$, we find k individuals in $D_i = 1$ that has close traits (X_i) to individual i, or k individuals with the smallest values of $||x_j - x_i||$. Then, we construct a counterfactual $Y_i(1)$ by taking a (weighted) average over the Y_i's of the k individuals found in the other group. The treatment effect would than be $Y_i(1) - Y_i$, where $Y_i(1)$ is calculated as in previous sentence.

Let's try with $k = 1$ - we find the one individual in the opposite treatment arm that has

the similar value of X_i. Then, the average treatment effect can be written as

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_i(1) - \hat{Y}_i(0))$$

where

$$\hat{Y}_i(1) = \begin{cases} Y_i & (D_i = 1) \\ \text{imputed value} & (D_i = 0) \end{cases}, \quad \hat{Y}_i(0) = \begin{cases} \text{imputed value} & (D_i = 1) \\ Y_i & (D_i = 0) \end{cases}$$

There are some caveat in this approach

- The covariate X_i that is used to find the closest match in the other treatment arm should not affect the assignment of the treatment. In other words, if the assignment of treatment vs. control depends on some subset of X_i, you should not use that subset as a standard for finding the closest match.
- The overlap condition becomes critical. If it is not satisfied, i.e. for some X_i, $p(X_i) = 1$ or 0, then we are unable to find a closest match in the other treatment arm because for individuals with that covariate value, all of them are either in control or treatment group and not spread around.
- Who are the neighbors? The answer might depend on what distance measure we use. Moreover, how many of those in the treatment can be considered neighbors? There is a trade-off in the sense that using larger k would force us to put someone who is not 'close' as neighbors and using small k may cause difficulty in imputing counterfactual values.

Example 2.1 (Di Tella, Gálvez, Schargrodsky (2021, NBER WP 29458))**.** *This paper tests the effect of an echo-chamber on how people form political opinions and preferences. The treatment mimics an exposure to a counter-attitudinal information related to politics (presidential debate vs neutral history of Argentine political institutions) on subjects under different degrees of echo-chamber (those who is allowed twitter vs not). The study finds that those who started inside the echo chambers are further polarized and stressed (measured through cortisol levels) when they are subject to both twitter treatment and being exposed to a counter-attitudinal data. This paper thus highlights how echo-chamber of being exposed to selective posts on social networking services may contribute to polarization.*

Example 2.2 (Baird, McIntosh, Özler (2011, QJE))**.** *This paper assesses the role of cash transfers conditional on attending school (over unconditional cash treatment and pure controls) on educational performance and teenage pregnancy and marriage rates using a field experiment on adolescent girls in Malawi. The experiment finds that those under CCT outperformed other groups in the English reading comprehension. However, pregnancy rates and marriage rates were lower in the UCT over CCT, which is apparently driven by girls who dropped out of school. This result highlights that while CCTs are more effective in achieving some behavioral change, UCTs do still improve outcomes for people who fail to meet the conditions.*
(Note: The same authors did a follow-up study for their 2019 JDE paper. CCT produced some sustained gains in education and fertility outcomes but not others. The health benefits from UCTs dried out once the financial support ended.)