Recitation 10: Semiparametrics, and treatment effects

Seung-hun Lee

Columbia University
Introduction to Econometrics II Recitation

April 11th, 2022

Semiparametrics

Happy medium between nonparametrics and parametrics

- Key is the tradeoff between robustness and efficiency
  - Nonparametrics is robust: CAN-ness of these estimation does not depend on getting the assumption about specification right, but converges very slowly
  - Parametrics does better in efficiency: Always converge to the true distribution at a nice speed of $\sqrt{n}$ but vulnerable to specification assumption
- Where does semiparametrics stand
  - By design, contains both semiparametrics and parametrics
  - Robust in larger class of distribution than in parametric estimation but not as much as in nonparametrics
  - Contains part of the estimation that is as efficient as parametrics, but also has inefficient parts arising from nonparametric aspects

Partially linear models: Split equation to parametric/nonparametrics

- Partition the covariates into two unoverlapping spaces - X and Z, where $E[\epsilon|X, W] = 0$.
- We work with the DGP of the following form

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

- Parameter of interest is: β and $g(\cdot)$
- Idea for β: Use the fact that

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

3-steps approach

- 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
- Conditions for identification
  - $\widehat{X}$ needs 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

β estimates can perform as well as parametrics

- It converges at rate $n^{-1/2}$ if $\dim(Z_i) < 4$
- Why? Kernel averaging
  - Since $\hat{\beta} = (\widehat{X}'\widehat{X})^{-1}(\widehat{X}'\widehat{Y})$, we have sums like

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

  - $\hat{\beta}$ involves averaging this over all possible values of X_i with $\frac{1}{n}\sum_i X_i \frac{\sum_j K\left(\frac{Z_j - Z_i}{h}\right) Y_j}{\sum_j K\left(\frac{Z_j - Z_i}{h}\right)}$
  - This is known to reduce 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 $n^{-\frac{1}{2}}$

$g(\cdot)$ resembles nonparametrics

- 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

Single index models: Parametrics used for dimension reduction

- 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 distribution
- Conceptual approach for β: Use the MRS idea
  - 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

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

  - We can identify β_j up to a scale: Pin down one β_j to 1 and calculate others using the ratio
  - Dependent on kernel estimation of conditional mean and choice of variable to normalize

Density weighted average derivative estimator

- It 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

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

Justifying density weighted average derivative estimator

- Since $\frac{m_j'(X)}{m_k'(X)} = \frac{\beta_j}{\beta_k}$ for all X, $\frac{\int m_j'(x)g(x)dx}{\int m_k'(x)g(x)dx} = \frac{\int \beta_j g(x)dx}{\int \beta_k g(x)dx}$
- Using integration by parts, we can write

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

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

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

- DWADE is the sample analogue

F_0 is the nonparametric part

- 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.

Treatment effect

We shift our attention to causal interpretation

- Now, our interest is in finding out whether some X variable **causes** Y
- 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.
- One key to causal relation is treatment assignment:
  - **Random Assignment**: Assignment to the treatment and control is determined by chance, rare in social science except RCTs, 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 (ignorability, unconfoundedness assumptions)
  - **Selection on Unobservables**: The assignment depends on unobservables, so we cannot break down the dependence structure of assignment using observed variables.

Observations vs Potential outcomes

- Define a variable D_i indicating treatment assignment s.t.

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

- i indexes the unit of the treatment - an individual, household, county, firm, and etc
- Outcome we observe can be broken into a sum of counterfactual potential outcomes

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

  Where we have
  - y_i: The observed outcome for *every* i
  - $y_i(1), y_i(0)$: Outcome for those whose $D_i = 1$ and 0, respectively
  - Counterfactual? If an i has $y_i(1)$, there is no $y_i(0)$ - fundamental problem of missing data

Outcome of interest: The treatment effect (TE)

- Ideally: $TE_i = y_i(1) - y_i(0)$
- TE averaged over those who share the common covariate value

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

- TE averaged over population of interest (ATE)

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

- Others: ATE for the treated/untreated, intent-to-treat (ITT, in case of imperfect treatment compliance)

Fundamental problem of missing data

- Take the ATE for example. We write

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

- 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

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

- We cannot do the same for $E[y_i(1)|D_i = 0]$ and $E[y_i(1)|D_i = 0]$.
- We are forced to make an assumption on the things that we cannot observe.

Random assignment: Untreated and treated are practically identical!

- Formally, we write

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

- This implies that (similarly for $E[y_i(0)]$)

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

- With this,

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

- ATE is now

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

Random assignment: ATE obtainable with OLS

- 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 ($E[e_i|D_i] = 0$)

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

- In the above context

$$E[y_i|D_i = 0] = \beta_0, \; E[y_i|D_i = 1] = \beta_0 + \beta_1$$
$$\implies E[y_i|D_i = 1] - E[y_i|D_i = 0] = \beta_1$$

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

Conditional Independence: Controlling for X_i

- Conditional on X_i, the outcomes and D_i are independent.

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

- Alternatively, write

$$y_i(d) = \mu(X_i, d) + \epsilon_i(d), \quad E[\epsilon_i(d)|X_i] = 0 \ \forall d \in \{0, 1\}$$

- The treatment assignment follows $D_i = \mathbb{I}[u_i < p(X_i)]$, where $p(X_i) \in (0, 1)$ is propensity score (overlap!), $[u_i < p(X_i)]$ determines size of the treatment region
- This gets us

$$(\epsilon_i(1), \epsilon_i(0)) \perp\!\!\!\perp u_i | X_i \tag{CIA2}$$

  unobserved assignment element is independent of unobserved outcome errors once X_i's are controlled

Treatment effect for $X_i = x$

- $TE(x) = \mu(x,1) - \mu(x,0)$
- Using CIA2 assumption, we can write

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

- If we stick to CIA, we get

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

Estimating $TE(x)$

- If CIA condition holds, $TE(x)$ can be easily estimated by

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

- 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 $E[y_i|D_i = 0, X_i = x]$
- 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

$$E[y_i|D_i = 0, X_i = x] = \beta_0 + x\gamma, \; E[y_i|D_i = 1, X_i = x] = \beta_0 + \beta_1 + x\gamma$$
$$\implies E[y_i|D_i = 1, X_i = x] - E[y_i|D_i = 0, X_i = x] = \beta_1$$

So β_1 captures our $TE(x)$.

Inverse probability weighting: More efficient version

- 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.

- An alternative, but more summation friendly version is in the notes!

Doubly robust estimator: Safety net

- We use this setup

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

- Slight deviation from the class notes: Justifiable and easier for intuition (also useful)

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

- 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 $d \in \{0, 1\}$

Consistency of just one parameter is enough!

- 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.

Matching: Find an observationally equivalent y_i from the other arm

- Impute $Y_i(0)$ for those in $D_i = 1$ and $Y_i(1)$ for those in $D_i = 0$ based on closest match of covariates
- General: To 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 based on smallest values of $||x_j - x_i||$
- 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 $Y_i(1) - Y_i$,
- Note: Overlap is especially crucial! If X_i determines the treatment assignment, avoid using this in finding the neighbors

RCT examples: So many!

- There are many RCTs you can find, especially in (micro)development and behavioral economics conducting field/lab-in-the-field/lab expermients
  - See Kremer, Miguel (2004 ECMA) on how deworming treatment affects education and health (and what to do for dealing with treatment externalities)
  - Many anti-poverty/personnel programs and are assessed in the field experiment: See Baird et al (2011 QJE, in the notes) or Bandiera et al (2021 QJE, Michael and Andrea are coauthors here)
  - For those interested in theory (especially information economics and preference formation): Experiments on how echo-chamber affects opinions, check Di Tella et al (2019 NBER WP, in the notes)
- Always clarify the balance of covariates between treated and the untreated: Ideal result is a balance that is 'as good as random'.