Introduction to Econometrics 2: Recitation 10

Seung-hun Lee

Columbia University

April 15th, 2020[1]

[1] Fun fact: After this recitation, my gov't mandated self-isolation is officially over.

Semiparametric Regression

Semiparametric Regression: Framework

- Semiparametrics can be thought of as a middle ground between nonparametric and parametric regression.
- Suppose that we partition the covariates into two unoverlapping spaces - X and W.
- Also assume that $E(\epsilon|X, W) = 0$.
- One example of a semiparametric regression is a **partially linear regression** which has the following form

$$Y_i = X_i\beta + g(W_i) + \epsilon_i$$

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

Semiparametric Regression

Semiparametric Regression: Estimation

- To estimate β, we use the fact that

$$E[Y_i|W_i] = E[X_i|W_i]\beta + g(W_i)$$

- Given this, we can write

$$Y_i - E[Y_i|W_i] = \{X_i - E[X_i|W_i]\}\beta + \epsilon_i$$

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

Semiparametric Regression

Semiparametric Regression: Estimation

- β follows the properties of parametric estimators (Converges at rate $n^{-1/2}$ regardless of the dimensions of X_i, W_i)
- Estimating $g(\cdot)$ follows the same properties as nonparametric estimators (Slower convergence rate, which becomes even slower with more dimensions of W_i)
- Caveat: Identification of β requires an exclusion restriction
  - None of the components in X_i is perfectly predictable by W_i components ($X_i \neq E[X_i|W_i]$)
  - This would effectively rule out including a constant in the X_i part of the regression

Semiparametric Regression

Examples

- Horowitz, Lee (2002): The paper shows that semiparametrics allow more flexibility than parametric modeling and more precision than nonparametric models.
  - For fun (at least for a baseball nerd like me), this paper tests this idea on a data of salaries, runs, tenure of baseball players in 1987.
- Ucal et al (2010): This paper analyzes whether and to what extent the inflow of FDI is affected before and after the occurence of a financial crisis in developing countries using generalized partial linear models.
  - The results indicate that FDI inflows decrease in the years after a financial crisis and an upturn in FDI inflows the year before a financial crisis hit the country.

Treatment Effects

Finding Causality

- We are looking to see whether X causes y
- $cov(X, Y) \neq 0$ is not enough 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.
- The key issue is the assignment of the treatment, which could be..
  - **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
  - **Selection on Observables**: The treatment assignment is effectively random once we condition on some observable covariates
  - **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.

Treatment Effects

Theoretical Setup

- Consider a binary treatment variable - whether individual i received a treatment or not
- Define a variable D_i s.t.

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

- i indexes the unit of the treatment
- For each unit i, there are two possible outcomes.
  - The outcome without treatment, $Y_i(0)$
  - The outcome with treatment, $Y_i(1)$
- This can be seen as a counterfactual framework - 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

Treatment Effects

Theoretical Setup

- A mathematical way to treat 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 everyone
- 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**

Treatment Effects

Outcome of interest

- We are interested in how an outcome for unit i changes between treatment and 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, which is

$$TE(x) = E(TE_i | X_i = x)$$

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

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

- And others: ATT, ATUT, etc.

Treatment Effects

Problem caused by missing data

- We need to make an assumption about the things we cannot observe.
- Take the ATE for example. We can 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

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

- 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 $E[Y_i(1)|D_i = 0]$, forcing us to make assumptions

Treatment Effects

Characterize TE in an econometrics-friendly way

- Define the counterfactual outcomes as

$$Y_i(D_i = d) = \mu(X_i, d) + \epsilon_i(d)$$

  where d can take either $0, 1$.
- 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

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

- Interpretation: If we shift everyone with $X_i = x$ from the control group to treatment, the average outcome increases by $TE(x)$.
- We can also calculate ATE as

$$\begin{aligned} ATE &= E[Y_i(1) - Y_i(0)] = E[\mu(X_i, 1) - \mu(X_i, 0) + \epsilon_i(1) - \epsilon_i(0)] \\ &= E\left[E[\mu(X_i, 1) - \mu(X_i, 0) + \epsilon_i(1) - \epsilon_i(0)|X_i]\right] = E\left[E[\mu(X_i, 1) - \mu(X_i, 0)|X_i]\right] \\ &= E[\mu(X_i, 1) - \mu(X_i, 0)] = E[TE(X_i)] \end{aligned}$$

Treatment Effects

Random Assignments

- A **random assignment** assumes that the outcome is independent of the treatment status. More formally

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

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

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

- Now what we are doing is to equate

$$\begin{aligned} E[Y_i|D_i = 1] &= E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0] \\ E[Y_i|D_i = 0] &= E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1] \end{aligned}$$

- 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 $E[Y_i|D_i = 0]$.

Treatment Effects

Conditional Independence Assumption

- Assume conditional on X_i, the outcomes and D_i are independent.
- Formally, we can write

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

- Alternatively: Define D_i as

$$D_i = 1(u_i < p(X_i))$$

  $u_i \equiv U[0,1]$ determines selection and $p(X_i)$ can be interpreted as a propensity score
- Then we can also say

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

Treatment Effects

Conditional Independence Assumption

- Why?

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

- The caveat, however, is that $p(x) \in (0,1)$.
  - If $p(x) = 1$, then for every possible u_i, $D_i = 1$ - everyone gets treated and no meaningful statement can be made about the $D_i = 0$ case.
  - In some textbooks, this is also known as **overlap** assumption

Treatment Effects

TE under CIA

- We can write

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

- Note that

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

- Thus, we can back out the ATE for $X_i = x$ using the observables by mapping $E[Y_i(1)|x]$ to $E[Y_i|1,x]$ and $E[Y_i(0)|x]$ to $E[Y_i|0,x]$

Treatment Effects

TE under CIA: Regression Adjustments

- This is called a **regression adjustment** method.
- The treatment effect for $X_i = x$ using a regression adjustment can be obtained by utilizing the fact that

$$\mu(x,1) = E[Y_i|D_i = 1, X_i = x],\ \mu(x,0) = E[Y_i|D_i = 0, X_i = x]$$

  and regressing on the subsample of each treatment arm to get $\hat{\mu}(x,1)$ and $\hat{\mu}(x,0)$. Thus

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{\mu}(x,1) - \hat{\mu}(x,0))$$

Treatment Effects

TE under CIA: Inverse Probability Weight

- We can obtain the ATE using a different approach.
- This is an **inverse probability weighting**. The steps are as follows
  1. Estimate the propensity score $p(X_i)$ by computing

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

  2. Use the magic formula to estimate $E(TE(x)|x \in A)$: That is,

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

where $a_i = \frac{1}{\hat{p}_n(x_i)}, b_i = \frac{1}{1-\hat{p}_n(x_i)}$. The above can also be written as

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

Treatment Effects

TE under CIA: Inverse Probability Weight

- Notice that

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

- Thus, we have

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

- Thus, $E\left[\frac{D_i Y_i}{p(X_i)}\right] = E[Y_i(1)], E\left[\frac{D_i Y_i}{p(X_i)}|X_i\right] = E[Y_i(1)|X_i]$.
- We can make the similar arguments for $Y_i(0)$.

Treatment Effects

TE under CIA: Inverse Probability Weight

- Thus, the ATE for $X_i \in A$ can be written as

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

- In most cases, the propensity score should be estimated.
- So we use the estimator involving the $\hat{p}$, which is

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

- It is said that normalizing the weights to one in finite samples improves the mean squared error properties of the estimator (Imbens, Rubin (2019) - *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*)

TE under CIA: Matching

- 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.

TE under CIA: Matching

- 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

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

TE under CIA: Matching

- 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.
  - 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.

Treatment Effects

TE under CIA: DID

- This involves a specific framework where we can clearly define a 'before and after' - denoted as t_0 and t_1.
- No one is treated at t_0 but there is a subset of people ($G_i = 1$) that are treated at t_1. Those in $G_i = 0$ are never treated in either time period.
- If we define

  $$Y_{i,t} = G_i Y_i(1,t) + (1 - G_i) Y_i(0,t) \quad (t \in \{t_0, t_1\})$$

  and impose

  $$Y_i(G_i, t_0) = Y_i(t_0) \text{ for both } G_i = 1 \text{ and } G_i = 0$$

  then we will be able to observe $(Y_{i,t_1}, Y_{i,t_0}, G_i.X_i)$ for every unit i.
- What we do not observe is $Y_i(1, t_1)$ for those in $G_i = 0$ and $Y_i(0, t_1)$ for those in $G_i = 1$.

Treatment Effects

TE under CIA: DID

- The analogue to the conditional independence assumption in this context is a **parallel trend assumption**, defined as

$$(Y_i(1,t_1) - Y_i(t_0), Y_i(0,t_1) - Y_i(t_0)) \perp\!\!\!\perp G_i | X_i$$

- To see why they are equal, consider a setting where $D_i = G_i$ and write

$$\begin{aligned} Y_i = Y_{i,t_1} - Y_{i,t_0} &= D_i(\underbrace{Y_i(1,t_0) - Y_i(1,t_0)}_{Y_i(1)}) + (1-D_i)(\underbrace{Y_i(0,t_1) - Y_i(0,t_0)}_{Y_i(0)}) \\ &= D_i Y_i(1) + (1-D_i) Y_i(0) \end{aligned}$$

- Therefore, we can rewrite the parallel trend assumption as

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i | X_i$$

- Testing: Select a time period $\tilde{t} < t_0$ and find out if the difference $y_i(t_0) - y_i(\tilde{t})$ is independent with G_i

Treatment Effects

TE under CIA: DID

- To apply this in regression, we can write

$$Y_{it} = \beta_0(X_i) + \beta_1(X_i) \cdot 1(t = t_1) + \beta_2(X_i) \cdot G_i + \beta_3(X_i) \cdot G_i \cdot 1(t = t_1) + \epsilon_{it}$$

  where X_i is a set of covariates, which can include a constant.
- In this context, the treatment effect for $X_i = x$ would be

$$\begin{aligned} TE(x) &= E[Y_i(1) - Y_i(0)|X_i = x] \\ &= E[(Y_i(1, t_1) - Y_i(1, t_0)) - (Y_i(0, t_1) - Y_i(0, t_0))|X_i = x] \\ &= x \cdot E\{[(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2)] - [(\beta_0 + \beta_1) - (\beta_0)]|X_i = x\} \\ &= x \cdot E[(\beta_1 + \beta_3) - \beta_1|X_i = x] \\ &= \beta_3 x \end{aligned}$$

  So β_3 would be our parameter of interest.