Topics in Econometrics

Regression Discontinuity Designs

Instructor: Ma, Jun

Renmin University of China

May 26, 2022

Regression discontinuity (sharp) design

- Very long history: Thistlethwaite and Campbell (1960).
- Triplet: score, threshold, treatment.
- Suppose $D = 1$ if $X \geq c$ and $D = 0$ if $X < c$, i.e., treatment if triggered by some score X.
- X could have causal effect on the outcome.
- X satisfies the unconfounded assumption but fails the overlap assumption.

Identification

- Hahn, Todd and Van Der Klaauw (2008) showed that the "conditional average treatment effect" (CATE) $\mathrm{E}\left[Y(1)-Y(0) \mid X=c\right]$ is identified under very weak assumptions.
- Note that $\mathrm{E}\left[Y \mid X=x\right]=\mathrm{E}\left[Y(0) \mid X=x\right]$ for $x<c$ and $\mathrm{E}\left[Y \mid X=x\right]=\mathrm{E}\left[Y(1) \mid X=x\right]$ for $x \geq c$.
- Assume that $\mathrm{E}\left[Y(0) \mid X=x\right]$ and $\mathrm{E}\left[Y(1) \mid X=x\right]$ are continuous in x. Then,

$$\lim_{\epsilon>0, \epsilon \downarrow 0} \mathrm{E}\left[Y(0) \mid X=c-\epsilon\right]=\mathrm{E}\left[Y(0) \mid X=c\right]$$

$$\lim_{\epsilon>0, \epsilon \downarrow 0} \mathrm{E}\left[Y(1) \mid X=c+\epsilon\right]=\mathrm{E}\left[Y(1) \mid X=c\right]$$

and

$$\begin{aligned} & \lim_{\epsilon>0, \epsilon \downarrow 0} \mathrm{E}\left[Y \mid X=c+\epsilon\right]-\lim_{\epsilon>0, \epsilon \downarrow 0} \mathrm{E}\left[Y \mid X=c-\epsilon\right] \\ = & \lim_{\epsilon>0, \epsilon \downarrow 0} \mathrm{E}\left[Y(1) \mid X=c+\epsilon\right]-\lim_{\epsilon>0, \epsilon \downarrow 0} \mathrm{E}\left[Y(0) \mid X=c-\epsilon\right] \\ = & \mathrm{E}\left[Y(1)-Y(0) \mid X=c\right] . \end{aligned}$$

Nonparametric regression

- Let $(Y, X) \in \mathbb{R}^2$ be a random vector. We are interested in estimating $g(x) = \mathrm{E}[Y \mid X = x]$.
- Let (Y_i, X_i), $i = 1, ..., n$, be an i.i.d. sample.
- If X is finitely discrete, then

$$\widehat{g}(x) = \frac{\sum_{i=1}^{n} 1(X_i = x) Y_i}{\sum_{i=1}^{n} 1(X_i = x)}$$

  is a nonparametric estimator. $\widehat{g}(x)$ is consistent and asymptotically normal.
- As k-NN, for estimation of $g(x)$ with continuous X, we take average of observations that are "close" to x.
- k-NN: random distance, fixed number of observations.
- Alternative approach: fixed distance, random number of observations.

- Fix h and consider all observations with $|X_i - x| \leq h$.
- For $h > 0$,

$$\widehat{g}\,(x) = \frac{\sum_{i=1}^{n} 1\left(|X_i - x| \leq h\right) Y_i}{\sum_{i=1}^{n} 1\left(|X_i - x| \leq h\right)}.$$

- $\widehat{g}\,(x)$ is discontinuous. We use continuous weights instead to get a continuous estimator.
- Let $K : \mathbb{R} \to \mathbb{R}$ be a symmetric probability density function. Then,

$$\widehat{g}\,(x) = \frac{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}$$

  is the Nadaraya-Watson estimator.
- For consistency, we need $nh \uparrow \infty$ and $h \downarrow 0$ as $n \uparrow \infty$.
- Large h: smaller variance, more bias.

Local linear estimator

- The Nadaraya-Watson estimator is also called a local constant estimator:

$$\widehat{g}(x) = \underset{c}{\operatorname{argmin}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(Y_i - c)^2 .$$

- Without the weights $K((X_i - x)/h)$, the estimator reduces to the sample mean.
- Instead of approximating g locally as a constant, the local linear estimation approximates g locally by a linear function.
- The local linear estimator of $g(x)$:

$$\left(\widehat{g}(x), \widehat{g}'(x)\right) = \underset{g_0, g_1}{\operatorname{argmin}} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)(Y_i - g_0 - g_1(X_i - x))^2 .$$

- The local linear estimator has better properties at the boundary than the Nadaraya-Watson estimator.

Local linear estimation for RD

- Fit linear regressions (Imbens and Lemieux, 2008) to the observations within an h ($h \downarrow 0$ as $n \uparrow \infty$) distance:

$$\min_{\alpha_-,\beta_-} \sum_{i=1}^{n} K\left(\frac{X_i - c}{h}\right) 1\left(X_i < c\right)\left(Y_i - \alpha_- - \beta_- \cdot (X_i - c)\right)^2$$

$$\min_{\alpha_+,\beta_+} \sum_{i=1}^{n} K\left(\frac{X_i - c}{h}\right) 1\left(X_i > c\right)\left(Y_i - \alpha_+ - \beta_+ \cdot (X_i - c)\right)^2.$$

- The local linear estimator of CATE is: $\widehat{\tau} = \widehat{\alpha}_+ - \widehat{\alpha}_-$.
- Alternatively, we can solve

$$\min_{\alpha,\beta,\tau,\gamma} \sum_{i=1}^{n} K\left(\frac{X_i - c}{h}\right) \times \left(Y_i - \alpha - \beta \cdot (X_i - c) - \tau \cdot D_i - \gamma \cdot (X_i - c)\, D_i\right)^2$$

which will numerically yield the same estimate $\widehat{\tau}$.

Bandwidth selection

- Bias and variance tradeoff in choice of h.
- Choice of h for estimation or inference (confidence interval)?
- Estimation: MSE-optimal choice, see Imbens and Kalyanaraman (2012). Their method restricts bandwidths on both sides to be equal. Arai and Ichimura (2018) investigates choice of bandwidths that could be different.
- Choices of h that are optimal for estimation (minimize MSE) lead to h that is too-large for bias of estimator to be negligible, resulting in confidence intervals that are not properly centered and with empirical coverage substantially below nominal coverage.
- Inference: Calonico, Cattaneo and Titiunik (2014), estimating the bias and using bias-corrected estimator, with new standard errors accounting for such estimation error. Calonico, Cattaneo and Farrell (2018), minimizing the coverage error.

Covariates

- Often there are additional covariates (Z) in addition to the score. These covariates can be used to improve estimation precision.
- The argument is analogous to that supports inclusion of covariates when analyzing experimental data.
- We solve

$$\min_{\alpha,\beta,\tau,\gamma,\delta} \sum_{i=1}^{n} K\left(\frac{X_i - c}{h}\right) \times \left(Y_i - \alpha - \beta \cdot (X_i - c) - \tau \cdot D_i - \gamma \cdot (X_i - c)\, D_i - Z_i^\top \delta\right)^2$$

- See Calonico, Cattaneo, Farrell and Titiunik (2018) for asymptotic properties of this method, including biases, asymptotic distributions and bandwidth selection...

Extensions

- Classical identification and estimation results are valid under the implicit assumption that the score is continuous.
- One may have a discrete score in applications:
  - A continuous latent score, of which only a discretized or rounded version is recorded in the data, e.g., age, weight... see Dong (2015).
  - Inherently only take on a limited number of values, e.g., the enrollment number of a school, the number of employees of a company ... see Kolesar and Rothe (2017).
- Discrete categorical outcome variable (e.g., Lee, 2008), see Xu (2017) for local maximum likelihood method.
- Discrete duration outcomes (e.g., the duration until recovery of a disease), see Xu (2018).

Falsification tests

- The RD model imposes weak identification assumptions, i.e., continuity of $\mathrm{E}\left[Y(0) \mid X=x\right]$ and $\mathrm{E}\left[Y(1) \mid X=x\right]$ as functions of x.
- This assumption is untestable. But in applications, we often reports results from two tests:
  - continuity of the density f_X of the score (manipulation test, McCrary, 2008);
  - continuity in the covariates.
- Suppose that X is a test score and the treatment is a scholarship. If the students know the policy and have the option of retaking the test, one may do so if his/her test score is just below the threshold. This leads to a discontinuity of the density at the threshold and possible discontinuity of $\mathrm{E}\left[Y(d) \mid X=x\right]$ ($d \in \{0,1\}$), since

$$\mathrm{E}\left[Y(d) \mid X=x\right] = \int y f_{Y(d)|X}(y \mid x)\,\mathrm{d}y = \frac{\int y f_{Y(d),X}(y,x)\,\mathrm{d}y}{f_X(x)}.$$

- The potential outcomes may also be affected by the covariates Z.
- If the distribution of Z is discontinuous at the threshold, $\mathrm{E}\left[Y\left(d\right) \mid X=x\right]$ $\left(d \in \{0,1\}\right)$ may also be discontinuous at the threshold.
- In applications, a common practice is to test

$$\lim_{x \downarrow c} \mathrm{E}\left[Z \mid X=x\right] = \lim_{x \uparrow c} \mathrm{E}\left[Z \mid X=x\right].$$

- For such a test, we can simply do the standard procedure with Y replaced by Z.

Fuzzy RD

- The probability of receiving the treatment changes discontinuously at the threshold, but not necessarily from 0 to 1. This is known as limited compliance in the literature.
- Suppose that incentive is assigned if $X \geq c$. Let $I = 1\left(X \geq c\right)$ denote whether one receives the incentive.
- Potential treatments with or without incentives: $\left(D_+, D_-\right)$. The observed treatment status $D = D_+ I + D_-\left(1 - I\right)$.
- In sharp RD, $\left(D_+, D_-\right) = (1, 0)$.
- Under continuity of conditional expectations and "no defiers" assumption $\Pr\left[D_- \leq D_+ \mid X = c\right] = 1$, an even narrower causal parameter is identified in the fuzzy RD model:

$$\begin{aligned}
&\mathrm{E}\left[Y\left(1\right) - Y\left(0\right) \mid D_+ > D_-, X = c\right] \\
&\qquad = \frac{\lim_{x \downarrow c} \mathrm{E}\left[Y \mid X = x\right] - \lim_{x \uparrow c} \mathrm{E}\left[Y \mid X = x\right]}{\lim_{x \downarrow c} \mathrm{E}\left[D \mid X = x\right] - \lim_{x \uparrow c} \mathrm{E}\left[D \mid X = x\right]}.
\end{aligned}$$

- The complier group is defined to be individuals with $D_+ > D_-$.