Técnicas Econométricas para Avaliação de Impacto

Regressão Descontínua

Guilherme Issamu Hirata
Centro Internacional de Pobreza (IPC/PNUD)

Brasília, 11 de junho de 2008.

Introdução

Estamos interessados no efeito causal de D sobre Y

D não é aleatoriamente distribuído

Porém, é conhecida uma variável X que, ao menos parcialmente, determina D

A idéia da regressão descontínua é explorar essa última relação

Alguns exemplos de aplicação

- Sindicalização: a relação empregador-funcionário (votação)
- Efeito de Incumbência: gastos federais nos estados americanos (votação)
- Bolsa de estudo: escolha da universidade (teste de conhecimento)
- CCT: frequência à escola (regras de elegibilidade)

A relação entre D e X pode ocorrer de duas formas

1. Determinística: $D = 1\{X \geq c\}$, sendo que c, o chamado ponto de corte, é conhecido
   - SRD – Sharp Regression Discontinuity Design

2. Probabilístico: D é uma variável aleatória dado X.

A probabilidade condicional $f(X) \equiv E[D \mid X = x] = P(D = 1 \mid X = x)$ é descontínua em c.

- FRD – Fuzzy Regression Discontinuity Design

Semelhança entre Sharp e Fuzzy: descontinuidade da função de probabilidade do tratamento em *c*

Diferença: no Fuzzy, existem outras variáveis que determinam o tratamento

- Votação X Bolsa de estudo

Identificação

Sharp Regression Discontinuity Design

Por definição:

$$\lim_{x \to c^+} E[D \mid X = x] = 1$$

$$\lim_{x \to c^-} E[D \mid X = x] = 0$$

Estamos interessados em:

$$\lim_{x \to c^+} E[Y \mid X = x] - \lim_{x \to c^-} E[Y \mid X = x]$$

Note o problema fundamental da avaliação: não se observa Y(1) para indivíduos com $X < c$, assim como não se observa Y(0) para indivíduos com $X \geq c$.

O que temos é:

$$E[Y \mid X] = E[Y \mid D = 0, X = x] \cdot P(D = 0 \mid X = x) +$$
$$E[Y \mid D = 1, X = x] \cdot P(D = 1 \mid X = x)$$

Lembrando aulas passadas, podemos escrever:

$$Y = Y_0(1 - D) + Y_1 D$$

$$Y = Y_0 + (Y_1 - Y_0)D$$

$$Y = \alpha + \beta \cdot D$$

Assim, supondo h um valor arbitrário pequeno,

$$E[Y \mid x = c + h] - E[Y \mid x = c - h] =$$

$$\{E[\alpha \mid x = c + h] - E[\alpha \mid x = c - h]\} +$$

$$\beta \cdot \{E[D \mid x = c + h] - E[D \mid x = c - h]\}$$

Dois pressupostos:

- β é constante
- $E[\alpha \mid X = x]$é contínuo em X no ponto c.

Temos:

$$\lim_{x \to c^+} E[Y \mid X = x] - \lim_{x \to c^-} E[Y \mid X = x] =$$

$$\beta \cdot \{\lim_{x \to c^+} E[D \mid X = x] - \lim_{x \to c^-} E[D \mid X = x]\}$$

Por definição, o termo multiplicando β é igual a 1

Portanto:

$$\beta_S = \lim_{x \to c^+} E[Y \mid X = x] - \lim_{x \to c^-} E[Y \mid X = x]$$

- β_S é o efeito de tratamento do SRD

Porém, e se

$$\lim_{x \to c^+} E[D \mid X = x] \neq 1 \quad \text{e}$$

$$\lim_{x \to c^-} E[D \mid X = x] \neq 0 \quad ?$$

A seleção ao tratamento não é mais determinística

Somado a isso, assuma que

$$\lim_{x \to c^+} E[D \mid X = x] \neq \lim_{x \to c^-} E[D \mid X = x]$$

Fuzzy Regression Discontinuity Design

$$\lim_{x \to c^+} E[Y \mid X = x] - \lim_{x \to c^-} E[Y \mid X = x] =$$

$$\beta \cdot \left\{ \lim_{x \to c^+} E[D \mid X = x] - \lim_{x \to c^-} E[D \mid X = x] \right\}$$

$$\beta_F = \frac{\lim_{x \to c^+} E[Y \mid X = x] - \lim_{x \to c^-} E[Y \mid X = x]}{\lim_{x \to c^+} E[D \mid X = x] - \lim_{x \to c^-} E[D \mid X = x]}$$

Simplificadamente,

$$\beta_F = \frac{Y^+ - Y^-}{D^+ - D^-}$$

- β_F é o efeito de tratamento do FRD

Mas e se β não for constante?

Um modo de tratar esse caso é considerar que o tratamento é determinístico, porém com funções distintas para diferentes indivíduos ou grupo de indivíduos (Imbens & Angrist, 1994): $D_i(X)$

Duas suposições:

1. Existe um e pequeno e arbitrário, tal que $D_i(c+e) \geq D_i(c-e)$ para todo $e > 0$ (monotonicidade)

2. β_i e $D_i(X)$ são conjuntamente independentes de X na vizinhança de c.

Lembrando a aula de variáveis instrumentais:

- *Complier*:

$$\lim_{x \to c_i^+} D_i = 1$$
$$\lim_{x \to c_i^-} D_i = 0$$

- *Always-taker*:

$$\lim_{x \to c_i^+} D_i = 1$$
$$\lim_{x \to c_i^-} D_i = 1$$

- *Never-taker*: $\lim_{x \to c_i^+} D_i = 0$

  $\lim_{x \to c_i^-} D_i = 0$

Isto é, no FRD, temos novamente a identificação de um efeito de tratamento somente para os *compliers* quando o efeito de tratamento não é constante.

$$\beta_F = \frac{\lim_{x \to c^+} E[Y \mid X = x] - \lim_{x \to c^-} E[Y \mid X = x]}{\lim_{x \to c^+} E[D \mid X = x] - \lim_{x \to c^-} E[D \mid X = x]} = \text{LATE}$$

Estimação

Basicamente, duas formas de estimação

<u>Regressão não-paramétrica na fronteira</u>

SRD:

$$\hat{\beta}_S = \hat{Y}^+ - \hat{Y}^- = \frac{\sum_{i:X \geq c} Y_i \cdot K(u)}{\sum_{i:X \geq c} K(u)} - \frac{\sum_{i:X < c} Y_i \cdot K(u)}{\sum_{i:X < c} K(u)}$$

FRD:

$Y = \alpha + D\beta + \upsilon$, usando $W = 1\{c < x < c + h\}$ como instrumento. O parâmetro estimado é equivalente a

$$\hat{\beta}_F = \frac{\hat{Y}^+ - \hat{Y}^-}{\hat{D}^+ - \hat{D}^-} = \frac{\dfrac{\sum_{i:X \geq c} Y_i \cdot K(u)}{\sum_{i:X \geq c} K(u)} - \dfrac{\sum_{i:X < c} Y_i \cdot K(u)}{\sum_{i:X < c} K(u)}}{\dfrac{\sum_{i:X \geq c} D_i \cdot K(u)}{\sum_{i:X \geq c} K(u)} - \dfrac{\sum_{i:X < c} D_i \cdot K(u)}{\sum_{i:X < c} K(u)}}$$

Problema: estimador de kernel não é consistente, devido às suas propriedades em relação a estimações próximas à fronteira.

Regressão Linear Local

SRD:

$$\min \sum_{i \in G} (Y - \alpha - \beta(X - c))^2$$

$$G = \{(c - h < X \leq c), (c < X < c + h)\}$$

Alternativamente,

$$\min \sum_{i \in F} (Y - \alpha - \lambda(X - c) - \beta_S D - \gamma(X - c)D)^2$$

$$F = (c - h, c + h)$$

FRD:

O mesmo procedimento, mas calculando também para a variável de tratamento

Alternativamente, pode-se utilizar o estimador de mínimos quadrados em dois estágios

$$Y = \alpha + 1\{X < c\}(X - c) + 1\{X \geq c\}(X - c) + D\beta + \upsilon$$

utilizando

$Z = 1\{X \geq c\}$ como instrumento.

Bandwidth

O método da regressão descontínua depende da escolha de *h*, o chamado *bandwidth*. Quanto maior *h* maior a variação captada pelo estimador; no entanto, quanto maior *h*, menor a variância do estimador. Assim, deve haver um bandwidth ótimo.

Não há consenso sobre como escolher h.

Uma regra de bolso é utilizar $h = N^{1/5}$, onde N é o número de observações.

Para o caso SRD, Imbens & Lemiex (2007) propõem utilizar o valor de h que minimiza

$$\frac{1}{N}\sum_{J}(Y - \hat{\mu}(X))^2$$

$$J = (q_{\delta}, q_{1-\delta}), \delta < 1$$

Já para o caso FRD, Imbens & Lemiex (2007) propõem o mesmo procedimento, desta vez calculando também um h que minimiza o desvio em relação a D, e tomar o valor mínimo dentre os dois mínimos.

Inferência

$$V = \frac{V_Y}{\beta_D^2} + \frac{\beta_Y^2}{\beta_D^4} V_D - 2\frac{\beta_Y}{\beta_D^3} Cov(Y, D)$$

Duas formas de estimar: Plug-in e MQ2S

Precauções ao utilizar a regressão descontínua

- Origem da descontinuidade:
  - Teste nas covariáveis
  - Teste de continuidade
- Escolha do Bandwidth
  - Teste de diferentes proporções do bandwidth escolhido