Avaliação de Políticas Públicas II: Métodos não experimentais

Paula Pereda (USP)

November 4, 2020

Aula 5

voltar

Resumo da Aula 4

1. **Propensity Score Matching - Inferência**: Distribuições assintótica e inferência usando bootstrap.

2. **Propensity Score Matching - Global** (Optimal Matching): Minimiza distância total entre os tratados e controle. Escolha entre 1-1, ..., m-m depende da estrutura de dados. Inferência usando bootstrap.

3. **Propensity Score estratificado**: Não exclui observações. Separa amostra em estratos. Possibilidade de ajustar viés remanescente (ajuste via regressão).
4. **Apresentação**: Matching com reposição como alternativa à falta de observações. Problemas de inferência (ver Abadie e Imbens 2002 para mais detalhes).

Propensity Score Weighting

- e(X) usado como pesos amostrais das unidades (Hirano e Imbens, 2001; Hirano et al, 2003).
- Pondera participantes a depender de suas probabilidades de receber o tratamento (inverso da probabilidade de seleção)
- Vantagens:
  1. flexível para várias análises multivariadas;
  2. retem toda a amostra; e
  3. não precisa que *y* seja contínua ou distribuída normalmente.
- Importante: hipóteses do mecanismo de seleção regular, especialmente $0 < e(X) < 1$

Propensity Score Weighting

- Método para criação dos pesos: Lousa

Propensity Score Weighting

Em resumo:

- Para tratados: peso $\frac{1}{e(x_i)}$ é o que iguala a esperança do resultado

$$E(y_{1i}) = E\left(y_{1i}.w_i.\frac{1}{e(x_i)}\right)$$

- Para não tratados: peso $\frac{1}{(1-e(x_i))}$

$$E(y_{0i}) = E\left(y_{0i}.(1-w_i).\frac{1}{[1-e(x_i)]}\right)$$

Propensity Score Weighting

- Assim, o Efeito Médio do Tratamento é, portanto,

$$\tau_{ATE} = E[y_{1i}] - E[y_{0i}] = E\left[\frac{w_i y_i}{e(X_i)} - \frac{(1-w_i)y_i}{1-e(X_i)}\right]$$

e a contrapartida amostral é

$$\begin{aligned}\hat{\tau}_{WEIGHT} &= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{w_i y_i}{e(X_i)} - \frac{(1-w_i)y_i}{1-e(X_i)}\right] = \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(w_i - e(X_i))y_i}{e(X_i)(1-e(X_i))}\right] = \\ &= \frac{1}{N}\sum_{i=1}^{N}(2w_i - 1)y_i\hat{\lambda}_i,\end{aligned}$$

em que $\hat{\lambda}_i = \frac{1}{1-e(X_i)}$, se $w_i = 0$ e $\hat{\lambda}_i = \frac{1}{e(X_i)}$ caso contrário (veja que E($\sum_{i:w_i=0} \lambda_i$) $= N$ e $E(\sum_{i:w_i=1} \lambda_i) = N$).

Propensity Score Weighting

- Reponderando os pesos e estimando o $e(x_i)$, temos o **Inverse Probability Treatment Weights (IPTW)**:

$$\hat{\tau}_{IPTW} = \frac{\sum \frac{w_i y_i}{\hat{e}(X_i)}}{\sum \frac{w_i}{\hat{e}(X_i)}} - \frac{\sum \frac{(1-w_i) y_i}{1-\hat{e}(X_i)}}{\sum \frac{1-w_i}{1-\hat{e}(X_i)}}.$$

- Imbens e Wooldridge (2009) mostram que $\hat{\tau}_{IPTW}$ é
  - equivalente à média amostral de uma amostra aleatória;
  - consistente para τ_{ATE} ; e
  - assintoticamente normalmente distribuído.

Propensity Score Weighting

- PSW: simples e fácil implementação, mantém toda a amostra, não requer grandes esforços de programação.
- Procedimento:
  1. criam-se os pesos (separadamente para tratamente controle); e
  2. faz-se qualquer modelo para y especificando o peso e chamando a função de peso na análise de y.
- $\hat{\tau}_{IPTW} \equiv$ PS ponderado $\equiv$ "weighting estimator para ATE.
- Checar balanceamento nas covariadas observadas: regressão simples ponderada de X contra w (logistica, se X binário).
  - Coeficiente da regressão estat. $= 0$.
  - Se forem significantes, é preciso gerar novo $\hat{e}(X)$.

Propensity Score Weighting

- O PSW é ótimo sob 3 condições:
  1. Participantes são i.i.d.;
  2. Seleção dos participantes é exógena; e
  3. Equação de seleção foi especificada corretamente (covariadas e relação funcional).
- Se essas condições não ocorrem:
  1. Aumenta-se o erro amostral das estimativas;
  2. Viesa-se para baixo os erros-padrão, mesmo que o mecanismo de seleção tenha sido bem entendido; e
  3. Pode-se aumentar o viés.
- Em resumo, $e(X)$ deve ser muito bem especificado para usá-lo como peso.

Propensity Score Weighting

- Fórmula dos pesos

$$w(W, X) = \frac{W}{\hat{e}(X)} + \frac{1 - W}{1 - \hat{e}(X)}.$$

Assim, temos

$$w(W, X) = \begin{cases} \frac{1}{\hat{e}(X)} & , \text{ se } W = 1 \\ \frac{1}{1-\hat{e}(X)} & , \text{ se } W = 0 \end{cases}. \tag{5}$$

- Se amostra complexa: Combinar pesos (Dugoff et al, 2014)
- No Stata, opção pweight, ou svy.

Alternativa: Duplo Robusto

- O estimador pode ser modificado para incorporar covariadas. Seja então

$$y_i = \alpha + \tau w_i + Z_i\beta + \varepsilon_i. \tag{6}$$

- Estimando por WLS com pesos

$$w_i = \begin{cases} \frac{1}{e(X_i)} & , \text{ se } w_i = 1 \\ \frac{1}{1-e(X_i)} & , \text{ se } w_i = 0 \end{cases}, \tag{7}$$

mostra-se que $\hat{\tau} \xrightarrow{p} \tau$. O estimador τ é chamado de **Duplo Robusto**.

- Berk e Freedman (2008): Evitar o uso do mesmo conjunto de covariadas na regressão de y e na equação de seleção para prever $e(X)$.

PSW e PSS

- O estimador de PSS é igual ao ponderado quando os pesos são baseados no seguinte $e(x)$:

$$\tilde{e}(X_i) = \sum B_{ij} \frac{N_{1j}}{N_j}$$

- que $\tilde{e}(X_i)$ é a média dentro dos estratos em vez de $\hat{e}(X_i)$ (o que aumenta o peso de valores baixos de PS e diminui valores altos).
- Comparação:
  - Não há muita diferença entre eles: Se $J \to \infty$, ambos de aproximam
  - Se $e(X)$ é o verdadeiro, então o PSW é não-viesado. Se $e(X)$ está errado, estimador é viesado (PSS reduz esse viés).
  - PSS tende a ter variância menor.
  - Duplo robusto para PSW usa mesmas covariadas (Z), PSS permite usar covariadas diferentes por bloco j.

Alternativas: Ajuste de regressão no PSM

Métodos complementares ao propensity score matching:

- Ajuste por regressão: Desbalanceamento em algumas variáveis pós-matching.
- Abadie e Imbens (2002): Matching com reposição ($\kappa_m(i) \geq 1$ é o número de vezes que a unidade i é usada em um par).
- Abadie e Imbens (2002): Procedimento útil quando se tem variáveis contínuas no matching.
- Aplicações: Hirano e Imbens (2001).
- Procedimento em 4 etapas

Alternativas: Ajuste de regressão no PSM

1. Regressão com os dados da amostra pareada: Como $\mu_w(X) = E[(Y(W = w)|X]$, roda-se reg y X para $w = 0$ e para $w = 1$ separadamente.
2. Escolha $\hat{\beta}_{w0}$ o intercepto e $\hat{\beta}_{w1}$ a inclinação que $Min_{\{\beta\}} \sum_{i:W_i=w} \kappa_m(i) \left(y_i - \beta_{w0} - \beta_{w1} X_i\right)^2$ e $\hat{\mu}_w(X) = \hat{\beta}_{w0} + \hat{\beta}_{w1} X$.
3. Podemos usar $\hat{\mu}_w$ para corrigir viés do estimador de matching simples supondo*:

$$\tilde{y}_i(0) = \begin{cases} y_i & , \text{ se } W_i = 0 \\ \frac{1}{\#J_M(i)} \sum_{l \in J_M(i)} [y_l + \hat{\mu}_0(X_i) - \hat{\mu}_0(X_l)] & , \text{ se } W_i = 1 \end{cases}$$

$$\tilde{y}_i(1) = \begin{cases} \frac{1}{\#J_M(i)} \sum_{l \in J_M(i)} [y_l + \hat{\mu}_1(X_i) - \hat{\mu}_1(X_l)] & , \text{ se } W_i = 0 \\ y_i & , \text{ se } W_i = 1 \end{cases}.$$

Os termos $\hat{\mu}_W(X_i) - \hat{\mu}_W(X_l)$ são os ajustes pelas covariadas.

4. Por fim, $\hat{\tau} = \frac{1}{N} \sum_{i=1}^{N} [\tilde{y}_i(1) - \tilde{y}_i(0)]$.

* $N = \sum_i K_M(i)$, e $J_M(i) = \{l = 1, ..., N | W_l = 1 - W_i, ||X_l - X_i|| * v \leq d_m(i)\}$.

PSM não paramétrico

- Heckman, Ichimura e Todd (1997, 1998)
- Objetivos: (i) suavizar funções não-conhecidas; e (ii) usar mais informações de controles mais próximos
- Hipóteses do matching com base em Kernel para estimar ATT:
  1. Substitui $(Y_0, Y_1) \perp (W|X)$ por $Y_0 \perp (W|X)$, i.e., apenas os resultados do controle são independentes do tratamento condicionado a X;
  2. Em vez de assumir independência completa, impõe-se independência na média condicional:

$$E(Y_0|W = 1, X) = E(Y_1|W = 0, X);$$

  3. Separabilidade (divide variáveis em observáveis e não-observáveis) e Restrições de Exclusão (isola variáveis que determinam Y e que determinam W)

PSM não paramétrico

- Com base nas 3 hipóteses, os autores propõem abordagem para estimar ATT como

$$ATT = E[(Y_1 - Y_0)|W = 1, X]$$

- considerando $U_0 \perp (W|X)$, portanto $u_0 \perp (W|e(X))$
- Com estas considerações, têm-se que

$$E[U_0|e(X)] = 0 = E[U_1|e(X)].$$

- Uma hipótese necessária é que a distribuição de U seja a mesma para $W = 1$ e $W = 0$ (condicionado a $e(X)$).
- Os autores comparam estimadores condicionados a X e a $e(X)$, mas não há consenso quanto a qual utilizar. Usam $e(X)$ por simplicidade.

PSM não paramétrico

Estimador Kernel para Matching

- Matching 1–m: Comparar o tratado com a média ponderada dos controles.
- Sejam I_0 e I_1 os conjuntos dos j e i participantes não-tratados e tratados. O estimador de ATT é

$$\hat{ATT} = \frac{1}{N_1} \sum_{i \in I_1} \left[y_{1i} - \sum_{j \in I_0} w(i,j) y_{0j} \right],$$

- em que n_1 é o núm. tratados e $w(.)$ pondera unidades controle com base na distância dos PS.

PSM não paramétrico

Estimador Kernel para Matching

- Regressão não-paramétrica para determinar $w(i,j)$, a função de ponderação dos controles para cada tratado.
- Suavização de curvas
  - Calcular médias locais em torno de um X_0
  - Definir janela (span) dentro da qual será calculada a média local dos pontos (ex., $\frac{1}{100}N$)
  - Média local: Média ponderada dos valores de y na mesma janela, mas com pesos diferentes dados pelo estimador Kernel

PSM não paramétrico

Estimador Kernel para Matching - Método

- Seja $Z_i = \frac{(X_i - X_0)}{h}$, em que $(X_i - X_0)$ é a distância entre o valor de X para a unidade i e o ponto focal X_0 e h é um fator de escala determinado pelo Kernel.
- Bandwidth: fraçao usada para determinar o número de observações dentro do span (se o span contém 50% de N, o bandwidth é 0,5.)
- Seja a função Kernel, $G(Z_i)$, o peso para o valor previsto de $\hat{f}(X_0)$ dentro de uma bandwidth para um ponto X_0. Dado $G(Z_i)$, podemos estimar

$$\hat{f}(X_0) = \hat{y}_{|X_0} = \frac{\sum_{i=1}^{N} G(Z_i) Y_i}{\sum_{i=1}^{N} G(Z_i)}.$$

PSM não paramétrico

Estimador Kernel para Matching - Escolha de G(.)

- Há diversas funções para $G(Z)$, com $Z_i = \frac{X_i - X_0}{h}$:

$$G(Z) = \begin{cases} \text{Tricube Kernel: } G_T(Z_i) = \begin{cases} (1 - Z_i^3)^3 & , \forall Z_i < 1 \\ 0 & , \text{c.c.} \end{cases} \\ \text{Normal Kernel: } G_N(Z_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} \\ \text{Outras: } \begin{cases} \text{Rectangular Kernel: mesmo peso} \\ \text{Epanechnikov: parabólica} \end{cases} \end{cases} .$$

- h na função Normal: desvio-padrão de uma normal centrada em X_0;
- na Tricube, é o número de observações dentro de um span centrado em X_0 (pesos seguem distribuição normal dentro de cada janela).

PSM não paramétrico

Estimador Kernel para Matching - Pesos

- Supondo $e(X)$ no lugar de X, a função Kernel definirá os pesos levando em conta a distância entre $e(X)$ de $i \in I_1$ para cada $j \in I_0$.
- Para cada i, temos que:
  1. existe $w(i,j)$ para todos $j \in I_0$;
  2. $w(i,j)$ é maior para quem está 'perto' (Se $|\frac{p_j - p_i}{h}|$ é pequeno, $G(.)$ é próximo de 1); e
  3. $w(i,j)$ é menor para quem está 'longe' (Se $|\frac{p_j - p_i}{h}|$ é grande, $G(.)$ é próximo de 0).

$$w(i,j) = \frac{G\left(\frac{p_j - p_i}{h}\right)}{\sum_{k \in I_0} G\left(\frac{p_j - p_i}{h}\right)},$$

- p_i se refere ao tratado em questão (ponto focal).
- Pode-se enxergar o estimador Kernel como a regressão ponderada de y_{0j} sobre intercepto considerando os pesos $w(i,j)$.

PSM não paramétrico

Regressão Linear Local (Lowess)

- Extensão do Kernel Matching: Regressão linear local suavizada

$$\min_{\{\beta_0,\beta_1\}} [y_i - \beta_0 - \beta_1(X_i - X_0)]^2 \, G\left(\frac{X_i - X_0}{h}\right),$$

  em que $G(.)$ é a função Kernel (X ou $e(X)$).
- Interessante quando controles são distribuídos assimetricamente em torno dos tratados (ex: concentração nas fronteiras).
- Observações
  - O estimador é sensível à escolha do bandwidth.
  - Heckman et al (1998) derivam distribuiçao assintótica do ATT.
  - A falta de suporte comum aumenta a variância (usar trimming).
  - Usa-se bootstrap para erro-padrão (alertas para o uso do bootstrap em Abadie et al (2004), Abadie e Imbens (2006)).

PS para Tratamento Contínuo

- Também conhecido por **Análise de Resposta à Dose**.
- Tratamento é contínuo
- Aplicações:

$$\begin{aligned} \text{Dose Remédio} &\Rightarrow \text{Efeito} \\ \text{Estratégias} &\Rightarrow \text{Resultado} \\ \text{Número de Cigarros} &\Rightarrow \text{Gasto com Saúde} \\ \text{Tamanho da Turma} &\Rightarrow \text{Desempenho de Alunos} \end{aligned}$$

- A abordagem de contrafactual é mais complicada: Há vários resultados potenciais.
- Hipóteses adicionais para identificação se fazem necessárias.

PS para Tratamento Contínuo

1. A definição original de PS não se aplica porque há múltiplas doses; cada participante pode ter múltiplos PS $\rightarrow e(X) \neq P(W = 1|X)$.

2. É preciso redefinir a distância entre tratados e controle porque, agora, pares similares nas covariadas podem diferentes nas dosagens.

PS para Tratamento Contínuo: GPS

Estimador Generalized Propensity Score (GPS) - Hirano e Imbens (2004)

- Seja uma a.a. $\{(y_i, T_i, X_i); i \in \{1, \ldots, N\}\}$ em que X representa as covariadas, T o tratamento (dose) recebido e y o resultado final.
- Seja também $\{y_i(t)\}_{t\in\gamma}$ o conjunto de resultados potenciais do indivíduo i para todos os valores do conjunto contínuo de tratamentos potenciais γ
- A função de resposta média à dose será denotada por $\mu(t) = E(y_i(t))$.
- $\{y_i(t)\}_{t\in\gamma}$, T_i e X_i são definidos no espaço de probabilidade comum.
- T_i é distribuído continuamente por uma medida de Lebesgue em γ
- Assim, temos que $y_i = y_i(T_i)$ é uma v.a. bem-definida.

PS para Tratamento Contínuo: GPS

Estimador GPS - Hirano e Imbens (2004)

- O GPS é a densidade condicional de T dadas as covariadas. Assim,

$$GPS = R = r(T, X),$$

- em que $r(t, x) = f_{T|X}(T|X)$. Essa função tem propriedades de balanceamento similares a $e(X)$.
- Hirano e Imbens (2004) provaram que:
  - $y(t) \perp T|X, \forall t \in \gamma$ implica que $f_T\left(t|r(t,x), y(t)\right) = f_T\left(t|r(t,x)\right), \forall t \in \gamma$
  - Portanto, GPS elimina o viés associado às diferenças em X.
  - $\beta(t,x) = E(y(t)|r(t,x) = r) = E(y|T = t, R = r)$
  - $\mu(t) = \int \beta(t, r(t,x)) f_X(x) \mathrm{d}x = E(\beta(t, r(t,x)))$, i.e., resultado médio no nível t, dadas as covariadas.

PS para Tratamento Contínuo: GPS

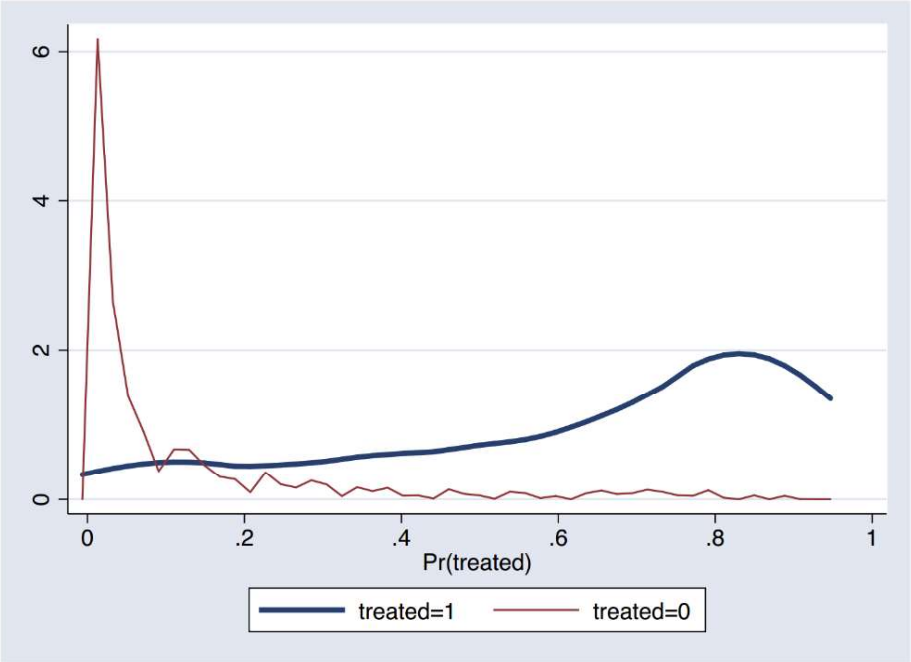
Método:

1. Modelar a distribuição condicional do tratamento dadas as covariadas: $g(T_i)|X_i \sim N(\beta_0 + X_i\beta, \sigma^2)$.
2. Estima-se o GPS baseado no modelo de regressão estimada: $\hat{R}_i = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\exp\{-\frac{1}{2\hat{\sigma}^2}[g(T_i) - X_i\hat{\beta}]^2\}$
3. Testa-se o balanceamento (procedimento em Bia e Matei, 2008): testa dentro de cada intervalo de g(.)
4. Estima-se a esperança condicional de y dado o tratamento e GPS pela aproximação quadrática: $E(y_i|T_i, R_i) = \alpha_0 + \alpha_1 T_i + \alpha_2 T_i^2 + \alpha_3 R_i + \alpha_4 R_i^2 + \alpha_5 T_i R_i$
5. Estima-se a função de resposta à dose por $\hat{E(y_i)} = \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2 + \hat{\alpha}_3 \hat{r}(t,x) + \hat{\alpha}_4 \hat{r}(t,x)^2 + \hat{\alpha}_5 t\hat{r}(t,x)$, que é o valor da função resposta à dose para o nível t de tratamento. O efeito da mudança de t para t' é $\mu(\mathbf{t}') - \mu(\mathbf{t})$.
6. Estima-se o erro-padrão por bootstrap.

Aplicações PS

- Dehejia e Wahba (1999): Efeito de treinamento sobre salários usando PSM e PSS. Resultados de acordo com a literatura.
- Dale e Krueger (2002): Efeito da qualidade da educação sobre retornos futuros
- Hirano e Imbens (2002): Estimador duplo robusto para estimar o efeito do procedimento de colocar catéter sobre sobreviviência dos pacientes
- Sianesi (2010): Efeito de treinamentos sobre salários usando o PSID americana

Efeito de treinamento sobre salários (EUA)



Efeito de treinamento sobre salários (EUA)