Introdução à Econometria Semiparamétrica

Aula 3 - Estimação Semiparamétrica

Luis A. F. Alvarez

15 de outubro de 2024

Exemplo

- Considere uma população de interesse em que definimos um tratamento individual binário, denotado por uma variável aleatória, $D \in \{0, 1\}$ e um resultado de interesse Y.
  - Os resultados potenciais, que descrevem o que acontece com um indivíduo caso ele seja alocado ao tratamento ou não, são dados por $(Y(0), Y(1))$, de modo que o resultado observado é dado por $Y = DY(1) + (1 - D)Y(0)$ e o efeito da política é $Y(1) - Y(0)$.
- Sejam X um vetor de características observáveis, tais que seja razoável supor que:

$$Y(0) \perp D|X$$

Estimação do ATT

- Sob a hipótese de identificação anterior, se supomos que $\mathbb{P}[D = 1] > 0$ e a seguinte hipótese de suporte comum (*overlap*):

$$\exists \epsilon > 0, \quad \mathbb{P}[\mathbb{P}[D = 0|X] \geq \epsilon] = 1$$

- Então é possível identificar o efeito médio do tratamento nos tratados (ATT),

$$\mathbb{E}[Y(1) - Y(0)|D = 1] = \frac{\mathbb{E}[DY]}{\mathbb{E}[D]} - \frac{1}{\mathbb{E}[D]}\mathbb{E}\left[\frac{\mathbb{P}[D = 1|X = 1]}{1 - \mathbb{P}[D = 1|X = 1]}(1 - D)Y\right]$$

- O problema é que a hipótese de suporte comum pode ser irrazoável em alguns contextos.

Combinações convexas de ATTs condicionais

- Considere, como alternativa, o estimando β^* que resolve

$$(\beta^*, g) = \text{argmin}_{b \in \mathbb{R}, h \in \mathcal{H}} \mathbb{E}[(Y - bD - h(X))^2],$$

onde $\mathcal{H}$ é um sub-espaço fechado de $L_2(\mathbb{P}_X)$.

- Nesse caso, é possível mostrar que, se $\mathbb{P}[D = 1|X] \in \mathcal{H}$ ou $\mathbb{E}[Y(0)|X] \in \mathcal{H}$, então:

$$\beta^* = \frac{\mathbb{E}[\mathbb{E}[Y(1) - Y(0)|X, D = 1]\mathbb{P}[D = 1|X](1 - m(X))]}{\mathbb{E}[\mathbb{P}[D = 1|X](1 - m(X))]}$$

onde $m(X) = \text{argmin}_{h \in \mathcal{H}} \mathbb{E}[(D - h(X))^2]$.

- Resultado é extensão direta de Angrist (1998). ▸ Detalhes
- Se $m(X) \in [0, 1]$ com probabilidade 1, estimando é uma combinação convexa de ATTs condicionais, dando mais peso para pontos do suporte com melhor sobreposição.
  - Combinação convexa mais fácil de se estimar eficientemente (Goldsmith-Pinkham, Hull e Kolesár, 2024), sob algumas hipóteses.

Estimando β^*

- Com base no resultado anterior e uma amostra aleatória da população, poderíamos tentar estimar β^* resolvendo o análogo amostral do problema.
  - Para implementar classes "complexas" $\mathcal{H}$, podemos alternar o estimador de MQO dos resíduos $(Y_i - \tilde{g}(X_i))$ em D_i e o estimador que projeta $Y_i - \tilde{\beta} D_i$ em $\mathcal{H}$ até convergência.
- Note, entretanto, que para a representação anterior valer, devemos escolher uma classe suficiente expressiva para representar ou $\mathbb{E}[Y(0)|X]$ ou $\mathbb{P}[D = 1|X]$.
  - Especificamente, se X contém variável contínua com suporte na reta que é relevante para a seleção, então classe linear não será capaz de satisfazer $m(X) \in [0, 1]$ com probabilidade 1, e não teremos combinação convexa de ATTs (pesos negativos).
  - Além disso, se há muitos possíveis controles, mas somente um subconjunto é relevante para explicar a seleção ao tratamento, deveríamos utilizar métodos de alta dimensão válidos sob esparsidade.
- Estimador resultante é dado por:

$$\hat{\beta} = \frac{\frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - \hat{g}(X_i))}{\frac{1}{n}\sum_{i=1}^{n} D_i^2}.$$

Aproximação assintótica do estimador

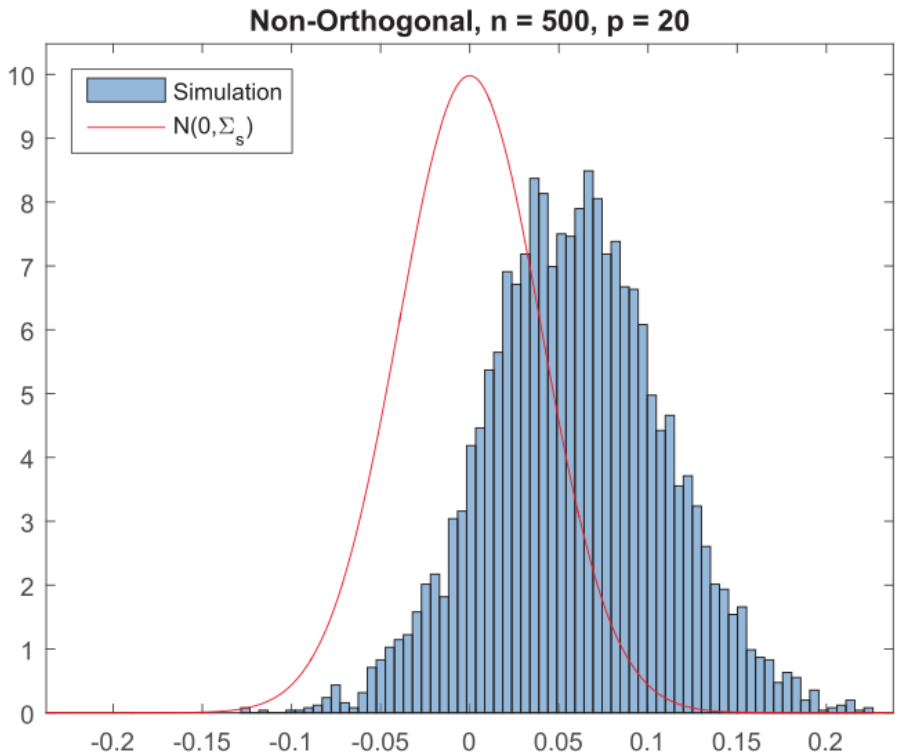
$$\sqrt{n}\left(\hat{\beta}-\beta^{*}\right)=$$

$$\underbrace{\left(\frac{1}{n}\sum_{i\in I}D_i^2\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I}D_iU_i}_{=:a}+\underbrace{\left(\frac{1}{n}\sum_{i\in I}D_i^2\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I}D_i\left(g\left(X_i\right)-\hat{g}\left(X_i\right)\right)}_{=:b}$$

onde $U_i = Y_i - \beta^* D_i - g(X_i)$.

- Termo a é bem comportado em amostras grandes $a \xrightarrow{d} N(0, \sigma^2)$.
- No entanto, termo b é dado por:
  $b = \left(E\left[D_i^2\right]\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i\in I} m\left(X_i\right)\left(g\left(X_i\right)-\hat{g}\left(X_i\right)\right) + o_P(1)$
- Esse termo não é bem comportado para estimadores não paramétricos, especialmente modernos.
  - Comparar EQM do MQO com outras alternativas que operam no trade-off viés variância.
    - No MQO, $\sqrt{n}(\hat{\delta}'x - \delta_0'x) \xrightarrow{d} N(0, x'\Sigma^2_{\hat{\delta}}x)$, de modo que esperamos que, em amostras grandes, $\infty > \mathbb{V}[b] > 0$, mas $\mathbb{E}[b] \approx 0$.
    - Para estimadores não paramétricos modernos que operam no tradeoff vié-variância $\mathbb{E}[b] = \frac{1}{n^{-1/2+\psi}}$, para $\psi < 1/2$ (ver exemplo do Lasso).
    - Distribuição arbitrariamente não centrada em β^* em amostras grandes.

Exemplo: viés na distribuição do estimador semiparamétrico "ingênuo"



Estimador alternativo

- É possível mostrar que parâmetro de interesse pode ser identificado como:

$$\beta^* = \frac{\mathbb{E}[(Y - g(X))(D - m(X))]}{\mathbb{E}[(D - m(X))D]}$$

- Considere então:

$$\check{\beta} = \left(\frac{1}{n} \sum_{i \in I} \hat{V}_i D_i \right)^{-1} \frac{1}{n} \sum_{i \in I} \hat{V}_i \left(Y_i - \hat{g}\left(X_i \right) \right) ,$$

onde $\hat{V}_i = D_i - \hat{m}(X_i)$.

Propriedades do estimador alternativo.

- Estimador escrever-se-á como:

$$\sqrt{n}(\check{\beta} - \beta^*) = a + b + c + o_{\mathbb{P}}(1)\,,$$

onde

1. $a^* = \left(E\left[V^2\right]\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} V_i U_i \xrightarrow{d} N(0, \sigma^2)$
   - Termo bem comportado, como antes
2. $b^* = \left(E\left[V^2\right]\right)^{-1} \frac{1}{\sqrt{n}} \sum_{i \in I} \left(\hat{m}\left(X_i\right) - m\left(X_i\right)\right)\left(\hat{g}\left(X_i\right) - g\left(X_i\right)\right)$
   - Agora, viés dos estimadores não paramétricos podem decair para zero a $n^{-1/4}$ e $b^* = o_{\mathbb{P}}(1)$.
3. $c^* =$
   $\left(E\left[V^2\right]\right)^{-1} \left[\frac{1}{\sqrt{n}} \sum_{i \in I} U_i \left(\hat{m}\left(X_i\right) - m\left(X_i\right)\right) + \frac{1}{\sqrt{n}} \sum_{i \in I} V_i \left(\hat{g}\left(X_i\right) - m\left(X_i\right)\right)\right]$
   - Termo captura a correlação dos estimadores com componente idiossincrático da amostra.
   - Caso haja algum sobreajuste, esse termo não desaparecerá, produzindo vieses.
   - **Solução:** estimar as funções não paramétricas em outra amostra $I^{\complement}$ tal que $I \cap I^{\complement} = \emptyset$.
   - Nesse caso, garantiremos que esse termo possui média zero (por quê?) e que $c^* = o_{\mathbb{P}}(1)$.

Debiased machine learning

- A estratégia anterior de utilização de aprendizagem estatística para a estimação de parâmetros consistiu de dois ingredientes cruciais.
  - Um ajuste na fórmula do estimador como forma de reduzir o viés devido a regularização de estimadores que operam no trade-off viés-variância para reduzir EQM.
  - *Sample-splitting* como forma de remover o viés de sobreajuste que seria gerado em modelos muito complexos.
- Esse método recebeu na literatura o nome de *debiased machine learning* (DML) (Chernozhukov, Chetverikov et al., 2018).
- No que segue, vamos estudar como encontrar a correção do viés genericamente.

Função influência

- Seja $\hat{\theta}$ um estimador de um parâmetro escalar de interesse.
  - Seja $\theta(P)$ o limite de probabilidade do estimador, quando a distribuição verdadeira dos observáveis, S, é P.
  - Vamos focar no caso *não-paramétrico*, em que P pode ser "qualquer" distribuição de probabilidade sobre S, a não ser por condições de regularidade.
- Nesse caso, definimos a função influência de $\hat{\theta}$ como o mapa $\psi(S; P)$ tal que, para "qualquer" distribuição H:

$$\frac{d}{d\tau}\theta(P_\tau)\Big|_{\tau=0} := \lim_{\tau\to 0}\frac{\theta(P+\tau(H-P))-\theta(P)}{\tau} = \int \psi(s;P)H(ds)\,,$$

onde $P_\tau = P + \tau(H - P)$.

  - Nome função influência vem do fato de que, heuristicamente, ela nos dá o efeito sobre o viés do estimador de se incluir uma observação "contaminada" na amostra.
  - Aplicações em estimação robusta (Rousseeuw et al., 1986) e na construção de estimadores eficientes em modelos semiparamétricos estritos (em que P não pode ser qualquer coisa) (Newey, 1990; Bickel et al., 1993).

Cálculo da função influência e suas propriedades

- Existe uma literatura bastante consolidada sobre como computar $\psi(S; P)$ na prática, que não entraremos por limitações de tempo.
  - Veja Ichimura e Newey (2022) e Kennedy (2023) para métodos.
- Somente notamos as seguintes propriedades de funções influência, que nos serão úteis a seguir.

$$\int \psi(s; P) P(ds) = 0$$

$$\int \frac{d}{d\tau} \psi(s; P_\tau)\Big|_{\tau=0} P(ds) = -\int \psi(s; P) H(ds)$$

Um problema de estimação semiparamétrica

- Nesta aula, vamos focar no seguinte problema de função semiparamétrica. O objetivo é estimar

$$\theta_0 := \mathbb{E}[m(S; \gamma_0)]\,,$$

onde γ_0 é uma função que desejamos aproximar não parametricamente (e.g. uma esperança condicional, quantil condicional ou densidade).

- Nós propomos estimar θ_0 partindo da seguinte condição de momento:

$$\mathbb{E}[m(S; \gamma_0) + \psi(S; P_0)] - \theta_0 = 0\,,$$

onde P_0 é a distribuição verdadeira, e $\psi(s; P_0)$ é a função influência de primeiro-estágio de se perturbar a parte não paramétrica:

$$\frac{d}{d\tau}\mathbb{E}[m(S; \gamma(P_0 + \tau(H - P_0)))]\Big|_{\tau=0} = \int \psi(s; P_0) H(ds)$$

Robustez local da condição de momento e estimador

- A condição de momento modificada é localmente robusta a "'vieses" na estimação não paramétrica da primeira etapa, no sentido que:

$$\frac{d}{d\tau}\mathbb{E}[m(S;\gamma(P_\tau))+\psi(S;P_\tau)]\Big|_{\tau=0}=0$$

- Metodologia de Chernozhukov, Chetverikov et al., 2018, particione amostra em $\mathcal{I}_1$ e $\mathcal{I}_2$, com tamanhos aproximadamente iguais.
  - Na parte $\mathcal{I}_1$, estime γ_0, $\hat{\gamma}^1$ e a função influência (que no geral depende de uma parte não paramétrica), $\hat{\psi}^1$.
  - Na partição $\mathcal{I}_2$, estimar θ como:

$$\hat{\theta}_2=\frac{1}{|\mathcal{I}_2|}\sum_{i\in\mathcal{I}_2}(m(S_i;\hat{\gamma}^1)+\hat{\psi}^1(S_i)) \tag{1}$$

- Para não perder observações, podemos trocar o papel de $\mathcal{I}_1$ e $\mathcal{I}_2$, calcular $\hat{\theta}_1$ analogamente e fazer o estimador *cross-fitted*:

$$\hat{\theta}=\frac{1}{2}\hat{\theta}_1+\frac{1}{2}\hat{\theta}_2$$

Inferência e extensões

- Resultado principal de Chernozhukov, Chetverikov et al., 2018: $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \sigma^2)$, onde $\sigma^2 = \mathbb{V}[m(S; \gamma_0) + \psi(S; P_0)]$.
  - Quantidade σ^2 pode ser estimada consistentemente usando a variância amostral do estimador $\hat{\theta}$.
- Inferência tornou-se imediata agora!
- Literatura enorme com extensões desse caso simples: GMM (Chernozhukov, Chetverikov et al., 2018; Chernozhukov, Escanciano et al., 2022), *expected shortfall* (Chetverikov, Liu e Tsyvinski, 2022), entre outros.

Derivação da representação do estimando

- Para uma v.a. arbitrária S que "mora" na mesma fonte de incerteza de $(Y(0), Y(1), D, X)$, seja $P_{\mathcal{H}}(S) = h^*(X)$, ond $h^* \in \text{argmin}_{h \in \mathcal{H}} \mathbb{E}[(Y - h(X))^2]$.
  - Como $\mathcal{H}$ é subespaço linear fechado do espaço de variáveis aleatórias com variância finita (fechado com respeito à norma $\|h(S)\|_2^2 = \int h(S)^2 \mathbb{P}(ds)$, onde $S = (Y(0), Y(1), D, X)$), $P_{\mathcal{H}}$ é um operador linear bem-definido, com $\mathbb{E}[(S - P_{\mathcal{H}}(S))h(X)] = 0$ para todo $h \in \mathcal{H}$ (ver Seções 3.2 e 3.3 de Kreyszig, 1991, para detalhes).
- Nesse caso, $g(X) = P_{\mathcal{H}}(Y - \beta^* D) = P_{\mathcal{H}}(Y(0)) + P_{\mathcal{H}}(D(Y(1) - Y(0))) - \beta^* P_{\mathcal{H}}(D)$
- Combinando a expressão acima à CPO que β^* deve satisfazer, temos:

$$\beta^* = \frac{\mathbb{E}[[D(Y(1) - Y(0)) + [Y_0 - P_{\mathcal{H}}(Y(0))] - P_{\mathcal{H}}(D(Y(1) - Y(0)))](D - P_{\mathcal{H}}(D))]}{\mathbb{E}[D(D - P_{\mathcal{H}}(D))]}$$

- Note que $\mathbb{E}[(D - P_{\mathcal{H}}(D))P_{\mathcal{H}}(D(Y(1) - Y(0)))] = 0$ pela definição do operador.
- Ademais, note que:

$$\mathbb{E}[[Y(0) - P_{\mathcal{H}}(Y(0))][D - P_{\mathcal{H}}(D)]] = 0$$

  se $\mathbb{E}[Y(0)|X] \in \mathcal{H}$ ou $\mathbb{P}[D = 1|X] \in \mathcal{H}$, pela lei das expectativas iteradas e a hipótese de identificação $Y(0) \perp D|X$.
- Representação segue então da lei das expectativas iteradas. ◂ Voltar

Bibliografia I

- Angrist, Joshua D. (1998). "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants". Em: *Econometrica* 66.2, pp. 249–288. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/2998558 (acesso em 10/10/2024).
- Bickel, Peter J et al. (1993). *Efficient and adaptive estimation for semiparametric models*. Vol. 4. Springer.
- Chernozhukov, Victor, Denis Chetverikov et al. (jan. de 2018). "Double/debiased machine learning for treatment and structural parameters". Em: *The Econometrics Journal* 21.1, pp. C1–C68. ISSN: 1368-4221. DOI: 10.1111/ectj.12097. eprint: https://academic.oup.com/ectj/article-pdf/21/1/C1/27684918/ectj00c1.pdf. URL: https://doi.org/10.1111/ectj.12097.

Bibliografia II


Chernozhukov, Victor, Juan Carlos Escanciano et al. (2022). "Locally Robust Semiparametric Estimation". Em: *Econometrica* 90.4, pp. 1501–1535. DOI: https://doi.org/10.3982/ECTA16294. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA16294. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16294.


Chetverikov, Denis, Yukun Liu e Aleh Tsyvinski (2022). *Weighted-average quantile regression*. arXiv: 2203.03032 [econ.EM].


Goldsmith-Pinkham, Paul, Peter Hull e Michal Kolesár (2024). *Contamination Bias in Linear Regressions*. arXiv: 2106.05024 [econ.EM]. URL: https://arxiv.org/abs/2106.05024.

Bibliografia III

- Ichimura, Hidehiko e Whitney K. Newey (2022). "The influence function of semiparametric estimators". Em: *Quantitative Economics* 13.1, pp. 29–61. DOI: https://doi.org/10.3982/QE826. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/QE826. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/QE826.
- Kennedy, Edward H. (2023). *Semiparametric doubly robust targeted double machine learning: a review*. arXiv: 2203.06469 [stat.ME].
- Kreyszig, Erwin (1991). *Introductory functional analysis with applications*. Vol. 17. John Wiley & Sons.
- Newey, Whitney K (1990). "Semiparametric efficiency bounds". Em: *Journal of applied econometrics* 5.2, pp. 99–135.
- Rousseeuw, Peter J et al. (1986). *Robust statistics: the approach based on influence functions*.