Double/debiased machine learning for treatment and structural parameters

Adriel Melo, Pedro Galera e Wesley Satelis

27 de junho de 2024

Sumário

Motivação

Resultados para, obteção de uma estimação raiz-N consistente, em que N é o tamanho da amostra, inferências a respeito de um parâmetro de interesse de baixa dimensão, θ_0, na presença de um parâmetro de alta dimensão, η_0.

- Tipicamente o parâmetro de interesse (θ_0) será um parâmetro de efeito causal ou de efeito de tratamento;
- Vamos considerar casos em que o parâmetro de perturbação, η_0, será estimado por métodos de machine learning (ML), como *Random Forest*, *Lasso*, *Neural Nets*, árvores de regressão, entre outros.
- Estes métodos são capazes de lidar com muitas covariáveis e fornecem estimadores naturais para parâmetos de perturbação (em nosso caso η_0) quando estes são altamente complexos.

Motivação

Considere o modelo de regressão linear parcial (PLR),

$$
\begin{aligned}
Y &= D\theta_0 + g_0(X) + U, \quad \mathrm{E}[U \mid X, D] = 0 \\
D &= m_0(X) + V, \quad \mathrm{E}[V \mid X] = 0,
\end{aligned}
$$

Em que Y é a variável de *outcome*, D é a variável de tratamento de interesse, o vetor

$$X = (X_1, ..., X_p)$$

consiste de outros controles, e U e V são perturbações.

- A primeira equação é a principal e θ_0 é o parâmetro de interesse. Se D é exógena condicional aos controles X, θ_0 tem a interpretação de efeito de tratameto;
- A segunda equação acompanha a dependência da variável de tratamento nos controles e é importante para remover vieses de regularização.

Motivação

$$Y = D\theta_0 + g_0(X) + U, \quad \mathrm{E}[U \mid X, D] = 0$$
$$D = m_0(X) + V, \quad \mathrm{E}[V \mid X] = 0,$$

- Os fatores de confusão afetam a variável de tratamento D pela função $m_0(X)$ e a variável de $outcome$ por $g_0(X)$;
- Em muitos casos a dimensão p do vetor X é grande em relação a N;
- Como p não é bem pequeno em relação a N, modelamos p como crescente com o tamanho da amostra, o que causa as suposições tradicionais que limitam a complexidade do espaço paramétrico de $\eta_0 = (m_0, g_0)$ a falharem.

Viés de regularização

Uma abordagem ingênua para estimar θ_0 usando ML seria, por exemplo, construir um estimador sofisticado $D\theta_0 + g_0(X)$. Suponha que partimos a amostra aleatoriamente em duas partes: a principal de tamanho n e a auxiliar $N - n$. Suponha que $\hat{g}_0$ é obtido usando a amostra auxiliar e que, dado $\hat{g}_0$, o estimador final de θ_0 é obtido com a amostra principal:

$$\widehat{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{n}\sum_{i\in I} D_i \left(Y_i - \widehat{g}_0\left(X_i\right)\right)$$

Este estimador terá taxa de convergência menor que $1/\sqrt{n}$. Isto é,

$$\left|\sqrt{n}\left(\hat{\theta}_0 - \theta_0\right)\right| \rightarrow_P \infty.$$

Viés de regularização

Como mostrado a seguir, o motivo para essa taxa de convergência é o viés em "aprender" g_0.
Para ilustrar o impacto do viés em g_0, podemos decompor o erro de estimação de θ_0 como

$$\sqrt{n}\left(\hat{\theta}_0 - \theta_0\right) = \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i U_i}_{:=a} + \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i\left(g_0\left(X_i\right) - \hat{g}_0\left(X_i\right)\right)}_{:=b}$$

a é bem comportado sob leves condições, $a \rightsquigarrow N(0, \bar{\Sigma})$ para algum $\bar{\Sigma}$, b é o viés de regularização, que não é centrado e diverge no geral.
O viés estimação de b é introduzido pelo método de regularização usado para estimar $\hat{g}_0$ (lasso, ridge, boosting, neural nets...), que controla a variância do estimador, mas em troca induz algum viés.

Superando viés de regularização com ortogonalização

Agora considere uma segunda construção que usa uma ortogonalização obtida parcializando o efeito de X de D, $V = D - m_0(X)$.

Especificamente, obtemos $\hat{V} = D - \hat{m}_0(X)$, em que $\hat{m}_0$ é um estimador ML obtido usando a amostra auxiliar. Estamos agora estamos resolvendo um problema de predição auxiliar para estimar a média condicinal de D dado X, então estamos fazendo *double prediction* ou *double machine learning*.

Depois de parcializar o efeito de X em D e obter um estimador preliminar de g_0 da amostra auxiliar, podemos formular o seguinte estimador não viesado (DML) para θ_0 usando a amostra principal,

$$\check{\theta}_0 = \left(\frac{1}{n} \sum_{i \in I} \widehat{V}_i D_i \right)^{-1} \frac{1}{n} \sum_{i \in I} \widehat{V}_i \left(Y_i - \widehat{g}_0 \left(X_i \right) \right).$$

Ortogonalizando D aproximadamente em respeito a X e aproximadamente removendo o efeito de confusão direto subtraindo uma estimativa de g_0, $\check{\theta}_0$ remove o efeito de viés de regularização.

Superando viés de regularização com ortogonalização

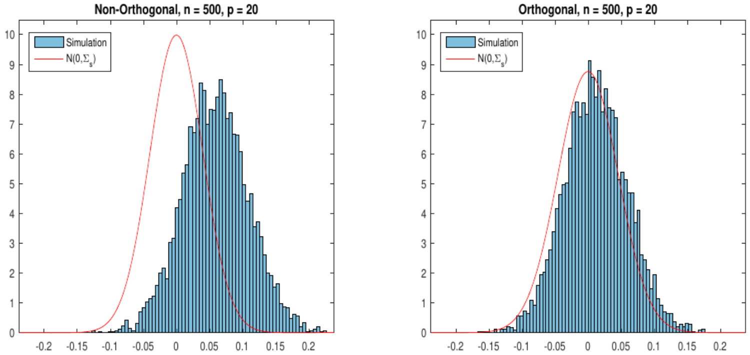

Figura: Esquerda: Comportamento de um estimador ML convencional (não ortogonal) $\hat{\theta}_0$. Direita: Comportamento de um estimador DML $\check{\theta}_0$. A distribuição simulada em azul, mostra que o estimador é não viesado, centrado em θ_0 e é aproximado pela distribuição normal. Fonte: [Chernozhukov et al., 2018]

Superando viés de regularização com ortogonalização

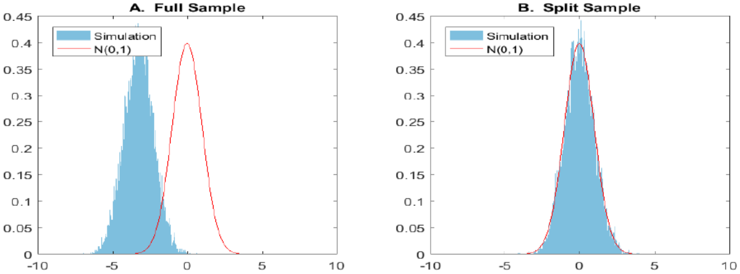

Figura: Mostra como o estimador viesado resultante de overfitting na estimação das perturbações pode causar viés no estimador $\check{\theta}_0$ e como cross-fitting elimina completamente o problema. Fonte: [Chernozhukov et al., 2018]

Sumário

Definições

Sejam

- W um elemento aletatório que toma valores em $\mathcal{W}$ no espaço de probabilidade $(\Omega_{\mathcal{W}}, \mathcal{A}_{\mathcal{W}}, P)$.
- θ_0 o valor verdadeiro do parâmetro de baixa dimensão de interesse $\theta \in \Theta$.
- η_0 o valor verdadeiro do parâmetro de perturbações $\eta \in T$, onde T é um subconjunto convexo de algum espaço vetorial normado $\|\cdot\|_T$.
- $\psi = (\psi_1, \ldots, \psi_{d_\theta})^\intercal$ um vetor de funções *score* conhecidas, tal que $\psi_j : \mathcal{W} \times \Theta \times T \to \mathbb{R}$.

Além disso, considere uma amostra aleatória $(W_i)_{i=1}^N$ com distribuição de W.

Condições de Ortogonalidade

Definição: Seja $\tilde{T} = \{\eta - \eta_0 : \eta \in T\}$ o conjunto formado pelas diferenças entre η e η_0. Considere as derivadas *pathwise* de $\mathsf{D}_r : \tilde{T} \to \mathbb{R}^{d_\theta}$

$$\mathsf{D}_r\left[\eta - \eta_0\right] := \partial_r\left\{\mathsf{E}_P\left[\psi\left(W; \theta_0, \eta_0 + r\left(\eta - \eta_0\right)\right)\right]\right\}, \; \eta \in T \tag{1}$$

para todo $r \in [0, 1)$.

- Essencialmente, trata-se da avaliação da variação de $[\eta - \eta_0]$ para um valor de η conforme um desvio r varia.

A partir disso, seja $\mathcal{T}_N \subset T$ o subconjunto formado pelos valores que os estimadores propostos para η_0 assumem com maior probabilidade, isto é, os valores mais "frequentes"de η próximos de η_0.

Condições de Ortogonalidade

Definição (Ortogonalidade de Neyman): O *score* ψ satisfaz a condição de ortogonalidade sob (θ_0, η_0) em relação ao subconjunto de controle observado $\mathcal{T}_N \subset T$ se

- θ_0 satisfaz a condição de momento

$$\mathsf{E}_P\left[\psi\left(W; \theta_0, \eta_0\right)\right] = 0 \tag{2}$$

- As derivadas *pathwise* $\mathsf{D}_r\left[\eta - \eta_0\right]$ existirem para todo $r \in [0, 1)$ e $\eta \in \mathcal{T}_N$;
- Para $r = 0$, a condição de ortogonalidade [Neyman, 1959] é satisfeita

$$\mathsf{D}_0\left[\eta - \eta_0\right] := \partial_\eta \mathsf{E}_P \psi\left(W; \theta_0, \eta_0\right)\left[\eta - \eta_0\right] = 0,\ \forall \eta \in \mathcal{T}_N \tag{3}$$

Logo, quando o desvio $r = 0$, a diferença $\left[\eta - \eta_0\right]$ para valores de η próximos a η_0 não varia.

- **Observação**: Para certos casos, é possível utilizar a condição de **Quase-Ortogonalidade**, que relaxa a suposição para casos em que a variação seja pelo menos menor que uma sequência $\left\{\lambda_N\right\}_{N \geqslant 1}$ de constantes positivas, com um decaimento de ordem $N^{-1/2}$.

Scores de Neyman

Resumindo, é necessário encontrar funções $score$ que satisfaçam:

$$\mathsf{E}_P\left[\psi\left(W;\theta_0,\eta_0\right)\right]=0 \quad e \quad \partial_\eta \mathsf{E}_P\psi\left(W;\theta_0,\eta_0\right)\left[\eta-\eta_0\right]=0$$

- Existem métodos que ortogonalizam funções de interesse, como a log-verossimilhança, em *scores* ortogonais de Neyman [Chernozhukov et al., 2015].
- Para a aplicação, focaremos em *scores* para abordar os modelos de regressão parcialmente linear e interativo.

Método DML

Definição de DML e suas propriedades básicas: assumindo uma amostra $(W)_{i=1}^{N}$ independentes e identicamente distribuídas (i.i.d.) e que N é divisível por K. O verdadeiro valor de η_0 do parâmetro η pode ser estimado por $\hat{\eta}_0$ usando uma parte dos dados $(W)_{i=1}^{N}$. O seguinte algoritmo define o DML cross-fitted simples:

DML1

- Partição dos dados em K subconjuntos $(I_k)_{k=1}^{K}$ tal que o tamanho seja $n = N/K$. Também, para cada $k \in [K] = 1, ..., K$, defina $I_k^c := 1, ..., N/I_k$.
- Para cada $k \in [K]$ construa um estimador de ML $\hat{\eta}_{0,k} = \hat{\eta}_0((W_i)_{i \in I_k^c})$ de η_0.
- Para cada $k \in [K]$ construa o estimador $\breve{\theta}_{0,k}$ como solução para a seguinte equação:

$$E_{n,k}[\psi(\mathbf{W}; \breve{\theta}_{0,k}, \hat{\eta}_{0,k})] = 0$$

  onde ψ é o *score* ortogonal de Neyman, e $E_{n,k}$ é a esperança empírica sobre o k-ésimo *fold* dos dados; isto é, $E_{n,k}[\psi(\mathbf{W})] = n^{-1} \sum_{i \in I_k} \psi(\mathbf{W})$. Se a realização de zero exato não for possível defina o estimador como uma solução aproximada.

Método DML

DML1

- Agregue os estimadores:
$$\tilde{\theta}_0 = \frac{1}{K} \sum_{k=1}^{K} \breve{\theta}_{0,k}.$$

Essa abordagem generaliza o método *cross-fitting* 50-50.

Agora definimos uma variação dessa abordagem básica de cross-fitting que pode se comportar melhor em pequenas amostras.

DML2

- Partição dos dados em K subconjuntos $(I_k)_{k=1}^{K}$ tal que o tamanho seja $n = N/K$. Também, para cada $k \in [K] = 1, ..., K$, defina $I_k^c := 1, ..., N/I_k$.
- Para cada $k \in [K]$ construa um estimador de ML $\hat{\eta}_{0,k} = \hat{\eta}_0((W_i)_{i \in I_k^c})$ de η_0.

Método DML

- Construa o estimador $\tilde{\theta}_0$ como a solução para
$$\frac{1}{K}\sum_{k=1}^{K} E_{n,k}[\psi(W;\tilde{\theta}_0,\hat{\eta}_{0,k})] = 0,$$
onde ψ é o *score ortogonal* de Neyman, e $E_{n,k}$ é a esperança empírica sobre o k-ésimo *fold* dos dados; isto é, $E_{n,k}[\psi(W)] = \frac{1}{n}\sum_{i\in I_k}\psi(W_i)$. Se a realização de zero exato não for possível, defina o estimador $\tilde{\theta}_0$ de θ_0 como uma solução aproximada.

Recomendações

- Valores moderados de K (4 ou 5) são geralmente melhores.
- DML2 é geralmente preferido sobre DML1 por sua maior estabilidade.

Suposições

Suposição 1: Scores Lineares com Ortogonalidade Aproximada de Neyman.
Suposição 2: Score regulatório e qualidade das estimativas de parâmetros de perturbação.

Resumo das Suposições 1 e 2

- Ortogonalidade Neyman (ou quase-ortogonalidade).
- Condição de identificação canônica.
- Estimativa de parâmetros perturbação numa vizinhança encolhendo de η_0.

Propriedades do DML

- **Hipóteses**: Suponha que as Suposições 1 e 2 sejam válidas e que $\delta_N \geq N^{-1/2}$ para todo $N \geq 1$.
- **Estimadores DML1 e DML2**:
  - Concentrados em uma vizinhança de $1/\sqrt{N}$ de θ_0.
  - Aproximadamente lineares e gaussianos centrados.
  - Fórmula:

$$\sqrt{N}\sigma^{-1}(\tilde{\theta}_0 - \theta_0) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \bar{\psi}(W_i) + O_P(\rho_N) \sim N(0, Id).$$

Estimador de Variância para DML

- **Hipóteses**: Suponha que as Suposições 1 e 2 sejam válidas e que $\delta_N \geq N^{-[(1-2/q)\wedge 1/2]}$ para todo $N \geq 1$.
- **Estimador da Matriz de Variância Assintótica**:

$$\hat{\sigma}^2 = \hat{J}_0^{-1} \frac{1}{K} \sum_{k=1}^{K} E_{n,k}[\psi(W; \tilde{\theta}_0, \hat{\eta}_{0,k}) \psi(W; \tilde{\theta}_0, \hat{\eta}_{0,k})'] (\hat{J}_0^{-1})'.$$

$$\hat{J}_0 = \frac{1}{K} \sum_{k=1}^{K} E_{n,k}[\psi_a(W; \hat{\eta}_{0,k})].$$

  onde ψ_a é parte da função de $\psi(W; \theta, \eta) = \psi_a(W; \eta)\theta + \psi_b(W; \eta)$ associada com θ.
- **Estimadores**:
  - $\tilde{\theta}_0$ é o estimador DML1 ou DML2.

Estimador de Variância para DML

- **Concentração**:
  - O estimador concentra-se em torno da verdadeira matriz de variância σ^2.

    $$\hat{\sigma}^2 = \sigma^2 + O_P(\varrho_N), \quad \varrho_N := N^{-[(1-2/q)\wedge 1/2]} + r_N + r'_N \lesssim \delta_N.$$

  - $\hat{\sigma}^2$ pode substituir σ^2 na fórmula da variância aproximada com o termo de resto atualizado:

    $$\rho_N = N^{-[(1-2/q)\wedge 1/2]} + r_N + r'_N + N^{1/2}\lambda_N + N^{1/2}\lambda'_N.$$

- **Construção de Regiões de Confiança**:
  - Utilizando as propriedades do DML e estimador de variância para DML podem ser usados para a construção do intervalo de confiança para o parâmetro de interesse θ_0.

Ajustes para Amostras Finitas para Incorporar Incerteza Induzida pela Divisão da Amostra

- A técnica de estimação depende de subamostras obtidas pela partição aleatória da amostra: uma amostra auxiliar para estimar as funções de perturbações e uma amostra principal para estimar o parâmetro de interesse.
- A divisão específica da amostra não impacta os resultados de estimação assintoticamente, mas pode ser importante em amostras finitas.
- É proposto repetir o estimador DML S vezes, obtendo as estimativas θ_s para $s = 1, ..., S$.
- As características dessas estimativas podem fornecer informações sobre a sensibilidade dos resultados à divisão da amostra.
- Recomenda-se relatar $\tilde{\theta}_{\text{median}}$ e $\hat{\sigma}^2_{\text{median}}$, pois são mais robustos a valores atípicos.

Incorporando o Impacto da Divisão da Amostra Usando Métodos de Média e Mediana

- Para estimativa pontual, definimos:

$$\tilde{\theta}_{\mathsf{mean}} = \frac{1}{S}\sum_{s=1}^{S}\tilde{\theta}_s$$

  ou

$$\tilde{\theta}_{\mathsf{median}} = \mathsf{median}\{\tilde{\theta}_s\}_{s=1}^{S},$$

- Para quantificar e incorporar a variação introduzida pela divisão da amostra, consideramos estimadores de variância:

$$\hat{\sigma}^2_{\mathsf{mean}} = \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\sigma}^2_s + (\hat{\theta}_s - \tilde{\theta}_{\mathsf{mean}})(\hat{\theta}_s - \tilde{\theta}_{\mathsf{mean}})^\top\right),$$

  e uma versão mais robusta,

$$\hat{\sigma}^2_{\mathsf{median}} = \mathsf{median}\left\{\hat{\sigma}^2_s + (\hat{\theta}_s - \tilde{\theta}_{\mathsf{median}})(\hat{\theta}_s - \tilde{\theta}_{\mathsf{median}})^\top\right\}_{s=1}^{S},$$

  onde a mediana seleciona a matriz com a norma do operador mediano, preservando a positividade definida.

Inferência em Modelos de Regressão Parcialmente Linear

- Seja o modelo de regressão parcialmente linear (PLR):

$$Y = D\theta_0 + g_0(X) + U, \quad \mathbb{E}[U|X, D] = 0,$$

$$D = m_0(X) + V, \quad \mathbb{E}[V|X] = 0.$$

- O parâmetro de interesse é o coeficiente de regressão θ_0. Se D é condicionalmente exógeno (tão bom quanto atribuído aleatoriamente condicionalmente aos covariáveis), então θ_0 mede o efeito causal médio/tratamento de D sobre os resultados potenciais.

Inferência em Modelos de Regressão Parcialmente Linear

- A primeira abordagem para inferência sobre θ_0 é empregar o método DML utilizando a função escore:

$$\psi(W;\theta,\eta) := \{Y - D\theta - g(X)\}(D - m(X)), \quad \eta = (g, m).$$

- É fácil ver que θ_0 satisfaz a condição de momento
  - $\mathbb{E}_P[\psi(W;\theta_0,\eta_0)] = 0$

  e também a condição de ortogonalidade
  - $\partial\eta\mathbb{E}_P[\psi(W;\theta_0,\eta_0)][\eta - \eta_0] = 0$

  onde $\eta_0 = (g_0, m_0)$.

Inferência em Modelos de Regressão Parcialmente Linear

A segunda abordagem emprega a função *score* no estilo de Robinson *partialling-out*

$$\psi(W;\theta,\eta) := \{Y - \bar{\mu}(X) - \theta(D - m(X))\}(D - m(X)), \quad \eta = (\bar{\mu}, m).$$

onde $\bar{\mu}_0(X) = \mathbb{E}[Y|X]$.

- É fácil ver que θ_0 satisfaz a condição de momento
  - $\mathbb{E}_P[\psi(W;\theta_0,\eta_0)] = 0$

  e também a condição de ortogonalidade
  - $\partial\eta\mathbb{E}_P[\psi(W;\theta_0,\eta_0)][\eta - \eta_0] = 0$

  onde $\eta_0 = (g_0, m_0)$.

Inferência sobre efeitos de tratamento no Modelo Interativo

Considere a estimativa de efeito médio do tratamento (ATE) quando esses forem totalmente heterogêneos e a variável de tratamento é binária, $D \in 0, 1$. Consideramos vetores (Y, D, X) tais que

$$Y = g_0(D, X) + U \quad \mathbb{E}_P[U|X, D] = 0,$$

$$D = m_0(X) + V, \quad \mathbb{E}_P[V|X] = 0.$$

Os dois parâmetros de interesse mais comuns neste modelo são o ATE e o ATTE:

$$\text{(ATE): } \theta_0 = \mathbb{E}_P[g_0(1, X) - g_0(0, X)],$$

$$\text{(ATTE): } \theta_0 = \mathbb{E}_P[g_0(1, X) - g_0(0, X)|D = 1].$$

Os fatores de confusão X afetam a variável de tratamento via o *propensity score* $m_0(X)$ e a variável de resultado via a função $g_0(D, X)$.

Inferência sobre efeitos de tratamento no Modelo Interativo

O Efeito Médio do Tratamento (ATE) refere-se à diferença média entre os resultados de um grupo que recebeu o tratamento e um grupo que não recebeu o tratamento.

- Estimativa de ATE:

$$\begin{aligned}\psi(W;\theta,\eta) :=& (g(1,X)-g(0,X)) + \frac{D(Y-g(1,X))}{m(X)} \\ &- \frac{(1-D)(Y-g(0,X))}{1-m(X)} - \theta,\end{aligned}$$

  onde o parâmetro de perturbações $\eta = (g, m)$ consiste de funções integráveis ao quadrado que mapeiam o suporte de (D, X) para $\mathbb{R}$ e o suporte de X para $(\epsilon, 1-\epsilon)$, respectivamente, para algum $\epsilon \in (0, 1/2)$.

Sumário

Exemplo do Bônus de Reemprego da Pensilvânia

Nos anos 80, o Departamento de Trabalho dos Estados Unidos realizou um experimento para testar os efeitos de benefícios alternativos para o "seguro-desemprego"no estado da Pensilvânia.

- Benefiários de seguro-desemprego foram aleatoriamente selecionados para um grupo controle ou para um dos cinco grupos de tratamento.
  - **Grupo Controle:** Foram aplicadas as regras e benefícios padrão do seguro-desemprego.
  - **Grupo Tratamento**: Foram oferecidos diferentes níveis de bônus em dinheiro para cada um dos cinco grupos, caso os indivíduos encontrassem um emprego e se mantivessem nele por um determinado período.

Nesse exemplo, foram considerados apenas os grupos controle e o nível de tratamento que ofereceu o maior bônus no maior período de tempo.

Exemplo do Bônus de Reemprego da Pensilvânia

Os dados coletados são:

- **Resposta** Y: O log do tempo que o benefiário ficou desempregado;
- **Tratamento** D: Indicadora se o benefiário recebeu o bônus;
- **Covariáveis** X: Faixa etária, gênero, raça, número de dependentes, trimestre do experimento, localização, expectativa de reemprego e tipo de ocupação.

Para estimar as funções referentes às covariáveis de controle, foram utilizados os seguintes métodos ML:

- *Random Forest*
- *Regression Tree*
- *Boosting*
- *Lasso*
- *Neural Net*

Exemplo do Bônus de Reemprego da Pensilvânia

Além disso, também foram aplicados dois métodos híbridos:

- *Ensemble*: Estimou as funções de perturbação fazendo a combinação ótima das médias ponderadas das estimativas do *Random Forest*, *Boosting*, *Lasso* e *Neural Net*.
- *Best*: Utiliza combinação dos métodos que tiveram a melhor performance de predição, incluindo *Ensemble*, para estimar diferentes funções de perturbação com os respectivos melhores métodos.

Exemplo do Bônus de Reemprego da Pensilvânia

Para obter as estimativas do efeito médio do bônus (ATE) na duração do desemprego, foi utilizado o algoritmo DML2 com o método da mediana.

- Modelos de Regressão Interativa (A) e Parcialmente Linear (B).
- Divisão da amostra em 2-*fold cross-fitting* e 5-*fold cross-fitting* ($K = 2, 5$).
- 100 repetições do algoritmo DML2 ($S = 100$).

Para cada método ML, foram reportados a estimativa pontual da mediana ($\tilde{\theta}_{\text{median}}$), o erro padrão calculado das 100 repetições e o ajustado para a variabilidade da divisão da amostra utilizando o método da mediana ($\hat{\sigma}^2_{\text{median}}$).

Exemplo do Bônus de Reemprego da Pensilvânia

	Lasso	Reg. Tree	Forest	Boosting	Neural Net.	Ensemble	Best
A. Interactive Regression Model							
ATE (2 fold)	-0.081	-0.084	-0.074	-0.079	-0.073	-0.079	-0.078
	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]	[0.036]
	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
ATE (5 fold)	-0.081	-0.085	-0.074	-0.077	-0.073	-0.078	-0.077
	[0.036]	[0.036]	[0.036]	[0.035]	[0.036]	[0.036]	[0.036]
	(0.036)	(0.037)	(0.036)	(0.036)	(0.036)	(0.036)	(0.036)
B. Partially Linear Regression Model							
ATE (2 fold)	-0.080	-0.084	-0.077	-0.076	-0.074	-0.075	-0.075
	[0.036]	[0.036]	[0.035]	[0.035]	[0.035]	[0.035]	[0.035]
	(0.036)	(0.036)	(0.037)	(0.036)	(0.036)	(0.036)	(0.036)
ATE (5 fold)	-0.080	-0.084	-0.077	-0.074	-0.073	-0.075	-0.074
	[0.036]	[0.036]	[0.035]	[0.035]	[0.035]	[0.035]	[0.035]
	(0.036)	(0.037)	(0.036)	(0.035)	(0.036)	(0.035)	(0.035)

Figura: Efeito estimado de bônus em dinheiro no tempo de desemprego.
Fonte: [Chernozhukov et al., 2018]

Sumário

Comentários

- A escolha do método ML para estimar as covariáveis de perturbação não afeta de forma significante a conclusão.
- Embora o erro padrão ajustado pelo método da mediana incorpore a variabilidade da divisão, não há diferenças significativas em relação ao obtido pelas repetições.

Sumário

Referências

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018).
Double/debiased machine learning for treatment and structural pa- rameters.
Econometrics Journal, 21:C1-C68.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015).
Post-selection and postregularization inference in linear models with very many controls and instruments.
Americal Economic Review: Papers and Proceedings, 105:486-490.

Neyman, J. (1959).
Optimal asymptotic tests of composite statistical hypotheses.
Probability and Statistics, 105:416-444.