On the Asymptotic Properties of Debiased Machine Learning Estimators

Amilcar Velez
Cornell University

XLIII Encuentro de Economistas del BCRP

Octubre 22, 2025

Introducción

- Parámetro de interés $\boldsymbol{\theta_0}$ satisface

$$E\big[m\big(W, \theta_0, \eta_0(X)\big)\big] = 0$$

  - W, X: vectores aleatorios (datos)
  - $\boldsymbol{\eta_0}(\cdot)$: función molestosa de características

- Estudio estimadores basados en DML (debiased machine learning)
  - DML: metodología para estimar $\boldsymbol{\theta_0}$ cuando $\boldsymbol{\eta_0}(\cdot)$ se estima usando *machine learning*

Motivacion

- DML1 y DML2
  - **ambos** dividen los datos de manera aleatoria en K bloques para estimar $\boldsymbol{\eta_0}$
  - **Difieren** en como estimar $\boldsymbol{\theta_0}$ usando las estimaciones de $\boldsymbol{\eta_0}$

- La teoria asintotica existente (donde K es fijo mientras $n \to \infty$) predice:
  - DML1 ~ DML2

- Simulaciones :
  - DML2 $\gtrsim$ DML1

(Q1) Por que deberíamos usar DML2?

(Q2) Como deberíamos elegir K para DML2?

Este paper:

- Estudia las propiedades de DML1 y DML2 bajo una nueva teoria asintotica

Considero secuencias $\boldsymbol{K_n}$ (e.g., $\boldsymbol{K_n = 5}$, $\boldsymbol{K_n \propto \sqrt{n}}$, $\boldsymbol{K_n = n}$)

Bajo esta nueva teoria asintotica explico las diferencias encontradas en simulaciones

Q1: Por que deberiamos usar DML2 ?

- DML2 $\succcurlyeq$DML1 en términos de bias, ECM, e inferencia
- DML2 es robusto a la elección de K

Q2: Como deberíamos elegir K para DML2? (bajo ciertas condiciones)

- K = n minimiza el bias y ECM asintótico para DML2

Outline

- Setup y notación
- Simulaciones
- Resultados principales
  - Por que y cuando DML2 $\succeq$ DML1
  - Como seleccionar K para DML2
- Conclusiones

Modelo Econométrico

- El parámetro de interés $\boldsymbol{\theta} \in \boldsymbol{R}$ satisface :

$$\boldsymbol{E}\left[\boldsymbol{m}\left(\boldsymbol{W}, \boldsymbol{\theta_0}, \boldsymbol{\eta_0}(\boldsymbol{X})\right)\right] = \mathbf{0}$$

donde

$$\boldsymbol{m}(\boldsymbol{W}, \boldsymbol{\theta}, \boldsymbol{\eta}) = \boldsymbol{\psi}^{\boldsymbol{b}}(\boldsymbol{W}, \boldsymbol{\eta}) - \boldsymbol{\psi}^{\boldsymbol{a}}(\boldsymbol{W}, \boldsymbol{\eta})\ \boldsymbol{\theta}$$

$$\boldsymbol{E}\left[\, \boldsymbol{\partial_\eta m}\left(\boldsymbol{W}, \boldsymbol{\theta_0}, \boldsymbol{\eta_0}(\boldsymbol{X})\right) \mid \boldsymbol{X} \right] = \mathbf{0}\ldots$$ (Condición de ortogonalidad de Neyman)

- Ejemplos estudiados en la literatura:
  - Average treatment effect (ATE)
    - Robins et al. (1994): augmented inverse probability weighted (AIPW) estimators
  - Average treatment effect on the treated in difference-in-differences (ATT-DID)
    - Sant'Anna and Zhao (2020): double-robust (DR) estimator
  - Local average treatment effect (LATE) with covariates
    - Tan (2006), Chernozhukov et al. (2018), Singh and Sun (2024): DR and DML estimator

Problema de estimación

- $$E\left[m\left(W,\theta_0,\eta_0(X)\right)\right] = 0 \ \text{ y } \ m(W,\theta,\eta) = \psi^b(W,\eta) - \psi^a(W,\eta)\ \theta$$

- **Objetivo**: estimador $\theta_0 \in \mathbf{R}$ usando una muestra aleatoria $\{W_i : 1 \leq i \leq n\}$

Setup implica

$$\theta_0 = \frac{E\left[\psi^b(W,\eta_0(X))\right]}{E\left[\psi^a(W,\eta_0(X))\right]}$$

- Un estimador ideal

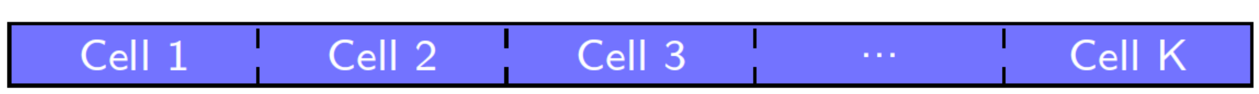
$$\hat{\theta}_n^* = \frac{\sum_{i=1}^n \psi^b(W_i,\eta_i)}{\sum_{i=1}^n \psi^a(W_i,\eta_i)}$$

donde $\eta_i = \eta_0(X_i)$

$\hat{\theta}_n^*$ *inviable* → estimamos $\hat{\eta}_i$ para η_i (1era-etapa) → estimamos θ_0 usando $\hat{\eta}_i$ (2da-etapa)

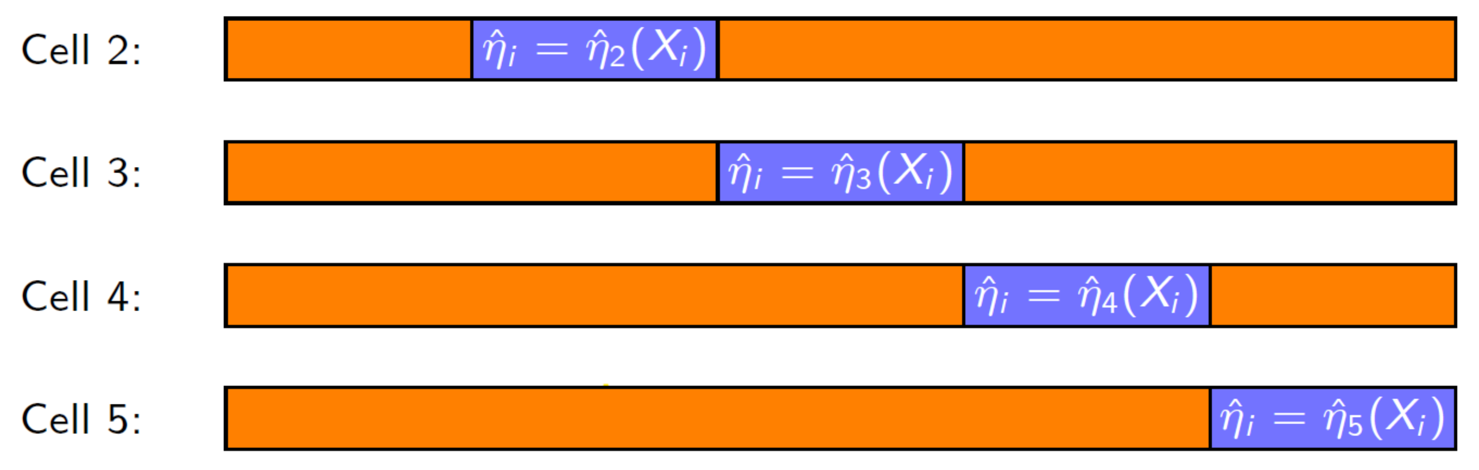
DML: 1era-etapa usa sample-splitting

1. Dividimos la muestra aleatoriamente en K bloques/celdas de igual tamaño



2. Usamos todos los bloques menos uno para estimar la función molestosa $\boldsymbol{\eta_0}$ y predecimos en el bloque restante

3. Repetimos el proceso para todos los bloques(ejemplo K=5)



DML: 2da-etapa

<u>DML1: promedio de estimadores sample-splitting</u>

$$\hat{\boldsymbol{\theta}}_{n,1}(\boldsymbol{K}) = \frac{\mathbf{1}}{\boldsymbol{K}} \underbrace{\sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}} \left(\frac{\sum_{i \in cell\, k} \boldsymbol{\psi}^{b}(\boldsymbol{W}_i, \hat{\boldsymbol{\eta}}_i)}{\sum_{i \in cell\, k} \boldsymbol{\psi}^{a}(\boldsymbol{W}_i, \hat{\boldsymbol{\eta}}_i)} \right)}$$

Estimador $\tilde{\theta}_k$ resuelve estimacion en cell k

$$\sum_{i \in cell\, k} m(W_i, \tilde{\boldsymbol{\theta}}_{\boldsymbol{k}}, \hat{\eta}_i) = 0$$

<u>DML2: estimador de promedio de modelos</u>

$$\hat{\boldsymbol{\theta}}_{n,2}(\boldsymbol{K}) = \underbrace{\frac{\sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}} \sum_{i \in cell\, k} \boldsymbol{\psi}^{b}(\boldsymbol{W}_i, \hat{\boldsymbol{\eta}}_i)}{\sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}} \sum_{i \in cell\, k} \boldsymbol{\psi}^{a}(\boldsymbol{W}_i, \hat{\boldsymbol{\eta}}_i)}}$$

Similar al estimador ideal pero reemplaza $\boldsymbol{\eta}_i$ por $\hat{\boldsymbol{\eta}}_i$

$$\sum_{\boldsymbol{k}=\mathbf{1}}^{\boldsymbol{K}} \sum_{i \in cell\, k} m(W_i, \hat{\boldsymbol{\theta}}_{\boldsymbol{n},\mathbf{2}}(\boldsymbol{K}), \hat{\eta}_i) = 0$$

DML1 y DML2 **difieren** en como estiman $\boldsymbol{\theta}_{\mathbf{0}}$ usando las estimaciones $\hat{\boldsymbol{\eta}}_i$'s

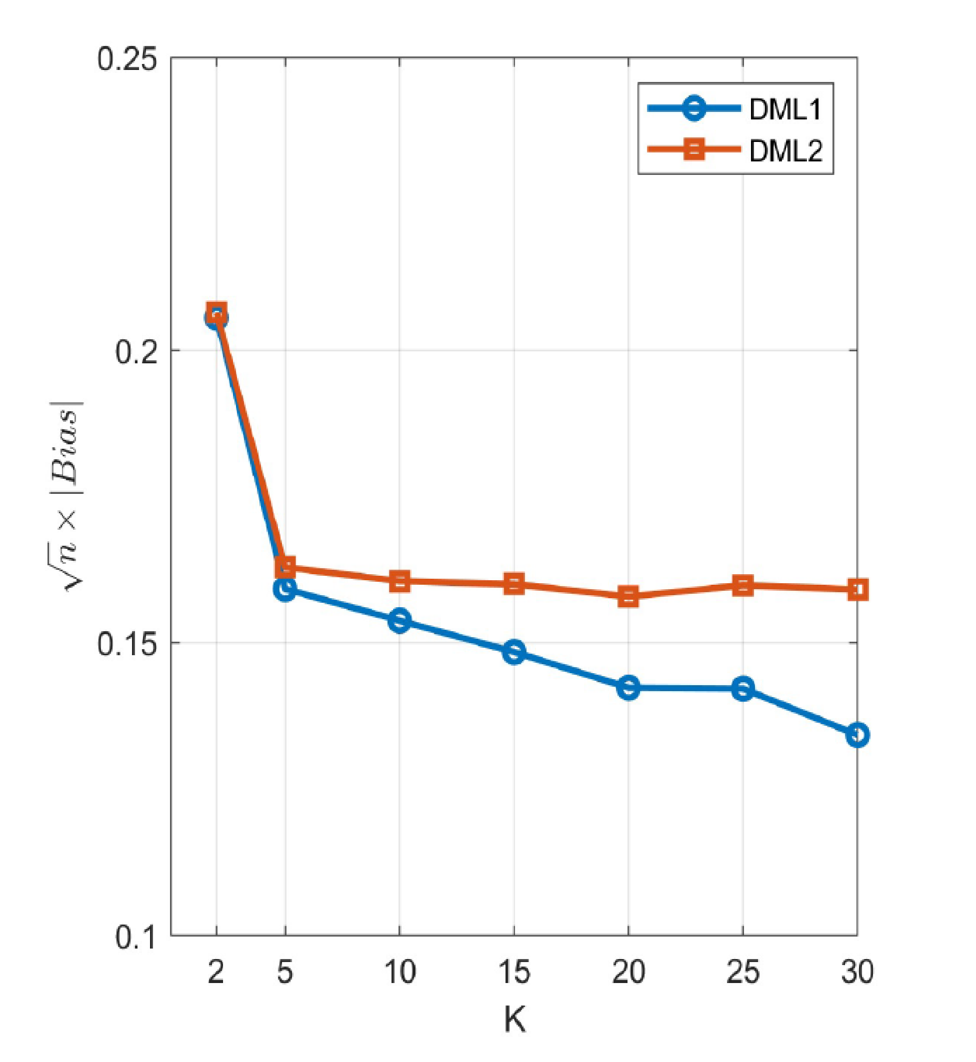
Comentarios

1. Teoría asintótica existente (donde K es fijo mientras $n \to \infty$):
   - Predice DML1 ~ DML2
     - $n^{1/2}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n,j}}(\boldsymbol{K}) - \boldsymbol{\theta_0}\right) \xrightarrow{d} N(0, \sigma^2)$ para j =1,2
   - Chernozhukov et al (2018) propuso intervalos de confianza (IC)
     - 95%-IC: $CI_{n,j} = \hat{\theta}_{n,j}(K) \pm 1.96\, \hat{\sigma}_{n,j}(K)/\sqrt{n}$ for j = 1,2
     - $\hat{\sigma}^2_{n,j}(K)$ estimador para σ^2
2. Simulaciones:
   - Sí DML1 ~ DML2 ....
   - .... pero también DML2 $\succ$ DML1

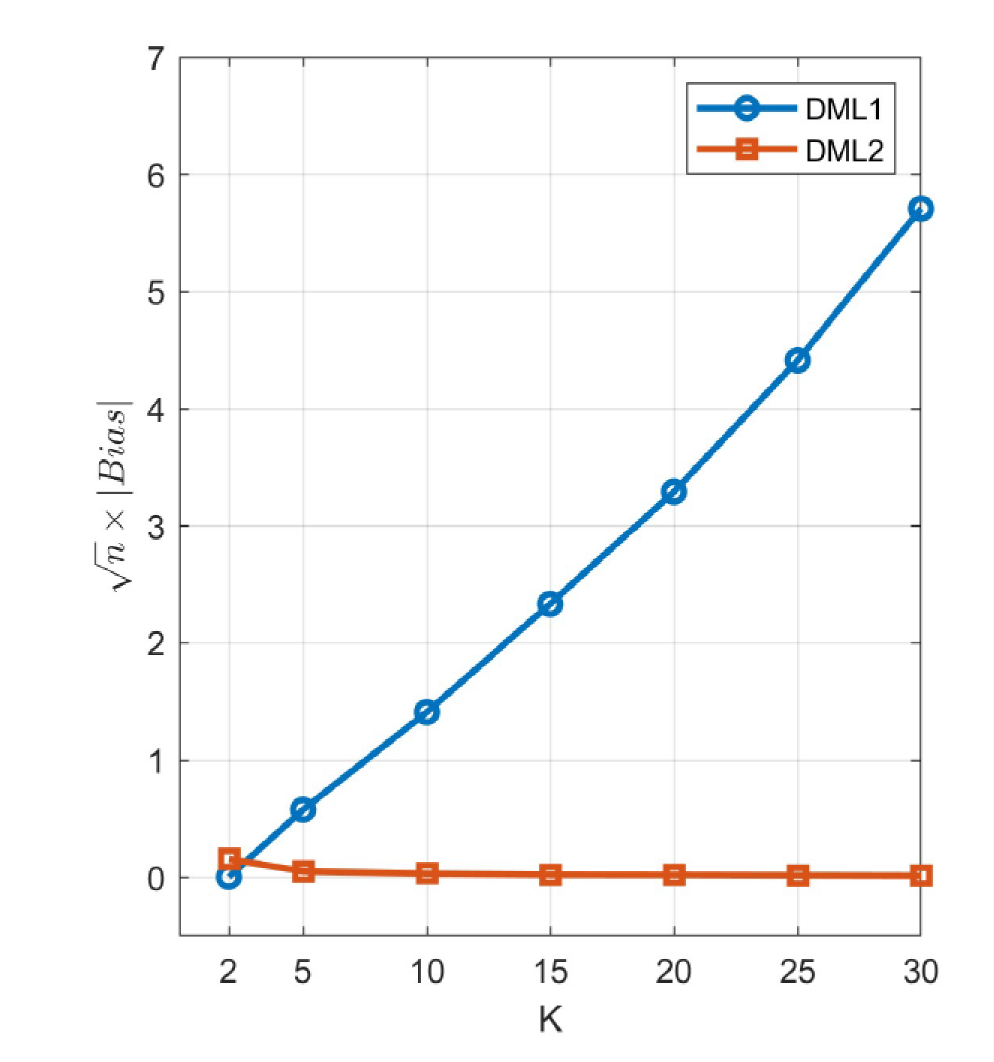
Outline

- Setup y notación
- Simulaciones
- Resultados principales
  - Por que y cuando DML2 $\succeq$ DML1
  - Como seleccionar K para DML2
- Conclusiones

$\sqrt{n}$ |Bias| (sesgo)

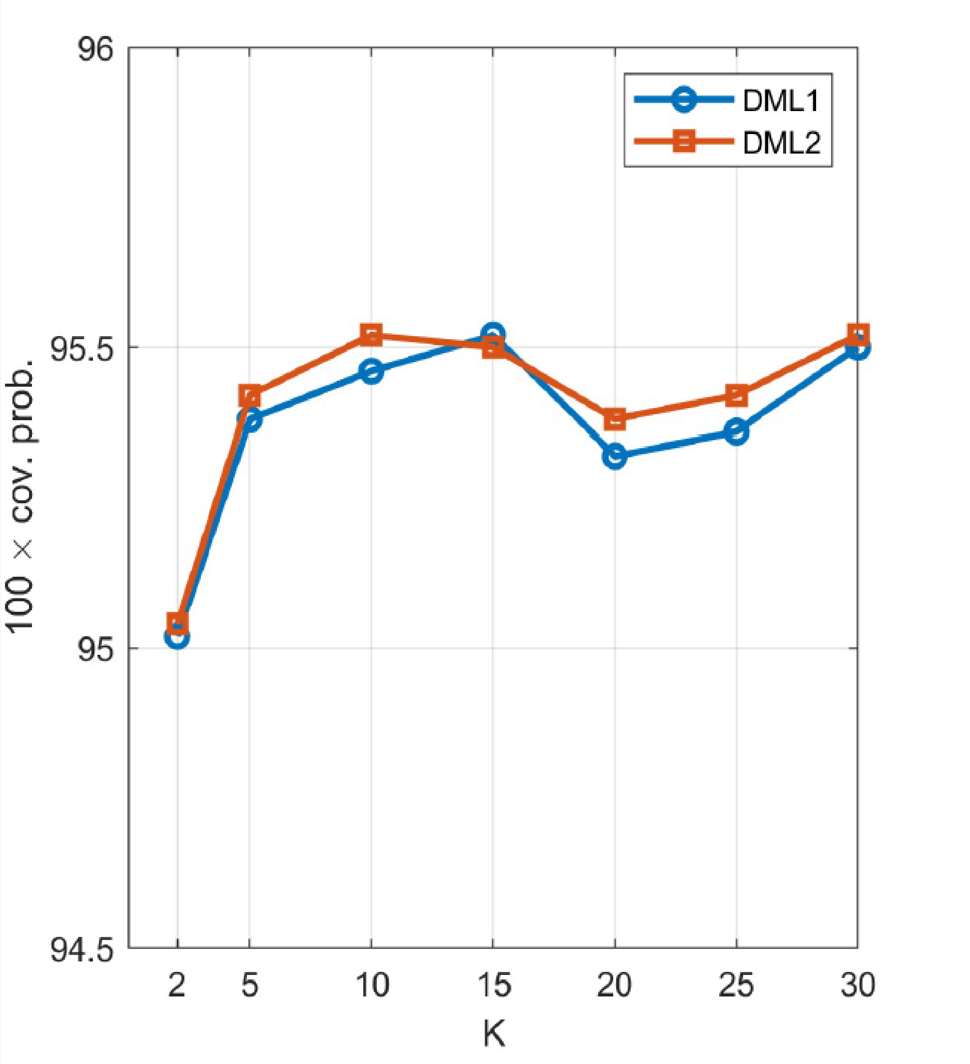


Modelo 1: ATT-DID, Sant'Anna & Zhao (2020)

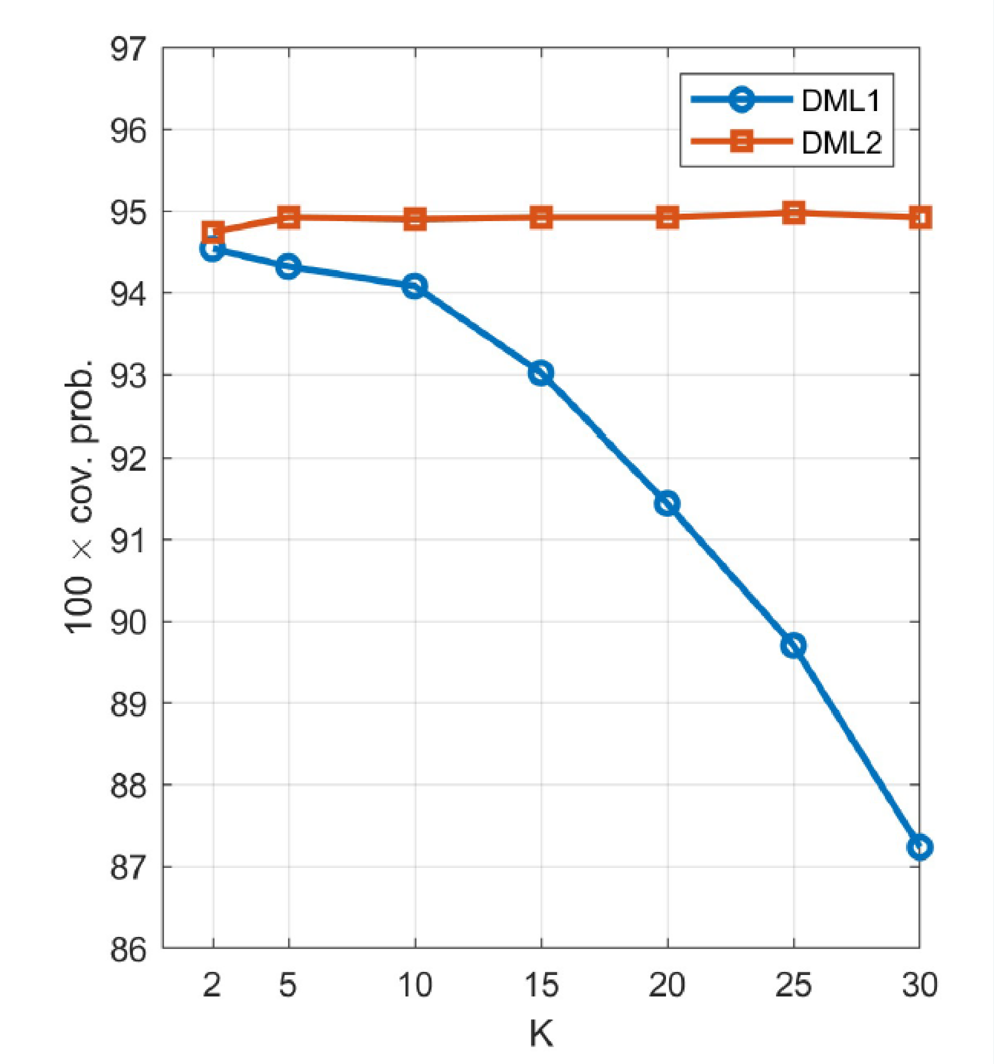


Modelo 2: LATE, Hong and Nekipelov (2010)

Probabilidad de cobertura (en %) de los ICs



Modelo 1: ATT-DID, Sant'Anna & Zhao (2020)



Modelo 2: LATE, Hong and Nekipelov (2010)

Comentarios

1. Simulaciones adicionales muestran:
   - DML1 $\sim$ DML2:
     - ATE, ATT, ATT-DID, PLM
   - DML2 $\succ$DML1
     - LATE, w-ATE, IV-PLM

2. Chernozhukov et al (2018) recomienda en su Remark 3.1 usar DML2
   - Stata package ddml (Ahrens et al. 2024) & R/python package DoubleML usan DML2

Este trabajo/paper:
- Explicación teórica DML2 $\succeq$ DML1
- Condiciones para seleccionar K para DML2

Outline

- Setup y notacion
- Simulaciones
- Resultados principales
  - Por que y cuando DML2 $\geq$ DML1
  - Como seleccionar K para DML2
- Conclusiones

Resultado principal para secuencias K_n

Teorema 1 (DML1):
Sea $\{K_n\}_{n\geq 1}$ una secuencia tal que (i) $K_n \leq n$ y (ii) $\lim_{n\to\infty} \frac{K_n}{\sqrt{n}} = \boldsymbol{c} \in [0,\infty)$. Asumamos que los supuestos de mi *paper* se cumplen. Entonces,

$$n^{1/2}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n},\mathbf{1}}(\boldsymbol{K_n}) - \boldsymbol{\theta_0}\right) \xrightarrow{d} N(\Lambda\, \boldsymbol{c}, \sigma^2)$$

donde

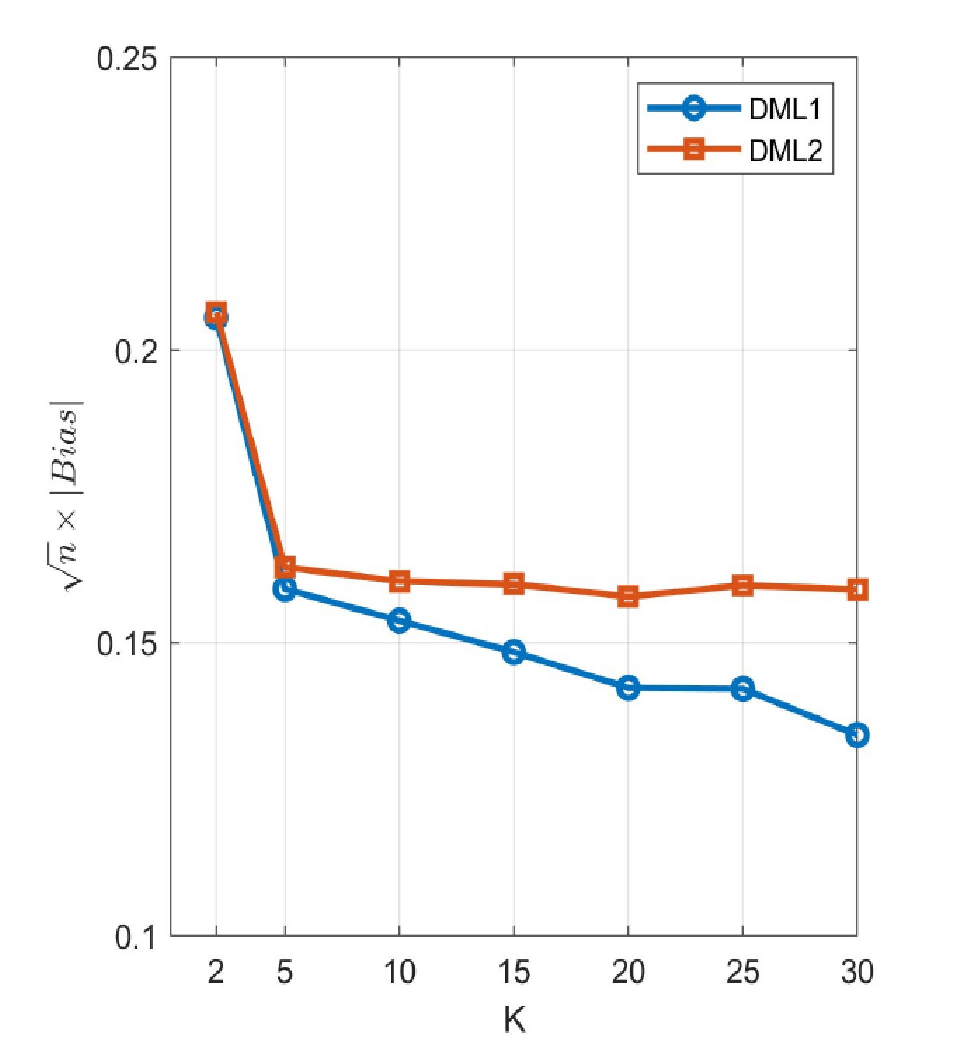
$$\boldsymbol{\Lambda} = \frac{\mathbf{Cov}[\mathbf{m}(\mathbf{W_i}, \boldsymbol{\theta_0}, \boldsymbol{\eta_i}), -\boldsymbol{\psi^a}(\boldsymbol{W_i}, \boldsymbol{\eta_i})]}{\boldsymbol{E}[\boldsymbol{\psi^a}(\boldsymbol{W_i}, \boldsymbol{\eta_i})]^2}$$

Teorema 2 (DML2):
Sea $\{K_n\}_{n\geq 1}$ una secuencia tal que $K_n \leq n$. Asumamos que los supuestos de mi *paper* se cumplen. Entonces,

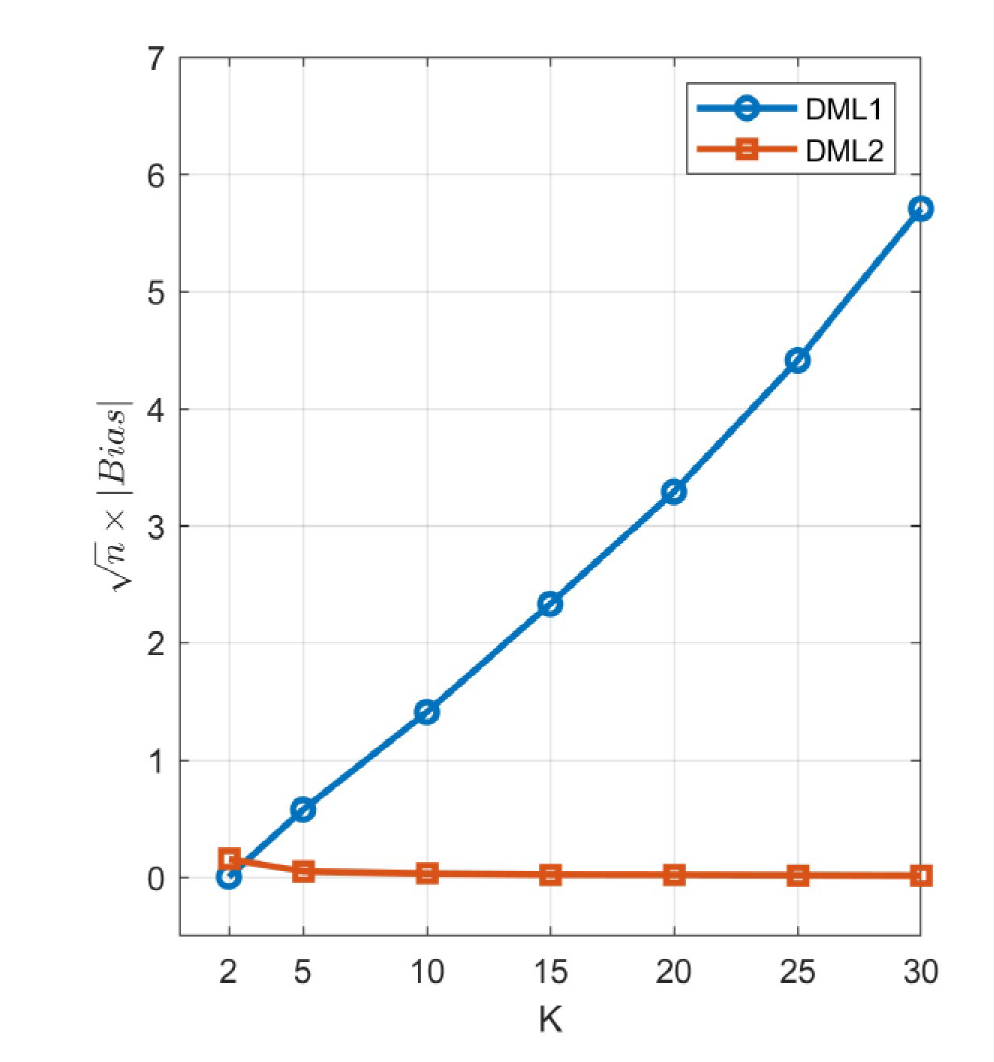
$$n^{1/2}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n},\mathbf{2}}(\boldsymbol{K_n}) - \boldsymbol{\theta_0}\right) \xrightarrow{d} N(0, \sigma^2)$$

- $\boldsymbol{\Lambda}$ y $\boldsymbol{c}$ explican DML2 $\succeq$DML1 y cuando DML2 $\succ$DML1
- Calculo de $\boldsymbol{\Lambda}$ solo depende del modelo econométrico; y podemos usar $\boldsymbol{c} \approx \boldsymbol{K_n}/\sqrt{n}$

$\sqrt{n}|\text{bias}|$ de DML1 es $\approx |\Lambda| c \approx |\Lambda|\, K/\sqrt{n}$

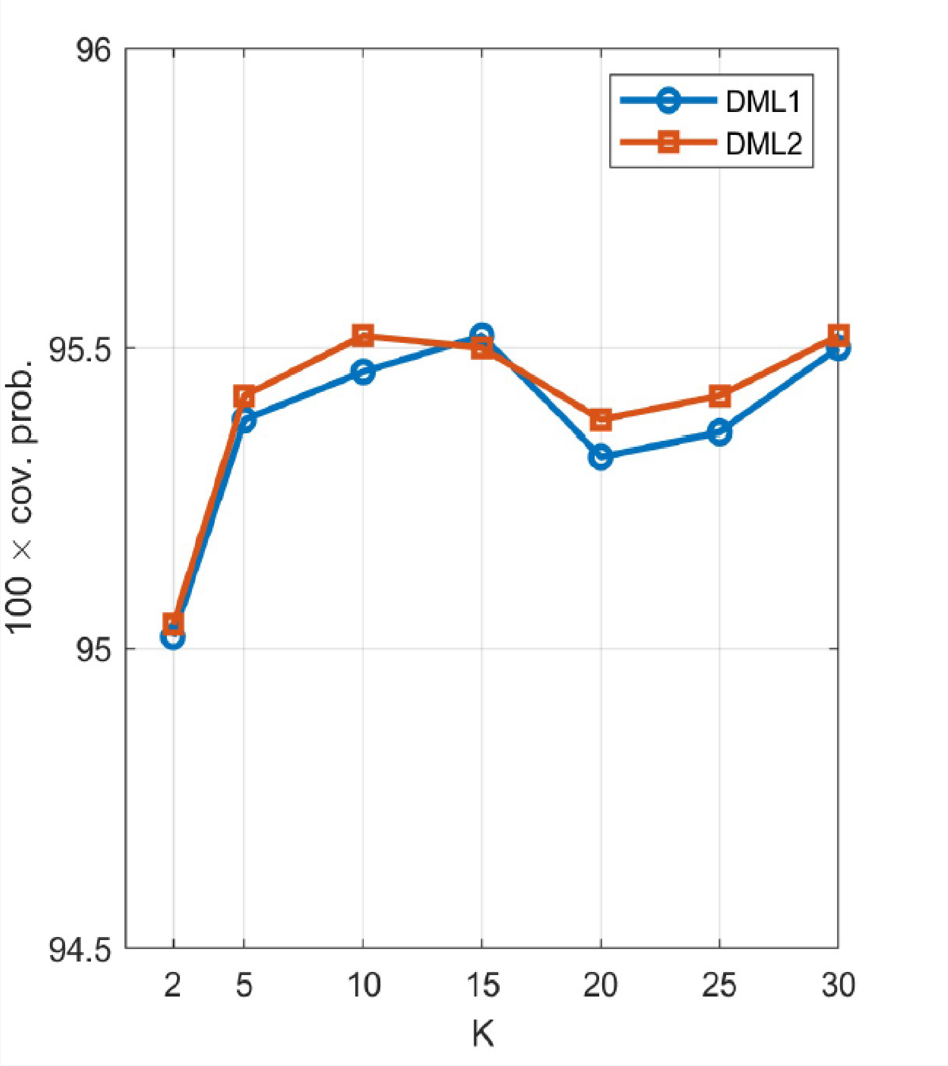


Modelo 1: ATT-DID ($\Lambda = 0$)

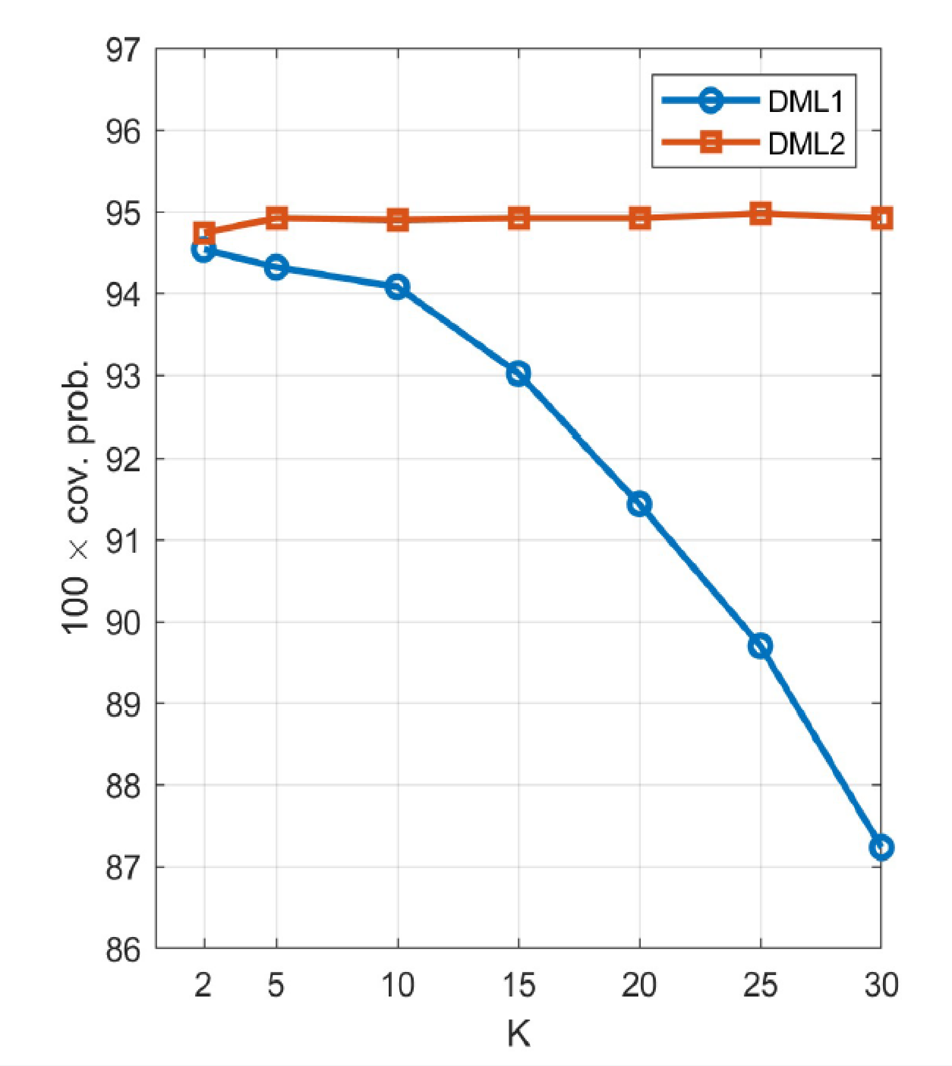


Modelo 2: LATE ($\Lambda \neq 0$)

Distribución de DML1 $\approx N(\boldsymbol{\Lambda K}/\sqrt{\boldsymbol{n}}, \boldsymbol{\sigma}^2)$



Modelo 1: ATT-DID ($\boldsymbol{\Lambda} = \mathbf{0}$)



Modelo 2: LATE ($\boldsymbol{\Lambda} \neq \mathbf{0}$)

Observaciones importantes

$$\boldsymbol{\Lambda} = \frac{\mathbf{Cov}[\mathbf{m}(\mathbf{W_i}, \boldsymbol{\theta_0}, \boldsymbol{\eta_i}), -\boldsymbol{\psi^a}(\boldsymbol{W_i}, \boldsymbol{\eta_i})]}{\boldsymbol{E}[\boldsymbol{\psi^a}(\boldsymbol{W_i}, \boldsymbol{\eta_i})]^2}$$

1. $\boldsymbol{\Lambda}$ es importante para explicar DML2 $\succeq$ DML1 en simulaciones (tomando $\boldsymbol{c} \approx \boldsymbol{K_n}/\sqrt{n}$)
   - $\boldsymbol{\Lambda}$ puede ser calculado para distintos modelos econométricos:
     - ATE, ATT, ATT-DID, y PLM tienen $\boldsymbol{\Lambda = 0}$ $\Rightarrow$ DML2$\sim$ DML1
     - LATE, w-ATE, y IV-PLM tienen $\boldsymbol{\Lambda \neq 0}$ $\Rightarrow$ DML2 $\succ$ DML1
2. Además,
   - DML2 es robusto a la elección de K: $n^{1/2}(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n,2}}(\boldsymbol{K_n}) - \boldsymbol{\theta_0}) \xrightarrow{d} N(0, \sigma^2)$
   - los intervalos de confianza usando DML2 continúan validos

$\Rightarrow$ Inferencia usando DML2 es valida y robusta a la elección de K

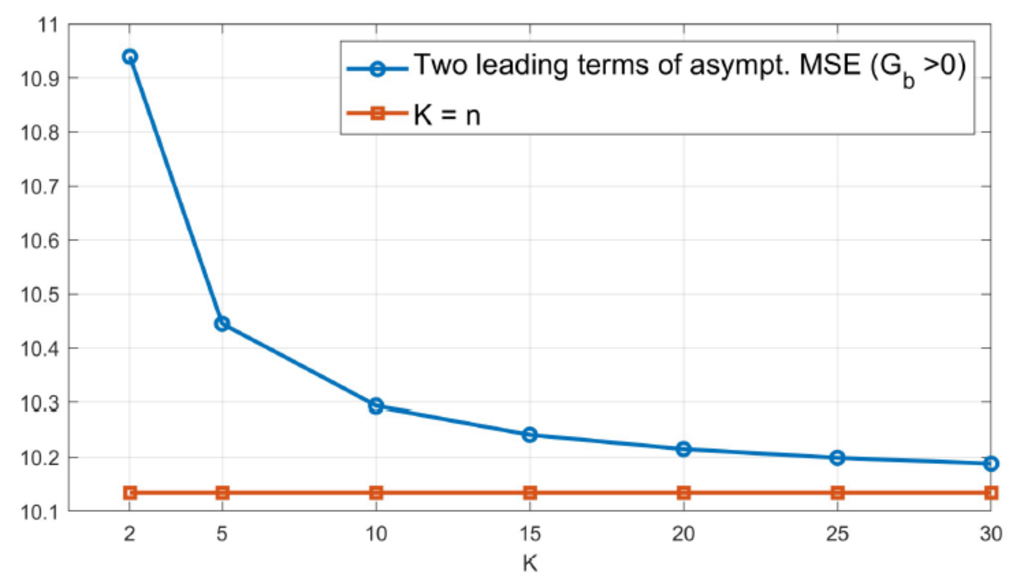
Outline

- Setup y notacion
- Simulaciones
- Resultados principales
  - Por que y cuando DML2 $\succeq$ DML1
  - Como seleccionar K para DML2
- Conclusiones

Aproximación de 2do-orden

$$\underbrace{E\left[n\ \left(\widehat{\boldsymbol{\theta}}_{n,2}(\boldsymbol{K_n})-\boldsymbol{\theta}_0\right)^2\right]}_{\text{n x ECM}} \approx \underbrace{\sigma^2 + G_b\left(1+\frac{1}{\boldsymbol{K_n}-1}\right)^{2\varphi-1/2} n^{1/2-2\varphi} + o(n^{1/2-2\varphi})}_{\text{Aproximación de segundo orden del ECM (SO-ECM)}}$$

Asumimos $\widehat{\boldsymbol{\eta}}$ tiene estructura tal que $\|\widehat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\|_2 = \boldsymbol{O}(\boldsymbol{n}^{-\varphi})$, donde $\varphi \in (\frac{1}{4},\frac{1}{2})$



⇒ DML2 con $K_n = 10$ mas preciso que DML2 con $K_n = 5$

⇒ DML2 con $K_n = n$ mas preciso que DML2 con cualquier K_n constante

Bajo SO-ECM, K_n =n es **optimo**, mientras que recomendación estandar (K_n =5 o 10) es subóptima

Conclusiones

- Este trabajo estudia DML1 y DML2 bajo una nueva teoria asintótica (secuencias K_n)
- Explica las diferencias encontradas en simulaciones
  - Introduce el parámetro Λ
- Brinda dos razones para elegir DML2 sobre DML1
  - DML2 domina asintóticamente a DML1 en términos de bias, MSE, e inferencia
  - DML2 es robusto a la selección de K
- Bajo ciertas condiciones adicionales, se recomienda seleccionar K=n para DML2

Gracias por su atención!