TESIS FINAL
MAESTRÍA EN CIENCIA DE DATOS

UNIVERSIDAD DE MONTEVIDEO

Aplicaciones de Machine Learning en Inferencia Causal: Estudio comparativo con Datos Simulados

---

Rodrigo González e Ignacio González

Tutores: PhD Ana Balsa & PhD Federico Veneri

22 de julio de 2025

Aplicaciones de Machine Learning en Inferencia Causal: Estudio comparativo con datos simulados

Rodrigo González Ignacio González

Junio 2025

Abstract

1. Introducción

La identificación y estimación precisa de efectos causales representa un desafío central en la investigación empírica, especialmente en el ámbito económico y social, donde entender las verdaderas relaciones causa-efecto tiene implicancias claves para el diseño de políticas públicas y estrategias empresariales que logren los efectos deseados. El análisis causal busca superar la simple detección de asociaciones o correlaciones, permitiendo inferir cómo intervenciones o tratamientos específicos modifican directamente a ciertas variables de interés. En este escenario, nos enfrentamos al problema fundamental de la inferencia causal: no es posible observar simultáneamente los resultados potenciales bajo tratamiento y control para un mismo individuo, lo que impide conocer directamente el efecto causal individual. Esta limitación fue formalizada en el marco de los resultados potenciales por Rubin (1974) y por Holland (1986), y constituye la base conceptual sobre la cual se construyen los enfoques modernos de inferencia causal (Imbens y Rubin 2015).

Por esta razón, al no poder observar directamente los efectos causales individuales, es necesario recurrir a diseños experimentales o cuasiexperimentales que permitan identificar dichos efectos. En entornos ideales, los experimentos aleatorios controlados (RCT) constituyen el estándar de oro para inferir causalidad al poder controlar los sesgos derivados de la selección al tratamiento mediante la aleatorización. Sin embargo, la implementación de estos experimentos no siempre es posible, enfrentándonos así muchas veces estudios observacionales [1], debiendo recurrir a métodos cuasiexperimentales como alternativa, siendo éstos ampliamente aplicados y sumamente útiles para evaluar efectos de tratamientos al aproximar las condiciones experimentales mediante el control de variables observables (Rosenbaum y Rubin 1983).

Por lo general, los métodos cuasiexperimentales requieren supuestos claros y específicos sobre las relaciones funcionales entre variables explicativas, tratamientos y resultados, lo que puede generar ciertas dificultades en contextos donde la forma funcional exacta es compleja o desconocida. En tales escenarios, la aplicación directa de técnicas paramétricas tradicionales podría llevar a resultados menos precisos debido a posibles errores de especificación.

Aun métodos que priorizan la robustez, como el enfoque de doble robustez (DR) o su aplicación específica en diferencias en diferencias desarrollado por Sant'Anna y Zhao (2020), requieren supuestos sobre la especificación correcta de modelos in-

[1] Un estudio observacional se refiere al escenario en donde el investigador no controla la asignación al tratamiento, sino que observa datos ya existentes sobre individuos que han sido tratados o no, según decisiones tomadas fuera del diseño del estudio. Esto contrasta con un RCT, donde el tratamiento se asigna de manera exógena y controlada por el investigador.

termedios que no siempre pueden asumirse con seguridad. En las conclusiones de su artículo, Sant'Anna y Zhao (2020) indican lo siguiente:

«*Nuestros estimadores propuestos permanecen consistentes para el ATT cuando cualquiera de los modelos, ya sea el modelo del puntaje de propensión o los modelos de regresión del resultado, están correctamente especificados (aunque no necesariamente ambos), y alcanzan la cota de eficiencia semiparamétrica cuando los modelos utilizados para las funciones auxiliares están correctamente especificados.*» [2]

Esta observación destaca como, si bien los métodos de doble robustez representan un avance crucial para obtener estimaciones insesgadas y robustas, continúan dependiendo críticamente de especificaciones correctas de modelos intermedios.

Frente a esta situación específica, el desarrollo reciente de métodos de machine learning (ML) ofrece un enfoque complementario para la inferencia causal, permitiendo modelar relaciones complejas de forma no paramétrica, directamente a partir de los datos y sin necesidad de asumir previamente una estructura paramétrica específica. Este argumento está ampliamente desarrollado tanto por Athey (2018) como por Chernozhukov et al. (2018) en sus respectivos trabajos.

Se debe que tener en cuenta que la inferencia causal y el ML tradicionalmente abordan objetivos distintos. Mientras el análisis causal busca identificar claramente la influencia de una variable específica, intentando aislar este efecto de otros factores mediante teorías económicas y métodos rigurosos de inferencia estadística, el ML se enfoca principalmente en la predicción, es decir, en construir modelos capaces de anticipar valores futuros o clasificar observaciones con alta precisión sin preocuparse necesariamente por entender la causa subyacente (Athey 2018). Por ejemplo, en el típico caso estudiado de modelos de predicción de default, el enfoque de ML busca anticipar si un cliente va a incumplir o no con los pagos, mientras que el enfoque causal busca entender por qué ocurre ese incumplimiento.

Sin embargo, en el mismo trabajo mencionado, Athey (2018) destaca también cómo estas dos disciplinas, aun teniendo un distinto objetivo, pueden complementarse significativamente. Aunque ML está orientado hacia la predicción y el reconocimiento de patrones en los datos, la gran precisión predictiva que se puede alcanzar mediante su aplicación puede ser sumamente útil en pasos intermedios en diseños de estudio de inferencia causal. Por ejemplo, modelos predictivos precisos pueden mejorar sustancialmente la estimación de funciones de regresión o de puntajes de propensión, elementos cruciales en metodologías causales como el estimador de do-

[2]Traducido de: Sant'Anna, P. H. C., & Zhao, J. B. (2020). *Doubly Robust Difference-in-Differences Estimators.*

ble robustez. De este modo, aunque el objetivo último de la inferencia causal difiere del de la predicción pura, la potencia predictiva de ML puede fortalecer considerablemente la precisión de las estimaciones causales (Chernozhukov et al. 2018).

Este trabajo busca explorar cómo técnicas avanzadas, especialmente el estimador de doble robustez, desarrollado originalmente por Bang y Robins (2005) y ampliado con aplicaciones de ML por Chernozhukov et al. (2018), pueden aprovechar esta flexibilidad predictiva de ML para complementar y mejorar las estimaciones causales en contextos específicos. Para ello, se realiza una revisión de los avances en la literatura, así como un ejercicio práctico controlado en el que se comparan estos nuevos métodos con enfoques econométricos tradicionales bajo distintos escenarios. Dado que en la mayoría de los contextos empíricos el efecto causal verdadero es desconocido se implementará un ejercicio basado en datos simulados en los cuales el efecto del tratamiento es conocido ex ante. De esta manera será posible evaluar con precisión la capacidad de cada método para recuperar el efecto causal verdadero y analizar su desempeño en términos de precisión y robustez bajo diferentes configuraciones del proceso de generación de datos (DGP).

2. Marco teórico y literatura existente

2.1. Fundamentos de la inferencia causal

La inferencia causal busca estimar el efecto de una intervención o tratamiento sobre una variable de interés, comparando lo que habría ocurrido con y sin dicha intervención. Como fue comentado en la introducción, esta comparación contrafactual se formalizó mediante el enfoque de resultados potenciales, desarrollado originalmente por Neyman (1990) en el contexto de experimentos aleatorizados, y popularizado y extendido por Rubin (1974) para estudios observacionales.

En esta sección se introducirán los fundamentos teóricos claves para comprender cómo se define y estima el efecto promedio del tratamiento (Average Treatment Effect, ATE), así como qué supuestos son necesarios para su identificación y por qué la inferencia causal presenta un problema fundamental diferente del análisis puramente predictivo. Esta base permitirá más adelante analizar cómo los métodos de ML pueden ser utilizados de manera integrada en estimaciones causales.

2.1.1. Resultados potenciales y el problema fundamental de la inferencia causal

Introducción y notación básica

Formalmente, es posible presentar el concepto de resultados potenciales utilizando la notación utilizada en el marco de los trabajos de Neyman (1990) y Rubin (1974). Frente a la potencial exposición de un tratamiento dado, se postula que para cada unidad i = 1, ..., n, existe un par de resultados posibles: uno correspondiente a la situación de haber recibido el tratamiento y otro correspondiente a no recibirlo. Formalmente, observamos un conjunto (X_i, Y_i, D_i) compuesto por:

- Un vector de características $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip}) \in \mathbb{R}^p$,
- Una respuesta observada $Y_i \in \mathbb{R}$,
- Una variable de asignación al tratamiento $D_i \in \{0, 1\}$, donde $D_i = 1$ indica que la unidad recibió el tratamiento.

A su vez, se define un par de variables de resultado potencial:

- $Y_i(1)$: resultado que obtendría la unidad i si recibiera el tratamiento,
- $Y_i(0)$: resultado que obtendría si no lo recibiera.

Dado que solo se puede observar el resultado asociado a la situación efectivamente vivida por la unidad i, se cumple que:

$$Y_i = Y_i(D_i),$$

es decir, el resultado observado coincide con uno de los dos resultados potenciales, dependiendo del valor de D_i.

Este marco permite definir de forma precisa un efecto causal individual como la diferencia entre los dos resultados potenciales, es decir, $Y_i(1) - Y_i(0)$.

La clave del planteo de Neyman y Rubin es que este enfoque se basa en un concepto contrafactual. Al definir el efecto causal como una comparación entre los resultados que una misma unidad habría tenido bajo distintas asignaciones se eliminarían los sesgos derivados a la heterogeneidad entre unidades.

En base a este marco, en un análisis causal, nuestro objetivo es estimar el ATE:

$$\tau = \mathbb{E}\left[Y_i(1) - Y_i(0)\right].$$

Esta definición de causalidad presenta una problemática evidente: no es posible observar simultáneamente ambos resultados potenciales para un mismo individuo, siendo éste el anteriormente mencionado problema fundamental de la inferencia causal, introducido por los anteriormente citados Neyman y Rubin.

Solo es posible observar uno de los resultados por individuo, dependiendo de si el individuo recibió el tratamiento ($D_i = 1$) o no ($D_i = 0$), lo que genera un problema de datos faltantes contrafactuales. Para un individuo i, podemos observar empíricamente su resultado potencial como control o su resultado potencial como tratado, pero nunca los dos al mismo tiempo. Esto obliga a realizar diseños experimentales y a utilizar herramientas estadísticas para poder estimar el efecto causal en cuestión.

2.1.2. Identificación del ATE

Randomized Control Trial (RCT)

La forma más sencilla de identificar el ATE en el marco de resultados potenciales es a través de un experimento aleatorizado.

En un RCT, si efectivamente el tratamiento fue asignado de manera aleatoria, entonces la asignación del tratamiento es independiente a los resultados potenciales de cada individuo (Rubin 1974), es decir, no hay ninguna asociación en este caso entre las características de los individuos y su asignación o no al tratamiento:

$$\{Y_i(0), Y_i(1)\} \perp D_i.$$

Esto implica que, bajo asignación aleatoria al tratamiento, los grupos de tratamiento y control son comparables en promedio y no existen diferencias sistemáticas entre ellos ya sea en variables observables o en características no observadas. En consecuencia, cualquier diferencia en los resultados puede atribuirse al efecto del tratamiento.

En estos casos, el ATE se puede estimar simplemente como la diferencia de medias entre el grupo tratado y el de control:

$$\tau = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$$

Por lo tanto, aunque nunca se observe $\tau_i = Y_i(1) - Y_i(0)$, se puede estimar consistentemente $\tau = \mathbb{E}[\tau_i]$ en un experimento aleatorizado.

Un diseño de RCT permite pasar de un escenario donde se define el ATE en función de resultados potenciales a un escenario donde planteamos el ATE como una diferencia de esperanzas condicionales que podemos observar.

Este enfoque es considerado según la literatura ocmo el "estándar oro" para la identificación de efectos causales (Imbens y Rubin 2015) dado que no requiere suponer ninguna forma funcional para las variables ni modelos para la asignación al tratamiento. La validez del estimador proviene directamente del diseño experimental

y no de supuestos estadísticos adicionales.

Sin embargo, no siempre es posible implementar un RCT para la evaluación causal. Como se mencionó previamente, existen numerosas situaciones en las que realizar un RCT no es factible, ya sea por restricciones presupuestarias, consideraciones éticas, o en casos donde simplemente contamos únicamente con datos observacionales para llevar adelante el estudio.[3]

2.1.3. Métodos cuasi-experimentales y el supuesto de independencia condicional

En escenarios donde no es posible implementar un RCT, se recurre a los llamados métodos cuasi-experimentales que buscan aproximar, dentro de lo posible, las condiciones de un RCT aun cuando se trabaja con datos observacionales. Estos métodos permiten estimar efectos causales bajo ciertos supuestos que buscan sustituir el rol de la aleatorización como mecanismo de identificación.

Una de las razones por las que estos supuestos son necesarios es que, en estudios observacionales, la asignación al tratamiento no suele ser aleatoria. Esto implica que el tratamiento puede estar correlacionado con características preexistentes de los individuos, generando sesgos sistemáticos en la estimación del efecto causal. Esta situación se conoce como confounding, y constituye uno de los principales desafíos para identificar correctamente el efecto promedio del tratamiento (Rosenbaum y Rubin 1983) .

Una estrategia comúnmente utilizada en estos escenarios consiste en el ajuste por covariables observadas antes del tratamiento las cuales, en la mayoría de los casos, suelen estar asociadas tanto a la probabilidad de recibir el tratamiento como al resultado de interés. La pregunta central es: ¿bajo qué condiciones es suficiente controlar por esas covariables para poder lograr estimar el ATE correctamente?

Rosenbaum y Rubin (1983) formaliza esta idea mediante el supuesto de ignorabilidad condicional o unconfoundedness. Bajo este supuesto, se asume que, al condicionar en un conjunto de covariables pretratamiento X, la asignación al tratamiento es independiente de los resultados potenciales del individuo. Formalmente:

$$\{Y_i(0), Y_i(1)\} \perp D_i \mid X_i$$

Este supuesto implica que, dentro de los grupos controlados por los valores de las covariables X, no existen diferencias sistemáticas entre aquellos a quienes se les asignó el tratamiento y a aquellos controles que no lo recibieron. En otras palabras, si

[3]En el trabajo de Imbens y Rubin (2015) se discuten en detalle las limitaciones que dificultan la implementación de experimentos aleatorizados en muchos contextos.

bien el diseño no es experimental, condicionar en las covariables observadas permite recrear un escenario donde el tratamiento sea "tan bueno como aleatorio" dentro de cada subgrupo definido por ciertos valores de las covariables X. Este supuesto es fundamental en estudios observacionales dado que nos permite interpretar la diferencia en resultados como un efecto causal, tal como se haría en un RCT.

Cabe aclarar que, para que este supuesto sea válido, es necesario que todas las variables que influyen tanto en la asignación del tratamiento como en el resultado estén incluidas en el vector X. Si alguna covariable relevante no es observada o no es correctamente medida, el supuesto de ignorabilidad se rompe y la estimación resultará sesgada. Como destacan Imbens y Rubin (2015), la ignorabilidad condicional es un supuesto fuerte y no contrastable directamente, pero ofrece una base clara para el análisis causal cuando los datos experimentales no están disponibles.

Es de extrema importancia al utilizar estos métodos la selección adecuada de las variables de control, principalmente, cuando se enfrentan situaciones con muchas covariables potencialmente relevantes. En este contexto, Belloni, Chernozhukov y Hansen (2014) introducen el método denominado Post Double Selection. A través del uso de regularizaciones del tipo LASSO, logran fortalecer la robustez del análisis de manera que errores en la selección inicial no se propaguen significativamente al resultado final, y alcanzan inferencias válidas aún tras una selección imperfecta de variables de control.

Así, bajo ignorabilidad condicional, el ATE puede identificarse en base a la siguiente formulación: [4]

$$\tau = \mathbb{E}_X \left[\mathbb{E}[Y_i \mid D_i = 1, X_i] - \mathbb{E}[Y_i \mid D_i = 0, X_i] \right]$$

donde las funciones condicionales de resultado $\mathbb{E}[Y_i \mid D_i = w, X_i]$ pueden ser estimadas utilizando distintas estrategias que pueden incluir modelos paramétricos tradicionales o técnicas más flexibles, como lo son modelos no paramétricos de aprendizaje automático.

2.2. Machine Learning e inferencia causal

La creciente necesidad de estimar funciones condicionales de forma precisa y flexible ha contribuido al creciente interés por incorporar herramientas de aprendizaje automático en el análisis causal. Sin embargo, esta integración no ha sido inmediata ni directa, y ha dado lugar a un debate importante sobre la validez del uso de métodos de ML en problemas de inferencia causal.

[4] Esta formulación se encuentra en Rosenbaum y Rubin (1983) y también en Imbens y Rubin (2015).

Athey y Imbens (2019) explica por qué los métodos de aprendizaje automático han tenido una adopción relativamente lenta en la economía en comparación con otras disciplinas. Particularmente, argumentan que esta brecha no se debe a una cuestión de capacidad técnica, sino a diferencias fundamentales en los objetivos y criterios que prioriza cada comunidad.

Mientras que en la econometría el énfasis está puesto en la inferencia estadística y tiene su foco en la estimación de efectos causales y la construcción de intervalos de confianza válidos basados en ciertas propiedades como la consistencia, la normalidad asintótica y la eficiencia, en la aplicación de ML se prioriza la capacidad predictiva y el diseño de algoritmos que minimicen el error en muestras de testeo.

Aun frente a un mismo problema como la estimación de una media condicional $E(Y/X = x)$ en un problema supervisado, el enfoque del de ML es distinto al de la econometría causal. Mientras que en econometría tradicional se impone una forma funcional con el objetivo de interpretar los coeficientes y extraer inferencias causales, ML no asume ninguna estructura paramétrica y el interés está puesto en obtener predicciones precisas fuera de muestra. En ML no se busca interpretar el efecto de una covariable específica en la esperanza condicional de la variable de resultado, sino construir un modelo que en su conjunto logre la mayor precisión posible minimizando el error predictivo, incluso si la función estimada es altamente no lineal o incluye interacciones complejas entre variables.

2.2.1. Problemas de ML para causalidad

La diferencia de foco entre el aprendizaje automático y la econometría tradicional ha generado advertencias acerca de la necesidad de precaución al interpretar directamente los resultados de modelos de ML como evidencia causal. Mullainathan y Spiess (2017) muestran como a partir de un mismo ejercicio es posible obtener múltiples funciones de predicción distintas para una misma variable de resultado en función de un conjunto de covariables, todas ellas con niveles similares de precisión fuera de muestra. Esto implica que, aunque un modelo prediga correctamente la variable de resultado a partir de un conjunto de variables explicativas, eso no garantiza que se haya identificado la verdadera relación estructural entre éstas.

Esta ambigüedad se debe a que muchos algoritmos de ML pueden generar predicciones precisas al utilizar diferentes subconjuntos de covariables especialmente cuando éstas están correlacionadas entre sí, lo que hace que la selección de variables dependa fuertemente de la muestra específica con la que se entrene el modelo. Al mismo tiempo, los autores explican como a diferencia de los modelos econométricos tradicionales, donde esta incertidumbre queda reflejada en errores estándar grandes,

en ML esta variabilidad no se muestra explícitamente y no es posible calcular errores estándar válidos de forma directa tras la selección automática del modelo, lo que quita validez a estos métodos para la inferencia causal de manera directa.

Otro punto de interés a tener en cuenta que detallan los autores es como la regularización, común en métodos como LASSO o en los métodos de 'poda de árboles' para disminuir la dimensionalidad de los modelos, contribuye a este problema. Al penalizar la complejidad del modelo, y buscar seleccionar las principales variables del mismo, puede llegar a seleccionar modelos más simples pero incorrectos, omitiendo variables relevantes e introduciendo sesgos.

Cabe destacar que, si bien la regularización también está presente en métodos como LASSO en regresiones lineales, donde penaliza los coeficientes de forma explícita, su impacto puede ser más controlable debido a la estructura lineal impuesta. En cambio en modelos no paramétricos de ML la regularización afecta la selección de variables de forma menos transparente y en un espacio funcional más amplio, lo que puede llevar a omisiones relevantes y sesgos más difíciles de detectar y corregir que en modelos paramétricos con una base teórica.

2.2.2. Aplicaciones válidas de ML al análisis causal

Tal como se ha descrito anteriormente, el principal valor del aprendizaje automático radica en su capacidad de predicción y no así en la identificación directa de relaciones causales. Sin embargo, esto no implica que el ML no tenga aplicaciones valiosas dentro del análisis causal. De hecho, en el trabajo previamente citado, Mullainathan y Spiess (2017) introducen el concepto de "predicción al servicio de la estimación", destacando cómo, en muchos contextos econométricos, el uso de técnicas de predicción logra mejorar significativamente la calidad de las estimaciones causales.

Desde esta perspectiva, no se debe ver al ML como algo que reemplace a la econometría tradicional, sino que actúa como complemento. Su poder predictivo puede aprovecharse para resolver pasos intermedios dentro de los diseños o metodologías de estimación causal, particularmente en entornos de alta dimensionalidad o con relaciones funcionales complejas. Un ejemplo concreto que mencionan los autores es su aplicación en modelos de variables instrumentales (IV), donde los algoritmos de ML pueden emplearse para mejorar la primera etapa del procedimiento de dos etapas (2SLS), al generar mejores predicciones de la variable endógena a partir de los instrumentos disponibles.

Esta lógica ha sido profundizada por otros autores. Por ejemplo, Chernozhukov et al. (2018) propone el marco de Double Machine Learning (DML), en el que tanto

el modelo de resultado como el propensity score son estimados mediante ML como pasos auxiliares y, luego, se utiliza esa información en una estimación robusta del parámetro causal diseñada para que los errores en esos pasos previos no sesguen significativamente el resultado final, resaltando nuevamente que el rol del ML se ubica en la fase predictiva, como una herramienta que alimenta y mejora los pasos posteriores del análisis causal.

2.3. Método de ajuste por regresión bajo el supuesto de independencia condicional

Cuando se cumple el supuesto de independencia condicional, es posible identificar el ATE a partir de las funciones de respuesta condicional:

$$\mu_{(w)}(x) = \mathbb{E}\left[Y_i \mid X_i = x, D_i = w\right],$$

lo que permite expresar el ATE como:

$$\tau = \mathbb{E}\left[\mu_{(1)}(X_i) - \mu_{(0)}(X_i)\right],^{5}$$

Esta expresión fue introducida inicialmente por Rubin (1974) y desarrollada con más detalle en el trabajo conjunto con Imbens (2015). Esta idea indica que, si fuera posible conocer para cada valor de las covariables X cuál sería el resultado esperado si una unidad es tratada y cuál si no lo es, entonces sería posible estimar el ATE comparando estas dos funciones.

Este resultado establece una estrategia concreta para estimar el ATE. En primer lugar, se debe estimar ambas funciones $\hat{\mu}_{(0)}(x)$ y $\hat{\mu}_{(1)}(x)$ a partir de los datos, usando sólo los controles para una y sólo los tratados para la otra. Luego, se calcula el promedio de la diferencia entre esas dos predicciones:

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i)\right).$$

En una primera instancia, se enfoca en la predicción del valor esperado de la variable de resultado en función de las covariables, con estimaciones por separado para las funciones correspondientes a las unidades tratadas y a las de control. En una segunda etapa, estas predicciones se utilizan para estimar el ATE como la diferencia esperada entre ambas funciones obtenidas.

[5]Esta formulación puede encontrarse en el capítulo 13 de Imbens y Rubin (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*

Este estimador será consistente siempre que sea posible predecir correctamente las funciones $\mu_{(d)}(x)$. Esto se logra si son utilizados modelos adecuados para capturar la relación entre las covariables y la variable de resultado, marcando la importancia que tiene el lograr una buena capacidad predictiva para capturar correctamente el efecto del tratamiento, así como alertando de los potenciales problemas que generarían una mala especificación en el modelo usado para predecir la variable de resultado en función de las covariables. Imponer una forma funcional incorrecta para esta predicción puede derivar en sesgos en la estimación del ATE.

2.4. Métodos basados en puntajes de propensión y el estimador IPW

Una estrategia alternativa para estimar efectos causales en estudios observacionales es el uso de puntajes de propensión. Esta metodología desarrollada por Rosenbaum y Rubin (1983), demostraron que bajo el supuesto de independencia condicional y en lugar de utilizar el conjunto completo de covariables, es posible utilizar la probabilidad condicional de recibir el tratamiento dado X, siendo este el puntaje de propensión $e(X) = \mathbb{P}(D = 1 \mid X)$.

Rosenbaum y Rubin (1983) muestran que, si el tratamiento es independiente de los resultados potenciales condicional en X, también lo es condicional en $e(X)$. Esto permite balancear las muestras de tratados y controles de forma similar a RCT, facilitando la estimación del efecto causal.

Una de las implementaciones más utilizadas de este enfoque es el estimador por ponderación inversa de la probabilidad de tratamiento (IPW). Este método asigna un peso a cada observación en función del valor inverso a la probabilidad obtenida de haber recibido el tratamiento con el objetivo de poder obtener una muestra balanceada. Bajo este enfoque, el ATE se puede estimar como:

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{D_i Y_i}{\hat{e}(X_i)} - \frac{(1 - D_i) Y_i}{1 - \hat{e}(X_i)} \right),$$

donde $\hat{e}(X_i)$ es una estimación del puntaje de propensión.

De manera similar al método de ajuste por regresión, este método plantea un mecanismo en dos partes para estimar un efecto causal en donde en la primera se debe estimar consistentemente $e(X)$ para luego utilizar esta función estimada en el cálculo del ATE. Nuevamente, además del supuesto de independencia condicional, es necesario que el modelo para $e(X)$ esté correctamente especificado para que este estimador sea consistente.

2.5. Método de Doble Robustez

Los métodos descritos previamente, el estimador basado en ajuste por regresión (OR) y el estimador por ponderación inversa del puntaje de propensión (IPW), presentan una limitación común: ambos requieren especificar correctamente un modelo para ser consistentes. En el caso del método OR, la consistencia depende de la correcta especificación del modelo de resultado condicional, mientras que en el método IPW, esta depende de especificar correctamente el modelo de puntaje de propensión. En la práctica, sin embargo, es frecuente que los modelos especificados sean incorrectos o aproximados, generando sesgos en la estimación del efecto causal.

Para mitigar este problema, Robins, Rotnitzky y Zhao (1994) introdujeron el estimador de doble robustez (DR), desarrollado luego por Bang y Robins (2005). Este estimador combina ambos métodos mencionados, OR e IPW, en una única estimación, proporcionando una garantía adicional frente a errores de especificación. Esto implica que el estimador es consistente siempre que al menos uno de los dos modelos auxiliares sea correcto.

Formalmente, el modelo DR se define como:

$$\hat{\tau}_{DR} = \frac{1}{N}\sum_{i=1}^{N}\left[\mu(1,X_i) - \mu(0,X_i) + \frac{D_i}{\hat{e}(X_i)}(Y_i - \mu(1,X_i)) - \frac{1-D_i}{1-\hat{e}(X_i)}(Y_i - \mu(0,X_i))\right], \tag{1}$$

La propiedad clave del estimador DR es que si al menos uno de los modelos, ya sea el de resultado o el de score de propensión está correctamente especificado, el estimador será consistente y asintóticamente normal. Por lo tanto, esta metodología ofrece dos oportunidades independientes para obtener estimaciones válidas del efecto causal, disminuyendo la vulnerabilidad del análisis frente a errores de especificación (Bang y Robins, 2005).

2.6. Double Debiased Machine Learning

Si bien los métodos de doble robustez reducen el riesgo de sesgos al requerir solo que uno de los modelos auxiliares esté correctamente especificado para asegurar la consistencia del estimador, este enfoque no elimina completamente el problema de la mala especificación. En la práctica, especialmente cuando las relaciones entre covariables, tratamiento y resultado son complejas, existe el riesgo de que ambos modelos estén mal especificados, afectando la validez de la estimación causal.

Ante este desafío, la incorporación de técnicas de ML representa una oportunidad para capturar relaciones complejas y reducir los errores de especificación mediante

modelos más flexibles. Sin embargo, Chernozhukov et al. (2018) advierte que integrar directamente ML en estimadores causales introduce un nuevo desafío: el sesgo por regularización. En particular, estimar la función de resultado condicional $g_0(X)$ con algoritmos de ML suele implicar un sesgo significativo debido a la regularización utilizada para evitar sobreajuste. Este sesgo no siempre disminuye lo suficientemente rápido con el aumento del tamaño muestral, lo que puede provocar que la tasa de convergencia del estimador causal sea más lenta que la tasa paramétrica habitual, dificultando la inferencia tradicional.

Para resolver este inconveniente, Chernozhukov et al. (2018) propone el método de Double Debiased Machine Learning (DML), que introduce dos innovaciones centrales: una estructura ortogonalizada respecto a las funciones auxiliares estimadas y el uso de *sample splitting*, es decir, dividir la muestra en partes independientes para estimar las funciones auxiliares mediante ML en una parte y, con esas predicciones, estimar el parámetro causal en la otra.

Citando a Bach et al. (2022), quien realiza una revisión exhaustiva del método DML (Chernozhukov et al. 2018) y su implementación práctica estandarizada en R, se destaca que el éxito de la inferencia causal mediante este método se basa en tres pilares fundamentales: la ortogonalidad de la función de *score*, el uso de *sample splitting* y la aplicación de algoritmos de ML de buena calidad, es decir, modelos que logran un adecuado equilibrio entre evitar el *overfitting* y el *underfitting*. Según el autor, la combinación de estos tres elementos garantiza que las estimaciones de efectos causales sean consistentes y válidas, incluso en contextos complejos o de alta dimensionalidad, posibilitando así una integración efectiva de técnicas de ML en el análisis causal empírico. La integración de estimadores ortogonales y del *sample splitting* introducen dos mecanismos de eliminación de sesgo, logrando este estimador *double debiased*.

En términos matemáticos, para el caso parcialmente lineal clásico, el estimador DML propuesto por Chernozhukov et al. (2018) se expresa como:

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

donde $\hat{V}_i = D_i - \hat{m}_0(X_i)$ representa el regresor ortogonalizado, $\hat{g}_0(X_i)$ es la predicción del resultado obtenida mediante ML, y $\hat{m}_0(X_i)$ es la probabilidad de recibir el tratamiento a partir de las covariables X_i (el *propensity score*). Esta construcción asegura que los errores de regularización de los algoritmos de ML no se transmitan directamente al estimador final, mejorando así su consistencia.

Ortogonalidad de Neyman

La ortogonalidad de Neyman es uno de los componentes fundamentales de este método, dado que es el mecanismo matemático que permite que el estimador sea insensible a pequeños errores en la estimación de las funciones auxiliares. Mientras los errores de ML no sean demasiado grandes en la estimación de la primera etapa, estos no afectan la validez de la inferencia sobre θ.

Esta propiedad es la base que permite que DML combine la potencia predictiva de ML para funciones auxiliares con la rigurosidad requerida en inferencia causal, tolerando errores inevitables de modelos no paramétricos o altamente flexibles y sin perder validez en la estimación final. De este modo, DML permite aprovechar la flexibilidad del ML sin perder solidez estadística, lo que resulta clave en contextos complejos o de alta dimensionalidad donde ningún modelo predictivo logra aproximar perfectamente la función real. Por lo explicado, se puede entonces utilizar algoritmos de ML de manera integrada en la fase de predicción para estimar las funciones auxiliares en problemas de inferencia causal, siempre y cuando el estimador causal que empleemos cumpla con la ortogonalidad de Neyman y el modelo predictivo arroje resultados con niveles de precisión aceptables.

Importancia del sample splitting y cross-fitting

Como se comentó anteriormente, para que el método DML sea válido, es fundamental evitar que el modelo de ML utilice la misma información tanto para ajustar las funciones auxiliares como para estimar el efecto causal. Si se utilizan las mismas observaciones para ambos pasos, el modelo puede sobreajustarse a los datos y trasladar ese sobreajuste a la estimación de θ, generando sesgos incluso si el estimador es ortogonal.

El sample splitting y el cross-fitting resuelven este problema al dividir la muestra para que en cada observación, las predicciones de las funciones auxiliares provengan de un modelo que no vio esa observación durante el entrenamiento. Así, se evita la dependencia artificial y se garantiza que la estimación causal sea robusta.

3. Evaluación comparativa sobre datos simulados

En esta sección se desarrollará un ejercicio comparativo entre el estimador DML y otros enfoques previamente descritos, con el objetivo de evaluar el desempeño de cada uno en la estimación ATE bajo diferentes escenarios. Se considerarán escenarios que varían en la forma funcional de las funciones auxiliares, siendo estas el modelo de resultado y el modelo de propensity scores, así como variaciones en el tamaño

muestral disponible. Para realizarlo, se utilizará un enfoque basado en datos simulados que permite contar con un entorno controlado en donde es posible conocer el valor verdadero del ATE. Esto posibilita medir con precisión el sesgo, la varianza y la sensibilidad de cada estimador ante errores de especificación o limitaciones en el tamaño de muestra, para obtener una comparación clara de sus fortalezas y debilidades en contextos observacionales realistas.

3.1. Proceso generativo de datos

Siguiendo el enfoque metodológico propuesto por Chernozhukov et al. (2018), y su implementación práctica en R por Bach et al. (2022), este trabajo construye su ejercicio empírico sobre un modelo parcialmente lineal. Este tipo de estructura es muy útil dado que permite separar de forma explícita el efecto causal de interés, representado por el parámetro estructural θ_0, de los componentes no paramétricos del modelo, que pueden ser estimados mediante técnicas de ML.

De forma específica, el proceso generativo utilizado se basa en el diseño del experimento de simulación contenido en la documentación oficial del paquete DoubleML (Bach et al. 2022), el cual, a su vez, está basado en el marco teórico propuesto por Chernozhukov et al. (2018).

Formalmente, el modelo se especifica como:

$$Y_i = \theta_0 D_i + g_0(X_i) + \varepsilon_i, \qquad D_i = m_0(X_i) + u_i$$

donde Y_i representa la variable de resultado, D_i la asignación al tratamiento (variable binaria que toma valor 1 si la unidad está tratada y 0 en caso contrario), X_i el vector de covariables, $g_0(X_i)$ la función de resultado condicional y $m_0(X_i)$ la función que modela la probabilidad de recibir el tratamiento. Los términos de error ε_i y u_i son independientes y distribuidos normalmente como $\mathcal{N}(0, 1)$.

Siguiendo el enfoque de los autores mencionados anteriormente, el parámetro causal verdadero se fija en $\theta_0 = 0{,}5$. El conocer este valor permite calcular con precisión el sesgo y el root mean square error (RMSE) de los estimadores considerados.

Replicando el diseño de Bach et al. (2022), las covariables X_i se generan como una muestra de una normal multivariada de dimensión p con media cero y matriz de covarianza Σ definida como:

$$\Sigma_{jk} = 0{,}7^{|j-k|}$$

3.1.1. Funciones auxiliares

Las funciones auxiliares $g_0(X)$ y $m_0(X)$ son componentes fundamentales del modelo parcialmente lineal que define el proceso generativo de datos (DGP). Estas funciones determinan respectivamente la forma en que las covariables afectan la variable de resultado y la probabilidad de asignación al tratamiento. Tal como se desarrolló en el marco teórico, la correcta especificación de estas funciones es crucial en métodos paramétricos como el estimador de doble robustez, cuya consistencia requiere que al menos uno de los dos modelos auxiliares esté correctamente especificado. En cambio, los enfoques que integran técnicas de ML ofrecen una mayor flexibilidad, al permitir aproximar estas funciones de forma no paramétrica y así capturar relaciones complejas entre las variables sin necesidad de imponer ex ante una forma funcional específica.

Funciones para el Modelo de Resultados: $g_0(X)$

Se consideran dos posibles formas funcionales para la función de resultado condicional, según el escenario:

- **Lineal:**

$$g_0(X) = 0{,}5\, x_1 - 0{,}3\, x_2 + 0{,}2\, x_3$$

- **No lineal:**

$$g_0(X) = \frac{\exp(x_1)}{1 + \exp(x_1)} + \frac{1}{4}\, x_3 + 0{,}1\, x_2^2$$

Se plantean dos posibles especificaciones funcionales para la función de resultado $g_0(X)$: una lineal, definida como una combinación lineal de tres regresores (x_1, x_2 y x_3), y otra no lineal, basada en el ejemplo utilizado por Chernozhukov et al. (2018) e implementado por Bach et al. (2022), la cual fue ampliada en este trabajo para incorporar explícitamente las mismas tres covariables consideradas en el caso lineal. Esta modificación permite realizar una comparación controlada entre ambos escenarios manteniendo constante el conjunto de variables relevantes entre ambos escenarios.

Funciones para el Propensity Score: $m_0(X)$

Se emplean también dos alternativas funcionales para la probabilidad de asignación al tratamiento:

- **Logística:**

$$m_0(X) = \frac{\exp(x_1 + x_2 + x_3)}{1 + \exp(x_1 + x_2 + x_3)}$$

- **No logística:**

$$m_0(X) = \sin(x_1 + x_2) + \log(|x_3| + 1) + 0{,}1\, x_2^2$$

Determinando la asignación al tratamiento como una variable binaria tal que si el valor de $m_0(X)$ sumado al error u_i esta por encima de un cierto umbral definido por la mediana de este puntaje de asignación de tratamiento, entonces va a estar definido como tratado $D_i = 1$, en caso contrario, como control, $D_i = 0$.

$$D_i = 1\left\{m_0(X_i) + u_i > q\right\}, \qquad u_i \sim \mathcal{N}(0, 1)$$

El hecho de usar la mediana del score de propensión como punto de corte para la determinación del tratamiento, permite asegurar un reparto balanceado de unidades tratadas y no tratadas.

Cabe aclarar que, este punto, se basa también en el enfoque propuesto por Bach et al. (2022) y Chernozhukov et al. (2018), aunque con una modificación en la forma funcional del propensity score no logístico. En lugar de utilizar la función original que incluía una transformación logística de uno de los regresores, se optó por una especificación alternativa que combina componentes no lineales, como funciones seno, logarítmica y cuadrática, sin involucrar explícitamente una estructura logística. Se realizó este cambio con el fin de evaluar la performance de los estimadores frente a escenarios donde la relación de las variables sea más compleja y más diferenciada de una forma logística, dado que ya se va a analizar un escenario con una forma funcional que siga esta distribución.

3.2. Escenarios de evaluación

Se comparan cuatro escenarios simulados que varían según la especificación de las funciones auxiliares $g_0(X)$ (modelo de resultado) y $m_0(X)$ (modelo de asignación al tratamiento):

- Escenario 1: $g_0(X)$ no lineal, $m_0(X)$ no logístico.
- Escenario 2: $g_0(X)$ lineal, $m_0(X)$ no logístico.
- Escenario 3: $g_0(X)$ no lineal, $m_0(X)$ logístico.
- Escenario 4: $g_0(X)$ lineal, $m_0(X)$ logístico.

3.2.1. Estimadores a comparar

En base a estos escenarios, se evaluará el desempeño de los distintos estimadores bajo especificaciones con distinto grado de complejidad en cuanto a la relación entre las variables y a las no linealidades incorporadas.

Los estimadores que serán evaluados en cada escenario son:

- DML (Double Debiased Machine Learning)
- DR Tradicional Paramétrico
- DR con Machine Learning plug-in

La comparación entre estos estimadores permite evaluar la capacidad del DML para adaptarse y estimar de forma flexible las funciones auxiliares $g_0(X)$ y $m_0(X)$ mediante técnicas de ML en contextos de distinta complejidad funcional. Tal como se explicó previamente, cada estimador combina de forma diferente tres características clave: el uso de ML para estimar funciones auxiliares, la ortogonalidad de Neyman y la implementación de sample splitting. La siguiente tabla resume estas propiedades:

Cuadro 1: Resumen de propiedades técnicas de los estimadores

Estimador	**Usa ML**	**Ortogonal**	**Sample Splitting**
DML	Si	Si	Si
DR Tradicional	No	Si	No
DR Plug-in (RF)	Si	Si	No

Se espera que el estimador DML, al reunir las tres propiedades, mantenga un buen desempeño en todos los escenarios, incluso en aquellos donde las funciones auxiliares presentan relaciones más complejas. El estimador DR tradicional debería funcionar bien solo cuando al menos una de las funciones está correctamente especificada, mientras que el DR plug-in con ML, al no cumplir con las condiciones necesarias para poder utilizar ML para inferencia causal descritas en Chernozhukov et al. (2018), podría presentar mayor sesgo o varianza cuando las funciones auxiliares son difíciles de estimar.

Para cada estimador se evaluará el sesgo frente al valor verdadero de tratamiento ($\theta_0 = 0{,}5$) y el RMSE para evaluar la precisión de cada estimador, así como los errores estándar para medir la dispersión en las estimaciones obtenidas.

3.2.2. Estadísticos descriptivos

Se realiza un análisis descriptivo para cada conjunto de datos generado que conforma cada escenario con $n = 10{,}000$ y 5 regresores. La idea de este análisis es poder verificar y mostrar con mayor claridad cómo las bases generadas siguen los patrones definidos con anterioridad. En cada caso se muestra la comparación entre la distribución empírica de las variables explicativas X_1 a X_5 frente a una distribución teórica $N(0, 1)$, que es la designada en el DGP para la generación de las variables, así como estadísticos de resumen como la media de su valor y la desviación estándar para la variable de resultado y las covariables, tanto por grupo de tratamiento como en forma global.

Escenario 1: función de resultado no lineal y asignación al tratamiento no logística.

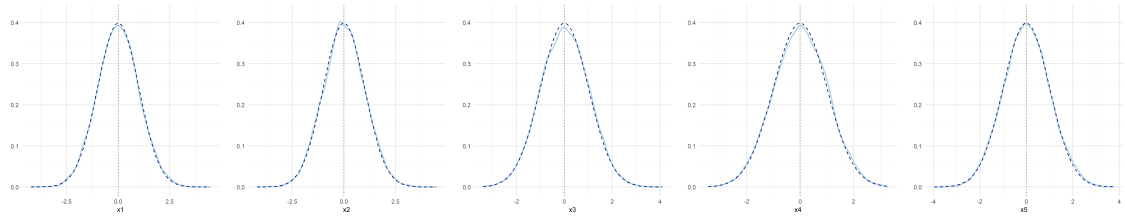


Figura 1: Distribuciones empíricas de las covariables X_1 a X_5 comparadas con una $N(0, 1)$. La línea azul representa la distribución observada; la línea punteada en rojo, la distribución normal teórica.

Variable	Y	X_1	X_2	X_3	X_4	X_5
Promedio (sd)	0.874 (1.135)	0.002 (1.005)	0.013 (1.005)	0.014 (1.001)	0.020 (1.006)	0.014 (1.004)

Grupo	n	Prop. (%)	X_1	X_2	X_3	X_4	X_5	Y
0	5000	50 %	-0.220 (0.914)	-0.218 (0.859)	-0.149 (0.842)	-0.087 (0.934)	-0.058 (0.962)	0.523 (1.071)
1	5000	50 %	0.224 (1.042)	0.245 (1.083)	0.177 (1.114)	0.127 (1.063)	0.086 (1.040)	1.225 (1.088)

Escenario 2: función de resultado lineal y asignación al tratamiento no logística.

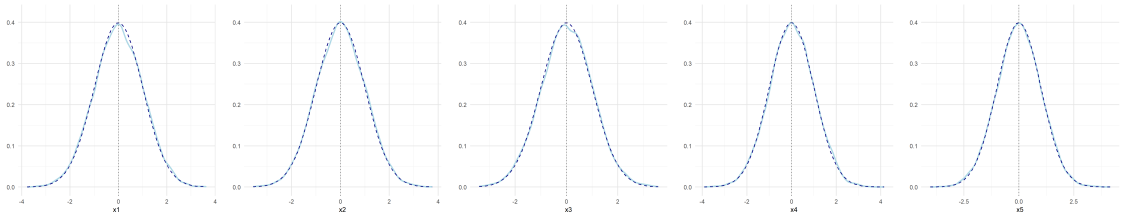


Figura 2: Distribuciones empíricas de las covariables X_1 a X_5 en el Escenario 2.

Variable	Y	X_1	X_2	X_3	X_4	X_5
Promedio (sd)	0.256 (1.142)	0.005 (1.012)	0.009 (1.003)	0.013 (1.008)	0.012 (1.004)	-0.001 (1.007)

Grupo	n	Prop. (%)	X_1	X_2	X_3	X_4	X_5	Y
0	5000	50 %	-0.221 (0.921)	-0.215 (0.849)	-0.159 (0.838)	-0.112 (0.923)	-0.099 (0.970)	-0.093 (1.091)
1	5000	50 %	0.231 (1.047)	0.234 (1.092)	0.186 (1.127)	0.136 (1.064)	0.096 (1.035)	0.605 (1.084)

Escenario 3: función de resultado no lineal y asignación al tratamiento logística.

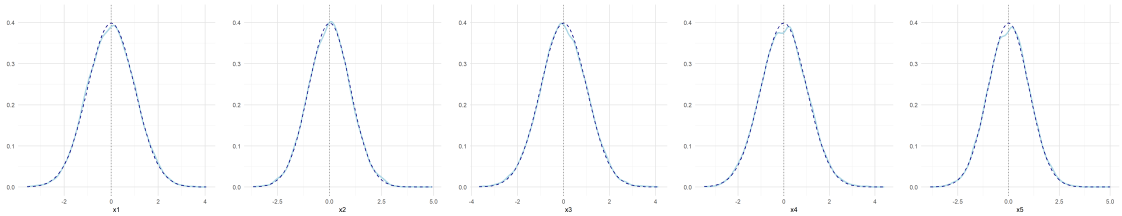


Figura 3: Distribuciones empíricas de las covariables X_1 a X_5 en el Escenario 3.

Variable	Y	X_1	X_2	X_3	X_4	X_5
Promedio (sd)	0.834 (1.126)	0.004 (1.000)	0.000 (1.005)	0.003 (1.012)	0.010 (1.008)	0.012 (1.006)

Grupo	n	Prop. (%)	X_1	X_2	X_3	X_4	X_5	Y
0	5000	50 %	-0.223 (0.982)	-0.244 (0.984)	-0.212 (0.997)	-0.150 (0.989)	-0.093 (0.999)	0.510 (1.066)
1	5000	50 %	0.231 (0.967)	0.245 (0.967)	0.218 (0.980)	0.170 (1.002)	0.117 (1.002)	1.158 (1.090)

Escenario 4: función de resultado lineal y asignación al tratamiento logística.

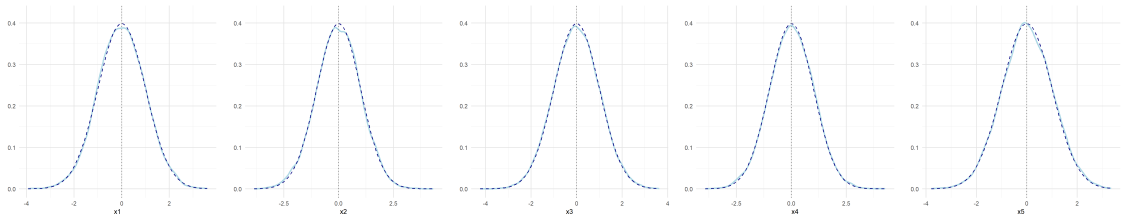


Figura 4: Distribuciones empíricas de las covariables X_1 a X_5 en el Escenario 4.

Variable	Y	X_1	X_2	X_3	X_4	X_5
Promedio (sd)	0.237 (1.126)	-0.001 (0.994)	-0.009 (1.007)	-0.002 (1.005)	0.010 (1.009)	0.016 (1.006)

Grupo	n	Prop. (%)	X_1	X_2	X_3	X_4	X_5	Y
0	5000	50 %	-0.212 (0.960)	-0.243 (0.969)	-0.221 (0.970)	-0.157 (0.980)	-0.090 (0.992)	-0.094 (1.065)
1	5000	50 %	0.211 (0.982)	0.226 (0.989)	0.217 (0.992)	0.176 (1.011)	0.122 (1.008)	0.568 (1.086)

3.3. Implementación en R y Reproducibilidad

3.3.1. Simulaciones Monte Carlo

3.4. Resultados

Resultados de las simulaciones por método y escenario

	DML			DR paramétrico			DR plug-in ML		
Escenario	$\hat{\theta}$	SE	RMSE	$\hat{\theta}$	SE	RMSE	$\hat{\theta}$	SE	RMSE
1	0.497	0.032	0.044	0.646	0.031	0.150	0.529	0.017	0.042
2	0.496	0.033	0.045	0.500	0.029	0.037	0.482	0.017	0.039
3	0.505	0.025	0.034	0.523	0.083	0.114	0.508	0.023	0.033
4	0.508	0.025	0.034	0.504	0.058	0.078	0.509	0.023	0.034

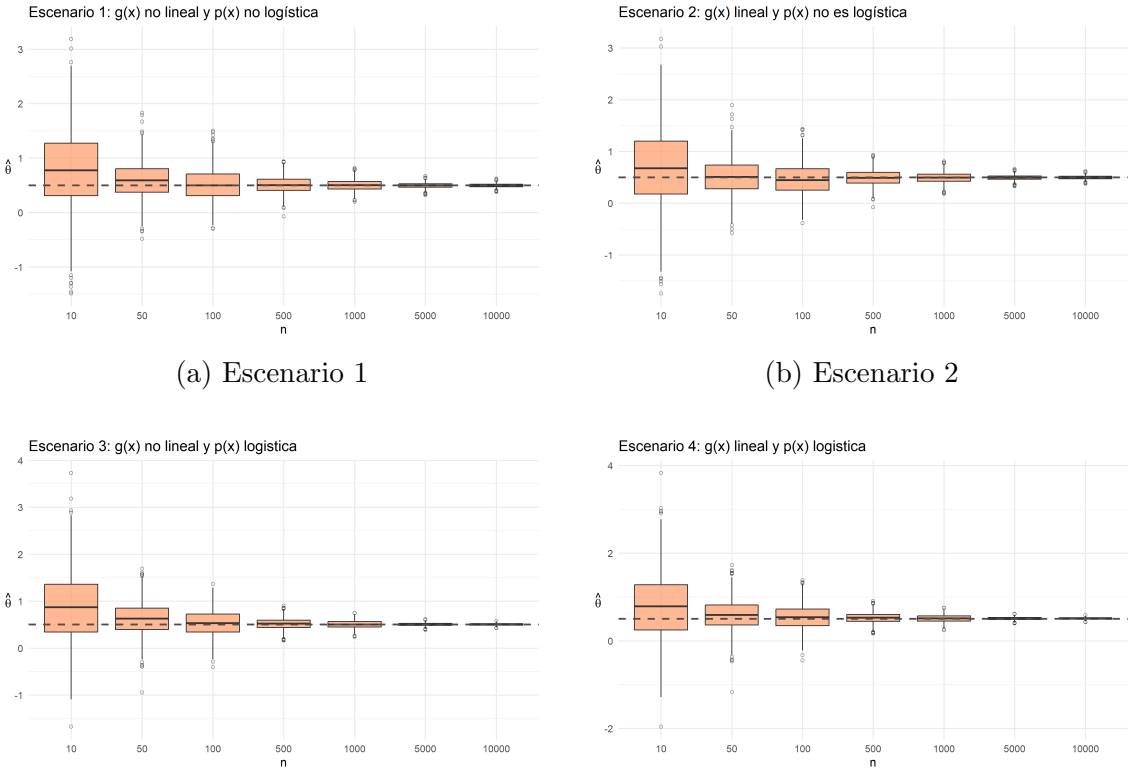
Cuadro 2: Estimaciones promedio de $\hat{\theta}$, error estándar (SE) y de la raíz del error cuadrático medio (RMSE) por método y escenario. Resultados reportados para $n =$ 10,000 observaciones y 1,000 simulaciones por escenario. El valor real es $\theta_0 = 0.5$.

Percentiles por escenario y tamaño muestral

Percentil	n=10	n=50	n=100	n=500	n=1,000	n=5,000	n=10,000
10 %	-0.131	0.186	0.159	0.326	0.361	0.436	0.453
20 %	0.192	0.323	0.268	0.387	0.413	0.457	0.469
30 %	0.417	0.414	0.357	0.426	0.449	0.472	0.479
40 %	0.599	0.511	0.427	0.466	0.477	0.484	0.488
50 %	0.777	0.591	0.500	0.503	0.501	0.497	0.497
60 %	0.961	0.684	0.582	0.540	0.530	0.510	0.504
70 %	1.142	0.756	0.660	0.588	0.557	0.523	0.513
80 %	1.407	0.854	0.768	0.637	0.589	0.537	0.526
90 %	1.736	1.008	0.887	0.702	0.635	0.562	0.539

Cuadro 3: Percentiles estimados por tamaño muestral. Escenario 1.

Distribuciones de $\hat{\theta}_0$ por tamaño muestral y escenario para DML



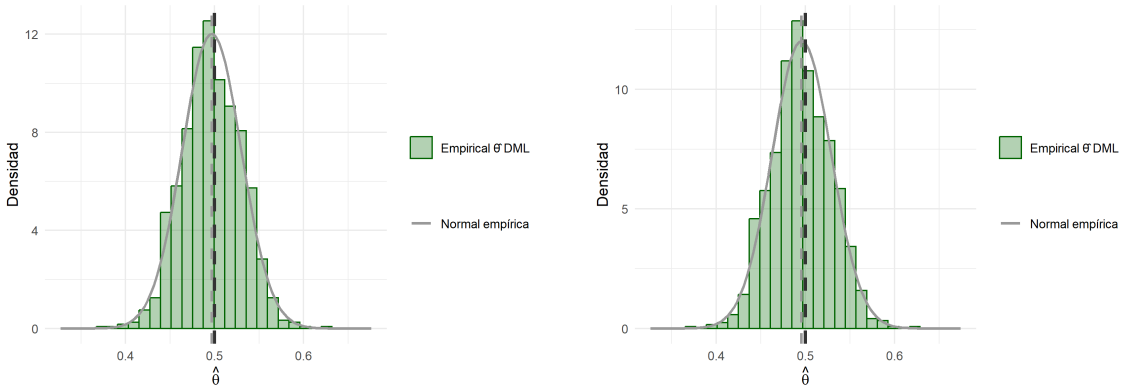
(a) Escenario 1

(b) Escenario 2

(c) Escenario 3

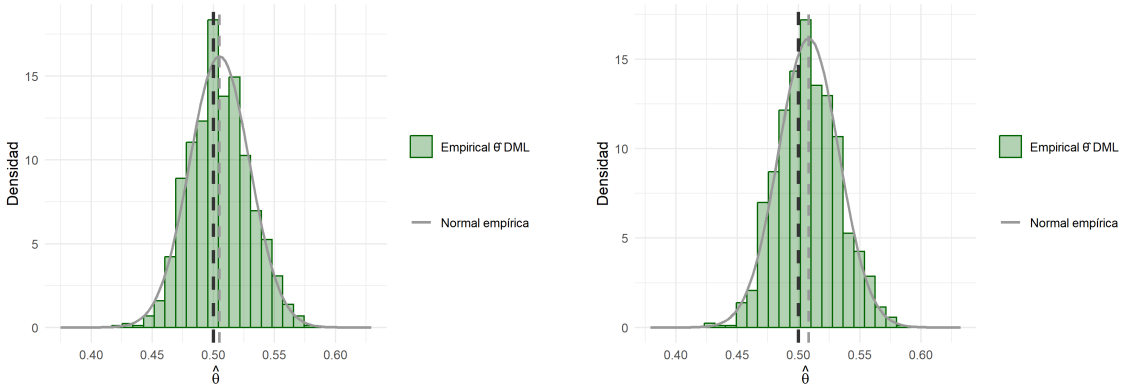
(d) Escenario 4

Figura 5: Distribución de estimaciones DML por tamaño muestral por escenario



(a) Escenario 1

(b) Escenario 2



(c) Escenario 3

(d) Escenario 4

Figura 6: Distribución empírica de las estimaciones $\hat{\theta}$ en cada escenario de simulación DML. La línea negra punteada indica el valor verdadero $\theta_0 = 0{,}5$, la gris indica la media empírica.

4. Conclusiones

Referencias

Athey, Susan (2018). «The impact of machine learning on economics». En: *The economics of artificial intelligence: An agenda*, págs. 507-547.

Bach, Philipp et al. (2022). «DoubleML – An Object-Oriented Implementation of Double Machine Learning in R». En: *Journal of Statistical Software* 107.4, págs. 1-62.

Bang, Heejung y James M. Robins (2005). «Doubly robust estimation in missing data and causal inference models». En: *Biometrics* 61.4, págs. 962-973.

Chernozhukov, Victor et al. (2018). «Double/debiased machine learning for treatment and structural parameters». En: *The Econometrics Journal* 21.1, págs. C1-C68.

Holland, Paul W. (1986). «Statistics and causal inference». En: *Journal of the American Statistical Association* 81.396, págs. 945-960.

Imbens, Guido W. y Donald B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Mullainathan, Sendhil y Jann Spiess (2017). «Machine learning: An applied econometric approach». En: *Journal of Economic Perspectives* 31.2, págs. 87-106.

Neyman, Jerzy (1990). «On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9». En: *Statistical Science* 5.4. Originally published in Polish in 1923; translated and with commentary by Dabrowska and Speed, págs. 465-472.

Rosenbaum, Paul R. y Donald B. Rubin (1983). «The central role of the propensity score in observational studies for causal effects». En: *Biometrika* 70.1, págs. 41-55.

Rubin, Donald B. (1974). «Estimating causal effects of treatments in randomized and nonrandomized studies». En: *Journal of Educational Psychology* 66.5, págs. 688-701.

Sant'Anna, Pedro H.C. y Jun B. Zhao (2020). «Doubly robust difference-in-differences estimators». En: *Journal of Econometrics*. Working Paper version May 2020.