UNIVERSIDAD DE MONTEVIDEO

UNIDAD DE MAESTRÍAS Y POSTGRADOS EN ECONOMÍA

Aplicaciones de Machine Learning en inferencia causal

Estudio comparativo con datos simulados

por

Rodrigo González e Ignacio González

Trabajo final de carrera presentado para optar al título de
Magíster en Ciencia de Datos

Directores de tesis
PhD Ana Balsa
PhD Federico Veneri

Montevideo, Uruguay
2025

UNIVERSIDAD DE MONTEVIDEO

Unidad de Maestrías y Postgrados en Economía

Rodrigo González e Ignacio González

APLICACIONES DE MACHINE LEARNING EN INFERENCIA CAUSAL
ESTUDIO COMPARATIVO CON DATOS SIMULADOS

Tutores: PhD Ana Balsa y PhD Federico Veneri

TFC presentado para aspirar al título de Magíster en Ciencia de Datos

Juicio del Tribunal:

Recomendación para su publicación en el Repositorio de la UM:

……

……

……

Presidente: ……
(Firma) (Aclaración)

Secretario: ……
(Firma) (Aclaración)

Vocal: …….
(Firma) (Aclaración)

Montevideo, 8 de diciembre de 2025

Descargo de responsabilidad:

El/Los autor/autores de este trabajo final de carrera declara(n) que es/son el/los único(s) responsable(s) de su contenido, y en particular de las opiniones expresadas en él, las que no necesariamente son compartidas por la Universidad de Montevideo; asimismo, declara(n) que no se infringe ningún derecho de terceros, ya sea de propiedad intelectual, industrial o cualquier otro. En consecuencia, es/son el/los único(s) responsable(s) y de manera exclusiva puede(n) asumir eventuales reclamaciones de terceros (personas físicas o jurídicas) que refieran a la autoría de la obra y a otros aspectos vinculados a ésta, incluido el reclamo por plagio.

Índice

1. Introducción **1**

2. Marco teórico y literatura existente **3**

2.1. Fundamentos de la inferencia causal 3

2.1.1. Resultados potenciales y el problema fundamental de la inferencia causal 4

2.1.2. Identificación del ATE 5

2.1.3. Métodos cuasiexperimentales y el supuesto de independencia condicional 6

2.2. Machine Learning e inferencia causal 7

2.2.1. Problemas de ML para causalidad 8

2.2.2. Aplicaciones válidas de ML al análisis causal 9

2.3. Trayectoria metodológica hasta el DML 9

2.3.1. Método de ajuste por regresión bajo el supuesto de independencia condicional 10

2.3.2. Métodos basados en puntajes de propensión y el estimador IPW . 11

2.3.3. Método de Doble Robustez 11

2.3.4. Double Debiased Machine Learning 12

3. Objetivos e hipótesis **14**

4. Metodología **16**

4.1. Proceso generativo de datos 16

4.1.1. Funciones auxiliares 17

4.1.2. Aplicación computacional del DGP 20

4.2. Estimación de modelos auxiliares 20

4.2.1. Outcome Regression 21

4.2.2. Propensity Score 21

4.3. Evaluación de modelos auxiliares 21

4.3.1. Convergencia y validez de los modelos auxiliares 23

4.4. Implementación métodos causales y evaluación 26

4.4.1. Estimador de Doble Robustez Paramétrico 26

4.4.2. Estimador DR plug in ML 26

4.4.3. DML 27

4.5. Métricas de comparación 28

4.6. Simulaciones 29

5. Resultados **30**
5.1. Resultados generales con muestras grandes 30
5.2. Resultados por tamaño muestral 31

6. Conclusiones **34**

7. Anexos **38**
7.1 Anexo I 38
7.2 Anexo II 40
7.3 Anexo III 43
7.4 Anexo IV 44
7.5 Anexo V 46

Resumen

Este trabajo tiene como objetivo examinar el uso de las técnicas de Machine Learning a la inferencia causal mediante un estudio comparativo basado en datos simulados. El objetivo central consiste en evaluar en que escenarios la integración de Machine Learning, en particular a través del método Double Debiased Machine Learning (DML) de Chernozhukov et al. (2018), mejora la estimación del efecto causal promedio frente a técnicas econométricas tradicionales. Para ello se comparan tres enfoques para estimar el efecto causal promedio: (i) un estimador de doble robustez paramétrico, (ii) una versión de doble robustez con Machine Learning aplicado de forma directa, sin ortogonalización ni sample splitting, y (iii) Double Debiased Machine Learning. En escenarios simulados, donde se conoce el efecto causal real, se utilizan funciones auxiliares tanto lineales como no lineales y distintos tamaños de muestra para comparar la precisión de cada método respecto al efecto causal verdadero.

Los resultados indican que, con un tamaño muestral amplio, Double Debiased Machine Learning brinda estimaciones consistentes al presentar estimaciones insesgadas y estables, aun cuando las funciones auxiliares son más complejas. Los resultados sugieren que supera consistentemente al estimador de doble robustez con aplicación directa de Machine Learning. La versión paramétrica del estimador de doble robustez solo es preferible en tanto al menos una de las funciones auxiliares se encuentre bien especificada. Por su parte, el estimador de doble robustez paramétrico resulta competitivo bajo correcta especificación, y se mantiene frente al DML como una alternativa más robusta cuando el tamaño muestral no es lo suficientemente grande para que los métodos de ML alcancen un nivel de precisión mínimo requerido. La aplicación directa de Machine Learning sin los ajustes propios del enfoque DML puede generar sesgos persistentes.

Los resultados de nuestro estudio confirman la importancia de la ortogonalización y el sample splitting para integrar Machine Learning en inferencia causal y destaca el potencial de DML como herramienta robusta para escenarios complejos cuando el tamaño de muestra es grande.

Palabras clave: Inferencia causal, Machine Learning, Double Debiased Machine Learning, Doble Robustez

Abstract

This study aims to examine the contribution of Machine Learning techniques to causal inference through a comparative analysis based on simulated data. The main objective is to evaluate the scenarios in which the integration of Machine Learning—particularly through the Double Debiased Machine Learning (DML) method of Chernozhukov et al. (2018)—improves the estimation of the average causal effect relative to traditional econometric techniques. To this end, three approaches to estimate the average causal effect are compared: (i) a parametric doubly robust estimator, (ii) a doubly robust version using Machine Learning applied directly, without orthogonalization or sample splitting, and (iii) Double Debiased Machine Learning. In simulated settings, where the true causal effect is known, both linear and nonlinear auxiliary functions and different sample sizes are used to assess each method's accuracy relative to the true effect.

The results indicate that, with a sufficiently large sample size, DML provides consistent estimates by yielding unbiased and stable results, even when auxiliary functions are more complex. The findings suggest that it consistently outperforms the doubly robust estimator with direct Machine Learning application. The parametric version of the doubly robust estimator is preferable only when at least one of the auxiliary functions is correctly specified. Under correct specification, the parametric estimator remains competitive and stands as a more robust alternative to DML when the sample size is not large enough for ML-based methods to achieve the required level of accuracy. Direct application of Machine Learning without the adjustments inherent to the DML framework may lead to persistent bias.

The results of our analysis confirm the importance of orthogonalization and sample splitting for integrating Machine Learning into causal inference, and highlight the potential of DML as a robust tool for complex scenarios when the sample size is large.

Keywords: Causal inference, Machine Learning, Double Debiased Machine Learning, Doubly Robust

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 ciertas variables de interés. En este escenario, el problema fundamental de la inferencia causal radica en que 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).

Al no poder observar directamente los efectos causales individuales, es necesario recurrir a diseños experimentales o cuasiexperimentales que permitan identificar los efectos promedios de la intervención. Los experimentos aleatorios controlados (RCT) constituyen el estándar de oro para inferir causalidad, ya que permiten controlar los sesgos derivados de la selección al tratamiento mediante la aleatorización. Sin embargo, la implementación de estos experimentos no siempre es posible, enfrentándonos los investigadores a estudios observacionales [1], debiendo recurrir a métodos cuasiexperimentales como alternativa. Los métodos cuasiexperimentales, ampliamente aplicados, buscan evaluar los 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 intermedios que no siempre pueden asumirse con seguridad. En las conclusiones de su artículo, Sant'Anna y Zhao (2020) indican lo siguiente:

[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.

«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]

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.

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. Este argumento está ampliamente desarrollado tanto por Athey (2018) como por Chernozhukov et al. (2018) en sus respectivos trabajos.

Se debe 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, 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).

Sin embargo, Athey (2018) destaca también cómo estas dos disciplinas, aun teniendo un distinto objetivo, pueden complementarse. Si bien 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 doble 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 como el estimador de doble robustez desarrollado originalmente por Bang y Robins (2005) puede aprovechar la flexibilidad predictiva de ML para complementar y mejorar las estimaciones causales en contextos específicos(Chernozhukov et al. 2018).

Para ello realizamos una revisión de los avances en la literatura. Luego, dado que

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

en la mayoría de los contextos empíricos el efecto causal verdadero es desconocido, se implementará un ejercicio de simulación sobre diferentes escenarios, basado en el análisis de Chernozhukov et al. (2018), 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).

El aporte central de este trabajo, respecto del ejercicio original de Chernozhukov et al. (2018) radica en incorporar una comparación entre tres enfoques: el método Double Debiased Machine Learning (DML), el estimador DR paramétrico tradicional y el estimador DR con aplicaciones de ML. Esta comparación se realiza en una variedad de escenarios que difieren tanto en la forma funcional subyacente del DGP como en el tamaño muestral. El objetivo es identificar con claridad en qué contextos cada método ofrece ventajas relativas, así como cuantificar la ganancia asociada a la incorporación de técnicas de ML dentro del marco DML frente a los métodos DR tradicionales. Esto permite derivar reglas prácticas de decisión sobre la selección metodológica adecuada según el escenario empírico considerado.

Los resultados del ejercicio muestran que, cuando el DR paramétrico está correctamente especificado en al menos una de las funciones subyacentes, el DML reproduce prácticamente el mismo resultado en términos de sesgo y variabilidad; pero cuando existe mala especificación, el DML evita el sesgo que afecta al DR tradicional. Al mismo tiempo, el DML domina de forma sistemática al estimador DR implementado con modelos de ML sin ortogonalización ni sample splitting, el cual tiende a sobreajustar y a generar mayor sesgo como se sugiere en Chernozhukov et al. (2018). Por estos motivos, el DML brinda la garantía de obtener estimaciones consistentes para efectos causales sin requerir una especificación previa de la relación funcional entre las variables, eliminando así el riesgo asociado a la mala especificación y asegurando mayor robustez en una amplia variedad de escenarios.

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. 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.

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

Introducción y notación básica

Es posible formalizar el concepto de resultados potenciales siguiendo la notación utilizada por Neyman (1990) y Rubin (1974).

Frente a la potencial exposición a un tratamiento dado, se postula que para cada individuo existe un par de resultados posibles: uno correspondiente a la situación de haber recibido el tratamiento y otro correspondiente a no recibirlo. Formalmente, para cada individuo $i = 1, \ldots, n$, observamos un único 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.

Para cada unidad i es posible definir los resultados potenciales:

- $Y_i(1)$: 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, siendo imposible observar el resultado que obtendría el individuo i en el estado que no se materializa.

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

Neyman (1990) y Rubin (1974) introducen el concepto de 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, el 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.

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 el efecto individual $\tau_i = Y_i(1) - Y_i(0)$, se puede estimar consistentemente el efecto promedio $\tau = \mathbb{E}[\tau_i]$ 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 como 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 cuasiexperimentales y el supuesto de independencia condicional

En escenarios donde no es posible implementar un RCT, es posible recurrir a los llamados métodos cuasiexperimentales. Estos métodos 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 es 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.

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) formalizan 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 bien el

[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.

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.

[4]

Así, bajo ignorabilidad condicional, el ATE puede identificarse en base a la siguiente formulación: [5]

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

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, Athey e Imbens (2019) explican explican que los métodos de aprendizaje automático han tenido una adopción relativamente lenta en la economía en comparación con otras disciplinas debido a las diferencias fundamentales en los objetivos y criterios que prioriza cada comunidad.

Aun frente a un mismo problema como la estimación de una media condicional $E(Y/X = x)$ en un problema supervisado, el enfoque de ML es distinto al de la econometría causal. Mientras que la econometría tiene su énfasis en la inferencia estadística y

[4] Cuando se enfrentan situaciones con muchas covariables potencialmente relevantes es de extrema importancia la selección adecuada de las variables de control. Una de las alternativas existentes para facilitar la ésto es el propuesto por Belloni, Chernozhukov y Hansen (2014), quienes 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.

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

su foco en la estimación de efectos causales, ML no asume ninguna estructura paramétrica y se centra en obtener predicciones precisas fuera de muestra.

Al utilizar 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 minimizar 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.

Como indican los autores, 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 buscar disminuir la dimensionalidad del modelo mediante estos métodos, se puede llegar a seleccionar modelos más simples pero incorrectos, omitiendo variables relevantes e introduciendo sesgos.

Si bien la regularización también puede estar presente en métodos paramétricos, su impacto puede ser más controlable debido a la estructura impuesta. En cambio, en modelos no paramétricos, la regularización afecta la selección de variables de forma menos transparente, lo que puede llevar a sesgos más difíciles de detectar.

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, 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 DML, en el que tanto el outcome regression model 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. Trayectoria metodológica hasta el DML

Este trabajo se centrará en el enfoque de DML dada su naturaleza que combina técnicas de ML con inferencia causal para obtener estimadores insesgados y consistentes en presencia de relaciones complejas entre las variables.

Dado que este método es el resultado de una evolución metodológica que parte de herramientas paramétricas clásicas y que fue incorporando avances para superar sus limitaciones, resulta útil repasar brevemente la evolución de los métodos que lo precedieron para comprender su origen.

En primer lugar, se presentarán los métodos de ajuste por regresión y los basados en puntajes de propensión, que constituyen los enfoques tradicionales para comparar grupos tratados y de control. Luego, se mostrará cómo ambos se combinan en el estimador de DR, el cual mejora la consistencia al combinar los modelos anteriores. Finalmente, se

introducirá el DML, el cual extiende la lógica del DR mediante aplicaciones de ML, ortogonalización y sample-splitting.

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

En escenarios no experimentales, una forma de abordar la identificación del efecto causal consiste en ajustar por las covariables relevantes para hacer comparables a las unidades tratadas y de control. El método de ajuste por regresión, cuando se cumple el supuesto de independencia condicional, permite 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:

$$\tau = \mathbb{E}\left[\mu_{(1)}(X_i) - \mu_{(0)}(X_i)\right],^{6}$$

Esta expresión fue introducida inicialmente por Rubin (1974) y desarrollada con más detalle en el trabajo conjunto con Imbens y Rubin (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.

En estudios observacionales, este método nos permite estimar correctamente el ATE siempre y cuando sean utilizados modelos adecuados para capturar la relación entre las covariables y la variable de resultado. Esto marca la importancia que tiene el hecho de

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

lograr una buena capacidad predictiva para capturar correctamente el efecto del tratamiento, así como alerta de los potenciales problemas que generaría 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 (Imbens y Rubin 2015).

2.3.2. 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 estimador, desarrollado por Horvitz y Thompson (1952) y adaptado por Robins y Greenland (1986) para inferencia causal, 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.

2.3.3. Método de Doble Robustez

Los métodos descritos previamente, el estimador basado en el outcome regression model (OR) y el estimador por 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 puntajes 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, Bang y Robins (2005) desarrollaron el estimador de doble robustez (DR). 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).

2.3.4. 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.

Siguiendo en el desarrollo de Chernozhukov et al. (2018), 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 en sentido de Neyman 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. Objetivos e hipótesis

Como señalamos previamente, Chernozhukov et al. (2018) propone el uso del DML para la estimación causal en la presencia de alta dimensionalidad de covariables o de relaciones funcionales complejas. En concreto, muestran que es posible utilizar Machine Learning en contextos de alta dimensionalidad o cuando la forma funcional es compleja, sin afectar la inferencia causal, siempre que se aplique ortogonalización y cross-fitting.

El objetivo de este trabajo es estudiar el comportamiento del método de DML de Chernozhukov et al. (2018) mediante simulaciones controladas, comparando el desempeño del DML frente a alternativas más tradicionales de inferencia causal.

Para esto, generamos datos simulados bajo el modelo parcialmente lineal de Chernozhukov et al. (2018) y diferentes estructuras funcionales, lineales y no lineales, de la función de resultado y del índice del propensity score, y luego comparamos el desempeño empírico de tres estimadores:

1. Doble robustez paramétrico (Bang y Robins 2005)
2. Doble robustez con Machine Learning plug-in (sin sample-splitting)
3. Double Debiased Machine Learning (con sample-splitting y scores ortogonales). (Chernozhukov et al. 2018)

El desempeño de estos estimadores es comparado en términos de sesgo, raíz del error cuadrático medio (RMSE) y error estándar, bajo distintos tamaños muestrales y distintas estructuras funcionales.

El objetivo de la comparación es poder indicar en qué condiciones, en cuanto a formas funcionales y tamaño muestral, conviene utilizar cada uno de los estimadores. En particular, nos interesa contrastar el desempeño del DR paramétrico frente al DML, para determinar bajo que condiciones puede ser beneficioso la incorporación de ML al análisis causal así como su diferencia respecto al DR plug-in con ML (sin sample-splitting).

Antes de pasar a la comparativa del ejercicio simulado es importante anticipar cuales son los resultados que esperaríamos siguiendo la naturaleza y la teoría detrás de cada estimador.

En primer lugar, el tamaño muestral juega un rol central en lo que esperamos observar en el ejercicio de simulación. La teoría indica que los métodos basados en ML requieren muestras relativamente grandes para estimar de forma estable las funciones auxiliares. Por este motivo, anticipamos que el DML mostrará una mejora progresiva en su desempeño a medida que aumenta n. Con tamaños muestrales pequeños es esperable que las funciones auxiliares estimadas mediante ML presenten alta variabilidad y no alcancen el nivel de precisión necesaria, lo que podría generar un sesgo apreciable tanto en el DML como en el DR plug-in con ML. Dada la estructura ortogonal del DML (Chernozhukov et al. 2018), esperamos que este método converja al valor verdadero del ATE a mayor velocidad que el DR plug-in con ML, alcanzando estimaciones insesgadas con muestras más pequeñas que las requeridas por el plug-in tradicional.

En contraste, el DR paramétrico, siguiendo a Bang y Robins (2005), debería mantenerse insesgado en la estimación del ATE siempre que al menos una de las dos funciones auxiliares esté correctamente especificada, incluso en tamaños muestrales reducidos. En tal caso, el tamaño muestral afectaría sobre todo la varianza, pero no el sesgo. Por lo tanto, bajo escenarios de correcta especificación, esperamos que el DR paramétrico permanezca centrado en el valor verdadero de θ para valores pequeños y medianos de n, a diferencia de los métodos basados en ML, cuyos errores deberían ser mayores en muestras reducidas y converger al valor verdadero únicamente cuando n aumenta.

Un segundo punto central es cómo deberían comportarse los estimadores bajo distintos escenarios de especificación funcional. La teoría establece que el DR paramétrico es consistente siempre que al menos uno de los dos modelos auxiliares (OR o PS) esté correctamente especificado (Bang y Robins 2005). Por lo tanto, esperamos observar un buen desempeño del DR paramétrico en los escenarios donde una de las funciones coincide con la forma funcional del DGP, incluso cuando la otra esté mal especificada. En un escenario en donde ambas funciones auxiliares esten mal especificadas se espera que la

estimación de ATE muestre un sesgo significativo.

En cuanto al DML, al no requerir especificar una forma funcional particular y al utilizar ortogonalización y sample-splitting, se espera que mantenga un desempeño sólido incluso en escenarios con funciones verdaderas no lineales o complejas, donde los modelos paramétricos fallan. Bajo funciones lineales correctamente especificadas, el DML debería ofrecer un rendimiento comparable al DR paramétrico, aunque posiblemente con algo más de variabilidad en muestras pequeñas. Finalmente, en comparación con el DR plug-in con ML, anticipamos que el DML muestre menor sesgo y menor variabilidad, dado que el plug-in sin sample-splitting es más propenso a caer en overfitting de las funciones auxiliares.

Buscamos así contrastar el desempeño relativo de los métodos en cada escenario, identificar si la evidencia práctica respalda las predicciones teóricas y, determinar si con la evidencia del ejercicio es posible observar puntos de corte que orienten la elección entre alternativas para una mejor estimación del ATE.

4. Metodología

4.1. Proceso generativo de datos

Siguiendo Chernozhukov et al. (2018) construimos el ejercicio sobre un modelo parcialmente lineal:

$$Y_i = \theta_0 D_i + g_0(X_i) + \varepsilon_i, \qquad P_i = m_0(X_i) + u_i, \qquad D_i = \mathbf{1}\{P_i > q\}\,.,$$

donde Y_i es la variable de resultado, D_i indica el estado de tratamiento, siendo $D_i = 1$ el individuo tratado y $D_i = 0$ el control que no recibió el tratamiento . X_i es el vector de covariables, $g_0(X_i)$ es la función de resultado condicional que establece la verdadera relación subyacente entre la variable de resultado y las covariables mientras que $P_i = \Pr(D_i = 1 \mid X_i) = m_0(X_i) \in (0, 1)$ es la probabilidad de tratamiento (propensity score). Los errores ε_i son independientes e idénticamente distribuidos con $\varepsilon_i \sim \mathcal{N}(0, 1)$. Fijamos $\theta_0 = 0{,}5$. Dado que el valor del parámetro es conocido, se puede calcular RMSE y sesgo (ver definiciones en sección 4.5).

Las covariables se generan como $X \sim \mathcal{N}(0, \Sigma)$ de dimensión d con

$$\Sigma_{jk} = 0.7^{|j-k|}$$

manteniendo así la estructura en la relación entre variables desarrollada por los autores en los que nos basamos.

Para asegurar una muestra balanceada en tamaño, el umbral de asignación se define como $q = \text{mediana}(P)$, de modo que aproximadamente el $50\,\%$ de las unidades queden tratadas y el $50\,\%$ en control.

4.1.1. Funciones auxiliares

Para definir diferentes escenarios en donde sea posible el contraste de cada método bajo diferentes especificaciones de ambos modelos auxiliares, propensity score y outcome regression model, definimos dos especificaciones para cada una de ellas.

Outcome regression model: $g_0(X)$ Para la función de resultados condicional, se utilizan dos especificaciones que difieren en su complejidad funcional: una primera especificación en donde la relación de la variable de resultado con las covariables es lineal y otra en la que es no lineal con un término cuadrático en una de las covariables.

- **Lineal:**

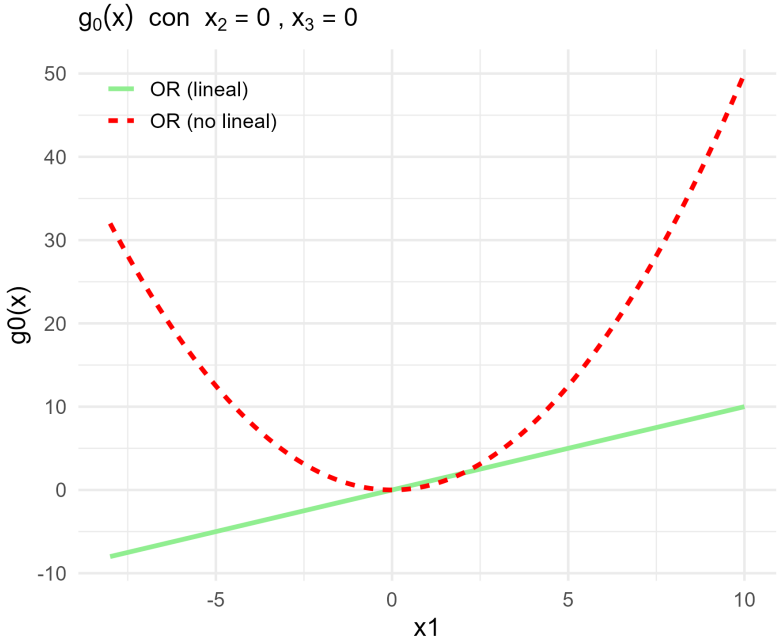
$$g_0(X) = X_1 + X_2 + X_3$$

- **No lineal:**

$$g_0(X) = 0.5\, X_1^2 + X_2 + X_3$$

Estas dos formas funcionales permiten representar distintos tipos de relaciones entre las variables explicativas y el resultado. En el caso lineal, se asume que cada covariable tiene un efecto constante sobre el resultado, sin importar su valor. En cambio, la forma no lineal permite que el impacto de una covariable varíe dependiendo de su nivel. La comparativa de estos casos es de importancia dado que en muchos contextos no sabemos ex ante cómo se relacionan las covariables con el resultado, y usar ambas especificaciones permite evaluar cómo se comportan los estimadores cuando la función está bien o mal especificada. La Figura 1 muestra estas dos funciones referidas.

Figura 1: Especificación lineal vs no lineal de OR



Propensity score: $m_0(X)$ Para la función de asignación al tratamiento se utilizó una función logística clásica de la forma:

$$m_0(X) = \frac{e^{z(X)}}{1 + e^{z(X)}}$$

donde $z(X)$ representa el índice latente de la función. Se consideran dos variantes funcionales para éste, determinando la relación :

- **Índice lineal:**

$$z(X) = X_1 + X_2 + X_3 + u$$

- **Índice no lineal:**

$$z(X) = \sin(X_1 + X_2 + X_3 + u)$$

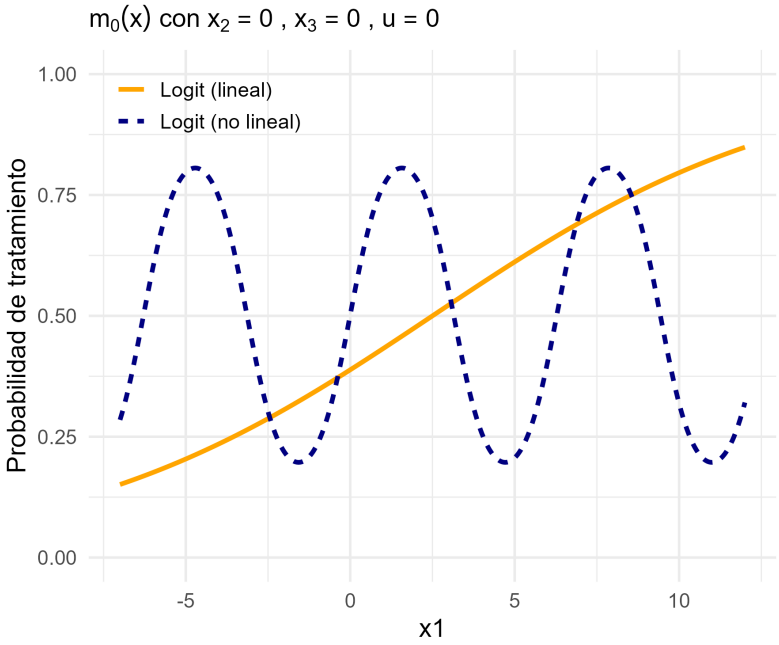
En ambos casos, $u \sim \mathcal{N}(0; 0.4^2)$ es un término de error que evita una asignación determinista al tratamiento. Además, se estandariza el índice utilizando la media y la desviación estándar de la suma $x_1 + x_2 + x_3$, con el fin de controlar la escala del índice logístico.

Esta estandarización resulta útil para disminuir el riesgo de que la probabilidad generada se vea afectada por valores extremos del índice, evitando valores cercanos a 0 o 1, así como también contribuye a mantener una asignación balanceada al momento de establecer el punto de corte de la probabilidad.

La incorporación de estas dos especificaciones funcionales nos permite modelar dife-

rentes situaciones. Bajo el índice lineal, la asignación al tratamiento es monótona en las covariables, lo que significa que a medida que aumentan los valores de X también la probabilidad de tratamiento se altera de forma continua y en una única dirección. En cambio, el índice no lineal que incorpora una estructura sinusoidal permite representar relaciones no monótonas, en las cuales la probabilidad de tratamiento varía en diferentes direcciones, pudiendo presentar aumentos y disminuciones alternados según los valores que toman las covariables (ver Figura 2). Esta distinción resulta útil para modelar escenarios en los que la relación entre las variables explicativas y la probabilidad de recibir tratamiento podría depender de ciertos tramos o rangos de los valores de X, y no necesariamente seguir una tendencia global estricta.

Figura 2: Especificaciones del propensity score



Escenarios de evaluación

Al combinar las dos formas para $g_0(X)$ (lineal o no lineal) y para $m_0(X)$ (logística con índice z_0 lineal o no lineal), obtenemos cuatro escenarios potenciales:

- **Escenario 1**: g_0 no lineal, m_0 logística con z_0 no lineal.
- **Escenario 2**: g_0 lineal, m_0 logística con z_0 no lineal.
- **Escenario 3**: g_0 no lineal, m_0 logística con z_0 lineal.
- **Escenario 4**: g_0 lineal, m_0 logística con z_0 lineal.

4.1.2. Aplicación computacional del DGP

El DGP y el resto del ejercicio se implementan en `R` utilizando funciones base y los paquetes `stats` y `MASS` para la simulación de las covariables.

Para cada escenario $(1, 2, 3, 4)$, se llevan a cabo los siguientes pasos para la simulación:

Algoritmo 1 - Generación de datos por escenario

1: Definir los parámetros necesarios: tamaño muestral n, la cantidad de covariables d, θ_0 y la matriz de covarianzas Σ.
2: Simular $X \sim \mathcal{N}(0, \Sigma)$ de dimensión $n \times d$.
3: Elegir $g_0(\cdot)$ y $m_0(\cdot)$ según el escenario elegido.
4: Generar $u \sim \mathcal{N}(0, 0{,}4^2)$
5: Calcular $P = m_0(X) + u$
6: Fijar $q = \text{mediana}(P)$ y asignar $D = \mathbf{1}\{P > q\}$.
7: Generar $\varepsilon \sim \mathcal{N}(0, 1)$ y computar $Y = \theta_0 D + g_0(X) + \varepsilon$.
8: Retornar (Y, D, X).

4.2. Estimación de modelos auxiliares

Si bien para la estimación del ATE mediante el DR paramétrico no necesitamos realizar una estimación de las funciones auxiliares al estar asumiendo una forma lineal para el OR y una logística con índice lineal para el propensity score, la estimación de los estimadores DR plug-in con ML y la del DML sí requiere una etapa previa de entrenamiento de modelos auxiliares.

Esto toma especial relevancia en los métodos de ML integrado dado que, como se comentó en la subsección 2.3.4, los métodos como DML operan correctamente si los modelos auxiliares no presentan grandes tasas de error. La propiedad de ortogonalidad en sentido de Neyman nos asegura que, si los estimadores son suficientemente buenos, potenciales pequeños errores en la estimación de las funciones auxiliares no se transmitirán a la estimación del ATE, logrando una inferencia válida. A continuación se detalla el proceso de estimación de los modelos auxiliares y su evaluación en muestras de testeo para verificar que las estimaciones intermedias cumplan la condición de ser lo suficientemente buenas.

Para esta estimación se opta por esquemas de boosting, con tuning interno y evaluación fuera de muestra mediante validación cruzada. La elección responde a su capacidad para capturar relaciones no lineales e interacciones con alta potencia predictiva, manteniendo control sobre el sobreajuste a través de tasas de aprendizaje, profundidad de interacción y early stopping.

4.2.1. Outcome Regression

El objetivo del Outcome Regression es aproximar $g(X) = \mathbb{E}[Y|X]$ con bajo error fuera de muestra y estabilidad. Para su estimación se considera un conjunto de modelos en caret y se selecciona el mejor según RMSE promedio de validación cruzada. En los escenarios analizados, los modelos de boosting resultaron competitivos frente a alternativas como random forest o SVM, ofreciendo mejor resultado en términos de sesgo-varianza.

Entrenamiento y tuning de hiperparámetros

Una vez definida la familia de modelos, el entrenamiento se realiza con validación cruzada repetida en 5 folds. Cabe destacar que, al repetir el proceso completo de validación cruzada dos veces con particiones aleatorias diferentes, se obtiene una estimación más estable y robusta del error fuera de muestra, lo que reduce la sensibilidad del modelo a una única partición específica de los datos y ayuda a mitigar el riesgo de sobreajuste, disminuyendo la variabilidad de las métricas de desempeño de nuestro modelo. La métrica objetivo utilizada en el entrenamiento es el RMSE.

4.2.2. Propensity Score

Para aproximar $m(X) = \mathbb{P}(D = 1|X)$ se utiliza un modelo de boosting para clasificación. En particular, se utiliza el método de Extreme Gradient Boosting (xgboost), desarrollado por Chen y Guestrin (2016), con validación cruzada repetida con 5 folds, optimizando la pérdida logística y monitoreando el área bajo la curva de ROC (AUC). Así como en el caso del OR, la evaluación se realiza en una muestra independiente, generada mediante el DGP con una semilla aleatoria distinta, reportando AUC y accuracy al punto de corte 0,5.

4.3. Evaluación de modelos auxiliares

Como se mencionó en la Sección 2.3.4, para obtener estimaciones insesgadas del ATE mediante DML, los modelos de ML utilizados para aproximar las funciones auxiliares deben alcanzar un nivel de precisión suficiente para que los errores de estimación no comprometan la validez del estimador final. Es importante destacar que, en esta sección, no se presentan los resultados finales del análisis causal. Los resultados aquí reportados corresponden únicamente a etapas intermedias, cuyo propósito es evaluar la validez de las aproximaciones obtenidas y determinar en qué medida los modelos empleados resultan adecuados para las funciones auxiliares.

Cuadro 1: Evaluación Outcome Regression

Estadístico	**Modelo lineal**	**Modelo no lineal**
R^2	0.870	0.841
NRMSE	0.362	0.399

Resultados obtenidos en una muestra de $n = 10{,}000$

Como se puede observar en Cuadro 1, tanto el modelo lineal como el no lineal muestran una elevada bondad de ajuste. El modelo lineal presentó un valor de R^2 de 87 %, mientras que el no lineal presenta una bondad de ajuste de 84.1 %, logrando así en ambos modelos explicar un alto porcentaje de la variabilidad total de la variable de resultado.

El NRMSE (definido como $\frac{RMSE}{sd(y)}$) es una medida adimensional que expresa el RMSE en relación con la variabilidad de de la variable de respuesta (y). Bajo esta definición, que el NRMSE sea igual a 1, implica que el modelo es tan bueno como usar un predictor basado en la media.En caso de que presente valores mayores a 1 esto es indicador de un peor desempeño que el predictor basado en la media, mientras que si presenta un valor menor a 1, indica un mejor desempeño en términos de error de predicción. En el escenario analizado, el modelo de OR lineal presenta un NRMSE de 0.362 y el modelo no lineal un NRMSE de 0.399. Ambos valores se encuentran significativamente por debajo del benchmark de la media, lo que sugieren desempeños predictivos sólidos, alineándose con interpretaciones donde valores inferiores indican una capacidad realativa para predecir con precisión.

Cuadro 2: Evaluación Modelo Propensity Score

Estadístico	**Índice lineal**	**Índice no lineal**
AUC	0.994	0.993
Accuracy (cutoff 0.5)	0.965	0.962

Resultados obtenidos en una muestra de $n = 10{,}000$

En el cuadro 2 podemos ver las métricas de evaluación de los modelos de propensity score. En cuanto al poder de discriminar entre clases que tienen los modelos, podemos ver en base al AUC que logran una discriminación casi perfecta entre las clases al presentar valores muy cercanos a 1. Esto significa que los modelos son capaces de asignar un valor de probabilidad alta a aquellos casos donde $T = 1$ y le asigna una probabilidad de ser tratado baja a aquellos casos de controles en el $99\,\%$ de las veces.

Por su parte, en cuanto a la precisión general, el modelo entrenado en base al logit no lineal alcanza una accuracy de 96.2 % mientras que el modelo entrenado en base al logit lineal alcanza una accuracy del 96.5 %. Estos valores muestran una gran precisión

de los modelos auxiliares en la base de testeo, logrando evitar tanto el overfitting como el underfitting.

Podemos observar en base a los resultados anteriores que los modelos muestran un alto grande de ajuste y que logran estimaciones muy precisas, tanto del outcome regression model como el propensity score. Sin embargo, ¿esto es suficiente para poder estar seguros que los modelos pueden usarse como pasos intermedios en la estimación causal sin sesgar las estimaciones? La respuesta a esta pregunta se desarrolla a continuación.

4.3.1. Convergencia y validez de los modelos auxiliares

Como se discute en Chernozhukov et al. (2018), el desempeño de los modelos de Outcome Regression y Propensity Score debe ser suficientemente bueno para que los errores de predicción en estas funciones no se propaguen al estimador final del ATE. El autor indica que esto no se refiere a un valor de precisión puntual de los modelos sino que, a la velocidad con la que los modelos auxiliares convergen al valor real a estimar cuando aumenta el tamaño muestral.

En términos formales, siguiendo el análisis de Chernozhukov, el requisito clave para la consistencia del estimador DML es que se cumpla la siguiente condición de convergencia:

$$\|\hat{g} - g_0\|_{P,2} \times \|\hat{m} - m_0\|_{P,2} = o_P(n^{-1/2})$$

donde g_0 y m_0 representan las verdaderas funciones del modelo estructural y $\hat{g}$ y $\hat{m}$ sus estimaciones. Esta condición implica que el producto de los errores de aproximación de los modelos auxiliares debe converger a cero más rápido que $n^{-1/2}$. Intuitivamente, si en conjunto el error absoluto de los modelos frente a las funciones reales decrece con n a una tasa mayor o igual de $n^{-1/2}$, los modelos auxiliares son lo suficientemente buenos para asegurarnos de poder estimar el ATE de forma insesgada mediante DML sin introducirle sesgo relativos a errores en las funciones de OR o PS. El autor advierte que, en funciones que no sean ortogonales frente a los errores de las funciones auxiliares, el requerimiento de convergencia individual es más estricto, siendo que cada modelo auxiliar debería converger a un ritmo más rápido que $n^{-1/4}$ para garantizar consistencia asintótica. Esto significa que, sin ortogonalidad, la estimación del ATE se vuelve mucho más sensible a la calidad de los modelos de Outcome Regression y Propensity Score.

Podemos comparar esto de manera empírica, más aun conociendo en este ejercicio las verdaderas formas de g0 y m0 al ser un ejercicio simulado, al comparar el producto empírico $\|\hat{g} - g_0\|_2 \times \|\hat{m} - m_0\|_2$ frente al umbral teórico $n^{-1/2}$, evaluado en distintas muestras o tamaños de simulación. Cuando este producto disminuye más rápido que $n^{-1/2}$ al incrementar n, se considera que los modelos auxiliares cumplen con la condición de

convergencia.

Podemos representar la tasa de convergencia empírica α [7] mediante la siguiente relación logarítmica:

$$\alpha = \frac{\log(E(L_1)) - \log(E(L_2))}{\log(L_2) - \log(L_1)}$$

donde $E(L_1)$ y $E(L_2)$ representan los productos empíricos de las normas $\|\hat{g} - g_0\|_2 \times \|\hat{m} - m_0\|_2$ obtenidos para dos tamaños muestrales distintos, L_1 y L_2. En este contexto, los valores de L_1 y L_2 corresponden directamente a los tamaños de muestra n_1 y n_2, con $n_2 > n_1$. Por ejemplo, en el ejercicio planteado se consideró $n_1 = L_1 = 1{,}000$ y $n_2 = L_2 = 10{,}000$.

Una tasa de $\alpha = 0.5$ indica convergencia al ritmo teórico mínimo exigido para consistencia $n^{-1/2}$, siendo que valores superiores a 0.5 reflejan una convergencia más rápida (mayor precisión en las funciones auxiliares), mientras que valores menores sugieren una convergencia más lenta y potencial necesidad de mayor tamaño muestral o mejor especificación de los modelos.

Cuadro 3: Tasa empírica de convergencia por escenario

Caso	α **empírico**
Referencia: $n^{-1/2}$	0.500
Escenario 1: OR no lineal $\times$ índice PS lineal	0.317
Escenario 2: OR lineal $\times$ índice PS no lineal	0.266
Escenario 3: OR no lineal $\times$ índice PS lineal	0.242
Escenario 4: OR lineal $\times$ índice PS lineal	0.227

Estimación de tasa de convergencia empírica de norma L2 general para el incremento del tamaño muestral de n=1,000 a n=10,000.

Como se puede ver en el cuadro 3, las tasas de convergencia empírica al pasar de una muestra de $n = 1{,}000$ a $n = 10{,}000$ no alcanzan el valor de referencia asintótico de 0.5. Si bien se observa una mejora en el ajuste de las funciones auxiliares g y m, dicha mejora ocurre a un ritmo que es menor al que sugiere la teoría. Sin embargo, es necesario tener en cuenta que la tasa utilizada como benchmark, es una tasa estrictamente asintótica y sería alcanzada en el escenario donde el límite de n tiende a infinito. En ese sentido, al tratarse de un análisis realizado en un rango finito de tamaños muestrales, aunque resulta relevante en términos empíricos, no garantiza la velocidad de convergencia asintótica del producto de normas L en cada uno de los escenarios.

[7] El parámetro α representa la tasa de convergencia con la cual $\|\hat{g} - g_0\|_2 \times \|\hat{m} - m_0\|_2$ decrece a medida que aumenta el tamaño muestral. Esta formulación se desarrolla en LeVeque (2007).

Si bien se trata de un análisis estrictamente finito muestral y, en consecuencia, no permite verificar de forma directa los resultados asintóticos, la evidencia gráfica del Anexo IV contribuye a evaluar la velocidad con que los errores de los modelos estimados se reducen conforme crece el tamaño muestral. En la Figura 12 se muestra para los cuatro escenarios simulados como evoluciona el producto de las normas $\|\hat{g} - g_0\| \cdot \|\hat{m} - m_0\|$ en relación a la caída de $n^{-1/2}$, lo que determina el nivel de disminución necesaria según la teoría asintótica. Se puede observar en estos gráficos que la pendiente empírica del producto de las normas de los errores presenta una pendiente similar, o incluso más pronunciada en algunos tramos que la del valor de referencia teórico. Esto sugiere que, al menos de un modo visual, el producto del error total de los modelos auxiliares converge a una velocidad razonablemente cercana al benchmark teórico.

En el análisis de métodos basados en scores no ortogonalizados, los cuales requieren una condición de convergencia más estricta[8], la realidad podría ser diferente. Mientras que la norma de $\hat{g}$ muestra un decrecimiento adecuado, con pendientes iguales o mayores que el benchmark teórico conforme varía el tamaño muestral, la norma de $\hat{m}$ no presenta el mismo comportamiento, especialmente en los escenarios 3 y 4 en donde su velocidad de convergencia es sustancialmente menor. Esto se explica ya que, en dichos escenarios, la norma parte desde un valor relativamente bajo incluso en muestras pequeñas, por lo que ve restringida su capacidad de mejora con los aumentos del tamaño muestral.

El ejercicio finito muestral nos brinda una aproximación útil para evaluar en qué medida cada estimador logra capturar la estructura subyacente del DGP, y poder así anticipar cómo los errores generados en las etapas intermedias podrían trasladarse a la estimación final del ATE. En base a esto y a los resultados observados, es razonable esperar que DML mantenga un buen desempeño en muestras grandes dentro de los escenarios presentados. La ortogonalización atenúa la sensibilidad frente a errores en cada modelo auxiliar, y esto se refleja en que, aun con convergencias individuales moderadas, el comportamiento conjunto de ambos modelos mejoraría a una tasa lo suficientemente veloz para que las estimaciones de ATE sean robustas a potenciales errores pequeños en las funciones auxiliares.

Por otra parte, para el DR plug-in con ML sin score ortogonalizado, el hecho de que la condición de convergencia individual no se verifique, al menos desde la evidencia visual presentada, indica que estos métodos podrían presentar sesgos más pronunciados al no alcanzar la velocidad de convergencia requerida por la teoría. Esto refuerza la expectativa de que los métodos sin ortogonalización sean más vulnerables a la propagación de errores de modelación hacia la estimación del ATE.

[8]Según Chernozhukov et al. (2018), estos métodos exigen que las normas L individuales de g y m decrezcan a una tasa de al menos $n^{-1/4}$ y no solo que su producto lo haga a $n^{-1/2}$

4.4. Implementación métodos causales y evaluación

Habiendo presentado la metodología general y los criterios teóricos que guían la evaluación del desempeño de las funciones auxiliares en cada escenario del DGP, corresponde ahora describir con precisión cómo se implementan los distintos métodos causales y la definición de las métricas con las cuales se analizarán los resultados.

4.4.1. Estimador de Doble Robustez Paramétrico

Para obtener el estimador DR paramétrico asumimos, en todos los casos, una especificación lineal para el OR y una logística con índice lineal para el PS. En consecuencia, en el escenario 1, donde la forma correcta del OR es no lineal con un término cuadrático en X_1 y, al mismo tiempo, el PS verdadero es logístico con índice no lineal que induce una relación no monótona entre $P(D = 1)$ y las covariables, es de esperar que la estimación DR resulte sesgada. En los demás escenarios, en cambio, al menos una de las dos funciones está correctamente especificada, por lo que no anticipamos sesgo. Para la implementación del estimador DR paramétrico se siguieron los siguientes pasos:

Algoritmo 2 Implementación del estimador de Doble Robustez Paramétrico

1: Estimar el outcome regression $Y \sim X$ mediante regresión lineal por separado para los tratados y los controles.

2: Obtener las predicciones $\hat{\mu}_1(X_i)$ y $\hat{\mu}_0(X_i)$ para toda la muestra usando los modelos anteriores.

3: Estimar el modelo de asignación al tratamiento $D \sim X$ mediante regresión logística (logit) en base a un índice lineal en X.

4: Obtener el propensity score estimado $\hat{p}(X_i)$ para cada observación.

5: Calcular la estimación individual de doble robustez para cada i:

$$\text{DR}_i = \frac{D_i \left(Y_i - \hat{\mu}_1(X_i)\right)}{\hat{p}(X_i)} - \frac{(1 - D_i) \left(Y_i - \hat{\mu}_0(X_i)\right)}{1 - \hat{p}(X_i)} + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

6: Estimar el ATE como el promedio muestral de DR_i.

7: Calcular el desvío estándar y el error estándar muestral.

Cabe aclarar que este procedimiento se implementa utilizando las funcionalidades base de R.

4.4.2. Estimador DR plug in ML

La estimación de este método sigue la misma lógica del caso anterior, pero sin suponer las formas funcionales de los modelos de resultados y el propensity score para estimarlos

utilizando un método de ML, en este caso, mantenemos la configuración óptima de los modelos presentadas anteriormente.

A diferencia del DML que se presenta luego, este procedimiento de DR plug-in no implementa ni sample splitting ni cross-fitting: las funciones auxiliares $\hat{\mu_1}(X)$, $\hat{\mu_0}(X)$ y $\hat{p}(X)$ se entrenan y se aplican sobre la misma muestra. Es interesante poder agregar este método de esta manera para poder evaluar en el mismo ejercicio como performa este estimador sin considerar el resampleo necesario para no caer en overfitting y ver en que medida esto puede afectar a la estimación del ATE.

Algoritmo 3 Implementación del estimador Doubly Robust plug-in con ML

1: Para cada escenario simulado, seleccionar los modelos auxiliares (modelo para OR y modelo para PS) preentrenados y optimizados según la estructura funcional del DGP.

2: Estimar el modelo de regresión del resultado $Y \sim X$ para el subgrupo de tratados ($D = 1$) usando un modelo de boosting. Utilizando este modelo obtener predicciones $\hat{\mu}_1(X_i)$ para toda la muestra.

3: Estimar el modelo de OR $Y \sim X$ para el subgrupo de controles ($D = 0$) usando boosting. Obtener predicciones $\hat{\mu}_0(X_i)$ para toda la muestra.

4: Estimar el modelo de PS $D \sim X$ mediante clasificación probabilística con un modelo de boosting de clasificación, utilizando los hiperparámetros óptimos hallados en el punto 1. Obtener $\hat{p}(X_i)$, la probabilidad estimada de tratamiento.

5: Calcular la estimación individual del estimador DR plug-in con ML:

$$\text{DR}_i = \frac{D_i \left(Y_i - \hat{\mu}_1(X_i)\right)}{\hat{p}(X_i)} - \frac{(1 - D_i) \left(Y_i - \hat{\mu}_0(X_i)\right)}{1 - \hat{p}(X_i)} + \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

6: Calcular el estimador del ATE como el promedio muestral de DR_i.

7: Calcular el desvío estándar y el error estándar muestral.

4.4.3. DML

El paquete DoubleML (Bach et al., 2024) ofrece una implementación en R del método de Double/Debiased Machine Learning propuesto por Chernozhukov et al. (2018), que permite estimar parámetros causales en modelos parcialmente lineales utilizando técnicas de aprendizaje automático, ortogonalización de los estimadores e inferencia estadística robusta en contextos de alta dimensionalidad.

En nuestro caso se emplea la clase DoubleMLPLR con el score IV-type, ya que esta especificación corresponde al modelo parcialmente lineal donde el tratamiento se incluye como regresor endógeno y su efecto se identifica a partir de la ortogonalización conjunta de las funciones auxiliares. Este método nos ayuda a corregir el potencial sesgo de regularización.

Cabe mencionar que en esta aplicación, se entrenan los modelos auxiliares internamente en los folds al hacer la validación cruzada, por lo que no podemos introducir directamente los modelos auxiliares previamente entrenados (boosting de regresión y clasificación). Para mantener la estructura de los modelos previos, definimos los learners con las mismas características, hiperparámetros y estructura de los modelos evaluados, a modo de poder reproducir el comportamiento de los modelos auxiliares empleados para aproximar las funciones OR y PS.

Aplicación de DML para obtener el ATE

Para la estimación del ATE seguimos la aplicación de este paquete de la manera descrita a continuación:

Algoritmo 4 Implementación del Estimador Double Machine Learning (DML)

1: Definir el objeto de datos DML del paquete `DoubleML` usando la función `double_ml_data_from_data_frame`.
2: Seleccionar como learners los modelos auxiliares previamente entrenados (modelos de boosting con hiperparámetros optimizados) para cada función de aprendizaje:
   - $E[Y|X]$: Modelo de regresión.
   - $E[D|X]$: Modelo de clasificación.
3: Crear las instancias del objeto estimador DML parcialmente lineal (`DoubleMLPLR`) con las funciones de aprendizaje seleccionadas, número de folds para cross-fitting y el tipo de score (`IV-type`), según Chernozhukov et al. (2018).
4: Ajustar el estimador DML sobre los datos.
5: Obtener el ATE estimado ($\hat{\theta}$), el error estándar (SE) y el intervalo de confianza al 95 % a partir del objeto ajustado.

4.5. Métricas de comparación

Para la comparativa de métodos entre escenarios y tamaños muestrales se presentarán tres métricas de resumen para evaluar el desempeño de los métodos: el error cuadrático medio (RMSE), el sesgo y el desvío estándar como medida de variabilidad.

Con estas métricas es posible determinar la precisión global así como la dispersión de las estimaciones obtenidas por cada método. En conjunto, ofrecen una visión completa del desempeño de los estimadores, mostrando no solo qué tan cercanas son las estimaciones al valor verdadero del parámetro, sino también su estabilidad y consistencia al variar las condiciones de simulación.

Formalmente, siendo N el tamaño muestral y n el individuo seleccionado:

- **Sesgo:** mide la diferencia promedio entre la estimación y el valor verdadero del parámetro:

$$\text{Sesgo} = \frac{1}{N}\sum_{n=1}^{N}(\hat{\theta}_n - \theta_0)$$

- **Desvío estándar:** cuantifica la variabilidad de las estimaciones en torno a su media:

$$\text{SD} = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(\hat{\theta}_n - \bar{\theta})^2}$$

- **Error estándar:** mide el error muestral de las estimaciones realizadas.

$$\text{SE} = \frac{\text{SD}}{\sqrt{n}}$$

4.6. Simulaciones

Para evaluar correctamente el desempeño de los estimadores implementados, se realizaron simulaciones repetidas de cada método (DML, DR paramétrico y DR plug-in ML) utilizando bases de datos generadas a partir del proceso de datos del ejercicio. En cada repetición se modificó la semilla aleatoria, lo que permite generar variaciones independientes en las muestras y obtener una comparación más robusta entre los métodos, evitando que los resultados dependan de realizaciones particulares del azar.

Para cada escenario y combinaciones de tamaño muestral se ejecutaron 100 simulaciones de la estimación para cada método. En cada iteración se estimó el parámetro de interés junto con su error estándar, y posteriormente se recopilaron las estimaciones obtenidas para analizar su desempeño agregado. Este enfoque permite evaluar el comportamiento de cada método en términos de sesgo promedio, desviación estándar y RMSE.

A partir de estas simulaciones, se calculó el sesgo promedio así como el SE y el RMSE asociados. Estas medidas de resumen permiten identificar qué método presenta mejor precisión y menor variabilidad en cada escenario, proporcionando una comparación más sólida y representativa del desempeño relativo de los distintos estimadores.

De este modo, estas medidas promedio de las simulaciones serán utilizadas para la comparativa de los métodos.

Siendo R el número total de simulaciones realizadas, definimos estas métricas como:

$$\overline{\text{Sesgo}} = \frac{1}{R}\sum_{r=1}^{R}\text{Sesgo}_r \qquad \text{SE} = \frac{SD(\hat{\theta}_r)}{\sqrt{R}} \qquad \text{RMSE} = \frac{1}{R}\sum_{r=1}^{R}\sqrt{\overline{(\hat{\theta}_r - \theta)}^2}$$

5. Resultados

5.1. Resultados generales con muestras grandes

Cuadro 4: Resultados de las simulaciones por método y escenario

	DR paramétrico			DR plug-in ML			DML		
Escenario	$\overline{\text{Sesgo}}$	SE	RMSE	$\overline{\text{Sesgo}}$	SE	RMSE	$\overline{\text{Sesgo}}$	SE	RMSE
1. g_{nl} y z_{nl}	-0.158***	0.0002	0.1596	-0.013***	0.0004	0.0413	-0.004	0.0003	0.0366
2. g_{lin} y z_{nl}	0.002	0.0002	0.0194	-0.013*	0.0007	0.0705	-0.004	0.0003	0.0325
3. g_{nl} y z_{lin}	-0.004	0.0005	0.0500	0.281***	0.0026	0.3838	-0.004	0.0006	0.0613
4. g_{lin} y z_{lin}	-0.004	0.0004	0.0435	0.407***	0.0020	0.4537	0.000	0.0059	0.0587

Promedio de sesgo, error estándar (SE) y raíz del error cuadrático medio (RMSE) por método y escenario, calculados sobre $R = 100$ simulaciones con $n = 10{,}000$ observaciones por simulación. Valor verdadero: $\theta_0 = 0.5$.
Asteriscos (*) según significancia del sesgo: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Este primer análisis nos permite ver como se comportan cada uno de los estimadores presentados frente a una muestra grande, de n=10,000. Los resultados obtenidos se alinean en términos generales con lo que predice la teoría, mostrando patrones coherentes con las propiedades de cada estimador. En promedio, el DML presenta un mejor desempeño que el estimador DR plug-in ML en todos los escenarios, siendo esta diferencia especialmente marcada en el escenario 3, donde el DR plug-in exhibe un sesgo promedio muy elevado, y en el escenario 4, donde, si bien el sesgo es menor, continúa siendo estadísticamente significativo. En contraste, el DML no muestra sesgos significativos al 5 % en ninguno de los escenarios, manteniendo estimaciones centradas en torno al valor verdadero del parámetro.

El DR paramétrico, por su parte, muestra un desempeño adecuado en la mayoría de los casos, salvo en el escenario 1, donde tanto la función de resultado como la función de propensión (PS) están mal especificadas, lo que se traduce en un sesgo significativo. El promedio de $\hat{\theta}$ en este caso es 0.342 con un error estándar de aproximadamente 0.0002, por lo que el intervalo de confianza al 95 % para θ es [0.3416 ; 0.3424], no conteniendo el valor verdadero $\theta_0 = 0.5$. En los demás escenarios (2–4), el sesgo promedio es bajo a moderado, y el estimador se comporta en línea con su propiedad de doble robustez, capturando correctamente el efecto cuando al menos una de las funciones auxiliares está bien especificada.

Un resultado interesante es que el estimador DR paramétrico no alcanza su máxima eficiencia en el escenario 4, donde ambas funciones auxiliares están correctamente especificadas, sino en el escenario 2, donde el OR es lineal y el PS corresponde a una

función logística con índice sinusoidal. Esto podría deberse a que, dentro del rango de los datos simulados, la función logística del escenario 2 adopta una forma muy similar a una logística lineal, reduciendo las diferencias prácticas entre ambos casos. También es posible que, en el escenario 4, el modelo logístico haya convergido a una parametrización menos informativa, lo que genera una leve pérdida de eficiencia. En cualquier caso, este resultado refleja la sensibilidad empírica del estimador a la forma funcional efectiva de las funciones auxiliares en muestras finitas.

Se adjunta en el anexo la comparativa de las distribuciones superpuestas de $\hat{\theta}$ obtenidas por cada método en los diferentes escenarios con un tamaño muestral de $n = 10{,}000$.

5.2. Resultados por tamaño muestral

Podemos ver para todos los escenarios un comportamiento donde el aumento del tamaño muestral disminuye significativamente la variabilidad de los resultados, quedando con estimaciones en promedio menos dispersas, cada vez más concentradas sobre el valor real del ATE (ver figuras del Conclusiones). A su vez, en los modelos que integran ML para estimar las funciones auxiliares, la estimación del ATE tiende a converger en la mediana hacia el valor verdadero de $\theta = 0.5$ a medida que crece n. En términos generales, a mayor n, las medianas se aproximan al parámetro de interés y las colas se acortan.

En el método DR paramétrico, en cambio, la estimación de θ se encuentra centrada en torno a su valor real de 0.5 aparece desde el inicio siempre que al menos una de las funciones auxiliares esté correctamente especificada (Figura 8). El estimador gana eficiencia al aumentar n, pero su ubicación central no cambia de manera significativa entre un tamaño muestral chico y grande. La excepción es el escenario 1, donde tanto $g_0(x)$ como $m_0(x)$ están mal especificadas y, por lo tanto, la distribución se centra de forma persistente por debajo del valor verdadero (en torno a 0.35), y el sesgo no se corrige ni siquiera con tamaños muestrales altos.

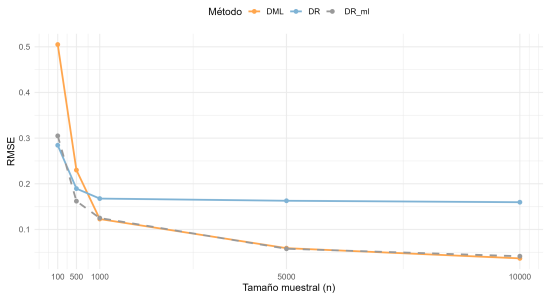
En cuanto al estimador DML (Figura 10), éste exhibe un desempeño muy estable en los cuatro escenarios cuando n es suficientemente grande. La mediana se alinea rápidamente con θ y la dispersión disminuye de forma marcada, sin requerir tamaños muestrales extraordinariamente grandes para que esto ocurra. Este resultado coincide con lo señalado por Chernozhukov et al. (2018) respecto a las velocidades de convergencia: la estructura ortogonal del DML, junto con el uso de *sample splitting* para evitar el sobreajuste, favorece una convergencia más rápida hacia el valor verdadero del parámetro de interés.

En contraste, el estimador DR plug-in ML (Figura 9), muestra una mayor sensibilidad según el escenario. En particular, en el escenario 3 falla de forma notoria en la convergencia de la mediana, que permanece alejada del valor verdadero de 0.5 incluso cuando n alcanza

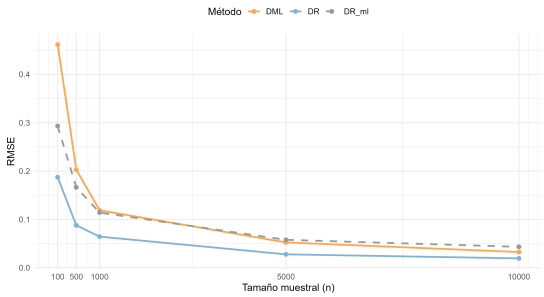
10,000. En el escenario 4, si bien la convergencia mejora, requiere tamaños muestrales más grandes para que la mediana se aproxime al valor real. En los escenarios 1 y 2 el DR plug-in se comporta razonablemente bien y puede asemejarse al DML, aunque su convergencia resulta más frágil ante errores de especificación en $g_0(x)$ y $m_0(x)$, lo que puede comprometer su insesgadez y aumentar los requerimientos de tamaño muestral para un desempeño adecuado.

El DML combina una buena convergencia de la mediana con una reducción rápida de la varianza a medida que crece n, mientras que el DR plug-in ML necesita más datos y presenta menor estabilidad en escenarios con mayor complejidad o desajustes de especificación. Por su parte, el DR paramétrico muestra una ventaja relativa frente a los métodos que integran ML en los escenarios donde las funciones auxiliares están correctamente especificadas y los tamaños muestrales son pequeños, manteniendo un centrado adecuado, incluso con un pequeño tamaño muestral.

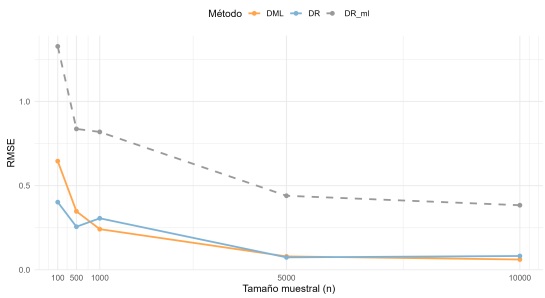
Figura 3: RMSE por tamaño muestral



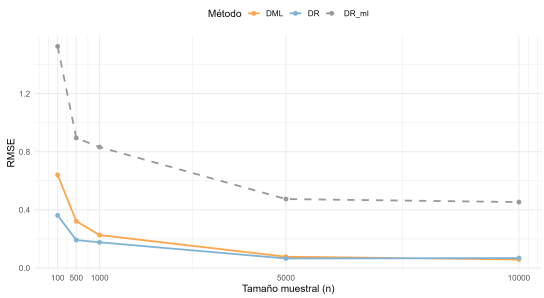
3.1 Escenario 1



3.2 Escenario 2

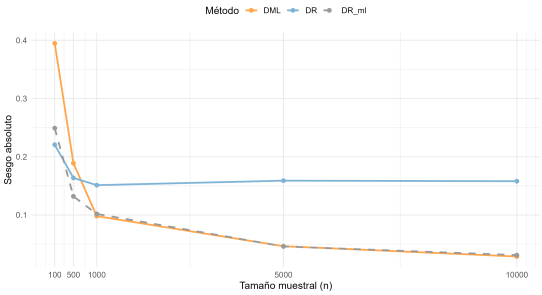


3.3 Escenario 3

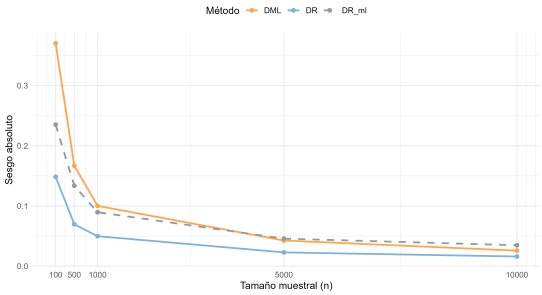


3.4 Escenario 4

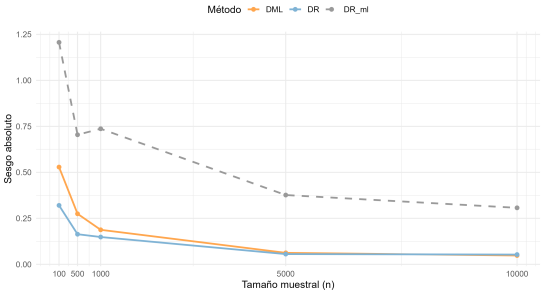
Figura 4: Sesgo absoluto promedio por tamaño muestral



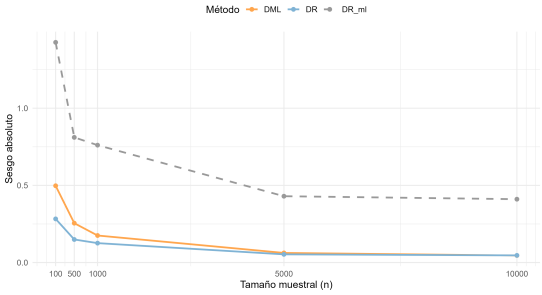
4.1 Escenario 1



4.2 Escenario 2



4.3 Escenario 3



4.4 Escenario 4

6. Conclusiones

En este trabajo se comparó el desempeño del método DML frente al estimador DR paramétrico y su versión con ML sin aplicar las condiciones fundamentales del marco de Chernozhukov et al. (2018), es decir, sin sample splitting ni ortogonalización. El objetivo fue identificar bajo qué condiciones cada método resulta preferible y cómo su rendimiento depende del tamaño muestral y de la complejidad funcional presente en los datos.

Los resultados obtenidos son consistentes con la literatura existente y sus predicciones teóricas. Al trabajar en escenarios con volúmenes muestrales altos, DML logra estimaciones consistentes, estables y con bajo sesgo aun cuando la relación entre las variables es compleja. El método mostró niveles de consistencia en todos los escenarios, comportándose de forma similar en cada uno de ellos y emulando como sería la estimación mediante un DR paramétrico dejando de lado los potenciales problemas de identificación de las funciones auxiliares, lo que destaca el potencial del DML para eliminar el riesgo de mala especificación. Por su parte, el DR con ML sin ortogonalización ni sample splitting exhibe sesgos significativos en todos los escenarios, reflejando que los sesgos en la estimación de los modelos auxiliares y el potencial sobreajuste de los mismos, se transmiten directamente al estimador y comprometen la validez causal.

En conjunto, estos resultados indican que las ventajas teóricas del DML sí se materializan en la práctica sobre el DR paramétrico, pero únicamente cuando se cumplen condiciones muestrales suficientes para garantizar la estimación adecuada de los modelos auxiliares. Cuando se trabaja con muestras pequeñas, el DML no garantiza una estimación consistente del ATE, siendo preferible el uso de alternativas paramétricas como el DR tradicional, siempre que su especificación funcional sea adecuada.

Por su parte, el DML supera de manera consistente al estimador DR con aplicación directa de ML en todos los escenarios evaluados. Este resultado reafirma el punto de Chernozhukov et al. (2018) mostrando como el uso de ML en aplicaciones causales no puede realizarse de forma directa, ya que sin ortogonalización ni sample splitting las estimaciones resultan sesgadas y no válidas.

Es necesario que tener en cuenta que el ejercicio presentado se realizó en un contexto simplificado en donde nos basamos en observar el comportamiento de los métodos centrando el análisis en las formas funcionales y el tamaño muestral que explicaban las bases utilizadas. En este contexto, los métodos no fueron evaluados bajo escenarios de variables relevantes omitidas o de variables irrelevantes incluidas en el modelo de datos, lo que podría afectar como cada método trata este problema para la estimación correcta del ATE. Al mismo tiempo, el ejercicio de simulación realizado contó con solo tres covariables, no evaluándose cada método frente a escenarios de alta dimensionalidad. Es necesario contemplar estos puntos para poder entender correctamente el alcance de estas conclusio-

nes, sus limitaciones y sentar las bases para futuras líneas de investigación en cuanto a la evaluación del desempeño de estos métodos en caso de extender el ejercicio simulado, relajando estos supuestos y aproximándose cada vez más a la realidad.

En resumen, los resultados confirman que el aprendizaje automático aporta valor en inferencia causal cuando se integra dentro de marcos metodológicos que preservan garantías estadísticas, como los que provee DML. La combinación entre la flexibilidad y capacidad predictiva del ML y las estructuras de identificación propias de los métodos tradicionales genera sinergias que permiten aprovechar lo mejor de ambos. Cuando se respetan las condiciones que garantizan consistencia, se amplía el conjunto de herramientas disponibles para estimar efectos causales, ofreciendo alternativas más robustas en escenarios donde los enfoques puramente paramétricos pueden fallar.

Referencias

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

Athey, Susan y Guido W. Imbens (2019). «Machine learning methods that economists should know about». En: *Annual Review of Economics* 11, págs. 685-725.

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

Belloni, Alexandre, Victor Chernozhukov y Christian Hansen (2014). «Inference on Treatment Effects after Selection among High-Dimensional Controls». En: *The Review of Economic Studies* 81.2, págs. 608-650.

Chen, Tianqi y Carlos Guestrin (2016). «XGBoost: A Scalable Tree Boosting System». En: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, págs. 785-794.

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.

Horvitz, D. G. y D. J. Thompson (1952). «A Generalization of Sampling Without Replacement from a Finite Universe». En: *Journal of the American Statistical Association* 47.260, págs. 663-685.

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

LeVeque, Randall J. (2007). *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*. Apéndice A.6. Philadelphia: Society for Industrial y Applied Mathematics.

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.

Robins, James M. y Sander Greenland (1986). «The Role of Models and Likelihood in Causal Inference from Observational Studies». En: *Biometrics* 42.4, págs. 1033-1058.

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.

Anexo I

Se presentan las distribuciones superpuestas por método de $\hat{\theta}$ con n = 10,000.

Figura 5: Comparación de densidades para DR paramétrico asumiendo OR lineal y PS logit-lineal.

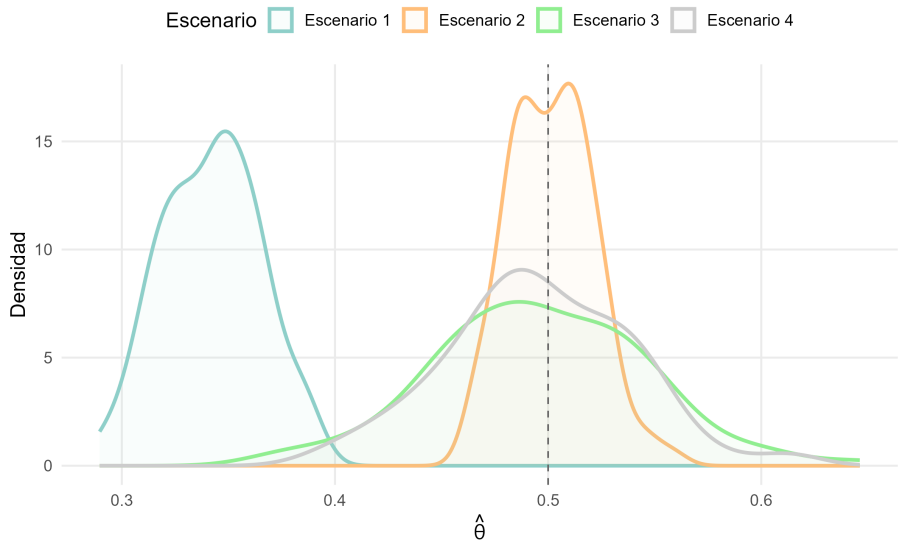


Figura 6: Comparación de densidades para DR con ML plug in.

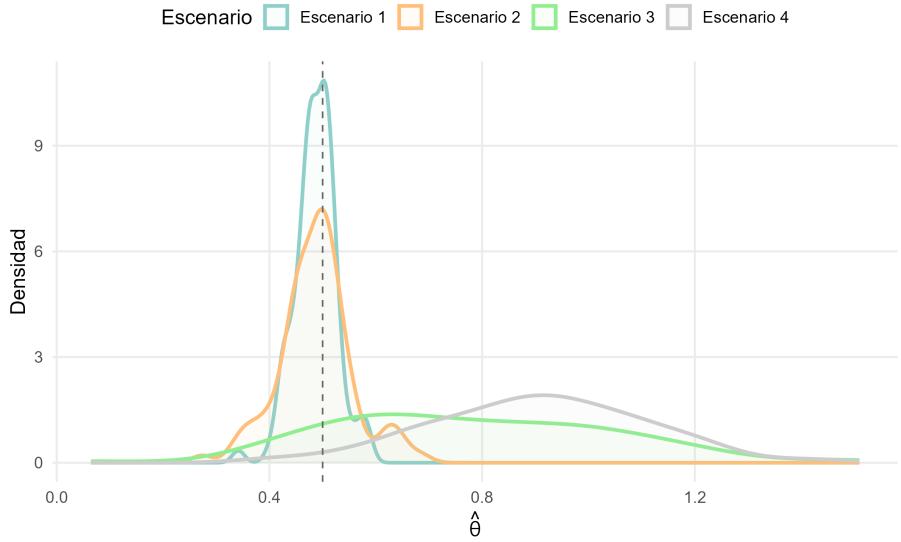
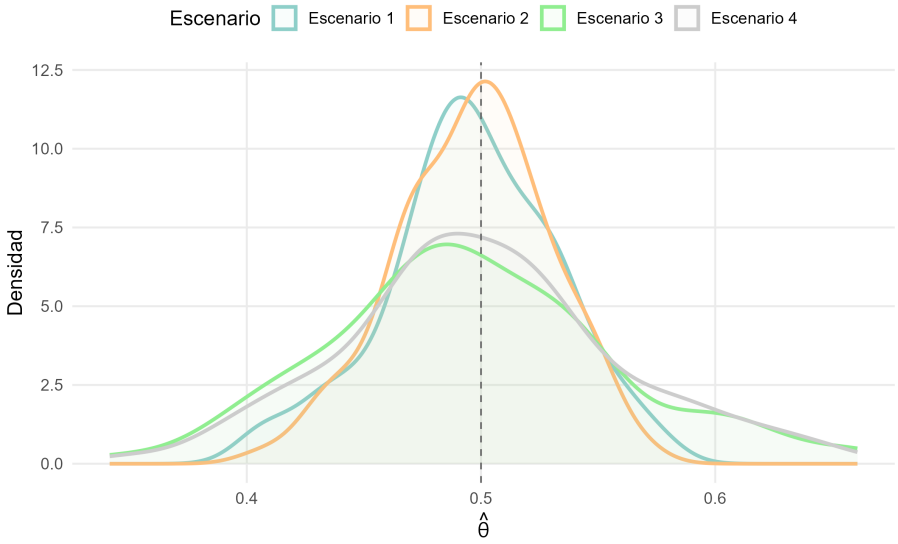


Figura 7: Comparación de densidades para DML.

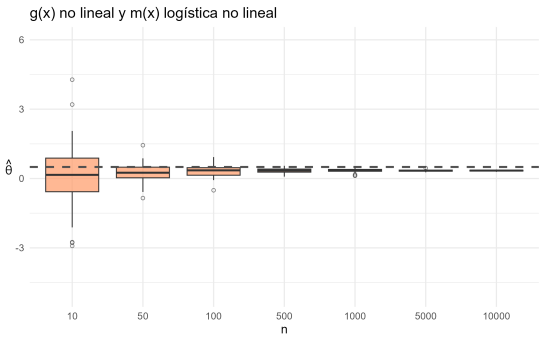


Anexo II

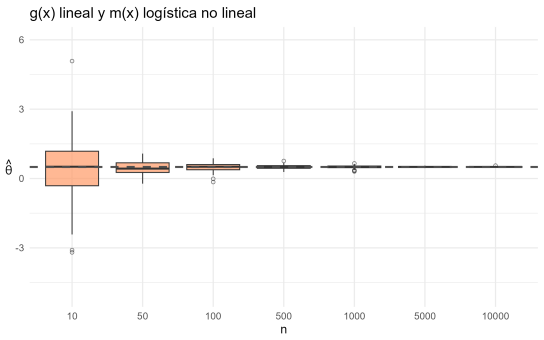
Se presentan los gráficos y tablas correspondientes a los resultados simulados para la obtención de $\hat{\theta}$ por método y escenario para diferentes tamaños muestrales.

Figura 8: Distribuciones de $\hat{\theta}_0$ por tamaño muestral y escenario para DR paramétrico

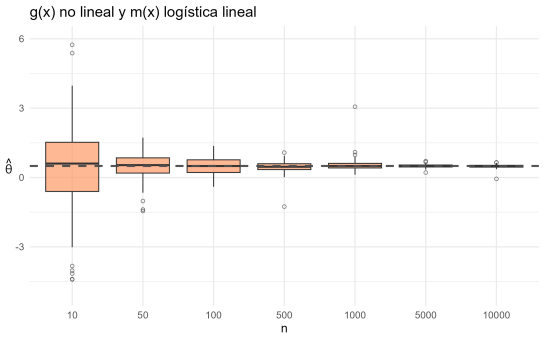
8.1 Escenario 1



8.2 Escenario 2



8.3 Escenario 3



8.4 Escenario 4

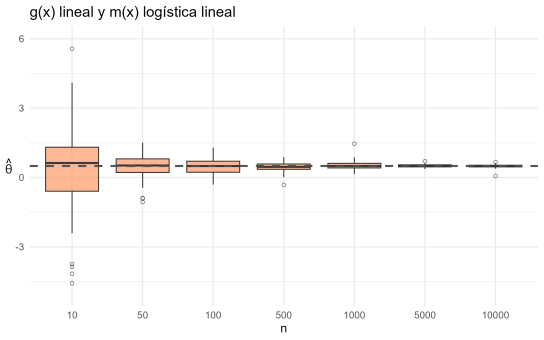
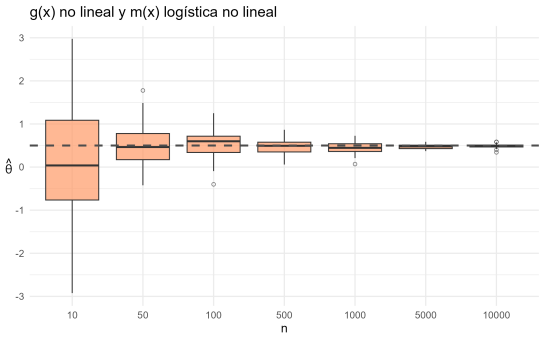
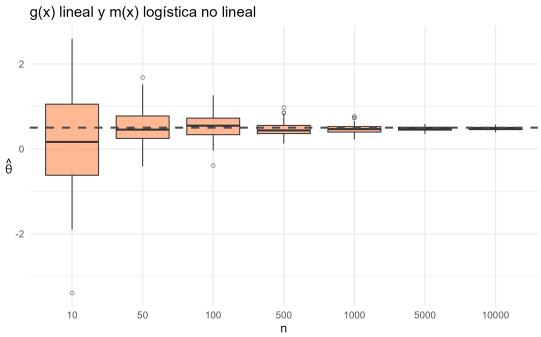


Figura 9: Distribuciones de $\hat{\theta}_0$ por tamaño muestral y escenario para DR plug in ML

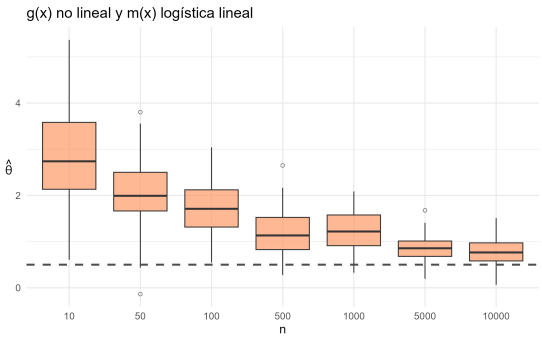
9.1 Escenario 1



9.2 Escenario 2



9.3 Escenario 3



9.4 Escenario 4

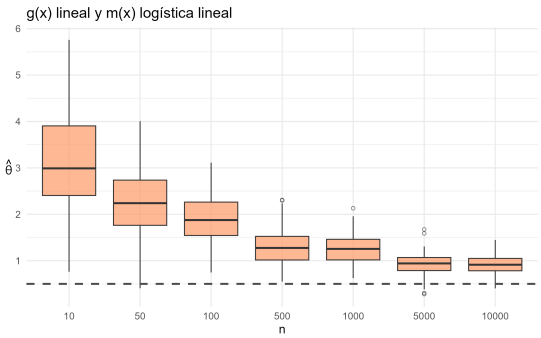
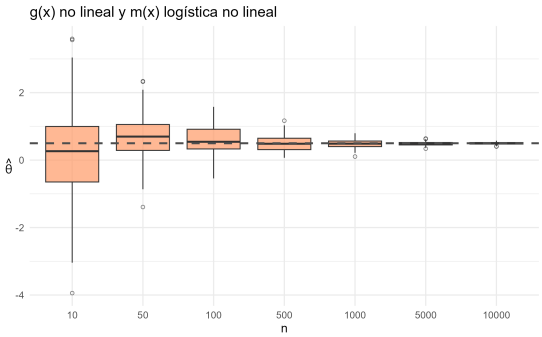
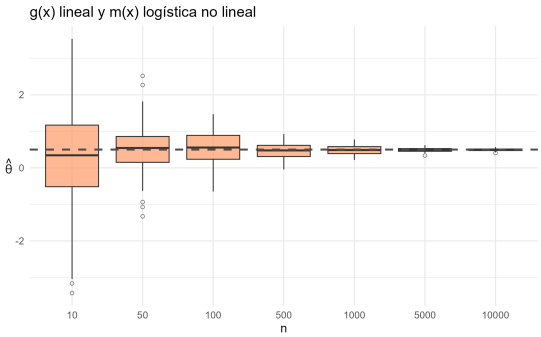


Figura 10: Distribuciones de $\hat{\theta}_0$ por tamaño muestral y escenario para DML

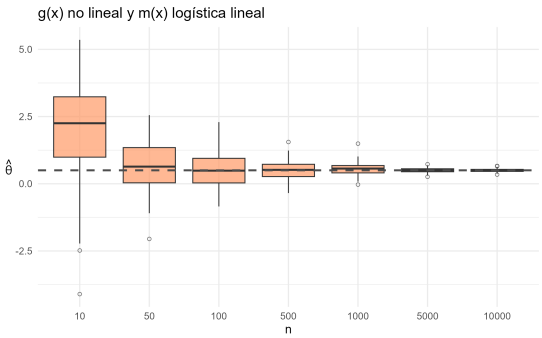
10.1 Escenario 1



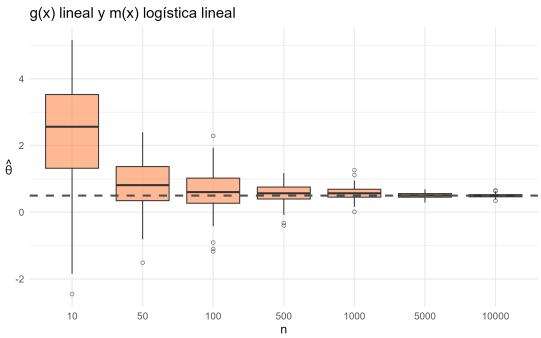
10.2 Escenario 2



10.3 Escenario 3



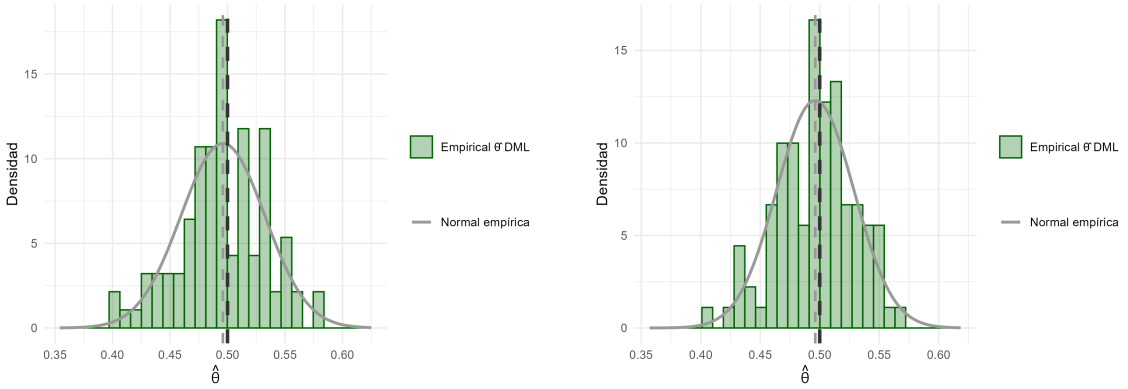
10.4 Escenario 4



Anexo III

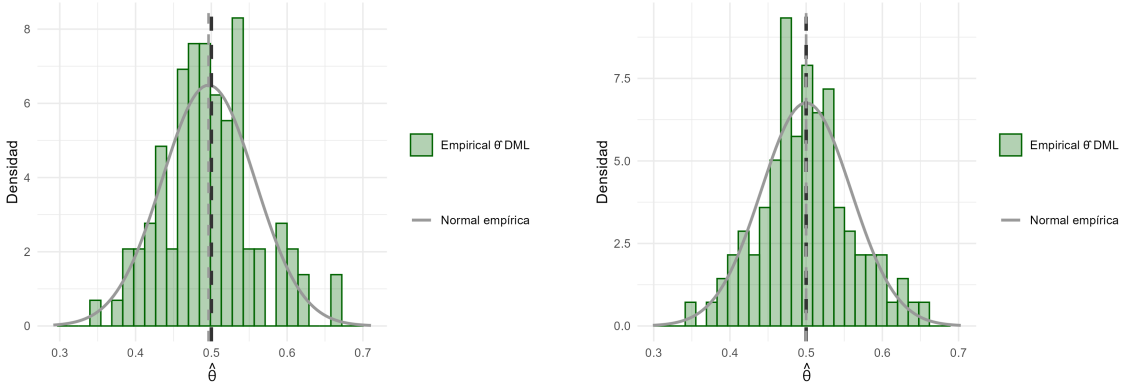
Se presentan histogramas para ver la comparativa de la distribución empírica de las estimaciones de θ mediante el método DML con respecto a una normal estandarizada.

Figura 11: Distribuciones de $\hat{\theta}_{DML}$



11.1 Escenario 1

11.2 Escenario 2



11.3 Escenario 3

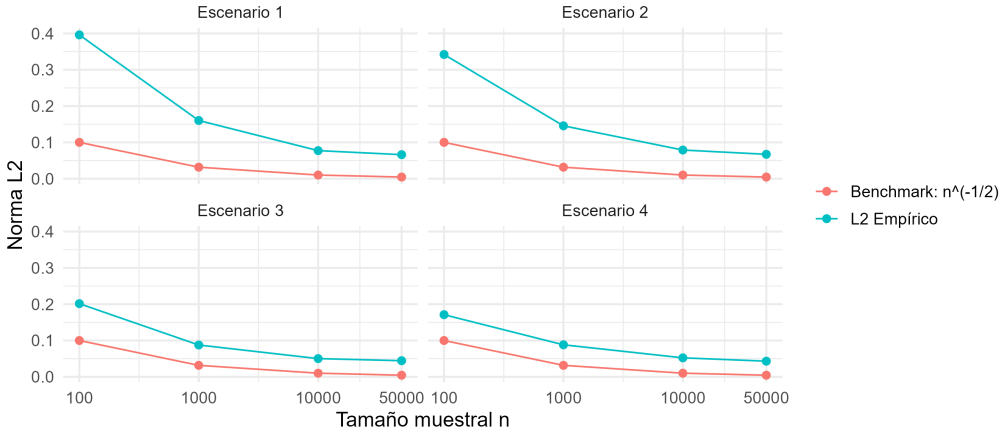
11.4 Escenario 4

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.

Anexo IV

Se presentan los gráficos de evolución de los errores de las estimaciones de las funciones auxiliares $\hat{g}$ y $\hat{m}$ en comparación con la referencia teórica en base al tamaño muestral siguiendo la teoría asintótica descrita en la sección 4.3.1 referida a la Tasa empírica de convergencia por escenario.

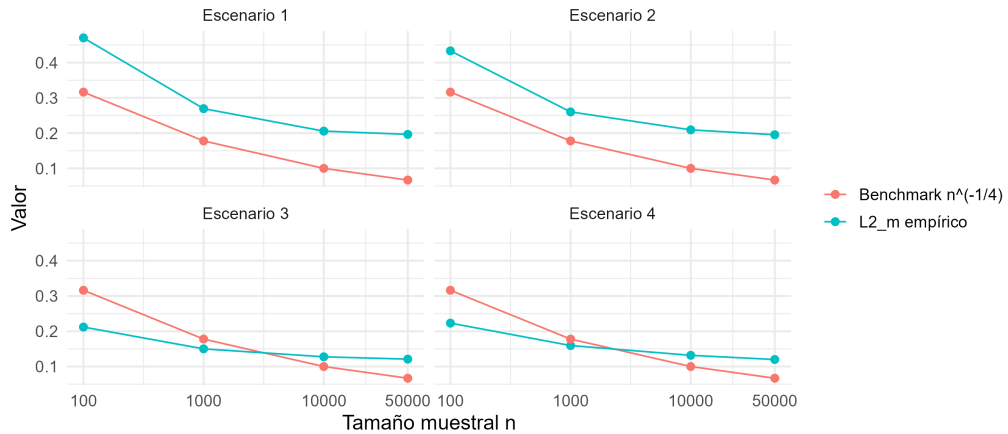
Figura 12: Convergencia de la norma L2 general



Comparación de de la norma de error L2 general: $\|\hat{m}(X) - m_0(X)\| \times \|\hat{g}(X) - g_0(X)\|$ para diferentes tamaños muestrales contra el benchmark en cuanto a la tasa de convergencia asintótica de modelos ortogonalizados $\frac{1}{\sqrt{n}}$ según Chernozhukov et al. (2018)

.

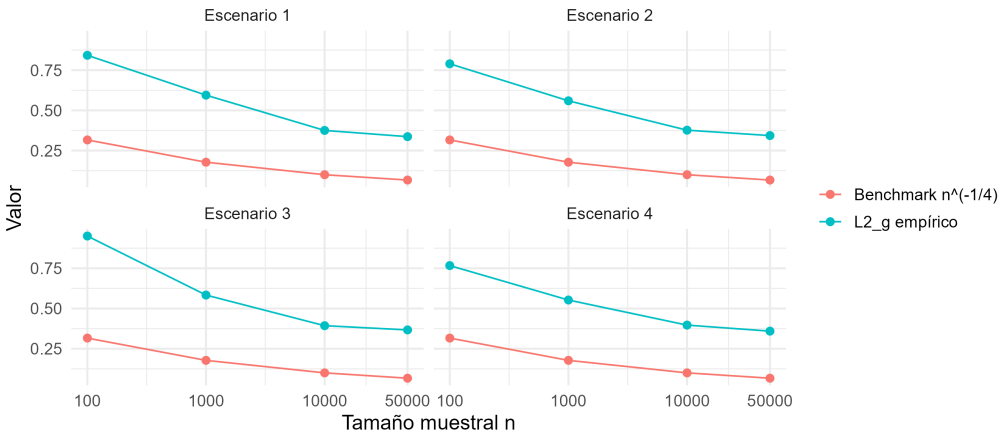
Figura 13: Convergencia de la norma L2 asociada a $\hat{m}$



Comparación de de la norma de error L2 individual para $\hat{m}$: $\|\hat{m}(X) - m_0(X)\|$ para diferentes tamaños muestrales contra el requerimiento en cuanto a la tasa de convergencia individual asintótica de modelos con scores no ortogonales: $\frac{1}{\sqrt[4]{n}}$ según Chernozhukov et al. (2018)

.

Figura 14: Convergencia de la norma L2 asociada a $\hat{g}$



Comparación de de la norma de error L2 individual para $\hat{m}$: $\|\hat{g}(X) - g_0(X)\|$ para diferentes tamaños muestrales contra el requerimiento en cuanto a la tasa de convergencia individual asintótica de modelos con scores no ortogonales: $\frac{1}{\sqrt[4]{n}}$ según Chernozhukov et al. (2018)

.

Anexo V

El código fuente completo utilizado para el ejercicio práctico simulado se encuentra disponible públicamente en el siguiente repositorio de GitHub:

`https://github.com/igonzalezsolano/Gonzalez_Gonzalez_TFM/tree/main`