CAPÍTULO III

Economía laboral y *big data:* panorámica sobre técnicas de regularización en la evaluación de efectos causales

Juan J. Dolado*

En este trabajo se ofrece una panorámica sobre las técnicas de regularización existentes en la literatura de *machine learning* para modelos lineales y no lineales, con controles exógenos y tratamientos endógenos, destinados a evaluar los efecto de deteterminadas políticas sobre variables del mercado de trabajo. Una aplicación empírica de dichas técnicas al conocido estudio de Angrist y Krueger (1991) acerca de los efectos de la educación sobre los salarios sirve para ilustrar su uso creciente en economía laboral.

Palabras clave: big data, machine learning, postselección simple, postselección doble, lasso, efecto causal.

* Estoy agradecido a los editores, Juan Carlos Escanciano y participantes en un seminario interno de UC3M por sus comentarios. Este artículo fue presentado en las jornadas de Funcas sobre *Nuevos Enfoques de Problemas Económicos con "Big Data"*. Se agradece el apoyo financiero del Ministerio de Economía y Competitividad (proyecto ECO2017-86009-P).

1. INTRODUCCIÓN

La disponibilidad de fuentes estadísticas con un gran número de variables relativas al tamaño de la muestra (en el argot: datos masivos o *big data*) es cada vez más común en economía aplicada. Por ello, el uso de técnicas de aprendizaje automático (*machine learning* o *ML*) avanza a pasos agigantados en la investigación económica[1]. La economía laboral no es una excepción a esta regla ya que muchas de las bases de datos administrativas que se utilizan en esta disciplina poseen dimensiones inherentemente elevadas, con multitud de características para cada observación disponible. Por ejemplo, en el caso de España, tanto el *Censo de Población* como la *Encuesta de Población Activa,* la *Muestra Continua de Vidas Laborales,* la *Encuesta de Salarios,* la *Encuesta Financiera de las Familias* o los datos de la Agencia Tributaria, entre otras, proporcionan información sobre cientos de variables en relación con empresas o trabajadores. Su relevancia es fundamental para contrastar las hipótesis derivadas de los principales modelos acerca de, por ejemplo, el funcionamiento de los mercados de trabajo o sobre los efectos de la acumulación de capital humano sobre los salarios. Además, incluso cuando el número de variables relevantes fuera reducido, los investigadores rara vez conocen la forma funcional exacta con que aparecen en el modelo, lo que supone enfrentarse a un gran conjunto de interacciones y transformaciones potenciales de las variables subyacentes.

Sin embargo, como señalan Angrist y Frandsen (2020), *a priori* no está claro que el uso de técnicas de ML se adapte fácilmente a las necesidades de la economía laboral. Tradicionalmente la mayoría de las cuestiones relevantes en esta área de investigación se refieren a las características de las distribuciones de las variables aleatorias de interés (como puede ser la forma funcional de la media condicional), más que a la precisión de las predicciones fuera de la muestra. En efecto, gran parte de la agenda de investigación en economía laboral está dirigida tanto a la estimación de efectos causales –por ejemplo, el efecto de la educación sobre los salarios o qué tipo de tratamiento a los parados resulta más efectivo para mejorar su empleabilidad– como a proporcionar evidencia descriptiva sobre el efecto que tienen los cambios tecnológicos o políticas impositivas sobre la desigualdad de rentas y riqueza. Para ello, las herramientas estadísticas que se han venido utilizando tradicionalmente han sido los métodos de regresión habituales, incluyendo el uso extensivo de variables instrumentales (IV). Por tanto, ya sean causales o descriptivas, rara vez las cuestiones relevantes en economía laboral se han centrado en problemas de predicción pura. Como apuntan acertadamente Mullainathan y Spiess (2017) en su panorámica sobre el uso de ML en economía, la distinción entre estimación de parámetros y predicción individual es paralela a la diferencia entre las pendientes estimadas en un modelo de regresión ($\hat{\beta}$) y el R^2. El objetivo de las técnicas de ML es mejorar la precisión de los valores ajustados ($\hat{y}$), en lugar de optimizar las propiedades del estimador de un determinado coeficiente, aparentemente lo contrario de lo que interesa a los economistas laborales, quienes raramente prestan atención a $\hat{y}$ como objeto central de su investigación.

[1] Panorámicas recientes sobre el uso de ML en economía puede encontrarse en Belloni, Chernozhukov y Hansen (2014a), Mullainathan y Spiess (2017), y Athey e Imbers (2019).

No obstante, como señalan Belloni, Chernozhukov y Hansen (2013) y Chernovzhukov, Hansen y Spindler (2015), la conexión entre ambos tipos de objetivos aparece mucho mas evidente en presencia de big data. En efecto, en cualquier aplicación empírica con multitud de controles resulta necesario evitar un ajuste excesivo intramuestral (*data mining*) que impida extraer conclusiones a muestras diferentes de las que se usan para estimar los parámetros de interés. Igualmente, en presencia de un gran número de intrumentos, la precision de los estimadores de los efectos causales de una variable sobre otra a través de mínimos cuadrados bietápicos (MCB) mejora sustancialmente cuando la estimación de la primera etapa se enfoca como un problema de predicción en el que, de nuevo, se evite un sobreajuste de las variables intrumentadas.

A la vista de estas consideraciones, este trabajo se centra en estos dos dominios (uso de MCO e IV para especificar relaciones con big data) en los que el ML podría desempeñar un papel muy relevante en la búsqueda de efectos en la economía laboral. Para ello, se ofrece una panorámica de procedimientos recientes de regularización (esto es, selección de controles e instrumentos en la especificación de modelos lineales y no lineales con variables exógenas y endógenas).

El resto del artículo está organizado de la siguiente forma. En la sección segunda se revisa la relación existente entre efectos causales y modelos de regresión. Las secciones tercera y cuarta están dedicadas a repasar el uso de procedimientos de regularización en modelos lineales con variables exógenas y endógenas, respectivamente. La quinta sección resume las propiedades estadísticas de los principales métodos de regularización en ML. La sección sexta extiende los resultados anteriores a modelos no lineales. En la sección séptima se ofrece una aplicación empírica de estos procedimientos para la estimación de los rendimientos salariales de la educación. Finalmente, la octava sección concluye.

2. REGRESIÓN Y EFECTOS CAUSALES EN ECONOMÍA LABORAL

Como es bien sabido, la regresión utiliza modelos (generalmente) lineales para describir funciones de expectativas condicionales. La esperanza condicional de una variable aleatoria y_i para un individuo i ($i = 1,2,...,n$), como función de los datos de un conjunto de variables, x_i, se puede escribir $\mathbb{E}[y_i \mid x_i = x]$ o, en forma abreviada, $\mathbb{E}[y_i \mid x]$. El símbolo "$\mathbb{E}$" denota un promedio de población, mientras que $\mathbb{E}[y_i \mid x]$ representa el promedio de y_i para todos aquellos individuos que poseen características x_i iguales a un valor particular, x. Así, el interés de los economistas laborales se ha centrado tradicionalmente en estimar en cuánto aumentan los salarios en promedio con cada etapa educativa completada. Por ejemplo, se compara $\mathbb{E}[y_i \mid x_i = 16]$, el salario promedio de los graduados universitarios, con $\mathbb{E}[y_i \mid x_i = 12]$, el ingreso de los bachilleres. Debido a que $\mathbb{E}[y_i \mid x_i = x]$ toma tantos valores como x, a menudo los economistas aplicados aspiran a simplificar el modelo de esperanza condicional para resumir sus características más importantes, donde la regresión de y_i sobre x_i proporciona la mejor aproximación lineal. Volviendo al ejemplo anterior, la pregunta clave se centra en establecer (si existe) relación causal entre completar un grado universitario (en vez de simplemente el bachillerato) y los ingresos de un determinado individuo. El hecho de completar la educación

superior se denomina variable de tratamiento, denotada de aquí en adelante como d_i. Idealmente uno estaría interesado en computar la diferencia entre los resultados potenciales, $y_{1i} - y_{0i}$, donde y_{1i} e y_{0i} son los ingresos del individuo *i* si se graduara ($d_i = 1$) y si no lo hiciera ($d_i = 0$) pero la dificultad evidente es que, en función del valor que tome d_i solamente se observa una de las dos situaciones: bien y_{1i} o y_{0i}. Por tanto, el analista aspira a medir un efecto causal medio como $\mathbb{E}\,[y_{1i} - y_{0i}]$, denominado efecto promedio del tratamiento *(average treatment effect, ATE)* o bien el efecto promedio condicionado al tratamiento, $\mathbb{E}\,[y_{1i} - y_{0i} \mid d_i = 1]$ *(average treatment on the treated, ATT)* o a su ausencia, $\mathbb{E}\,[y_{1i} - y_{0i} \mid d_i = 0]$ *(average treatment on the non- treated, ATNT)*.

El vínculo entre inferencia causal y regresión se ve facilitado en un contexto de efectos causales homogéneos en el que se subraya el problema del sesgo de selección muestral pasando por alto la distinción entre diferentes tipos de promedios causales. El modelo subyacente se puede escribir en la forma siguiente:

$$\begin{aligned} y_{0i} &= \mu + \upsilon_i, \\ y_{1i} &= \alpha + y_{0i}, \end{aligned} \qquad [1]$$

donde, en la primera ecuación [1], μ es la media de y_{0i} mientras que υ_i representa su desviación individual respecto a dicha media. La segunda ecuación expresa que el efecto causal del tratamiento, $y_{1i} - y_{0i}$, es homogéneo e igual a α. Utilizando la relación existente entre los resultados observados y los contrafactuales a través de la identidad $y_{1i} \equiv y_{0i} + (y_{1i} - y_{0i})\, d_i$, dicho modelo puede reescribirse en términos de una única regresión como:

$$y_i = \mu + \alpha d_i + \upsilon_i \qquad [2]$$

La ecuación [2] plantea el problema del sesgo de selección en términos de υ_i, que se asemeja a un término de error de regresión. Sin embargo, a diferencia de una regresión, donde por definición los residuos no están correlacionados con los regresores, υ_i puede estar correlacionado con d_i excepto que el tratamiento se aplique de manera completamente aleatoria (ver abajo). Con datos observacionales, las soluciones al problema del sesgo de selección se basan en el supuesto clave de *independencia condicional en media* (ICM). En concreto, se supone la existencia de un amplio conjunto de características observables del individuo tales que:

$$\mathbb{E}\,(\upsilon_i \mid d_i = 1, \mathrm{x}_i = \mathrm{x}) = \mathbb{E}\,(\upsilon_i \mid d_i = 0, \mathrm{x}_i = \mathrm{x}), \qquad [3]$$

donde x_i es un vector de *p* controles que toman un valor particular igual a x. En otras palabras, en la población con $\mathrm{x}_i = \mathrm{x}$, la comparación de los ingresos de individuos con diferentes niveles de educación es un contraste de "manzanas con manzanas" en vez de "manzanas con peras". El supuesto clave es que $\mathbb{E}\,(\upsilon_i \mid d_i, \mathrm{x}_i) = \mathbb{E}\,(\upsilon_i \mid \mathrm{x}_i)$, de manera que, si la media condicional de y_{0i} es una función lineal de x_i, los controles han de ser "variables predeterminadas al tratamiento", es decir, no pueden ser resultados en sí mismos. Ello implica que $\mathbb{E}\,(\upsilon_i \mid \mathrm{x}_i = \mathrm{x}) = \beta' \mathrm{x}$, o $\upsilon_i = \beta' \mathrm{x} + \varepsilon_i$, con $\mathbb{E}\,(\varepsilon_i \mid \mathrm{x}) = 0$. Combinando estos supuestos se obtiene el tradicional modelo de regresión con interpretación causal:

$$y_i = \mu + \alpha d_i + \beta' \mathrm{x}_i + \varepsilon_i, \qquad [4]$$

que suele emplearse para obtener estimaciones insesgadas del efecto causal de interés, α. Nótese que, aunque generalmente el vector de coeficientes de los controles, β, no suele ser objeto de interés por parte del analista, resulta crucial incluir las variables x_i en la regresión como diagnóstico sobre la plausibilidad de [4]. Claramente, cuando p sea muy grande, será imprescindible aplicar métodos de regularización que permitan acotar el conjunto de controles relevantes para los que se cumple el supuesto de ICM.

No obstante, en la práctica, normalmente los modelos teóricos no especifican todas las variables que deben controlarse al estimar una relación, además de que puede resultar complicado medirlas con precisión incluso cuando la especificación es correcta. Por ejemplo, este es el caso de la habilidad intelectual innata de un individuo en una ecuación *minceriana* de salarios[2]. Una solución al problema de las variables omitidas es asignar de forma aleatoria a los participantes en los grupos de tratamiento y de control, de manera que la participación en el programa no esté correlacionada con los factores personales o sociales omitidos.

Sin embargo, los experimentos aleatorios no siempre son factibles, incluso condicionando en observables, como ocurre en [4]. No resultaría ético obligar a un grupo de personas a asistir a la escuela un año más al tiempo que se excluye a otro, de la misma forma que no parece razonable asignar el valor del salario mínimo al azar entre diferentes regiones de un país. Sin embargo, sí que resulta posible identificar un grado de variación exógena en variables como la escolaridad. Las variables instrumentales ofrecen una posible solución en el contexto de experimentos naturales.

En el caso de la educación, la teoría del capital humano sugiere que las personas eligen su nivel de educación comparando los costes y beneficios de las diferentes alternativas a que se enfrentan. Por tanto, una posible fuente de instrumentos podría estar en las diferencias en las políticas de préstamos, becas u otros subsidios que varían independientemente de la habilidad o el potencial de ingresos, o en la existencia de las limitaciones institucionales en la edad de acceso a la educación obligatoria. En este último caso, un famoso trabajo de Angrist y Krueger (1991), que analizaremos en detalle posteriormente, elige un amplio conjunto de variable instrumentales, denominadas z_i, basadas en la regla de que, en varios distritos escolares de EE. UU., los niños ingresan en la educación obligatoria en otoño del año en que cumplen 6 años, mientras que a todos se les permite abandonar la escuela en el momento de cumplir 16 años. Por ello, los alumnos nacidos a principios de año acceden a la escuela a una edad más avanzada que aquellos nacidos a finales de año, de manera que alcanzan la edad legal de abandono escolar tras haber obtenido menos años de educación. En esencia, la combinación de las políticas sobre la edad de inicio de la escuela y las leyes de escolarización obligatoria crean un experimento natural en el que los niños se ven obligados a asistir a la escuela durante

[2] Una ecuación minceriana capta el rendimiento de la educación en términos salariales como el coeficiente de los años de educación en un modelo de regresión donde la variable dependiente es el logaritmo de los salarios por hora y a la que se añaden otros regresores, como la antigüedad laboral, para capturar la formación en el puesto de trabajo; véase Heckman, Lochner y Todd (2006).

periodos de tiempo diferentes, de acuerdo con su fecha de nacimiento, lo que repercute en su educación y por tanto en sus ingresos futuros.

En el caso de IV, cuando el tratamiento no sea aleatorio y pueda depender del término de error, el modelo causal relevante sería el siguiente:

$$\begin{aligned} y_i &= \alpha d_i + \beta' \mathrm{x}_i + \varepsilon_i, \\ d_i &= \gamma' z_i + e_i, \\ \mathrm{z}_i &= \varphi' \mathrm{x}_i + \omega_i, \end{aligned} \qquad [5]$$

donde el supuesto de ICM implicaría $\mathbb{E}(\varepsilon_i \mid \mathrm{z}_i, x_i) = 0$, siendo z_i un conjunto de instrumentos válidos proporcionados por algún experimento natural, como el discutido en la sección séptima[3]. De nuevo, cuando las dimensiones de los vectores de parámetros β, γ y φ sean muy elevadas, la aplicación de métodos de regularización de ML resultará imprescindible para logran un buen estimador del parámetro de interés α.

3. SELECCIÓN DE CONTROLES EN MODELOS LINEALES CON VARIABLES EXÓGENAS (MCO)

Empezaremos considerando el caso del modelo [4] donde existen p controles incorrelacionados con el término de error ε_i pero correlacionados con la variable de interés, d_i. Lo relevante en presencia de big data es que la dimensión p puede ser muy elevada (incluso muy superior al tamaño muestral), por las razones que se apuntaban anteriormente: por una parte, la creciente disponibilidad de infinidad de características individuales cuyo efecto sobre la distribución de d_i no puede descartarse *a priori* y, por otra, el hecho de que la forma funcional de las esperanzas condicionales sea desconocida y muy flexible. Como punto de partida, supondremos que la media condicional es lineal, de manera que [4] corresponde a:

$$y_i = \alpha d_i + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad \mathbb{E}(\varepsilon_i \mid \mathrm{x}_i, z_i) = 0, \qquad [6]$$

donde, para simplificar la notación, el término constante de la regresión se ha incluido en el conjunto de p controles.

El siguiente paso se centra en escoger un método de regularización para seleccionar los controles más relevantes en [6]. La práctica más habitual consiste en seleccionar dicho subconjunto de regresores en [6] usando lo que se ha denominado el método de *Post Selección (PS)*. Funciona de la siguiente manera. Primero, se incluye un determinado regresor x_{ij} si resulta ser un predictor significativo de y_i en la ecuación [6], habiendo excluido d_i. Para ello se puede utilizar un contraste conservador dentro del ambito clásico (tests t o F) cuando

[3] Nótese que en la segunda ecuación de [5] se pueden incluir el control x_i, de forma que $d_i = \gamma' \mathrm{z}_i + \phi' \mathrm{x}_i + e_i$. Se supone que $\phi = 0$ para simplificar los cálculos sin que cambien cualitativamente los resultados derivados en secciones posteriores.

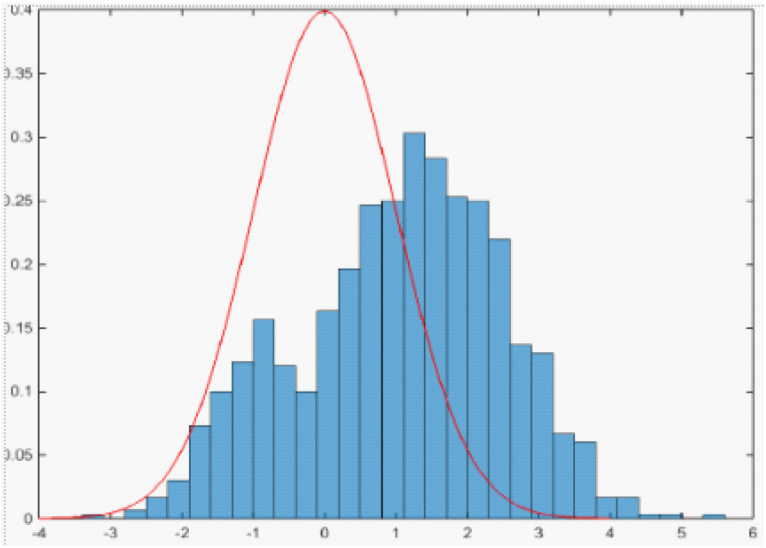
$p < n$ o alguna técnica más moderna dentro del conjunto de herramientas habituales de ML (Lasso y sus variantes, Random Trees, Random Forests, Boosting, Bagging, Neural Networks, etc.)[4]; en caso contrario, se excluyen dichos regresores. Una vez seleccionados los controles x_{ij} que son relevantes para predecir y_i, se vuelve a ajustar el modelo, esta vez incluyendo d_i en la regresión. De esta forma se estima el parámetro de interés α utilizando intervalos de confianza estándar. Sin embargo, como veremos a continuación, este procedimiento puede fallar si $|\beta|$ es cercano (pero no igual) a cero, esto es, formalmente cuando $|\beta| \approx 1/\sqrt{n}$. Para entender los problemas derivados de PS, es conveniente analizar un proceso de generación de datos (PGD en adelante) muy sencillo propuesto por Belloni *et al.* (2014a y b), donde hay un único control ($p = 1$) que está correlacionado con el tratamiento d_i pero no con la perturbación ε_i, es decir:

$$\begin{aligned} y_i &= \alpha d_i + \beta x_i + \varepsilon_i, \\ d_i &= \gamma x_i + \upsilon_i, \end{aligned} \qquad [7]$$

donde se simulan los datos para $n = 100$ considerando los siguientes valores de los parámetros: $\alpha = 0$, $\beta = 0{,}2$, $\gamma = 0{,}8$, $\varepsilon_i \sim N(0,1)$, $x_i \sim N(0,1)$, $\upsilon_i \sim N(0{,}0{,}32)$ y $\mathbb{E}(\varepsilon_i \upsilon_i) = 0$. Como se muestra en la figura 1, tras implementar el método PS en 1.000 simulaciones de Monte Carlo, se rechaza la hipotesis nula $H_0 : \alpha = 0$ en alrededor del 50 % de los casos con un nivel de significación nominal del 5 % para el constraste t, y lo mismo ocurre con Lasso (se puede demostrar que cuando $p \ll n$ ambos procedimientos propocionan resultados muy similares).

Figura 1.

(Post Selección (PS): t-ratio al 5 %)



Fuente: Elaboraciòn propia.

[4] Para un excelente compendio de técnicas de ML, véase Hastie, Tibshirani y Friedman (2009).

Otra posibilidad sería usar *Bootstrap*, pero tampoco funciona bien puesto que simplemente replica la distribución $N(0,1)$ del término de error ε_i.

Por contra, la aplicación de un procedimiento alternativo, denominado *Post Selección Doble (PSD),* sí que logra que el nivel de significación efectivo de los contrastes coincida con el nivel nominal (véase Belloni *et al.* (2014a y b). Dicho procedimiento consiste en los siguientes pasos:

(i) Se incluye x_{ij} como control en la regresión si resulta ser un predictor significativo tanto de y_i como de d_i con cualquiera de los procedimientos clásicos o de ML señalados anteriormente.

(ii) Una vez seleccionado o descartado el control, se ajusta el modelo utilizando intervalos de confianza estándar, o alternativamente, se regresa el residuo obtenido para y_i sobre el residuo de d_i, ambos obtenidos a partir de (i).

Nótese que el procedimiento PSD es equivalente al uso del teorema de parcialización de Frisch-Waugh-Lovell para el cómputo de los coeficientes de MCO en el modelo de regresión lineal estándar. En resumen, se incluye x_{ij} en la regresión siempre y cuando ayude a predecir tanto la variable dependiente como el tratamiento, a diferencia del método PS donde la inclusión de x_{ij} en la primera ecuación de [7] solo depende de si predice bien y_i.

El origen del problema de utilizar PS puede entenderse de manera intuitiva computando la forma reducida de y_i en el modelo ilustrativo [7], esto es:.

$$y_i = (\alpha\gamma + \beta)\, x_i + (\varepsilon_i + \alpha v_i) = \pi\, x_i + \eta_i, \qquad [8]$$

donde $\pi = \alpha\gamma + \beta$, y $\sigma_\eta^2 = \sigma_\varepsilon^2 + \alpha^2\sigma_v^2$. A partir de [8] se observa como el método PS solo tenderá a escoger x_i como regresor relevante en [7] cuando su coeficiente π alcance un valor suficientemente elevado, mientras que se descartará x_i cuando el tamaño de su coeficiente sea reducido. Sin embargo, en este último caso, el hecho de descartar un control que presenta un fuerte poder predictivo para d_i (p. ej. en la simulación anterior $\gamma = 0{,}8$), puede acarrear un importante *sesgo de variables omitidas* (SVO) en el estimador de α cuando el coeficiente β de esta variable en la ecuación a estimar sea pequeño (en el PGD previo, $\beta = 0{,}2$). Intuitivamente, el efecto de cualquier control con un impacto directo moderado sobre la variable y_i se atribuirá incorrectamente al efecto del tratamiento d_i, con el que está fuertemente correlacionado. En consecuencia, la variable x_i quedará excluida de la regresión. Igualmente, si se aplicara un método de selección de variables para predecir d_i en la segunda ecuación de [7], se excluiría x_i siempre que γ fuera reducido, lo cual sería incorrecto en caso de que el tamaño de β sea elevado. De nuevo, tal tipo de omisión de variables puede producir un SVO no despreciable.

En otras palabras, para aminorar el SVO en la estimacion de α, resulta crucial incluir en la primera regresión de [7] todos aquellos controles que sean que resulten útiles para predecir tanto y_i como d_i, procediendo a continuación a regresar (en este caso por MCO) y_i sobre d_i y la unión de todos los controles preseleccionados en la primera etapa.

Las consecuencias de omitir d_i en la regresión [7] cuando su impacto directo sobre y_i es reducido (*i.e.* β es reducido) pueden apreciarse formalmente derivando el SVO a partir de la expresión [8], esto es[5]:

$$\sqrt{n}\left(\hat{\alpha}-\alpha\right)=\left(\sum_{i=1}^{n}\frac{d_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{d_i\varepsilon_i}{\sqrt{n}}+\sqrt{n}\left(\sum_{i=1}^{n}\frac{d_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{x_i^2}{n}\beta\gamma:=\left(A\right)+\left(B\right).$$

Bajo el supuesto de exogeneidad de x_i, el término (A) en la expresión anterior es asintóticamente $N(0,E\left(d_i^2\right)^{-1})$. Nótese que el procedimiento PS descartará correctamente el control x_i siempre que su coeficiente β sea suficientemente pequeño, lo que formalmente ocurre cuando $\beta = O\left(\frac{1}{\sqrt{n}}\right)$. No obstante, incluso en dicho caso, el término (B) puede no anularse ya que asintóticamente se comporta como:

$$\sqrt{n}\beta\gamma\approx\sqrt{n}\,O\left(\frac{1}{\sqrt{n}}\right)\gamma\nrightarrow 0 \text{ si } \gamma\neq 0,$$

Por contra, el método de PSD solamente descartará x_i si no aparece como predictor descritivo relevante tanto para y_i como para d_i. Al igual que con β, ello ocurrirá si el tamaño del coeficiente γ es reducido, esto es, cuando $\gamma = O\left(\frac{1}{\sqrt{n}}\right)$. En dicho caso, el término (B) se convierte en :

$$\sqrt{n}\beta\gamma\approx\sqrt{n}O\left(\frac{1}{\sqrt{n}}\right)O\left(\frac{1}{\sqrt{n}}\right)\rightarrow 0,$$

por lo que esta vez el SVO desaparece.

4. SELECCIÓN DE CONTROLES EN MODELOS LINEALES CON TRATAMIENTO NO ALEATORIO (IV)

Adicionalmente, cabe señalar que los argumentos en el ejemplo anterior se generalizan fácilmente al caso de IV, cuando el tratamiento d_i no sea asignado aleatoriamente (condicional en x_i) sino que se administre en función de una variable instrumental z_i correlacionada con x_i pero ortogonal a la perturbación ε_i. En este caso, el PGD corresponde al siguiente sistema de tres ecuaciones:

$$\begin{aligned} y_i &= \alpha d_i + \beta x_i + \varepsilon_i, \\ d_i &= \gamma z_i + e_i, \\ z_i &= \varphi x_i + \omega_i, \end{aligned} \qquad [9]$$

donde ahora $\mathbb{E}(\varepsilon_i\, e_i) \neq 0$, $\mathbb{E}(\varepsilon_i\, \omega_i) = 0$ y $\mathbb{E}(\varepsilon_i \mid x_i) = \mathbb{E}(\omega_i \mid x_i) = 0$. Al igual que en [7], el sistema se puede reeescribir en forma reducida, de manera que el vector de variables (y, d, z) dependa

[5] Recuérdese que en una regresión lineal con variables exógenas estimada por MCO $y = \beta_1 x + \beta_2 z + u$, el sesgo de β_1 al excluir z es $\mathbb{E}\left(\hat{\beta}_1\right) - \beta_1 = \beta_2 \mathbb{E}(x,z) / \mathbb{E}\left(x^2\right)$.

del control exógeno x_i. En este caso, la aplicación del procedimiento PSD consistirá en escoger x_i siempre que sea un buen predictor de cada una de las tres variables en el sistema, aplicando posteriormente MCB a la primera ecuación de [9] con el fin de estimar α.

Pese a haber analizado inicialmente el caso sencillo con $p = 1$ a efectos ilustrativos, en la práctica el caso más realista en presencia de big data es aquel donde p es muy grande, dependiendo posiblemente del tamaño muestral n, de manera que $p = p_n$, donde $p \approx n$ o p » n. En estas circunstancias, la aplicación de MCO o MCB no es factible y, por ello, se requiere el uso de métodos de regularización basados en ML. Para cubrir los casos analizados previamente en términos de modelos predictivos, denotemos como $\mathbf{w}_i$ al vector de los datos $(y_i, d_i, z_i)'$, de manera que dichos modelos pueden representarse de la siguiente forma:

$$\mathbf{w}_i = \sum_{j=1}^{p} \phi_j x_{ij} + \xi_i;\ \mathbb{E}\left(\xi_i \mid \mathrm{x}_{ij}\right) = 0,\ j = 1,2,...,p. \qquad [10]$$

Para proceder a la regularización de los coeficientes en [10] se necesitan dos supuestos clave en ML: (I) *Parsimonia (approximate sparsity)* en el conjunto de parámetros ϕ en [10], lo que conlleva la existencia de un subconjunto de dichos coeficientes, de dimensión $s_n \ll p_n$, que son relevantes mientras que los restantes no lo son tanto; por ejemplo, tras ordenar los coeficientes por tamaño, esta condición se verifica si $|\phi_j| \leq Aj^{-a}$ para $j = 1,2,...,p_n$, siendo A una constante y $a > 1$, y (II) *Isometria Resgringida,* una propiedad de álgebra lineal aplicada a la matriz de covarianzas de los controles que implica la existencia de pequeños grupos de regresores que son cuasi-ortonormales, es decir, con dependencia muy reducida.

5. MÉTODOS DE REGULARIZACIÓN EN ML

Los supuestos descritos anteriormente subyacen a la mayoría de los procedimientos de selección de variables mediante ML, entre los que se encuentran los métodos de regularización más populares en econometría. La idea básica de estos procedimientos es que un predictor mejor fuera de la muestra puede conllevar un aumento de las sumas de los cuadrados de los residuos en [10]. En consecuencia, se añade un término de regularización que se encargue de la elimininación de los coeficientes más pequeños y, con ello, del diseño de modelos más parsimoniosos, esto es, con una menor dimensión que la contemplada en un ajuste por MCO del modelo sin restricciones. Idealmente, los estimadores minizan la siguiente función de pérdidas (donde, a efectos ilustrativos, nos centramos en la ecuación que determina y_i en [10]):

$$\min_b \left[\sum_{i=1}^{n} \frac{\left(y_1 - \Sigma_{j=1}^{p} b_j x_{ij}\right)^2}{n} + \frac{\lambda}{n} \sum_{j=1}^{p} 1\{b_j \neq 0\} \right], \qquad [11]$$

donde 1 {·}es una función indicador y el parámetro λ penaliza la dimensión del modelo. Dicho objetivo incluye a los criterios de información de Akaike y Schwartz. Desafortunadamente, este problema resulta prohibitivo en términos computacionales (problema NP) cuando p es muy grande, ya que requiere efectuar $\Sigma_{j \leq n} \binom{p}{j}$ regresiones. La solución propuesta por el

método Lasso es convexificar la función de pérdidas anterior sustituyendo el tamaño de los coeficientes por su valor absoluto, lo que da lugar a la siguiente función alternativa de pérdidas (Tibshirani, 1996):

$$\min_b \left[\sum_{i=1}^{n} \frac{\left(y_1 - \Sigma_{j=1}^{p} b_j x_{ij}\right)^2}{n} + \frac{\lambda}{n} \sum_{j=1}^{p} |b_j| \right], \qquad [12]$$

o en el caso de Lasso adaptativo

$$\min_b \left[\sum_{i=1}^{n} \frac{\left(y_1 - \Sigma_{j=1}^{p} b_j x_{ij}\right)^2}{n} + \frac{\lambda}{n} \sum_{j=1}^{p} \varpi_j |b_j| \right],$$

donde los pesos ϖ_j representan las penalizaciones heterogéneas de los coeficientes b_j que, junto a λ, han de elegirse antes de aplicar Lasso, bien mediante procedimientos de validación cruzada (CV) o a través de un estimador complentario (*plug-in*). Bajo los dos supuestos enunciados previamente, Bickel, Ritov y Tsybakov (2009) demuestran que una elección adecuada del parámetro *plug-in* λ es $\lambda = 2\sigma_\xi^2 \sqrt{2n1n(pn)}$ donde la varianza del ruido σ_ξ^2 puede calcularse de forma iterativa; p. ej., inicializando el proceso a través del cómputo de la varianza de los datos originales, σ_y^2 y procediendo de forma recursiva. Otra posibilidad es aplicar el procedimiento denominado *Root Lasso* propuesto por Belloni, Chernozhukov y Wang (2011), donde la función anterior de pérdidas pasa a ser:

$$\min_b \left[\sqrt{\sum_{i=1}^{n} \frac{\left(y_1 - \Sigma_{j=1}^{p} b_j x_{ij}\right)^2}{n}} + \frac{\lambda}{n} \sum_{j=1}^{p} |b_j| \right], \qquad [13]$$

de manera que el criterio a minimizar se convierte en pivotal respecto a σ_ξ^2, con $\lambda = \sqrt{2n1n(pn)}$. Cuando cualquiera de estos criterios se aplican para regularizar la regresión [10] bajo los dos supuestos anteriores (parsimonia en los parámetros e isometra en los regresores) se obtienen los siguientes resultados (véase Belloni *et al.,* 2014a y b) y Chernovzhukov, Hansen y Spindler (2015) para los detalles):

- Lasso y Root Lasso identifican modelos de tamaño óptimo s_n (véase la definición de *Parsimonia* arriba) que, en el caso de $|\phi_j| \leq A_j^{-a}$, resulta ser $s_n = n^{\frac{1}{2a}}$,
- El uso de ambos procedimientos como primera etapa de PSD en la regresión [10] de nuevo identifica modelos de tamaño óptimo que, en el caso de $|\phi_j| \leq A_j^{-a}$, resulta ser $s_n = \sqrt{\frac{s}{n}} \log(pn)$.
- En este último caso, se verifica que $\hat{\sigma}_n^{-1} \sqrt{\hat{\alpha} - \alpha} \rightsquigarrow N(0,1)$, donde σ_n es la fórmula convencional del estimador MCO de α en [6] y "$\rightsquigarrow$" denota convergencia débil en distribución.

Para ilustrar la utilidad de estos resultados, resulta conveniente considerar un PGD con controles exógenos similar al simulado previamente para $p = 1$, pero esta vez con $p \gg n$ y

donde se cumplen los dos supuestos clave señalados antes. En concreto, siguiendo a Belloni *et al.* (2014), consideremos la siguiente generalización del PGD anterior, ahora con un gran número de controles:

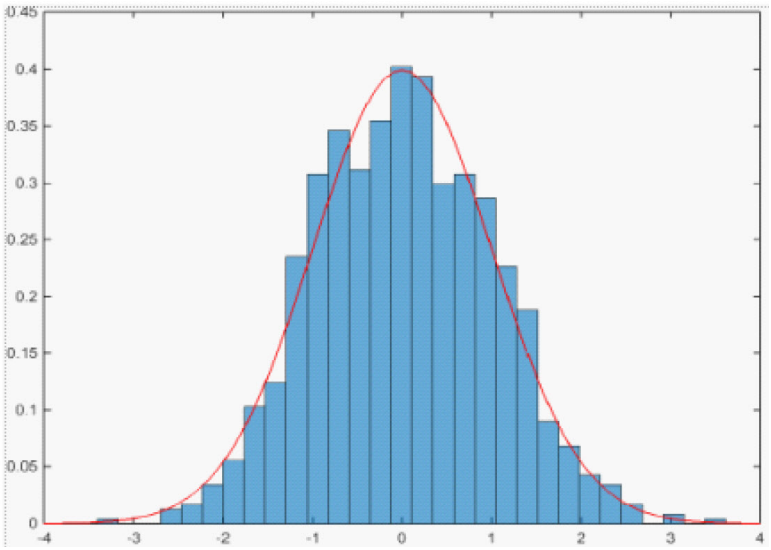
$$y_i = \alpha d_i + \sum_{j=1}^{p} \beta_i x_{ij} + \varepsilon_i,$$

$$d_i = \sum_{j=1}^{p} \gamma_i x_{ij} + \nu_i, \qquad [14]$$

donde $p = 200$, $n = 100$, $\alpha = 0$, $|\beta_j|$ y $|\gamma_j| = O\left(\frac{1}{j^2}\right)$, $\varepsilon_i \sim N(0,1)$, $\nu_i \sim N(0,0,32)$, $\mathbb{E}(\varepsilon_i \nu_i) = 0$, y x $\sim N(0,\Omega)$ con $\Omega_{kj} = (0,5)^{|j-k|}$. Utilizando 1.000 simulaciones, el resultado de contrastar $H_0 : \alpha = 0$, tras regularizar los parámetros β_j y γ_j mediante Lasso y aplicar PSD con un nivel nominal del 5 %, es una tasa de rechazo del 5,3 %, muy cercana al nivel nominal del contraste. Por contra, cuando se usa PS para regularizar únicamente la forma reducida de y_i, el rechazo efectivo es del 47 % (figura 2), corroborando de esta forma los resultados obtenidos para $p = 1$.

Figura 2.

(Post Selección Doble (PSD): Lasso al 5%)



Fuente: Elaboraciòn propia.

6. REGULARIZACIÓN EN MODELOS NO LINEALES

Si bien los resultados anteriores se cumplen para modelos lineales, su generalización a modelos no lineales no resulta excesivamente complicada. La lógica subyacente es similar a la

anterior. No obstante, ahora aparecen algunos nuevos resultados que permiten relajar alguno de los supuestos previos, como son las restricciones inherentes al supuesto de parsimonia en el número de parámetros verdaderos (la dependencia entre p y n). Como ejemplo ilustrativo en esta sección, usaremos un modelo lineal parcial (Robinson, 1988). En este modelo existe un parámetro de interés, α_0, asociado a un tratamiento o al cambio en una determinada política, d, esta vez en presencia de un gran número de controles, x, cuya forma funcional sea desconocida (en vez de ser lineal) y potencialmente complicada, tanto en la primera como en la segunda ecuación. En concreto, el DGP ilustrativo en este caso es el siguiente

$$\begin{aligned} y &= \alpha_0 d + g_0(\mathrm{x}) + \varepsilon, \\ d &= m_0(\mathrm{x}) + v, \end{aligned} \qquad [15]$$

donde x es un vector (n x p) de controles (con $p \simeq n$ o $p \gg n$) cuya inclusión es necesaria para que se cumpla $\mathbb{E}(\varepsilon|d,\mathrm{x}) = \mathbb{E}(v|\mathrm{x}) = 0$. El subindice "0" en los parámetros anteriores indica su verdadero valor y, en consonancia con los estudios observacionales, se supone que $m_0(\mathrm{x}) \neq 0$. De forma similar, se denota como $l_0(x)$ a la verdadera función de x que predice y en la forma reducida.

Un procedimiento habitual para estimar α_0 consiste en utilizar un método de regularización de ML (Lasso o cualquier otro entre los mencionados) de forma iterativa. Por ejemplo, usando un estimador incial de $\alpha_0^{(0)}$, se puede computar el residuo $\left(y - \hat{\alpha}_0^{(0)} d\right)$ y utilizar técnicas de ML en la regresión de dicho residuo sobre x para estimar $g_0^{(1)}(\mathrm{x})$ de forma no paramétrica A continuación se computa un nuevo residuo $\left(y - g_0^{(1)}(\mathrm{x})\right)$ que se regresa sobre d, obteniendo otro estimador $\hat{\alpha}_0^{(1)}$, y así sucesivamente hasta que el procedimiento iterativo converja. Este método resulta similar a la aplicación de PS en el modelo lineal que, como vimos, funciona mal. De nuevo, la intuición es que ML produce excelentes predicciones pero aumenta el SVO que es lo que importa a la hora de estimar la primera ecuación en [15].

La alternativa consiste en aplicar PSD de la siguiente manera. Primero, se utiliza ML para predecir d e y dado x, y de esta manera estimar las esperanzas condicionales $E(y|\mathrm{x}) = l_o(\mathrm{x})$ y $E(d|\mathrm{x}) = m_o(\mathrm{x})$. A continuación, se obtienen los residuos $\hat{\varepsilon} = y - \widehat{E(y/x)}$ y $\hat{v} = d - \widehat{E(d/x)}$. Finalmente, se regresa $\hat{\varepsilon}$ sobre $\hat{v}$ por MCO para obtener $\hat{\theta}$. Al igual que en el modelo lineal, el procedimiento PSD funciona correctamente, siendo parecida la intuición de por qué lo hace. No obstante aparecen nuevos resultados procedentes de la no linealidad que tienen interés. Para entender los argumentos en este caso, conviene analizar las condiciones de momentos que subyacen a los dos procedimientos anteriores (PS y PSD), las cuales vienen dadas por:

$$\mathbb{E}\left[\left(y - \alpha_0 d - g_0(\mathrm{x})\right)d\right] = 0 \text{(PS)},$$
$$\mathbb{E}\left[\left(y - E(y|\mathrm{x}) - \left(d - \left(E(d|z)\alpha_0\right)\right)\left(d - E(d|\mathrm{x})\right)\right] = 0 \text{(PSD)}.$$

En el caso de aplicar PS, la primera condición de momentos implica que:

$$\hat{\alpha}_{PS} = \left(\sum_{i=1}^{n} \frac{d_i^2}{n}\right)^{-1} \sum_{i=1}^{n} \frac{d_i\left(y_i - \hat{g}_0(x)\right)}{\sqrt{n}}$$

es decir:

$$\sqrt{n}(\hat{\alpha}_{PS}-\alpha_0)=\left(\sum_{i=1}^{n}\frac{d_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{d_i\varepsilon_i}{\sqrt{n}}+\left(\sum_{i=1}^{n}\frac{d_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{d_i\left(g_0(x)-\widehat{g_0}(x)\right)}{\sqrt{n}}:=(A)+(B). \quad [16]$$

Una vez más, mientras que el término (A) converge asintóticamente a una distribución $N(0, E(d^2)^{-1})$, el término (B) difiere de 0. De hecho, (B) diverge ya que la tasa de convergencia de los estimadores no paramétricos resulta ser más lenta que la de los paramétricos: $n^{-\varphi}$ con $0{,}25 < \varphi < 0{,}5$, en vez de $n^{-0{,}5}$. Por tanto, en el límite, dicho término equivale a:

$$(B)\approx\sqrt{n}n^{-\varphi}\to\infty,$$

con lo que, al igual que en la regresión lineal, la aplicación de PS a este modelo pueda acarrear un SVO muy elevado. Por contra, la condición de momentos para PSD implica que:

$$\hat{\alpha}_{PDS}=\left(\sum_{i=1}^{n}\frac{\hat{v}_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{\hat{\varepsilon}_i\hat{v}_i}{\sqrt{n}},$$

tal que $\hat{\varepsilon}=y-\hat{l}_0(\mathrm{x})$ y $\hat{v}=d-\hat{m}_0(\mathrm{x})$ Por consiguiente, dado que $\hat{\varepsilon}=\varepsilon-(\varepsilon-\hat{\varepsilon})$ y $\hat{v}=v-(v-\hat{v})$ se obtiene:

$$\begin{aligned}&\sqrt{n}(\hat{\alpha}_{PDS}-\alpha_0)=\\&\left(\sum_{i=1}^{n}\frac{v_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{\varepsilon_i v_i}{\sqrt{n}}+\left(\sum_{i=1}^{n}\frac{v_i^2}{n}\right)^{-1}\sum_{i=1}^{n}\frac{\left(l_0(\mathbf{x})-\hat{l}_0(x)\right)\left(m_0(\mathbf{x})-\hat{m}_0(\mathbf{x})\right)}{\sqrt{n}}+r(\mathbf{x}) \quad [17]\\&:=(A)+(B)+(C)\end{aligned}$$

donde de nuevo el término (A) es asintóticamente normal. En cuanto a (B), a diferencia de lo que ocurría al emplear PS, ahora sí que converge a 0 pues cada uno de los errores de predicción de d e y son de orden $n^{-\varphi_k}$ (k=l,m); por tanto:

$$(B)\approx\sqrt{n}n^{-(\varphi_m+\varphi_l)}\to 0.$$

Finalmente, para conseguir un buen comportamiento del estimador PSD queda por demostrar que el término (C) es asintóticamente despreciable, esto es, $o(1)$. Nótese que (C) captura un término residual, $r(\mathrm{x})$, que depende de los productos cruzados de ε y v con $(\varepsilon-\hat{\varepsilon})$ y $(v-\hat{v})$. Belloni *et al.* (2014b) demuestan que este resultado se puede alcanzar mediante la partición de la muestra total de n observaciones en al menos dos submuestras independientes donde, en una de ellas se usa ML para estimar $l_0(\mathrm{x})$ y $m_0(\mathrm{x})$ mientras que en la otra se regresa el residuo de predicción $\hat{\varepsilon}$ sobre $\hat{v}$, ambos obtenidos a partir de la primera submuestra. De hecho, dicho procedimiento permite relajar parcialmente el supuesto de *Parsimonia* necesario en la aplicación de PSD a modelos lineares con un gran número de controles.

Todos los resultados anteriores se pueden enmarcar de forma general en términos de la denominada *Condición de Ortogonalidad* de Neyman (1979) que pasamos a discutir brevemente. Por una parte, sea W la matriz de datos, θ_0 un conjunto de dimensión reducida con los

parámetros de interés, y $\eta_0 = (l_0, m_0)$ el vector de las verdaderas funciones predictivas de un conjunto p-dimensional de controles x sobre d e y (también sobre z si se requiriera un instrumento). Por otra parte, sea $\Psi\ (W, \theta_0, \eta_0)$ la función que captura las condiciones de momentos que permiten identificar θ_0 como la solución de la minimización de $\mathbb{E}\ (\Psi\ (W, \theta_0, \eta_0) = 0$.

La cuestión abordada por Neyman es cómo modificar el estimador de θ_0 de forma que su distribución asintótica no se vea afectada por pequeños cambios en η_0, dado que sus componentes son desconocidos y habrán de reemplazarse en la anterior condición de momentos por estimaciones no paramétricas de los mismos. Esta última propiedad se traduce intuitivamente en la siguiente condición de ortogonalidad:

$$\partial_\eta\, \mathbb{E}(\Psi(W, \theta_0, \eta)\, \downarrow_{\eta=\eta_0}=0, \qquad [18]$$

donde ∂ es la derivada parcial convencional cuando la función Ψ sea diferenciable (en términos más generales, se tomaría la derivada direccional o de Fréchet). En el caso del modelo lineal parcial, se puede comprobar fácilmente que la condición de momentos verifica la siguiente igualdad:

$$\mathbb{E}(\Psi(W, \theta_0, \eta_0) = \mathbb{E}[(y\text{-}l(x)\text{-}\theta_0(d\text{-}m(\mathrm{x}))(d\text{-}m(\mathrm{x})) = 0,$$

en la que, diferenciando respecto a los dos componentes de η y sustituyendo en $\eta = \eta_0$, se cumple [18]. Nótese que dicha condición de ortogonalidad es similar a la usada en el método $C(\alpha)$ de Neyman (1979) que permite realizar inferencia sobre la estimación de θ_0 con independencia de η_0, siempre y cuando Ψ se interprete como una funcion de verosimilitud.

Finalmente, con algunos supuestos adicionales, los argumentos esgrimidos anteriormente pueden generalizarse a modelos completamente no lineales en las variables pero aditivos en las perturbaciones, del tipo:

$$y = h_0\ (d,\mathrm{x}) + \varepsilon,$$

$$d = m_0\ (\mathrm{x}) + v, \qquad [19]$$

con $\mathbb{E}(\varepsilon \mid d, \mathrm{x}) = \mathbb{E}(v \mid \mathrm{x}) = 0$, donde $h_0(d, \mathrm{x})$ es una función desconocida de la variable de interés d y de los controles x, que puede estimarse con métodos semi o no paramétricos (véase Chernovzhukv *et al.,* 2020). Este modelo engloba a los discutidos previamente (inchuyendo a aquellos estimados por IV en los que basta añadir otra ecuación relacionando d con el instrumento z, como en [9]). Un caso particular de este tipo de modelos puede utilizarse para estimar el efecto medio de un tratamiento binario, $d_i = \{0, 1\}$, sobre la variable y. Puesto que d no es aditivamente separable de x, este modelo permite la existencia de efectos heterogéneos en el tratamiento. En este caso, la interpretación natural del parámetro θ_0 se corresponde con el efecto promedio del tratamiento (ATE), definido por:

$$\theta_0 = \mathbb{E}\ [h_0(1,\mathrm{x}) - h_o(0,\mathrm{x})],$$

el cual puede expresarse alternativamente (utilizando la distribución conjunta de d y x) como:

$$\theta_0 = \frac{d\,y}{\Pr(d=1|\,\mathrm{x})} - \frac{(1-d)y}{1-\Pr(d=1|\,\mathrm{x})},$$

donde $Pr\,(d{=}1 \mid \mathrm{x})$ es la probabilidad de recibir el tratamiento por parte de un individuo con características x, lo que en la literatura se conoce como *propensity score*. Combinando ambas definiciones de θ_0, se puede utilizar la siguiente condición de momentos:

$$\Psi(W,\theta,\eta) = h(1,x) - h(0,\mathrm{x}) + \frac{d\,(y - h(1,\mathrm{x}))}{m(\mathrm{x})} - \frac{(1-d)(y - h(0,\mathrm{x}))}{1-m(\mathrm{x})} - \theta,$$

que verifica la condición de ortogonalidad de Neyman con $\eta = (h(1,x),\ h(0,\mathrm{x}),\ m(\mathrm{x}))$. Chernozhukov *et al.* (2017) proponen un método de PSD consistente en utilizar una partición de la muestra de tamaño n en K submuestras, denominadas I_k (k = 1,2,...,K), con n/K observaciones cada una. Para cada I_k, se utilizan las observaciones de su conjunto complementario de submuestras restantes, denotado I_k^c para estimar los componentes del vector η por ML, mientras que I_k se usa para obtener $\hat{\theta}_0$ como la solución de la condición de momentos $n^{-1}\ \Sigma_{i\in I_k^c}\Psi\ (W,\theta,\ \eta(I_k^c))=0$. Finalmente, se construye el estimador promedio $\hat{\theta}_{0K} = K^{-1}\ \Sigma_{k=1}^{K}\theta_0\left(I_k, I_k^c\right)$. Bajo condiciones menos restrictivas de parsimonia e isometría, dicho estimador satisface el mismo resultado que se obtuvo para los modelos total o parcialmente lineales, esto es:

$$\sigma^{-1}\sqrt{n}\left(\hat{\theta}_{0K} - \theta_0\right) \rightsquigarrow N(0,1),$$

donde $\sigma^2 = \mathbb{E}[\Psi^2\ (\mathrm{w},\ \theta_0,\ \eta_0\ (\mathrm{x})]$.

Otro método que resulta más sencillo de implementar en el caso de tratamientos heterogéneos en contextos no experimentales es el denonominado *método de controles modificados (MCM)*. A través de una estrategia de identificación basada en la selección de observables, MCM parte de un caso particular de la función $h_0\ (d,\mathrm{x})$ en [19], consistente en el siguiente modelo con interacciones:

$$y_i = \beta'\,\mathrm{x}_i + d_i\,\delta'\,\mathrm{x}_i + \varepsilon_i, \qquad [20]$$

donde, para simplificar, se supone que el tratamiento (condicional en los observables x_i) se administra aleatoriamente a un 50 % de la muestra de individuos, de forma que $\mathbb{E}(\varepsilon_i \mid d_i, \mathrm{x}_i) = 0$. Por tanto el primer término de la derecha en [20] representa una aproximación lineal de la esperanza condicional de la variable y para los no participantes en el programa, $\mathbb{E}[y_{0i} \mid \mathrm{x}] = \beta'\,\mathrm{x}$, mientras que el segundo término proporciona otra aproximación lineal del ATE, $\mathbb{E}[y_{1i} - y_{0i} \mid \mathrm{x}] = \delta'\,\mathrm{x}$. Cuando p sea muy grande, el problema de regularización de los coeficientes se agudiza al duplicarse el conjunto de regresores. Tian *et al.* (2014) proponen abordar la regularización en dos etapas. Para ello, usan la transformación de la variable indicador d_i en la forma $T_i = 2d_i - 1$, de manera que el modelo de interacciones [20] pasa a ser:

$$y_i = \beta'\,\mathrm{x}_i + \frac{T_i}{2}\delta'\,x_i + \varepsilon_i.$$

Dado que $d_i \in \{0,1\}$, se cumple $T_i/2 \in \{-0{,}5,0{,}5\}$, por lo que $\mathbb{E}(T_i) = 0$. Esta modificación, que no altera el vector de coeficientes de interés δ, consigue que los dos conjuntos de regresores en [20] sean ortogonales, ya que $Cov\,(x_{ij}, T_i x_{ik}) = Cov\,(x_{ij}, x_{ik})\,\mathbb{E}(T_i) = 0$ para $j,k \in \{1,2,...,p\}$; nótese que la primera igualdad procede de la asignación aleatoria del tratamiento y la segunda de $\mathbb{E}(T_i) = 0$. Ello permite abordar la estimación de los coeficientes δ independientemente de los coeficientes β en un modelo más parsimonioso que [20], dado por :

$$y_i = \frac{T_i}{2}\delta' x_i + \varepsilon_i,$$

que constituye la regresión básica del MCM a la que se puede aplicar Lasso o cualquier otro procedimiento de ML para seleccionar aquellos controles que presenten efectos heterogéneos. Pese a haber utizado el supuesto de asignación aleatoria en aras a la simplificación, Chen *et al.* (2017) han demonstrado que la idea básica del MCM puede extenderse a estrategias de identificación no experimentales (en contextos observacionales) mediante su combinación con procedimientos de ponderación de probabilidad inversa (IPW).

7. APLICACIÓN EMPÍRICA

A continuación se proporciona una ilustración empírica de los procedimientos de regularización discutidos previamente mediante la aplicación de PSD en el modelo Angrist y Krueger (1991, AK en adelante) comentado en la sección segunda. Recordemos que estos autores encuentran relaciones estadísticamente significativas entre el trimestre del año en que uno nace, el nivel educativo y los ingresos para las cohortes de las décadas de 1920, 1930 y 1940 en EE. UU. Recordemos que los nacidos durante el primer trimestre del año obtienen menos educación y tienen menores ingresos que los nacidos durante los restantes trimestres del año ya que las regulaciones sobre asistencia escolar obligatoria en EE. UU. típicamente exigen que los estudiantes comiencen el primer grado en el otoño del año en que cumplen 6 años y que permanezcan en la escuela hasta que cumplan los 16 años. Por consiguiente, las personas nacidas a principos de año generalmente ingresan en el primer grado cuando tienen cerca de 7 años de edad y cumplen los 16 a mediados del décimo grado. Por contra, las personas nacidas en el tercer o cuarto trimestre generalmente comenzarán la escuela justo antes o justo después de cumplir 6 y terminarán el décimo grado antes de cumplir los 16. El modelo a estimar es el siguiente:

$$\begin{aligned} w_i &= \alpha s_i + \beta' \mathrm{x}_i + \varepsilon_i, \qquad & \mathbb{E}(\varepsilon_i \mid \mathrm{x}, \mathrm{z}) = 0, \\ s_i &= \gamma' \mathrm{z}_i + \phi' \mathrm{x}_i + v_i, \qquad & \mathbb{E}(v_i \mid \mathrm{x}, \mathrm{z}) = 0, \end{aligned} \qquad [21]$$

donde w_i es el logaritmo del salario del individuo i, s_i denota los trimestres de educación obligatoria, x_i es un vector de p controles, y z_i es un vector de m variables instrumentales ($m > 1$) que afectan a la educación pero no directamente al salario. Los datos proceden del censo de EE. UU. de 1980 y contienen observaciones para casi 330.000 hombres nacidos entre

1930 y 1939[6]. En concreto, z es un conjunto de 510 variables formado por: una constante, indicadores (*dummies*) de 9 años de nacimiento, 50 dummies del Estado donde nacieron, y 450 interacciones de los dos conjuntos anteriores de dummies. En concreto, AK usan los siguientes instrumentos: tres dummies del trimestre de nacimiento, sus dobles interacciones con el estado de procedencia y año de nacimiento y la triple interacción de todas ellas a la vez, es decir un total de 1.530 instrumentos potenciales. Se remite al lector interesado a AK (1991) para obtener el resto de detalles de su estimación. El coeficiente de interés es α, el cual recoge el impacto causal de la educación sobre los ingresos.

En la literatura se encuentan dos opciones básicas para estimar [21]: (i) usar como instrumentos solamente las tres dummies del trimestre de nacimiento o, alternativamente, (ii) utilizar 180 instrumentos resultantes de las tres dummies de trimestre de nacimiento y sus interacciones con las 9 de años de nacimiento y las 50 de estado de procedencia (excluyendo triples interacciones). Hansen, Hausman y Newey (2008) argumentan que el uso del conjunto de 180 instrumentos en estimaciones por MCB presentan un sesgo sustancial pero mayor precisión que cuando solo se usan tres instrumentos por el problema de "instrumentos débiles". Una posible solución a este problema es usar el estimador LIML de Fuller (1977), denominado FULL, consistente en corregir el estimador por MCB por sesgos de orden n^{-1}. Por ejemplo, suponiendo (para simplificar) que $\beta = \phi = 0$ en [21] , FULL proporcional el estimador $\hat{\alpha}_{FULL}$ dado por:

$$\hat{\alpha}_{FULL} = \arg\min_{\alpha} \frac{(\mathrm{x}-\alpha s)' Q_z (\mathrm{x}-\alpha s)}{(\mathrm{x}-\alpha s)'(\mathrm{x}-\alpha s)} \Rightarrow \hat{\alpha}_{FULL} = \left(s' Q_z s - \tilde{k} s' s\right)^{-1} \left(s' Q_z \mathrm{w} - \tilde{k} s' \mathrm{w}\right),$$

$$\tilde{k} = [\hat{k} - \left(1-\hat{k}\right)\frac{c}{n}][1-\left(1-\hat{k}\right)\frac{c}{n}], Q_z = z\left(z'z\right)^{-1} z',$$

tal que $c \geq 4$ es un parámetro a elegir y $\hat{k} = \hat{\varepsilon}' Q_z \hat{\varepsilon} / \hat{\varepsilon}' \hat{\varepsilon}$, donde $\hat{\varepsilon}$ son los residuos en el modelo [21].

El cuadro 1 presenta estimaciones de los rendimientos a la educación mediante MCB y FULL para diferentes conjuntos de instrumentos. Las tres primeras filas corresponden a las

Cuadro 1.

Estimación del rendimiento de la educación en AK (1991)

nº IVs	*MCB*	*FULL*
3	0.106 (0.019)	0.109 (0.021)
180	0.096 (0.010)	0.103 (0.014)
1.530	0.069 (0.005)	0.108 (0.042)
1(*)	0.087 (0.031)	---
12(**)	0.088 (0.014)	0.089 (0.014)

Nota: Estimadores MCB y FULL del parámetro α en el modelo de regresión [21]. Desviación típica entre paréntesis. (*) Lasso *plug-in,* (**) Lasso CV.
Fuente: Elaboración propia mediante los comandos *poivgress* y *lasso linear* de Stata.

[6] Los datos están disponibles en la web: https://economics.mit.edu/faculty/angrist/data1/data/angkru1991

agrupaciones naturales de los instrumentos comentados anteriormente de tamaño 3.180 y 1,5, respectivamente. Las dos últimas filas ofrecen los resultados basados en el uso de LASSO para seleccionar instrumentos con niveles de penalización dados por Lasso *plug-in* y CV en 10 subconjuntos de la muestra. De acuerdo con el primer valor de λ, Lasso únicamente selecciona la dummy de haber nacido en el cuarto trimestre, mientras que con CV elige 12 instrumentos entre los que se encuentran las dummies de haber nacido en tercer y cuarto trimestres. Todos los estimadores de α se obtienen utilizando PDS como técnica de regularización.

El primer resultado a destacar es que, con 180 o 1.530 instrumentos, existen algunas diferencias entre las estimaciones por MCB y FULL. Sin embargo, desaparecen al usar PSD con un número pequeño de instrumentos (1 o 10), lo cual indica que este procedimiento evita el sobreajuste en la primera etapa de IV. Además, aunque Lasso desconoce la relevancia de las dummies de trimestre de nacimiento entre los 180 o 1.530 instrumentos, siempre incluye alguna de ellas en el conjunto seleccionado (especialmente la variable dummy del cuarto trimestre, es decir la de los individuos más favorecidos por el tratamiento). Con este instrumento se estima el rendimiento anual de la escolarización en 0,087 con una desviación típica estimada de 0,031, mientras que con 180 y 1.530 se encuentra alrededor de 0,11 (MCB) y 0,10 (FULL). En general, estos resultados demuestran que la selección de instrumentos por PSDS es factible, produciendo estimaciones sensatas y comparables con las disponibles en esta literatura.

8. CONCLUSIONES

En este trabajo se ofrece una panorámica de los métodos de regularización disponibles en la literatura de ML que pueden ser útiles para abordar preguntas relevantes en economía laboral. Se destaca que los procedimientos de post selección doble (PSD) ofrecen resultados muy superiores a los de post selección (PS) sencilla. Estas ventajas se producen tanto en la especificación del conjunto de controles para estimar el efecto de un determinado tratamiento basado en variables observables cuando hay datos masivos en contextos no experimentales, como cuando se usa un gran número de variables instrumentales para tratamientos endógenos. Además, estos resultados son válidos para para modelos lineales y no lineales.

Referencias

Angrist, J. y Frandsen, B. (2020). Machine Labour. *NBER WP,* 26584.

Angrist, J. y Krueger, A. (1991). Does Compulsory School Attendance Affect Schooling and Earnings? *The Quarterly Journal of Economics,* 196, pp. 979-1014.

Athey, S. e Imbens, G. (2019). Machine Learning Methods Economists Should Know About. *Annual Reviews of Economics,* 11, pp. 685-725.

Belloni, A., Chen, D., Chernozhukov, V. y Hansen, CH. B. (2012). Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain. *Econometrica,* 80, pp. 2369–2429.

Belloni, A., Chen, D., Chernozhukov, V. y Hansen, CH. B. (2013). Inference for High-Dimensional Sparse Econometric Models. *Advances in Economics and Econometrics. 10th World Congress,* Vol. 3, edited by D. Acemoglu, M. Arellano y E. Dekel, pp. 245–295. Cambridge University Press.

Belloni, A., Chen, D., Chernozhukov, V. y Hansen, CH. B. (2014a).Inference on Structural and Treatment Effects with High-Dimensional Data. *Journal of Economic Perspectives,* 2014.

Belloni, A., Chen, D., Chernozhukov, V. y Hansen, CH. B. (2014b). Inference on Treatment Effects with High-Dimensional Controls, with Application to Abortion and Crime. *Review of Economic Studies,* 81, pp. 608-650.

Belloni, A., Chernozhukov, V. y Wang, L. (2011). Square-root Lasso: Pivotal Recovery of Space Signal via Conic Programming. *Biometrika,* 98, pp. 791-806.

Bickel, P. J., Ritov, Y. y Tsybakov, A. B. (2009). Simultaneous Analysis of Lasso and Dantzig Selector. *Annals of Statistics,* 37, pp. 1705 –1732.

Chen, S., Tian, L., Cai, T. y Yu, M. (2017). A General Statistical Framework for Subgroup Identification and Comparative Treatment Scoring. *Biometrics,* 73, pp. 1199-1209.

Chernozhukov, V., Chetverikov, D., Demiert, M., Dufflo, E., Hansen, CH. B. y Newey, W. (2017).Double/Debiased/Neyman Machine Learning of Treatment Effects. *American Economic Review P&P,* 107, pp. 261-265.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K. y Robins, J. (2020). *Locally Robust Semiparametric Estimation.* Mimeo.

Chernozhukov, V., Hansen, CH. B. y Spindler, M. (2015). Post-Selection and Post-Regularization Inference: An Elementary, General Approach. *Annual Review of Economics,* 7, pp. 649-688.

Fuller, W. A. (1977). Some Properties of a Modification of the Limited Information Estimator. *Econometrica,* 45, pp. 939-954.

Hansen, CH. B., Hausman, J. y Newey, W. K. (2008). Estimation with Many Instrumental Variables. *Journal of Business & Economic Statistics,* 26, pp. 398–422.

Hastie, T., Tibshirani, R. y Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York, NY: Springer.

Heckman, J., Lochner, L. y Todd, P. (2006). Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond. En: Eric Hanushek y Finis Welch (eds.), *Handbook of Education Economics,* Vol. 1, chapter 7. Elsevier.

Mullainathan, S. y Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. *Journal of Economic Perspectives,* 31, pp. 87-106.

Neyman, J. (1979). C(α) Tests and their Use. *Shankhya: The Indian Journal of Statistics,* 41, pp. 1-21.

Robinson, P. M. (1988). Root-N-Consistent Semiparametric Regression. *Econometrica*, 56, pp. 931–954.

Tian, L., Alizadeh, A. A., Gentles, A. J. y Tibshirani, R. (2014). A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates. *Journal of the American Statistical Association,* 109, pp. 1517-1532.