Grado en Análisis de Negocios (Business Analytics)

Trabajo de Final de Grado

Modelos Teóricos de economía basados en matemáticas y econométricos de análisis de datos

Author

Álvaro Lastra Aragoneses

Supervised by

Riccardo Ciacci

Madrid

June 2023

Modelos Teóricos de economía basados en matemáticas y econométricos de análisis de datos

Autor: Lastra Aragoneses, Álvaro

Director: Ciacci, Riccardo

Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

Palabras Clave: Método de los Mínimos cuadrados ordinarios, Variables instrumentales, econometría causal, álgebra, análisis estadístico

Resumen del Proyecto

La econometría causal es una rama de la econometría que tiene por objetivo encontrar las relaciones causales entre distintas variables econométricas. El desarrollo de buenos modelos para explicar fenómenos en econometría es esencial para explicar la efectividad de políticas públicas, toma de decisiones en negocios y fenómenos macroeconómicos. La cantidad de datos econométricos ha crecido de manera exponencial en los últimos veinte años, esto ha desembocado en una gran relevancia de los estudios centrados en econometría causal. La evolución de nuevos métodos estadísticos y el avance computacional también han ampliado del conocimiento en este campo.

La literatura utiliza varios modelos para encontrar relaciones causales, entre los que destacan mínimos cuadrados ordinarios (OLS) y variables instrumentales (IV). Estos métodos son usados en varios artículos para encontrar relaciones causales

entre las variables sociales y económicas.

Este trabajo profundiza en la literatura de la econometría causal para realizar un resumen de cómo son usados estos métodos, qué problemas resuelven y, sobretodo, cuál es el criterio de selección entre ambos métodos. Como veremos a lo largo del trabajo, sólo hay un artículo centrado en comparar estos dos métodos para encontrar la verdadera relación causal que esconden las variables econométricas.

Tras el resumen de la literatura, se ampliará el conocimiento para poder comparar formalmente estos dos estimadores. Se realizará una nueva fórmula matricial basada en la fórmula del artículo de Ciacci, para posteriormente aportar un test estadístico que permita crear un marco teórico con el que futuros investigadores puedan basar su criterio de selección.

Theoretical models of economics based on mathematics and econometrics of data analysis

Author: Lastra Aragoneses, Álvaro

Supervisor: Ciacci, Riccardo

Collaborating Entity: ICAI – Universidad Pontificia Comillas

Keywords: Method of ordinary least squares (OLS), Instrumental variables (IV), causal econometrics, algebra, statistical analysis

Abstract

Causal econometrics is a branch of econometrics that aims to find causal relationships between different econometric variables. The development of good models to explain phenomena in econometrics is essential to explain the effectiveness of public policies, business decision making and macroeconomic phenomena. The amount of econometric data has grown exponentially in the last twenty years, which has led to a great relevance of studies focused on causal econometrics. The evolution of new statistical methods and computational progress has also led to the expansion of knowledge in this field.

The literature uses several models to find causal relationships, among which ordinary least squares (OLS) and instrumental variables (IV) stand out. These methods are used in several articles to find causal relationships between social and economic variables.

This paper delves into the causal econometric literature to summarize how these methods are used, what problems they solve and, above all, what is the selection criterion between the two methods. As we will see throughout the paper, there is only one article focused on comparing these two methods to find the true causal relationship hidden by the econometric variables.

After the summary of the literature, the knowledge will be extended in order to formally compare these two estimators. A new matrix formula based on the formula in Ciacci's article will be developed, and then a statistical test will be provided to create a theoretical framework on which future researchers can base their selection criteria.

Contents

List of Figures

1 Introducción

La econometría causal es una rama de la econometría que se centra en el estudio de las relaciones causales entre variables econométricas. El desarrollo de buenos modelos de econometría causal es fundamental para explicar la efectividad de políticas públicas, toma de decisiones de negocio y grandes fenómenos económicos.

La econometría causal ha tomado relevancia desde finales de siglo XX cuando la cantidad de datos económicos empezó a crecer exponencialmente. La evolución de nuevos métodos estadísticos y el avance computacional también ha ampliado el conocimiento en este campo.

La literatura en este campo ha hecho uso de varios estimadores, entre los cuales se destacan el Estimador de Mínimos Cuadrados Ordinarios (OLS) y el Estimador de Variables Instrumentales (IV). Estos dos estimadores son ampliamente populares en trabajos del campo de la econometría para establecer relaciones causales entre variables económicas y sociales.

No obstante, a pesar de los progresos en la investigación, la literatura aún adolece de un enfoque riguroso y unificado que permita comparar efectiva y sistemáticamente estas dos metodologías. Hasta el momento, solo se ha desarrollado un artículo que aborda esta laguna metodológica, el cual ha sido elaborado por el tutor del presente trabajo de fin de grado. Dicho artículo[1], realizado por el profesor Ciacci, propone un marco teórico y empírico que provee una base sólida para el estudio comparativo de los estimadores OLS e IV en la econometría causal.

El propósito de este trabajo de fin de grado es ofrecer un resumen de la literatura

vigente en el campo de la econometría causal, enfocándose especialmente en la comparación entre los estimadores OLS e IV. Asimismo, se examinarán las implicaciones de las investigaciones actuales y se identificarán posibles áreas de estudio en el futuro en este dominio. Mediante un análisis riguroso de los principales artículos académicos, se espera arrojar luz sobre los desafíos y oportunidades que enfrenta la econometría causal en la búsqueda de métodos eficaces para establecer relaciones causales en el ámbito económico y social.

2 Contexto del estudio

2.1 Método de los mínimos cuadrados ordinarios (OLS)

Los mínimos cuadrados ordinarios (Ordinary Least Squares OLS) son un método estadístico empleado en el análisis de regresión lineal para estimar los parámetros poblacionales de un modelo. Este enfoque busca minimizar la suma de los cuadrados de las distancias verticales, denominados residuos, entre las respuestas observadas en la muestra y las respuestas generadas por el modelo. Los parámetros resultantes pueden expresarse mediante una fórmula polinómica sencilla, especialmente en el caso de un único regresor. Para la realización de un buen modelo se deben de tener en cuenta una serie de supuestos clave, como la exogeneidad de los regresores y la ausencia de multicolinealidad perfecta. El método OLS es consistente y óptimo entre los estimadores lineales cuando los errores son homocedásticos y no presentan autocorrelación. Además, si se asume que los errores siguen una distribución normal, el estimador OLS equivale al estimador de máxima verosimilitud.[2]

El principal problema en econometría causal surge al enfrentarse al problema de endogeneidad. Este problema surge cuando una variable está correlacionada con el término de error del modelo. En el caso de incluir una variable endógena en nuestro modelo, nuestro modelo puede estar sesgado y ser inconsistente. Esto quiere decir que nuestro coeficiente de la regresión de mínimos cuadrados ordinarios (OLS) va a estar sesgado. La endogeneidad puede surgir por diversos factores como error de medición, auto regresión con autocorrelación de errores, simultaneidad y variables

omitidas.[3]

Un ejemplo claro de endogeneidad es la variable precio al intentar predecir la demanda de un producto. Un cambio en la demanda del producto también cambia el precio, es decir, el cambio en el precio de un producto cambia la demanda y un cambio en la demanda cambia el precio. Este fenómeno que es muy frecuente en economía se denomina causalidad inversa. Para realizar un buen modelo que sea capaz de predecir la demanda debemos encontrar una variable exógena, como puede ser un cambio en el gusto o las necesidades de los consumidores.

Veamos el ejemplo de omitir una variable [3]. Supongamos que nuestro modelo viene dado por la siguiente ecuación:

$$y_i = \alpha + \beta x_i + \gamma z_i + u_i \tag{1}$$

En el caso de omitir la variable z_i nuestro modelo vendría dado por la ecuación:

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{2}$$

En la que el término de error es:

$$\varepsilon_i = \gamma z_i + u_i \tag{3}$$

Si la correlación entre x y z no es 0, y z está correlacionado con y lo que implica que $\gamma \neq 0$, entonces x estará correlacionado con el término de error. Por lo tanto,

x no es una variable exógena ya que no consigue explicar la variable dependiente a través de beta, porque y también depende de z y gamma.

Ahora veamos el caso en el que tenemos error de medición el los datos para realizar nuestro modelo de regresión. En este caso, no obtenemos una medida perfecta de nuestra variable independiente y, por lo tanto, tenemos un ruido de medición en cada una de nuestras muestras. Es decir, que en lugar de observar x_i observamos $x_i = x_i^* - v_i$ donde v_i es el ruido de medición. Por lo tanto, cuando intentamos realizar la regresión de nuestro modelo univariado:

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{4}$$

Realmente estamos realizando la siguiente regresión lineal:

$$y_i = \alpha + \beta(x_i^* - v_i) + \epsilon_i \tag{5}$$

$$y_i = \alpha + \beta x_i^* + \epsilon_i - \beta v_i \tag{6}$$

$$y_i = \alpha + \beta x_i + u_i \tag{7}$$

Donde u_i es igual a $\epsilon_i - \beta v_i$. Por lo tanto el término de error está correlacionado con nuestra variable independiente surgiendo el problema de endogeneidad.

2.2 Variables Instrumentales

Uno de los métodos más usados para intentar solventar los problemas de endogeneidad es usar variables instrumentales (IV). Una buena variable instrumental

permite obtener una estimación consistente cuando las variables explicativas se correlacionan con los términos de error. Una variable instrumental es una variable que no pertenece a la ecuación explicativa pero que se relaciona con variables explicativas endógenas [4]. Para que un instrumento sea adecuado se deben dar dos cualidades de este: el instrumento debe estar correlacionado con las variables explicativas endógenas y no puede estar correlacionado con el término de error. En caso de surgir el segundo problema estaríamos de nuevo ante el problema que intentamos resolver.

Si intentamos predecir una variable dependiente y a partir de una variable explicativa independiente de tratamiento x que es endógena, podemos usar la variable z como instrumento que, de manera ideal, afecta a y sólo a través de x.

El método computacional más conocido para el método de variables instrumentales es el de mínimos cuadrados en dos etapas (2SLS). Se denomina de esta manera ya que el proceso de regresión consta de dos etapas. En la primera fase, se realiza una regresión para predecir las variables endógenas del modelo con todas las variables exógenas de la ecuación de interés y los instrumentos excluidos. Finalmente, en esta fase se obtienen los valores predichos por esta regresión. Regresión de cada variable contenida en la matriz X en Z. Modelo: $X = X\delta + error$.

$$\hat{\delta} = (Z^T Z)^{-1} Z^T X \tag{8}$$

Y obtenemos los valores predichos por nuestras regresión:

$$\hat{X} = Z\hat{\delta} = Z(Z^T Z)^{-1} Z^T X \tag{9}$$

La segunda etapa corresponde a una regresión normal de la ecuación de interés, sin embargo, en vez de usar las variables independientes endógenas de X las sustituimos por las variables predichas por nuestro modelo de regresión anterior $\hat{X}$. Regresamos Y a partir de los valores predichos en la etapa anterior $\hat{X}$.

$$Y = \hat{X}\beta_{IV} + error \tag{10}$$

El estimador IV corresponde a β_{IV}.

3 Justificación del tema

La elección de este tema se debe a la necesidad de encontrar un método formal para comparar los estimadores IV y OLS. Esto ayuda a ampliar el conocimiento en econometría y ayuda a encontrar estimadores verdaderos para los fenómenos que se estudian que ayudarán a crear políticas económicas y sociales efectivas.

Los métodos OLS e IV ayudan a comprender los efectos de las variables sobre nuestra variable dependiente. La correcta interpretación de estos efectos es fundamental y el tener estimadores sesgados impide lograr este objetivo. La correcta interpretación de estos estimadores es fundamental sobretodo cuando tenemos problemas de endogeneidad, donde los estimadores dan resultados distintos dependiendo del método.

Debido a la carencia de una metodología que permita la comparación de ambos estimadores, la motivación de este trabajo es intentar resolver este problema o al menos dar una visión clarificada del estado del arte actual. Como veremos en la revisión de la literatura de este trabajo, la actual comparación que se realiza entre los estimadores IV y OLS es una mera comparación de tamaño, sin tener en cuenta otros datos como la selección de no observables y un estudio profundo de endogeneidad.

Como veremos a continuación, numerosos artículos usan estos modelos de regresión para encontrar relaciones causales en sus trabajos de econometría, pero ninguno de ellos muestra una argumentación clara para seleccionar el estimador verdadero. Consideramos fundamental arrojar luz sobre este campo del conocimiento para

obtener relaciones causales que expliquen fenómenos de la manera más fiel posible.

4 Objetivos

Actualmente no existe una metodología formal con la que permita comparar y seleccionar los estimadores obtenidos a partir de OLS e IV. Este trabajo contribuirá a la literatura existente al proporcionar una visión más clara y unificada de estas dos metodologías y su aplicación en la econometría causal. Para ello, se va a realizar una revisión de la literatura existente sobre los principales artículos que usan OLS e IV.

Por otro lado, este trabajo tiene por objetivo extender el artículo de Ciacci[1]. Para ello vamos a desarrollar la ecuación de Ciacci en forma matricial para obtener una ecuación más completa y generalizable a otros problemas.

Otro objetivo de este trabajo de final de grado es realizar un test estadístico de la ecuación (5) del artículo de Ciacci [1], el cual nos permita desarrollar una mejor interpretación de la ecuación.

Finalmente, se desarrollará una pequeña guía metodológica que ayude a futuros investigadores interesados en comparar IV y OLS en sus trabajos. Mediante esa guía, podrá interpretar sus estimadores obtenidos por ambos métodos y ser capaces de decidir cuál es el estimador verdadero para su problema propuesto.

Como se puede observar en los objetivos descritos, este trabajo va a consistir en investigar a fondo el marco teórico en este área de conocimiento de la econometría para después realizar una aportación teórica. Por lo tanto, el objetivo último

de este trabajo de final de grado es implementar las habilidades algebraicas y estadísticas para ampliar el conocimiento en la econometría causal de manera teórica.

5 Resumen de la literatura

En este apartado se va a realizar un resumen de la literatura de los artículos que utilizan tanto OLS como IV para realizar sus estudios de econometría causal. También hablaremos de aquellos artículos que exploran la correlación entre las variables observables y no observables que servirán de base para el artículo de Ciacci[1]. Veremos que en todos ellos, pese a hacer uso de ambos métodos, no existe una manera formal de comparar los dos resultados.

5.1 Using instrumental variables to establish causality

Como introducción al método de variables instrumentales se va a exponer el artículo de Sascha O. Becker[5], que hace uso de una variable instrumental para el estudio del efecto que tienen de los años de estudio sobre los ingresos en el futuro de otra persona. En este estudio podemos ver que el modelo hecho con OLS estima que los estudiantes ingresan hasta un 10% más por cada año extra de estudio, sin embargo, estos resultados pueden estar sesgados al alza ya que los estudiantes con más habilidad encuentran más fácil estudiar y pueden ganar más dinero más fácilmente sin la necesidad de estudiar un año extra. Por lo tanto, la correlación OLS incluye una prima por habilidad que sobreestima la causalidad. El modelo OLS no es informativo ya que no incluye la variable habilidad en su modelo, esto es debido a que la variable habilidad no es observada por el investigador y, por lo tanto, tenemos un sesgo por variable omitida.

Como los experimentos en educación son prácticamente imposibles ya que no podemos asignar aleatoriamente a diferentes personas un nivel de educación y ver sus

efectos en los ingresos, debemos hacer uso de experimentos naturales que alteran la decisión de estudio de un grupo de personas. Un ejemplo de experimento natural sería el cambio de ley de edad mínima de abandono escolar. Sascha O. Becker[5] se aprovecha del experimento natural que acaeció en 1947 con el cambio de ley en Reino Unido para la edad mínima de abandono escolar. Esta ley incrementó la edad mínima de 14 a 15 años afectando a todos los niños nacidos desde 1933 en adelante. Este cambio afecta a todos los estudiantes independientemente de su nivel de habilidad, es decir, no está correlacionado con la variable omitida y por lo tanto tampoco lo está con el error.

Con el cambio de edad mínima para abandono escolar muchos estudiantes serán forzados a permanecer un año más, mientras que alumnos con habilidades similares de años anteriores no serán forzados. Esta ley puede ser usada como instrumento ya que afecta a los ingresos sólo a través de los años de educación. El instrumento está correlacionado con los años de educación porque si te afecta la ley, cambia la probabilidad de los años estudiados, pero no está correlacionado con los ingresos de una persona. El nuevo modelo realizado con el método de variables instrumentales obtiene el resultado de que el efecto causal de estudiar un año extra de estudio sobre tus ingresos es de un 3%.

En resumen, tenemos una variable omitida (habilidad) que afecta tanto a nuestra variable de tratamiento (años de educación) y a nuestra variable dependiente (salario), por lo tanto, nuestro modelo sufre del sesgo de autoselección. Nuestro modelo también puede ser sesgado a cero si tenemos error de medición. Sin embargo, todos estos problemas pueden ser solucionados con un buen instrumento que sea relevante y cumpla con la restricción de exclusión. Un instrumento es rel-

evante si el instrumento guarda una fuerte correlación con la variable tratamiento. Si la restricción de exclusión es satisfecha, el instrumento sólo afecta a la variable dependiente a través de la variable de tratamiento.

El instrumento de Sascha O. Becker es relevante, esto se puede comprobar mediante un F-test en el first stage de la regresión IV. El instrumento afecta a los años de educación y lo hace de forma aleatoria, ya que es en base a la fecha de nacimiento. Además, guarda una fuerte correlación con la variable de tratamiento, podemos denominar nuestro instrumento entonces como "instrumento fuerte". Existen test para estudiar si nuestro instrumento es débil o fuerte, Sascha O. Becker usa uno de ellos para evidenciar la fortaleza del mismo. En el caso de realizar un modelo IV con un instrumento débil puede introducir sesgos más grandes que en modelo OLS.

Sin embargo, evidenciar la fuerte correlación del instrumento con la variable de tratamiento es relativamente sencillo, el verdadero reto es encontrar un instrumento que cumpla la segunda condición (restricción de exclusión), que es imposible testear estadísticamente porque no se puede tener en cuenta todas las variables que pueden afectar al modelo, se puede proporcionar evidencia empírica. Normalmente esta condición se suele apoyar en una buena narrativa que explique porque el instrumento sólo afecta a la variable dependiente a través de la variable tratamiento. En el caso de estudio de Sascha O. Becker se puede dar el contraargumento de que la subida en los salarios se debe a que los trabajadores de 14 años son malos sustitutos y por lo tanto los empresarios tuvieron que enfrentarse a una baja oferta de empleados. Debido a esto, los empresarios tuvieron que subir los salarios para intentar atraer a nuevos empleados. En definitiva, la validez del

instrumento depende del contexto de cada caso de estudio.

5.1.1 LATE

En muchas ocasiones nuestro modelo con IV no explica un efecto causal sobre toda la población, si no sobre un grupo específico de gente que es la afectada por el instrumento. En el modelo OLS se obtiene la media de la diferencia en salarios por estudiar un año más a cualquier nivel ya que est.

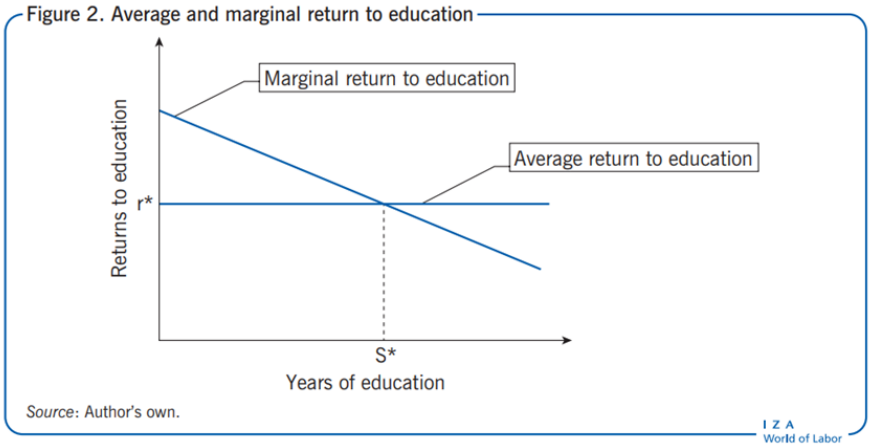

Figure 1: Average and Maginal return to education[5]

En cambio, en el caso de estudio de Sascha O. Becker el coeficiente representa sólo el incremento en el salario por estudiar un año más con 14 años, en econometría, la literatura indica que la ganancia marginal por estudiar un año más decrece con los años, como indica la Figura 1. Es decir, aprender a leer en tus primeros años de colegio tiene más impacto en tu salario que estudiar un año de doctorado. En resumen, puede que nuestro modelo IV tenga alta validez interna para el grupo a tratar (niños de 14 años estudiando un curso más), pero tenga baja validez externa ya que no aplique a otros grupos de población (persona que quiere estudiar un año de máster).

La dificultad de interpretar nuestro modelo IV como LATE es identificar que grupo específico es el afectado por nuestro instrumento. En el estudio de Sascha O. Becker[5] sobre la educación y los salarios, al estar ante un instrumento binario es más sencillo. Están los estudiantes que abandonan la educación a los 14 y los que abandonan más tarde. Existen también otras dos clases de personas, los que fueron afectados por la ley y los que no. Esto nos da un total de cuatro grupos de personas, los que siempre dejan la educación a los 14 aunque les afecte la ley ('never-takers'), los que siempre dejan la educación más tarde de los 14 independientemente de que les afecte la ley ('always-takers'), los que estudian forzados un año más por el cambio de ley ('compliers') y un último grupo que se les denomina 'defiers' porque estudiarían un año más con el modelo antiguo, pero con la nueva ley hacen lo contrario y abandonan los estudios antes de tiempo desafiando a la nueva ley. Para ser capaces de interpretar el IV como LATE se asume que nadie es 'defier' y, por lo tanto, como es lógico pensar, todo el mundo que estudiaban un año más con el modelo antiguo siguen cumpliendo la ley. Esta suposición se denomina monotonocidad. Con esta suposición se puede extraer el efecto que tiene la variable tratamiento sobre los 'compliers', es decir, sobre aquellos forzados a estudiar un año más, pero no se puede identificar el efecto que tiene un año más de estudio sobre los 'always-takers' y los 'never-takers'. Sólo se puede recuperar con IV el efecto medio de la variable tratamiento en toda la población si los grupos 'always-takers' y 'never-takers' fueran negligibles. Un aspecto positivo, es que los legisladores pueden aprovecharse de IV para ver qué efecto ha tenido sus leyes sobre la población.

En definitiva, nuestros modelos IV tienen la limitación de sólo explicar el efecto

causal de un grupo específico de la población. Para esta limitación el autor Sascha O. Becker[5] propone la solución de usar varios instrumentos que identifiquen a varios subgrupos de la población. Propone un primer instrumento que indique si un niño que ha estudiado durante la segunda guerra mundial tiene un padre que participó de la guerra y un segundo instrumento que indique la educación del padre. El primer instrumento afecta negativamente a la educación del niño, mientras que el segundo instrumento afecta positivamente ya que padres inteligentes ayudan a sus hijos a ser más inteligentes. El fin de usar estos dos instrumentos es acaparar todo el espectro de la población, el primer instrumento recupera el efecto sobre la población pobre que fue afectada negativamente por la ausencia del padre debido a la guerra, mientras que el segundo instrumento recupera los efectos de la población con más ingresos que son la población con padres con más educación. Los resultados dicen que se aumentan los ingresos un 4,8% por cada año de educación del padre y un 14% por cada año de estudio de un hijo con padre en la guerra. Esto muestra claramente la gran heterogeneidad que existe en educación y que ya adelantábamos en la figura 2.

Desde un punto de vista más técnico, IV tiene otras limitaciones ya que nuestro modelo es consistente pero sesgado. La consistencia implica que a medida que nuestras muestras crecen nuestro modelo es capaz de converger hacia el verdadero efecto causal. No sesgado implica que, aunque tengamos tamaños muestrales pequeños, de media, somos capaces de recuperar siempre el verdadero efecto causal. Nuestros modelos IV con tamaños muestrales pequeños son muy poco probables de recuperar el verdadero efecto causal y, por ello, están sesgados. Sin embargo, según crece el tamaño muestral nuestra predicción está más cerca del verdadero efecto

causal.

5.2 The Colonial Origins of Comparative Development: An Empirical Investigation

En el artículo escrito por Daron Acemoglu, Simon Johnson, y James A. Robinson (ASJ)[6] ("The Colonial Origins of Comparative Development: An Empirical Investigation") tiene por objetivo ver de forma empírica el efecto de las instituciones coloniales sobre el desarrollo económico de un país. Para ello hace uso de variables instrumentales, y utiliza como instrumento la mortalidad de los colonizadores europeos para modelar las instituciones. La intuición detrás de este estudio es que mejores instituciones llevan a un país a invertir más en capital humano y esto repercute en la bonanza económica de un país. Los autores evidencian de forma lógica debido a sucesos históricos, como con divisiones artificiales de países como Korea y Alemania tras la segunda guerra mundial, que las instituciones afectan al desarrollo económico de un país.

Aunque este efecto es obvio, se busca obtener un estimador óptimo que cuantifique el verdadero efecto de las instituciones sobre la economía de un país. Sin embargo, nos encontramos con problemas de endogeneidad como que los países más ricos se pueden permitirse instituciones mejores, lo que da lugar a causalidad inversa ya que la variable dependiente afecta a la variable tratamiento. Para realizar una buena estimación del efecto de las instituciones debemos aislar el efecto de las instituciones encontrando un instrumento exógeno que sea capaz de predecir nuestra varible de tratamiento.

Los ASJ proponen en su estudio el uso de la mortalidad de los primeros colonizadores como instrumento para estimar las instituciones actuales. Como hemos mencionado en el anterior estudio de Sasha O. Becker, la narrativa para explicar por qué este instrumento es válido y cumple con la restricción de exclusividad es muy importante. Para ello, ASJ desarrolla la lógica expuesta en la Figura 2 para interpretar el efecto de la mortalidad (variable instrumental) sobre las instituciones actuales (variable de tratamiento), estando esta última correlacionada con el desempeño económico actual (variable dependiente).

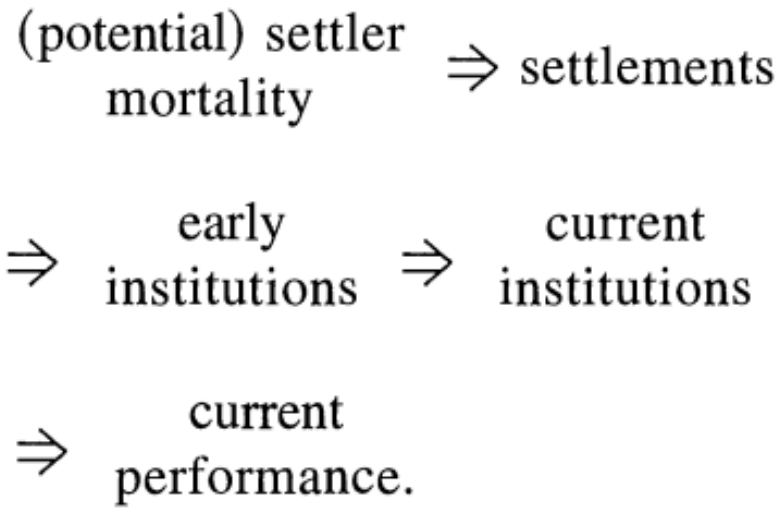

Figure 2: Correlación entre los instrumentos con la variable tratamiento[6]

ASJ dicen que hay dos tipos de colonias; las extractivas, cuyo objetivo era extraer la máxima riqueza para enriquecer al colonizador, y las Neo-Europeas, en las que el colonizador intenta replicar un estado europeo mediante instituciones en el país colonizado, es el ejemplo de Australia, Nueva Zelanda o Estados Unidos. ASJ dice que esta decisión se hacía en parte debido a la factibilidad de los asentamientos europeos. En países donde la mortalidad europea fuera alta, debido a múltiples

factores como enfermedades, era más improbable que los colonizadores decidiesen crear un estado Neo-Europeo. Estas instituciones creadas durante la colonización aún perduran hasta nuestros días y, por lo tanto, modelan las instituciones actuales, que estas a su vez afectan al desempeño económico actual de las naciones.

Por otro lado, la mortalidad de hace cien años no tiene ningún efecto sobre la economía actual, salvo a través de las instituciones, por lo tanto, cumple la restricción de exclusividad. La mayor preocupación para que se cumpla la restricción de exclusividad sería que las tasas de mortalidad pasadas estén correlacionadas con las enfermedades actuales, que puede afectar al desempeño económico actual. Sin embargo, esta correlación es poco plausible ya que la mayoría de los colonos fallecieron debido a la malaria y a la fiebre amarilla, enfermedades que tienen muy poco impacto actualmente y que los indígenas actuales han desarrollado inmunidad.

Una vez explicada la narrativa del instrumento, se toma como variable instrumental la mortalidad de curas, soldados y navegantes de las colonias. Para representar a las instituciones de forma cuantitativa se toma el valor de Average Expropiation Risk. Para la variable independiente se toman los valores de PIB per cápita del país.

ASJ arroja datos de su estudio muy interesantes como por ejemplo que “Africa es más pobre que el resto del mundo, no por razones geográficas si no porque tiene peores instuciones.” (p.5). Esta conclusión la obtienen al incluir en el modelo variables como la distancia al ecuador y la variable dummy África, observando que ninguna de las dos es significativa.

ASJ menciona que sus resultados son válidos cumpliendo con la restricción de exclusividad ya que al incluir variables de control a su modelo los resultados de los estimadores no cambian.

El primer modelo que desarrollan ASJ arroja resultados bastante interesantes. El modelo simple usando sólo el estimador para modelar el PIB per cápita nos muestra una fuerte correlación entre las instituciones y el PIB. Atendiendo a R cuadrado más del 50% de los movimientos en el PIB per cápita son explicados por nuestras medidas de las instituciones. El modelo con las variables dummy de continente y la latitud mejora el modelo anterior y todas las variables son significativas. Destaca el hecho de que ser un país africano o asiático tiene un impacto negativo en el PIB del país. Sin embargo, pese a las correlaciones fuertes que existen en este modelo, ASJ advierte de no interpretar estos estimadores como los verdaderos efectos causales debido a la posibilidad de causalidad inversa como explicamos anteriormente. Por otro lado, también advierte de la alta probabilidad de existencia de variables omitidas que se correlacionen de manera natural con las instituciones, dando lugar a una estimación sesgada. Otros problemas que puede tener el modelo es que la medida de las instituciones está hecha ex post, lo que puede llevar a ver mejores instituciones en países ricos. También el artículo advierte del problema de error de medición, puede que la medida usada para representar las instituciones no sea representativa del verdadero conjunto de las instituciones que tienen impacto sobre la economía. Esto quiere decir que nuestro estimador estaría sesgado a la baja. Todos estos problemas se resuelven si encontramos un buen instrumento.

Para demostrar que la mortalidad está correlacionada con las instituciones ac-

tuales y que esta correlación es fuerte, ASJ realiza varias regresiones. Realiza una regresión entre la mortalidad y el asentamiento europeo (porcentaje de población europea en 1900), que muestra que alrededor del 50% del asentamiento viene dado por la mortalidad. Realiza la regresión de asentamiento europeo con la medida de democracia en 1900 para verificar la causalidad entre los asentamientos europeos y las instituciones pasadas. Esta causalidad también la explica con la regresión de asentamientos europeos sobre la restricción al ejecutivo. Todas las variables mencionadas en este párrafo son además usadas para modelar una regresión OLS sobre la variable Average Expropiation Risk que representa las instituciones actuales. Mediante todas estas regresiones el objetivo del autor es demostrar su narrativa de que existe una correlación fuerte entre la mortalidad y las instituciones actuales, así como cual es el proceso que ha llevado a relacionar estas dos variables (Mortalidad, asentamiento europeo, instituciones pasadas, instituciones actuales).

Una vez demostrada la correlación, ASJ desarrolla su modelo final basado en mínimos cuadrados en dos etapas (2SLS). Para ello, primero realiza el modelo de regresión de la mortalidad (Log Mortalidad para normalizar los valores atípicos) sobre las instituciones actuales (Average Expropiation Risk). Las predicciones de este modelo son usadas para realizar el modelo OLS sobre el desempeño económico actual (Log GDP). En el modelo más simple que solo incluye la variable tratamiento la correlación entre las instituciones actuales y el PIB per cápita es de 0.94, aproximadamente el doble que el estimador formado mediante la regresión OLS (0.52). Esto nos indica que probablemente tengamos error de medición y en el estimador OLS tengamos sesgo a la baja. El número de instituciones que afecta

a la ejecución económica es muy grande y, por lo tanto, una sola variable para capturar el verdadero efecto no es suficiente. Otros resultados interesantes que arroja el estudio es que la variable de control latitud ya no es significativa para el modelo. Es posible que esta variable estuviese relacionada con las instituciones o alguna otra variable exógena. El estimador calculado mediante 2SLS muestra que es bastante consistente a cambios en el tamaño de la muestra. Al excluir valores atípicos como son los de Canada, Estados Unidos o Nueva Zelanda vemos que nuestro estimador prácticamente no cambia (1.21). El estudio de ASJ también realiza el modelo 2SLS incluyendo dummies continentales, y los resultados indican que los estimadores de estas dummies son insignificantes como adelantábamos al principio. Además, nuestro estimador del efecto de las instituciones permanece constante. Tras este modelo parece evidente la fuerte correlación entre las instituciones y el desempeño económico, y que ahora las variables de control al usar el instrumento mortalidad no tienen significancia.

Durante el resto del artículo los autores investigan la robustez del estimador. Para ello ASJ hace una resvisión de la literatura para comprobar las distintas variables que diversos autores consideran que tienen una relación causal sobre las instituciones actuales. Para ello comienza estudiando si la identidad del colonizador influye en el desempeño económico, tal y como sugiere La Porta et al. (1999). ASJ hace uso de las dummies para colonias británicas y francesas y la dummy omitida para el resto. Los resultados indican que no son variables significativas y que el estimador no cambia. Von Hayek (1960) y La Porta et al. (1999) tambiés sugieren que el origen legal influye, sin embargo, al añadir la dummy de origen legal francés (origen legal británico omitida) se puede ver que no son significativas

y el estimador no cambia. Cambiando el tamaño muestral para incluir diferentes grupos religiosos tampoco tiene impacto sobre nuestro modelo, contraargumentando a Max Weber. ASJ también incluye variables naturales al modelo como la temperatura o la humedad, con los que de nuevo nos encontramos que no tiene influencia sobre el modelo. Unas últimas variables de control como la malaria son añadidas en el modelo, pero ninguna de ellas es significativa. Otro resultado significativo del estudio es que también construye un modelo 2SLS con el instrumento de la malaria, que, aunque tiene menos variabilidad que la mortalidad de los colonizadores es evidente que no está relacionado con el desempeño económico actual ya que la fiebre amarilla está actualmente erradicada. El estimador resultante usando la fiebre amarilla como instrumento es 0.91, muy similar al estimador usando la mortalidad como instrumento.

Tras este estudio de la robustez de nuestro estimador ASJ sugiere a que estamos ante un estimador muy válido para nuestro modelo. Finalmente, para cerrar su artículo ASJ realiza una prueba de sobre identificación. Mediante esta prueba de chi cuadrado, ASJ muestra que ninguno de los modelos rechaza la hipótesis nula de que los estimadores del modelo 2SLS normal y el estimador 2SLS pero incluyendo en la segunda etapa la variable mortalidad como exógena son iguales. Por lo tanto, los resultados no muestran evidencia de que la mortalidad y el PIB tenga una relación directa salvo a través de la variable de tratamiento.

En conclusión, en el artículo "The Colonial Origins of Colonial Development", se sostiene que las disparidades en instituciones y políticas estatales constituyen la razón fundamental detrás de las diferencias en ingresos per cápita entre países. Los autores proponen que las variaciones en las experiencias coloniales es un he-

cho histórico que puede considerarse como una fuente de diferencias exógenas en instituciones. Se establecen tres argumentos principales para explicar este experimento natural: (1) los colonizadores europeos implementaron estrategias de colonización muy diversas con instituciones asociadas distintas; (2) la estrategia de colonización fue influenciada parcialmente por la viabilidad del asentamiento europeo, con fuerte correlación con la mortalidad, ya que, en zonas con elevadas tasas, los europeos no podían establecerse y, por ende, eran más propensos a instaurar estados extractivos; (3) las instituciones iniciales han perdurado hasta la actualidad. Los investigadores emplean dichas diferencias como fuente de variación exógena para estimar el impacto de las instituciones en el rendimiento económico. Toda esta narrativa se sustenta gracias a que se observa una correlación significativa entre las tasas de mortalidad de los colonos, los asentamientos europeos, las instituciones tempranas y las contemporáneas. Se puede concluir entonces que la mejora de las instituciones podría generar beneficios económicos considerables. Sin embargo, el estudio no aborda cuestiones relacionadas con los pasos concretos para mejorar dichas instituciones, dando lugar a futuras investigaciones en este ámbito. En general, se resalta la importancia del uso de técnicas apropiadas para afrontar los problemas de endogeneidad, así como la aplicación de múltiples niveles de tratamiento en modelos econométricos complejos.

5.3 Salvaging Falsified Instrumental Variable Models

El artículo "Salvaging Falsified Instrumental Variable Models"[7] escrito por los investigadores Matthew A. Masten y Alexandre Poirier, publicado en 2020, habla sobre el problema de los modelos con variables instrumentales falsificados y, además,

propone un método para recuperar los estimadores verdaderos de este modelo. Se refiere al concepto de modelos falsificados como aquellos modelos que hacen uso de variables instrumentales no válidas que no proporcionan relaciones causales reales. Las variables instrumentales no son válidas cuando presentan problemas de endogeneidad y de selección de variables.

Ante este desafío, Masten y Porier proponen un método innovador que busca recuperar estimadores válidos en modelos falsificados mediante el uso del método de mínimos cuadrados. Su enfoque novedoso implica el empleo de variables instrumentales falsificadas y variables exógenas para maximizar la varianza explicada de la variable endógena, brindando así una herramienta efectiva para abordar la endogeneidad en la estimación de modelos económicos.

Para respaldar su propuesta teórica, los autores complementan su trabajo con un estudio empírico en el cual aplican su metodología. Utilizan datos obtenidos de la Encuesta Nacional de Examen de Salud y Nutrición (NHANES) con el objetivo de investigar la relación causal entre la obesidad y la presión arterial. Mediante la aplicación de su método, logran analizar y cuantificar de manera más precisa dicha relación, ofreciendo una contribución significativa al campo de la econometría en términos de identificación de causas subyacentes y análisis de variables instrumentales.

En resumen, el trabajo propuesto presenta un enfoque novedoso y útil en situaciones donde no es posible identificar la causa subyacente de la endogeneidad, lo que nos obliga a trabajar con las variables instrumentales disponibles. La metodología propuesta por Masten y Porier ofrece una alternativa valiosa para abordar este de-

safío y proporciona nuevas herramientas para mejorar la calidad de la estimación en modelos econométricos.

5.4 Instrumental variables regression with weak instruments

En el trabajo "Instrumental variables regression with weak instruments"[8] escrito por los autores Douglas Staiger y James H. Stock, publicado en 1997, aborda el problema de estimación de modelos de regresión con variables instrumentales débiles. Las variables instrumentales débiles son definidas como aquellas que mantienen una baja correlación con la variable de tratamiento endógena. Estas variables tienen menor capacidad para eliminar el sesgo de endogeneidad en el modelo. El problema de eliminar la endogeneidad se acrecenta cuando el tamaño muestral es pequeño.

En el ámbito de la econometría causal, es fundamental encontrar un instrumento adecuado cuando nos enfrentamos a problemas de endogeneidad en nuestros estudios. Sin embargo, en muchas ocasiones nos encontramos con la dificultad de que los instrumentos propuestos no presentan una correlación fuerte con la variable de tratamiento, lo que puede conducir a resultados sesgados en las nuevas regresiones.

El artículo propone un método de estimación de la regresión de variables instrumentales. Para ello, hace uso de la máxima verosimilitud y corrección de la matriz de varianzas y covarianzas, de esta manera, se tiene en cuenta la debilidad de las variables instrumentales. Finalmente, el trabajo propone un test estadístico para probar la debilidad de las variables instrumentales. La propuesta teórica del

artículo va acompañada de un ejemplo con datos reales, realizan una regresión sobre los salarios. Los autores concluyen que usando máxima verosimilitud y corrección de la matriz de varianzas y covarianzas se obtiene un estadístico más veraz que usando otros métodos de estimación para variables instrumentales débiles.

En definitiva, el trabajo realizado por los autores Douglas Staiger y James H. Stock ofrece un método formal para abordar el desafío de las variables instrumentales débiles y resalta la importancia de realizar pruebas de validez para las variables instrumentales utilizadas.

5.5 Base teórica artículo Ciacci[1]: "Unobservable Selection and Coefficient Stability: Theory and Evidence"

Emiliy Oster en su artículo sobre la relación entre observables y no observables[9] ("Unobservable Selection and Coefficient Stability: Theory and Evidence"), extiende la teoría ya escrita por Altonji, Eder y Taber (2005)[10] y presenta un enfoque para evaluar lo robustez de los resultados obtenidos en estudios empíricos frente al sesgo por variables omitidas. La metodología de Oster (2019)[9] utiliza información de la regresión OLS, como la inclusión de controles, tamaño de varianzas y movimiento de R cuadrado, para estimar un conjunto de valores en los que debería encontrarse el verdadero efecto del tratamiento. Un enfoque común en la literatura previa a Oster es analizar cómo se mueven los coeficientes al incluir observables en el modelo OLS, como acabamos de ver en el artículo escrito por ASJ[6] sobre el efecto de las instituciones sobre el desempeño económico. Oster argumenta que para tener en cuenta el sesgo por variables omitidas es necesario tanto tener en cuenta los movimientos de los coeficientes como los movimientos

e R cuadrado. Para ello, Oster[9] desarrolla una estimación consistente del sesgo por variables omitidas haciendo uso de los coeficientes y R cuadrado. El método también indica que el investigador debe atender a las varianzas de la variable de tratamiento y de la variable dependiente.

Oster[9] en su artículo habla sobre la estabilidad de los coeficientes. Oster pone de ejemplo un modelo de regresión lineal en educación, en el que el modelo viene definido por la ecuación lineal descrita en 11.

$$Y = \beta X + W + C \tag{11}$$

En la ecuación 11, W y C son dos componentes ortogonales de habilidad y X es la educación. Menciona que la varianza de W es mucho mayor que la varianza de C, ambas variables se tienen la misma correlación con X. Pongámonos ahora en la situación en la que C es la variable observable y W es la no observable. Si incluimos C en nuestro modelo nuestro coeficiente será estable ya que nuestra variable es menos explicativa para la variable dependiente, pese a que el sesgo sea grande.

Para verlo mejor Oster supone que beta es igual a cero y construye la tabla 1 donde muestra las varianzas y el valor de R^2.

Table 1: Breakdown of Variable Maintenance Costs

	Coeficiente var. Controlada [R^2]	**Coeficiente Adicional [R^2]**
Control alta varianza observado	0.202 [0.04]	0.002 [0.990]
Control baja varianza observado	0.202 [0.04]	0.200 [0.013]

En la segunda fila de la tabla 1 podemos ver claramente como el coeficiente parece mucho más estable al incluir la variable con baja varianza, sin embargo, el verdadero efecto es 0 en ambos casos. La clave en ambas filas es atender a los movimientos de R cuadrado para poder valorar la calidad de nuestra variable. La variable con baja varianza cambia poco el coeficiente, pero también cambia poco R cuadrado.

La conclusión que obtiene Oster a partir de esta intuición es que el sesgo por variables omitidas es proporcional a los movimientos de los coeficientes, pero sólo si estos son escalados a los movimientos de R cuadrado. Oster añade que esta consideración es ignorada por lo general en la literatura ya que de 27 artículos seleccionados sólo dos prestan atención a los movimientos de R cuadrado, en el resto, como en el artículo de ASJ[6], se pasa por alto.

5.5.1 Construcción del estimador

Para esta sección del trabajo se va a hacer uso de el trabajo "Comments on 'Unobservable Selection and Coefficient Stability: Theory and Evidence' and 'Poorly Measured Confounders are More Useful on the Left Than on the Right" [11] escrito por los autores Giuseppe De Luca, Jan R. Magnus y Franco Peracchi, publicado en 2018. Este trabajo hace un resumen de los dos trabajos mencionados en el título y nos proporciona una visión resumida del trabajo de Oster que nos ayuda a desarrollar el marco teórico de esta sección.

El trabajo realizado por Giuseppe De Lucca, Jan R. Magnus y Franco Peracchi presenta un resumen de la literatura junto con una crítica exhaustiva del trabajo de Oster que nos ayuda a la interpretación del mismo para una futura implementación

en nuestro trabajo. El trabajo hace uso de un análisis cuidadoso de las variables, así como de su impacto y su significado en nuestro modelo.

Oster define la ecuación 12:

$$Y = \beta X + \Psi\prime\omega o + W2 + \varepsilon \tag{12}$$

En la que X es la variable escalar de tratamiento, omega cero es el vector que contiene las variables observables y W2 son todas las variables omitidas. Para nuestra comodidad y concordancia para desarrollar la idea que Ciacci desarrolla en su artículo vamos a reescribir la ecuación del modelo de tal manera que $\beta X = \beta d$, $\Psi'\omega o = \theta X$ y $W2 = \gamma\omega$. Oster pone en relación d, X y ω a través de la ecuación de proporción de selección:

$$\delta\frac{Cov\,(d,\ X)}{Var\,(X)} = \frac{Cov\,(d,\ \omega)}{Var\,(\omega)} \tag{13}$$

Para algún valor de coeficiente de proporcionalidad δ.

Por otro lado, debemos definir el coeficiente de la regresión restringida, resultante de la regresión de Y sobre d, como β_r y a su R cuadrado como R_r. Por otro lado, la ecuación intermedia de Y sobre d y X como β_u y a su R cuadrado como R_u. Finalmente definimos Rmax como el R cuadrado resultante de realizar la regresión de Y sobre d, X y ω. El objetivo de Oster es encontrar un estimador de β para la ecuación 1. Para ello, busca representar la inconsistencia $b_u = \beta_u - \beta$. Aunque este estimador viene con muchas suposiciones, ha atraído el interés de muchos

investigadores gracias a su simplicidad y a su uso para análisis de sensibilidad para obtener estimadores no sesgados de β.

Para realizar su estimador, Oster se basa en varias suposiciones:

- Suposición 1: la correlación entre la variable tratamiento y los observables no es cero.
- Suposición 2: las variables de control y las no observables no están correlacionadas.
- Suposición 3: las variables de control están no correlacionadas entre sí.
- Suposición 4: la ecuación de proporción de selección es cierta para $\delta = 1$.
- Suposición 5: Considerando una regresión de d sobre X y nombramos al vector de coeficientes como $(\mu_1, \ldots, \mu_J)$. Los coeficientes de la regresión de Y sobre d y X son φ_i. Asumimos que $\varphi_i/\varphi_J = \mu_i/\mu_J \; \forall \; i, j$.

Las suposiciones que Oster menciona de manera explícita son la 4 y la 5. Estas suposiciones son fundamentales para el desarrollo del trabajo de Ciacci que explicaremos al final del trabajo. Las otras tres primeras suposiciones son incluidas en base a la lectura realizada sobre los comentarios del artículo de Oster escrito por Giussepe de Lucca.

La suposición 1 se puede inferir de la lectura del artículo de Oster. La suposición cuatro indica que tanto las variables omitidas como las observables guardan la misma correlación con la variable tratamiento. La última suposición más restricitiva y como Oster resalta en su trabajo, con múltiples variables de control es muy

difícil que se mantenga. Es el caso también de la suposición 2, la cual es muy difícil que se mantenga, Oster también menciona en su artículo que es probable que exista correlación entre las variables omitidas y las de control.

Bajo las cinco suposiciones mencionadas anteriormente, Oster desarrolla su primera propuesta:

$$b_u = (\beta_r - \beta_u) \frac{R^2_{max} - R^2_u}{R^2_u - R^2_r} \tag{14}$$

Dado que $b_u = \beta_u - \beta$ esto implica que la ecuación se puede expresar como:

$$\frac{(\beta_u - \beta)}{(\beta_r - \beta_u)} = \frac{R^2_{max} - R^2_u}{R^2_u - R^2_r} \tag{15}$$

Lo que implica que tal y como menciona Oster en su artículo, "el ratio de los movimientos de los coeficientes es igual al ratio de los movimientos en R cuadrado" (Oster, p.7). Como $\left(\hat{\beta}_r - \hat{\beta}_u\right)$ es consistente para $(\beta_r - \beta_u)$, para el caso en el que R^2_{max} es conocido se obtiene el estimador de sesgo corregido de β:

$$\beta = \beta_u - (\beta_r - \beta_u) \frac{R^2_{max} - R^2_u}{R^2_u - R^2_r} \tag{16}$$

Sin embargo, Oster avisa que debido a las suposiciones tan restrictivas necesarias para obtener el estimador no es aconsejable usarlo directamente. Teniendo en cuenta todas las suposiciones, se puede confirmar que añadir observables a nuestro modelo decrece nuestro sesgo en estimar β. Si no tomamos en cuenta la suposición

número 4 de misma selección, la ecuación para el cálculo del estimador de sesgo corregido vendría dada por:

$$\beta \approx \beta_u - \delta \left(\beta_r - \beta_u\right) \frac{R^2_{max} - R^2_u}{R^2_u - R^2_r} \tag{17}$$

Esta ecuación es útil para desarrollar la intuición de los investigadores ya que se puede calcular de forma sencilla con los datos que normalmente incluye una table de una regresión estándar y, por lo tanto, es una manera rápida de calcular la robustez de los resultados ya que nos da una buena aproximación del estimador de la segunda propuesta de Oster.

La segunda propuesta de Oster viene dada cuando no tomamos en cuenta las suposiciones 4 y 5. Esta propuesta es más generalista pero más difícil de aplicar. Esta propuesta dice que β_u viene dada por una de las raíces cúbicas de la siguiente ecuación:

$$a_3 z^3 + a_2 z^2 + a_1 z + a_0 = 0 \tag{18}$$

Cuyos coeficientes reales son:

$$a_0 = \ \delta \sigma_1^2 \sigma_y^2 \left(R^2_{max} - R^2_u\right) \left(\beta_r - \beta_u\right) \tag{19}$$

$$a_1 = \delta \left(\sigma_1^2 - \sigma_v^2\right) \sigma_y^2 \left(R^2_{max} - R^2_u\right) - \sigma_v^2 \left(\sigma_y^2 \left(R^2_u - R^2_r\right) + \sigma_1^2 \left(\beta_r - \beta_u\right)^2\right), \tag{20}$$

$$a_2 = (\delta - 2)\sigma_1^2(\beta_r - \beta_u)\sigma_v^2, \tag{21}$$

$$a_3 = (\delta - 1)\left(\sigma_1^2 - \sigma_v^2\right)\sigma_v^2, \tag{22}$$

Donde σ_y^2 y $\sigma_v^2 = \sigma_1^2 - \sigma_{21}' \sum_{22}^{-1} \sigma_{21}$ son las varianzas de la variable dependiente y y de la regresión $v = x_1 - \mu' X_2$. En el caso de relajar solo la suposición 5, $a_3 = 0$ y, por lo tanto, b_r, definido por $b_r = \beta_u - \beta$, es una raíz de la ecuación cuadrática:

$$a_2 z^2 + a_1 z + a_0 = 0 \tag{23}$$

En la segunda propuesta mas general de Oster[9] podemos ver claramente que la inconsistencia del estimador OLS sin restringir depende de nuevo de las diferencias $\beta_r - \beta_u$, $R_{max}^2 - R_u^2$ y $R_u^2 - R_r^2$. Sin embargo, esta propuesta es mucho menos intuitiva que la primera y las ecuaciones presentan varias raíces lo que crea confusión a la hora de elegir una. Oster[9] añade una nueva suposición para solventar esto (Assumption 3).

En resumen, Oster[9] en su artículo obtiene varias conclusiones y nos proporciona herramientas para ser capaces de evaluar el impacto de incluir observables en nuestro modelo. Nos muestra un marco teórico que explica que el sesgo de nuestro modelo depende tanto del movimiento de los estimadores como de los movimientos de R cuadrado, estos últimos movimientos siempre pasados por alto en los estudios econométricos. Este artículo también nos proporciona herramientas muy útiles

para el desarrollo del artículo de Ciacci[1] que veremos a continuación. Primero las suposiciones 4 y 5 las toma Ciacci[1] para construir un modelo intuitivo. Por otro lado, también nos proporciona la ecuación de selección de observables y no observables. En definitiva, nos da otra visión de cómo encontrar el estimador óptimos, así como un marco de suposiciones útiles para construir estimadores intuitivos.

5.6 Ciacci: A Matter of Size: Comparing IV and OLS estimates

Durante todo el trabajo se ha expuesto varios ejemplos tanto de IV como de OLS, así como algunas de las técnicas y líneas de investigación para comprobar la robustez de nuestros resultados y el impacto que tienen las variables observables y no observables sobre nuestro modelo. Sin embargo, en toda la literatura, bajo mi conocimiento, sólo existe un artículo que compare formal y objetivamente los estimadores obtenidos a partir de los métodos de OLS y IV. Este artículo, escrito por el tutor de este trabajo de fina de grado Riccardo Ciacci, propone un método basado en la ecuación de proporción de Oster para comparar objetivamente los estimadores en base al tamaño de la muestra. La comparación de estos dos enfoques es fundamental para entender las ventajas y desventajas de cada uno y cómo se aplican a diferentes situaciones en la investigación empírica. Hasta la fecha, los investigadores lo que realizan en sus trabajos de investigación para comparar los estimadores es calcular la proporción de tamaño que existe entre ambos estimadores. Sin embargo, esta técnica no es del todo correcta para ciertas ocasiones, tal y como Ciacci ejemplifica en su Artículo. Ciacci sugiere que valores más grandes

del coeficiente de proporcionalidad pueden ser evidencia en contra de la validez del instrumento, y viceversa.

En el artículo de Ciacci, se presenta la función de regresión de población (PRF), en la que se considera una variable de tratamiento escalar (d) y dos tipos de variables de control: las observables (X) y las no observables (w). La ecuación de queda definida como:

$$y_{ih} \;=\; \alpha_1 + \beta_1 d_{ih} \;+\; \gamma\omega_{ih} \;+\; \theta_1 X_{ih} \;+\; \varepsilon_{1ih} \tag{24}$$

Siendo ih la variable para la unidad i en un tiempo h. Sin embargo, dado la naturaleza de las variables omitidas solo podemos correr la regresión para la ecuación:

$$y_{ih} \;=\; \alpha_2 + \beta_2 d_{ih} \;+\; \theta_2 X_{ih} \;+\; \varepsilon_{2ih} \tag{25}$$

A continuación, Ciacci supone un modelo de una sola variable en el que la variable no observada es la única variable de control y no es incluida en el modelo. Con esto el autor logra simplificar la intuición detrás de su método, sin embargo, no deja de lado el análisis multivariable, incluido en el Apéndice A del articulo[1]. En la próxima sección, haremos uso de las ecuaciones multivariables para escribir de forma matricial la ecuación del coeficiente de correlación de Ciacci. Por lo tanto, suponiendo que estamos en un escenario univariado, nuestro estimador introduce un sesgo que viene representado por:

$$\hat{\beta}_2 \ = \beta_1 \ + \ \gamma \frac{Cov\left(d_{ih}, \omega_{ih}\right)}{Var\left(d_{ih}\right)} \tag{26}$$

Ese sesgo que suma a beta uno se puede interpretar como la prima por no incluir la variable omitida. En una estimación del efecto de los años de estudio sobre los salarios, como en el estudio de Sasha O. Becker[5], sería la prima por habilidad, que sesga nuestro estimador al alza. El signo de γ indica hacia donde sesga nuestra variable omitida: al alza, positivo; a la baja, negativo.

Siendo beta uno el estimador verdadero que proviene de la ecuación PRF. Tomando la ecuación de proporción de selección que menciona Oster[9] en su artículo que define la proporción entre las varianzas de las variables omitidas y las de control observables, podemos insertar en la ecuación anterior y despejar para δ. Esto da lugar a la ecuación:

$$\delta = (\hat{\beta}_2 - \beta_1) \frac{Var\left(d_{ih}\right) Var\left(X_{ih}\right)}{\gamma Var\left(\omega_{ih}\right) Cov\left(d_{ih}, X_{ih}\right)} \tag{27}$$

Mediante esta ecuación podemos calcular el tamaño que el coeficiente de proporcionalidad necesita tener para racionalizar el candidato a estimador. Tamaños muy grandes del coeficiente de proporcionalidad implica que se necesita una gran cantidad de variables omitidas para racionalizar el estimador y esto genera dudas sobre la robustez de el estimador. Pude que el estimador no sea válido o que exista efectos heterogéneos y que el estimador de variables instrumentales (IV) represente el efecto causal de un grupo de población.

La metodología se aplica a dos conjuntos de datos observacionales diferentes, uno

de Ciacci (2018)[1] y otro de Acemoglu et al. (2001)[6]. En el primer ejemplo, la metodología sugiere que un coeficiente de proporcionalidad δ relativamente pequeño es suficiente para racionalizar la gran diferencia entre las estimaciones MCO e IV. Aunque la literatura hasta la fecha considera esta diferencia entre los estimadores como muy grande ya que el estimador IV es trece veces del tamaño del otro, según la visión de Ciacci la diferencia en términos de observables para explicar este nuevo estimador no es muy grande. De hecho, con una cantidad de 1.16 veces no observables en comparación con observables se puede racionalizar este estimador.

En el segundo ejemplo, ningún coeficiente de proporcionalidad δ entre -1 y -1000 podría racionalizar la diferencia entre las estimaciones MCO e IV, poniendo en duda la credibilidad de las estimaciones IV en ese caso. Pese a que la diferencia entre los estimadores OLS y IV no es muy grande proporcionalmente, se necesita gran cantidad de observables parar explicar esta diferencia. Por lo tanto, pese a que la literatura hasta la fecha habría considerado estos dos estimadores como similares y que no representa un gran cambio, Ciacci propone otra visión y dice que la diferencia entre los estimadores es mucho mayor de la que pensamos debido a que hace falta muchos más controles para explicar esta diferencia.

Para el segundo conjunto observacional de Acemoglu et al. (2001), Ciacci[1] usa otras dos regresiones más del artículo que incluyen nuevas variables de control. De nuevo, para este conjunto de datos, pese a disminuir un poco los estimadores, no hay coeficiente de proporcionalidad entre -1 y -1000 que pueda racionalizar el estimador IV. Por lo tanto, esta diferencia entre los estimadores de IV y OLS es muy grande.

Sin embargo, Ciacci[1] nos informa de que la interpretación de los resultados requiere un conocimiento previo del entorno específico que se está analizando, ya que el investigador debe determinar si los supuestos realizados para utilizar esta metodología son plausibles. La metodología desarrollada en este trabajo no puede descartar la posibilidad de que los efectos sean heterogéneos para la subpoblación afectada por el instrumento, y puede ser necesario seguir investigando para abordar esta cuestión. El autor menciona que las variables de control pueden no ser informativas sobre las variables omitidas, aunque de ser así el caso, no tendría sentido introducir nuevas variables de control para estudiar la robustez.

En conclusión, Ciacci[1] en su artículo comparte una nueva visión de comparación de tamaño de los estimadores de OLS y IV, en los que demuestra que la diferencia de tamaño puede ser mayor de la que pensamos según la literatura actual. Ciacci también habla de que el estimador OLS puede ser un buen informativo para la relación causal que el investigador intenta hallar. La interpretación que se le debe dar al coeficiente de proporcionalidad según el autor es que, a bajos valores del coeficiente, se puede afirmar que no hay gran diferencia de tamaño. Sin embargo, a grandes coeficientes el estimador de IV debe ser revisado y complementado con otras técnicas para comprobar la robustez del mismo o ver si existen efectos heterogéneos. En definitiva, el trabajo de Ciacci proporciona una nueva visión sobre la comparación de tamaño de los estimadores OLS e IV, subrayando la necesidad de complementar el análisis con otras técnicas y enfoques para una evaluación más completa de la relación causal en cuestión.

6 Aportación al artículo de Ciacci

En esta sección del trabajo, vamos a aportar la ecuación (5) de Ciacci en forma matricial, así como un test para su evaluación. Finalmente, se expondrá una guía metodológica para la interpretación de los resultados.

6.1 Ecuación Ciacci en forma matricial

En el trabajo de Ciacci[1], la fórmula que se aporta está escrita en forma escalar. Por lo tanto, para extender el artículo de Ciacci vamos a escribir de forma matricial la ecuación del cálculo de coeficiente de proporción. Para ello comenzamos reescribiendo la ecuación PRF en forma matricial. Para ello vamos a especificar previamente as dimensiones de nuestra nueva fórmula:

- Y (n x 1): matriz de la variable dependiente, n observaciones.
- D (n x 1): matriz del tratamiento.
- W (n x 1): matriz de controles no observados.
- X (n x k): matriz de controles observados con k variables observables y n observaciones.
- β_1, β_2 (1 x 1): coeficientes escalares del tratamiento.
- θ_1, θ_2 (k x 1): vectores de coeficientes de los controles observados.
- γ (1 x 1): coeficiente escalar del control no observado.

Nuestra ecuación se va a reescribir tal que asi:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} D_1 \\ \vdots \\ D_n \end{bmatrix} \beta_1 + \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} \gamma + \begin{bmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \tag{28}$$

Que de forma simplificada se puede expresar como:

$$Y \ = D\beta_1 \ + \ W\gamma \ + \ X\theta_1 \ + \ \varepsilon_1 \tag{29}$$

Por la naturaleza de las variables omitidas nosotros solo podemos correr la regresión que viene dada por la ecuación:

$$Y \ = D\beta_2 \ + \ X\theta_2 \ + \ \varepsilon_2 \tag{30}$$

Además, vamos a exponer las dimensiones de las matrices de covarianza y varianza que serán usadas más adelante para escribir de forma matricial la ecuación 5 del artículo de Ciacci[1]:

- Σ_{DX} (1 x k): matriz de covarianza cruzada entre D y X, es decir, es el vector que contiene las covarianzas entre la variable de tratamiento y todas las variables observables del modelo.
- Σ_{DW} (1 x 1): matriz de covarianza entre D y W, que es equivalente al escalar Cov(D, W).
- Σ_D (1 x 1): matriz de varianza de D.

- Σ_X (1 x k): corresponde a la diagonal de la matriz de varianza-covarianza de X. Es decir, contiene las k varianzas de las variables observables.

- Σ_W (1 x 1): matriz de varianza de W.

Ahora tenemos un coeficiente de proporcionalidad por cada variable de control para que cumpla la ecuación de proporcionalidad con la variable omitida, por lo tanto, la ecuación de proporción de selección viene dada por:

$$\delta_j \frac{Cov\,(D,\ X_j)}{Var\,(X_j)} = \frac{Cov\,(D,\ \omega)}{Var\,(\omega)},\ para\ \forall j \in 1, \ldots,\ k \tag{31}$$

Siendo Xj el vector con n observaciones de la variable en la columna j de la matriz de observables X. La diferencia entre Σ_{DX} y Σ_{DXj} es que Σ_{DXj} solo contiene la Covarianza entre D y la variable observable X_j, mientras que Σ_{DX} contiene todas las covarianzas entre D y todas las observables. Lo mismo sucede para Σ_X y Σ_{Xj}. En notación matricial quedaría escrito como:

$$\delta_j \Sigma_{DXj} \Sigma_{Xj}^{-1} = \Sigma_{DW} \Sigma_W^{-1} \tag{32}$$

Siendo: δ (1 x k): la matriz con todos los coeficientes de proporción, hay tantos como variables observables tenemos. δ_j es el valor de delta para la observable X_j.

La ecuación de Ciacci del apéndice (A.2) dice que el sesgo por variable omitida viene dado por:

$$\hat{\beta}_2 \;=\; \beta_1 \;+\; \frac{\gamma\; Cov\left(d_{ih},\; w_{ih}\right) \;-\tau_1\; Cov\left(X_{ih},\; w_{ih}\right)}{Var\left(\;\widetilde{d}_{ih}\right)} \tag{33}$$

Teniendo en cuenta que hemos asumido que las observables y las no observables son ortogonales entre sí, ya que tomamos las suposiciones hechas por Oster, $Cov\left(X_{ih},\; w_{ih}\right) = 0$. Por otro lado, $\widetilde{d}$ son los residuales de la regresión de d sobre X, es decir, $d_{ih} = \tau_0 + \tau_1 X_{ih} + \widetilde{d}_{ih}$, que es la ecuación A.1 en el artículo de Ciacci[1]. Reescribimos la ecuación de regresión de d sobre X en forma matricial:

$$D = XT_1 + \widetilde{D} \tag{34}$$

Teniendo D y X las dimensiones mencionadas antes y $\widetilde{D}$ dimensión n x 1. El vector de coeficientes 1 tiene dimensiones k x 1. Reescribimos la ecuación A.2 en forma matricial, para ello primero debemos definir: $\Sigma_{\widetilde{D}}$ (1 x 1): matriz de varianza de $\widetilde{D}$. Σ_{XW} (k x 1): matriz de covarianza cruzada entre X y W, es decir, es el vector que contiene las covarianzas entre las variables observables y todas las variables omitidas. Por lo tanto la ecuación A.1 queda definida de forma matricial como:

$$\hat{\beta}_2 = \beta_1 + (\gamma\Sigma_{DW} - T_1^t\Sigma_{XW})\Sigma_{\widetilde{D}}^{-1} \tag{35}$$

Teniendo en cuenta que hemos asumido que las variables omitidas son ortogonales a las variables de control, sus varianzas serán 0 y por lo tanto la ecuación simplificada A.1 en forma matricial es:

$$\hat{\beta}_2 \;=\; \beta_1 \;+\; (\gamma\ \Sigma_{DW})\,\Sigma_{\widetilde{D}}^{-1} \tag{36}$$

Finalmente, si inyectamos la ecuación de proporción de selección en forma matricial en la ecuación anterior y despejamos para δ_j obtenemos la ecuación final de Ciacci escrita de forma matricial:

$$\delta_j = \left(\hat{\beta}_2 - \beta_1\right) \frac{1}{\gamma} \Sigma_{\widetilde{D}} \Sigma_{Xj} \Sigma_{DXj}^{-1} \Sigma_W^{-1} \tag{37}$$

Una vez hemos obtenido la ecuación anterior, vamos a generalizar para obtener un solo coeficiente de proporcionalidad. En la ecuación anterior tenemos δ_j, es decir tenemos un delta por cada variable no observable. En la ecuación 37 debemos sustituir Σ_{Xj} y $_{DXj}$ generalizando para todas las variables de control de X. De esta manera, Σ_{Xj} lo sustituimos por Σ_X, una matriz de 1 x k que contiene las varianzas de las k variables de control:

$$\begin{bmatrix} Var(X_1) & ... & Var(X_j) \end{bmatrix} \tag{38}$$

Σ_{DXj} lo sustituimos por Σ_{DX}, que es la matriz de covarianza cruzada entre D y X, que es de dimensión 1 x k. Como en la fórmula se requiere la matriz inversa y estamos ante una matiz no cuadrada debemos usar la pseudo inversa que representamos con el símbolo $^+$. Finalmente, llegamos con estos cambios a la ecuación final en forma matricial con un único coeficiente de proporcionalidad:

$$\delta = \left(\hat{\beta}_2 - \beta_1\right) \frac{1}{\gamma} \Sigma_{\widetilde{D}} \Sigma_X \Sigma_{DX}^{-1} \Sigma_W^{-1} \tag{39}$$

Tras todo este desarrollo llegamos a la ecuación 39 que es la ecuación 5 de Ciacci escrita de forma matricial y generalizable a un mayor número de casos.

6.2 Metodología para la interpretación de IV vs. OLS

En este apartado se aporta un test para probar estadísticamente el coeficiente de selección obtenido a partir de la ecuación de Ciacci [1], además de una metodología acompañada del test Durbin Wu Hausman [12]. En este apartado se quiere aportar una metodología formal que ayude a los investigadores a descubrir si el estimador obtenido a partir de variables instrumentales es el verdadero.

Nuestro método va a constar de dos fases. La primera de ellas va a consistir en escoger un valor umbral para el coeficiente de correlación y probar estadísticamente si nuestro coeficiente es mayor que ese umbral. La segunda fase consiste en usar el test de Durbin Wu Hausman [12] para comprobar la endogeneidad de nuestra variable de tratamiento. Dependiendo de los resultados, crearemos una tabla para poder interpretar los resultados obtenidos.

6.2.1 Test de la ecuación de Ciacci

El valor de δ en la ecuación (5) de Ciacci representa el grado de selección en no observables en comparación con la selección en observables. Un valor de δ mayor que 1 sugeriría que la selección en no observables es más fuerte que la selección en observables y, por lo tanto, supone un argumento en contra del estimador IV.

Además, en el paper de Oster [9] se escoge también este valor de referencia. Por lo tanto, aunque este valor umbral se deja a elección del investigador, es recomendable usar 1 por los argumentos expuestos anteriormente.

Para desarrollar un test estadístico teórico para cuando δ es mayor que 1, seguimos estos pasos:

- Hipótesis: Desarrollamos la hipótesis nula (H_0), $\delta \leq 1$ que sugiere que la selección de no observables es menor que la de observables. La hipótesis alternativa (H_1) es que $\delta > 1$, lo que sugiere que la selección de no observables es mayor que la de observables.

- Estadístico: Desarrollamos un estadístico de prueba adecuado. El estadístico de prueba que vamos a usar es directamente el valor de δ de la ecuación (5) de Ciacci. Además, como vamos a usar un test Z, vamos a necesitar estandar el estadístico de prueba restando la media bajo la hipótesis nula y dividiendo por la desviación estándar. De esta manera el estadístico de prueba resulta ser $Z = (\delta - 1)/\sigma_\delta$, donde σ_δ resulta ser al desviación estándar de δ.

- Distribución de muestreo: para la realización de este test asumimos que δ sigue una distribución normal, de tal manera que Z sigue una distribución estándar normal bajo la hipótesis nula.

- Nivel de significancia: De nuevo a elección del investigador, sin embargo, se recomienda usar el valor $\alpha = 0.05$ para una fiabilidad del 95 %.

- Decisión final: Para tomar una decisión comparamos nuestro estadístico de prueba con la distribución de muestreo bajo la hipótesis nula. En el caso en

el que el valor de Z es mayor que el valor crítico de la distribución estándar normal para el nivel de significancia marcado α, es decir $Z_{1-\alpha}$, rechazamos la hipótesis nula y concluimos que δ es mayor que 1 para un nivel de significancia de α. El rechazo de la hipótesis nula implica que es necesario una mayor selección de no observables que de observables para explicar el estimador IV. Tal y como menciona Ciacci en su artículo [1], esto es un indicador negativo hacia el supuesto de que el estimador IV es el efecto causal verdadero. En el apartado 6.2.3 se desarrollará más en profundidad las interpretaciones de los test expuestos.

Hay que resaltar el hecho de que para el desarrollo de este estadístico se ha asumido un distribución de muestreo normal. Esto es debido a la falta de conocimiento sobre la distribución de δ. Para conocer la distribución precisa del estadístico se podrían usar técnicas de remuestreo para estimar la distribución de muestreo del estadístico de prueba, como por ejemplo una de las más comunes: bootstrap.

6.2.2 Test de Durbin Wu Hausman para comprobar la endogeneidad

En este apartado se desarrolla la segunda fase de nuestra metodología para comprobar la endogeneidad de la variable de tratamiento a partir del test Durbin Wu Hausman. Este test fue desarrollado para evaluar la coherencia de un estimador cuando se compara con un estimador alternativo menos eficiente que ya se sabe que es coherente.

Para realizar este test debemos primero calcular el estadístico de prueba de Durbin Wu Hausman que en su forma para un sólo estimador es la diferencia entre los estimadores OLS e IV, dividida por su error estándar. Para los el estimador OLS,

$\hat{\beta}^{OLS}$, y para el estimador IV, $\hat{\beta}^{IV}$, el estadístico de prueba es:

$$DWH = \frac{\hat{\beta}^{OLS} - \hat{\beta}^{IV}}{se(\hat{\beta}^{OLS} - \hat{\beta}^{IV})} \tag{40}$$

Donde $se(\hat{\beta}^{OLS} - \hat{\beta}^{IV})$ es el error estándar de la diferencia entre los dos estimadores.

Para el caso general, usamos la ecuación general 41 cuando tenemos más de un estimador. La fórmula 41 es la forma matricial del estimador de Durbin Wu Hausman.

$$DWH = (\hat{\beta}^{OLS} - \hat{\beta}^{IV})^t [se(\hat{\beta}^{OLS} - \hat{\beta}^{IV})]^{-1} (\hat{\beta}^{OLS} - \hat{\beta}^{IV}) \tag{41}$$

Donde $se(\hat{\beta}^{OLS} - \hat{\beta}^{IV})$ es el error estándar de la diferencia entre los dos vectores de estimadores.

Finalmente, debemos tomar una decisión en base al estadístico de prueba. Si el estadístico de prueba sigue asintóticamente la distribución chi-cuadrado, se sostiene bajo la hipótesis nula.

La hipótesis nula (H_0) indica que la variable no está correlacionada con el término de error en el modelo de regresión, es decir el estimador OLS es eficiente y consistente. La hipótesis alternativa (H_1) indica que la variable de tratamiento está correlacionada con el error en el modelo de regresión. Es decir, el estimador de OLS es sesgado e inconsistente, mientras que el estimador IV es consistente.

El test de Durbin Wu Hausman compara los estimadores de OLS e IV, de tal manera que si la diferencia entre ellos es muy grande se determina que el estimador

OLS es inconsistente. De esta manera rechazamos la hipótesis nula y concluimos que la variable de tratamiento es endógena. En caso contrario, no rechazamos la hipótesis nula y determinamos que la variable de tratamiento es exógena.

Hausman demostró que para el caso de la hipótesis nula se cumple la siguiente igualdad:

$$se(\hat{\beta}^{OLS} - \hat{\beta}^{IV}) = se(\hat{\beta}^{OLS}) - se(\hat{\beta}^{IV}) \tag{42}$$

Hay que tener en cuenta que el test asume variables aleatorias independientes e idénticamente distribuidas que para tamaños muestrales pequeños no funciona correctamente. Al asumir variables aleatorias independientes una condición muy fuerte se plantea otro test alternativo para poder testear la endogeneidad de una variable:

- El primer paso de este test alternativo es regresar la variable independiente de tratamiento (d) en función del instrumento (z). Regresamos la ecuación: $d = \alpha_1 z + v$ .Con el modelo obtenido realizamos la predicción de la variable de tratamiento y obtenemos $\hat{v}$ que son los residuos estimados.

- El segundo paso consiste en realizar la regresión lineal de interés de la variable dependiente (y) en función de la variable de tratamiento, el instrumento y la predicción $\hat{v}$ de la variable de tratamiento. Regresamos la ecuación: $y = \beta_1 d + \beta_2 z + \theta_1 \hat{v} + u$

Procedemos a analizar los resultados para probar la endogeneidad de la variable de interés:

- Hipótesis nula (H_1): Si la covarianza entre la variable de tratamiento y el error es cero ($cov(d, u) = 0$), $plim\ cov_N(\hat{v}, u) = plim\ cov_n(\hat{v}, y) = 0$ y la $\theta_1 = 0$. En este caso se concluye que la variable de interés es exógena.

- Hipótesis alternativa (H_0): Si la covarianza entre la variable dependiente y el error es igual a cero ($cov(y, u) \neq 0$), plim $cov(\hat{v}, y) \neq 0$ y el coeficiente de $\hat{v}$ en el segundo paso es significativo, es decir, $\theta_1 \neq 0$. En este caso, el segundo paso es como añadir a la regresión original la variable omitida que captura la correlación entre y y u. En este caso se concluye que la variable de interés es endógena.

Este método, además de ser más sencillo computacionalmente nos permite obviar la condición de variables aleatorias independientes e idénticamente distribuidas, que en algunas ocasiones para tamaños muestrales pequeños no se cumple. Para probar si el estimador de los residuos es estadísticamente distinto de cero en este método usaremos un Z test, con nivel de significancia de 0.05 como hemos usado en apartados anteriores.

6.2.3 Decisión en base a los tests expuestos

En la tabla 2 podemos ver una guía de decisión en base a los resultados de los dos test expuestos anteriormente.

Como por cada test podemos aceptar o rechazar la hipótesis nula, tenemos cuatro escenarios posibles que un investigador se puede encontrar. El escenario más fácil de interpretar de todos es cuando aceptamos la hipótesis nula del test de Ciacci y rechazamos la hipótesis nula del test de Durbin Wu Hausman. En este caso

	Test Ciacci aceptar H0	**Test Ciacci rechazar H0**
Test de Durbin-Wu-Hausman aceptar H0	OLS estimador verdadero, IV posible estimador LATE	OLS estimador verdadero
Test de Durbin-Wu-Hausman rechazar H0	IV estimador verdadero	No podemos afirmar que IV o OLS son verdaderos

Table 2: Decisiones para el test de Durbin-Wu-Hausman y el test de la ecuación 5 de Ciacci

necesitamos menos variables no observables que variables observables y determinamos que la variable de tratamiento es endógena. Por lo tanto, es evidente que el estimador verdadero es desarrollado por el modelo IV.

Por otro lado, tenemos el escenario en el que rechazamos la hipótesis nula del test de Ciacci y aceptamos la hipótesis nula del test de Durbin Wu Hausman. En este caso, consideramos la variable de tratamiento exógena y necesitamos más variables no observables que observables para racionalizar el estimador IV. Por lo tanto, consideramos que el estimador verdadero viene dado por el modelo OLS.

Finalmente, tenemos los dos últimos escenarios que son menos evidentes que los dos anteriores. El escenario en el que aceptamos tanto la hipótesis nula del test de la ecuación de Ciacci y el test de Durbin Wu Hausman, estamos diciendo que la variable es exógena y que hace falta menos variables no observables que observables para racionalizar el estimador IV. En este caso aceptamos OLS como el verdadero ya la variable de tratamiento es exógena, sin embargo, el estimador IV puede tener una interpretación LATE que puede servir de utilidad dependiendo del caso de estudio del investigador. El último escenario escenario es el más negativo de todos ya que no podemos confirmar ningún estimador como válido. Esto sucede cuando

rechazamos la hipótesis nula de ambos tests. Para este caso, lo recomendable sería buscar un nuevo estimador o usar una de las técnicas expuestas en la revisión de la literatura para mejorar los resultados del modelo.

7 Conclusiones

Este trabajo ha proporcionado una visión detallada de dos de los métodos más importantes en la econometría causal: el Estimador de Mínimos Cuadrados Ordinarios (OLS) y el Estimador de Variables Instrumentales (IV). A través de un análisis exhaustivo, hemos explorado las fortalezas y debilidades de cada método, así como las posibles formas de decidir entre uno y otro en el estado del arte actual.

Hemos discutido cómo usar IV y que problemas resuelve con respecto a OLS. Así mismo, también hemos revisado las condiciones que debe cumplir un instrumento para poder ser usado en un modelo: restricción de exclusión y fuerte correlación con la variable tratamiento. También hemos discutido el concepto de LATE y cómo un efecto causal puede variar par distintos grupos poblacionales.

A pesar de los avances en investigación, sólo se cuenta con un método riguroso para comparar ambos estimadores. Este trabajo aporta una fórmula matricial para esa ecuación, que el autor Ciacci[1] escribe de forma escalar. Mediante el desarrollo matricial de esta ecuación se permite una generalización de la ecuación. Es urgente la necesidad de crear un marco teórico que se pueda aplicar de forma empírica para evaluar ambos estimadores y descubrir cuál es el estimador verdadero.

Finalmente, se expone un test estadístico para la interpretación del resultado obtenido a partir de la ecuación. Este test se complementa con el conocido tes

Durbin Wu Hausman el cual nos permite probar la endogeneidad de las variables de un modelo. Mezclando estos dos test llegamos a una metodología que ayuda al investigador a interpretar sus resultados para finalmente lograr obtener el estimador verdadero, que es el objetivo último en econometría causal.

Como siempre, es esencial tener en cuenta las limitaciones de cada método y considerar cuidadosamente las implicaciones de nuestras elecciones metodológicas. Cada estudio depende de un contexto y objeto de análisis distinto para el que puede ser mejor aplicar un método. En un mundo en constante evolución, donde los desafíos y las preguntas de investigación varían, es crucial reconocer que cada estudio depende de un contexto y objeto de análisis distinto. Además, debemos ser conscientes de que no existe un enfoque único que se ajuste perfectamente a todas las situaciones. De esta manera, podremos obtener resultados confiables y relevantes que contribuyan al avance del conocimiento en nuestro campo de estudio.

Como futuras líneas de investigación, es obvio que las aportaciones realizadas a la econometría causal se han realizado de manera teórica únicamente. Por lo tanto, sería bueno realizar un estudio empírico de la aplicación de la forma matricial y de los test propuestos. De esta manera se podría realmente ver el impacto de las conclusiones y aportaciones obtenidas de este trabajo. Como hemos visto en la mayoría de la literatura, los artículos citados que expanden el conocimiento o proponen nuevas metodologías lo acompañan con un estudio empírico para probar sus teorías. En este sentido, sería de gran valor agregar un estudio empírico a este trabajo. Sin embargo, debido a la limitación temporal del proyecto, no ha sido posible incluir un estudio empírico en esta etapa.

References

[1] Riccardo Ciacci. "A Matter of Size: Comparing IV and OLS estimatess". In: (May 20, 2021).

[2] *Ordinary least squares*. `https://en.wikipedia.org/wiki/Ordinary_least_squares`. Accessed: 2023-05-25.

[3] *Endogeneidad Economía*. `https://es.wikipedia.org/wiki/Endogeneidad_(economa)`. Accessed: 2023-05-25.

[4] *Variable instrumental*. `https://es.wikipedia.org/wiki/Variable_instrumental`. Accessed: 2023-05-25.

[5] Sascha O. Becker. "Using instrumental variables to establish causality". In: (April 2016).

[6] DARON ACEMOGLU - SIMON JOHNSON - JAMES A. ROBINSON. "The Colonial Origins of Comparative Development: An Empirical Investigation". In: (Dec., 2001).

[7] Matthew A. Masten and Alexandre Poirier. "Salvaging Falsified Instrumental Variable Models". In: (January 7, 2020).

[8] Douglas Staiger y James H. Stock. "Instrumental variables regression with weak instruments". In: (1997).

[9] Emily Oster. "Unobservable Selection and Coefficient Stability: Theory and Evidence". In: (August 9, 2016).

[10] Joseph G. Altonji - Todd E. Elder - Christopher R. Taber. "SELECTION OF OBSERVED AND UNOBSERVED VARIABLES: ASSESSING THE EFFECTIVENESS OF CATHOLIC SCHOOLS". In: (August 2000).

[11] Giuseppe De Luca - Jan R. Magnus - Franco Peracchi. "Comments on "Unobservable Selection and Coefficient Stability: Theory and Evidence" and "Poorly Measured Confounders are More Useful on the Left Than on the Right"". In: (September 12, 2018).

[12] *Durbin–Wu–Hausman test*. `https://en.wikipedia.org/wiki/DurbinWuHausman_test`. Accessed: 2023-05-25.