FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

MODELOS DE INFERENCIA CAUSAL EN ECONOMETRÍA

Autor: Jaime González Lara

5º E-3 Analytics

Tutor: Riccardo Ciacci

Madrid

Junio 2023

RESUMEN

La endogeneidad es un problema común en la econometría que puede conducir a estimaciones sesgadas e incorrectas de los parámetros del modelo, lo que dificulta establecer relaciones causales precisas. Para abordarla en todas sus versiones, se utilizan técnicas como las Variables Instrumentales. Los coeficientes estimados mediante el método de Mínimos Cuadrados Ordinarios se han utilizado como indicadores para evaluar la validez del instrumento utilizado. Sin embargo, hasta la fecha, no existe una metodología formal para comparar de manera objetiva estas dos estimaciones. Este trabajo se propone abordar dicha carencia y presenta una metodología formal para comparar las estimaciones obtenidas.

PALABRAS CLAVE

Endogeneidad, Variables Instrumentales, Mínimos Cuadrados Ordinarios, Econometría, Inferencia Causal, Estadística.

ABSTRACT

Endogeneity is a common problem in econometrics that can lead to biased and incorrect estimates of model parameters, making it difficult to establish accurate causal relationships. To address it in all its versions, techniques such as Instrumental Variables are used. The coefficients estimated by the Ordinary Least Squares method have been used as indicators to assess the validity of the instrument used. However, to date, there is no formal methodology to objectively compare these two estimates. This paper sets out to address this shortcoming and presents a formal methodology for comparing the estimates obtained.

KEY WORDS

Endogeneity, Instrumental Variables, Ordinary Least Squares, Econometrics, Causal Inference, Statistics.

ÍNDICE

1. INTRODUCCIÓN

1.1 Presentación del estudio

La **endogeneidad** es un problema común en la econometría que surge debido a la presencia de correlación entre las variables explicativas y los términos de error en un modelo econométrico. Esta correlación puede conducir a estimaciones sesgadas e incorrectas de los parámetros del modelo, lo que dificulta establecer relaciones causales precisas.

La endogeneidad puede manifestarse de diferentes maneras. Una de ellas es a través de la simultaneidad, donde las variables de interés se influyen mutuamente de forma bidireccional. Por ejemplo, en un modelo que relaciona la inversión y el crecimiento económico, la inversión puede influir en el crecimiento económico, pero a su vez, el crecimiento económico puede afectar la inversión. Esta interacción simultánea dificulta discernir la dirección de la causalidad.

Otra fuente de endogeneidad es la presencia de variables omitidas o no observadas. Estas variables pueden tener un impacto tanto en las variables explicativas como en la variable dependiente, pero no se incluyen en el modelo debido a su falta de disponibilidad o medición. Este problema puede conducir a una correlación espuria entre las variables incluidas y los términos de error, generando un sesgo en las estimaciones. Si bien realizar una investigación de la manera más completa y exhaustiva puede ser lo más deseable, en muchas ocasiones es extremadamente difícil, costoso e incluso imposible tener acceso a todas las variables relevantes en un estudio. En una investigación sobre el efecto del ejercicio en la salud cardiovascular podemos recopilar fácilmente datos sobre la cantidad de ejercicio que hacen las personas así como las mediciones de su presión arterial. Sin embargo, puede ser que no tengamos información sobre la dieta la cual afecta tanto al ejercicio como a la presión arterial. Del mismo modo, al realizar un trabajo sobre el efecto de la educación en el salario muy probablemente se vayan a omitir variables como la motivación, por lo difícil y costoso que podría ser medirla, la cual influye también tanto en la educación como en el salario.

La endogeneidad también puede surgir debido a la autocorrelación, que ocurre cuando los términos de error están correlacionados a lo largo del tiempo. Esta correlación puede

distorsionar las estimaciones y dificultar la interpretación causal adecuada de los resultados. Por ejemplo, podemos tratar de investigar el efecto del café en el nivel de energía de las personas recopilando diariamente los niveles de energía de las personas y el consumo de café. No obstante, el nivel de energía a lo largo del tiempo también se relaciona con el nivel de energía del día anterior siendo difícil identificar el impacto del café en el nivel de energía.

Para abordar la endogeneidad en todas sus versiones se utilizan distintos tipos de técnicas. Técnicas como las variables instrumentales, buscan identificar variables que estén correlacionadas con las variables endógenas de interés pero no estén correlacionadas con los términos de error. Estas variables instrumentales permiten controlar la endogeneidad y obtener estimaciones consistentes y no sesgadas.

En el campo de la econometría, las técnicas de variables instrumentales son ampliamente utilizadas para abordar problemas fundamentales relacionados con la endogeneidad y el error de medición en los análisis económicos (Ciacci, 2021). Estas técnicas ofrecen herramientas valiosas para estimar relaciones causales precisas y confiables. En particular, los coeficientes estimados mediante el método de Mínimos Cuadrados Ordinarios (OLS) y las variables instrumentales (IV)[1] se han utilizado como indicadores para evaluar la validez del instrumento utilizado.

Sin embargo, hasta la fecha, no existe una metodología formal para comparar de manera objetiva estas dos estimaciones. Aunque algunos estudios han intentado realizar comparaciones empíricas, carecemos de un enfoque sistemático para abordar este problema (Ciacci, 2021).Por tanto, surge la necesidad de desarrollar una metodología que permita realizar estas comparaciones de manera rigurosa.

En este contexto, este trabajo se propone abordar dicha carencia y presenta una metodología formal para comparar las estimaciones obtenidas mediante las técnicas de variables instrumentales y Mínimos Cuadrados Ordinarios. Siguiendo la propuesta de Oster (2019) (Ciacci, 2021), se utiliza información proveniente de la regresión OLS, como la inclusión de variables de control, el tamaño de las varianzas y el coeficiente de determinación (R^2), para establecer un rango de valores dentro del cual el efecto del

[1] [1] Por su traducción del inglés *Ordinary Least Squares – OLS* e *Instrumental Variables - IV*

tratamiento verdadero debería encontrarse. Este rango dependerá de la capacidad informativa de las variables observables respecto a las variables no observables.

La contribución principal de este trabajo radica en proporcionar una metodología rigurosa que permita realizar comparaciones objetivas entre las estimaciones obtenidas mediante las técnicas de variables instrumentales y Mínimos Cuadrados Ordinarios (Ciacci, 2021). Además, se enmarca en la línea de investigación de otros estudios que han abordado la comparación de estas estimaciones sin contar con una metodología formal adecuada. Asimismo, se relaciona con la literatura que utiliza variables observables para evaluar el sesgo generado por las variables no observables en los análisis realizados con Mínimos Cuadrados Ordinarios.

1.2 Justificación del tema

La elección de este tema de trabajo se basa en la importancia de abordar de manera rigurosa la comparación entre las estimaciones obtenidas mediante las técnicas de variables instrumentales y Mínimos Cuadrados Ordinarios (IV y OLS en adelante). Estas técnicas son ampliamente utilizadas en la econometría para abordar problemas de endogeneidad y error de medición en los análisis económicos. Sin embargo, a pesar de su uso generalizado, no existe una metodología formal establecida para realizar comparaciones objetivas entre las estimaciones obtenidas con ambos métodos. Esta falta de una metodología sistemática limita nuestra capacidad para evaluar la validez de los instrumentos utilizados y comprender plenamente los efectos del tratamiento en diferentes contextos y escenarios.

Analizar esta cuestión es de gran interés por varias razones. En primer lugar, la correcta identificación de las relaciones causales en economía es fundamental para tomar decisiones informadas y diseñar políticas efectivas. Al comparar las estimaciones obtenidas mediante IV y OLS, podemos obtener una comprensión más profunda de la validez de los instrumentos utilizados y evaluar si las diferencias observadas en los coeficientes son indicativas de una endogeneidad real o simplemente reflejan sesgos de medición.

La capacidad de realizar comparaciones objetivas entre las estimaciones de IV y OLS es especialmente relevante en escenarios donde la endogeneidad es un problema importante,

como en estudios de impacto de políticas públicas, análisis de causalidad en economía laboral o investigaciones en áreas como la salud y la educación. Comprender la validez del instrumento utilizado y las diferencias entre las estimaciones de IV y OLS nos permite obtener resultados más sólidos y confiables, evitando inferencias erróneas que podrían conducir a decisiones políticas o acciones inapropiadas.

Además, al desarrollar una metodología formal para realizar estas comparaciones, podemos avanzar en la literatura existente y brindar una base sólida para futuras investigaciones. Esto permitiría a los investigadores y analistas económicos contar con herramientas más precisas y confiables para evaluar los efectos del tratamiento y tomar decisiones basadas en información respaldada y fundamentada. La metodología propuesta por Oster (2019) se basa en la utilización de información proveniente de la regresión OLS, como la inclusión de variables de control, el tamaño de las varianzas y el coeficiente de determinación R^2, para establecer un rango de valores dentro del cual el efecto del tratamiento verdadero debería encontrarse. Esta aproximación proporciona una forma sistemática de evaluar las diferencias entre las estimaciones de IV y OLS, permitiendo una mejor comprensión de las relaciones causales subyacentes.

Las conclusiones que se pueden extraer de este análisis son valiosas tanto para la teoría económica como para la práctica. Al establecer un rango de valores dentro del cual el efecto del tratamiento verdadero debería encontrarse, podemos evaluar la robustez de los resultados obtenidos con diferentes metodologías. Además, al considerar la capacidad informativa de las variables observables respecto a las variables no observables, podemos obtener una estimación más precisa y confiable del efecto del tratamiento en la población de interés.

Por lo tanto, el análisis y la comparación rigurosa entre las estimaciones obtenidas mediante IV y OLS contribuyen a mejorar nuestra comprensión de las relaciones causales en economía y a fortalecer la base empírica de las investigaciones. Esto tiene implicaciones tanto teóricas como prácticas, permitiendo una toma de decisiones más informada y la mejora de las políticas económicas. Además, proporciona una herramienta valiosa para los investigadores y analistas económicos que deseen evaluar la validez de los instrumentos utilizados y obtener estimaciones más confiables del efecto del tratamiento.

La elección de este tema de trabajo, por tanto, se justifica en la falta de una metodología formal para comparar las estimaciones obtenidas mediante IV y OLS. Al analizar esta cuestión de manera rigurosa, podremos obtener conclusiones sólidas sobre la validez de los instrumentos utilizados y mejorar nuestra comprensión de los efectos del tratamiento en diferentes contextos. Estas conclusiones tienen implicaciones tanto teóricas como prácticas, lo que nos permitirá tomas decisiones de manera informada y bajo un pretexto sólido y poder aplicar estos modelos o políticas de manera más eficiente y práctica.

1.3 Objetivos de la investigación

Los objetivos de esta investigación son múltiples y se centran en abordar el problema de la endogeneidad en la econometría y en proporcionar una metodología formal para comparar los estimadores de OLS y IV.

En primer lugar, vamos a tratar de desarrollar una metodología para comparar OLS y IV, proponiendo una metodología formal que permita comparar los estimadores de OLS y IV. Esto implica utilizar información derivada de la regresión OLS, como la inclusión de controles, el tamaño de las varianzas y los cambios en el coeficiente de determinación (R^2), para establecer un conjunto de valores plausibles donde el verdadero efecto tratado debe estar. Esta metodología nos proporcionará un enfoque objetivo y sistemático para evaluar y comparar los estimadores OLS y IV.

En segundo lugar, otro objetivo clave de esta investigación es el de utilizar la metodología propuesta para evaluar la validez del instrumento utilizado en el modelo de IV. La idea es que los valores más grandes (o más pequeños) del coeficiente de proporcionalidad, que mide la relación relativa entre la selección basada en variables observables y no observables, van a poder proporcionar evidencia en contra (o a favor) de la validez del instrumento. Esto permitiría a los investigadores tener una medida objetiva para poder evaluar si el instrumento utilizado es adecuado y confiable para estimar el efecto del tratamiento.

En tercer y último lugar, es importante comparar los resultados obtenidos a través de los estimadores OLS y IV utilizando la metodología propuesta. Esto permitirá evaluar si existen diferencias significativas entre los estimadores y analizar las implicaciones de

estas diferencias en la inferencia causal. Al hacerlo, los investigadores podrán obtener conclusiones más sólidas y robustas sobre el efecto del tratamiento objeto de estudio.

Por todo ello, este trabajo trata de proporcionar un enfoque más riguroso y estructurado para poder evaluar la validez del instrumento y realizar inferencias causales más sólidas. Dicho lo cual, este trabajo ayudaría a añadir calidad y confiabilidad a los resultados obtenidos en estudios econométricos. Además, al utilizar medidas objetivas, como el coeficiente de proporcionalidad, estaríamos reduciendo el sesgo subjetivo en la evaluación de la validez del instrumento. Esto sería de ayuda a los investigadores a la hora de tomar decisiones más fundamentadas y a evitar interpretaciones erróneas basadas únicamente en diferencias entre los estimadores OLS y IV. La metodología propuesta podrías ser aplicada más adelante en una variedad de contextos y problemas de investigación donde la endogeneidad y la comparación de estimadores son relevantes. Esto amplía la utilidad y aplicabilidad de los resultados de esta investigación en diversos campos de la econometría y las ciencias sociales.

1.4 Contexto del estudio

En la actualidad, el mundo está experimentando una creciente digitalización en la cual diversas ramas, como la estadística en combinación con el aprendizaje automático (*machine learning*), están dando pasos de gigante. La capacidad de modelización y creación de modelos predictivos en diferentes campos se ha vuelto omnipresente. Sin embargo, en este contexto, a menudo nos encontramos con el desafío de la endogeneidad en nuestros modelos, el cual puede afectar la validez de las conclusiones obtenidas. Afortunadamente, existen técnicas como el uso de variables instrumentales que nos permiten abordar esta problemática.

Sin embargo, hasta el día de hoy no existe una metodología formal que nos permita rechazar o aceptar la validez del instrumento. Ante el interrogante acerca de cómo evaluar la validez de las variables instrumentales utilizadas, el trabajo de Ciacci ha propuesto una metodología formal basada en la ecuación del coeficiente de proporcionalidad, que puede resultar útil para evaluar y comparar las estimaciones obtenidas mediante el uso de regresión de Mínimos Cuadrados Ordinarios (OLS) y regresión con variables instrumentales (IV).

La ecuación del coeficiente de proporcionalidad propuesta por Ciacci permite realizar una comparación sistemática entre los resultados de OLS y IV, con el objetivo de evaluar la validez del instrumento empleado. Al aplicar esta metodología, se pueden realizar inferencias más sólidas acerca de la efectividad de las variables instrumentales utilizadas en la resolución del problema de endogeneidad.

En este contexto, resulta crucial abordar el desafío de la validación de instrumentos utilizados en regresiones instrumentales, con el fin de otorgar robustez y confiabilidad a las estimaciones obtenidas. Es en este sentido que se hace relevante el trabajo de Ciacci y la aplicación de la ecuación del coeficiente de proporcionalidad.

El enfoque propuesto por Ciacci proporciona una metodología formal que nos permite evaluar la validez de los instrumentos empleados en las regresiones instrumentales. Mediante el análisis y la comparación sistemática de las estimaciones obtenidas mediante OLS y IV, utilizando la ecuación del coeficiente de proporcionalidad, podemos obtener una mayor comprensión de la efectividad de los instrumentos utilizados y validar su uso en el proceso de modelización estadística.

Al contar con una metodología formal para validar los instrumentos, se fortalece la confianza en los resultados obtenidos a través de las regresiones instrumentales. Esto implica una mejora significativa en la calidad de las estimaciones, permitiendo tomar decisiones más fundamentadas en base a los análisis realizados.

En conclusión, en un entorno de creciente digitalización y uso de técnicas estadísticas y de *machine learning*, resulta esencial contar con una metodología formal que permita validar los instrumentos utilizados en las regresiones instrumentales. El enfoque basado en la ecuación del coeficiente de proporcionalidad, propuesto por Ciacci, se presenta como una herramienta valiosa para fortalecer la robustez de las estimaciones obtenidas y otorgar mayor confiabilidad a los resultados de los modelos predictivos y de modelización en general.

2. REVISIÓN DE LA LITERATURA

2.1 Marco teórico

La inferencia causal es un concepto con el que todos estamos familiarizados y que al mismo tiempo es fundamental en la econometría. La **inferencia causal** no es otra cosa que el proceso por el que establecemos relaciones de causa y efecto entre eventos y cosas. Este concepto no solo es innato de las personas, podríamos establecer que de todos los seres vivos, pues es gracias a nuestra capacidad de entender las relaciones causales de las cosas más básicas que podemos perpetuarnos como especies.

La inferencia causal, por tanto, es un concepto fundamental en nuestra comprensión del mundo que nos rodea. Como hemos comentado, nos permite establecer relaciones de causa y efecto, aunque tan solo sean mentalmente, lo que a su vez nos ayuda a tomar decisiones informadas, predecir resultados y solucionar problemas de manera efectiva, siendo aplicable a infinidad de situaciones. Por ejemplo, cuando uno empieza a trabajar en un sitio nuevo al principio le puede costar adaptarse a la rutina y sin embargo tras unos pocos días uno sabe perfectamente a que hora pasará el bus por las mañanas y no tendrá que esperarle 15 minutos o que horas la cantina de la empresa estará menos ajetreada y podrá comer más tranquilo o como comportarse con cada uno de sus compañeros y de que hablar con cada uno. Pues uno mentalmente esta continuamente estableciendo relaciones de causa y efecto. Es por ello por lo que, en la capacidad de la inferencia causal de proporcionarnos esa información más profunda de los fenómenos que observamos en nuestra vida cotidiana y en otros campos de estudio, radique su importancia.

En campos como el de la salud la inferencia causal ha sido fundamental en el establecimiento de conexiones entre ciertos comportamientos y la aparición de enfermedades. Mediante los correspondientes estudios y experimentos, los investigadores han podido inferir que el consumo excesivo de azúcar esta relacionado con un mayor riesgo de desarrollar enfermedades como la diabetes tipo 2. En otras ocasiones, si observamos que una persona fuma regularmente, podemos inferir casualmente que es más probable que esa persona desarrolle problemas respiratorios o enfermedades pulmonares. Esta inferencia causal en el ámbito de la medicina ha permitido a las autoridades sanitarias promocionar estilos de vida más saludables desincentivando el consumo excesivo de azúcar o del tabaco como en estos ejemplos.

La inferencia causal desempeña un papel de vital importancia en campos de estudio como es el de la econometría. En el campo de la econometría, es fundamental establecer relaciones causales entre variables económicas. Por ejemplo, supongamos que queremos investigar el impacto del aumento del salario mínimo en el empleo. Utilizando modelos de inferencia causal, podremos diseñar estudios y recopilar datos relevantes para analizar si existe una relación causal entre estos dos factores. Controlando otras variables y utilizando las técnicas estadísticas apropiadas, podremos determinar si el aumento del salario mínimo efectivamente tiene un impacto en el empleo y en qué medida. Del mismo modo, podremos investigar el efecto de un programa de incentivos fiscales en el crecimiento de las empresas o los efectos que las políticas monetarias tienen en la inflación.

Tener esta comprensión causal es esencial a la hora de evaluar la efectividad de las políticas económicas y poder tomar las decisiones de manera razonada e informada para promover el crecimiento económico y la eficiente asignación de los recursos. Si bien nuestro conocimiento innato de la inferencia causal nos puede ser de ayuda en muchos aspectos, a la hora de valorar políticas de este estilo, y conocer con mayor profundidad que variables son las que afectan nuestro objetivo en mayor o menor medida, no es conocimiento suficiente y por ello, dado que es sumamente importante, debemos trasladar este conocimiento innato al lenguaje de las matemáticas.

Por ello, la inferencia causal se basa en una sólida notación matemática que, junto con la **regresión lineal**, nos permite analizar, comprender, cuantificar y medir los efectos de una variable sobre otra, estableciendo relaciones de causa y efecto de manera rigurosa (Hernan & Robins, 2020). En el contexto de la inferencia causal, la regresión lineal nos ayuda a identificar y cuantificar los efectos causales individuales de las variables independientes sobre la variable dependiente y de este modo priorizar unas variables sobre otras.

A continuación, y utilizando ejemplos muy parecidos a los anteriormente expuestos, podemos aplicar de manera simplificada la notación matemática y la regresión lineal a casos reales. Supongamos que queremos investigar el efecto de la educación en los ingresos de las personas. Para ello, podríamos utilizar la notación matemática y representar la variable de educación como "E" y los ingresos como "Y". Para medir el efecto causal individual de la educación sobre los ingresos, podemos utilizar una

regresión lineal simple, donde la ecuación sería: $Y = \alpha + \beta E + \varepsilon$. Siendo α la intersección de la línea de regresión, β el coeficiente de la variable educación y ε el término de error. Esta ecuación, si bien es la versión más simple posible de la misma, es la base de este trabajo y la desarrollaremos más adelante.

Con un modelo de estas características ya estaríamos en posición de estimar el efecto causal individual de la educación sobre los ingresos al analizar la relación entre las dos variables. Si el coeficiente β fuese positivo y significativo, sería entonces un indicativo de que a medida que aumenta la educación, los ingresos también tienden a aumentar. Por ejemplo, si β= 3.000 podríamos decir que un año más de educación, en promedio, implican 3.000 euros más de salario. En este caso, un valor de α=15.000 nos indicaría que, en promedio, cuando el nivel de educación es 0 los ingresos son de 15.000 euros. Podríamos, por tanto, fundamentar en estos resultados que existe un efecto causal positivo de la educación en los ingresos.

Los **mínimos cuadrados ordinarios**, OLS, son una técnica estadística comúnmente utilizada para estimar los parámetros de una ecuación de regresión como la del ejemplo anterior y determinar la relación entre las variables independientes y la variable dependiente. Esta técnica, al minimizar la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo, nos permite obtener una estimación precisa del efecto causal individual de la educación sobre los ingresos al ajustar una línea de regresión que mejor se ajuste a los datos observados, generando los valores de α y β que mejor se ajustan a los datos (Mendenhall, Beaver, & Beaver, 2006).

Sin embargo, los mínimos cuadrados ordinarios tienen una limitación importante. Ofrecen estimaciones consistentes solo si los términos de error son asintóticamente ortogonales a los regresores. Estas condiciones necesarias no se pueden verificar directamente, ya que la propiedad de ortogonalidad de los mínimos cuadrados asegura que, independientemente de si los términos de error están correlacionados o no con los regresores, los residuos de los mínimos cuadrados serán ortogonales a los regresores. Esto significa que, incluso si las estimaciones de los mínimos cuadrados son sesgadas e inconsistentes, los residuos de los mínimos cuadrados no proporcionarán evidencia de que haya un problema (Davidson & MacKinnon, 1993). En otras palabras, los OLS solo nos ofrecerán estimaciones consistentes si los errores en nuestras predicciones no están relacionados o no dependen de las cosas que estamos tratando de estudiar o predecir. Por

ejemplo, si al estudiar cómo afecta la cantidad de tiempo de estudio a las calificaciones de los alumnos las cosas que no podemos controlar, como el cansancio, las distracciones o la suerte, no tienen una relación directa con la cantidad de tiempo de estudio y por tanto no afectan consistentemente las calificaciones de los estudiantes. Sin embargo, hay muchas ocasiones en las que estos resultados pueden estar sesgados al estar las variables independientes correlacionadas con otras no observadas, generando un problema de endogeneidad.

Aquí es donde entran en juego las **variables instrumentales**. Las IV nos permiten manejar estas situaciones y obtener estimaciones consistentes y confiables. Básicamente son variables que se utilizan como "instrumentos" para estimar el efecto causal de una variable de interés sobre otra variable. La idea principal detrás de las variables instrumentales es encontrar una variable que esté correlacionada con la variable de interés pero no esté correlacionada con los errores del modelo. Esta variable instrumental actúa como un "proxy" o sustituto para la variable de interés y nos permite estimar su efecto causal sin verse afectado por la endogeneidad de los errores. Por ejemplo, en el caso anterior puede ser el caso de que los resultados estén sesgados por estar las horas de estudio correlacionadas con otros factores no observados que también influyan en las calificaciones, como podría ser la motivación. Para resolver este problema necesitaríamos encontrar una variable instrumental que esté correlacionada con las horas de estudio pero no tenga efecto directo en las calificaciones como podría ser la disponibilidad de tutorías académicas gratuitas. De esta manera nos sería posible estimar el efecto causal de las horas de estudio en las calificaciones, controlando el posible sesgo causado por la endogeneidad.

Vamos a tratar de darle un enfoque práctico y matemático a estos dos conceptos que acabamos de explicar. Pongamos esta vez como ejemplo que estamos tratando de elaborar un modelo para estudiar el efecto causal de recibir un programa educativo o master sobre los salarios.

Utilizando notación matemática a un caso de este estilo se representaría de la siguiente manera: el grupo de tratamiento (que recibió el programa) como "T" y el grupo de control

(que no recibió el programa) como "C"[2]. El salario promedio de los participantes del grupo de tratamiento sería denotado como "Y(T)" y el salario promedio del grupo de control como "Y(C)".

Sin embargo, al realizar estos experimentos y ejercicios, muchas veces nos encontramos con el problema de la endogeneidad. Como ya hemos explicado anteriormente, este problema puede surgir por una serie distinta de motivos (simultaneidad, variables omitidas, etc). Una de las herramientas a nuestro alcance para hacer frente a este problema es el de las variables instrumentales. En el caso que estamos utilizando como ejemplo, puede ser que tengamos este problema al omitir variables que, por ejemplo, en este caso no nos es posible conseguir.

Para el caso del programa educativo y su efecto causal en los salarios, introduciremos una variable instrumental ("Z") que está correlacionada con la participación en el programa educativo pero no tiene una relación directa con los salarios. Supongamos que esta variable instrumental es la distancia geográfica al lugar donde se imparte dicho programa.

A continuación, y recapitulando la notación matemática que hemos ido incorporando a nuestra explicación sobre la inferencia causal, tenemos la ecuación de regresión que hemos trabajado hasta el momento:

$$Y(T) - Y(C) = \gamma + \alpha(T) + \beta(D) + \varepsilon$$

Donde:

- Y(T) representa el salario promedio de los participantes del grupo de tratamiento.
- Y(C) es el salario promedio del grupo de control.

[2] Es materialmente imposible contar con toda la población deseada para realizar un estudio. Supongamos que queremos investigar el efecto de un programa educativo en el salario promedio de los participantes. Lo ideal sería poder comparar los cambios en el salario de una persona que ha recibido el curso, y el salario que tendría si no lo hubiese hecho. Naturalmente esto no es posible, una vez que asignamos a una persona al grupo de personas que realizarán el programa no podremos ya saber que hubiera ocurrido si no lo hubiese hecho. Para medir estos efectos, necesitamos comparar el grupo de personas que recibió el programa de entrenamiento con un grupo de control que no lo recibió. De esta forma, el grupo de control nos proporciona una medida de referencia o ”lo que hubiera sucedido en ausencia del tratamiento”. Al comparar los resultados entre el grupo de tratamiento (que recibe el tratamiento) y el grupo de control (que no lo recibe) podemos aislar el efecto causal del tratamiento al controlar otros factores que podrían influir en los resultados.

- γ es la intersección de la línea de regresión.
- α es el coeficiente de la variable de tratamiento (programa de entrenamiento).
- β es el coeficiente de la variable de tiempo.
- D es la variable de participación en el programa de entrenamiento (1 si pertenece al grupo de tratamiento, 0 si pertenece al grupo de control).
- ε es el término de error.

Ahora, para incorporar la variable instrumental Z, podemos modificar la ecuación de regresión de la siguiente manera:

$$Y(T) - Y(C) = \gamma + \alpha(T) + \beta(D) + \delta Z + \varepsilon$$

Donde:

- δ es el coeficiente de la variable instrumental Z.

Al incluir la variable instrumental Z en el modelo, estamos controlando el posible sesgo debido a la endogeneidad. La variable instrumental, en este caso, actúa como una fuente de variación exógena que afecta la participación en el programa de entrenamiento pero no a los salarios directamente.

La validez de una variable instrumental es fundamental, ya que asegura que cumpla su propósito de desligar la relación entre la variable de tratamiento y la variable dependiente de posibles sesgos y confusión causada por factores no observados.

Volvamos al ejemplo anterior en el que, tratando de investigar el efecto causal de un programa educativo en los salarios hemos introducido una variable instrumental puesto que hemos considerado que hay factores no observados como es la motivación individual o las habilidades innatas de cada uno. Para ello hemos introducido una variable instrumental como es la distancia geográfica de la casa del individuo al centro. Esta variable instrumental cumple con la relevancia, ya que está relacionada con la participación en el programa educativo (las personas más cerca del centro tienen más probabilidades de participar) y tiene un impacto directo en esta última.

Además de la relevancia, la validez de una variable instrumental se basa en su exogeneidad, lo cual significa que la variable instrumental debe ser independiente de los factores de error en el modelo de regresión. Continuando con el ejemplo anterior, para que la distancia geográfica sea una variable instrumental válida, debe ser independiente de los factores no observados que puedan afectar tanto la participación en el programa educativo como los salarios, como la motivación individual o las habilidades innatas. Esto asegura que la variable instrumental no se vea afectada por sesgos endógenos y contribuya a estimaciones no sesgadas del efecto causal. Otro aspecto importante es la exclusión, que implica que la variable instrumental no debe tener un efecto directo sobre los salarios.

Como ya sabemos, un método ampliamente utilizado para comprender la dirección (positiva o negativa) y la magnitud de la influencia de la variable independiente sobre la variable dependiente es el de los Mínimos Cuadrados Ordinarios (OLS). Las estimaciones OLS se utilizan para interpretar la relación entre las variables y hacer inferencias sobre los efectos causales. Los coeficientes obtenidos a través de OLS pueden ayudar a responder preguntas como "¿cómo afecta un cambio en la variable independiente al valor esperado de la variable dependiente?" y "¿qué tan fuerte es la relación entre las variables?".

Sin embargo, diferencias significativas en las estimaciones de mínimos cuadrados ordinarios y variables instrumentales, pueden sugerir la falta de validez del instrumento. Podría haber problemas de relevancia si la correlación entre la distancia geográfica y la participación en el programa es débil, lo que significa que el instrumento no captura de manera efectiva la variación en la participación que no está relacionada con otros factores no observados.

Otra posible razón es la violación del supuesto de restricción de exclusión, lo que significa que la distancia geográfica tiene un efecto directo en los salarios, aparte de su influencia en la participación en el programa. Esta situación introduciría sesgo en las estimaciones IV y cuestionaría la validez del instrumento utilizado.

Además, si la distancia geográfica como instrumento es débil, es decir, tiene una baja capacidad para explicar la variación en la participación, esto puede conducir a diferencias

significativas entre las estimaciones OLS y IV y plantear dudas sobre la validez del instrumento.

Como señala Ciacci en su artículo hay una falta de metodología formal que nos permita resolver las dudas sobre la validez del instrumento comparando las estimaciones de OLS e IV. Sin embargo, estas estimaciones, y como trataremos de llevar a cabo más adelante, se deberían comparar objetivamente, a través del **coeficiente de proporcionalidad**, para facilitar pruebas que respalden o descarten la validez del instrumento utilizado en nuestro análisis. Más adelante, desarrollaremos este concepto.

2.2 Principales estudios relacionados

Parte fundamental en la elaboración de este trabajo ha sido el de leer y comprender la literatura actual al respecto del objeto de estudio de este trabajo. Sin una base sólida del estado actual del conocimiento en la materia sería muy complicado profundizar en el estudio de este tema. Para ello, y por recomendación de mi profesor, hemos analizado los siguientes estudios de los cuales me gustaría hacer una revisión de lo más fundamental para poder poner en contexto este trabajo así como señalar el estado actual de la materia y en que sentido, el objetivo de este trabajo, es aportar un valor añadido a estos estudios.

A Matter of Size: Comparing IV and OLS estimates (Ciacci, 2021)

El trabajo de Ricardo Ciacci publicado en mayo de 2021, es la base de este. En este estudio se aborda una cuestión fundamental en la econometría, y que ya hemos mencionado anteriormente, que es la comparación de dos técnicas estadísticas ampliamente utilizadas para la estimación de coeficientes: la regresión lineal ordinaria (OLS) y la regresión instrumental (IV).
Según el trabajo citado, en el campo de la econometría es común considerar diferencias significativas entre los coeficientes estimados mediante Mínimos Cuadrados Ordinarios (OLS) y Variables Instrumentales (IV) como una señal de que el instrumento utilizado puede no ser válido. Esta estrategia se basa en la idea intuitiva de que la regresión OLS puede proporcionar información sobre el verdadero efecto que el investigador desea estimar. Sin embargo, se señala que no existe una metodología formal en la literatura para comparar directamente estas dos estimaciones.

Se cita el trabajo de Oster (2019), que comentaremos más adelante, en el cual se aprovecha la información obtenida de la regresión OLS, como la inclusión de variables de control, el tamaño de las varianzas y los cambios en el coeficiente de determinación (R^2), para estimar un rango de valores en el cual se espera que se encuentre el efecto real del tratamiento. El tamaño de este rango dependerá de la relevancia de las variables observables en relación con las variables no observables, según el criterio del investigador. Esta metodología permitiría al investigador calcular un parámetro llamado coeficiente de proporcionalidad, el cual mide la magnitud relativa de la proporcionalidad entre la influencia de las variables observables y no observables.

El autor de este trabajo sostiene que este enfoque puede ser útil para realizar comparaciones objetivas entre las estimaciones obtenidas mediante variables instrumentales y mínimos cuadrados ordinarios (Ciacci, 2021). Concretamente, dado que las estimaciones IV miden el efecto únicamente para la población cuya elección de tratamiento es afectada por el instrumento, argumenta que valores más altos o bajos del coeficiente de proporcionalidad son evidencia en contra o a favor de la validez del instrumento, respectivamente. La contribución principal de Ciacci (2021) radica en proponer una metodología formal que permita a los investigadores comparar de manera sistemática las estimaciones IV y OLS. Esta propuesta se relaciona de igual manera con otros estudios recientes que han comparado el tamaño relativo de estas dos estimaciones sin utilizar una metodología formal, como con una corriente de investigación que emplea variables observables para evaluar el sesgo ocasionado por variables no observables en entornos de regresión OLS.

Como decíamos anteriormente, un **coeficiente de proporcionalidad** bajo se interpreta como evidencia a favor de las estimaciones IV. En otras palabras, un valor reducido de δ, teniendo como referencia la siguiente ecuación:

$$\delta = (\widehat{\beta_2} - \beta_1)\frac{\mathrm{Var}(d_{ih})\mathrm{Var}(X_{ih})}{\gamma \mathrm{Var}(w_{ih})\mathrm{Cov}(d_{ih}, X_{ih})}$$

indica que se requiere poca influencia de variables no observables para respaldar la idea de que el efecto real coincide con el estimado mediante la regresión IV. Básicamente, δ nos informaría sobre la cantidad relativa de selección en variables no observables en comparación con las observables que se necesita para respaldar la afirmación de que el

"efecto real" tiene la magnitud de las estimaciones IV. Dado que este análisis toma en consideración la inclusión de variables de control, el tamaño de las varianzas y los cambios en el coeficiente de determinación (R^2), entre otros factores, valores altos de δ pueden indicar que el instrumento no es válido o que existen efectos heterogéneos y las estimaciones IV los están capturando para una subpoblación específica. En este tipo de situaciones (es decir, cuando se obtiene un coeficiente de proporcionalidad alto), para distinguir entre ambos casos, el autor menciona que esta metodología podría complementarse utilizando el enfoque propuesto por Masten y Poirier (2018) y realizando un análisis similar al llevado a cabo por Bhuller et al. (2020) para investigar si existen evidencias empíricas en los datos que respalden la idea de que las estimaciones IV están capturando efectos heterogéneos para una subpoblación determinada.

Unobservable Selection and Coefficient Stability: Theory and Evidence (Oster, 2016)

El trabajo de Oster constituye un punto fundamental en nuestro trabajo, ya que sienta las bases y sirve prácticamente de punto de partida para las contribuciones realizadas por Ciacci. Por consiguiente, es igualmente un estudio relevante en nuestro trabajo que merece un análisis detallado del mismo.

En este trabajo, Emily Oster, aborda el problema de la selección no observada en modelos econométricos y su impacto en la estabilidad de los coeficientes estimados a lo largo del tiempo. La autora comienza explicando que la selección no observada ocurre cuando una variable relevante que influye en el resultado de interés no está disponible o no se mide correctamente. Este problema puede ser especialmente relevante en estudios longitudinales, donde la selección a lo largo del tiempo puede afectar los resultados.

El documento discute las inquietudes relacionadas con el sesgo de variable omitida, algo común en la mayoría de los estudios no experimentales en economía. Se sugiere abordar estas preocupaciones al incluir controles observables en el análisis. No obstante, cabe destacar que en muchos casos los controles observados son solo una aproximación incompleta de la variable omitida real. Se menciona en el trabajo que algunos investigadores argumentan que ciertos controles capturan completamente una variable omitida específica. Sin embargo, esto suele ser la excepción y no la norma.

El enfoque comúnmente utilizado consiste en evaluar la sensibilidad de los efectos del tratamiento al incluir los controles observados. Si el coeficiente del tratamiento se mantiene estable después de incluir estos controles, se considera que el sesgo de variable omitida es limitado. La idea subyacente es que si los controles observados, aunque imperfectos, no tienen un impacto significativo en el coeficiente, es probable que las variables no observadas tampoco lo tengan.

No obstante, el trabajo señala que esta intuición no se deriva directamente de las suposiciones básicas del modelo lineal. Para obtener inferencias válidas sobre el sesgo de variable omitida a partir de los controles observados, se requieren suposiciones adicionales sobre las propiedades de covarianza de los dos conjuntos de variables. Incluso en el caso más optimista, donde el sesgo de variable omitida se puede identificar por completo mediante los controles observados, los cambios en el coeficiente por sí solos no son suficientes para calcular el sesgo.

El trabajo propone un método desarrollado por Altonji, Elder y Taber para evaluar la robustez de los resultados, asumiendo que se puede recuperar la relación entre el tratamiento y las variables no observadas a partir de la relación entre el tratamiento y las variables observadas. Este método utiliza un estadístico de prueba que evalúa la importancia relativa de las variables no observadas en comparación con las observadas para eliminar el efecto observado. Sin embargo, este enfoque no se utiliza ampliamente en la investigación empírica en economía.

Adicionalmente, el trabajo extiende la metodología para evaluar la robustez del sesgo de variable omitida bajo la suposición de que la relación entre el tratamiento y las variables no observadas se puede recuperar a partir de la relación entre el tratamiento y las variables observadas proporcionando un estimador consistente del efecto del tratamiento ajustado por sesgo y propone estándares de robustez basados en datos aleatorizados. En el trabajo se realizan validaciones empíricas del método propuesto utilizando datos reales.

Para analizar cómo la selección no observada afecta la estabilidad de los coeficientes estimados en el tiempo, la autora propone un marco teórico basado en los "límites de selección". Estos límites representan la distribución de las variables no observadas que influyen en la selección a lo largo del tiempo.

La aplicación empírica se enfoca en la relación entre la exposición al plomo durante la infancia y los resultados educativos posteriores en los Estados Unidos. Utilizando datos longitudinales de una cohorte de niños nacidos en las décadas de 1970 y 1980, Oster encuentra que la selección no observada puede ser un problema importante en este contexto además de realizar pruebas de robustez confirmando la consistencia de los resultados en diferentes especificaciones del modelo y en diferentes cohortes de niños. De esta forma, se resalta la importancia de considerar cuidadosamente la selección no observada en modelos econométricos y cómo esto puede afectar la estabilidad de los coeficientes estimados a lo largo del tiempo.

Estimating and Testing Models with Many Treatment Levels and Limited Instruments (Moretti & Lochner, 2014)

En este trabajo, Lochner y Moretti abordan diversos aspectos significativos. En primer lugar, destacan que, en la aplicación práctica, las estimaciones obtenidas mediante los métodos de mínimos cuadrados ordinarios (OLS) y variables instrumentales (IV) suelen ser diferentes. Esta disparidad de resultados puede resultar sorprendente teniendo en cuenta las suposiciones plausibles basadas en la teoría económica acerca del sesgo de endogeneidad.

El estudio se enfoca en un tipo específico de modelos que involucran un regresor endógeno discreto con un conjunto finito de valores, regresores exógenos que pueden ser separados de forma aditiva y coeficientes que no varían entre individuos. Estos modelos son ampliamente utilizados en la investigación empírica para analizar los efectos de tratamientos con múltiples valores, en niveles de dosificación de medicamentos y logros educativos.

Un aspecto relevante que analizan los autores, es considerar la posibilidad de que los efectos del tratamiento por unidad varíen según los niveles de tratamiento. Dado el creciente enfoque en la identificación de efectos causales en economía, muchos investigadores recurren a métodos de variables instrumentales para estimar modelos de este tipo. Sin embargo, debido a la disponibilidad limitada de instrumentos válidos, es común asumir efectos uniformes de tratamiento por unidad, incluso cuando es probable

que estos efectos varíen según los niveles de tratamiento, como sugieren especificaciones más generales estimadas mediante OLS.

Lochner y Moretti demuestran que, en este escenario, los estimadores OLS y IV identifican promedios ponderados diferentes de todos los efectos por unidad, lo cual puede conducir a conclusiones incorrectas sobre la endogeneidad al utilizar una prueba convencional de Hausman.

La principal contribución de este trabajo radica en el desarrollo de una generalización sencilla de la prueba de Hausman. Esta prueba permite evaluar si la ponderación diferencial y los efectos de tratamiento por unidad variable pueden explicar la discrepancia entre los estimadores OLS y IV. Dentro del conjunto de modelos considerados, esta prueba funciona como una prueba de especificación para evaluar la exogeneidad bajo condiciones razonables. Una ventaja de esta prueba es que solo requiere un único instrumento, lo que la hace útil en diversas aplicaciones.

The Colonial Origins of Comparative Development: An Empirical Investigation (Acemoglu, Johnson, & James, 2001)

Este trabajo aborda la pregunta de por qué algunos países son más ricos y prósperos que otros. Los autores argumentan que la razón subyacente es la forma en que los países fueron colonizados. En particular, argumentan que los países que fueron colonizados por potencias europeas que establecieron instituciones políticas y económicas inclusivas, en las que las elites locales podían participar y compartir el poder, han tenido un mejor desempeño económico que aquellos que fueron colonizados por potencias que establecieron instituciones extractivas, en las que el poder y los recursos fueron monopolizados por la élite colonial.

En este estudio, se argumenta que las diferencias en la experiencia colonial podrían ser una fuente de variación exógena en las instituciones. A la hora de estimar el impacto de las instituciones en el desempeño económico, se aprovechan estas diferencias como una fuente de variación exógena. Y lo cierto es que se demuestra que podemos encontrar una alta correlación entre las tasas de mortalidad enfrentadas por soldados, obispos y marineros en las colonias y los asentamientos europeos, entre los asentamientos europeos y las primeras medidas de instituciones, y entre las instituciones tempranas e instituciones

actuales. Mediante esta fuente de variación, se estiman efectos significativos de las instituciones en el ingreso per cápita. Además, se demuestra en este trabajo que esta relación no está impulsada por valores atípicos y es robusta al controlar la latitud, el clima, el entorno de enfermedades actual, la religión, los recursos naturales, la calidad del suelo, la fragmentación etnolingüística y la composición racial actual.

Es importante destacar que los hallazgos no implican que las instituciones actuales estén predeterminadas por las políticas coloniales y no puedan cambiarse. Los autores enfatizan que la experiencia colonial es uno de los muchos factores que influyen en las instituciones. Puesto que las tasas de mortalidad enfrentadas por los colonos se consideran exógenas, estas se utilizan como instrumento para aislar el efecto de las instituciones en el desempeño económico. De hecho, se sugiere que los resultados indican ganancias económicas sustanciales derivadas de la mejora de las instituciones, como se observó en casos como la Restauración Meiji en Japón o Corea del Sur durante la década de 1960.

Sin embargo, este análisis plantea algunas preguntas que no se abordan en este trabajo. Las instituciones se tratan en gran medida como una "caja negra": los resultados indican que reducir el riesgo de expropiación (o mejorar otros aspectos del "conjunto de instituciones") resultaría en ganancias significativas en el ingreso per cápita, pero no señalan qué pasos concretos llevarían a una mejora en estas instituciones. Se llega a la conclusión de que es importante realizar un análisis más detallado del efecto de instituciones más fundamentales en los derechos de propiedad y el riesgo de expropiación, como por ejemplo, el sistema presidencial versus parlamentario, que pueden cambiarse. Un análisis más detallado de cómo estas instituciones fundamentales afectan los derechos de propiedad y el riesgo de expropiación es sin duda un área importante para futuros estudios.

Salvaging Falsified Instrumental Variable Models (Masten & Poirier, 2020)

Frente a la refutación de su modelo base, Masten y Poirier se proponen dar una respuesta sistemática y constructiva. En lugar de tratar la falsificación como un inconveniente a ignorar o como una falla fatal que condena un estudio, se enfoca en lo que se puede aprender de los modelos falsificados. Para ello, en el trabajo se presentan cuatro recomendaciones las cuales mencionaremos a continuación.

En primer lugar, sugieren medir el grado de falsificación mediante la definición de relajaciones continuas de las suposiciones clave de identificación de interés, y relajando estas suposiciones hasta que el modelo ya no sea falsificado. El conjunto de puntos en los cuales esto ocurre se denomina "falsification frontier".

En segundo lugar, proponen presentar el "falsification adaptive set", que es el conjunto identificado para el parámetro de interés bajo la suposición de que el verdadero modelo se encuentra en algún punto de la "falsification frontier". Afirman además que esta recomendación es una generalización de la práctica estándar para modelos base no refutados y no requiere que el investigador seleccione o calibre parámetros de sensibilidad.

Además, sugieren presentar el conjunto identificado para puntos de interés específicos en la "falsification frontier". Y por último, proponen presentar conjuntos identificados para puntos más allá de la "falsification frontier" como un análisis de sensibilidad adicional.

Estas cuatro recomendaciones se ilustran en dos modelos de variables instrumentales sobreidentificados diferentes. En el primer modelo, se imponen efectos de tratamiento homogéneos pero se permiten tratamientos continuos, mientras que en el segundo modelo se permiten efectos de tratamiento heterogéneos pero se enfoca en tratamientos binarios. En ambos modelos, se observan múltiples instrumentos y las suposiciones clave de identificación son la exclusión y exogeneidad de cada instrumento. Se consideran relajaciones continuas de estas suposiciones y se caracteriza la "falsification frontier", el "falsification adaptive set" y los conjuntos identificados para puntos más allá de la frontera.

En las aplicaciones empíricas realizadas, se destaca que el "falsification adaptive set" complementa la tradicional prueba de sobreidentificación en términos de valores *p*. También resumen directamente el rango de estimaciones correspondientes a modelos alternativos no falsificados a la vez que enfatizan la importancia de controlar los instrumentos posiblemente inválidos al considerar modelos alternativos.

Using instrumental variables to establish causality (Becker, 2016)

En este artículo se explora el uso de variables instrumentales como una herramienta para establecer relaciones de causalidad en la investigación económica. Sascha, acertadamente

plantea que en muchas situaciones del mundo real resulta difícil llevar a cabo experimentos controlados que permitan determinar las relaciones causales entre variables, por lo que se recurre a modelos econométricos basados en datos observados. Sin embargo, como se mencionó al inicio de nuestro trabajo, estos modelos pueden enfrentar problemas de endogeneidad y sesgo de selección, lo que puede conducir a estimaciones incorrectas de las relaciones causales.

Sascha propone que al utilizar variables instrumentales de manera adecuada, es posible obtener estimaciones más precisas de las relaciones causales entre las variables observables en este tipo de situaciones. A través de ejemplos de casos conocidos en la literatura económica, como los estudios sobre la relación entre educación e ingresos o la inversión en infraestructura y crecimiento económico, Sascha profundiza y refuerza esta idea.

Además, se discuten los problemas que pueden surgir al utilizar variables instrumentales, como la selección de instrumentos inadecuados o la falta de validez de los supuestos del modelo. Este artículo resalta la importancia de emplear variables instrumentales en la investigación económica y proporciona orientación útil para la selección y validación de instrumentos apropiados en este tipo de análisis.

Selection on observed and unobserved variables: assessing the effectiveness of catholic schools (Altonji, Elder, & Taber, 2000)

El trabajo se centra en el uso de variables instrumentales para establecer relaciones de causalidad en la investigación económica, específicamente en el efecto de las escuelas católicas. El estudio aborda tres premisas fundamentales. En primer lugar, se cuestiona la confiabilidad de las restricciones de exclusión utilizadas en estudios anteriores para identificar el efecto de las escuelas católicas. En segundo lugar, se destaca la importancia de contar con un conjunto amplio de variables de control y un grupo de estudiantes que no difieran significativamente en su asistencia a escuelas católicas de secundaria. Esto les lleva a centrarse en estudiantes de octavo grado de escuelas católicas. Esta elección permite evitar preocupaciones sobre la falta de comparabilidad entre el reducido grupo de estudiantes de escuelas primarias públicas que asisten a escuelas católicas de secundaria y otros estudiantes. Además, permite aislar el efecto de las escuelas católicas de secundaria del efecto de las escuelas de primaria. La tercera premisa sostiene que el

grado de selección en las variables observables proporciona información sobre la selección en las características no observables. El trabajo se enfoca en formalizar el uso de esta información y proporcionar una forma de evaluar cuantitativamente el sesgo de selección.

Los autores, en este trabajo, llegan a varias conclusiones. Primero, nos encontramos con que asistir a una escuela católica de secundaria aumenta significativamente las tasas de graduación. Se estima que el efecto de las escuelas católicas en la graduación de la escuela secundaria es robusto y no se explica por los resultados de octavo grado ni por antecedentes familiares. Sin embargo, se reconoce que las estimaciones que tratan la asistencia a escuelas católicas como exógena probablemente sobreestiman el efecto real de las escuelas católicas, y se considera que el grado de selección en las variables no observables tendría que ser mucho más fuerte que en las variables observables para explicar completamente el efecto.

Segundo, hay poca evidencia de que las escuelas católicas de secundaria mejoren los puntajes de lectura. De hecho, la mayoría de las estimaciones muestran efectos negativos. Las estimaciones de una sola ecuación indican un efecto positivo en los puntajes de matemáticas de aproximadamente 0.1 desviaciones estándar, pero dada la presencia de error de muestreo y evidencia de sesgo de selección positivo, no se tienen suficientes pruebas de que las escuelas católicas de secundaria mejoren los puntajes de las pruebas tanto como las tasas de graduación.

Tercero, se observa que la asistencia a escuelas católicas de secundaria aumenta sustancialmente la probabilidad de graduación de la escuela secundaria para estudiantes de minorías urbanas. Las estimaciones de una sola ecuación también indican un impacto muy grande en la asistencia a la universidad, pero el grado de selección positiva en las variables observables que determinan la asistencia a la universidad es lo suficientemente grande como para no descartar el sesgo de selección como explicación total del efecto de las escuelas católicas en la asistencia a la universidad. Se menciona que el tamaño de la muestra de minorías urbanas que asistieron a escuelas católicas de octavo grado no es suficientemente grande para realizar un análisis similar.

En general, se concluye que parte del efecto de las escuelas católicas en la graduación de la escuela secundaria y la asistencia a la universidad es probablemente real, pero la evidencia es menos sólida en el caso de la asistencia a la universidad.

Comments on 'Unobservable Selection and Coefficient Stability: Theory and Evidence' and 'Poorly Measured Confounders are More Useful on the Left Than on the Right (De Luca, Magnus, & Peracchi, 2018)

Se menciona en este artículo que los artículos de Oster y PPS contribuyen al desarrollo de métodos de inferencia sobre efectos causales en situaciones desafiantes y empíricamente relevantes en las que el proceso de generación de datos desconocido no está incluido en los modelos de regresión considerados por el investigador. Mientras que Oster analiza el sesgo de variables omitidas en la estimación de un efecto causal de interés y la estabilidad de los coeficientes en los modelos de regresión; por el otro lado, PPS analiza las propiedades de poder de dos estrategias alternativas para probar la consistencia del estimador OLS del efecto causal cuando los controles en el modelo intermedio están sujetos a errores de medición. Ambos artículos comparan el sesgo o la varianza muestral de los estimadores OLS en modelos incorrectamente especificados con diferentes conjuntos de regresores.

Al final del artículo, se destaca que la proposición de Oster ofrece un resultado preciso pero requiere supuestos fuertes y conocimiento del parámetro clave R^2_{max}. Reformulando esta proposición, Oster consigue debilitar algunos de estos supuestos pero también requiere conocimiento de R^2_{max} y del parámetro adicional ϕ. A pesar de los supuestos fuertes, la caracterización del sesgo del estimador OLS no restringido como una raíz de una ecuación cúbica es útil, aunque cuando esta ecuación tiene tres raíces, no está claro cuál seleccionar. El trabajo no resuelve este problema ni ofrece otras formas de corregir el sesgo, pero se espera que ayude a clarificar la naturaleza de los supuestos de Oster y a evaluar correctamente los resultados de su rutina en Stata. En cuanto a las estrategias de prueba en PPS, también requieren restricciones, pero en este caso los supuestos son menos y más transparentes, lo que facilita a los profesionales verificar si se cumplen realmente.

2.3 Relación de aportaciones

En el ensayo del profesor Ciacci, se examina en detalle el concepto de endogeneidad en los modelos econométricos, que se refiere a la existencia de una relación simultánea entre la variable dependiente y una o más variables independientes. Por otro lado, en el ensayo de la profesora Oster se aborda el concepto de selección no observada, que se refiere a la presencia de factores no medidos que pueden influir tanto en la variable dependiente como en la elección de las variables independientes.

La conexión fundamental entre ambos trabajos radica en que tanto la endogeneidad como la selección no observada pueden tener un impacto significativo en la estimación de los parámetros en modelos econométricos. Estos fenómenos pueden introducir sesgos en las estimaciones y conducir a resultados inconsistentes o poco confiables.

La endogeneidad puede surgir cuando las variables independientes están correlacionadas con el término de error del modelo, lo que dificulta determinar la verdadera relación causal entre las variables. Por su parte, la selección no observada puede surgir cuando existe una heterogeneidad no capturada entre los individuos o unidades de estudio, lo que puede conducir a una distorsión en las estimaciones de los parámetros.

Ambos problemas requieren una atención cuidadosa y la aplicación de métodos apropiados para mitigar su efecto en los resultados de los modelos econométricos. Es crucial buscar enfoques que permitan abordar estos desafíos de manera rigurosa y confiable.

Sin embargo, al examinar la literatura existente en el campo de la econometría, se observa una notable ausencia de una metodología formal ampliamente aceptada para abordar de manera sistemática la endogeneidad y la selección no observada. Aunque muchos estudios reconocen y enfrentan estos desafíos, no se ha establecido un enfoque estándar que proporcione una solución definitiva.

En consecuencia, existe una necesidad apremiante de desarrollar y promover una metodología formal que permita resolver de manera efectiva los problemas de endogeneidad y selección no observada en los modelos econométricos. Esto contribuiría a mejorar la calidad y confiabilidad de las estimaciones, así como a fortalecer las conclusiones y las inferencias extraídas de los análisis econométricos.

En resumen, tanto la endogeneidad como la selección no observada representan desafíos importantes en la estimación de los parámetros en modelos econométricos. Aunque se han realizado esfuerzos para abordar estos problemas en la literatura, no se ha establecido una metodología formal ampliamente aceptada para resolverlos de manera sistemática. Es necesario fomentar la investigación y el desarrollo de enfoques rigurosos que permitan superar estos desafíos, a fin de mejorar la calidad y confiabilidad de los análisis econométricos y las conclusiones derivadas de ellos.

3. DE ECUACIÓN A MATRIZ

3.1 Análisis crítico del artículo

Una vez hemos repasado el marco teórico de este trabajo así como la literatura relevante al mismo, podemos proceder a desarrollar matemáticamente nuestro trabajo, o la parte más importante del mismo.

Anteriormente, haciendo un repaso del marco teórico y para empezar a familiarizarnos con la notación matemática, hemos presentado una ecuación de regresión lineal muy sencilla. Esta ecuación, si bien no es tan sencilla como la primera será nuestra la ecuación en la que nos apoyaremos para desarrollar nuestra regresión:

$$y_{ih} = \alpha_1 + \beta_1 * d_{ih} + \gamma * w_{ih} + \theta_1 * X_{ih} + \varepsilon_{1ih}$$

Donde:

- *i* representa la unidad
- *h* representa el tiempo
- *d* es el escalar que hace de tratamiento
- *w* el vector de los controles no observables. Esto no podrá formar parte de la ecuación al ser calculada
- *X* representa el vector que contiene los controles observados

En el caso univariante, la regresión consiste en encontrar una línea recta que relacione una variable dependiente con una variable independiente X. Esta línea recta se caracteriza por dos parámetros: α, que representa el punto de intersección con el eje *y*, y β_1, que representa la pendiente de la recta. Además de la variable independiente *X*, estos parámetros son parte de la solución obtenida mediante la regresión. Gráficamente, la regresión se representa mediante una recta que se ajusta a los datos. A partir de esta regresión, se pueden realizar diversas aplicaciones, como se ha observado en investigaciones anteriores.

Recordando lo ya visto sobre las asunciones expuestas por Oster, llegamos a una relación de selección proporcional como sigue:

$$\delta \frac{Cov(d_{ih}, X_{ih})}{Var(X_{ih})} = \frac{Cov(d_{ih}, X_{ih})}{Var(X_{ih})}$$

La cual se cumplirá siempre que δ sea diferente a 0. Sabiendo que *w* es el único control, encontramos el sesgo de la variable omitida:

$$\hat{\beta}_2 = \beta_1 + \gamma * \frac{Cov(d_{ih}, w_{ih})}{Var(w_{ih})}$$

Si seguimos la misma lógica para el caso de regresión lineal múltiple, teniendo en cuenta que d_{ih} (cuyo significado explicaremos en el siguiente apartado) es igual a $\mathrm{d} - \tau + \tau \mathrm{X}$ que el sesgo de variable omitida es:

$$\hat{\beta}_2 = \beta_1 + \gamma * \frac{Cov(d_{ih}, w_{ih}) - \tau * Cov(X_{ih}, w_{ih})}{Var(\tilde{d}_{ih})}$$

Tomando esta ecuación e introduciendo las conclusiones de la relación de selección proporcional expuestas por Oster anteriormente llegamos, en nuestro caso de regresión lineal múltiple, a el coeficiente de proporcionalidad:

$$\boldsymbol{\delta = (\hat{\beta}_2 - \beta_1) + \frac{Var(d_{ih}) * Var(X_{ih})}{\gamma * Var(w_{ih}) * Cov(d_{ih}, X_{ih})}}$$

De esta manera, obtenemos δ. Este coeficiente de proporcionalidad, al que ya nos hemos referido anteriormente, es el centro de este trabajo pues sería la herramienta necesaria con la cual poder comparar las estimaciones obtenidas de OLS e IV y poder así dar objetivizar e instaurar una metodología formal con la que evaluar la validez del instrumento.

El coeficiente de proporcionalidad (δ) es una medida de la necesidad de selección de variables no observables para respaldar las estimaciones de la regresión de variables instrumentales (IV) en un estudio empírico. Básicamente, δ nos informaría sobre la cantidad relativa de selección en variables no observables en comparación con las observables que se necesita para respaldar la afirmación de que el "efecto real" tiene la

magnitud de las estimaciones IV. Dado que este análisis toma en consideración la inclusión de variables de control, el tamaño de las varianzas y los cambios en el coeficiente de determinación (R^2), entre otros factores, valores altos de δ pueden indicar que el instrumento no es válido o que existen efectos heterogéneos y las estimaciones IV los están capturando para una subpoblación específica.

En contextos empíricos, es de vital importancia determinar el signo de δ, ya que esto nos permite calcular los conjuntos identificados. En términos simplificados, el signo de δ está condicionado, en parte, por la suposición que hagamos sobre el signo de γ.

La razón subyacente es que las varianzas siempre son positivas y tanto la diferencia de betas como la covarianza pueden estimarse utilizando los datos disponibles en el estudio en cuestión. Una vez que conocemos el signo de δ, podemos calcular los conjuntos identificados para los coeficientes de proporcionalidad correspondientes y examinar cómo varían los límites del conjunto a medida que el coeficiente de proporcionalidad cambia. En la siguiente sección, pondremos esto en práctica.

Por último, es importante mencionar el valor que elegimos para R^2, ya que es crucial para estimar los conjuntos identificados. Este valor seleccionado, conocido como R_{max}según Oster (2019), se elige en función del conocimiento previo que tenemos sobre el entorno del estudio. Si el investigador considera de manera objetiva que la regresión puede explicar completamente la variable de resultado, entonces R_{max} se establece en 1.

R_{max} es un valor utilizado en el contexto de la estimación de conjuntos identificados en análisis empíricos. Se refiere al valor máximo que se elige para el coeficiente de determinación R^2 en un modelo de regresión.

En términos más simples, R^2 es una medida que indica qué tan bien se ajusta un modelo de regresión a los datos observados. Toma valores entre 0 y 1, donde 1 indica un ajuste perfecto del modelo a los datos. Sin embargo, en algunos casos, especialmente en estudios empíricos con datos limitados, puede ser difícil alcanzar un ajuste perfecto.

Por lo tanto, se introduce el concepto de R_{max}, que representa el valor más alto que se considera plausible o realista para R^2 en el contexto específico del estudio. Este valor se elige según el conocimiento previo o las suposiciones sobre el entorno en el que se realiza la investigación.

Al establecer un límite superior para R^2a través de R_{max}, se reconoce que no se espera que el modelo de regresión explique la totalidad de la variabilidad de la variable de resultado. Al limitar R^2, se tiene en cuenta la incertidumbre inherente a los datos y se evitan conclusiones excesivamente optimistas o infladas sobre la capacidad explicativa del modelo.

3.2 Identificación de puntos de expansión

Después de llevar a cabo un exhaustivo repaso de la literatura académica, se ha observado que a lo largo del tiempo se han realizado avances en la introducción de metodologías para abordar los problemas de endogeneidad y selección no observada en modelos econométricos. También, hemos podido comprobar que la metodología propuesta por Ciacci no ha sido ampliamente adoptada por los estudiosos en la actualidad.

En la mayoría de los casos, los enfoques utilizados permiten determinar de manera general si el estimador es creíble en términos de representar un efecto verdadero. Sin embargo, no proporcionan una precisión detallada en esta evaluación. Por lo tanto, se reconoce la necesidad de desarrollar una metodología que, aprovechando los últimos avances en la materia y basándose en la ecuación de Ciacci, simplifique su aplicación y facilite su puesta en práctica.

En este trabajo, no se propone una metodología completamente nueva para abordar la endogeneidad y la selección no observada. En cambio, se busca ofrecer una presentación más accesible y una forma más sencilla de aplicar la metodología existente, teniendo en cuenta los avances recientes. El objetivo es hacer que la metodología basada en la ecuación de Ciacci sea más fácilmente implementable por los investigadores, brindando un enfoque práctico y concreto para abordar los desafíos metodológicos en la estimación de los parámetros en modelos econométricos.

Al hacerlo, se espera fomentar una mayor adopción y aplicación de la metodología de Ciacci, brindando a los estudiosos una herramienta más accesible y robusta para enfrentar los problemas de endogeneidad y selección no observada en sus investigaciones. Esta simplificación y adaptación de la metodología existente permitirá a los investigadores obtener resultados más confiables y precisos, fortaleciendo así las bases teóricas y empíricas de sus estudios en el campo de la econometría.

3.3 Beta en forma matricial

El enfoque de Ciacci se basa en la premisa de que, al utilizar variables instrumentales, es fundamental contar con un instrumento válido y relevante que no esté correlacionado con el término de error en el modelo. La ecuación del coeficiente de proporcionalidad permite analizar la relación entre las estimaciones de los parámetros obtenidos a través de IV y OLS, brindando así una forma de evaluar la validez del instrumento utilizado en el modelo.

Para avanzar en nuestro estudio, proponemos desarrollar y reescribir la ecuación del coeficiente de proporcionalidad en forma matricial. Esta reescritura nos permitirá utilizar herramientas y técnicas de álgebra lineal para analizar y comparar las estimaciones obtenidas mediante IV y OLS de manera más eficiente y precisa.

Al adoptar esta metodología formal y presentarla en una forma matricial, esperamos proporcionar a los investigadores una herramienta más sólida y accesible para evaluar la validez del instrumento, así como para realizar comparaciones rigurosas entre las estimaciones de los parámetros obtenidas a través de IV y OLS.

En nuestro enfoque, estableceremos inicialmente las dimensiones de la nueva fórmula que utilizaremos. Dichas dimensiones se especifican de la siguiente manera:

- La matriz Y (n x 1) representa la variable dependiente, con n observaciones.
- La matriz D (n x 1) se refiere al tratamiento.
- La matriz W (n x 1) corresponde a los controles no observados.
- La matriz X (n x k) representa los controles observados, con k variables observables y n observaciones.
- Los coeficientes $\beta 1$ y $\beta 2$ (1 x 1) son escalares que reflejan los efectos del tratamiento.
- Los vectores de coeficientes $\theta 1$ y $\theta 2$ (k x 1) representan los coeficientes de los controles observados.
- El coeficiente γ (1 x 1) es un escalar que refleja el efecto del control no observado.

Con esta especificación de dimensiones, estaremos preparados para desarrollar la fórmula en cuestión y aplicarla en nuestros análisis posteriores. Una vez dicho lo cual, podemos pasar a presentar nuestra ecuación:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \mathrm{D}_1 \\ \vdots \\ \mathrm{D}_n \end{bmatrix} \beta_1 + \begin{bmatrix} \omega_1 \\ \vdots \\ \omega_n \end{bmatrix} \gamma + \begin{bmatrix} X_{11} & \cdots & X_{1k} \\ \vdots & \ddots & \vdots \\ X_{n1} & \cdots & X_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

De manera simplificada podemos mostrar la ecuación matricial así:

$$Y = D\beta_1 + \omega\gamma + \mathrm{X}\theta_1 + \varepsilon_1$$

Debido a la presencia de variables omitidas, nos vemos limitados a realizar la regresión que se define mediante la siguiente ecuación:

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

Además, es importante tener en cuenta las dimensiones de las matrices de covarianza[3] y varianza[4], las cuales se definen de la siguiente manera:

- Σ_{DX} (1 x k): Esta matriz representa la covarianza cruzada entre D (variable de tratamiento) y X (variables observables). Es un vector que contiene las covarianzas entre la variable de tratamiento y todas las variables observables del modelo.
- Σ_{DW} (1 x 1): Esta matriz representa la covarianza entre D (variable de tratamiento) y W (controles no observados). Es equivalente al escalar Cov(D, W).
- Σ_D (1 x 1): Esta matriz representa la varianza de D (variable de tratamiento).
- Σ_X (k x k): Esta matriz representa la matriz de varianza-covarianza de X (controles observados). Tiene dimensiones k x k, donde k es el número de variables observables.

[3] La covarianza se utiliza para analizar la asociación entre una variable independiente (X) y la variable dependiente (Y) La covarianza nos proporciona información sobre cómo se mueven conjuntamente X e Y, y si tienden a variar en la misma dirección (covarianza positiva) o en direcciones opuestas (covarianza negativa) La covarianza se utiliza para calcular el coeficiente de correlación, que indica la fuerza y la dirección de la relación lineal entre X e Y. Si la covarianza es positiva y alta, significa que a medida que los valores de X aumentan, los valores de Y también tienden a aumentar, lo que indica una relación directa. Por otro lado, si la covarianza es negativa y alta, significa que a medida que los valores de X aumentan, los valores de Y tienden a disminuir, lo que indica una relación inversa.

[4] La varianza en una regresión nos proporciona información sobre la dispersión de los datos alrededor de la línea de regresión y nos ayuda a evaluar la calidad del ajuste del modelo y la precisión de las predicciones. Es una medida fundamental para comprender la variabilidad y la confiabilidad de los resultados en el análisis de regresión lineal.

- Σ_W (1 x 1): Esta matriz representa la varianza de W (controles no observados).

Con estas dimensiones definidas, procederemos a introducir un coeficiente de proporcionalidad adicional para cada variable de control, a fin de cumplir con la ecuación de proporcionalidad con la variable omitida. En consecuencia, la ecuación de proporción de selección se expresa de la siguiente manera:

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

Considerando X_j as el vector con n observaciones de la variable en la columna j de la matriz de observables X, podemos expresarlo de manera matricial de la siguiente forma:

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

En este contexto, la matriz δ (1 x k) está compuesta por todos los coeficientes de proporción correspondientes a las variables observables. Cabe destacar que el número de coeficientes en esta matriz es igual al número de variables observables, es decir, cada coeficiente se relaciona con una variable específica.

La ecuación propuesta por Ciacci en el apéndice (A.1) establece que el sesgo causado por una variable omitida se expresa de la siguiente manera:

$$\hat{\beta}_2 = \beta_1 + \frac{\gamma\ Cov(d_{ih}, \omega_{\mathrm{ih}}) - \tau_1\ Cov(X_{ih}, w_{ih})}{Var(\tilde{d}_{ih})}$$

Teniendo en cuenta las suposiciones de ortogonalidad entre las variables observables y no observables, asumidas según las premisas de Oster $Cov(X_{ih}, w_{ih}) = 0$ podemos proceder a reescribir la ecuación de regresión de d sobre X en forma matricial. Para ello, consideramos que $\tilde{d}$ son los residuales de la regresión de d sobre X, es decir, $d_{ih} = \tau_0 + \tau_1 X_{ih} + \tilde{d}_{ih}$.

La ecuación de regresión de d sobre X en forma matricial se expresa como $D = X\mathrm{T}_1 + \widetilde{D}$, donde D y X tienen las dimensiones mencionadas anteriormente y $\widetilde{D}$ tiene una dimensión de n x 1. El vector de coeficientes T_1 tiene dimensiones de k x 1.

Al reescribir la ecuación A.1 en forma matricial, es necesario definir $\Sigma_{\tilde{D}}$(1 x 1), que representa la matriz de varianza de $\widetilde{D}$, y Σ_{XW}(k x 1), que representa la matriz de covarianza cruzada entre X y W, es decir, el vector que contiene las covarianzas entre las variables observables y todas las variables omitidas.

De esta manera, la ecuación A.1 se redefine de forma matricial como

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

Considerando que hemos asumido que las variables omitidas son ortogonales a las variables de control y, por lo tanto, sus varianzas son cero, la ecuación A.1 simplificada en forma matricial queda:

$$\hat{\beta}_2 = \beta_1 + (\gamma \Sigma_{DW})\Sigma_{\tilde{D}}^{-1}$$

Finalmente, al combinar la ecuación de proporción de selección en forma matricial con la ecuación anterior y despejar para δ_j, obtenemos la ecuación final de Ciacci en forma matricial:

$$\boldsymbol{\delta_j = (\hat{\beta}_2 - \beta_1)\frac{1}{\gamma}\Sigma_{\tilde{D}}\Sigma_{Xj}\Sigma_{DXj}^{-1}\Sigma_W^{-1}}$$

3.4 Ventajas de la matriz frente a la ecuación

La formulación matricial de la ecuación delta en una regresión multivariante presenta diversas ventajas y mejoras en comparación con su formulación en forma de ecuación.

En primer lugar, al expresar la ecuación delta en forma matricial, se logra una visión más estructurada y concisa de las relaciones entre las variables involucradas. La representación matricial revela una estructura de datos más clara y facilita la comprensión de las interacciones entre las diferentes variables. Esta forma matricial nos permite apreciar la naturaleza multivariante del problema y obtener una perspectiva más completa de las relaciones entre las variables explicativas y la variable dependiente.

Además, al tener la ecuación en forma matricial, se pueden aprovechar las propiedades algebraicas de las matrices, lo que conlleva a realizar manipulaciones matemáticas y cálculos de manera más eficiente. Podemos aplicar técnicas y métodos específicos para el análisis de matrices, como la descomposición espectral o la diagonalización, lo que nos proporciona herramientas adicionales para comprender y analizar la estructura de la ecuación.

Otra ventaja significativa de la representación matricial es la capacidad de tratar conjuntamente múltiples observaciones y estimaciones. Al agrupar todas las observaciones y coeficientes de regresión en matrices, se pueden realizar análisis estadísticos más completos. Se pueden calcular fácilmente la matriz de covarianza, realizar pruebas de hipótesis conjuntas y evaluar la calidad del ajuste mediante el análisis de residuos y otras medidas de bondad de ajuste.

Además, la formulación matricial permite extender el marco de análisis a situaciones más complejas, como modelos de regresión multivariante con más de una variable dependiente. Al emplear matrices, se puede abordar de manera más eficiente y sistemática problemas que involucran múltiples variables de respuesta y múltiples variables explicativas.

En resumidas cuentas, la formulación matricial de la ecuación delta en una regresión multivariante ofrece una representación más clara, estructurada y eficiente de las relaciones entre las variables. Proporciona herramientas adicionales para el análisis estadístico, facilita el cálculo y la manipulación algebraica, y permite una comprensión más profunda y completa del modelo. Por lo tanto, esta formulación matricial es preferible a la formulación en forma de ecuación, ya que nos brinda una mayor capacidad analítica y nos permite abordar problemas más complejos en el contexto de la regresión multivariante.

4. CONCLUSIONES

A lo largo de este trabajo nos hemos centrado en abordar el desafío de la endogeneidad en la estadística y en los modelos de inferencia causal en econometría. Para enfrentar el desafío que este problema muchas veces representa, se utilizan, entre otras, técnicas como el de las variables instrumentales. Sin embargo, durante nuestra revisión exhaustiva de la literatura, nos hemos dado cuenta de que aún no existe una metodología formalmente establecida para evaluar la validez de los instrumentos utilizados en este contexto.

Tradicionalmente, se ha considerado que la existencia de diferencias significativas entre las estimaciones obtenidas mediante el método de Mínimos Cuadrados Ordinarios (OLS) y los estimadores de Variables Instrumentales (IV) es suficiente para cuestionar la validez del instrumento utilizado. Sin embargo, la falta de una metodología formal subyacente ha planteado preocupaciones acerca de la solidez de esta práctica.

En este trabajo nos hemos propuesto abordar dicha carencia y presentar una metodología formal para comparar las estimaciones obtenidas mediante las técnicas de variables instrumentales y Mínimos Cuadrados Ordinarios.

Con el fin de abordar esta limitación y contribuir al avance del campo, nos hemos basado en el trabajo de Ciacci y su artículo titulado "A Matter of Size: Comparing IV and OLS estimates" en el cual se hace referencia al mismo tiempo al artículo de Oster (2019). Oster, se aprovecha la información obtenida de la regresión OLS, como la inclusión de variables de control, el tamaño de las varianzas y los cambios en el coeficiente de determinación (R^2), para estimar un rango de valores en el cual se espera que se encuentre el efecto real del tratamiento. El tamaño de este rango dependerá de la relevancia de las variables observables en relación con las variables no observables, según el criterio del investigador. De esta metodología se sirve Ciacci para calcular un parámetro, el coeficiente de proporcionalidad, con el cual medir la magnitud relativa de la proporcionalidad entre la influencia de las variables observables y no observables, lo cual puede proporcionar evidencia a favor o en contra de la validez del instrumento utilizado.

Considerando que las estimaciones de Variables Instrumentales (IV) capturan el efecto únicamente para aquellos individuos cuya elección de tratamiento es influenciada por el instrumento, se puede argumentar que valores más altos o más bajos del coeficiente de

proporcionalidad son indicios contrarios o favorables, respectivamente, a la validez del instrumento utilizado. En este sentido, el aporte principal de Ciacci (2021) radica en su propuesta de una metodología formal que posibilite a los investigadores realizar comparaciones sistemáticas entre las estimaciones obtenidas mediante IV y el método de Mínimos Cuadrados Ordinarios (OLS). De la ecuación del coeficiente de proporcionalidad:

$$\delta = (\widehat{\beta_2} - \beta_1)\frac{\mathrm{Var}(d_{ih})\mathrm{Var}(X_{ih})}{\gamma \mathrm{Var}(w_{ih})\mathrm{Cov}(d_{ih}, X_{ih})}$$

podemos interpretar como evidencia a favor del instrumento un coeficiente bajo de δ. En términos más simples, cuando el valor de δ es bajo, significa que se necesita una menor influencia de variables no observables para respaldar la hipótesis de que el efecto real coincide con la estimación obtenida mediante la regresión IV. En esencia, el valor de δ nos proporciona información acerca de la proporción relativa de selección basada en variables no observables en comparación con las variables observables necesarias para respaldar la afirmación de que el "efecto real" tiene una magnitud similar a las estimaciones obtenidas mediante el método de Variables Instrumentales (IV).

Para mejorar la aplicabilidad y utilidad de la metodología propuesta por Ciacci y del coeficiente de proporcionalidad, hemos realizado una transformación de la ecuación en una forma matricial, transformando la anterior ecuación en la siguiente:

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

Esta representación matricial presenta varias ventajas importantes. En primer lugar, simplifica la implementación y comprensión de la ecuación en futuros estudios. La forma matricial permite lidiar de manera más eficiente y clara con problemas de alta dimensionalidad, al tiempo que facilita la inclusión de múltiples variables instrumentales y endógenas en el análisis.

Además, la formulación matricial del coeficiente de proporcionalidad abre la puerta a la aplicación de técnicas y herramientas matemáticas avanzadas para su evaluación y análisis. Esto incluye métodos de álgebra lineal, descomposiciones matriciales y técnicas de optimización, que pueden enriquecer y ampliar el rango de enfoques metodológicos

disponibles para evaluar la validez de los instrumentos utilizados en el contexto de la regresión instrumental.

En definitiva, nuestros hallazgos resaltan la necesidad de contar con una metodología formal para evaluar la validez de los instrumentos utilizados en la resolución del problema de la endogeneidad. La ecuación propuesta por Ciacci para el coeficiente de proporcionalidad se presenta como una prometedora alternativa en este sentido, y su formulación matricial aporta ventajas en términos de aplicabilidad y análisis. Asimismo, la representación matricial puede resultar de gran utilidad para futuros estudios, al permitir el aprovechamiento de técnicas matemáticas avanzadas en el campo de la estadística y la econometría, en aras de robustecer y mejorar la evaluación de la validez de los instrumentos.

5. BIBLIOGRAFÍA

Acemoglu, D., Johnson, S., & James, R. (2001). *The Colonial Origins of Comparative Development: An empirical investigation.* The American Economic Review.

Altonji, J., Elder, T., & Taber, C. (August de 2000). *Selection on observed and unobserved variables: assessing the efectiveness of catholic schools.* National Bureau of Economic Research.

Becker, S. (April de 2016). *Using instrumental variables to establish causality.* University of Warwick.

Ciacci, R. (20 de Mayo de 2021). *A Matter of Size: Comparing IV and OLS estimates.*

Davidson, R., & MacKinnon, J. (1993). *Estimation and Inference in Econometrics.* Nueva York: Oxford University Press.

De Luca, G., Magnus, J., & Peracchi, F. (2018). *Comments on 'Unobservable Selection and Coefficient Stability: Theory and Evidence' and 'Poorly Measured Confounders are More Useful on the Left Than on the Right.*

Hernan, M., & Robins, J. (2020). *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC.

Masten, M., & Poirier, A. (07 de January de 2020). *Salvaging Falsified Instrumental Variable Models.* Auburn University.

Mendenhall, W., Beaver, R., & Beaver, B. (2006). *Introduction to Probability and Statistics.* Cengage Learning.

Moretti, E., & Lochner, L. (14 de January de 2014). *Estimating and Testing Models with Many Treatment Levels and Limited Instruments.* University of California-Berkely.

Oster, E. (09 de August de 2016). *Unobservable Selection and Coefficient Stability: Theory and Evidence*. Brown University.

Staiger, D., & Stock, J. (1997). *Instrumental variables regression with weak instruments.* Econometrica .