FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

MODELOS DE INFERENCIA CAUSAL EN ECONOMETRÍA

Autor: Blanca Beamonte Ramiro

5º E-3 Analytics

Tutor: Riccardo Ciacci

Madrid

Junio 2023

Resumen

Este trabajo de investigación se enfoca en el análisis de modelos de inferencia causal en el campo de la econometría, con el objetivo de establecer relaciones causales en fenómenos económicos. El estudio examina diferentes técnicas y metodologías utilizadas en la inferencia causal, como el modelo de regresión lineal, los métodos de variables instrumentales y el diseño de experimentos. A través de una revisión de la literatura, se analiza el estado actual del conocimiento en este campo y se identifican los estudios clave relacionados. El desarrollo del estudio se centra en la aplicación de los modelos de inferencia causal en econometría, utilizando casos de estudio específicos para presentar los resultados obtenidos. Este trabajo contribuye al avance del conocimiento en econometría al proporcionar una comprensión más profunda de las técnicas de inferencia causal y su aplicación en el análisis económico.

Palabras Clave: Endogeneidad, delta, beta, variables instrumentales, OLS, coeficiente de proporcionalidad, Durbin-Wu-Hausman

Abstract

This research work focuses on the analysis of causal inference models in the field of econometrics, aiming to establish causal relationships in economic phenomena. The study examines different techniques and methodologies used in causal inference, such as linear regression models, instrumental variable methods, and experimental design. Through a literature review, the current state of knowledge in this field is analyzed, and key related studies are identified. The study's development centers on the application of causal inference models in econometrics, utilizing specific case studies to present the obtained results. This research contributes to the advancement of knowledge in econometrics by providing a deeper understanding of causal inference techniques and their application in economic analysis.

Key Words: Endogeneity, delta, beta, instrumental variables, OLS, proportionality coefficient, Durbin-Wu-Hausman

ÍNDICE

1. INTRODUCCIÓN

1.1 Contexto del estudio

Con el objetivo de introducir el tema, definamos primero los conceptos fundamentales de este trabajo para poder crear un marco teórico sólido más adelante. Así, entendemos econometría como la aplicación de los métodos estadísticos al estudio de la economía.

La inferencia causal es el proceso de establecer relaciones causales entre variables en un contexto de investigación. Se refiere a la capacidad de determinar si un cambio en una variable causa un cambio en otra variable, y no solo si están correlacionadas o asociadas de alguna manera.

En definitiva, nuestro trabajo gira en torno al estudio de las variables que, a través de la estadística, nos permiten extraer conclusiones en el campo de la economía. Dentro de este tipo de investigaciones podemos encontrar problemas que nos alejen de una fiabilidad de resultados por diversos motivos, como puede ser un sesgo de selección o multicolinealidad de las variables.

La endogeneidad plantea un desafío habitual en este contexto, la cual es causada por la existencia de una correlación entre las variables explicativas y los términos de error en un modelo económico que no se tiene en cuenta. Esta relación puede resultar en estimaciones sesgadas e incorrectas de los parámetros del modelo, lo que complica la identificación de relaciones causales precisas.

Esta puede adoptar diversas formas, una de las cuales es la simultaneidad, en la cual las variables de interés se influyen mutuamente en una relación bidireccional. Por ejemplo, en un modelo que vincula la inversión y el crecimiento económico, la inversión puede tener un impacto en el crecimiento económico, al mismo tiempo que el crecimiento económico puede afectar la inversión. Esta interacción simultánea dificulta identificar en qué dirección se produce la causalidad.

La segunda forma que puede adoptar la endogeneidad puede ser debido a la presencia de autocorrelación, la cual se produce cuando los términos de error están correlacionados a

lo largo del tiempo. Esta correlación puede generar distorsiones en las estimaciones y complicar la correcta interpretación causal de los resultados obtenidos. Un ejemplo de endogeneidad en este contexto podría ser un modelo que examina la relación entre el nivel de desempleo y la tasa de inflación a lo largo del tiempo. Si los términos de error en el modelo están correlacionados temporalmente, es decir, si los períodos de altos niveles de desempleo tienden a estar asociados con períodos de alta inflación y viceversa, las estimaciones de los efectos de una variable sobre la otra podrían verse distorsionadas y dificultar una interpretación causal precisa.

El tercer tipo de endogeneidad puede originarse por la existencia de variables omitidas o no observadas. Estas variables, aunque pueden influir tanto en las variables explicativas como en la variable dependiente, no se incluyen en el modelo debido a limitaciones en su disponibilidad o medición. Esta situación puede resultar en una correlación engañosa entre las variables incluidas y los términos de error, lo que genera un sesgo en las estimaciones. Por ejemplo, un modelo que relacione el consumo de café con el rendimiento académico de los estudiantes. Si no se incluye una variable como la calidad del sueño, que puede afectar tanto el consumo de café como el rendimiento académico, se podría obtener una estimación sesgada de la relación entre estas dos variables (Hendry, 2014).

Es importante abordar estos desafíos de manera adecuada, aplicando técnicas y métodos apropiados para mitigar su impacto en la estimación de los parámetros y obtener resultados más confiables. Aquí es donde nos encontramos con la técnica de las variables instrumentales, las cuales tienen como objetivo identificar variables que se encuentren correlacionadas con las variables endógenas de interés pero que no tengan correlación con los términos de error. Así, estas variables instrumentales permiten controlar la endogeneidad y obtener estimaciones coherentes y libres de sesgo.

Una estrategia común en econometría consiste en considerar las discrepancias significativas entre los coeficientes estimados mediante Mínimos Cuadrados Ordinarios (OLS) y Variables Instrumentales (IV) como indicios de que el instrumento utilizado podría no ser válido (Ciacci, 2021). Estas técnicas son valiosas para obtener estimaciones confiables y precisas de relaciones causales. Sin embargo, hasta el estudio que realizó el Prof. Ciacci, no había una metodología formal para comparar de manera objetiva estas

dos estimaciones. Aunque algunos estudios han intentado hacer comparaciones empíricas previamente, ninguno abordaba el problema desde esa perspectiva.

En este contexto, este estudio se propone abordar esta novedad y extender las aportaciones de Prof. Ciacci para simplificar, haciendo uso de matrices, la metodología formal propuesta. Siguiendo los primeros pasos de Oster (2019) (Ciacci, 2021), se utiliza información derivada de la regresión OLS, como la inclusión de variables de control, el tamaño de las varianzas y el coeficiente de determinación (R2), para establecer un rango de valores en el cual se espera que se encuentre el verdadero efecto del tratamiento. Este rango dependerá de la capacidad informativa de las variables observables en relación con las variables no observables.

1.2 Justificación del tema

La elección de este tema de investigación se basa en la importancia de abordar de manera rigurosa la comparación entre las estimaciones obtenidas mediante las técnicas de variables instrumentales (IV) y Mínimos Cuadrados Ordinarios (OLS). Como comentamos en el apartado anterior, estas técnicas son ampliamente utilizadas en econometría para resolver problemas relacionados con la endogeneidad y el error de medición en análisis económicos. Aunque son de suma importancia, han sido objeto de una escasa atención. Esto sin duda limita la capacidad de evaluar la validez de los instrumentos utilizados y comprender plenamente los efectos del tratamiento en diferentes contextos y escenarios.

El análisis de 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 variables instrumentales y Mínimos Cuadrados Ordinarios, 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 proporcionar una base sólida para futuras investigaciones. Esto permite a los investigadores y analistas económicos contar con herramientas más precisas y confiables para evaluar los efectos del tratamiento y tomar decisiones basadas en evidencia sólida. Se verá que en un apartado del trabajo, expondremos los principales estudios relacionados y observaremos la evolución y relación que tienen aquellos en lo que nos interesa.

La contribución principal de este estudio radica en desarrollar una metodología rigurosa que permita realizar comparaciones objetivas entre las estimaciones obtenidas mediante las técnicas de variables instrumentales y Mínimos Cuadrados Ordinarios de una manera más sencilla. 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.

1.3 Objetivos de la investigación

Los objetivos de esta investigación son diversos, pero siempre enfocados a abordar el desafío de la endogeneidad en la econometría, así como en desarrollar la metodología formal propuesta anteriormente para contrastar los estimadores de Mínimos Cuadrados Ordinarios (MCO) y Variables Instrumentales (VI).

En primer lugar, se introduce un marco teórico más amplio, para tocar los dos métodos principales y diferentes conceptos útiles para una mejor comprensión del trabajo.

Después, se busca exponer al lector una síntesis de estudios relacionados, con el objetivo de introducirle en la problemática y su desarrollo a través de las investigaciones existentes que abordan el tema. Una vez resumidas, ser capaces de destacar sus aspectos esenciales

y su evolución cronológica. Es evidente que los estudios posteriores beberán, aunque a veces menos de lo que sería óptimo, de los anteriores descubrimientos.

Al establecer una metodología para realizar la comparación entre MCO y VI, proponiendo un enfoque sistemático y objetivo. Esto implica emplear información derivada de la regresión MCO, como la inclusión de variables de control, el tamaño de las varianzas y los cambios en el coeficiente de determinación (R2), con el fin de establecer un rango plausible de valores donde se espera encontrar el verdadero efecto del tratamiento. Esta metodología proporciona un enfoque riguroso para evaluar y contrastar los estimadores MCO y VI.

En segundo lugar, otro objetivo fundamental de esta investigación es ofrecer la posibilidad de evaluar la validez del instrumento utilizado en el modelo de VI utilizando la metodología propuesta. La idea es que los valores más altos o más bajos del coeficiente de proporcionalidad, que mide la relación relativa entre la selección basada en variables observables y no observables, pueden ofrecer evidencia en contra o a favor de la validez del instrumento. Esto permite a los investigadores contar con una medida objetiva para evaluar si el instrumento utilizado es adecuado y confiable para estimar el efecto del tratamiento.

En tercer y último lugar, desarrollaremos un test variación del Durbin-Wu-Hausman, el cual hemos elaborado en función de delta para añadir otro punto técnico a nuestro trabajo. De ser útil, permitiría llegar a través de un camino alternativo a conclusiones sobre la endogeneidad de un estudio, además de darnos datos sobre delta y lo que ello implica.

Este trabajo tiene como objetivo proporcionar un enfoque más riguroso y estructurado para evaluar la validez del instrumento y realizar inferencias causales más sólidas. Esto contribuye a la calidad y confiabilidad de los resultados obtenidos en estudios econométricos. Además, al utilizar medidas objetivas, como el coeficiente de proporcionalidad, se reduce el sesgo subjetivo en la evaluación de la validez del instrumento, lo cual ayuda a los investigadores a tomar decisiones más fundamentadas y evitar interpretaciones erróneas basadas únicamente en diferencias entre los estimadores MCO y VI.

La metodología propuesta puede ser aplicada en una variedad de contextos y problemas de investigación donde la endogeneidad y la comparación de estimadores sean relevantes, ampliando así 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

Con el fin de lograr los objetivos planteados, se seguirá una metodología basada en una revisión crítica y un análisis detallado del artículo del Prof.Ciacci, permitiendo un avance que simplifique la ecuación a tener en cuenta para facilitar su ecuación.

Se realizará una lectura minuciosa del estudio, identificando los elementos clave relacionados con la conversión de la ecuación beta en función de delta y adaptación de un test estadístico. Se estudiarán las propiedades de las matrices para poder llevar a cabo este trabajo. Se pondrá en contexto cada tecnicismo expuesto, se explicarán las metodologías de manera sencilla y comprensiva al igual que los diferentes estudios en los que nos hemos apoyado para basar el tema.

De esta manera, el objetivo principal es desde una versión todo lo sencilla posible, hacer entender al lector la importancia de estas metodologías y de su facilidad de aplicación, después de darles un contexto sobre el tema.

Creemos que siendo un vocabulario sencillo donde lo técnico es explicado, se puede entender por un gran rango de lectores, animándolos así a aplicar importantes cuestiones de validación que ni expertos en determinados casos tienen en cuenta a la hora de presentar sus conclusiones.

En resumen, el objetivo principal de esta investigación es examinar la metodología propuesta por Ciacci en su artículo "*A Matter of Size: Comparing IV and OLS Estimates*", centrándose en la transformación de la ecuación beta en función de delta en un contexto multivariable, acompañado de una propuesta estadística.

2. REVISIÓN DE LA LITERATURA

2.1 Marco teórico

El enfoque de Mínimos Cuadrados Ordinarios (OLS) y el método de Variables Instrumentales (IV) son dos métodos utilizados en econometría para lidiar con el problema de endogeneidad, visto anteriormente, en los modelos de regresión.

Así, vamos a explicar sumariamente en qué consisten:

2.1.1 OLS

El Análisis de Regresión es ampliamente utilizado en las ciencias sociales como una herramienta estadística fundamental. Su objetivo principal es evaluar las relaciones entre dos o más atributos de entidades en un contexto dado. Mediante la identificación y medición de estas relaciones, se busca obtener un mejor entendimiento de lo que ocurre en un lugar, predecir eventos futuros en ese lugar o examinar las causas que contribuyen a que algo suceda en un lugar específico.

La técnica más reconocida en el campo de la regresión es conocida como Mínimos Cuadrados Ordinarios (OLS). Esta técnica, que sirve como punto de partida en todos los análisis de regresión espacial, proporciona un modelo global de la variable o proceso que se desea comprender o predecir. A través de la construcción de una ecuación de regresión simple, se busca representar y analizar dicho proceso de manera cuantitativa.

De esta manera, nos encontramos con un método que estima los coeficientes de las ecuaciones que describen la relación entre una o varias variables, a las que llamamos independientes (Xs) y una variable dependiente (Y), ya que como su propio nombre indica, depende de ellas.

La regresión de los mínimos cuadrados se denomina así porque se basa en minimizar el error cuadrático. Además de esta técnica, también podemos utilizar el método del máximo de verosimilitud y el estimador del método generalizado de momentos.

Dentro de este modelo de regresión, nos encontramos dos tipos: la regresión lineal simple y la múltiple. El primero se compone de dos variables estadísticas, denominadas X e Y. En este modelo, se considera una única variable independiente o explicativa, X, y una variable dependiente o respuesta, Y, asumiendo que la relación entre ambas es lineal. En la regresión lineal simple, se parte del supuesto de que X e Y están relacionadas mediante una función lineal, donde β1 y β0 son estimadores que representan los coeficientes de la relación.
Su representación es:

$$Y = \beta_0 + \beta_1 * X$$

Respecto al múltiple, que nos interesa más en tanto en cuanto habla de formas matriciales, es la que permite analizar la relación entre dos o más variables a través de ecuaciones. En la investigación estadística, es frecuente encontrar variables que están interrelacionadas de alguna manera, lo que implica que una de las variables pueda ser expresada matemáticamente en función de otras variables.

En el caso del modelo de regresión lineal múltiple, se considera la presencia de múltiples variables predictoras, lo que conlleva la existencia de varios parámetros. Para este tipo de regresión, se asume que la variable de respuesta, denotada como Y, está relacionada con las variables predictoras X0, X1, ..., Xm mediante una relación funcional.

Matricialmente, quedaría representada:

$$Y = X\beta$$

2.1.2 IV

El método de Variables Instrumentales se emplea para obtener estimaciones de relaciones causales cuando no es factible realizar experimentos controlados o cuando no se puede administrar un tratamiento de manera exitosa a todas las unidades en un experimento aleatorio.

En términos intuitivos, las variables instrumentales se utilizan cuando una variable explicativa de interés está correlacionada con el término de error, lo que puede llevar a sesgos en los resultados obtenidos mediante el uso de mínimos cuadrados ordinarios y

ANOVA. Un instrumento válido provoca cambios en la variable explicativa sin tener un efecto independiente en la variable dependiente, lo que permite al investigador descubrir el efecto causal de la variable explicativa en la variable dependiente.

Este método permite una estimación consistente cuando las covariables se correlacionan con los términos de error en una regresión. En este escenario, la regresión lineal genera estimaciones que presentan sesgo e inconsistencia. Sin embargo, si se dispone de un instrumento, es posible obtener estimaciones consistentes. Un instrumento se refiere a una variable que no forma parte directamente de la ecuación explicativa, pero que se correlaciona con las variables explicativas endógenas, considerando las demás covariables presentes.

Para poder usar esta metodología, necesitamos que se cumplan dos requisitos:

Primero, que exista correlación entre el instrumento y las variables explicativas endógenas. Si esta relación es fuerte, entonces el instrumento tiene una primera etapa fuerte. Si es débil, término que aparecerá a lo largo del trabajo, puede conducir a inferencias incorrectas sobre la estimación de los parámetros y los errores estándar, lo que supone un problema de confianza en el análisis.

El segundo requisito impide que el instrumento esté correlacionado con el término error en la ecuación explicativa. Es decir, no puede este último sufrir el mismo problema que la variable original. Cuando este requisito se cumple, se dice que el instrumento satisface la restricción de exclusión.

En resumen, si no podemos asignar variables aleatorias siempre a los resultados, usamos los datos observados en seis pasos:

1. El instrumento tiene que estar relacionado con la variable dependiente
2. Asumir que el instrumento no tiene un efecto causal en la variable dependiente
3. Asumir que el instrumento sí que lo tiene en el tratamiento
4. Asumir que el instrumento es asignado de manera aleatoria

5. Por el punto anterior, el efecto causal del instrumento en el tratamiento es la correlación en los datos
6. Como el instrumento es aleatorio, no hay más relación que con el tratamiento

De esta manera, si el instrumento de manera clara tiene una relación con la variable dependiente, es porque el tratamiento ha tenido efecto en ella.

2.1.3 Relación entre ambas

El método OLS asume que las variables explicativas no están correlacionadas con el término de error del modelo, lo que se conoce como exogeneidad. Sin embargo, cuando hay endogeneidad, es decir, cuando las variables explicativas están correlacionadas con el término de error, el método OLS produce estimaciones sesgadas e inconsistentes.

Por otro lado, el método de Variables Instrumentales también se utiliza como técnica cuando se sospecha que existe endogeneidad en el modelo. Este método busca variables instrumentales que cumplan dos condiciones importantes: estar correlacionadas con las variables endógenas, pero no estar correlacionadas con el término de error. Estas variables instrumentales se utilizan para estimar las relaciones causales entre las variables explicativas y la variable dependiente de manera consistente.

La idea central del método IV es utilizar las correlaciones entre las variables instrumentales y las variables endógenas para descorrelacionar las variables explicativas del término de error, obteniendo así estimaciones no sesgadas y consistentes. Para lograr esto, se emplean diferentes técnicas de estimación como el Método de Dos Etapas (2SLS) o el Estimador de Mínimos Cuadrados en Dos Etapas (2SLS).

En resumen, mientras que el método OLS se utiliza cuando se cumple la suposición de exogeneidad, el método de Variables Instrumentales se emplea cuando hay endogeneidad y se busca obtener estimaciones no sesgadas y consistentes al descorrelacionar las variables explicativas del término de error mediante el uso de variables instrumentales (Wooldridge, 2009)

2.2 Estado actual del conocimiento

En este apartado, únicamente se quiere explicar que, vistos los dos métodos que existen para lidiar con la endogeneidad de manera teórica en detalle, ahora nos adentramos en un número concreto de lecturas.

El motivo es que, hasta ahora, ha sido simplemente una exposición de conocimientos teóricos puros, en la cual hemos entendido los diferentes métodos existentes y posiblemente parte de su uso.
A continuación, a través de aportaciones y ejemplos de diferentes estudiosos, comprobamos que los mismos van evolucionando y añadiendo nuevas ideas para un mejor desarrollo de los estudios econométricos.

Es de gran importancia tener en cuenta estos aspectos. En un mundo que cada vez se desarrolla más rápido donde la curva de aportaciones y aprendizajes es exponencial, son muchos los estudios que pretenden avanzar el estado del conocimiento a través de sus aportaciones. Pero como decíamos, al igual que lo anterior son conocimientos consolidados, hay muchos autores que se quedan desconocidos o inutilizados, resultando más torpe para los nuevos estudiosos el avanzar.

Cuanta más literatura científica tengamos, más fácil será utilizar las anteriores, esfuerzos ya conseguidos y demostrados o no, pero diferentes perspectivas, para apoyarse en ellas y realizar nuestras conclusiones.

De esta manera, queremos mostrar los diferentes estudios en los que nos hemos apoyado, para que el lector pueda tanto apoyarse en ellos hasta llegar a nuestras conclusiones o simplemente observar una evolución o si hay aspectos relevantes no tenidos en cuenta por nosotros. Además, escribiremos una pequeña relación entre los textos.

Así, hasta llegar las aportaciones a una novedosa que constituye el núcleo de nuestro trabajo, el cual tiene la intención de ser una nueva aportación a la ciencia debido a las grandes ventajas que supone, como las que explicaremos más adelante.

2.3 Principales estudios relacionados

Para una mayor comprensión de la evolución del tema a tratar, expondremos diferentes artículos en orden cronológico. Estos son las principales fuentes en las que nos hemos apoyado para crear un marco teórico sobre el que construir nuestro trabajo. De esta manera, observamos cómo los estudios más recientes se han ido apoyando en las conclusiones que investigadores anteriores han ido aportando.

Resumiremos los puntos clave que estos estudios abordan, lo que ayudará a los lectores a comprender mejor la contribución específica de nuestro trabajo y a identificar las deficiencias que hemos identificado tanto en la teoría como en la aplicación de los estudios existentes.

Instrumental variables regression with weak instruments

El artículo "*Instrumental variables regression with weak instruments*" de Douglas Staiger y James H. Stock, publicado en Econometrica en 1997, se enfoca en abordar el desafío de estimar modelos de regresión que involucran variables instrumentales débiles.

Los autores definen las variables instrumentales débiles como aquellas que presentan una correlación baja con las variables endógenas en la regresión, lo cual limita su capacidad para mitigar el sesgo de endogeneidad. Este problema se vuelve especialmente problemático en casos donde la muestra de datos es pequeña, lo que reduce aún más la efectividad de las variables instrumentales para controlar la endogeneidad.

En su estudio, proponen un método de estimación para la regresión de variables instrumentales que emplea el enfoque de máxima verosimilitud y una corrección en la matriz de varianzas y covarianzas para tener en cuenta la debilidad a la que nos referimos. Además, proporcionan un estadístico de prueba que permite evaluar la validez de las variables instrumentales utilizadas.

Los autores aplican su método a un ejemplo de regresión salarial y descubren que el enfoque de máxima verosimilitud con la corrección en la matriz de varianzas y

covarianzas es más efectivo para abordar el problema de las variables instrumentales débiles en comparación con otros métodos de estimación.

En resumen, el trabajo de Staiger y Stock presenta una solución al desafío de las variables instrumentales débiles en modelos de regresión con variables instrumentales. Además, resalta la importancia de evaluar la validez de las variables instrumentales al estimar modelos de regresión.

Para lo que nos interesa, resaltamos que ya desde este estudio se puede observar cómo los estudiosos, además de utilizar el método de variables instrumentales, dan importancia a evaluar la validez de estas a través de la matriz de varianas-covarianzas.

Selection on observed and unobserved variables: assessing the effectiveness of catholic schools

El trabajo "*Selection on observed and unobserved variables: assessing the effectiveness of catholic schools*" de Josep G. Altonji, Todd E. Elder y Christopher R. Taber, publicado en la revista Journal of Political Economy en 2005, aborda el problema de la selección de variables observadas y no observadas en la evaluación del impacto de la educación en escuelas católicas en la vida de los estudiantes.

Los autores utilizan datos de la Encuesta Nacional de Educación de 1972 y una estrategia de variables instrumentales para controlar la posible selección de estudiantes en escuelas católicas en función de variables no observadas, como habilidades no medidas o antecedentes familiares. El enfoque de variables instrumentales se basa en la utilización de una variable instrumental, en este caso la densidad de católicos en la zona de residencia del estudiante, para estimar el efecto causal de la educación católica en la probabilidad de graduación de la escuela secundaria y la participación universitaria.

De esta manera encuentran que los estudiantes que asisten a escuelas católicas tienen una probabilidad significativamente mayor de graduarse de la escuela secundaria y de asistir a la universidad en comparación con los estudiantes que asisten a escuelas públicas. Estos resultados se mantienen robustos incluso después de controlar la posible selección de variables no observadas.

En resumen, el trabajo a través de un ejemplo cómo abordar el problema de la selección de variables observadas y no observadas en la evaluación del impacto de programas educativos utilizando variables instrumentales y destaca la importancia de controlar adecuadamente los factores que influyen en la selección de estudiantes en diferentes tipos de escuelas.

De este caso, a través de otro camino diferente pero también haciendo uso de las variables instrumentales, resaltamos lo mismo que en el anterior. Es decir, el uso de métodos para evaluar problemas de selección de variables.

The Colonial Origins of Comparative Development: An Empirical Investigation

El artículo "*The Colonial Origins of Comparative Development: An Empirical Investigation*" de Daron Acemoglu, Simon Johnson y James A. Robinson, publicado en diciembre de 2001 en The American Economic Review, se ocupa de investigar las razones por las cuales algunos países son más ricos y prósperos que otros.

Los autores sostienen que la causa subyacente radica en la forma en que los países fueron colonizados. Específicamente, argumentan que aquellos países colonizados por potencias europeas que establecieron instituciones políticas y económicas inclusivas, donde las élites locales podían participar y compartir el poder, han experimentado un mejor desempeño económico en comparación con aquellos colonizados por potencias que establecieron instituciones extractivas, en las cuales el poder y los recursos eran monopolizados por la élite colonial.

Esta conclusión es soportada por evidencia empírica, utilizando datos de diversos países alrededor del mundo. Descubren que los países colonizados por potencias europeas con instituciones inclusivas muestran un mejor desempeño económico en indicadores como el ingreso per cápita, la tasa de alfabetización y la esperanza de vida, en comparación con aquellos colonizados por potencias con instituciones extractivas.

Además, hacen énfasis en que el colonialismo no es el único factor que afecta a las instituciones, si no que los autores también resaltan que las diferencias persisten hasta

hoy debido a la inercia institucional, resultando las instituciones extractivas difíciles a las que se agarran difíciles de cambiar.

En resumen, en este caso mostramos a través de un ejemplo, como los autores desarrollan un estudio que a través del método OLS explicado anteriormente extraen conclusiones y encaran el problema de la endogeneidad de las instituciones utilizando el método de las variables instrumentales.

Estimating and Testing Models with Many Treatment Levels and Limited Instruments

El artículo titulado "*Estimating and Testing Models with Numerous Treatment Levels and Limited Instruments*" de Lance Lochner y Enrico Moretti, publicado en enero de 2014, aborda el desafío de estimar y evaluar modelos económicos que involucran múltiples niveles de tratamiento y un número restringido de instrumentos.

Los autores se centran en el análisis empírico de los efectos de la educación en los ingresos laborales, utilizando datos provenientes de la Encuesta de Ingreso y Participación Programática de Canadá. El modelo considera distintos niveles educativos y se emplea un conjunto limitado de instrumentos para controlar la posible endogeneidad entre la educación y los ingresos.

Para hacer frente a esta problemática, los autores proponen una metodología de "regresión ampliada" que permite la inclusión de múltiples niveles de tratamiento y un número reducido de instrumentos. Esta aproximación también permite incorporar otras variables relevantes en el modelo.

Los resultados del estudio indican que la inclusión de múltiples niveles educativos y un número limitado de instrumentos puede mejorar la precisión de las estimaciones y brindar información adicional sobre los efectos de la educación en los ingresos laborales. Además, los autores llevan a cabo pruebas de robustez para evaluar la validez de los resultados y encuentran consistencia en diferentes especificaciones del modelo.

En resumen, este trabajo que se centra en la inclusión de tratamientos para poder tenerlos en cuenta en la regresión de forma más “directa” resalta la importancia de utilizar técnicas

estadísticas adecuadas para abordar los desafíos de la endogeneidad y la inclusión de múltiples niveles de tratamiento en modelos económicos complejos.

Unobservable Selection and Coefficient Stability: Theory and Evidence

El artículo titulado "*Unobservable Selection and Coefficient Stability: Theory and Evidence* " escrito por Emily Oster, publicado en agosto de 2016, se centra en el desafío de abordar la selección no observada en modelos económicos y cómo esto puede influir en la persistencia de los coeficientes estimados a lo largo del tiempo.

La misma discute que una aproximación común para evaluar la resistencia al sesgo por omisión de variables consiste en observar cómo se modifican los coeficientes después de incluir controles adicionales. No obstante, esta información es relevante únicamente si la selección de variables observables proporciona una información válida sobre la selección de variables no observables. Aunque esta relación se reconoce en la teoría por otros anteriores (por ejemplo, el anteriormente comentado de Altonji, Elder y Taber), son muy pocos los estudios empíricos que abordan formalmente esta cuestión.

La autora comienza explicando que la selección no observada puede surgir cuando una variable relevante que afecta el resultado de interés no está disponible o no se mide de manera adecuada, tal y como explicamos en la introducción. Este problema puede ser particularmente significativo en estudios longitudinales donde la selección en el tiempo puede tener impacto en los resultados.

Por ello, Emily propone un marco teórico para analizar cómo la selección no observada afecta la persistencia de los coeficientes estimados en el tiempo. Este enfoque se basa en el concepto de "límites de elección" que representan la distribución de las variables no observadas que influyen en la selección a lo largo del tiempo.

La autora aplica este marco teórico a través de un estudio empírico que investiga la relación entre la exposición al plomo durante la infancia y los logros educativos posteriores en los Estados Unidos. Utiliza datos longitudinales de una cohorte de niños nacidos en las décadas de 1970 y 1980, y encuentra que la selección no observada puede ser un problema relevante en este contexto.

Al igual que los autores anteriores, realiza pruebas de robustez y descubre que los resultados son consistentes en diferentes especificaciones del modelo y en distintas cohortes de niños. En general, este estudio subraya la importancia de considerar detenidamente la selección no observada en modelos económicos y cómo esto puede afectar la persistencia de los coeficientes estimados a lo largo del tiempo.

Así, para lo que nos interesa, Emily desarrolla una extensión de la teoría que establece una conexión explícita entre el sesgo y la estabilidad de los coeficientes. Se muestra que es necesario considerar los cambios en los coeficientes y en el R-cuadrado. Además, se presenta un argumento formal de acotación. Se llevan a cabo dos ejercicios de validación y se discute la aplicación en la literatura económica.

Este estudio desempeña un papel fundamental en nuestra investigación, ya que sienta las bases y sirve de inspiración para la contribución realizada por el Prof. Ciacci. Por tanto, es esencial seguir una secuencia cronológica adecuada para comprender la evolución de las aportaciones presentadas.

Using instrumental variables to establish causality

Sascha O. Becker escribió un artículo titulado "*Using instrumental variables to establish causality* " en la revista Journal of Economic Surveys en 2005.

El autor explica que se utilizan modelos econométricos basados en datos observados porque en muchas situaciones de la vida real no es posible realizar experimentos controlados para determinar la relación causal entre dos variables. Sin embargo, los problemas de endogeneidad y sesgo de selección pueden afectar estos modelos, lo que puede resultar en estimaciones erróneas de la relación causal.

Así, propone emplear variables instrumentales, que son variables independientes de las variables observadas pero que guardan una relación causal con las variables endógenas. Al utilizar las variables instrumentales correctas, se logra obtener estimaciones más precisas de la relación causal entre las variables observadas. Como sabemos, este es uno de los principales objetivos que persigue el método IV.

El autor ilustra el uso de variables instrumentales con varios ejemplos de la literatura económica, como la relación entre educación e ingresos, así como la relación entre inversión en infraestructura y crecimiento económico. También aborda los desafíos que pueden surgir al emplear variables instrumentales, como la elección inadecuada de instrumentos o la falta de validez de los supuestos del modelo.

En resumen, el trabajo enfatiza la importancia de utilizar variables instrumentales en la investigación económica para establecer la causalidad entre las variables observadas, y ofrece una guía valiosa para la selección y validación de instrumentos adecuados.

Comments on 'Unobservable Selection and Coefficient Stability: Theory and Evidence' and 'Poorly Measured Confounders are More Useful on the Left Than on the Right

El estudio *Comments on 'Unobservable Selection and Coefficient Stability: Theory and Evidence' and 'Poorly Measured Confounders are More Useful on the Left Than on the Right* realizado por Giuseppe De Luca, Jan R. Magnus y Franco Peracchi en 2018, examina dos investigaciones previas relacionadas con el problema de la selección no observada y los confusores mal medidos en modelos econométricos.

En primer lugar, los autores analizan el ya mencionado trabajo de Emily Oster, donde se investiga el impacto de la selección no observada en la estimación de modelos de regresión. De Luca, Magnus y Peracchi discuten el marco teórico utilizado por Oster y presentan una crítica a su enfoque empírico, argumentando que la relevancia de la selección no observada puede variar en diferentes situaciones, y que se requiere un análisis minucioso de las variables y su relevancia para el modelo en cuestión.

En segundo lugar, los autores comentan el trabajo de James J. Heckman y Vytlacil titulado "*Poorly Measured Confounders are More Useful on the Left Than on the Right*", donde se examina el impacto de los confusores mal medidos en la estimación de modelos de regresión. De Luca, Magnus y Peracchi discuten la validez de los supuestos utilizados por Heckman y Vytlacil en su trabajo, argumentando que los confusores mal medidos pueden tener un impacto más significativo en la estimación de los efectos marginales de las variables independientes que en las estimaciones de los coeficientes de las variables.

En resumen, como hacemos referencia en la evolución de aportaciones apoyadas en las anteriores, el estudio de De Luca, Magnus y Peracchi ofrece un análisis crítico detallado de los enfoques utilizados en los trabajos anteriores sobre la selección no observada y los confusores mal medidos, enfatizando la importancia de realizar un análisis exhaustivo de las variables y su relevancia para el modelo en cuestión.

Salvaging Falsified Instrumental Variable Models

El artículo "*Salvaging Falsified Instrumental Variable Models*" escrito por Matthew A. Masten y Alexandre Poirier, y publicado en enero de 2020, se ocupa de abordar el desafío que plantean los modelos de variables instrumentales falsificados, proponiendo un enfoque para extraer la información válida de estos modelos.

Como sabemos, los modelos de variables instrumentales son ampliamente utilizados en el campo de la econometría para establecer relaciones causales entre variables observables. Sin embargo, estos modelos a menudo enfrentan problemas de endogeneidad y selección de variables. En ciertos casos, los modelos pueden estar falsificados, lo que implica que las variables instrumentales utilizadas no son válidas y no proporcionan una identificación causal precisa.

En su investigación, los autores desarrollan un enfoque novedoso que permite recuperar la información válida de los modelos falsificados mediante un modelo de ajuste de mínimos cuadrados. En lugar de intentar identificar la causa subyacente del problema de endogeneidad, el enfoque propuesto se enfoca en maximizar la varianza explicada por la variable endógena utilizando una combinación de variables instrumentales falsificadas y variables exógenas.

Los autores demuestran la efectividad de este enfoque para recuperar información válida de los modelos falsificados, presentando pruebas empíricas basadas en datos de la Encuesta Nacional de Examen de Salud y Nutrición (NHANES), donde evalúan la relación causal entre la obesidad y la presión arterial.

En resumen, este trabajo ofrece una solución innovadora y prometedora para recuperar información válida de modelos de variables instrumentales falsificados, lo que resulta especialmente útil en situaciones donde no es posible identificar la causa subyacente del problema de endogeneidad.

A Matter of Size: Comparing IV and OLS estimates

El artículo " *A Matter of Size: Comparing IV and OLS estimates*" escrito por Riccardo Ciacci y publicado en mayo de 2021, tiene como objetivo principal contrastar las estimaciones obtenidas mediante dos técnicas estadísticas ampliamente utilizadas: la regresión lineal ordinaria (OLS) y la regresión instrumental (IV).

El autor comienza destacando las diferencias entre estas dos técnicas y su aplicabilidad en diferentes escenarios. La regresión OLS es ampliamente empleada en estudios empíricos para analizar la relación entre una variable dependiente y múltiples variables independientes. Por otro lado, la regresión IV se utiliza cuando se presenta endogeneidad en una o más variables independientes, es decir, cuando estas variables están correlacionadas tanto con la variable dependiente como con algún factor no observado.

El estudio se enfoca en comparar las estimaciones obtenidas mediante estas dos técnicas utilizando un conjunto de datos simulados que varían en tamaño de muestra y niveles de endogeneidad. Los resultados obtenidos revelan que, en la mayoría de los casos, la regresión IV proporciona estimaciones más precisas que la regresión OLS, especialmente cuando existe una alta endogeneidad y la muestra es pequeña. No obstante, el autor también señala que la regresión IV puede verse afectada por errores de medición en las variables instrumentales, lo cual constituye una limitación a considerar.

En resumen, este trabajo resalta la importancia de seleccionar la técnica estadística apropiada según el escenario y los datos disponibles. Además, sugiere que la regresión IV puede ser una opción más adecuada en situaciones de alta endogeneidad y muestras pequeñas. Estas valiosas contribuciones servirán como base central para el desarrollo de nuestro propio estudio, ampliando su contenido para abordar casos multivariantes y explicar sus implicaciones, así como aplicarlos en otros apartados.

2.4 Relación de aportaciones

Después de ver en orden cronológico las diferentes aportaciones, que generalmente también iban acompañados de muestras empíricas, podemos hacernos a la idea de cómo ha sido la evolución en el campo a grandes rasgos hasta llegar a hoy.

Podemos ir viendo cómo al principio, ya se le buscaba una solución al problema de las variables instruméntales débiles en sus modelos de regresión, y la importancia que desde aquí se le da a evaluar la validez de un estudio.

En el caso del segundo, el problema de selección de variables mismo tiene un enfoque diferente, ya que se centra en los factores que se eligen para lo mismo. Más adelante, el estudio es a través del método OLS, pero también se encara a la endogeneidad, como sabemos que es necesario, a través del método de IV.

Siguiendo por este camino, Lochner y Moretti intentan demostrar la importancia de la selección de más tratamientos y menos variables. Así, se aborda la estimación de modelos con múltiples niveles de tratamiento y un número limitado de instrumentos. Se propone una extensión del enfoque de regresión lineal ordinaria (OLS) para abordar este problema, ya que las suposiciones del OLS pueden no cumplirse en casos de múltiples niveles de tratamiento, lo que puede llevar a resultados sesgados o inconsistentes.

Los autores introducen el Estimador de Mínimos Cuadrados con Instrumentos Ampliados (ACLS), que utiliza instrumentos adicionales para controlar los efectos no observados y eliminar el sesgo de selección. Este estimador permite una estimación más precisa y consistente de los efectos causales de los diferentes niveles de tratamiento.

Además, se pueden encontrar implicaciones para la regresión instrumental (IV) en el trabajo de Lochner y Moretti. La regresión instrumental se utiliza para abordar la endogeneidad, que ocurre cuando una variable independiente está correlacionada con el error del modelo. En el contexto de modelos con múltiples niveles de tratamiento, puede ser desafiante encontrar instrumentos válidos que cumplan con los supuestos necesarios para la regresión instrumental.

Para superar estas limitaciones, los autores proponen un enfoque que combina múltiples instrumentos, lo que mejora la precisión y consistencia de las estimaciones en modelos con muchos niveles de tratamiento y un número limitado de instrumentos válidos.

Sascha a su vez utiliza en método de IV, dando un esquema relevante sobre la selección y validación de instrumentos. El siguiente, analiza dos estudios, entre otros el de Emily, en el que que entraremos en detalle más tarde, para seguir avanzando con el estado de la ciencia, dando mucha importancia al análisis que hay que hacer de las variables que se introducen en este tipo de estudios.

Después, Masten y Porier desarrollan un enfoque para sacar información válida de las variables falsificadas a partir de la maximización de su varianza.

Dicho esto, en esta sección, es relevante establecer la conexión entre los dos ensayos principales que han sido objeto de análisis y estudio en la elaboración de este trabajo*: A Matter of Size: Comparing IV and OLS estimates* (R.Ciacci) y *Unobservable Selection and Coefficient Stability: Theory and Evidence* (E.Oster).

La metodología propuesta por Oster (2019) utiliza información derivada de la regresión OLS, como la inclusión de variables de control, el tamaño de las varianzas y el coeficiente de determinación (R2), para establecer un rango de valores en el cual se espera que se encuentre el verdadero efecto del tratamiento. 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.

El ensayo del Prof. Ciacci se enfoca esencialmente en el concepto de endogeneidad, mientras que el de la Prfª. Oster se centra específicamente en la selección no observada. Un fenómeno se encuadra dentro del otro, donde su consecuencia resulta en la posibilidad de afectar la estimación de los parámetros y conducir a resultados sesgados o inconsistentes en modelos econométricos.

De esta manera, pueden generar resultados sesgados e inconsistentes en los modelos econométricos. La endogeneidad puede ser causada por la presencia de variables

omitidas, lo que conlleva a una correlación entre la variable explicativa y el error del modelo, generando sesgos en la estimación de los parámetros.

La selección no observada entonces puede surgir cuando hay variables relevantes que no se incluyen en el modelo, lo que puede afectar tanto a la variable dependiente como a las variables explicativas. Al no tener en cuenta estas variables no observadas, la estimación de los parámetros puede estar sesgada.

Aunque como hemos visto, el estudio de la economista Oster fue criticado con posterioridad, sienta unas bases y conclusiones esenciales en el campo.

En resumen, estos ensayos proporcionan un análisis de conceptos y métodos relacionados con la estimación de parámetros en modelos econométricos. Abordan la endogeneidad y la estimación en modelos con múltiples niveles de tratamiento, proponiendo enfoques para abordar estos desafíos y mejorar la precisión de las estimaciones en la investigación económica.

2.5 Identificación de lagunas en la literatura

Una vez vistos los principales estudios, observamos que los autores, hasta que llegó Riccardo, hacen uso de una metodología u de otra (u OLS u IV) de las posibles para estos estudios econométricos.

Pero, ¿nadie se había planteado cómo se podían usar ambas y decidir cuál es la más adecuada para el caso concreto? Cada estudioso, en función de las características de su estudio, elige una metodología para llevarlo a cabo.

Es muy posible, que al planteárselo pudiesen surgir dudas sobre cuál utilizar, sin embargo, no existía una metodología formal que permitiese compararlas hasta la propuesta del Prof. Ciacci. Para ello, nuestro autor introduce una novedosa metodología en la cual, a partir de la delta de Oster, expone una manera de comparar los diferentes rangos y adecuación de ambas metodologías.

Al leerla, pensamos que es una aportación más que útil, por lo que nuestro objetivo en este trabajo, además de introducir la información previa existente y cómo está relacionada, es poder introducir una simplificación de la ecuación dada, ofreciendo una mayor facilidad para los estudiosos en su aplicación.

Además, no queremos olvidar el mencionar que incluso en el tiempo de avances tecnológicos que corren, una computadora funciona mucho mejor con matrices más sencillas que con operaciones de factores de gran intensidad.

Sin embargo, hacer notar que, en los diferentes estudios analizados, como en estudios de los que se habla en sus páginas, existe todavía un peligro que provoca grandes sesgos y errores en los resultados, al no hacer siempre los autores uso de diferentes métricas, para interpretar y validar sus conclusiones. Así, se causan muchos sesgos que son ajenos a nuestro trabajo, pero de importancia capital.

3. DE ECUACIÓN A MATRIZ

3.1 Análisis crítico del artículo

Con ánimo de ir abordando el tema de forma comprensiva, comencemos con los conceptos básicos para mostrar la regresión sobre la que nos apoyaremos:

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

En este caso, cada variable tiene sus subíndices unidad (i) y tiempo (h). Además, d es el escalar que hace de tratamiento, w el vector de los controles no observables (por lo que no podrá formar parte de la ecuación al ser calculada) y X el vector que contiene los controles observados.

En el caso univariante, la regresión es aquella en la que queremos averiguar la recta y, fruto de la solución que contiene alfa como punto de corte con el eje y beta sub-uno como la pendiente de la recta, además de la variable independiente X.
Así, obtenemos tanto una solución como una recta sencilla, que gráficamente podemos ver. A partir de esta regresión, surgen muchas aplicaciones, tal y como hemos observado en estudiosos en los apartados anteriores.

Para lo que nos interesa en este caso y siguiendo las asunciones expuestas por Oster, llegamos a una relación de selección proporcional dada para el caso:

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

La cual se cumplirá siempre que delta 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 $\tilde{d}_{ih}$ (cuyo significado explicaremos en el siguiente apartado) es igual a $\mathrm{d} - \tau + \tau \mathrm{X}$, el sesgo de variable omitida que resulta es:

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

Si cogemos esta e introducimos las conclusiones de la relación de selección proporcional expuestas por Oster anteriormente, llegamos, en nuestro caso de regresión lineal múltiple, a nuestro coeficiente de proporcionalidad:

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

De esta manera, obtenemos delta. Este coeficiente de proporcionalidad es el núcleo de nuestro estudio. ¿Qué significa su resultado?

El coeficiente de proporcionalidad (δ) es una medida que muestra la necesidad de selección de variables no observables para respaldar las estimaciones de la regresión de variables instrumentales (IV) en un estudio empírico. Un valor alto de δ indica una mayor dependencia de factores no observables para justificar que el efecto real coincide con el efecto estimado mediante la regresión de IV. De la misma manera ocurre, al contrario, un valor bajo de δ sugiere que la variable instrumental es más válida y proporciona una evidencia más sólida para el efecto estimado.

La magnitud de δ indica la proporción de selección de variables no observables en relación con las observables necesaria para que el "efecto real" sea similar en tamaño a las estimaciones de IV. En este análisis, además, es de suma importancia considerar y evaluar diversos factores, como la inclusión de controles, el tamaño de las varianzas y el movimiento de R^2. Esto se debe a que valores altos de δ podrían en ciertos casos indicar simplemente la invalidez del instrumento o la existencia de efectos heterogéneos, donde las estimaciones de IV capturan los efectos para una subpoblación específica.

En entornos empíricos, es fundamental establecer el signo de δ. Esto nos permitirá calcular los conjuntos identificados. En una formulación simplificada, el signo de δ está condicionado en parte por la asunción que hagamos del signo de γ.

El motivo que subyace es que las varianzas son siempre positivas, y tanto la resta de betas como la covarianza pueden estimarse a partir de los datos disponibles en el estudio en cuestión. Una vez que se conoce el signo de δ, se pueden 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. Esto lo pondremos en práctica en el siguiente apartado.

Por último, hacer especial mención al valor que escogemos para R^2. Este indica la cantidad proporcional de variación en la variable de respuesta y, explicada según las variables independientes X en el modelo de regresión. Por ello, además de lo mencionado anteriormente, es crucial para estimar los conjuntos identificados.

Este valor escogido, que por Oster (2019) es llamado R_{max}, se elige en función del conocimiento previo que se tiene del entorno del estudio. Si el sujeto que lo realiza estima objetivamente que la regresión puede explicar la variable de resultado de manera total, entonces R_{max} se establece con el valor 1.

3.2 Identificación de puntos de expansión

Al estudiar los diferentes artículos, vemos como poco a poco se va introduciendo una metodología que, al implementarse, permite afirmar si el estimador es creíble de un efecto verdadero, en vez de simplemente comparar las estimaciones de OLS y IV según su tamaño relativo.

Pero al igual que observamos eso, también que la metodología que propone el Prof.Ciacci no es usada por los estudiosos actualmente. La mayoría de los estudios estadísticos dejan de lado estas importantes métricas para descubrir sesgos, resultando conclusiones no del todo precisas. Así, el objetivo del presente trabajo no es introducir una metodología nueva, si no con los últimos avances en el campo, presentar una manera más sencilla de poder ponerlos en práctica.

De esta forma, propondremos para el caso multivariante una matriz que permita simplificar los cálculos, y expondremos los pasos para llegar a ello.

Iremos explicando a continuación paso a paso lo que hacemos, pero queremos resaltar, como veremos, la simplificación que hacemos de w para llegar al final con cálculos más sencillos al igual que la decisión de coger por columnas la matriz X.

3.3 Explicación detallada de la metodología propuesta

Para llegar a la forma matricial de la ecuación vista en el apartado 3.2, de delta en función de beta, vamos a ir paso a paso explicando el procedimiento seguido, además de ir aclarando los diferentes elementos que lo componen por el camino.

Paso 1: Definición de las matrices y conceptos relevantes

El primer paso para poder realizar este cambio es comenzar escribiendo de forma matricial la ecuación PRF en forma matricial y explicar sus componentes y dimensiones. Esta se escribiría tal que así:

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

Con ánimo de simplificarla, podemos expresarla como:

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

Como explicamos a continuación, al tener en cuenta la naturaleza de las variables omitidas, nosotros solo podemos calcular la siguiente regresión, que expresamos de manera simplificada:

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

En el contexto de una ecuación de regresión multivariante en forma matricial, tenemos los siguientes elementos clave que serán trasladados a nuestra nueva ecuación:

- Y_{ih}: Matriz de variables dependientes. Esta matriz es el resultado de la ecuación resuelta, cuyo nombre indica que se obtiene en función del resto de factores al otro lado del signo del igual. Sus dimensiones son nx1, donde n es el número de observaciones, ya que a cada observación se le asignará un valor de Y.

$$Y_{ih} = [y_1; y_2; \ldots; y_n]$$

- X_{ih}: Matriz de variables independientes. Esta matriz tiene dimensiones n x k, donde n es el número de observaciones y k es el número de variables independientes. Cada fila de la matriz Xih corresponde a una observación, y cada columna corresponde a una variable independiente. Podemos representar esta matriz como:

$$X_{ih} = [x1,1\ x1,2\ \ldots\ x1,k;\ x2,1\ x2,2\ \ldots\ x2,k;\ \ldots\ \ldots\ \ldots\ \ldots;\ xn,1\ xn,2 \ldots xn,k]$$

- $\beta\hat{}$: Coeficiente escalar estimado del tratamiento. Este tiene dimensiones 1 x 1.

- β: Este es el coeficiente verdadero del tratamiento para el conjunto de variables independientes. También tiene 1x1 dimensiones.

- d_{ih}: Matriz de tratamiento. Este vector tiene dimensiones n x 1. El término "tratamiento" se refiere a una variable o factor que se introduce como parte de un estudio o experimento para investigar su efecto sobre una variable dependiente. Su objetivo en la regresión es determinar si existe una relación causal entre el tratamiento y la variable dependiente, y determinar qué magnitud y dirección tiene esta relación. Podemos representarlo en forma matricial como:

$$d_{ih} = [d1;\ \ d2;\ \ \ldots;\ \ dn]$$

- $\tilde{d}_{ih}$: Residuo de una regresión de d en X. Representa la diferencia entre los valores reales de la variable dependiente d y los valores estimados por el modelo de regresión. Matemáticamente, se define como $d_{ih} = \tau_0 + \tau_1 X_{ih} + \tilde{d}_{ih}$. Su análisis es fundamental para evaluar la calidad del modelo y detectar posibles problemas en la relación entre las variables. En forma matricial resulta:

$$\tilde{d}_{ih} = [\tilde{d}1; \quad \tilde{d}2; \quad \dots; \quad \tilde{d}n]$$

- w_{ih} : Matriz de los controles no observables. La matriz de controles no observables en un modelo de regresión múltiple es una matriz de variables explicativas o covariables que no se observan directamente en los datos, pero que se utilizan como controles para controlar el efecto de posibles variables de confusión. La dimensión de esta matriz depende del número de observaciones y del número de variables de control no observables que se incluyan en el modelo.

  Si tienes "n" observaciones y "p" variables de control no observables, la dimensión de la matriz de controles no observables será (n, p). Cada fila de la matriz representa una observación y cada columna representa una variable de control no observable. Cada fila de la matriz w_{ih} corresponde a una observación, y cada columna corresponde a una variable no observable. Podemos representar esta matriz como:

$$\begin{aligned} w_{ih} = [&w1{,}1\ w1{,}2\ \dots\ w1,p; \\ &w2{,}1\ w2{,}2\ \dots\ w2,p; \\ &\dots\ \dots\ \dots\ \dots; \\ &wn,1\ wn,2 \dots wn,p] \end{aligned}$$

  Sin embargo, en nuestro trabajo hemos decidido simplificar su estudio, y utilizar esta matriz con dimensiones n x 1, asumiendo que solo contamos con una variable de control no observable, pero pudiendo ampliar su estudio en el

caso que se quiera realizar como definida *supra*, donde lo único que hay que hacer es replicar la metodología y tratar con cuidado las dimensiones.

Así, nuestra w_{ih} va a ser definida como:

$$w_{ih} = [w1;\ w2;\ \dots\ ; wn]$$

- γ: Coeficiente escalar del control no observado. Tiene dimensiones 1x1 como su propia definición indica.

- $\theta 1$, $\theta 2$: Son los vectores que contienen los coeficientes de los controles observados. Su dimensión es kx1 al ser matrices con una columna y n filas.

$$\theta_{ih} = [\theta_1; \theta_2;\ \dots;\ \theta_n]$$

Paso 2: Definición de matrices de varianzas y covarianzas

Con ánimo de hacer el estudio lo más comprensible posible, vamos a explicar además de las matrices, el significado de la varianza y covarianza que está implícito en las mismas.

La varianza es una medida estadística que cuantifica la dispersión o la variabilidad de un conjunto de datos con respecto a su media. Es una medida de dispersión que muestra qué tan dispersos están los valores individuales alrededor de la media. Una varianza más alta indica una mayor dispersión de los datos alrededor de la media, mientras que una varianza más baja indica una menor dispersión y una mayor agrupación de los datos.

En nuestra ecuación, tenemos tres varianzas, por lo que estamos teniendo en cuenta como inputs para nuestro delta la dispersión tanto del residuo de la regresión de d en X, como de las variables independientes, como de los no observables.

En cuanto a la covarianza, una medida estadística que evalúa la relación lineal entre dos variables aleatorias, en nuestro caso los errores y las variables independientes. Mide cómo

se mueven conjuntamente y si tienden a variar en la misma dirección (covarianza positiva) o en direcciones opuestas (covarianza negativa).

Ahora, para convertir esta ecuación a forma matricial, debemos representar los términos $Cov(dih, Xih)$ y $Cov(dih, wih), Var(dih), Var(Xih)\ y\ Var(wih)$, en sus respectivas formas. Así, las dimensiones que encontramos en este tipo de matrices son, respecto a las covarianzas:

- Σ_{DX}: Esta matriz es la covarianza cruzada de nuestras matrices $dih\ y\ Xih$. La misma contiene las covarianzas entre la variable del tratamiento (d) y todas las variables observables del modelo (x). Su dimensión es 1 x k

- Σ_{DW}: Esta es la covarianza entre la matriz tratamiento y la de controles no observados. Como hemos explicado antes, hemos simplificado las dimensiones para facilitar cálculos, y por lo tanto tenemos como resultado un escalar formado por la covarianza de ambos $dih\ y\ Wih$

Respecto a las varianzas, es más sencillo, donde vemos que:

- Σ_D: Es matriz de varianza de dih. Sus dimensiones son 1x1

- Σ_W: Esta es la matriz de varianza de Wih, que también tiene dimensiones 1x1

- Σ_X: Esta es un poco más complicada, ya que introduce el término de varianza-covarianza dadas las dimensiones de X. Es una matriz cuadrada de dimensión k x k que recoge las varianzas en la diagonal principal y las covarianzas en los elementos de fuera de la diagonal principal.
  La técnica tiene múltiples aplicaciones en diversos campos. Es ampliamente utilizada en econometría, especialmente en el cálculo matricial de los coeficientes de regresión lineal a través del método de Mínimos Cuadrados Ordinarios. Además, se emplea en otras áreas como las finanzas, donde se utiliza para obtener una visión general de la volatilidad de los activos financieros, entre otros usos.

Paso 3: Coeficiente de proporcionalidad en forma matricial

Visto esto, volvamos al coeficiente de proporcionalidad visto en el apartado 3.1. En este caso, hemos decidido hacer uno por cada variable de control, haciendo uso de cada una de las columnas de X_{ih} para así lograr que funcione con la variable omitida, de manera que tenemos:

$$\delta_j \frac{Cov(d_{ih}, X_j)}{Var(X_j)} = \frac{Cov(d_{ih}, w_{ih})}{Var(w_{ih})}$$

Que nos sirve para todas las columnas que existan entre 1 y k. De esta manera, siendo X_j el vector con n observaciones de la variable de cada columna como decíamos. Vemos que, aunque hemos hablado de delta anteriormente, no la hemos definido para el caso, el cual sería:

- δ: Es una matriz, de dimensiones 1 x k, que contiene todos los coeficientes de proporción. Como podemos observar, tendrá tantas columnas como variables observables tengamos en la ecuación de partida.
  En la ecuación que acabamos de escribir, lo ponemos en función de j porque estamos operando por columnas con ánimo de simplificar cálculos.

Con esto visto, podemos afirmar que en notación matricial nos quedaría:

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

Hacer notar que hemos jugado con las propiedades matriciales para, en vez de dividir, multiplicar por la inversa.

Paso 4: Sesgo de variable omitida en forma matricial

Esta ecuación ya fue presentada en el apartado 3., igual que hablamos de los residuales de la regresión de d sobre X en el paso 1 de las definiciones. Necesitamos, para seguir desarrollando nuestra matriz, esta última expresión en forma matricial, la cual quedaría:

$$d_{ih} = X_{ih} * T + \tilde{d}_{ih}$$

Sabiendo ya las dimensiones de todos los factores salvo de T, lo definimos tal que:

- T: vector de coeficientes, tiene dimensiones de k x 1

Más allá de eso y volviendo a la expresión que nos determina el sesgo, los faltarían dos matrices que definir. Digo dos, y no tres, porque al asumir que las observables y las no observables son ortogonales entre sí, al igual que hacía Oster, podemos ignorar $Cov(X_{ih}, w_{ih})$, ya que es igual a cero.

Así, definimos:

- $\Sigma_{\tilde{D}}$: La matriz de varianza de $\tilde{d}_{ih}$, que teníamos pendiente anteriormente y cuyas dimensiones son 1x1

- Σ_{XW}: Siguiendo la misma lógica que para la matriz Σ_{DX}, la definimos como la 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. Tiene dimensiones k x 1

Vistas estas nuevas definiciones, la ecuación que habla del sesgo por variable omitida en forma matricial es representada como:

$$\hat{\beta}_2 = \beta_1 + (\gamma \Sigma_{DW} - \mathrm{T}_1^t \Sigma_{XW})\Sigma_{\tilde{D}}^{-1}$$

Donde también hemos hecho uso de las propiedades de las matrices para multiplicar por la inversa. Pero tenemos más lugares por los que simplificar, ya que al haber asumido que las variables omitidas son ortogonales a las de control, sus varianzas serán cero ($\Sigma_{XW} = 0$) por lo que podemos expresar aquello como:

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

Paso 5: Juntar resultados y ordenar

De esta manera tenemos dos ecuaciones en forma matricial, resultado del paso 3 y del paso 4. Por ello, lo único que tenemos que hacer es despejar para que quede en un lado solo δ_j, resultando la ecuación final en forma matricial:

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

3.4 Ventajas de la matriz vs. ecuación

La formulación de la ecuación delta en función de la matriz beta en una regresión multivariante, representada de manera matricial, presenta una serie de ventajas y mejoras significativas en comparación con su formulación en forma de ecuación.

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

Además, al tener la ecuación en forma matricial, podemos aprovechar las propiedades algebraicas de las matrices, lo que nos permite 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 importante de la representación matricial es que nos permite tratar conjuntamente múltiples observaciones y estimaciones. Al tener todas las observaciones

y coeficientes de regresión agrupados en matrices, podemos realizar análisis estadísticos más completos y exhaustivos. Podemos 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 representación matricial de la ecuación delta nos permite extender el marco de análisis a situaciones más complejas, como modelos de regresión multivariante con más de una variable dependiente. Al emplear matrices, podemos abordar de manera más eficiente y sistemática problemas que implican múltiples variables de respuesta y múltiples variables explicativas.

En resumen, 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. DESARROLLO DE UN TEST ESTADÍSTICO

4.1 Introducción a la prueba de Durbin–Wu–Hausman

La prueba de Durbin-Wu-Hausman es una prueba estadística utilizada en econometría para evaluar la presencia de sesgo de endogeneidad en un modelo de regresión. Como hemos comentado a lo largo del trabajo, el sesgo de endogeneidad ocurre cuando una variable explicativa está correlacionada con el término de error del modelo, lo que puede conducir a estimaciones sesgadas e inconsistentes de los coeficientes.

La idea básica detrás de la prueba de Durbin-Wu-Hausman es comparar las estimaciones de dos modelos diferentes: uno que asume que no hay endogeneidad (modelo consistente) y otro que permite la presencia de endogeneidad (modelo inconsistente). La prueba se basa en la comparación de las diferencias entre las estimaciones de los coeficientes de los dos modelos (Greene, 2012).

Así, bajo la hipótesis nula, ambas estimaciones son consistentes y no hay sesgo de endogeneidad en los coeficientes estimados del modelo de regresión, y bajo la alternativa, solo una es consistente. El procedimiento general de la prueba de Durbin-Wu-Hausman se puede resumir en una serie de pasos:

1. Estimar el modelo consistente (también conocido como modelo "reducido") utilizando un método que controle la endogeneidad, como el método de variables instrumentales (IV).

2. Estimar el modelo inconsistente (también conocido como modelo "completo") que incluye todas las variables explicativas, incluida la variable potencialmente endógena.

3. Calcular el estadístico de prueba. Para ello, necesitamos las diferencias entre las estimaciones de los coeficientes en ambos modelos, a las que vamos a denominar *diff*. Además, también es requisito calcular las varianzas de ambos coeficientes. Una vez tenemos estos dos datos, se estima el estadístico de prueba de Durbin-

Wu-Hausman, que es el de Wald basado en los datos que acabamos de calcular, cuya fórmula general es:

$$DWHA = (diff)' * (Var(\beta_0) - Var(\beta_1))^{-1} * diff$$

4. Distribución del estadístico. Comparar el valor del estadístico de prueba con una distribución de referencia, generalmente la distribución chi-cuadrado, para determinar si hay evidencia de endogeneidad.

5. Toma de decisiones. Si el valor del estadístico de prueba es significativamente diferente de cero, se rechaza la hipótesis nula de no endogeneidad, lo que sugiere la presencia de sesgo de endogeneidad en el modelo.

En resumen, la prueba de Durbin-Wu-Hausman es una herramienta útil para detectar y corregir el sesgo de endogeneidad en los modelos de regresión, lo que ayuda a mejorar la precisión y confiabilidad de las estimaciones de los coeficientes.

4.2 Elaboración propia de un test estadístico

Visto el test anterior, nos ha parecido interesante adaptarlo para nuestro trabajo, con el objetivo de evaluar la validez del coeficiente de proporcionalidad (δ) en el contexto de la regresión lineal múltiple con variables instrumentales.

Así, describimos a continuación, con la estructura utilizada en el anterior apartado, los pasos a seguir para estudiar la hipótesis nula, que viene a ser igual que en anterior apartado:

1. Estimar el modelo consistente: en esta, se estima el modelo de regresión utilizando el método de variables instrumentales (IV) para obtener los coeficientes estimados, incluyendo el coeficiente de proporcionalidad delta (δ) que nos interesa. Al coeficiente que resulta le denominamos *delta_IV*.

2. Estimar el modelo inconsistente: en este segundo paso, se estima el mismo modelo, pero utilizando el método de mínimos cuadrados ordinarios (MCO) para

obtener otro conjunto de coeficientes estimados, incluyendo el coeficiente de proporcionalidad, al que denominaremos *delta_MCO.*

3. Calcular el estadístico de prueba. Para ello, calculamos la diferencia entre los estimadores IV y MCO del coeficiente de proporcionalidad delta, al que denominaremos *delta_diff.* Una vez tenemos esto, calculemos las varianzas de *delta* tanto en el modelo IV como en el MCO para así tener todos los datos necesarios para calcular el estadístico de prueba DWH:

$$DWHA_{2=}\, delta\_diff' * ((Var(delta\_IV) - Var(delta\_MCO))^{-1} * delta\_diff$$

4. Distribución del estadístico. Bajo la hipótesis nula comentada, delta es consistente y no hay problemas de endogeneidad, donde DWH seguiría una distribución chi-cuadrado con k grados de libertad. Estos grados se calcularían como la diferencia en el número de restricciones entre los modelos IV y MCO.

5. Toma de decisiones. Así, se compara el valor obtenido de la estadística de prueba DWH con el valor crítico de la distribución chi-cuadrado correspondiente al nivel de significancia deseado. Si el valor obtenido de DWH es mayor que el valor crítico, se rechaza la hipótesis nula y se concluye que hay evidencia de que delta es mayor a 1, lo que sugiere una dependencia significativa de factores no observables en relación con los observables.

4.3 Diferencias y justificación

Ya hemos explicado con anterioridad la razón de ser de nuestra base, el DWH. Al estudiarlo, vimos que con lo que opera en su caso son los coeficientes, y dado que como hemos visto en el apartado previo, delta depende de ellos y viceversa, se nos ocurrió adaptar este estudio para comenzar por otro punto de partida que pensamos que puede generar los mismos resultados.

Sabiendo la importancia que tiene que delta sea mayor que uno, hemos preferido poner el enfoque en este aspecto y desarrollar nuestro propio test.

Creemos que era una forma sencilla de adaptarlo ya que beta y delta tienen las mismas dimensiones, por lo que los cálculos no tenían que hacer otra cosa que dejarse estables.

Aunque compartamos hipótesis nula con el DWH, podemos afirmar que nos da información ligeramente diferente, ya que nos permite evidenciar el tamaño de delta y las conclusiones que eso mismo conlleva, como sabemos ya.

5. CONCLUSIONES

5.1 Recapitulación de los objetivos de investigación

Como explicábamos en un principio, nuestro trabajo tenía lo que podemos reorganizar como tres objetivos fundamentales.

El primero de ellos, era dar al lector una introducción lo más amplia posible, pero en términos sencillos de todo el recorrido hasta llegar al corazón del trabajo, para una mayor compresión de este. Pensamos que esta explicación detallada permite que no solo los principales estudiosos, si no cualquier persona que se lo proponga, pueda hacer uso de los conocimientos y aportaciones que aquí se exponen.

Así, comenzamos explicando primero la importancia que tiene para el investigador el uso de estos métodos y explicamos ambos, el de variables instrumentales y el de mínimos cuadrados ordinarios. Una vez vistos los conceptos básicos, pasamos a otra fase que también consideramos de gran relevancia: meternos en los principales estudios que abordan el tema en relación con el nuestro, y ver de manera suscinta la evolución que se ha ido produciendo a lo largo del tiempo. Después, identificamos las lagunas de la literatura en las que nos vamos a apoyar para llegar al segundo objetivo.

Este es escribir la ecuación que el Prof. Ciacci tiene en el apéndice de su trabajo para obtener delta en función de beta en forma matricial. En este apartado vamos explicando los diferentes componentes de esta y la lógica que hemos seguido detrás, para llegar a una forma simplificada de ecuación de la que exponemos las ventajas seguidamente.

El tercer objetivo era añadir otro componente técnico, y así proponer una nueva perspectiva para el test de Durbin, que nos permitiese comprender su utilidad y poder enfocarlo a través de delta, al ser este último esencial a lo largo de todo nuestro texto.

Nuestro cuarto objetivo rodea todo el trabajo, y es el de concienciar a las personas de la importancia del uso de este tipo de métodos. Por ello, queríamos que todo el mundo pudiese entenderlo, a la par que estableciámos una metodología sencilla a través de la cual facilitásemos su aplicación.

5.2 Resumen de los hallazgos obtenidos

En este caso, nosotros hemos aprendido personalmente la importancia de controlar la validez de un estudio y todos los factores además de métodos que se tienen por detrás para ello.

Además, yendo por partes, hemos podido investigar los dos diferentes métodos en los cuales nos hemos centrado. Después, las diferentes aplicaciones que los diferentes autores les daban, y dentro de estos, cómo lo aplicaban a casos prácticos para sacar sus conclusiones.

Leyéndolos, nos hemos dado cuenta de que no siempre los posteriores siguen los hallazgos o recomendaciones de los anteriores, pero todo lo que sean aportaciones nuevas debieran ser estudiadas por los últimos en introducir conclusiones, para no trabajar sobre cosas que ya se han estudiado, o incluso para refutarlas y cambiarlas de rumbo, tal y como vimos que se hizo con ciertas conclusiones de Oster.

En lo relativo al núcleo de nuestro trabajo, nuestro máximo hallazgo ha sido el punto tercero, en el cual desarrollamos la ecuación hasta llegar a su forma matricial. Así, al ir desglosando paso por paso, hemos ido pensando en las diferentes dimensiones de las matrices, utilizando sus propiedades y entendiendo el uso que tiene delta para este tipo de estudios.

Hemos buscado diferentes formas de simplificar nuestros cálculos, esencialmente a partir de dos premisas: la primera, que la matriz W, aquella que contiene los controles no observables, la empezamos planteando como una matriz n x p, como explicamos, pero al hacer los diferentes cálculos había dimensiones que no nos dejaban continuar con ellos.

Decidimos, ya que seguía teniendo sentido, darle las dimensiones que hemos utilizado, y así se simplificaron. El segundo punto que nos ayudó enormemente fue enfocar, vamos a decir “por partes” la forma de la matriz X. Descubrimos que, haciéndolo por columnas, se simplificaban los cálculos de gran manera y era más sencillo seguirlos.

Aprendimos las implicaciones de las varianzas, covarianzas, y la matriz varianza-covarianza en este contexto, y a decir verdad, refrescamos muchos conocimientos de una estadística que vimos pero que teníamos un poco oxidada, además de aprender muchas cosas nuevas que no son enseñadas en la universidad.

Respecto el estudio estadístico propuesto, creemos que es posible que sea de aplicación y el potencial que tiene para que se le extraigan conclusiones valiosas.

5.3 Limitaciones del estudio y posibles áreas de mejora futura

Creemos que, dada la complejidad y tecnicismos del asunto, aunque hayamos intentado explicarlo de forma sencilla, igual al ser un trabajo de esta envergadura debiese de haber sido adoptado desde un punto de vista más técnico en su totalidad.

Si hubiese sido así, es verdad que no hubiese sido entendido por todo el público que lo quiera leer, cosa que ha sido uno de nuestros objetivos desde el primer momento.

Aun así, si lo hubiésemos querido hacer más complejo, creemos que aun teniendo diferentes estudios y herramientas, con una base más solida de conocimientos al respecto podríamos haber sacado alguna otra conclusión relevante después de dedicarle tanto tiempo a entender los textos.

A su vez, podemos ver diferentes áreas de mejora, que son aplicables a nuevos estudios. La primera, es el haber elaborado un test estadístico para probar delta, el cual se ha intentado, pero del cual decidimos prescindir al no estar del todo seguros de lo que aportaba o si era correcto lo que estábamos midiendo.

Además, habiendo hecho tanto hincapié en la ecuación que hemos desarrollado, creemos que podríamos haber elaborado, como fue propuesto por el profesor, una tabla de deltas que sirviese de ejemplo para nuestras conclusiones. La mayor dificultad podía surgir en este caso en encontrar un buen artículo que sirviese como base.

Otra área de mejora, desde una perspectiva diferente, hubiese sido incluir otras formas de desarrollar la ecuación, cosa que también pensamos, pero que descartamos al pensar que

aportaba más a la parte matemática y menos al objetivo de simplificar la ecuación para darle un mayor uso.

Dentro de esto último, nos aventuramos a explorar el estadístico de DWH y proponer, con su misma hipótesis una variación. Éramos bastante conscientes de su complejidad por lo que, aunque a priori estemos satisfechos con el resultado, entendemos que pueden derivarse limitaciones del mismo.

5.4 Principales problemas encontrados

Como hemos mencionado antes, personalmente ha sido un reto grande enfrentarme a este trabajo de fin de grado. Creo que se necesitan unos conocimientos base para poder desarrollarlo y entenderlo en profundidad, no solo relativo a la econometría, sino a toda la matemática que lleva detrás.

Así, al principio, dediqué una gran parte del tiempo a entender cada una de las palabras que salían en los diferentes estudios, de las que estoy convencida que algún concepto se me sigue escapando.

Pero entre otras cosas, eso fue una de las razones que me llevaron a comenzar a escribir y explicar poco a poco con el nivel de detalle que consideraba que otros lectores inexpertos como yo podían llegar a entender.

Luego, me centré en la ecuación. Desarrollé como mínimo cinco formas diferentes, y hasta llegar a la que se expone en el trabajo, ninguna de las ideas que intentaba conseguían darme el resultado que buscaba. Estoy convencida y sé que hay varias formas de hacerlo, pero a decir verdad esta era la más sencilla y la que acabó resultando.

Cuando pensaba dar por terminado el trabajo, me sentí impulsada a examinar la metodología aplicada por Durbin, y con ello proponer una manera que no variara mucho pero que pudiese ser útil.

Así, no tengo demasiado claro dónde dediqué más tiempo, si al principio del trabajo en entender los conceptos, o al final peleándome con la ecuación o el estadístico para conseguir sacar algo que tuviese sentido matemáticamente.

5.5 Relevancia y aplicaciones prácticas de los resultados obtenidos

Como se ha repetido bastantes veces a lo largo del trabajo, pensamos que un trabajo de fin de grado tan sencillo, pero con una aportación tan práctica a la vez, puede tener enormes utilidades.

Entre estas, permite a las personas menos enteradas hacerse una idea sumaria de lo que es la econometría, y que dos métodos nos ayudan a resolver los problemas de endogeneidad. Además, propone conocer a diferentes estudiosos y sus aportaciones, y ver cómo el estado de la ciencia va evolucionando con el paso del tiempo.

No contentos con eso y por si hay alguna persona que le interesa profundizar un poco más, nos adentramos en el artículo del Prof.Ciacci para desgranarlo y proponer una ecuación de su apéndice en forma matricial, que puede servir para simplificar mucho los cálculos, sobre todo en el mundo de calculadora-ordenador en el que vivimos. Además, propusimos una variación del test de Durbin, de la cual pensamos posible extraer conclusiones factibles.

Por todo ello y por la relevancia que tienen unos estudios que busquen ser lo más veraces posibles a la realidad, podemos afirmar haber contribuido con un grano de arena a que en estudios posteriores, los investigadores tanto más eruditos como menos, tengan en cuenta la importancia de esto y se le pueda aplicar la metodología propuesta con dicho propósito.

6. BIBLIOGRAFÍA

Acemoglu, D., Johnson, S., & Robinson, J. A. (2001). *The Colonial Origins of Comparative Development: An Empirical Investigation*. The American Economic Review, 91(5), 1369-1401.

Altonji, J. G., Elder, T. E., & Taber, C. R. (2000). *Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools.* National Bureau of Economic Research, Working Paper 7831.

Becker, S. O. (2016). *Using Instrumental Variables to Establish Causality.* Universities of Warwick (UK) and IZA (Germany).

Ciacci, R. M. (2021). *A Matter of Size: Comparing IV and OLS Estimates.* Universidad Pontificia de Comillas, Department of Economics and Business Management.

De Luca, G., Magnus, J. R., & 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".* Universities of Palermo, Amsterdam, and Georgetown.

Greene, William H. (2012). *Econometric Analysis* (7th ed.). Pearson. pp. 379–380, 420. ISBN 978-0-273-75356-8.

Hendry, D. F., & Mizon, G. E. (2014). Endogeneity in Econometric Practice. Cambridge University Press.

Lochner, L., & Moretti, E. (2014). *Estimating and Testing Models with Many Treatment Levels and Limited Instruments.* Universities of Western Ontario and California-Berkley.

Masten, M. A., & Poirier, A. (2020). *Salvaging Falsified Instrumental Variable Models.* Universities of Duke and Georgetown.

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

Staiger, D., & Stock, J. H. (1994). *Instrumental Variables Regression With Weak Instruments.* National Bureau of Economic Research, Working Paper 151.

Wooldridge, J. M. (2009). *Introductory Econometrics: A Modern Approach*. South Western Educational Publishing.