FACULTAD DE CIENCIAS EMPRESARIALES

MODELOS DE INFERENCIA CAUSAL EN ECONOMETRÍA

Autor: José Loring Meneses de Orozco

5º E3 Analytics

TFG Business Analytics

Tutor: D. Riccardo Ciacci

Madrid

Junio 2023

RESUMEN

El siguiente Trabajo de Fin de Grado pretende establecer un procedimiento estandarizado que permita comparar la precisión de los métodos de estimación de Mínimos Cuadrados Ordinarios (MCO) y Variables Instrumentales (VI) en el contexto de relaciones de causalidad en modelos econométricos. El trabajo parte de las bases establecidas por el autor R. Ciacci en su publicación *A Matter of Size: Comparing IV and OLS Estimates*. El objetivo del trabajo consiste en transformar la ecuación planteada en la publicación de R. Ciacci, expresarla en su forma matricial y exponer una serie de conclusiones al respecto.

Palabras clave: endogeneidad, regresión, Mínimos Cuadrados Ordinarios, Variables Instrumentales, coeficiente de proporcionalidad.

ABSTRACT

The following Final Degree Project aims to establish a standardised procedure to compare the accuracy of Ordinary Least Squares (OLS) and Instrumental Variables (IV) estimation methods in the context of causality relations within econometric models. The paper is based on the groundwork set out by author R. Ciacci in his publication *A Matter of Size: Comparing IV and OLS Estimates*. The aim of the paper is to transform the equation set out in R. Ciacci's publication, express it in its matricial form and present a series of conclusions with respect to the issue.

Keywords: endogeneity, regression, Ordinary Least Squares, Instrumental Variables, coefficient of proportionality.

ÍNDICE

1. INTRODUCCIÓN

1.1 Contexto del estudio

El siguiente Trabajo de Fin de Grado para el grado de Business Analytics trata sobre una serie de factores que afectan de manera negativa a los resultados obtenidos en modelos econométricos. El problema de la endogeneidad en modelos econométricos surge cuando existe correlación entre las variables explicativas y los términos de error. La correlación puede resultar en que las estimaciones de los parámetros del modelo estén sesgadas y devengan ineficaces para establecer una relación de causalidad correcta.

La endogeneidad surge cuando la relación entre dos variables es bidireccional, por lo que resulta complicado establecer una relación clara de causa-efecto entre ellas. Desde una perspectiva práctica, en un contexto económico, la endogeneidad suele producirse en situaciones donde una variable es simultáneamente determinada por otras variables, que a su vez causa un efecto sobre las mismas. Este problema puede presentar desafíos para la inferencia causal en la investigación empírica al no poder definirse de forma precisa las variables involucradas y los factores de causalidad.

El ejemplo epítome de la endogeneidad es el análisis de la relación entre el nivel de educación y el salario de una persona. El escenario inicial parte de que si una persona ha recibido mucha educación de calidad a lo largo de su vida académica es lógico que a posteriori pueda acceder a un puesto de trabajo altamente remunerado. A partir de esta afirmación se podría deducir que la educación es la causa principal del aumento salarial. Esta deducción sería incorrecta ya que no se está teniendo en cuenta la influencia simultánea del salario en la decisión de obtener más educación: una persona con un salario alto dispone de los recursos necesarios para invertir más en educación.

En situaciones como el ejemplo anterior, los métodos estadísticos tradicionales pueden derivar resultados sesgados o incorrectos. Para mitigar las impresiones fruto de las relaciones bidireccionales, los académicos utilizan técnicas de estimación y modelos econométricos más sofisticados, como el uso de variables instrumentales o la regresión en dos etapas, que permiten controlar la endogeneidad y obtener estimaciones más precisas de las relaciones causales.

1.2 Justificación del tema

En econometría, un enfoque común consiste en interpretar las diferencias considerables entre el tamaño de los coeficientes de los Mínimos Cuadrados Ordinarios (OLS, por sus siglas en inglés) y las Variables Instrumentales (IV) como prueba de validez del instrumento (*Ciacci, 2021*).

Este trabajo parte de la base de que la regresión OLS puede proporcionar una gran cantidad de información sobre el efecto que se pretende estimar. La justificación de este tema en concreto reside en el hecho de no existe una metodología formal para comparar las dos estimaciones mencionadas supra.

La autora Oster, en su publicación de 2019, utiliza la información obtenida a través de la regresión OLS para estimar el conjunto de valores donde efectivamente debería situarse el verdadero efecto del tratamiento. El tamaño de dicho conjunto dependerá de la cantidad de información que proporcionen los observables sobre los no observables.

En consecuencia, esta metodología (que se desarrollará más adelante, vid. 1.4 Metodología empleada), permite calcular un parámetro para desarrollar un argumento formal de limitación, conocido como *coeficiente de proporcionalidad*, que mide el tamaño relativo de la proporcionalidad entre la selección de observables y no observables (*Ciacci, 2021*).

La elección del tema desarrollado a lo largo de este trabajo se justifica por lo tanto en esa ausencia de una metodología formal estandarizada para comparar las estimaciones obtenidas a través de Mínimos Cuadrados Ordinarios y Variables Instrumentales. Al estructurar un rango de valores dentro del cual debería encontrarse el efecto del tratamiento se permite evaluar la precisión de los resultados obtenidos.

A través de las consideraciones planteadas en el trabajo, se pueden estructurar una metodología sistemática, lo que permite analizar una gran variedad de supuestos donde dicha metodología sea pertinente de aplicación, tanto desde una perspectiva teórica como en el plano práctico, especialmente para una mayor comprensión de las relaciones causales de carácter económico.

1.3 Objetivos de la investigación

La investigación se desarrolla en el contexto de la endogeneidad y las limitaciones que produce a la hora de realizar estimaciones en un modelo econométrico. El objetivo principal del estudio es establecer una metodología formal que permita evaluar los estimadores de Mínimos Cuadrados Ordinarios (OLS) y Variables Instrumentales (IV).

El punto de partida es utilizar la información obtenida de la regresión OLS para determinar el rango de valores donde se localice el efecto del tratamiento. Este proceso permite evaluar de manera estandarizada la precisión de los resultados obtenidos a través de los estimadores OLS e IV.

La siguiente fase del análisis consiste en utilizar la metodología planteada para poder obtener conclusiones relativas a la validez del instrumento utilizado en el modelo de IV. El objetivo es que los valores mayores (o menores) del *coeficiente de proporcionalidad*, que mide el tamaño relativo de la proporcionalidad entre la selección de observables y no observables, pueden proporcionar evidencia en contra (o a favor) de la precisión de los resultados obtenidos (*Ciacci, 2021*).

Por último, en fundamental la exhaustividad del análisis conducido entre los resultados obtenidos a través de los estimadores OLS e IV siguiendo el proceso planteado. A medida que se identifique un mayor número de diferencias entre los estimadores se podrá determinar con mayor exactitud que implicaciones suponen dichas diferencias en el modelo de inferencia causal. En consecuencia, las conclusiones obtenidas sobre el efecto del tratamiento en cuestión gozarán de un mayor grado de precisión.

Por otro lado, aparte de estructurar el proceso sistemático para la comparación de estimadores, se desarrollará un análisis sobre los estudios principales relacionados con la problemática en cuestión, de tal manera que se puedan, por un lado, determinar las bases de partida del trabajo y, por otro lado, conocer en profundidad las visiones de otros autores, así como la evolución doctrinal, en materia de la evaluación del sesgo generado por variables no observables a través de OLS en modelos de inferencia causal en econometría.

1.4 Metodología empleada

Para la realización de este Trabajo de Fin de Grado se ha utilizado como base la publicación del autor R. Ciacci "*A Matter of Size: Comparing IV and OLS Estimates*" del año 2021. A partir de este trabajo, se analiza la metodología para comparar la precisión ofrecida por dos estimadores diferentes en el contexto de relaciones causales en modelos econométricos, tanto desde una aproximación teórica como en la práctica, en este caso, entre los métodos de Mínimos Cuadrados Ordinarios (OLS) y Variables Instrumentales (IV).

De manera paralela, se analiza la publicación de la autora E. Oster "*Unobservable Selection and Coefficient Stability: Theory and Evidence*" del año 2019. La importancia de este trabajo reside en que sirvió como precedente para el trabajo de R. Ciacci, por lo que resulta de gran importancia comprender, tanto su fundamentación, como la evolución entre ambos trabajos y las diferentes aportaciones de los mismos.

En el estudio de R. Ciacci, los resultados y las conclusiones vienen datos por una ecuación multivariable de beta en función de delta. La verdadera aportación de este trabajo consiste en transformar dicha ecuación y representarla en forma matricial. Una vez se haya obtenido la matriz, se extraerán una serie de conclusiones evaluando las diferentes ventajas de ambas metodologías y la aplicabilidad de las mismas en situaciones diferentes.

En el siguiente apartado, se desarrollarán los principales conceptos técnicos en torno a los cuales gira el desarrollo del trabajo con el objetivo de asegurar una comprensión adecuada por parte del lector (*vid*. 2.1 Marco teórico). Una vez se hayan asentado las bases teóricas del trabajo, se procederá a realizar un análisis exhaustivo de toda la literatura relacionada con la causalidad en modelos econométricos y los factores que afectan a la precisión de los resultados. El objetivo de este apartado (*vid*. 2.3 Principales estudios relacionados) consiste en exponer de manera clara y concisa al lector como ha ido evolucionando la línea de pensamiento de la doctrina académica.

Finalmente, se realizarán los cálculos pertinentes para expresar la ecuación de beta en función de delta en su forma matricial y se expondrán las conclusiones finales del trabajo.

2. REVISIÓN DE LA LITERATURA

2.1 Marco teórico

En este apartado se analizan los principales conceptos que deben tenerse en cuenta para una correcta comprensión del estudio. En primer lugar, se tratará el concepto de *regresión* como la técnica principal utilizada en modelos de carácter econométrico para analizar y obtener conclusiones sobre una relación entre variables.

En segundo lugar, continuaremos con el concepto de *endogeneidad*. Dentro de la relación de variables en modelos econométricos, la endogeneidad hace referencia a la relación entre variables explicativas y variables dependientes. Se trata de una consecuencia que viene dada cuando las variables explicativas están correlacionadas con el término de error del modelo, generando sesgo en los resultados y disminuyendo la precisión de los estimadores.

En tercer lugar, se desarrollarán los dos métodos de estimación sobre los que recae el estudio del trabajo: por una parte, el método de los *Mínimos Cuadrados Ordinarios (OLS)*, como método para estimar los parámetros en una regresión de tal manera que los valores que se asignen a los coeficientes minimicen la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos y, por otro lado, el método de *Variables Instrumentales (IV)*, como método para mitigar los efectos negativos de la endogeneidad a través de variables exógenas correlacionadas con las variables endógenas, pero no con el término de error. De esta forma, los resultados obtenidos sobre las estimaciones son más consistentes y menos sesgados.

Por último, se analizará la relación entre ambos métodos de estimación, estimadores complementarios que ofrecen diferentes enfoques acerca de causalidad en un modelo. El método OLS es efectivo bajo la asunción de que las variables son exógenas mientras que, por su parte, el método IV se plantea como solución para mitigar el sesgo producido por la endogeneidad del modelo. Una vez se han desarrollado dichos conceptos, así como la relación que existe entre ellos, el planteamiento sobre una estructura capaz de comparar de manera sistemática ambos métodos cobra sentido.

(A) *Regresión*

En primer lugar, analizaremos el concepto de regresión. La regresión se trata de una técnica estadística comúnmente utilizada en el contexto de modelos econométricos con el objetivo de analizar y modelar relaciones entre variables. La idea fundamental del concepto reside en identificar una función matemática que se ajuste de manera adecuada a un conjunto de datos observados. De esta manera, se permite elaborar predicciones obteniendo conclusiones sobre la relación existente entre las variables.

Dentro de sus implicaciones principales, la regresión permite determinar la relación entre una variable dependiente y una o varias variables explicativas. Para facilitar la comprensión del lector, la variable dependiente se puede definir como aquella variable que se pretende explicar o predecir, mientras que, por otro lado, las variables explicativas se pueden definir como aquellas variables que aportan la información determinante de la variable dependiente. En consecuencia, la regresión permite estimar que efectos tienen las variables explicativas sobre la variable dependiente en un modelo econométrico.

Cuando se trabaja a través de un modelo de regresión hay que tener en cuenta una serie de asunciones que permitirán estimar los parámetros del modelo y realizar las inferencias pertinentes sobre la relación entre las variables. El principal factor que se debe considerar es el hecho de que se basa en supuestos sobre la forma funcional de la relación entre las variables. En consecuencia, se asume que la relación puede expresarse mediante una función matemática: (i) línea recta (regresión lineal); o (ii) curva más compleja (regresión no lineal).

La elaboración de un modelo de regresión implica la necesidad de estimar los coeficientes de regresión. Dichos coeficientes representan la magnitud y dirección del impacto de las variables explicativas sobre la variable dependiente, además de proporcionar información de gran relevancia acerca de la relación entre las variables que permite realizar comparaciones de carácter cuantitativo. Por otra parte, la regresión proporciona medidas que permiten evaluar la adecuación del modelo a los datos observados. Dentro de estas medidas de control podemos destacar el coeficiente de determinación (R^2) y el error estándar de la estimación. A partir de estas medidas es posible evaluar la calidad del modelo y su capacidad para determinar la variabilidad de la variable dependiente.

En un contexto más específico, en el campo de un modelo de carácter econométrico, la regresión es ampliamente utilizada para diversas finalidades. Una de sus aplicaciones de mayor relevancia es la estimación de la relación causal entre variables económicas. Con el objetivo de visualizar dichas funcionalidades en un plano más práctico, se puede utilizar la regresión para analizar el impacto del ingreso en el consumo, el efecto de las políticas monetarias en la inflación y demás implicaciones de carácter macroeconómico.

Continuando en la línea de las aplicaciones prácticas de la regresión, resulta muy útil para realizar pronósticos o predicciones. Utilizando el método de estimación más adecuado para el modelo en cuestión, se pueden utilizar los valores de las variables explicativas para predecir el valor de la variable dependiente en el futuro. En consecuencia, dichas predicciones permiten tomar decisiones informadas y planificar estrategias a partir de las tendencias observadas en los datos.

Por otra parte, hemos de tener en cuenta las limitaciones y desafíos que presenta la regresión en modelos econométricos. Dentro de dichas limitaciones podemos destacar la necesidad de cumplir con una serie de condiciones predeterminadas en el modelo, principalmente la relación lineal entre variables y la ausencia de errores de especificación. Del mismo modo, hemos de distinguir y tener en cuenta que la regresión no implica *per se* una relación de causalidad directa, sino que proporciona evidencia estadística de que existe una cierta asociación entre las variables del modelo.

En conclusión, la regresión no deja de ser una técnica fundamental en el contexto de modelos econométricos para analizar y determinar relaciones entre variables. No solo permite estimar los efectos de las variables explicativas sobre la variable dependiente, sino que además permite realizar predicciones y evaluar la precisión de las estimaciones del modelo.

Aunque se deba analizar de forma previa la concurrencia de una serie de factores mencionada *supra*, su aplicación adecuada puede proporcionar información de gran valor que puede ayudar a comprender y a obtener conclusiones precisas de en una variedad de supuestos, como por ejemplo sobre corrientes y ciclos en el campo de la macroeconomía.

(B) *Endogeneidad*

A continuación, trataremos otro concepto fundamental en el marco de nuestra investigación, el concepto de endogeneidad. Este concepto hace referencia a la relación entre las variables explicativas y la variable dependiente en un modelo económico.

A la hora de analizar el grado de causalidad entre las variables de un modelo, la endogeneidad se da en situaciones cuando las variables explicativas están correlacionadas con el término de error del modelo. De manera práctica, puede generar estimaciones sesgadas y conducir a conclusiones incorrectas o poco precisas.

Los factores que producen una situación de endogeneidad en un modelo pueden ser de diversa índole. Una de las principales causas que pueden ocasionar dicho resultado se trata de la omisión de variables relevantes en el modelo. En el caso de omitir variables que estuviesen correlacionadas tanto con la variable dependiente como con las variables explicativas que definen el modelo, estaríamos generando un escenario de endogeneidad por omisión, uno de los supuestos más comunes que ponen de manifiesto la importancia de seleccionar el método de estimación apropiado.

Otro de los factores que puede conducir a una situación de endogeneidad es la simultaneidad. Este concepto tiene lugar cuando existe una relación de retroalimentación entre la variable dependiente y las variables explicativas. En este caso en concreto, la causalidad puede ser bidireccional – los cambios en una variable afectan al resto y viceversa, lo que dificulta en gran medida determinar una dirección clara de inferencia causal en un modelo.

Para identificar la endogeneidad en un modelo, se ha de atender a circunstancias y características muy concretas que deberán ser analizadas *ex ante* de manera individual para anticiparse a la falta de precisión de las estimaciones y prevenir resultados inconsistentes. Por una parte, se deberá prestar atención a los factores indicativos de presencia de endogeneidad en las variables y, en ese caso, por otra parte, se deberá analizar si el modelo cuenta con los supuestos clave, necesarios para llevar a cabo una serie de técnicas de corrección.

De manera enunciativa y no exhaustiva podemos enumerar los siguientes factores:

(i) *Correlación entre las variables explicativas y el término de error*;

En el supuesto de que las variables explicativas estén correlacionadas con el término de error, las estimaciones de los coeficientes pueden derivar en resultados sesgados e inconsistentes, cuyas conclusiones sean erróneas al haber malinterpretado la relación causal entre las variables;

(ii) *Problemas de inferencia*;

Un análisis correcto y anticipado sobre la presencia de endogeneidad puede permitirnos deducir *ex ante* que el método de estimación de Mínimos Cuadrados Ordinarios (OLS) resultaría inconsistente, a la par de ineficaz, al no ser capaz de tratar de forma apropiada modelos con presencia de endogeneidad en las variables, llegando a poder aceptar hipótesis nulas o a descartar hipótesis alternativas idóneas; y

(iii) *Dificultad para establecer causalidad*.

Como se ha mencionado *supra*, en supuestos de causalidad bidireccional, donde varias variables influyen mutuamente entre sí, la existencia de dicha relación de retroalimentación dificulta en gran medida la determinación de una dirección causal clara en el modelo, sin poder diferenciar si los cambios en las variables explicativas son la causa de los cambios en la variable dependiente o viceversa.

De la misma forma, resulta pertinente recordar las principales limitaciones que experimenta un modelo con presencia de endogeneidad en sus variables y enumerar ciertos factores que deben tenerse en cuenta para mitigar sus resultados negativos.

(i) *Necesidad de técnicas de corrección*;

El uso de técnicas de corrección para abordar supuestos de endogeneidad puede contribuir a la obtención de resultados más consistentes y menos sesgados. Algunas de estas soluciones más apropiadas en estas situaciones pueden venir dadas por el uso del método de estimación de Variables Instrumentales (IV) o estructurar un modelo de ecuaciones simultáneas – un sistema de ecuaciones que relaciona las variables entre sí, teniendo en cuenta tanto las relaciones directas como las indirectas, teniendo en cuenta su interdependencia y los posibles efectos de retroalimentación;

(ii) *Supuestos clave*;

El éxito de las técnicas de corrección de endogeneidad queda supeditado a la concurrencia de una serie de supuestos clave. Dentro de dichos supuestos, podemos destacar la validez de los instrumentos utilizados o, en el caso mencionado supra acerca de las ecuaciones simultáneas, su correcta especificación dentro del modelo. En el caso de inexistencia de los supuestos clave los resultados obtenidos continuarán aflorando inconsistencias; y

(iii) *Datos limitados*.

Por último, puede darse la situación donde la disponibilidad de datos idóneos que permitan abordar la endogeneidad pueda ser limitada. Dicha carencia puede dificultar la aplicación de técnicas de corrección y limitar su efecto mitigador sobre las conclusiones obtenidas.

En conclusión, queda expuesta la gran relevancia de la endogeneidad en el marco de modelos de carácter econométrico, haciendo referencia a la correlación entre las variables explicativas y el término de error de un modelo económico. Resulta de vital importancia conducir un análisis exhaustivo de forma preventiva para evitar la obtención de resultados no deseados. La identificación de endogeneidad en las variables del modelo supone razón suficiente para seleccionar un método de estimación u otro, por ejemplo, descartar en uso de Mínimos Cuadrados Ordinarios (OLS) y decantarse por emplear Variables Instrumentales (IV).

(C) *Mínimos Cuadrados Ordinarios (OLS)*

El método de estimación de Mínimos Cuadrados Ordinarios (OLS) es una técnica empleada en el campo de la econometría con el objetivo de estimar los parámetros de un modelo de regresión. Su funcionalidad se basas en identificar los valores de los coeficientes que minimicen la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo. Desde una perspectiva práctica, en un supuesto de regresión lineal, el proceso consiste en determinar una línea recta que se ajuste de manera óptima a los datos observados. Para alcanzar dicho objetivo, se deben identificar los valores de los coeficientes que minimicen la suma de los errores al cuadrado, lo que implica identificar la línea de mayor proximidad a los puntos observados.

El método de estimación OLS cuenta con una serie de características fundamentales que se deberán tener en cuenta para determinar la adecuación del mismo. Dentro de las principales propiedades podemos enumerar:

(i) *El método OLS proporciona estimaciones lineales de los coeficientes*;

En otras palabras, el objeto de estimación reside en relaciones lineales entre las variables independientes y la variable dependiente. Esta linealidad facilita el cálculo de los coeficientes y las inferencias estadísticas asociadas.

(ii) *Las estimaciones proporcionadas no presentan sesgo con respecto a los coeficientes y las conclusiones derivadas serán consistentes, siempre y cuando se cumplan una serie de condiciones*; y

Dichas condiciones incluyen la linealidad de la relación, la ausencia de errores de especificación, la exogeneidad de las variables explicativas y la ausencia de multicolinealidad.

(iii) *El funcionamiento reside en el principio de identificar los coeficientes que minimicen la suma de las diferencias entre los valores observados y los valores predichos por el modelo*.

En un supuesto de regresión lineal, la línea recta que mejor se ajuste a los datos permitirá aumentar la precisión de las estimaciones y alcanzar resultados de mayor consistencia.

Sin embargo, es importante conocer las limitaciones que presenta el método OLS a la hora de realizar estimaciones en un modelo. Existen una serie de factores que, de estar presentes en el modelo, disminuirán en gran medida la veracidad de las estimaciones, por lo que resultará fundamental examinarlos *ex ante*:

(i) *Presencia de endogeneidad en las variables*;

Se trata de la principal limitación del método OLS. Como ya se ha mencionado con anterioridad, este fenómeno ocurre cuando las variables explicativas están correlacionadas con el término de error del modelo. En consecuencia, este hecho conducirá a estimaciones incorrectas y, por ende, a relaciones de causalidad imprecisas.

(ii) *Presencia de errores de especificación*; y

Los errores de especificación surgen cuando el modelo no es capaz de capturar de forma adecuada la verdadera relación causal que existe entre las variables. Como cualquier otra imprecisión en un modelo, conduce a resultados sesgados e inconsistentes.

(iii) *Presencia de multicolinealidad*;

Se trata de un supuesto de la alta correlación entre las variables explicativas. En este caso, la estimación de los coeficientes se vuelve inestable y difícil de interpretar. Por otra parte, el método OLS también es sensible a la presencia de valores atípicos o influencia de puntos. Estos valores extremos pueden distorsionar las estimaciones y afectar la interpretación de los resultados.

En conclusión, el método de Mínimos Cuadrados Ordinarios (OLS) conforma una del as principales técnicas utilizadas para estimar parámetros en modelos de regresión lineal. Proporciona estimaciones lineales y no sesgadas de los coeficientes, condicionado al cumplimiento de ciertos supuestos claves enumerados *supra*. Siempre hay que tener presente las limitaciones que presenta el método para realizar una correcta interpretación de los resultados y que las conclusiones obtenidas sean válidas y consistentes.

(D) *Variables Instrumentales (IV)*

El método de estimación de Variables Instrumentales (IV) es una técnica utilizada en situaciones de presencia de endogeneidad en las variables del modelo, ya que permite obtener estimaciones consistentes de los parámetros en modelos de regresión. Su funcionamiento consiste en el uso de variables instrumentales, entendidas como variables exógenas que, estando correlacionadas con las variables endógenas, no lo están con el término de error.

De la misma forma que el método OLS, la eficacia del método VI queda condicionada a una serie de condiciones fundamentales. Básicamente, se requiere que (i) *las variables instrumentales sean relevantes*, es decir, que estén correlacionadas con las variables endógenas, y que (ii) *las variables instrumentales cumplan con el supuesto de exogeneidad*, lo que implica que no deben estar correlacionadas con el término de error del modelo.

La estructura del método de IV, por sus características y peculiaridades, implica la construcción de un modelo en dos etapas. La primera etapa consiste en la estimación de un modelo auxiliar para las variables endógenas empleando las variables instrumentales como variables explicativas. A continuación, a partir de dicho modelo secundario, se extraen las estimaciones ajustadas de las variables endógenas.

Posteriormente, en una segunda etapa, se estima el modelo de regresión principal utilizando las predicciones ajustadas de las variables endógenas obtenidas en la primera etapa, junto con las variables explicativas restantes. A partir de estas estimaciones se obtienen los coeficientes de interés ajustados por el sesgo causado por la endogeneidad.

Continuando con la misma línea de argumentación, es vital considerar en todo momento las limitaciones del método IV a lo largo de la construcción del modelo. La gran dificultad que presenta esta metodología se trata de (i) *la disponibilidad de variables instrumentales adecuadas*; debido a que, en numerosas ocasiones, el hecho de disponer de variables idóneas puede resultar complicado al ser necesario un conocimiento sustancial del contexto económico y teórico. Por otra parte, se requiere (ii) *la validez de las variables instrumentales*, además de (iii) *la exclusión de otras formas de endogeneidad*.

En el supuesto de que el número de variables instrumentales sea limitado o las variables endógenas estén débilmente correlacionadas con las variables instrumentales, podemos afirmar que el método OLS sería más apropiado que el método IV para la obtención de estimaciones consistentes. En caso contrario, en el supuesto de presencia de endogeneidad en las variables del modelo con una fuerte correlación con las variables instrumentales, los resultados obtenidos con el método IV serán de mayor precisión.

En conclusión, el método de estimación de Variables Instrumentales (IV) aporta soluciones pragmáticas ye efectivas para abordar el problema de la endogeneidad en modelos econométricos. El uso de variables instrumentales proporciona una fuente de variación exógena que se traduce en estimaciones consistentes de los parámetros. No debemos olvidar que dicha metodología cuenta con limitaciones en términos de disponibilidad de variables instrumentales idóneas, su validez y el grado de precisión en comparación con el método de MCO.

(E) *Relación entre OLS e IV*

La relación entre los métodos de estimación de Mínimos Cuadrados Ordinarios (OLS) y Variables Instrumentales (IV) se basa en su uso complementario para abordar problemas específicos en modelos econométricos. Ambos métodos se utilizan para obtener estimaciones de los parámetros en modelos de regresión, pero cada método requiere de una serie de condiciones y circunstancias que lo harán más o menos apropiados en el caso concreto.

Por una parte, el método de MCO se utiliza bajo la asunción de que todas las variables explicativas son exógenas y no están correlacionadas con el término de error del modelo. Bajo esas circunstancias, proporciona estimaciones eficientes y lineales de los coeficientes de regresión. Sin embargo, ante la presencia de endogeneidad – correlación entre las variables explicativas y el término de error – las estimaciones obtenidas resultan sesgadas e inconsistentes.

Por otra parte, el método de IV se utiliza a través de variables instrumentales con el objetivo de mitigar los problemas de la endogeneidad y obtener estimaciones consistentes de los parámetros. Las variables instrumentales al estar correlacionadas con las variables endógenas, pero no con el término de error, permiten estructurar una fuente de variación exógena, a través de la cual se estiman los efectos causales de las variables explicativas en la variable dependiente.

La principal ventaja del método de IV sobre el método OLS es la capacidad para mitigar los efectos negativos provocados por la presencia de endogeneidad en el modelo y proporcionar estimaciones precisas, incluso en presencia de variables endógenas. En consecuencia, el método IV permite controlar el sesgo causado por la endogeneidad y obtener estimaciones válidas de los coeficientes de interés.

Sin embargo, el método de IV resulta más estricto a la hora de su condicionamiento a la presencia de una serie de factores, como la validez de las variables instrumentales y la exclusión de otras formas de endogeneidad. En contraste, el método de MCO no requiere la disponibilidad de variables instrumentales ni supuestos adicionales más allá de los supuestos clásicos de regresión lineal (*vid*. 2.1 Marco teórico, Sección (C)). En

consecuencia, por norma general, el método OLS resulta más fácil de implementar y proporciona estimaciones eficientes de los coeficientes en ausencia de endogeneidad, aunque un análisis previo adecuado sobre las características del modelo será en todo caso fundamental para determinar el método que se debe implementar.

En conclusión, los métodos de Mínimos Cuadrados Ordinarios y Variables Instrumentales presentan dos enfoques diferentes – y complementarios – para realizar estimaciones sobre los parámetros en modelos de regresión. Por una parte, el método MCO se utiliza bajo el contexto de variables explicativas exógenas y, por otra parte, el método IV ofrece una solución al problema de la endogeneidad. Ambos métodos tienen ventajas e inconvenientes diferentes y su elección dependerá del contexto y de los desafíos específicos del modelo econométrico en cuestión.

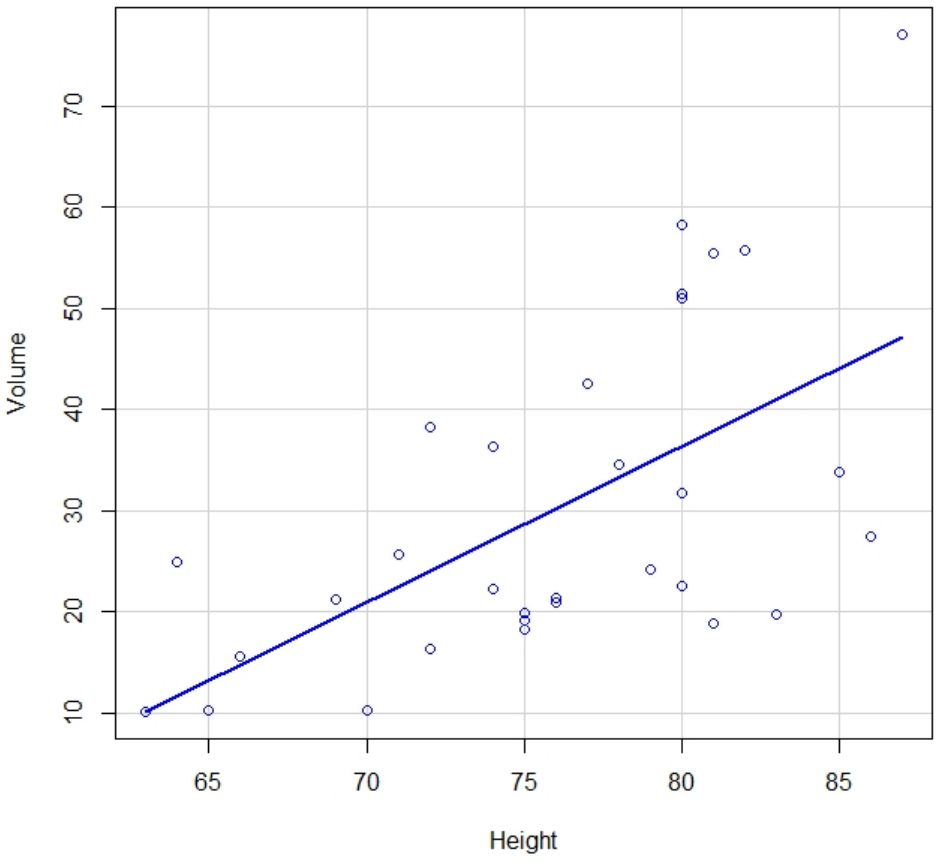

Imagen 1: Ejemplo gráfico de regresión lineal con el método de Mínimos Cuadrados Ordinarios (OLS)

Fuente: Hospital Universitario La Paz, Artículo de Manuel Molina, Junio 2020.

2.2 Estado actual del conocimiento

En el contexto de un modelo econométrico, donde existe una relación de causalidad entre las variables, existen numerosos métodos para conducir estimaciones. Como ha quedado expuesto, la determinación del método más adecuado para obtener los resultados más sólidos y precisos dependerá de las características de las variables que integran el modelo y el tipo de estimaciones que se desee realizar. Cada método de estimación ofrece una serie de ventajas diferentes y, de igual manera, presentan ciertos inconvenientes en la práctica que pueden traducirse en un resulta sesgado o inconsistente.

Por un lado, el método de Mínimos Cuadrados Ordinarios (OLS), resulta de gran utilidad y precisión bajo una serie de condiciones de las características de las variables. Dentro de estas condiciones podemos destacar: (i) ausencia de multicolinealidad – fuerte correlación entre varias variables independientes bajo un modelo de regresión; (ii) presencia de exogeneidad – falta de correlación entre una variable explicativa y el término de error del modelo; y (iii) ausencia de errores de especificación y linealidad entre las variables.

Por otro lado, el método de Variables Instrumentales (IV), requiere la coincidencia de una serie de supuestos para asegurar la consistencia de las estimaciones, entre las que podemos destacar: (i) relevancia de las variables instrumentales; (ii) exogeneidad de las variables instrumentales; y (iii) correlación entre las variables exógenas y a las variables endógenas, pero no con el término de error.

En ocasiones, la complejidad del modelo puede suponer escasa transparencia acerca de la información y características de las variables del mismo lo que puede dificultar el ejercicio de determinar que método de estimación resulta más adecuado. En este contexto, una vez que existe extensa doctrina acerca de las funcionalidades, ventajas e inconvenientes de cada uno de los métodos (*vid*. 2.3 Principales estudios relacionados), resulta lógica y necesaria la elaboración de un proceso sistemático que, de manera estandarizada, permita obtener la información necesaria para resolver la indeterminación acerca de la adecuación de los métodos para el caso concreto.

En conclusión, las aportaciones de este trabajo sobre el estudio previo de Ciacci (2021), en consonancia con el resto de los trabajos relacionados, resulta de gran utilidad.

2.3 Principales estudios relacionados

En este apartado se analizarán y expondrán de manera concisa las principales ideas extraídas de las lecturas utilizadas para la realización de este trabajo. En primer lugar, se comenzará tratando la lectura que supone la parte *core* de este trabajo. "A Matter of Size: Comparing IV and OLS estimates" de R. Ciacci, supone la base a partir de la cual se estructura y se desarrolla este trabajo.

En el estudio de R. Ciacci (2021), el extracto principal se podría resumir como, una vez analizados ambos métodos de estimación, teniendo en cuenta las ventajas e inconvenientes de cada uno, es fundamental saber elegir la técnica más adecuada según el tipo de datos que se vayan a utilizar en el modelo. En este escenario de comparación de métodos, la necesidad de ofrecer una solución estandarizada que pueda resolver la elección según la adecuación de cada método en el caso concreto resulta lógica. Por último, uno de los objetivos de este Trabajo de Fin de Grado consiste en, a partir de la ecuación multivariable de beta en función de delta (*vid. Ciacci, 2021*, *Appendix A, pág. 21*), expresarla en forma matricial y extraer una serie de conclusiones.

A continuación, se tratarán los trabajos de E. Oster, "Unobservable Selection and Coefficient Stability: Theory and Evidence", y L. Lochner y E. Moretti, "Estimating and Testing Models with Many Treatment Levels and Limited Instruments". Los mencionados estudios, junto al trabajo de R. Ciacci, componen la principal fuente de información sobre las que construye y se inspira este trabajo.

El trabajo de E. Oster (2016) resulta de vital importancia, al tratarse del antecedente de A Matter of Size, donde R. Ciacci lo cita y hace referencia al mismo de manera recurrente a lo largo de su trabajo. Oster utiliza la información obtenida a través de la regresión OLS (inclusión de controles, tamaño de varianzas y movimiento de R^2) a partir de la cual se estima un conjunto de valores donde debería situarse el verdadero efecto del tratamiento.

Mas allá de los tres trabajos principales, se analizarán los trabajos de D. Acemoglu, S. Johnson y J. A. Robinson (2001); M. A. Masten y A. Poirier (2020); S. O. Becker (2005); J. G. Altonji, T. E. Elder y C. R. Taber (2005); D. Staiger y J. H. Stock (1997); y G. De Luca, J. R. Magnus y F. Peracchi (2018).

(A) *A Matter of Size: Comparing IV and OLS estimates*

En primer lugar, comenzaremos resumiendo la obra principal sobre la que se construye este trabajo: "A Matter of Size: Comparing IV and OLS estimates" de Riccardo Ciacci, publicada en 2021. Este estudio se centra en realizar una comparativa sobre las estimaciones obtenidas a través de los dos métodos de estimación que se han expuesto a lo largo de este Trabajo de Fin de Grado, regresión a través de Mínimos Cuadrados Ordinarios (OLS) y de Variables Instrumentales (IV).

R. Ciacci introduce el estudio mediante una explicación a cerca de las diferencias entre estas dos técnicas y su uso en diferentes escenarios. La regresión OLS es ampliamente utilizada en estudios de carácter empírico, con el objetivo de analizar la relación entre una variable dependiente y varias variables independientes. Por otro lado, en escenarios de endogeneidad en una o más de las variables independientes, donde las variables están correlacionadas con la variable dependiente y con algún factor no observado, el autor defiende que es más preciso utilizar la regresión IV, en línea con el presente trabajo.

El objetivo principal del estudio reside en comparar las estimaciones obtenidas a través de ambos métodos en un conjunto de datos simulados, con diferentes tamaños de muestra y distintos niveles de endogeneidad. Los resultados que arroja el estudio muestran que la regresión IV es más precisa que la regresión OLS en la mayoría de los casos, especialmente cuando la endogeneidad es alta y la muestra es pequeña. Sin embargo, otra conclusión interesante es que la regresión IV también puede ser vulnerable a la presencia de errores de medición en las variables instrumentales.

En términos generales, la idea principal que se puede derivar del trabajo de R. Ciacci es la importancia de elegir el método de estimación adecuado teniendo en cuenta las circunstancias del modelo y los datos disponibles. El método de estimación IV puede ser una opción más apropiada y eficaz en casos de alta endogeneidad y muestras pequeñas. Por otra parte, el método IV requiere de unas circunstancias más estrictas, como la validez de las variables instrumentales o la exclusión de otras formas de endogeneidad, por lo que, en muchas ocasiones el método OLS resulta más práctico. En conclusión, es necesaria la estructuración de un proceso estandarizado que permita evaluar el nivel de adecuación de ambos métodos para un modelo concreto.

(B) *Unobservable Selection and Coefficient Stability: Theory and Evidence*

En segundo lugar, analizaremos el antecedente de la publicación de R. Ciacci, el trabajo "Unobservable Selection and Coefficient Stability: Theory and Evidence" de Emily Oster, publicado en 2016, al que hace referencia de manera recurrente a lo largo de su estudio. Este trabajo se centra en el problema de la selección no observada en modelos econométricos y cómo este factor puede afectar a la estabilidad de los coeficientes estimados en el tiempo.

El principal argumento que analiza Oster se trata del hecho de que la selección no observada puede en situaciones donde una variable relevante que influye en el resultado de interés no está disponible o no se mide correctamente. Desde una perspectiva práctica, esta circunstancia puede resultar problemática en el campo de estudios longitudinales, donde la selección en el tiempo puede afectar de manera negativa a los resultados.

La aportación principal del estudio consiste en proponer un marco teórico que sea capaz de analizar cómo la selección no observada influye en la estabilidad de los coeficientes estimados en el tiempo. Esta estructura se basa en el concepto de *límites de selección* que representan la distribución de las variables no observadas que afectan la selección en el tiempo. A lo largo del estudio se exponen una serie de casos prácticos aplicados.

Oster conduce una serie de análisis de aplicaciones reales para probar las aplicaciones de su trabajo. Uno de los estudios expuestos consiste en analizar los resultados de validación para el análisis del impacto del comportamiento materno sobre el peso al nacer y el cociente intelectual del niño. Los efectos de referencia incluyen sólo controles para el sexo del niño, variables ficticias de edad en el caso del cociente intelectual y semana de gestación en el caso del peso al nacer. Por último, muestra la supervivencia de los resultados no aleatorizados y aleatorizados bajo el ajuste de selección proporcional.

En conclusión, el trabajo destaca la importancia de considerar cuidadosamente la selección no observada en modelos econométricos y cómo semejante factor puede afectar a la estabilidad de los coeficientes estimados en el tiempo. Este trabajo supone una de las principales fuentes del estudio de R. Ciacci y, por ende, una de las bases de este trabajo.

(C) *Estimating and Testing Models with Many Treatment Levels and Limited Instruments*

El siguiente trabajo, "Estimating and Testing Models with Many Treatment Levels and Limited Instruments" de Lance Lochner y Enrico Moretti, publicado en 2014, consiste en la tercera pieza de la base sobre la que se construye este Trabajo de Fin de Grado. La obra de Lochner y Moretti se centra en el problema de estimar y probar modelos económicos que involucran muchos niveles de tratamiento y un número limitado de instrumentos.

Los autores se enfocan en el análisis empírico y utilizan una serie de casos prácticos para exponer sus estimaciones y ponderaciones mediante el método OLS. Los casos que se plantean son, por un lado, los efectos de la escolarización en la probabilidad de encarcelamiento, con una división entre varones blancos y varones negros, y, por otro lado, los efectos de la escolarización de la madre en la probabilidad de bajo peso del niño al nacer y en la probabilidad de parto prematuro.

El modelo incluye diferentes niveles de escolarización y se utiliza un conjunto limitado de instrumentos para abordar la posible endogeneidad entre la educación y el resto de las variables. Para abordar este problema, los autores proponen un enfoque de *regresión extendida* que permite la inclusión de múltiples niveles de tratamiento, de otras variables relevantes en el modelo y de un número limitado de instrumentos.

Los resultados del estudio demuestran que el hecho de tener en cuenta múltiples niveles de escolarización y utilizar un número limitado de instrumentos puede incrementar la precisión de las estimaciones y proporcionar información adicional sobre los efectos de la educación en (i) la probabilidad de encarcelamiento y (ii) la probabilidad de bajo peso al nacer o parto prematuro. Los autores también realizan pruebas de robustez para evaluar la validez de los resultados y encuentran que los resultados son consistentes en diferentes especificaciones del modelo.

El extracto fundamental del trabajo de Lochner y Moretti reside en destacar la importancia de utilizar técnicas estadísticas apropiadas para abordar los problemas de endogeneidad y como afecta el hecho de incluir múltiples niveles de tratamiento en modelos econométricos con un alto grado de complejidad.

(D) *The Colonial Origins of Comparative Development: An Empirical Investigation*

A continuación, analizaremos las fuentes de información secundarias del trabajo. Se tratan de obras cuyas lecturas han proporcionado diferentes perspectivas acerca de la causalidad entre variables en modelos econométricos, pero que no han sido utilizados como base para plantear la hipótesis de este Trabajo de Fin de Grado.

En primer lugar, el trabajo "The Colonial Origins of Comparative Development: An Empirical Investigation" de Daron Acemoglu, Simon Johnson y James A. Robinson, publicado en 2001 en la revista The American Economic Review, plantea el análisis de por qué algunos países son más ricos y prósperos que otros y que factores afectan en mayor medida a la situación actual de los países.

El factor principal que analizan los autores es los procesos de *colonización* y sus agentes. 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 otros países que fueron colonizados por potencias con instituciones extractivas, en las que el poder y los recursos fueron monopolizados por la élite colonial.

Los autores proporcionan evidencia empírica para apoyar su argumento utilizando datos de una amplia gama de países en todo el mundo. La conclusión que deriva de su estudio es que los países que fueron colonizados por potencias europeas con instituciones inclusivas cuentan con un mejor desempeño económico, mayor ingreso per cápita, mayor tasa de alfabetización y mayor esperanza de vida, en comparación con aquellos países que fueron colonizados por potencias con instituciones extractivas.

En adición, los autores argumentan que estas diferencias persisten hasta el día dehoy debido a la inercia institucional, donde las instituciones extractivas establecidas durante el período colonial son difíciles de erradicar. En general, el trabajo ha tenido un impacto significativo en la literatura sobre desarrollo económico, generando un gran interés en la importancia de las instituciones políticas y económicas, en la medida en la que afectan de manera determinante al crecimiento económico a largo plazo.

(E) *Salvaging Falsified Instrumental Variable Models*

En siguiente lugar, el trabajo "Salvaging Falsified Instrumental Variable Models" de Matthew A. Masten y Alexandre Poirier, publicado en 2020, se centra en el problema de los modelos que presentan variables instrumentales falsificados, proponiendo una solución plausible para recuperar la información válida de estos modelos.

Los modelos de Variables Instrumentales (IV), tratados de forma recurrente a lo largo de este Trabajo de Fin de Grado, suponen la medida más efectiva para establecer relaciones causales entre variables observables, pero a menudo se enfrentan a problemas de endogeneidad y selección de variables (*vid*. 2.1 Marco teórico, Sección (D)). En ocasiones particulares, los modelos pueden ser falsificados, lo que conlleva a que las variables instrumentales utilizadas no sean válidas y no proporcionen una identificación causal adecuada, derivando en estimaciones sesgadas e inconsistentes.

A lo largo de su estudio, los autores desarrollan un enfoque novedoso para recuperar la información válida de los modelos falsificados utilizando un modelo de ajuste a través de Mínimos Cuadrados Ordinarios (OLS). El modelo de ajuste consiste en que, en lugar de tratar de identificar la causa subyacente del problema de endogeneidad, se centra en maximizar la varianza explicada por la variable endógena utilizando una combinación de variables instrumentales falsificadas y variables exógenas.

A partir de los resultados y conclusiones obtenidos del trabajo, se puede afirmar que dicho enfoque puede ser efectivo para recuperar información válida de modelos falsificados. En su estudio particularmente, proporcionan pruebas empíricas utilizando datos de la *Encuesta Nacional de Examen de Salud y Nutrición* (NHANES) para evaluar la relación causal entre la obesidad y la presión arterial.

En conclusión, el trabajo proporciona una metodología innovadora, ofreciendo una solución para el problema de recuperar información válida de modelos de variables instrumentales falsificados. La principal ventaja de esta metodología resulta el hecho de poder aplicarse en situaciones donde no es factible identificar la causa subyacente del problema de endogeneidad.

(F) *Using instrumental variables to establish causality*

En el siguiente trabajo, "Using instrumental variables to establish causality" de Sascha O. Becker, publicado en la revista Journal of Economic Surveys en 2005, se analiza el uso de variables instrumentales en el contexto de modelos econométricos con el objetivo de determinar la causalidad en las relaciones económicas.

La línea de argumentación que presenta el autor gira en torno a que, en muchas situaciones de la vida real, no es posible realizar experimentos controlados para establecer la relación causal entre dos variables, por lo que se deben utilizar modelos econométricos basados en datos observados. Sin embargo, estos modelos pueden verse afectados por problemas de endogeneidad y sesgo de selección, lo que puede llevar a estimaciones incorrectas de la relación causal, proporcionando resultados inconsistentes.

El autor plantea un modelo a través del uso de variables instrumentales, siguiendo el método de estimación IV, para mitigar las imprecisiones fruto de la presencia de endogeneidad en la relación entre las variables del modelo. Al utilizar variables instrumentales adecuadas, con un grado de validez suficiente, se pueden obtener estimaciones más precisas de la relación causal entre las variables observables.

El autor proporciona varios ejemplos de cómo se han utilizado variables instrumentales en la literatura económica, incluyendo la relación entre la educación y el ingreso, y la relación entre la inversión en infraestructura y el crecimiento económico. Por otro lado, se discuten las complicaciones que pueden surgir al utilizar variables instrumentales, además de las limitaciones que presenta dicho método de estimación, como la presencia de otras formad de endogeneidad o la falta de validez de las variables instrumentales.

En términos generales, el trabajo de Becker hace referencia a la importancia del uso de variables instrumentales en la investigación económica, debido a las ventajas que presenta en modelos con presencia de endogeneidad, para poder determinar con precisión la causalidad entre variables observables y proporcionar indicaciones útiles para la selección y validación de los instrumentos que permitan obtener estimaciones insesgadas y conclusiones contundentes.

(G) *Selection on observed and unobserved variables: assessing the effectiveness of catholic schools*

A continuación, se examinará 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. Este estudio aborda las variaciones según la selección de variables observadas y no observadas en la evaluación del impacto que tiene el hecho de haber recibido educación en escuelas católicas en la vida de los estudiantes.

Los autores estructuran su hipótesis a partir de los datos obtenidos de la Encuesta Nacional de Educación de 1972, empelando una estrategia de variables instrumentales con el objetivo de controlar la posible selección de estudiantes en colegios católicos en función de una serie de variables no observadas, como habilidades no medidas o antecedentes familiares.

Esta metodología de variables instrumentales en concreto se basa en la utilización de una variable instrumental, en este caso la densidad de católicos en la zona de residencia del estudiante, con el objetivo estimar el grado de afectación causal que implica la educación católica en la probabilidad de graduarse en educación secundaria y de, posteriormente, matricularse en la universitaria.

Los autores deducen que los estudiantes que estudian en colegios católicos tienen una probabilidad significativamente mayor de graduarse en educación secundaria y de matricularse en la universidad, en comparación con los estudiantes que asisten a colegios públicos. Estos resultados demuestran un alto grado de consistencia incluso *a posteriori* de ejercer un control sobre la posible selección de variables no observadas.

En conclusión, el trabajo proporciona un ejemplo práctico de cómo tratar la problemática de la selección de variables observadas y no observadas en la evaluación del impacto del nivel de formación académica a través de variables instrumentales, destacando la importancia de controlar adecuadamente los factores que influyen en la selección de estudiantes según el tipo de colegio.

(H) *Instrumental variables regression with weak instruments*

El siguiente trabajo de nuestra bibliografía, se trata del estudio "Instrumental variables regression with weak instruments" de Douglas Staiger y James H. Stock, publicado en la revista *Econometrica* en 1997. Los autores indagan sobre la problemática de la estimación de modelos de regresión con variables instrumentales débiles y como su falta de validez conllevan estimaciones sesgadas e inconsistentes.

El concepto de *variables instrumentales débiles* se define, atendiendo a Staiger y Stock, como aquellas variables que tienen una correlación baja con las variables endógenas en la regresión. Este factor limita la capacidad para eliminar el sesgo de endogeneidad, provocando que las variables instrumentales débiles puedan ser particularmente problemáticas en supuestos donde la muestra es pequeña, reduciendo aún más la capacidad de las variables instrumentales para mitigar la endogeneidad.

Staiger y Stock plantean un clásico método de estimación de regresión a través de variables instrumentales, pero que utiliza un enfoque de máxima verosimilitud y una corrección de la matriz de varianzas y covarianzas para tener en cuenta la debilidad de las variables instrumentales. Como aportación fundamental, se proporciona un estadístico de prueba que permite evaluar la validez de las variables instrumentales y, por ende, analizar la potencial precisión de las estimaciones.

Con el objetivo de testar su hipótesis en el plano práctico, los autores aplican su método a un ejemplo de regresión de salarios y determinan que el enfoque de máxima verosimilitud con corrección de la matriz de varianzas y covarianzas es más efectivo para tratar el problema de variables instrumentales débiles que lo que pueden llegar a serlo otros métodos de estimación diferentes.

En términos generales, el trabajo de Staiger y Stock proporciona una solución novedosa para contrarrestar las dificultades que presentan las variables instrumentales débiles en la regresión de variables instrumentales y se destaca la importancia de evaluar *ex ante* la validez de las variables instrumentales en la estimación de modelos de regresión para poder determinar el grado de adecuación del método empleado.

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

Por último, se tratará el trabajo “Comments on 'Unobservable Selection and Coefficient Stability: Theory and Evidence' and 'Poorly Measured Confounders are More Useful on the Left Than on the Right'” de Giuseppe De Luca, Jan R. Magnus y Franco Peracchi, publicado en 2018. Este trabajo continúa con la doctrina académica planteada en dos trabajos anteriores en materia de la problemática de la selección no observada y los *confusores* – variables que influyen, tanto en la variable dependiente, como en las variables dependiente generando una causalidad errónea – mal medidos en modelos de carácter econométrico.

En primer lugar, los autores comentan el trabajo de Emily Oster "Unobservable Selection and Coefficient Stability: Theory and Evidence", en el que, como ya se ha analizado *supra*, se desarrolla el impacto de la selección no observada en la estimación de modelos de regresión. Los autores discuten el marco teórico utilizado por Oster y presentan una crítica sobre su enfoque empírico, argumentando que la selección no observada puede ser menos relevante en algunas situaciones y que es necesario un análisis cuidadoso de las variables para determinar su adecuación al 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. Los autores analizan la validez de los supuestos utilizados por Heckman y Vytlacil, llegando a la conclusión de 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 conclusión, el trabajo de De Luca, Magnus y Peracchi plantea una crítica detallada y fundamentada de los enfoques utilizados en los dos trabajos anteriores que se utilizan como base del estudio atendiendo a la selección no observada y a los *confusores* mal medidos, destacando la importancia de un análisis cuidadoso y preventivo de las variables que permita evaluar su compatibilidad con el modelo en cuestión.

2.4 Relación de aportaciones

En este apartado, resulta pertinente comenzar estableciendo la relación que existe entre los dos principales ensayos que se han empleado como objeto de análisis y de estudio para la realización de este Trabajo de Fin de Grado: *A Matter of Size: Comparing IV and OLS estimates* (R. Ciacci) y *Unobservable Selection and Coefficient Stability: Theory and Evidence* (E. Oster).

En el ensayo del R. Ciacci se analiza el concepto de endogeneidad, mientras que en el de la autora E. Oster, el concepto de la selección no observada. El punto principal de conexión entre ambos trabajos reside en que ambos conceptos se tratan de fenómenos que pueden afectar a la estimación de los parámetros y conducir a resultados sesgados o inconsistentes en modelos econométricos.

La endogeneidad hace referencia a la relación entre una variable explicativa y el término de error del modelo. Tiene lugar cuando una variable explicativa está correlacionada con el término de error, lo que implica que no se puede establecer una relación causal precisa entre la variable explicativa y la variable dependiente. Esto puede suceder debido a la existencia de variables omitidas o a la presencia de simultaneidad entre las variables.

La selección no observada, por otro lado, se refiere a la falta de información sobre ciertas variables relevantes que afectan tanto a la variable dependiente como a las variables explicativas. Estas variables no observadas pueden suponer imprecisiones en la estimación de los parámetros del modelo.

La relación entre la endogeneidad y la selección no observada radica en el hecho de que ambas pueden dar lugar a resultados sesgados e inconsistentes en los modelos econométricos. La endogeneidad puede ser causada por la presencia de variables omitidas que afectan tanto a la variable dependiente como a las variables explicativas, lo que lleva a una correlación entre la variable explicativa y el término de error. Esto a su vez puede generar problemas en la estimación de los parámetros. La selección no observada puede surgir cuando hay variables relevantes que no se incluyen en el modelo y que afectan tanto a la variable dependiente como a las variables explicativas.

En conclusión, las variables no observadas influyen de manera determinante en los resultados y, de no tenerlas en cuenta, la estimación de los parámetros puede devenir sesgada. Por otro lado, el ensayo *Estimating and Testing Models with Many Treatment Levels and Limited Instruments* (L. Lochner y E. Moretti) se centra en la aplicación de métodos econométricos para estimar y probar modelos con múltiples niveles de tratamiento y un número limitado de instrumentos. Esta investigación tiene implicaciones tanto para la regresión lineal ordinaria (OLS) como para la regresión instrumental (IV). En el trabajo se propone una extensión de la regresión lineal ordinaria para abordar el problema de los múltiples niveles de tratamiento. El OLS asume que las variables independientes están linealmente relacionadas con la variable dependiente y que no hay errores de medición en las variables independientes. Sin embargo, cuando se tienen múltiples niveles de tratamiento, es posible que estas suposiciones no se cumplan y que los resultados del OLS sean sesgados e inconsistentes.

Lochner y Moretti introducen un nuevo estimador, llamado *Estimador de Mínimos Cuadrados con Instrumentos Ampliados* (ACLS, por sus siglas en inglés), que incorpora instrumentos adicionales para controlar los efectos no observados y eliminar el sesgo de selección. El ACLS permite estimar los efectos causales de los diferentes niveles de tratamiento de manera más precisa y consistente. De igual manera, se pueden identificar en el trabajo implicaciones para la regresión de variables instrumentales (IV). Como ya se ha mencionado *supra*, la regresión instrumental es un método utilizado para abordar el problema de la endogeneidad, es decir, cuando una variable independiente está correlacionada con el término de error del modelo. En el contexto de modelos con múltiples niveles de tratamiento, puede ser difícil encontrar instrumentos válidos que cumplan los supuestos necesarios para llevar a cabo una regresión instrumental.

Partiendo de dicho escenario, los autores tratan de superar estas limitaciones al introducir un enfoque que permite la estimación y prueba de modelos con muchos niveles de tratamiento y un número limitado de instrumentos válidos. Su enfoque se basa en la combinación de instrumentos múltiples y débiles, de tal manera que aumenten la precisión y consistencia de las estimaciones. De esta forma, se establece una secuencia y un significado en el orden de los ensayos empleados como objeto de análisis y estudio, principalmente, entre las tres primeras lecturas (*vid*. Bibliografía), sobre las que se ha construido la estructura principal de la tesis de este Trabajo de Fin de Grado.

3. DE ECUACIÓN A MATRIZ

3.1 Análisis crítico del artículo de R. Ciacci (2021)

El punto de partida de nuestro estudio comienza la *Función de Regresión de la Población* (PRF, por sus siglas en inglés) que se plantea al inicio del estudio de la obra de R. Ciacci (2021). Comenzaremos por definir cada una de las variables que componen la ecuación:

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

Dentro de cada variable sus respectivos subíndices hacen referencia a la unidad (***i***) y al tiempo (***h***). Por otra parte, ***d*** es el escalar que supone el tratamiento de nuestro interés, ***w*** es el vector de los controles no observables y, por último, ***X*** es el vector que contiene los controles observados. Debido a la naturaleza del vector de los controles no observables, ***w*** deberá ser omitido de la ecuación (*Ciacci, 2021*).

En un supuesto univariante, el objetivo de la regresión reside en averiguar la recta ***y***. En este caso, el valor de ***alfa sub-1*** representa el punto de corte con el eje, mientras que **beta sub-1** representa la pendiente de la recta y, por último, ***X*** representa la variable independiente. De esta manera, logramos obtener una recta a partir de la solución. Como se ha venido desarrollando a lo largo del trabajo, un modelo de regresión presenta numerosas funcionalidades aplicables a diferentes campos, por ejemplo, la predicción de tendencias de carácter macroeconómico.

En el supuesto de tomar por válidas y partir de las asunciones expuestas por E. Oster en su trabajo, podemos plantear la relación de selección proporcional en línea con nuestro planteamiento:

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

Esta ecuación se sostiene y se cumple para cualquier supuesto en el que ***delta*** sea distinto de 0. Sobre la estructura univariante, partiendo de la base de que ***w*** es el único control y es omitido de la ecuación de regresión, deducimos el sesgo de la variable omitida:

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

En el caso de extrapolar las premisas planteadas en este apartado a un supuesto de regresión lineal múltiple, definiendo en valor residual de una regresión de ***d*** en ***X*** como $\tilde{\boldsymbol{d}}_{\boldsymbol{ih}} = \boldsymbol{d} - \boldsymbol{\tau} + \boldsymbol{\tau X}$, el sesgo de la variable omitida quedaría estructurado de la siguiente forma:

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

Teniendo en cuenta la línea del planteamiento propuesto hasta el momento, asumiendo una relación ortogonal entre las variables observables y las variables no observables, tal y como propone Oster en su obra, obtenemos el llamado *coeficiente de proporcionalidad*:

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

En el supuesto, bajo la premisa de que el instrumento sea válido, el coeficiente obtenido a través de variables instrumentales es capaz de estimar de manera consistente el valor de ***beta sub-1***. Por lo tanto, se puede calcular el tamaño óptimo del *coeficiente de proporcionalidad*, denominado como ***delta***, para soportar la diferencia de tamaño entre la estimación mediante Mínimos Cuadrados Ordinarios (OLS), ***beta sub-2***, y la estimación mediante Variables Instrumentales (IV) para el verdadero efecto, ***beta sub-1***.

El *coeficiente de proporcionalidad* (***δ***) hace referencia a la selección de variables no observables pertinentes que sean capaces de respaldar las estimaciones obtenidas a través de la metodología IV en un supuesto de carácter empírico.

A medida que incrementa el valor de ***delta***, representa que aumenta la dependencia de variables no observables que justifican que el verdadero efecto coincide con el efecto estimado como resultado del modelo de regresión a través del método de estimación IV. En consecuencia, a medida que disminuye el valor de ***delta***, se cumple en mayor medida

la condición de validez de la variable instrumental, mencionada *supra* como una de las condiciones fundamentales para obtener estimaciones precisas y consistentes.

En conclusión, el tamaño del valor que adopte ***delta*** hace referencia a la proporción de selección de variables no observables con respecto de las variables observables que resulta imperativa para que el verdadero efecto guarde relación con el valor de las estimaciones obtenidas mediante la metodología IV. Un valor alto de ***delta*** podría significar la falta de validez de la variable instrumental o la presencia de otra clase de endogeneidad que conllevaría la inconsistencia de las estimaciones obtenidas.

Continuando con las aportaciones de Oster en su ensayo, a la hora de conducir este análisis es fundamental tener en cuenta una serie de factores adicionales como la inclusión de controles, el tamaño de las varianzas y el movimiento de ***R cuadrado*** (*Ciacci, 2021*). En un supuesto empírico, el signo de ***delta*** adopta una gran relevancia. En la última ecuación planteada *supra*, el signo de ***delta*** únicamente depende de la asunción que se haga sobre el signo de $\boldsymbol{\gamma}$. El racional detrás de esta afirmación reside en que las varianzas siempre son positivas, lo que nos permite estimar a través de los datos el resto de los elementos de la ecuación.

En consecuencia, conociendo el signo de la resta entre ***beta sub-2*** y ***beta sub-1***, así como el de la covarianza del escalar ***d*** y el vector de controles no observados ***X***, se puede identificar el signo de delta en función del signo que estimemos que ostenta la variable omitida en la ecuación de regresión principal (*Ciacci, 2021*). En este caso, conocer el signo de ***delta*** nos permite calcular el conjunto identificado para los *coeficientes de proporcionalidad* que compartan dicho signo y determinar la variación de los límites del conjunto a medida que se modifica el *coeficiente de proporcionalidad*.

En adición, resulta de gran relevancia la selección del valor que ha de adoptar el estadístico ***R cuadrado*** a la hora de estimar el conjunto identificado en un modelo de regresión con variables observables y variables no observables (regresión completa). Haciendo referencia de nuevo al ensayo de Oster, dicho estadístico adopta el nombre de ***R max***. Su elección se fija en función del conocimiento *ex ante* que se tenga del entorno. Sobre la premisa de que la regresión completa pueda explicar en su totalidad la variable de resultado, el estadístico adoptaría el valor 1.

3.2 Beta en forma matricial

En este apartado, se desarrollará la aportación principal de este Trabajo de Fin de Grado. La bibliografía utilizada como precedente pone de manifiesto la importancia de la necesidad de estructurar un proceso estandarizado que permita evaluar que método de estimación resulta más apropiado atendiendo a las circunstancias del modelo, así como los principales avances en esta materia.

A lo largo del análisis de la doctrina académica, se puede observar que en numerosas ocasiones no se realiza un ejercicio apropiado acerca de la adecuación del método de estimación empleado, conduciendo muchas veces a resultados inconsistentes. En conclusión, la aportación de este trabajo no consiste en solucionar dicha problemática, sino retomar el estado actual de la cuestión y proponer una forma diferente de testarlo.

A continuación, partiendo de la última ecuación expuesta en el análisis del apartado anterior, que hace referencia a la ecuación del *Appendix A* del supuesto multivariante del trabajo de R. Ciacci (2021), construiremos una matriz que permita simplificar el proceso y que contribuya al propósito para el que se plantea este trabajo, que no es otro que reivindicar la importancia de que si las estimaciones de un modelo no están bien planteadas, nunca se podrán concluir resultados representativos y contundentes.

El punto de partida se trata de la mencionada *Función de Regresión de la Población* (PRF). En primer lugar, se deben determinar *ex ante* las dimensiones de la nueva ecuación que se pretende construir:

$\boldsymbol{Y\ (n\ x\ 1)}$	Matriz de la variable dependiente que cuenta con ***n*** observaciones;
$\boldsymbol{D\ (n\ x\ 1)}$	Matriz del escalar que supone el tratamiento de nuestro interés;
$\boldsymbol{W\ (n\ x\ 1)}$	Matriz de controles no observados;
$\boldsymbol{X\ (n\ x\ k)}$	Matriz de controles observados, ***n*** observaciones, ***k*** variables observables;
$\boldsymbol{\beta_1, \beta_2\ (1\ x\ 1)}$	Coeficientes escalares del tratamiento;
$\boldsymbol{\theta_1, \theta_2\ (k\ x\ 1)}$	Vectores de coeficientes de los controles observados; y
$\boldsymbol{\gamma\ (1\ x\ 1)}$	Coeficiente escalar del control no observado.

La ecuación formulada en su forma matricial adoptaría la siguiente forma:

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

La ecuación en su modalidad simplificada seguiría la siguiente estructura:

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

Atendiendo a la naturaleza de las variables omitidas, solo resulta factible ejecutar la regresión a partir de la siguiente ecuación:

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

El siguiente paso radica en fijar las dimensiones de las matrices de varianza y covarianza:

$\boldsymbol{\Sigma_{DX}}$ ***(1 x k)***	Matriz de covarianza cruzada entre ***D*** y ***X***;
$\boldsymbol{\Sigma_{DW}}$ ***(1 x 1)***	Matriz de covarianza entre ***D*** y ***W***;
$\boldsymbol{\Sigma_{D}}$ ***(1 x 1)***	Matriz de varianza de ***D***;
$\boldsymbol{\Sigma_{X}}$ ***(k x k)***	Matriz de varianza-covarianza de ***X***; y
$\boldsymbol{\Sigma_{W}}$ ***(1 x 1)***	Matriz de varianza de ***W***.

Este planteamiento consiste en asignar un *coeficiente de proporcionalidad* – ***delta*** – a cada una de las variables de control de tal manera que se pueda cumplir la ecuación de proporcionalidad con la variable omitida. En consecuencia, la ecuación de proporción de selección quedaría estructurada de la siguiente forma, en adelante, *Ecuación de Proporción de Selección*:

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

En este supuesto, $\boldsymbol{X_j}$ equivaldría al vector de la variable en la columna ***j*** de la matriz de variables observables en ***X***, que estaría compuesto por ***n*** observaciones. En forma matricial quedaría estructurada de la siguiente forma:

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

Prosiguiendo con el planteamiento, ***δ (1 x k)*** representaría la matriz que contiene la totalidad de los coeficientes de proporcionalidad. El número total de coeficientes de proporcionalidad equivale al número total de variables observables del modelo.

Partiendo de la Ecuación que representa el sesgo de la variable omitida (*vid*. 3.1 Análisis crítico del artículo de R. Ciacci (2021), ecuación núm. 4), en adelante, *Ecuación del Sesgo de la Variable Omitida*:

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

Beneficiándonos de las asunciones planteadas por E. Oster, tomando en consideración el hecho de haber validado la premisa de que las variables observables y las variables no observables guardan una relación ortogonal, $Cov(X_{ih}, w_{ih}) = 0$.

En siguiente lugar, ***d̃*** hace referencia al valor residual de una regresión de ***d*** en ***X*** de tal manera que $d_{ih} = \tau_0 + \tau_1 X_{ih} + \tilde{d}_{ih}$. A continuación, la ecuación de regresión de ***d*** en ***X*** representada en forma matricial quedaría estructurada de la siguiente forma:

$$D = X\mathrm{T}_1 + \widetilde{D}$$

En este supuesto, partiendo de ***D*** [de dimensiones ***(n x 1)***] y ***X*** [de dimensiones ***(n x k)***], $\widetilde{\boldsymbol{D}}$ es de dimensiones ***(n x 1)***, mientras que el vector de coeficientes $\boldsymbol{T_1}$ es de dimensiones ***(k x 1)***. En siguiente paso consiste en estructurar la *Ecuación del Sesgo de la Variable Omitida* en su forma matricial, por ende, resulta necesario definir *ex ante* las siguientes matrices:

$\boldsymbol{\Sigma_{\widetilde{D}}}$ ***(1 x 1)***	Matriz de la varianza de $\widetilde{\boldsymbol{D}}$; y
$\boldsymbol{\Sigma_{XW}}$ ***(k x 1)***	Matriz de la covarianza cruzada entre ***X*** y ***W***.

La segunda matriz hace referencia al vector que contiene las covarianzas entre las variables observables y la totalidad de las variables omitidas. En consecuencia, la *Ecuación del Sesgo de la Variable Omitida* queda estructurada en forma matricial de la siguiente forma:

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

Siguiendo la línea de argumentación de E. Oster, partiendo de la presunción de que las variables omitidas guardan una relación de carácter ortogonal con las variables de control, es decir, que no existe ningún tipo de correlación o interferencia alguna entre las mismas, implica que sus varianzas toman el valor de 0 y, por ende, la forma matricial simplificada de la *Ecuación del Sesgo de la Variable Omitida* queda expresada de la siguiente manera:

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

En última instancia, si consolidamos la *Ecuación de Proporción de Selección* en su forma matricial con la *Ecuación del Sesgo de la Variable Omitida* en su forma matricial simplificada y despejamos para δ_j, finalmente logramos obtener la representación matricial de la ecuación de partida, es decir, la representación matricial de la ecuación multivariante de beta en función de delta del trabajo de R. Ciacci (2021), Appendix A:

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

3.3 Explicación concisa de la metodología propuesta

En resumen, el procedimiento planteado a lo largo de este Apartado 3, trata de realizar una contribución a la doctrina prexistente en materia de análisis de la adecuación de los métodos de estimación en el campo de los modelos de carácter econométrico, partiendo de los pilares asentados por R. Ciacci y E. Oster en sus trabajos.

El proceso ha consistido en analizar el planteamiento de la función multivariante de beta en función de delta, de tal manera que, a partir de una serie de asunciones, sea posible logar su expresión en forma matricial aportando una perspectiva distinta y novedosa que se construya una visión más estructurada del modelo pertinente.

4. CONCLUSIONES

El Trabajo de Fin de Grado se ha fundamentado en exponer de manera clara y objetiva las distintas funcionalidades que ofrecen los métodos de estimación de Mínimos Cuadrados Ordinarios (OLS) y de Variables Instrumentales (IV), así como sus principales ventajes e inconvenientes, y aquellos supuestos que, debido a las circunstancias particulares del modelo, provocan que uno resulte más apropiado y preciso que el otro.

El extracto principal que pretendo demostrar al lector no debe enfocarse como concluir si un método resulta mejor que el otro, sino que lo fundamental reside en realizar un análisis apropiado de las peculiaridades de las variables que conforman el modelo para elegir el método óptimo que nos conduzca a unas estimaciones de mayor precisión y consistencia.

En cuanto a la principal aportación del trabajo, la formulación de la ecuación delta en función de la matriz beta en una regresión multivariante representada de manera matricial, ofrece un nuevo enfoque más práctico, estructurado y manejable que su ecuación base. A continuación, se expondrán de forma clara y concisa las principales cualidades a su favor.

En primer lugar, la representación matricial de una ecuación siempre va a aportar una perspectiva más compacta, de tal manera que se observen mejor las relaciones entre las variables involucradas en el modelo. En un supuesto como el nuestro, donde se trabaja una ecuación multivariante, la relación de causalidad entre la variable dependiente y las variables explicativas resulta más identificable.

En segundo lugar, desde una perspectiva operativa, las matrices resultan tener un mayor grado de interoperabilidad, al beneficiarse de ciertas propiedades algebraicas. Dichas propiedades pueden proporcionar información valiosa acerca de la simetría, la dominancia, la estabilidad o la singularidad, entre otros factores determinantes.

En tercer lugar, la simplificación de la notación que conlleva la expresión matricial nos facilita el tratamiento de información simultánea, por ejemplo, de varias estimaciones y observaciones. El agrupamiento de los coeficientes de regresión permite conducir un análisis con mayor exhaustividad, por lo que las características del modelo resultan apreciables, facilitando la tarea de determinación de la precisión de los resultados.

En resumen, el planteamiento en forma matricial de la ecuación conlleva una serie de ventajas implícitas que permiten simplificar la determinación de la inferencia causal en modelos econométricos. El uso de la regresión está ampliamente extendido en una gran variedad de campos. Sobre esta base, la notación matricial permite la aplicación de métodos y técnicas propias de cada disciplina, simplificando la modelización y el análisis de problemas complejos. En definitiva, para asegurar la precisión de las estimaciones derivadas de un modelo, la notación matricial contribuye a mitigar los efectos negativos de ciertos factores, por ejemplo, la endogeneidad o la falta de validez de las variables instrumentales. Constituye una estructura simple, compacta y operable que facilita el planteamiento general sobre el que se ha construido este trabajo, defendiendo la premisa de que cuanto mayor sea el grado de conocimiento sobre las características de un modelo y las variables que lo componen, más fácil resulta la aplicación de las técnicas necesarias para llegar a las estimaciones óptimas que aborden unos resultados de mayor consistencia.

5. BIBLIOGRAFÍA

Ciacci, R. M., "A Matter of Size: Comparing IV and OLS estimates", *Universidad Pontificia de Comillas, Department of Economics and Business Management*, 20 de mayo de 2021.

Oster, E., "Unobservable Selection and Coefficient Stability: Theory and Evidence", *Brown University and NBER*, 9 de agosto de 2016.

Lochner, L. y Moretti, E., "Estimating and Testing Models with many Treatment Levels and Limited Instruments", *Universities of Western Ontario and California-Berkely*, 14 de junio de 2014.

Acemoglu, D.; Johnson, S.; y Robinson, J. A.; "The Colonial Origins of Comparative Development: An Empirical Investigation", *The American Economic Review, Vol. 91, No. 5, pp. 1369-1401* (Dec. 2001).

Masten, M. A. y Poirier, A., "Salvaging Falsified Instrumental Variable Models", *Universities of Duke and Georgetown*, 7 de enero de 2020.

Becker, S. O., "Using instrumental variables to establish causality", Universities of Warwick (UK) and IZA (Germany), abril de 2016.

Altonji, J. G.; Elder, T. E.; Taber, C. R.; "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools", *National Bureau of Economic Research, Working Paper 7831*, agosto de 2000.

Staiger, D. y Stock, J. H., "Instrumental Variables Regression With Weak Instruments", *National Bureau of Economic Research, Working Paper 151*, enero de 1994.

De Luca, G.; Magnus, J. R.; Peracchi, F.; "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*, 12 de septiembre de 2018.