Fecha de recepción: 2 de abril de 2020.
Fecha de aprobación: 5 de junio de 2020.

doi: https://doi.org/10.35997/runacv13n24a9

Resolviendo endogeneidad sin instrumentos. Una aplicación desde Lewbel

Solving endogeneity without instruments. An application from Lewbel

Uvenny Quirama Estrada[1] - Paula Andrea Forero Delgadillo[2]
Diego Fernando Montañez Herrera[3] - Diana Marcela Mena Serna[4]
Henner Andrés Solarte[5]

Resumen

Hasta el día de hoy, se ha venido presentando una motivación académica importante por comprender las relaciones causales para hacer inferencia estadística robusta. No obstante, es muy común enfrentarse a problemas de endogeneidad y en la práctica encontrar instrumentos suele ser complejo. Aunque la endogeneidad puede deberse a diferentes mecanismos de interacción y relaciones entre los regresores y las variables respuesta con el error, el efecto que causa es la inconsistencia en la estimación, lo que significa que los resultados no están respondiendo de forma adecuada a resolver el problema propuesto. Parte de allí entonces, la motivación de investigaciones y metodologías que buscan corregir la endogeneidad sin recurrir al uso de instrumentos externos. En este documento de trabajo, se parte por considerar que la inconsistencia se debe a un error de medida en la variable endógena, y se exploran mecanismos de solución para corregir el sesgo. Dada la naturaleza de los datos la metodología elegida es la propuesta por Lewbel (1997). Finalmente, se realiza una aplicación a un ejercicio empírico empleando las bases de datos provistas por Stock and Watson (2007), concluyendo con la solución a los problemas de endogeneidad y la diferencia existente entre la solución encontrada en la investigación y otros estudios econométricos.

Palabras clave: Endogeneidad, econometría, análisis comparativo, variables instrumentales. **JEL: C13, C26, C36**.

Abstract

To this day, there has been an important academic motivation for understanding causal relationships to make robust statistical inference. However, it is very common to face endogeneity problems and in practice finding instruments is usually complex. Although endogeneity may be due to different interaction mechanisms and relationships between the regressors and the error response variables, the effect it causes is the inconsistency in the estimation, which means that the results are not responding adequately to solving the problem. proposed. Hence, the motivation for research and methodologies that seek to correct endogeneity without resorting to the use of external instruments. In this working paper, we start by considering that the inconsistency is due to a measurement error in the endogenous variable, and solution mechanisms are explored to correct the bias. Given the nature of the data, the chosen methodology is that proposed by Lewbel (1997). Finally, an applica-

[1] Magister en Administración, Corporación Universitaria Americana, Docente/ Business Intelligence /Ciencias Económicas y Administrativas, Medellín, Colombia. Corporación Universitaria Americana – uquirama@amerciana.edu.co. ORCID: 0000-0002-8930-0450
[2] Economista y Estudiante de la Maestría en Economía, Asistente de investigación/Escuela de economía y finanzas, Medellín, Colombia. Universidad EAFIT - paforerod@eafit.edu.co. ORCID: 0000-0001-8461-7873
[3] Economista, Integrante/ Grupo de Estudios en Economía y Empresa/Escuela de Economía y Finanzas, Medellín, Colombia. Universidad EAFIT – dfmontaneh@eafit.edu.co, ORCID: 0000-0003-2326-7405
[4] Economista, Universidad Antioquia, Asistente de Investigación Centro de Investigaciones Económicas y Financieras / Escuela de Economía y Finanzas, Medellín, Colombia. Universidad EAFIT - dmmenas@eafit.edu.co, ORCID: 0000-0002-6246-7189
[5] Economista, Asistente de Investigación/Escuela de Administración, Medellín, Colombia. Universidad EAFIT – hmosque1@eafit.edu.co, ORCID: 0000-0001-5576-6554

tion is made to an empirical exercise using the databases provided by Stock and Watson (2007), concluding with the solution to endogeneity problems and the difference between the solution found in the research and other econometric studies.

Keywords: Endogeneity, econometrics, comparative analysis, instrumental variables. **JEL: C13, C26, C36**.

Introducción

Dentro de los supuestos del conocido modelo de regresión por Mínimos Cuadrados Ordinarios-MCO, uno de los más fuertes es la exogeneidad de las variables independientes, lo que implica que estas no contengan información relevante para la predicción de los errores. Cuando esta afirmación no se cumple, los estimadores son sesgados e inconsistentes.

Este problema está latente en distintas aplicaciones empíricas como variables ficticias endógenas (efectos tratamiento), modelado económico, financiero y de marketing, modelos dinámicos, por mencionar algunos. En general, se presenta en situaciones donde existen variables que no son observadas y/o que tienen omisión de variables.

Diversos han sido los métodos expuestos para la resolución del problema de endogeneidad, el más trabajado, es el asociado a variables instrumentales, el cual consiste en el uso de una variable alternativa que está relacionada con otra que genera el problema de endogeneidad pero que no está correlacionada con las perturbaciones.

No obstante, en algunas ocasiones se presenta inconsistencia, cuando los instrumentos poseen una correlación con el termino error en la ecuación de interés, o cuando el problema está relacionado con la elección de instrumentos "débiles" que son predictores "pobres" de la variable endógena en la estimación por primera etapa. Por consiguiente, la dificultad de encontrar instrumentos adecuados ha despertado el interés de los investigadores por encontrar formas de explicar la endogeneidad en los datos de observación sin necesidad de usar instrumentos observados. (Papies, Ebbes & Heerde, 2017).

De esta manera, el siguiente trabajo pretende dar respuesta al interrogante: ¿Cuál es el mejor método alternativo para la solución de endogeneidad en los modelos econométricos que no implique el uso de variables instrumentales? Esto, a partir de la implementación de una metodología alternativa, que brinde una solución robusta, práctica y estadísticamente significativa al problema de investigación, con el desarrollo y la implementación de modelos seleccionados de la búsqueda de la literatura basados en datos estadísticos que se exponen en el documento.

Este artículo consta de seis secciones adicional a esta. En primera instancia, se presenta una revisión sistemática de la literatura de cómo se ha intentado resolver la endogeneidad sin instrumentos, para continuar con un desarrollo metodológico del modelo Higher Moments y mostrando su aplicación comparando con estimación por instrumentos usando una base de datos específica. Posteriormente, teniendo en cuenta que los problemas de endogeneidad son una amenaza para inferir efectos causales (Papies et al., 2017) se presenta una aplicación adicional que resuelve el problema de endogeneidad de forma estándar con variables instrumentales, bajo su enfoque más habitual de mínimos cuadrados en dos etapas (2SLS), que se puede calcular en dos pasos simples y se compara con nuestra propuesta metodológica alternativa, teniendo como base los datos de Stock y Watson (2007), sobre dos casos específicos: el rendimiento en las pruebas en colegios de California y el consumo de cigarrillo en los 48 estados contiguos de Estados Unidos. En seguida, se presentan los resultados y se expone una breve sección sobre la discusión para finalizar con las conclusiones.

Revisión de la Literatura

La dificultad de encontrar instrumentos adecuados ha despertado el interés de los investigadores por identificar formas de explicar la endogeneidad en los datos de observación sin necesidad de usar instru-

mentos observados. En este sentido, dentro de la literatura se han explorado diversas técnicas para contribuir a la solución de endogeneidad en distintas aplicaciones. Un punto de partida, consiste en analizar las causas que ocasionan dicha endogeneidad, con el fin de imaginar las diferentes alternativas que pueden corregirlas. Estas causas se dividen en dos: error de medida y error de especificación.

El error de medida está relacionado con la medición imprecisa de una de las variables independientes, generando un ruido de medición, Carroll (2016) describe que las medidas de errores en las covariables tienen tres efectos: i) causa sesgo en los parámetros, ii) conduce una pérdida de poder para detectar relaciones entre variables y iii) encubre características de los datos lo cual dificulta el análisis de modelos gráficos. Por su parte el error de especificación, se da por una forma funcional incorrecta, las más conocidas incluyen: i) omisión de variables relevantes, ii) inclusión de variables irrelevantes al modelo, lo cual hace que los estimadores sean sesgados e inconsistentes.

Tradicionalmente, para corregir la endogeneidad se utiliza el método de Variables Instrumentales (IV), la idea general es que la variación observada en la variable independiente puede descomponerse en una parte exógena y una parte endógena (Papies et al., 2017). Tras verificar cuál de los regresores está generando la correlación con el error, se toma otra variable alterna que este fuertemente relacionada con el regresor pero que no con el termino error, cumpliéndose la exogeneidad requerida. El estimador IV más común es el enfoque de mínimos cuadrados en dos etapas (2SLS), que se puede calcular en dos pasos simples (Wang, 2015). El enfoque más cercano en naturaleza a 2SLS es el llamado enfoque de función de control (CF) (Ebbes et al., 2011; Petrin y Train 2010; Wooldridge 2015; Quiroga 2018)

La dificultad de encontrar instrumentos adecuados (Rossi, 2014) ha impulsado la creciente literatura en métodos estadísticos alternativos para solucionar la endogeneidad en el modelado. Se destacan al menos cuatro enfoques en la literatura: el método de variables instrumentales latentes (LIV), cópulas gaussianas, Multiple Indicator Solution (MIS) y el que usaremos en este documento la metodología de error de medida mediante la estimación que explota los momentos de los datos, para nuestro caso, los terceros momentos y usándolos como instrumentos para la estimación de 2SLS, como se muestra en la Figura 1

Figura 1: **Cuatro formas de solucionar endogeneidad sin instrumentos**

Fuente: Elaboración de los autores

Los trabajos de Ebbes y col (2005); Papies et al. (2017); desarrollan el método de variables instrumentales latentes (LIV) que proporciona identificación a través de componentes discretos en los regresores endógenos. Similar al enfoque IV observado, el enfoque LIV comparte la idea subyacente de que el regresor endógeno es una variable aleatoria que se puede separar en dos componentes, la variación exógena y la variación endógena. El componente endógeno se correlaciona con el término de error de la ecuación de regresión principal a través de una distribución normal bivariada.

Lo que se refiere al segundo enfoque, el trabajo de Park y Gupta (2012) titulado: "Handling endogenous regressors by joint estimation using copulas", introducen un método que modela directamente la correlación entre el regresor endógeno y el error utilizando cópulas gaussianas. En pocas palabras, la cópula conecta las distribuciones marginales de dos o más variables que siguen cualquier distribución (por ejemplo, normal, no normal). Park y Gupta (2012)

agregan un término de cópula al modelo que representa la correlación entre la variable endógena y el término de error. Al incluir este término, el efecto del regresor endógeno se puede estimar consistentemente. Tanto la cópula latente IV como la gaussiana explotan la no normalidad en el regresor endógeno y la normalidad de los términos de error. Otro trabajo reciente en esta misma línea es el de Tran y Tsionas (2015) titulado: Endogeneity in stochastic frontier models: Copula approach without external instruments" dónde agregan una estimación de los parámetros del modelo utilizando la máxima verosimilitud. Encuentran que las simulaciones de Monte Carlo se utilizan para evaluar y comparar los rendimientos de muestras finitas de los procedimientos de estimación propuestos.

Ahora bien, lo que concierne al método de la Solución de Indicadores Múltiples (MIS) que está ganando popularidad en el campo por los trabajos recientes como los de Guevara y Polanco (2013) que muestran que el método de indicadores múltiples soluciones (MIS) no requiere instrumentos para corregir la endogeneidad, y lo extiende a modelos de elección discreta, para finalmente simular Monte Carlo para ilustrar la eficacia y la eficiencia de los métodos de MIS y CF en modelos Logit, y para estudiar el impacto del fracaso de sus respectivos supuestos; el de Guevara, Tirachini, Hurtubia, y Dekker (2018) titulado "Correcting for endogeneity due to omitted crowding in public transport choice using the Multiple Indicator Solution (MIS) method", que muestran que se puede utilizar MIS para controlar una amplia gama de atributos omitidos en los datos de Solution Partial (SP). También discuten la posible aplicación de este enfoque a los modelos de transporte público de Preferencias Reveladas (RP) haciendo preguntas específicas a los usuarios después del viaje. Se aplicaron dos variaciones de MIS a este estudio de caso de SP y ambas proporcionaron resultados que fueron superiores a los del modelo restringido. Encuentran que pueden surgir problemas potenciales en presencia de interacciones descuidadas y si los indicadores solo se correlacionan débilmente con el atributo omitido. Para el estudio de caso de SP analizado, solo el primer problema parece jugar un papel en los resultados.

Finalmente, el método que se desarrollará en el documento será el presentando en el trabajo de Lewbel (1997) titulado "Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R D", propone la metodología de estimación explotando los terceros momentos de los datos, y usándolos como instrumentos para la estimación de 2SLS.Tras el cumplimiento de algunos supuestos sobre valores esperados entre los instrumentos generados, errores y desviaciones las estimaciones son consistentes. Adicional, esta propuesta no supone una distribución específica de los errores. La aplicación para este trabajo se basa en datos empíricos sobre el retorno a escala de las patentes en I+D, puesto que existe una variedad de estudios que estiman que los rendimientos a escalas decrecen, mientras que otros afirman que son constantes, la hipótesis del autor es que este tipo de datos tienen problemas de error de medida, por lo cual es importante reducirlo. Haciendo uso de los terceros momentos se realiza la estimación, versus la estimación por 2SLS tradicional, donde el gasto en I+D por empresa es el instrumento. Comparando los resultados de estos modelos, se concluye que los rendimientos a escala son más cercanos a uno.

Metodología

Lewel (1997): Construcción de instrumentos internos, para regresores con error de medición sin datos adicionales disponibles

Arthur Lewbel en su paper "Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R D", expuso teóricamente, (bajo el cumplimiento de ciertos supuestos) que el método de estimacion Higher Moments es muy apropiado cuando no se tiene instrumentos externos. Básicamente éste enfoque, indica que diversas trasnformaciones y momentos estadísticos de las variables en desviaciones, son instrumentos válidos para corregir el error de medida.

Esta forma de estimación, tambien resalta el autor que funciona particularmente bien en grandes muestras y con asimetría en la variable endógena, y como se verá más adelante en nuestro ejercicio de estudio.

Se parte por considerar el siguiente modelo:

$$y_i = a + b'W_i + cX_i + e_i \quad (1)$$

Una ventaja asociada al modelo como se menciono anteriormente, es que los errores no deben presentar una distribución específica, es decir, que no se requiere ni se supone normalidad en los datos, sin embargo, esta metodología por lo general, es desarrollada en varios conjuntos de datos de sección transversal, ya que las estimaciones pueden ser erráticas en muestras muy pequeñas.

$$Z_i = d + X_i + V_i \quad (2)$$

Donde i va desde 1 a n observaciones indexadas. Los parametros a,b,c y d son constantes W_i y b son J vectores de elementos W_{ji} y b_j, todas las otras variables y constantes son escalares. Los datos observados son Y_i, W_i, Z_i para i=1,...,n, mientras X_i, e_i y V_i son no observados. Donde e_i y $d + V_i$ es la medida del error (con d siendo el error de medición medio, por tanto la media de V_i es cero).

La ecuación (1) y (2) implica

$$Y_i = \alpha + b'W_i + cZ_i + \varepsilon_\iota \quad (3)$$

Donde $a = a - cd$ y $\varepsilon_i = e_1 - cV_i$. Sin embargo, la estimación para b y c aplicando OLS es inconsistente, porque el error ε_i esta correlacionado con Z_i , ya que ambos dependen del error de medición V_i.

Sea $\acute{S}$ la media muestral de una variable S_i y sea $G_t = G(W_i)$ para cualquier función dada G, que puede ser transformada en: x^2, x^3, $ln(x)o$ en $1/x$. Entonces

$$(4.a) \quad q_{1i} = (G_i - \bar{G})$$

$$(4.b) \quad q_{2i} = (G_i - \bar{G})(Z_i - \bar{Z})$$

$$(4.c) \quad q_{3i} = (G_i - \bar{G})(Y_i - \bar{Y})$$

$$(4.d) \quad q_{4i} = (Y_i - \bar{Y})(Z_i - \bar{Z})$$

Son todos instrumentos válidos, que se forman de las relaciones en desviación de la variable exógena: (q_{1i}); de las variables exógenas y la endógena: (q_{2i}) ; de las exógenas y la variable respuesta: (q_{3i}) o la variable respuesta y las variables endogenas (q_{4i}). Estos instrumentos tendran correlación con la variables inobservada X_i que depende en el tercer momento de la distribucción conjunta de X_i, W_i y G_i.

Cuando los el error de medida y el error en el modelo, puede garantizarse que son simétricamente distribuidos, entonces pueden emplearse tambien los instrumentos:

$$(4.e) \quad q_{5i} = (Z_i - \bar{Z})^2$$

$$(4.f) \quad q_{6i} = (Y_i - \bar{Y})^2$$

q_{5i} plantea como instrumento a la varianza de la variable endógena, mientras q_{6i} es la varianza de la variable respuesta. Una ventaja asociada al enfoque es que no supone simetría en la distribución de los datos, por el contrario se ajusta mejor en presencia de asimetría. Además, los errores no deben presentar una distribución específica, es decir, que no se requiere ni se supone normalidad en los datos, no obstante, esta metodología por lo general, es más aplicada en varios conjuntos de datos de sección transversal, ya que las estimaciones pueden ser erráticas en muestras muy pequeñas. Se requiere para la consistencia, el cumplimiento de los siguientes supuestos (5):

$$E(qe) = 0$$

$$E(qv) = 0$$

$$E(q\tilde{z}_i) \neq 0$$

Donde $\check{z}_i$ es el residual de la proyeccion de Z. Teniendo en cuenta que $\check{z}_i = \dot{X}_i + v_i$ donde $\dot{X}_i$ es el residual de la proyección de X. Finalmente, la condición $E(q\check{z}_i) \neq 0$ significa que al menos un elemento de $E(q\check{z}_i)$ no es cero.

Ejercicio teórico

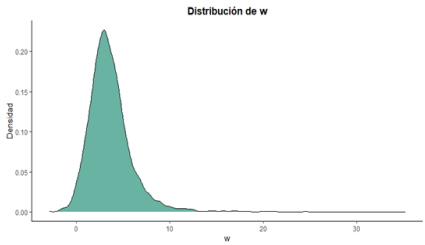
La base de datos para el ejercicio considera 5000 observaciones. La variable respuesta Y está en función de las variables $X1$, $X2$ *y* w. Rutinariamente, el modelo a estimar es:

$$Y_i = \beta_0 + \beta_1 X1 + \beta_2 X2 + \beta_3 w + \epsilon_i$$

Bajo el conocimiento de que hay endogeneidad en el regresor w, se asume que es el único que presenta error de medida, estos estimadores no son consistentes, puesto que la correlación de los regresores con el error están generando un sesgo como resultado en error de medida; y éste sesgo es el causante de la inconsistencia. Una vez corregida la endogeneidad, es de esperar que haya cambios no solo en el regresor endógeno, también en los otros regresores porque el error compuesto $\varepsilon_i = e_1 - cV_i$ definido en (3) es corregido.

De los cuatro instrumentos válidos propuestos por Lewel (1997), se tomaron dos: la relacion en desviaciones de la variable respuesta con la variable endógena (q_{4i}) y la relación de la variable exogena X2 y la variable respuesta (q_{3i}) y se especifica la siguiene transformación para la variable exogena mencionada: $G(W_i) = 1/W_i$. Adicionalmente, se resalta que el tamaño muestral es grande , y la variable endogena (w) presenta un coeficiente de asimetría de 2.81. El *(5000obs)* comportamiento asimetrico de w se puede apreciar en la siguiente grafica:

Figura 2: **Distribución de la variable endógena**



Fuente: Elaboración de los autores.

La tabla 1 muestra las estimaciones por MCO y Higher Moments usando como instrumentos (q_{4i}) y (q_{3i}) respectivamente, a los cuales se le asignaron los nombres HM-YP y HM-GY. Siguiendo a Carrol (2006) el error de medida produce estimadores sesgados e inconsistentes y, dichos estimadores estarán atenuados hacia cero. Siguiendo esa idea y, partiendo del conocimiento previo de que la variable w presenta problemas de endogeneidad, la estimación por MCO será entonces sesgada e inconsistente. Al aplicar el método propuesto se puede apreciar la corrección de la atenuación tanto en los coeficientes del modelo HM-YP como en el modelo HM-GY.

Se destaca el cambio en el coeficiente estimado de la variable endógena, una vez corregido el sesgo, su valor bajo el instrumento (q_{4i}) paso de 0.44 a 0.52 y bajo el instrumento (q_{3i}) de 0.44 a 0.618. Se validaron los supuestos de consistencia descritos por Lewbel (1997), los cuales se cumplen satisfactoriamente. Para el modelo HM-YP las covarianzas estimadas fueron $E(q\varepsilon) = 3.088786e - 14$; $E(qv) = 0.068$ y $E(q\check{z}_i) = 24.12$. Y Para el modelo HM-GY: $E(q\varepsilon) =)=-3.152539e$; $E(qv) = 0.007$ y $E(q\check{z}_i) = -1.17$

La diferencia entre los estimadores del modelo HM-YP y HM-GY no pasa inadvertida, a pesar de que ambas corrigen la atenuación en la dirección correcta, el coeficiente asociado a la variable w difiere considerablemente en ambos modelos. Por lo tanto, se realizó el test diagnóstico de Instrumentos débiles, esta es una prueba F sobre los instrumentos en la primera etapa. La hipótesis nula es esencialmente

que tenemos instrumentos débiles, por lo que un rechazo significa que nuestros instrumentos no son débiles.

De acuerdo con los resultados anteriores, se observa que el p-valor asociado al modelo que tiene el instrumento GY es 0.0985, mientras que el p-valor del modelo asociado al instrumento YP es un valor menor a 2e-16.

Aunque en ambos modelos no se acepta la hipotesis nula , se debe resaltar que ante la hipótesis nula de tener instrumentos debiles, con el instrumento GY hay significancia al **0.1**, y con el instrumento YP la significancia es al **0.001**. Esto nos indica que el modelo bajo el instrumento YP, presenta resultados mucho mas robustos.

Siguiento nuevamente a Carroll (2006), el sesgo y la inconsistencia en el coeficiente asociado a la variable endógena w, es producida por un factor de atenuación:

$$\lambda = \frac{\sigma_w^2}{\sigma_e^2 + \sigma_w^2}$$

Al realizar esta corrección, se obtiene que el coeficiente asociado a la variable endógena w es igual a 0.5191,el cual resulta bastante similiar al coeficiente obtenido bajo el modelo HM-YP **(0.523)**. Esto podría indicar que la correción bajo el modelo HM-YP es la mas adecuada. Por otro lado, al realizar el test de Wu-Hausman para los modelos HM-YP y HM-GY, se encuentra que en el modelo HM-YP no se acepta la hipotesis nula de consistencia y eficiencia, mientras que en el modelo HM-GY se acepta la hipótesis nula, *(p-valor:0.2848)*.

Finalmente, por ser un ejercicio de simulación, es difícil concluir si éste resultado económicamente tiene sentido, o si es posible hablar de una relación causal. Por tanto, en la sección siguiente se hará una validación empírica donde podremos realizar inferencia y comparar los resultados aplicando la estimación por variable instrumental (IV) y Higher Moments.

Tabla 1: **Resultados de Regresión**

	Variable dependiente: Y		
	MCO	HM-YP	HM-GY
	(1)	(2)	(3)
X1	-0.962^***	-1.011^***	-1.068^***
	(0.011)	(0.011)	(0.119)
X2	0.351^***	0.412^***	0.485^***
	(0.011)	(0.012)	(0.149)
w	0.444^***	0.523^***	0.618^***
	(0.004)	(0.006)	(0.193)
Intercepto	5.223^***	4.918^***	4.556^***
	(0.021)	(0.028)	(0.742)
Observations	5,000	5,000	5,000
R^2	0.792	0.773	0.704

Adjusted R^2	0.792	0.773	0.704
Residual Std. Error (df = 4996)	0.665	0.693	0.792
Note:	*p<0.1; **p<0.05; ***p<0.01		

Resultados

Para la aplicación empírica se tomaron los datos de Stock & Watson (2007), disponibles en el paquete (AER) en el software R. La base contiene datos de corte transversal sobre el rendimiento de las pruebas, las características de la escuela y los antecedentes demográficos de los estudiantes de los distritos escolares de California para 1998. En total son 420 distritos escolares (n = 420) y 14 variables.

Con análisis cross-section, se investiga cómo el ratio alumno/maestro afecta el puntaje medio de lectura.

read es la variable respuesta que mide el puntaje medio de lectura. La variable que mide la proporción estudiante/maestro, es la variable endógena. Esto porque podría estar correlacionada con factores no observados, como los salarios de los maestros o las condiciones de trabajo del maestro, que no son observados, pero pueden afectar el puntaje de lectura de los estudiantes.

El instrumento externo para esta variable es: ***expenditure*** y mide el gasto por estudiante agregado a nivel de distrito. Esto es posible ya que está correlacionado con la proporción alumno/maestro, pero no explica directamente los resultados en las pruebas de puntaje de lectura de los alumnos. Por lo tanto, resulta ser un instrumento idóneo, y es usado en la regresión de Minimos Cuadrados en Dos Etapas (2SLS)

Las otras variables que influyen en ***read*** son ***english***: porcentaje de aprendices de inglés; ***lunch***: porcentaje que tienen un almuerzo a precio reducido; ***income***: ingresos; ***calworks***: si el condado califica para asistencia de ingresos; y ***county*** al igual que ***grades***: son variables ficticias. El modelo a estimar será:

$$read_i = \beta_0 + \beta_1 stratio + \beta_2 english + \beta_3 grades + \beta_4 lunch + \beta_5 calworks + \beta_6 county + \beta_7 grades + \epsilon_i$$

Los resultados y las comparaciones de los estimadores por OLS, IV y Higher Moments, son resumidos en la tabla (2). Para el modelo HM, se usó como instrumento la trasformación q_2, y se realizó la transformación $G = w^3$ sobre la variable *lincome*

Como puede observarse, el coeficiente asociado a ***stratio***, la variable endógena, bajo MCO es de -0.30. La estimación bajo el instrumento propuesto por los autores arroja un coeficiente de -1.137, mientras que la estimación bajo la metodología propuesta, HM, arroja un coeficiente de -1.308, corrigiendo el sesgo en el mismo sentido y cercano a los resultados bajo IV. Respecto a los coeficientes de los otros regresores se observa que bajo IV como con HM las estimaciones no difieren en gran medida.

También se replica el ejercicio propuesto por Stock & Watson en ***Introducton to Econometrics*** sobre demanda de cigarrillos, los datos tambien fueron tomados del paquete estadistico (AER) del software R. De este panel se hizo análisis cross-section al comparar las estimaciones solo para 1995. El ejercicio en este caso busca estimar la elasticidad precio de la demanda de cigarrillos. El modelo a estimar es el siguiente:

$$lquant_i = lprice + lincome + \epsilon_i$$

La cantidad consumida *lquant* es la variable respuesta, y se mide por las ventas anuales de cigarrillos; *lprice* es una variable que mide el precio promedio de cigarrillo minorista por paquete durante el año fiscal, incluidos los impuestos, y asumimos que es la variable endógena. Se consideran dos instrumentos: *tdiff* : que representa el impuesto percápita sobre cigarrillos y *tax/cpi* que denota el impuesto real sobre los cigarrillos . *lincome* representa el ingreso percápita y es considerada una variable exógena. Los resutados y comparaciones de la estimación son aportados en la tabla (3).

El instrumento usado para la estimación del modelo HM es q_2, y la transformación sobre la variable exógena *lincome* es $G(W_i) = W^3_i$. La regresión por MCO muestra que el coeficiente de la variable endógena *lprice* es -1.407; al corregir la endogeneidad bajo los instrumentos propuesto por los autores, el coeficiente pasa a -1.227, y bajo la metodología propuesta, el estimador es -1.162. Por su parte, el coeficiente de la variable exógena lincome bajo ols es 0.344, mientras que bajo IV y HM es bastante similar: 0.280 y 0.224 respectivamente.

Quizá una explicación a la diferencia entre los coeficientes en la estimación IV y HM, se sustenta por el hecho de que en la metodología propuesta por Lewbel (1997) los resultados son mas robustos en corregir el sesgo por el error de medida, cuando la muestra es grande. En este caso la base de demanda para cigarrillo solo tiene 48 observaciones, y podría explicar la sobre-estimación que el modelo HM hace en comparación con el modelo por IV. Un comentario similar puede hacerse sobre las estamaciones del primer caso mostrato en la tabla 2, donde la muestra, a pesar de tener un buen tamaño, $n = 420$, pueda que la estimación de la variable *stratio* este subestimada debido a que dicha variable no presenta con coeficiente de asimetria alto **(-0.025)**. En ambos modelos la incertidumbre se ve reflejada en el error estandar, el cual resulta considerablemente alto.

Tabla 2: **Resultados de Regresión**

Variable dependiente: *read*			
	MCO	2SLS	HM
	(1)	(2)	(3)
stratio	-0.300	-1.137^**	-1.308
	(0.258)	(0.535)	(2.731)
english	-0.206^***	-0.214^***	-0.216^***
	(0.038)	(0.038)	(0.047)
lunch	-0.387^***	-0.394^***	-0.395^***
	(0.037)	(0.038)	(0.044)
grades	-1.913	-1.892	-1.888
	(1.359)	(1.378)	(1.388)
income	0.716^***	0.625^***	0.606^*
	(0.098)	(0.112)	(0.313)
calworks	-0.053	-0.050	-0.049
	(0.062)	(0.062)	(0.064)
Intercepto	683.453^***	700.479^***	703.956^***
	(9.562)	(13.581)	(56.183)
Observations	420	420	420

R	0.877	0.873	0.872
Adjusted R	0.860	0.856	0.855
Residual Std. Error (df = 369)	7.515	7.621	7.668
Note:	*p<0.1; **p<0.05; **p<0.01		

Table 3: **Resultados de Regresión**

Variable dependiente: *lquant*			
	MCO	2SlS	HM
	(1)	(2)	(3)
lprice	-1.407^***	-1.277^***	-1.162
	(0.251)	(0.263)	(1.357)
lincome	0.344	0.280	0.224
	(0.235)	(0.239)	(0.697)
Constant	10.342^***	9.895^***	9.496^*
	(1.023)	(1.059)	(4.730)
Observations	48	48	48
R	0.433	0.429	0.421
Adjusted R	0.408	0.404	0.395
Residual Std. Error (df = 45)	0.187	0.188	0.189
Note:	*p<0.1; **p<0.05; ***p<0.01		

Discusión

En la literatura se ha evidenciado que existe un creciente interés por nuevas metodologías que diversifiquen los tipos de estimación, teniendo en cuenta las especificaciones y retos que imponen los datos reales al momento de hacer inferencia estadística que es usada en la toma de decisiones. Este documento agrega evidencia de la aplicación del método de Higher Moments, el cual se enmarca dentro de las alternativas de la resolución de problemas de endogeneidad sin el uso tradicional de instrumentos externos para variables con alto grado de asimetría.

La endogeneidad puede deberse a múltiples factores, variables omitidas, simultaneidad o error de medida. Centrándonos en el error de medida, este documento agrega evidencia a favor de la aplicación del método de Higher Moments el cual se enmarca en las alternativas de la resolución de problemas de endogeneidad sin el uso tradicional de instrumentos externos y alta asimetría en la variable endógena. La ventaja de Higher Moments, respecto a otros métodos de estimación, es que no supone una forma específica ni modela el error, además pueden emplearse diversos instrumentos y transformaciones entre ellos. El método requiere garantizar la correlación entre la variable inobservable que tiene el error de medida con el tercer

momento de la distribución conjunta entre los regresores y la endógena.

Finalmente, la investigación evidencia que esta metodología se ajustó bastante bien a la base de datos trabajada donde la variable endógena presentaba un alto grado de asimetría, y el tamaño muestral era grande. Al aplicar la metodología en dos bases de datos diferentes pudo observarse que los resultados fueron muy similares a los obtenidos bajo instrumentos externos. Sin embargo, en una base se subestimó el resultado porque el tamaño muestral era bajo y en otro caso se sobreestimó debido a que la baja asimetría de la variable endógena, corroborando los resultados hallados por otros estudios presentes en la literatura.

Conclusiones

En el análisis de relación causal, es muy importante ocuparse por la robustez de las estimaciones. No obstante, ante problemas de endogeneidad, en la práctica suele ser muy complejo encontrar instrumentos apropiados que estén correlacionados con el regresor endógeno y no con el error. De ahí la importancia y el reto de hallar instrumentos internos que puedan ser usados para corregir la endogeneidad cuando los instrumentos externos no están disponibles.

En este documento, se parte por considerar que la inconsistencia se debe a un error de medida en la variable endógena, y se exploran mecanismos de solución para corregir el sesgo. Inicialmente, con una base de datos que no posee instrumentos y posteriormente se probó la metodología propuesta con las bases de datos de Stock y Watson, las cuales presentan endogeneidad y tiene instrumentos externos. El objetivo es comparar los estimadores IV con los construidos metodológicamente.

De la revisión de literatura, la metodología de Lewbel (1997) tuvo más afinidad. Los supuestos que deben cumplirse, la construcción de los instrumentos internos y las transformaciones posibles, fueron discutidos en secciones previas. Para la primera base de datos se encontró que la estimación de Higher Moments por los instrumentos q_3 y q_4, así como la transformación $1/W_i$ fue la más acertada.

Para la aplicación empírica, la metodología propuesta fue comparada con dos bases de datos y modelos diferentes tomados de Stock y Watson: en los cuales se busca estimar la elasticidad precio-demanda de cigarrillos y el efecto del ratio estudiantes/ profesores sobre el test de lectura en escuelas de California.

Para el estudio sobre las escuelas en California se evidenció que la metodología propuesta fue cercana a la obtenida por IV. No obstante, se subestimó el resultado, esto puede atribuirse a la poca asimetría que presenta la variable endógena. Un problema similar ocurre en la estimación del ejercicio sobre la demanda de cigarrillos, pero en este caso la sobre estimación se atribuye a un tamaño muestral pequeño (n=48).

Lista de Referencias

Amsler, C., Prokhorov, A., Schmidt, P. (2016). Endogeneity in stochastic frontier models. Journal of Econometrics, 190(2), 280-288.

Carroll, R. J., Ruppert, D., Stefanski, L. A., Crainiceanu, C. M. (2006). Measurement error in nonlinear models: a modern perspective. CRC press.

Duncan, G. J., Magnuson, K. A., Ludwig, J. (2004). The endogeneity problem in developmental studies. Research in human development, 1(1-2), 59-80

Ebbes, P., Wedel, M., Böckenholt, U., Steerneman, T. (2005). Solving and testing for regressor-error (in) dependence when no instrumental variables are available: With new evidence for the effect of education on income. Quantitative Marketing and Economics, 3(4), 365-392.

Greene, W. H. (2008). The econometric approach to efficiency analysis. The measurement of productive efficiency and productivity growth, 1(1), 92-250.

Greene, W. H. (2003). Econometric analysis. Pearson Education India.

Guevara, C. A., Polanco, D. (2016). Correcting for endogeneity due to omitted attributes in discrete-choice models: the multiple indicator solution. Transportmetrica A: Transport Science, 12(5), 458-478.

Hamilton, B. H., Nickerson, J. A. (2003). Correcting for endogeneity in strategic management research. Strategic organization, 1(1), 51-78.

Hausman, J. A. (1978). Specification tests in econometrics. Econometrica: Journal of the econometric society, 1251-1271.

Huang, J., Zhao, P., Huang, X. (2019). Instrumental variable based SEE variable selection for Poisson regression models with endogenous covariates. Journal of Applied Mathematics and Computing, 59(1-2), 163-178.

Lehmann, E. L. (2004). Elements of large-sample theory. Springer Science Business Media.

Lewbel, A. (1997). Constructing instruments for regressions with measurement error when no additional data are available, with an application to patents and R D. Econometrica: journal of the econometric society, 1201-1213.

Lin, W., Wooldridge, J. M. (2019). Testing and correcting for endogeneity in nonlinear unobserved effects models. In Panel Data Econometrics (pp. 21-43). Academic Press.

Martínez, R., Lyssenko, N. (2011). Correcting for the endogeneity of pro-environment behavioral choices in contingent valuation. Ecological Economics, 70(8), 1435-1439.

Nakamura, A., Nakamura, M. (1998). Model specification and endogeneity. Journal of Econometrics, 83(1-2), 213-237

Papies, D., Ebbes, P., Van Heerde, H. J. (2017). Addressing endogeneity in marketing models. In Advanced methods for modeling markets (pp. 581-627). Springer, Cham.

Quiroga, B. F. (2018). Addressing Endogeneity Without Strong Instruments: A Practical Guide to Heteroskedasticity-Based Instrumental Variables. Available at SSRN 3293789.

Rossi, P. E. (2014). Even the rich can make themselves poor: A critical examination of IV methods in marketing applications. Marketing Science, 33(5), 655-672.

Shee, A., Stefanou, S. E. (2015). Endogeneity corrected stochastic production frontier and technical efficiency. American Journal of Agricultural Economics, 97(3), 939-952.

Stock, J. H., Watson, M. W. (2015). Intro-

duction to econometrics.

Terza, J. V. (1998). Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. Journal of econometrics, 84(1), 129-154.

Tran, K. C., Tsionas, E. G. (2015). Endogeneity in stochastic frontier models: Copula approach without external instruments. Economics Letters, 133, 85-88.

Wang, C. J. (2015). Instrumental variables approach to correct for endogeneity in finance. In Handbook of financial econometrics and statistics (pp. 2577-2600). Springer USA.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. Journal of Human Resources, 50(2), 420-445.

Wasserstein, R. L., Schirm, A. L., Lazar, N. A. (2019). Moving to a world beyond "p 0.05".