RICYDE. REVISTA INTERNACIONAL DE CIENCIAS DEL DEPORTE
http://www.ricyde.org
doi:10.5232/ricyde

RICYDE. Revista Internacional de Ciencias del Deporte
Volumen XII - Año XII
Paginas:335-337 ISSN:1885-3137
Número 46 - Octubre - 2016

EDITORIAL
http://dx.doi.org/10.5232/ricyde2016.046ed

Endogeneidad, diferencia de medias y regresión [Endogeneity, mean difference and regression]

José A. Martínez
Universidad Politécnica de Cartagena

La endogeneidad es un concepto estadístico que se refiere a que la relación entre una variable explicativa *x* y otra que queremos explicar *y* viene determinada por otras variables que no se han tenido en cuenta *e* y que covarían con *x*. De este modo, $cov(x,e)\neq 0$.

A pesar de que es un tema muy comentado en los libros de econometría (ej. Greene, 2008; Stock y Watson, 2007; Wooldridge, 2003), donde se discute sobre las causas de ese problema, las consecuencias de no tratarlo adecuadamente, y las posibles soluciones para minimizar sus efectos perversos sobre las estimaciones, en ciertas ocasiones los autores y revisores olvidan considerarlo en las investigaciones en ciencias del deporte. Las razones de por qué ocurre esto en este campo de conocimiento son seguramente complejas, aunque es posible que algunas vengan determinadas por la más estrecha relación entre las ciencias del deporte y la metodología en psicología y sociología con respecto a economía (econometría). Esa discrepancia entre econometría y la psicología/sociología en cuanto al tratamiento de la endogeneidad es comentada en Antonakis y col. (2010).

La no consideración de la endogeneidad provoca estimadores inconsistentes y sesgados, lo que significa que los resultados son erróneos. Por tanto, es una grave amenaza a la validez de cualquier investigación que no lo tenga en cuenta en su análisis.

La diferencia de medias y la regresión

Es un error común también pensar que realizar una prueba de diferencia de medias (t-Student, ANOVA) no puede enmarcarse en un modelo lineal general de estimación, que en su versión más sencilla, no es más que un modelo de regresión con una única covariable. La palabra "modelo" es fundamental, porque los investigadores tienen que plantear que los datos que manejan se generan a partir de un modelo.

Los modelos son una representación de la realidad que debe ser testada sobre los datos empíricos para comprobar su validez. Boos y Stefanski (2013) defienden que todos los modelos son aproximaciones y que los modelos propuestos son tentativos, sujetos a la re-evaluación y modificación en el curso del análisis de datos. Pero esos datos no se generan por azar (sólo el momento del decaimiento nuclear radiactivo se considera aleatorio), por lo que es obligación del investigador planear una o varias alternativas de modelización para explicar los datos empíricos. Es ahí donde entran a formar parte las hipótesis y teorías que se plantean para explicar los datos, considerando que en cualquier modelo propuesto están implícitas tanto la especificación de las variables y su relación funcional, como las asunciones (Spanos, 2007).

Por tanto, cuando se plantea la siguiente sencilla ecuación para una muestra de *i* individuos;

$$yi=a+\beta 1xi+ei \quad (1)$$

donde *y* es la variable que queremos explicar, *x* es una variable dicotómica que representa dos grupos de individuos distintos (como en el caso de un tratamiento experimental vs control), y *e* representa un conjunto de variables que desconocemos que podrían explicar *y,* se está suponiendo también que existe una relación lineal entre *y* y *x*, y que $cov(x,e)=0$, entre otras asunciones. Esas son restricciones que forman parte del modelo y deben ser testadas.

Simplificando al máximo, si en una investigación en ciencias del deporte hemos diseñado un experimento para ver la eficacia de un determinado tratamiento *x*, entonces $\beta 1$ representa el cambio en la esperanza de *y* cuando *x* varía una unidad (Pearl, 2000), o lo que es lo mismo, lo que cambia la media de *y* entre el grupo de control y el experimental, que es exactamente la misma información que provee una prueba t-Student y un ANOVA.

Sin embargo, las ventajas de plantear el modelo de esta forma, Eq (1), es que se tiene un marco mucho más amplio para testar las asunciones del modelo, por ejemplo al estimar por mínimos cuadrados ordinarios. Y además, se entiende más claramente que esa ecuación representa el modelo planteado por los autores, con las restricciones que ello supone.

En el estudio de Antonakis y col. (2010) se puede encontrar una explicación muy extensa sobre cómo plantear estos modelos y solucionar diferentes amenazas a la validez de los mismos.

Diseño experimental vs. no experimental

En un diseño completamente aleatorio, en el que la muestra es aleatoria y la asignación de los grupos de control y experimental tambien lo es, la Eq (1) puede ser suficiente para explicar los datos empíricos. Sin embargo, cuando los tamaños de los grupos no son muy grandes, la introducción de covariables es deseable, ya que esas covariables a veces son precisamente las variables omitidas en el análisis que sesgarian el coeficiente *ß1*, al ser el tamaño muestral pequeño.

Por tanto, es conveniente en diseños experimentales totalmente aleatorizados considerar covariables que reduzcan la probabilidad de que aparezca un problema de endogeneidad. De nuevo el marco del modelo lineal general de regresión ofrece una mayor flexibilidad que las especificación de ANOVA de varios factores o ANCOVA.

En cualquier caso, esas asunciones son testables. Tanto las especificación, como la asuncion de linealidad y de exogeneidad de *x* son testables vía análisis de los residuos (ver Spanos, 2007), y empleando herramientas como el test RESET (ver Antonakis y col., 2010).

Sin embargo, la mayor parte de las investigaciones son no experimentales, y trabajan con datos observacionales o con diseños en los que la asignación de casos a tratamientos no es aleatoria. Aquí es donde cobra mayor relevancia si cabe el tratamiento de la endogeneidad, empleando lo que a veces se denomina variables de control, que no son más que covariables que correlacionan con *x* y que "aislan" su efecto sobre y. La palabra aislamiento significa que la estimacion del efecto *ß1* es consistente e insesgada.

Antonakis y col. (2010) proveen un sencillo gráfico donde nos muestran en que condiciones *ß1* seria consistente (Figura 1). En el caso A, *ß1* es consistente porque *x* no correlaciona con *e*. En el caso B *ß2* es inconsistente porque *z* correlaciona con *e*, es decir, *cov(z,e)≠0* En el caso C, aunque *x* es exógeno *ß1* es inconsistente porque *z* correlaciona con *e* y *x* y sesga el coeficiente de *x*. En el caso D, *ß1* es consistente incluso si *ß2* no lo es, porque *x* y *z* no correlacionan.

Por tanto, a efectos prácticos, nuestro modelo debe incluir todas las covariables *z* que correlacionen con *x* y que sean teoricamente relevantes. Las consecuencias de esta afirmación son de nuevo contundentes: (1) Las variables que no covarian con *x* pueden omitirse en el modelo sin afectar a *ß1*; (2) La teoría es fundamental para identificar las posibles covariables *z* que deben ser incluidas en el análisis.

La primera consecuencia implica que si omitimos variables *z* relevantes para explicar *y* pero que no covarían con *x*, seguiríamos obteniendo una estimación válida *ß1*, aunque la varianza explicada de *y*, es decir, R^2, se vería afectada. Pero en muchas ocasiones, el principal interés de la investigación radica en testar el efecto de una intervención *x* sobre *y*, por lo que aunque globalmente el modelo sea poco explicativo, el efecto de interés está correctamente estimado y se puede interpretar.

La segunda consecuencia implica que los modelos se construyen a partir de la teoria, y que cualquier afirmacion causal o simplemente correlacional del modelo planteado se basa en la fortaleza de la teoria subyacente (Pearl, 2000). Si hay varias teorías, se pueden plantear modelos alternativos y comparar su ajuste a los datos empiricos.

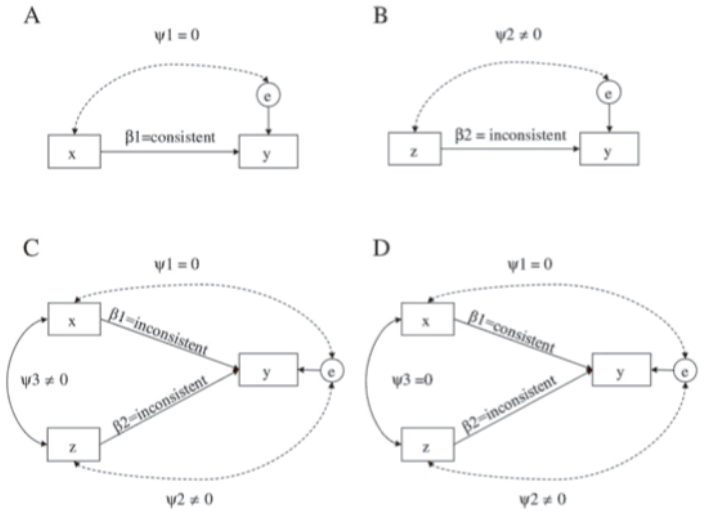

Figura 1. Cómo la endogeneidad afecta a la consistencia (Antonakis y col., 2010).

Cuando *x* no es dicotómica o categórica sino contínua, el modo de conceptualizar este problema y de operar con él es exactamente el mismo.

Conclusión

En los últimos años y de manera acertada se está poniendo mucho énfasis en la investigación en ciencias del deporte en conceptos tales como la potencia estadística o el tamano de efecto, en aras de evitar los errores derivados de interpretar únicamente la significación estadística (Barriopedro, 2015).

Sin embargo, autores y revisores deberían/deberíamos también prestar mas atención al problema de la endogeneidad en la modelización y analisis de datos, ya que es una grave amenaza a la validez. Una guía sencilla sobre como tratar este problema para diferentes contextos de modelizacion puede verse en Antonakis y col. (2010), así como sugerencias de actuación y consejos extremadamente prácticos para los investigadores aplicados.

De este modo, en ciencias del deporte sería conveniente un mayor uso del modelo lineal general, que en su versión mas sencilla, Eq (1), se sitúa como punto de partida para desarrollar modelos más amplios con la inclusión de covariables, ya sea para testar una simple diferencia de medias o para establecer cualquier otro tipo de análisis. Esto ayudaría a los autores a descartar el análisis "por separado" de diferentes variables y su influencia sobre la variable que quieren explicar, algo que, desafortunadamente, constituye un error muy común.

La idea es, por tanto, plantear desde el comienzo modelos completos que pueden ser testados contra los datos, con todo lo que implica en cuanto a desarrollar test de mala especificación (Spanos, 2007) para estudiar las asunciones. Si hay varios modelos alternativos, éstos se pueden comparar, y si ocurre alguna re-especificación, ésta debería ser testada con nuevos datos, no sólamente con los actuales. Aunque los modelos sean aproximaciones de la realidad según Boos y Stefanski (2013), todo puede ser "modelado", incluyendo en los modelos especificaciones relativas a variables omitidas, error de medida, sesgo de método común u otro tipo de potenciales efectos (Hayduk, 1996). En este sentido los modelos de ecuaciones estructurales presentan cierto atractivo para manejar esa complejidad aunque siempre teniendo en cuenta que deben ajustarse con el test de la chi-cuadrado (Antonakis y col., 2010; Hayduk, 2014), y no por cualquier otro tipo de índice (RMSEA, CFI, TLI, etc.).

Agradecimientos

Este trabajo es el resultado de la actividad desarrollada en el marco del Programa de Ayudas a Grupos de Excelencia de la Región de Murcia, de la Fundación Séneca, Agencia de Ciencia y Tecnología de la Región de Murcia proyecto 19884/GERM/15. Asimismo, el autor agradece la financiación recibida del proyecto ECO2015-65637-P (MINECO/FEDER).

El uso de estimaciones en mínimos cuadrados en dos etapas con variables instrumentales, los modelos de Heckman para corregir el sesgo de autoselección en diseños no experimentales, o la apropiada distinción entre modelos de efectos fijos y aleatorios, se pueden emplear incluso en los casos más sencillos del modelo lineal general, tal y como sugiere Antonakis y col. (2010). Las herramientas metodológicas, por tanto, están a disposición de los investigadores en ciencias del deporte para superar varios de los errores comunes que surgen de la no consideración de la endogeneidad.

Referencias

Antonakis, J.; Bendahan, S. ; Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly, 21*, 1086-1120, http://dx.doi.org/10.1016/j.leaqua.2010.10.010

Barriopedro, M. I. (2015). La significación estadística no es suficiente. *RICYDE. Revista Internacional de Ciencias del Deporte, 40*(11), 101-103. http://dx.doi.org/10.5232/ricyde2015.040ed

Boos, D. D. & Stefanski, L. A. (2013). *Essential Statistical Inference: Theory and Methods*. New York: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4614-4818-1

Greene, W. H. (2008). *Econometric analysis*. 6th ed. Upper Saddle River, N.J.: Prentice Hall.

Hayduk, L. A. (2014). Shame for disrespecting evidence: the personal consequences of insufficient respect for structural equation model testing. *BMC Medical Research Methodology*, *14*, 124. http://dx.doi.org/10.1186/1471-2288-14-124

Hayduk, L. A. (1996). LISREL: Issues, debates, and strategies. Baltimore: Johns Hopkins University Press.

Pearl, J. (2000). Causality: Models, Reasoning and Inference. Cambridge, UK: Cambridge University Press Spanos, A. (2007). Curve fitting, the reliability of inductive inference, and the error-statistical approach. *Philosophy of Science, 74*, 1046–1066. http://dx.doi.org/10.1086/525643

Stock, J. H. & Watson, M. W. (2007). *Introduction to econometrics*. 2nd edition. Boston: Pearson Addison Wesley.

Wooldridge, J. M. (2003). *Introducción a la econometría: Un enfoque moderno.* Thomson.

Jose A. Martínez
Universidad Politécnica de Cartagena (España).
Facultad de Ciencias de la Empresa.
Email: josean.martinez@upct.es