Variables Endógenas en los Modelos Econométricos

CONFERENCIA INAUGURAL
XLIX REUNION ANUAL Asociación Argentina de Economía Política
Facultad de Ciencias Económicas -Universidad Nacional de Misiones
12 de Noviembre de 2014

Hildegart Ahumada
UTDT

Introducción

- Endogeneidad y exogeneidad son conceptos que han sido ampliamente discutidos desde hace mucho tiempo atrás.

- Aunque con otros nombres (por ejemplo, “autonomía” han estado presentes desde los años 30 en los trabajos de Frisch, Timbergen, Haavelmo, Koopman , Granger, Sims, Engle, Hendry, Heckman, entre otros .

- No parece que puede hablarse de que haya consenso sobre
  - su definición sobre la base de errores no observables o sobre la distribución de probabilidad de las variables observables, por ejemplo, Engle, Hendry and Richard (1983)
  - ni sobre los varios caminos que se han sugerido para su consideración en la modelización empírica como el uso y para algunos “abuso” de variables instrumentales, por ejemplo, Imbens (2010) sobre los trabajos de Deaton (2009) and Heckman and Urzúa (2009)

Introducción

- Esto parecería contrastar con la aceptación, aparentemente sin mayor discusión, de la tendencia reciente de

- englobar en el concepto de ***endogeneidad*** un gran número de casos que pueden implicar estimadores ***inconsistentes*** *y que lleva a preguntarnos*

- **¿no tenemos hoy demasiadas variables endógenas en los modelos econométricos**?

Las raíces del concepto ampliado de Endogeneidad

En Wooldrige *"Econometric Analysis of Cross Section and Panel Data" (2010) p. 54:*

An explanatory variable x_j *is said to be* ***endogenous*** *…*

if it is correlated with u*. You should not rely too much on the meaning of "endogenous " from other branches of economics. In traditional usage, a variable is endogenous if it is determined within the context of a model. The usage in econometrics, while related to traditional definition, has evolved to describe any situation where an explanatory variable is correlated with the disturbance.*

If x_j *is uncorrelated with* u*, then* x_j *is said to be* ***exogenous…"***

Los casos correspondientes al concepto ampliado de Endogeneidad

- Se enumeran los siguientes casos en econometría aplicada, en orden,
  - ✓ **Variables Omitidas** (incluyendo las no observables como *self- selection*)
  - ✓ **Error de medición**
  - ✓ **Simultaneidad**
- Se cita el trabajo de Deaton (1995) para la discusión de estos tipos de **endogeneidad** en el campo de desarrollo económico, pero Deaton los considera como casos en que los estimadores obtenidos por MCO son inconsistentes (ver sección 2.15 *Regression Bias*, p. 1823).
- A estos 3 casos se agregan otros problemas de selección muestral (por ej., por "truncamiento") en Baltagi (2013) y Greene (2012).

Tres cuestiones relativas a este concepto de Endogeneidad

1. Si no es discutible la idea de evolución: se aleja del concepto de otras ramas de la economía, e incluso de otras ciencias cuando vemos que las ciencias se mueven en un camino cada día más interdisciplinario.

2. Si son realmente los 3 casos enunciados de endogeneidad:

- el 3ro lo es sin duda (asociado a la cuestión de identificación de modelos estructurales);
- el 2do podría serlo, pero
- el 1ro parece ser el más cuestionable. Por ejemplo, en Hansen (2000) se consideran los otros dos pero no el de variables omitidas.

3. Si todo lo que requiere IVE necesita asociarse a un problema de endogeneidad

El concepto de Endogeneidad en otras ciencias

- La distinción entre variables endógenas/exógenas no está presente en matemática.
- Los conceptos relevantes son variables dependientes/independientes

- Sí se encuentran en las ciencias biológicas o de materiales.
- En ellas está siempre presente la idea básica de un **sistema** en las que las endógenas están generadas dentro de él mientras la acción de las exógenas provienen de fuera del sistema.

El concepto de Endogeneidad en Econometría

- En Angrist and Pischke (2009, p.109) reservan el concepto de endógeno al más tradicional de sistemas de ecuaciones (SEM) donde las variables **endógenas** son determinadas conjuntamente dentro del sistema y las **exógena**s fuera del mismo.

- En Wooldridge:

Endogeneidad $Cov[x_j\ u] \neq 0$

Exogeneidad $Cov[x_j\ u] = 0$

El concepto de Endogeneidad en Econometría

- Y en consecuencia todo rechazo de la igualdad llevaría a un caso de endogeneidad (y a inconsistencias de los estimadores MCO). En cambio en la literatura pionera (Koopmans and Hood, 1953), retomada en Sims (1977) y discutida entre otros por Engle, Hendry and Richard (1983),

Exogeneidad ("estricta")

$$Cov\left[x_{ji}\ u_{i+l}\right] = 0\ para\ todo\ l$$

Predeterminada

$$Cov\left[x_{ji}\ u_{i+l}\right] = 0\ para\ todo\ l \geq 0$$

El concepto de Endogeneidad en Econometría

Para Wooldridge no habría diferencia entre "exogeneidad estricta" y "predeterminación "

Porque en la definición de su modelo de regresión poblacional (el PGD) la variable a explicar y

y las diferentes variables explicativas $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, ... $\boldsymbol{x_k}$ son

- *"... observable random scalars (that we can observe them in a random sample of the population..." (p. 53)*
- **y por consiguiente, son independientes e idénticamente distribuidos *(iid)***

Suponer este PGD para las variables económicas y no pensarlo como ***procesos*** *aleatorios* (como lo define, entre muchos otros, Greene 2012, p 53) no es un tema menor,

en particular si la definición de endogeneidad/exogeneidad no quiere ser acotada a determinados casos particulares de PGD.

El concepto de Endogeneidad en Econometría

¿Tiene sentido hablar de endogeneidad en forma diferente según el tipo de series económicas en las que estemos pensando? Y si es no

¿Es válido el supuesto PGD?

- Decididamente, no lo es para **series temporales** debido a la persistencia de la series económicas
- Ni para **datos de panel**. Por ejemplo, en Arellano Bond (1991) cuando define la matriz óptima de instrumentos para las estimaciones de GMM para paneles dinámicos de acuerdo a si las explicativas son estrictamente exógenas o predeterminadas. Tampoco cuando pensamos en los casos de *cross dependence* en los paneles de T grande y N grande)
- Ni siquiera para **corte transversal** si pensamos en la interacción entre unidades geográficas (o de otro tipo) estudiadas en la econometría espacial (ver por ejemplo Elhorst, 2014).

Casos de **Endogeneidad**: **SIMULTANEIDAD**

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Como

$$x_i = f(y_i) = g(u_i)$$

Siendo f y g funciones lineales y en consecuencia

$$Cov\,[x_i\ u_i] \neq 0$$

Si pensamos en precios y cantidades de equilibrio en mercados, tendremos numerosos ejemplos de posible endogeneidad.

Casos de **Endogeneidad** : **ERROR DE MEDICION?**

$$y_i = \beta_0 + \beta_1 x_i^* + u_i \qquad u_i \sim \boldsymbol{iid}(0, \sigma_u^2)$$

Donde x_i^* es la variable explicativa relevante siendo

$$E[x_i^* u_i] = 0$$

pero es no observable y la observable x_i está sujeta a un error de medición w_i,

$$x_i = x_i^* + w_i \qquad w_i \sim \boldsymbol{iid}(0, \sigma_w^2)$$

Y remplazando

$$y_i = \beta_0 + \beta_1 x_i + (u_i - \beta_1 w_i)$$

Entonces, el error está correlacionado con la explicativa observada,

$$E[x_i(u_i - \beta_1 w_i)] = E[(x_i^* + w_i)(u_i - \beta_1 w_i)] = -\beta_1 \sigma_w^2 \neq 0$$

Casos de **Endogeneidad : VARIABLES OMITIDAS???**

En la formulación de Wooldridge (simplificada)

$$y = \beta_1 x_1 + \gamma q + v$$

$$E(v \backslash x_1, q = 0)$$

Donde v es el "error estructural" pero remplazando,

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

$$\varepsilon = v + \gamma\, q$$

El MODELO es = al PGD *si y solo si q fuera* ***iid***

Lo es dado el supuesto sobre x_1 y el comportamiento de

$$q = \delta_1 x_1 + r$$

Casos de **Endogeneidad** : **VARIABLES OMITIDAS???**

Esta interpretación tendría dos problemas relacionados:

1) No distingue el modelo estimado (incorrecto)
   del PGD correcto (que contiene a la/las variables omitidas)
   ya que la variable omitida va al término del error poblacional !

2) Un error como ε implica además **otros problemas de especificación** si se adoptan procesos más generales y realistas **que suponer a las variables explicadas y explicativas como *iid*.** Por ejemplo, tendríamos un problema de autocorrelación serial superpuesto al de omisión de variables.

Casos de **Endogeneidad : VARIABLES OMITIDAS???**

En la vieja formulación relacionamos el coeficiente de una variable **estimado en un modelo incorrecto** (correspondiente a una "regresión corta"),

$$y_i = b\, x_i + e_i \qquad i = 1 \ldots N$$

porque que omite una o más variables que se encuentran en el PGD (correspondiente a la "regresión larga")

$$y_i = \beta\, x_i + \gamma q_i + v_i \qquad i = 1 \ldots N \qquad v_i \sim \boldsymbol{iid}(0, \sigma_v^2)$$

Donde es $e_i = y_i - b\, x_i$, el término residual (no el ε_i poblacional en la formulación de W) el que contiene a la/s variable/s omitida/s y donde b es estimado (por MCO) como

$$b = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i (\beta\, x_i + \gamma q + v_i)}{\sum x_i^2} = \beta + \gamma \frac{\sum x_i\, q_i}{\sum x_i^2} + \frac{\sum x_i\, v_i}{\sum x_i^2}$$

$$plim\, b = \beta + \gamma\delta$$

Un punto muy interesante que el sesgo en b puede derivarse incluso para variables supuestas como **no aleatorias**,

$$\mathrm{E}[\, b] = \beta + \gamma\delta$$

VARIABLES OMITIDAS y IV ?

- La propuesta de usar IV para Variables Omitidas iniciada en Angrist and Krueger (1991, quienes usan al trimestre de nacimiento como IV para estimar los beneficios económicos de la escolaridad), no dependería de cómo formulemos el problema.

- Para obtener el coeficiente de x_i en la *long regression* (el PGD) cuando estimamos la *short regression* "evitamos" usar x_i porque depende de q_i a través de δ (que es justamente lo que hacemos cuando estimamos esta ecuación por MCO) usando z_i.

- Esta variable z_i es un instrumento que, para cumplir su (nueva) función no debe depender de q_i ($\sum z_i\, q_i$=0)

- pero que obviamente debe explicar bien a x_i para no reemplazar un problema de inconsistencia por variables omitidas por uno de inconsistencia por instrumentos débiles (Bound, Jagger and Baker, 1995)

VARIABLES OMITIDAS e IVE ?

$$b_{iv} = \frac{\sum z_i y_i}{\sum z_i x_i} = \frac{\sum z_i(\beta x_i + \gamma q_i + v_i)}{\sum z_i x_i} = \beta + \gamma \frac{\sum z_i q_i}{\sum z_i x_i} + \frac{\sum z_i v_i}{\sum z_i x_i}$$

$$plim\, b = \beta$$

En consecuencia no sería necesario incluir el caso de caso de variables omitidas como de endogeneidad para la propuesta de utilizar IV

REFLEXIONES FINALES

En términos de la modelación econométrica

"**general a particular**" (*Gets*).

- La modelización econométrica debe lidiar con variables económicas i**nterdependientes** y que por la definición misma de modelo **debe tener** variables omitidas.

- Si vamos a modelar $\boldsymbol{y_i, x_{1i}, x_{2i}, \ldots x_{ki}}$ podemos "marginalizar" todo lo restante (las variables omitidas $\boldsymbol{x_{ji}}$ $\boldsymbol{j > k}$) para obtener estimadores (relativamente) consistentes de las variables que incluimos en nuestro modelo.

REFLEXIONES FINALES

- Si esto es así, luego nos podremos preguntar si

  en **el sistema** $y_i, x_{1i}, x_{2i}, \ldots x_{ki}$ son las x_{ji} exógenas (con distintos requisitos de acuerdo a nuestro objetivo en el uso del modelo y nuestros parámetros de interés) para poder simplificar nuestro modelo a uno condicional, tal vez instrumentado aquellas que no lo fuesen o estén mal medidas.

- "**Marginalizar**" y "**Condicionalizar**" son dos acciones distintas en la modelización econométrica que se asocian con superar los problemas de **"variables omitidas"** y **"variables endógenas"**, respectivamente.

¡Muchas gracias!