JAMES HECKMAN, EL SESGO DE SELECCIÓN MUESTRAL.

Cristina Sánchez Figueroa
Pedro Cortiñas Vázquez
Iñigo Tejera Martín
Senda del Rey
UNED. Madrid
csanchez@cee.uned.es, pcortinas@cee.uned.es, itejera@cee.uned.es

El objetivo de este trabajo es hacer una breve revisión del trabajo elaborado por James Heckman (1979) que tiene en cuenta el problema del sesgo de selección muestral. Este problema, esencial a lo hora de obtener conclusiones acertadas, no había sido estimado en los modelos tradicionales desarrollados hasta ese momento.

Es importante resaltar que la corrección en la selección muestral se puede realizar tanto en modelos con variable dependiente continua como en modelos con variable dependiente discreta. En este trabajo se considera únicamente el caso de variable dependiente continua, y se analizan los métodos de estimación que existen; considerando principalmente el método de Heckman en dos etapas.

Al presentarse el problema de la selección muestral los modelos de estimación deben recurrir, además de la ecuación objetivo que se pretende estimar, a una segunda ecuación que se le suele denominar ecuación de selección. La ecuación de selección corresponde a un modelo de variable dependiente discreta y mide la probabilidad de estar en la muestra. El ejemplo típico considerado por Heckman en su trabajo es el mercado laboral. En este caso las personas que trabajan son una submuestra de la población potencialmente activa, que puede trabajar.

1. Antecedentes.

Heckman, James J. (1944-). Economista norteamericano nacido en Chicago (Illinois) en 1944. Estudió matemáticas en la universidad de Colorado y después economía en la universidad de Princeton, donde, en 1971, consiguió el doctorado. En 1985 obtuvo el cargo de profesor de economía en la cátedra Henry Schultz, de la universidad de Chicago, donde se le elevó a catedrático con mención por servicios distinguidos en 1995.

Su labor docente se ha desarrollado en la universidad de Columbia, Yale y Chicago, y la ha compatibilizado con cargos en la oficina nacional norteamericana de investigación económica o en el centro de investigación económica del centro nacional norteamericano de investigación de opinión. También ha estado vinculado a diferentes organizaciones (como la academia nacional de las ciencias) y a publicaciones de carácter económico, y ha sido miembro de diversas comisiones de la academia nacional de las ciencias.

Heckman se ha especializado en el estudio *estadístico de la microeconomía*, es decir la parte de la economía que describe el comportamiento de los individuos, las familias y las empresas ante diversos incentivos de mercado y de gobierno. A esta combinación se le ha denominado microeconometría. Y surge como consecuencia de la creciente disponibilidad y accesibilidad de información de tipo individual proveniente de encuestas, así como del espectacular desarrollo de los medios de cálculo necesarios para procesar dicha información.

De esta manera, Heckman ha orientado sus análisis a temas relacionados con la economía laboral, como la decisión de aceptar un empleo, los ingresos del trabajo, la duración del desempleo, los programas gubernamentales para desempleados y menos capacitados, la fecundidad y la discriminación.

Utilizando como base estos análisis su principal contribución ha sido metodológica, al resolver problemas frecuentemente encontrados en la aplicación de la economía y conocidos en el lenguaje técnico como *"sesgo de selección" y "autoselección"*. Así, el problema del sesgo de selección fue desarrollado por James Heckman en su trabajo *Sample selection bias as a especification error (1979*). Hasta la publicación del trabajo este problema, de gran importancia, no era considerado en los análisis de los economistas, desconociendo a su vez que la corrección del mismo resulta fundamental para obtener conclusiones acertadas con estimadores insesgados, consistentes y eficientes sobre las características de la población o la muestra en estudio. Actualmente resulta ser uno de los problemas que más se deben tener en cuenta a la hora de la estimación de un modelo.

Fue galardonado con el Premio Novel de Economía en el año 2000. Este hecho supone un reconocimiento a sus trabajos, pioneros en la resolución de problemas que tienen relevancia desde el punto de vista social. La línea de investigación iniciada por Heckman ha generado numerosas contribuciones centradas en la estimación de los efectos de políticas activas del mercado de trabajo. Su trabajo ha permitido establecer ventajas e inconvenientes de utilizar datos experimentales en la evaluación de políticas públicas.

2. El sesgo de selección muestral.

El sesgo de selección muestral surge cuando las muestras a disposición de los investigadores no son "aleatorias", es decir no representan adecuadamente la población que se desea estudiar. Dentro de los sesgos de selección existen diferentes modalidades que pueden depender de los criterios del analista, de la decisión de los agentes económicos, etc. En base esto podemos decir que el propio analista, al decidir el diseño muestral, puede realizar una mala selección de los grupos que se comparan, o bien, lo que se puede dar es un problema de autoselección, cuando los individuos deciden autoselecionarse para pertenecer a un determinado grupo.

En primer lugar comentamos brevemente en qué consiste la autoselección, que correspondería al caso en el que la observabilidad de la variable dependiente está en función del valor que tome otra variable. El caso más típico, desarrollado por Heckman, es analizar cómo las muestras de participantes en el mercado laboral no son el resultado de una selección aleatoria sino de la autoselección de los individuos derivada de un proceso de maximización de utilidad. Al mismo tiempo diversos factores, como la educación, afectan al salario que puede conseguir un individuo en el mercado laboral. Si el estudio sólo considera a los individuos que trabajan y además están educados, se obtiene una muestra incompleta de la población ("autoseleccionada" de acuerdo con la decisión de las personas por educarse) lo que conduciría a conclusiones erróneas ("sesgadas") sobre el efecto de la educación. En este caso, al estar "sobrerepresentada" la población educada en la muestra, se tiende a subestimar el efecto de la educación.

Por otro lado los datos pueden no ser seleccionados de forma aleatoria por decisiones del propio analista. Un ejemplo son los estudios con datos de panel, una muestra será seleccionada por el analista si existe estabilidad en la unidad familiar durante varios periodos de análisis.

Para obtener estimaciones no sesgadas, debe considerarse este hecho. Por tanto, teniendo en cuenta este problema y la complejidad de las estimaciones, se han desarrollado métodos computacionales muy sencillos que son utilizados por la mayor parte de los investigadores a la hora de obtener conclusiones acertadas en los estudios realizados.

3. El Modelo de Heckman.

Ante la presencia de sesgo de selección, existen métodos de corrección que tienen como objetivo solucionar este problema. Para obtener estimaciones en modelos de variable dependiente continua, los métodos de corrección que se pueden utilizar son el propuesto por Heckman en 1979 y el método de Máxima Verosimilitud de Amemiya 1981 En este trabajo se considera principalmente el primero, aunque se hace un pequeño

apunte al de Máxima Verosimilitud, pues es uno de los más utilizados gracias al desarrollo de los programas computacionales.

El método propuesto por Heckman permite aislar el sesgo de selección muestral que se deriva de trabajar con modelos, ya sean de ingresos u horas de trabajo, de los individuos en el mercado laboral. Tal sesgo es producto de la autoselección de los individuos que deciden estar ocupados, de manera que cuando se utilizan los métodos clásicos –Mínimos Cuadrados Ordinarios (MCO), por ejemplo- los coeficientes obtenidos por este procedimiento están sesgados por el hecho de que la población ocupada constituye un segmento de la población total que paso por un proceso de autoselección para ingresar al mercado laboral. El procedimiento sugerido por Heckman para tratar con este tipo de problemas es conocido como el método bietapico.

El método consiste en estimar en un primer paso un modelo tipo probit para calcular la probabilidad (dadas ciertas variables de interés que determinen tal decisión) de que un individuo decida o no estar ocupado, de esta estimación se obtiene el estadístico conocido como la razón inversa de Mills que captura la magnitud de dicho sesgo. Posteriormente al calculo del modelo probit, la razón de Mills estimada se incorpora al modelo de regresión original (estimado por MCO) para ser añadido como un regresor más, de esta manera la significatividad de este coeficiente indica la magnitud de sesgo en que se incurriría si no se hubiese incorporado a la regresión explicativa de la desigualdad salarial. De esta manera, los coeficientes estimados por MCO añadiendo la variable λ ,que capta la magnitud del sesgo, son consistentes. Los estimadores obtenidos en por MCO con el método bietapico aunque consistentes, presentan problemas de eficiencia tal como demostró Maddala (1983). Este hecho hace que surja el método por máxima verosimilitud en el cual la estimación se realiza de manera conjunta.

1. Definición del modelo de Heckman.

$$y_{2i} = z_i\delta + v_{2i} \qquad \text{(a)}$$

$$y_{1i} = x_i\beta + u_{1i} \quad \text{si } y_{2i} > 0 \qquad \text{(b)}$$

y_{1i} no se observa si $y_{2i} \leq 0$

D_{2i} 1 si $y_{2i} > 0$

D_{2i} 0 si $y_{2i} \leq 0$

La ecuación para y_{1i} es una ecuación de regresión común. Sin embargo, bajo ciertas condiciones no observamos la variable dependiente de esta ecuación. Denotaremos si observamos o no esta variable mediante una variable dummy D_{2i} .

La observación de la variable dependiente y_{1i} es función del valor de otra regresión: la ecuación de selección que relaciona la variable latente y_{2i} con algunas características observadas z_i .

Para simplificar la exposición podemos decir que se consideran dos ecuaciones en el modelo, una ecuación de interés que corresponde a la ecuación que se busca estimar, de la que buscamos extraer conclusiones, y una ecuación de selección o participación (regresión auxiliar) que corresponde a un modelo de elección discreta (Probit o Logit), que mide la probabilidad de estar en la muestra, en esta última ecuación se pueden incluir las variables independientes de la ecuación de interés y a su vez esta ecuación deberá contener al menos una variable continua que sea determinante en el proceso de pertenecer o no a la muestra pero que a su vez no resulte relevante para determinar la variable dependiente, lo anterior con el fin de no caer en problemas de identificación.

Además, se asume la existencia de una distribución normal bivariada de los errores en las ecuaciones (a) y (b) con la siguiente estructura:

$$\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \approx N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix}\right]$$

De esta forma, la ecuación de selección se convierte en un modelo Probit. Por su parte, recordemos que la varianza de la distribución en la ecuación Probit puede ser normalizada a uno sin pérdida de información ya que la escala de la variable dependiente no es observada.

De esta manera, usando el supuesto de normalidad y las propiedades de la normal bivariada truncada podemos calcular:

$$\begin{aligned} E(y_1 / y_2 > 0) &= x\beta + E\left[v_1 / v_2 > -z\delta\right] \qquad \text{(c)} \\ &= x\beta + \rho\sigma_1 \lambda\left[\frac{-z\delta}{1}\right] \\ &= x\beta + \rho\sigma_1 \frac{\phi(-z\delta)}{1-\Phi(-z\delta)} \\ &= x\beta + \rho\sigma_1 \frac{\phi(z\delta)}{\Phi(z\delta)} \end{aligned}$$

Consideramos que la razón inversa de Mills siempre es positiva, la regresión de y sobre x está sesgada dependiendo del valor de ρ

Así la magnitud del sesgo dependerá de la magnitud de la correlación entre los errores (ρ), la varianza relativa del error (σ_1) y la severidad del truncamiento (la razón inversa de Mills es mayor cuando $z\delta$ es menor). Así, si ρ=0 entonces no habrás sesgo de selección

2. Estimación del modelo de Heckman.

Así utilizando la siguiente especificación:

$$E(y_1 / y_2 > 0) = x\beta + \rho\sigma_1 \frac{\phi(z\delta)}{\Phi(z\delta)}$$

El objetivo es estimar β en la ecuación (b) por MCO incluyendo en dicha ecuación la medida $\frac{\phi(z\delta)}{\Phi(z\delta)}$. Con este fin Heckman (1979) sugiere realizar los siguientes pasos:

1. Estimar δ consistentemente usando un probit para la probabilidad de observar los datos en función de z.
2. Calcular su valor ajustado para la función índice o variable latente $\hat{y}_{2i} = \hat{z}_i \delta$ y calcular la razón inversa de Mills $\hat{\lambda}_i$ como función de $\hat{y}_{2i}$.
3. Incluir $\hat{\lambda}_i$ en la regresión de y_{1i} sobre x_i para aproximar $\lambda(z_i\delta)$. El coeficiente de $\hat{\lambda}_i$ será una medida de $\rho\sigma_1$ y de esta forma una estimación de ρ y de σ_1 puede ser obtenida a partir de allí.

Los valores resultantes (estimadores) de $\beta, \rho \; y \; \sigma_1$ son consistentes pero asintóticamente ineficientes bajo el supuesto de normalidad. La gran importancia de este método es su sencillez , puesto que sólo se necesita realizar un probit y un MCO.

No obstante y una vez establecido este método, existen por lo menos tres aspectos que se deben considerar con respecto a este estimador en dos etapas:

1. El estimador del error estándar convencional en (a) es inconsistente pues el modelo de regresión en (c) es intrínsecamente heterocedástico debido a la selección. Una forma de solucionar esto es mediante el uso de los estimadores de los errores estándar robustos los cuales son, al menos consistentes.
2. El método no impone la condición que $|\rho| \leq 1$ lo cual está implícitamente asumido en el modelo. Esta condición es a menudo no respetada.
3. El supuesto de normalidad es necesario para la consistencia de los estimadores.

3. Estimación por Máxima Verosimilitud (ML).

En el método por máxima verosimilitud lo primero que debemos hacer es especificar el modelo, tal como hemos visto en las ecuaciones (a) y (b), para una vez especificado realizar la estimación de manera conjunta.

En este caso al considerar el sesgo de selección cada grupo tendrá diferente función de verosimilitud. Como tenemos dos tipos de observaciones:

1. Aquellas donde y_1 es observada para lo cual sabemos que se cumple que $y_2 > 0$. Para estas observaciones la función de verosimilitud es la probabilidad del evento y_1 y que también ocurra que $y_2 > 0$.

$$
\begin{aligned}
P(y_{1i}, y_{2i} > 0 / x, z) &= f(y_{1i}) \quad P(y_{2i} > 0 / y_{1i}, x, z) \\
&= f(v_{1i}) \quad P(v_{2i} > -z_i\delta / v_{1i}, x, z) \\
&= \frac{1}{\sigma_1}\phi\left(\frac{y_{1i} - x_i\beta}{\sigma_1}\right) \int_{-z_i\delta}^{\infty} f(v_{2i} / v_{1i})\, dv_{2i} \\
&= \frac{1}{\sigma_1}\left(\frac{y_{1i} - x_i\beta}{\sigma_1}\right) \int_{-z_i\delta}^{\infty} \phi\left[\frac{v_{2i} - \frac{\rho}{\sigma_1}(y_{1i} - x_i\beta)}{\sqrt{1-\rho^2}}\right] dv_{2i} \\
&= \frac{1}{\sigma_1}\phi\left(\frac{y_{1i} - x_i\beta}{\sigma_1}\right)\left[1 - \Phi\left(\frac{z_i\delta + \frac{\rho}{\sigma_1}(y_{1i} - x_i\beta)}{\sqrt{1-\rho^2}}\right)\right] \\
&= \frac{1}{\sigma_1}\phi\left(\frac{y_{1i} - x_i\beta}{\sigma_1}\right)\left[\Phi\left(\frac{z_i\delta + \frac{\rho}{\sigma_1}(y_{1i} - x_i\beta)}{\sqrt{1-\rho^2}}\right)\right]
\end{aligned}
$$

2. Aquellas donde y_1 no es observada para lo cual sabemos que se cumple que $y_2 \leq 0$ del manera, no tenemos información independiente para y_1.

$$P(y_2 \leq 0) = P(v_{2i} \leq -z_i\delta)$$
$$= \Phi(-z_i\delta)$$
$$= 1 - \Phi(z_i\delta)$$

De esta manera considerando la función de verosimilitud para todos los elementos de la muestra obtendríamos la siguiente expresión:

$$\log L(\beta, \delta, \rho, \sigma_1, datos) = \sum \log(1 - \Phi(z_i\delta)) +$$
$$\sum \left[-\log\sigma_1 + \log\phi\left(\frac{y_{1i} - x_i\beta}{\sigma_1}\right) + \log\Phi\left(\frac{z_i\delta + \frac{\rho}{\sigma_1}(y_{1i} - x_i\beta)}{\sqrt{1}}\right)\right]$$

Estos estimadores serán consistentes y asintóticamente eficientes bajo el supuesto de normalidad y homocedasticidad de los términos de error no censurados. Aunque unos de los problemas que tiene la estimación por ML es que la función no es estrictamente cóncava y en consecuencia no necesariamente existe una única solución.

BIBLIOGRAFIA.

- Gonzalez Espitia, Carlos G.Sesgo de selección muestral con STATA
- Gujarati (2010). Econometría. México. Mc Graw Hill.
- Heckman James J. (enero 1979) «Sample selection bias as a specification error»*Econometrica. Journal of the Econometric Society* (47): pp.153-161.
- MADDALA G.S. (1983), Limited Dependent and Qualitative Variables in Econometrics, Econometric. Society Monographs.
- Murray, M. (2006). Econometrics: a modern introduction. Ed. Pearson.
- Pérez L. César (2006). Problemas resueltos de econometría, Ed. Thomson.
- STATA CORP (2008). User's Guide, Reference Manual Release 10. Stata Press.
- • Wooldridge, J. (2006). Introducción a la Econometría. Un enfoque moderno. Ed.Thomson.