MODELOS DE SESGO DE SELECCIÓN

1. Introducción

1.1 Sesgo de Selección y Datos Faltantes[1]

El problema de selección surge en la estimación de modelos estructurales con resultados potenciales solo parcialmente observados. Pero el problema es más general y puede surgir siempre que se utilice esquemas de muestreo diferentes al aleatorio para obtener muestras de la población subyacente objeto del análisis.

Una muestra seleccionada por cualquier regla que no sea equivalente al muestreo aleatorio produce una descripción de características de la población que no describe adecuadamente las de las verdaderas, no importando el tamaño de la muestra. Debe observarse que las reglas de selección distorsionadas pueden surgir tanto de decisiones de diseño muestral, como de decisiones económicas de autoselección, lo cual da lugar a que los economistas solo observemos submuestras seleccionadas.

Existen dos caracterizaciones útiles del problema de selección:

a) Los estadísticos caracterizan la regla de muestreo como la aplicación de un ponderador a la distribución de una población hipotética para producir las distribuciones observadas.

b) Los econometristas lo ven como un problema de datos faltantes y, esencialmente, utilizan las variables observadas para imputar los valores de las variables no observadas.

En lo que se refiere al primer enfoque debe señalarse que cualquier modelo de sesgo de selección puede describirse en términos de distribuciones ponderadas. Sea Y variables a explicar, y **X** un vector de variables explicativas o de control. Sean **F(X,Y)** y ***f*(y,x)** la distribución poblacional y la densidad de (**Y,X**), respectivamente. Cualquier regla de muestreo es equivalente a una ponderación **w(Y,X)** que altera la densidad de la población. Sean **(y*,x*)** las variables aleatorias producidas por el muestreo. La densidad de los datos muestrales **w(y*,x*)** es:

$$g\,(y^*,x^*) = w\,(y^*,x^*)\,f(y^*,x^*) \,/\, (\int w(y^*,x^*)\,(y^*, x^*)\,dy^*dx^*) \qquad (\bullet)$$

[1] "La Microeconometría y la obra de James J. Heckman", Fernando Butler Silva, 2004

Muestreo Aleatorio Simple w(y*,x*)=1

Muestras Truncadas o Censuradas w(y*,x*)=0 para algunas observaciones.

Dada una muestra de datos es imposible el recuperar la verdadera densidad *f*(y,x) sin conocer la regla de ponderación. Si esta no se conoce es imposible (excepto bajo fuertes supuestos) el determinar el que si los datos faltantes en ciertos valores de la población (y,x) se debe a si el plan de muestreo, o a que la densidad poblacional no existe en dichos puntos.

Para observar el análisis econométrico del problema de selección veamos el siguiente ejemplo. Supongamos que tenemos tres posibles funciones de resultados:

$$Y_0 = g_0(X) + U_0$$
$$Y_1 = g_1(X) + U_1$$
$$Y_2 = g_2(X) + U_2$$

Donde las $\mathbf{Y_i}$ son variables latentes que solo pueden ser observadas imperfectamente. Recordemos que en las teorías neoclásicas de oferta laboral y de búsqueda, el salario de reserva (precio de reserva a cero horas de trabajo) juega un papel central, ya que nos dice el precio necesario para inducir al agente a ofrecer la primera hora de trabajo. Representemos la función de salario de reserva por $\mathbf{Y_0}$, y sea $\mathbf{Y_1}$ la función de salario de mercado. En ausencia de costos fijos, una persona trabaja **(D=1)** si:

$$Y_1 > Y_0 \Leftrightarrow D = 1$$

Las horas potenciales de trabajo $\mathbf{Y_2}$, también son generadas por las mismas preferencias que producen el salario de reserva. Los términos $\mathbf{U_0}$, $\mathbf{U_1}$, y $\mathbf{U_2}$, buscan tomar en cuenta las variables no observadas que explican porque personas observacionalmente idénticas (con la misma **X**) toman diferentes decisiones $\mathbf{Y_0}$ y $\mathbf{Y_1}$ son resultados potenciales, y $\mathbf{Y_2}$ es la utilidad latente, con reglas de decisión:

$$Y_2 > 0 \Leftrightarrow D = 1 \text{ e } Y_1 \text{ se observa}$$
$$Y_2 < 0 \Leftrightarrow D = 0 \text{ e } Y_0 \text{ se observa}$$

Por lo anterior, la Y observada es igual a:

$$Y = DY_1 + (1-D)Y_0$$

Sean Z variables que afectan las elecciones, y X variables que afectan los resultados.

Entonces las Y esperadas son:

$$\begin{aligned} E(Y \mid X, Z, D = 1) &= E(Y_1 \mid X, Z, D = 1) \\ &= \mu_1(X) + E(U_1 \mid X, Z, D = 1) \\ E(Y \mid X, Z, D = 0) &= E(Y_0 \mid X, Z, D = 0) \\ &= \mu_0(X) + E(U_0 \mid X, Z, D = 0) \end{aligned}$$

Las medias condicionales de $\mathbf{U}_0$ y $\mathbf{U}_1$ son las funciones de control o sesgo (Heckman 1980). Como podemos observar, las medias de los resultados observados son generadas por los resultados potenciales y un término de sesgo. Definamos **P(z)=Pr(D=1 / Z=z)**. Debido a las reglas de decisión de este modelo, se puede demostrar (Heckman 1980) que, bajo ciertas condiciones generales, las medias de los resultados se pueden expresar como:

$$\begin{aligned} E(Y \mid X, Z, D = 1) &= \mu_1(X) + K_1(P(Z)) \\ E(Y \mid X, Z, D = 0) &= \mu_0(X) + K_0(P(Z)) \end{aligned}$$

Donde $\mathbf{K_1(P(Z))}$ y $\mathbf{K_0(P(Z))}$ son funciones de control que dependen de Z exclusivamente a través de **P**. El valor de **P** se relaciona con la magnitud del sesgo de selección.
Conforme las muestras son más representativas, $\mathbf{P(z) = 1}$ y $\mathbf{K_1(P) = 0}$, ya que la probabilidad de que cualquier tipo de persona este incluida en la muestra es la misma **(P=1)**.

En general, regresiones en muestras que experimentan sesgo de selección, producirán resultados sesgados para $\mu_1(X)$.

1.2 Aplicaciones a la Economía Laboral

El marco analítico esbozado permite analizar en forma consistente fenómenos microeconómicos con muestras de datos con problemas de auto-selección, y ha sido aplicado a una amplia variedad de problemas económicos, además del análisis de la oferta laboral, cambiando radicalmente la forma en que interpretamos los datos económicos y sociales, y evaluamos la efectividad de las políticas sociales.

Por ejemplo: ¿Existe mejoras en la situación económica de la población negra norteamericana dadas las políticas sociales de acción afirmativa? Como podemos observar el cociente de los salarios de la población negra a la blanca se incremento desde 1940 a 1980, y a partir de ese momento se ha estabilizado. Sin embargo, en dicho periodo los negros se han estado retirando de la fuerza laboral (disminución de **P(Z)**). Tomando en cuenta esta retirada selectiva de la fuerza laboral de los trabajadores negros de bajos salarios, el progreso económico de la población negra con respecto a la blanca se reduce virtualmente a cero (Heckman, Lyons y Todd 2000).

¿Existe mejores salarios y menos desigualdad en los mercados laborales de Europa comparados con los de Estados Unidos? Al respecto, debe señalarse que en Europa los desempleados no están tomados en cuenta en el cálculo de los salarios, y que esta práctica subestima la desigualdad salarial y sobrestima los niveles salariales de la población al tomar en cuenta solo a los empleados (Blundell, Reed y Stoker 1999).

Adicionalmente, en Europa se sobrestima el crecimiento de los salarios, ya que si se toma en cuenta que la proporción de personas trabajando ha declinado, se reduce drásticamente el nivel y la tasa de crecimiento de los salarios reales (Bils 1985).

1.3 Aplicaciones a la Evaluación de Políticas Sociales

¿Cuales son los principios básicos para evaluar de manera correcta una política social?

Preguntas tales como: ¿Debemos aumentar los impuestos?, ¿Cuáles serían los efectos?, ¿Quiénes soportarían los gravámenes?, ¿Debemos subsidiar la educación?, ¿Quiénes resultan beneficiados y quienes resultan perdedores?, ¿Debemos reubicar a los trabajadores que han sido desplazados ó a aquellos con bajo nivel de habilidades?, son preguntas rutinarias que exigen respuesta. El trabajo empírico para responder a estas preguntas desempeña un papel muy importante cuando esta basado en la teoría económica para cuantificar los beneficios y los costos, y su distribución. Esto no necesariamente resolverá todos los conflictos, aunque si los reducirá mucho y proporcionara una orientación con mayores bases. Un sistema de evaluación objetiva, evita conflictos y propicia que la discusión pública quede por encima de discusiones políticas.

J. Heckman señala que los elementos para un esquema de evaluación exitosa son:

a) Establecer los diversos criterios de interés en los que la sociedad no esté necesariamente de acuerdo, y que no resulten fáciles de estimar,

b) Crear un sistema bien diseñado de recolección de datos que produzca estimaciones confiables, ya que los datos objetivos reducen los elementos subjetivos en el diseño de decisiones políticas, y

c) Establecer métodos para evaluar los resultados de políticas hipotéticas que hagan explícitos todos los supuestos.

Por ejemplo, en su articulo “Legislación Laboral y Empleo: Lecciones de América Latina y el Caribe”, Heckman señala que los estudios del mercado laboral en los países en desarrollo están afectados por problemas de los datos. Por esta razón, las variables laborales contenidas en la bases de datos de secciones cruzadas de diferentes países sufren de falta de comparabilidad y credibilidad. Intentando superar estos problemas, Heckman y Pagés analizaron un nuevo grupo de información que incluye los países de la OCDE y de AL. Concluyen que:

a) Los beneficios de los programas financiados con una contribución nominal obligatoria, deben ser ponderados contra de sus costos en términos de empleo. El fortalecer el vínculo entre los pagos y los beneficios, contribuye a cambiar los costos de dichos programas hacia los trabajadores (en menos en el largo plazo). También se debe

tomar en cuenta la capacidad de forzar el cumplimiento de la regulación. Cuando esta capacidad es limitada, la distribución de los costos de los beneficios obligatorios esta distribuido desigualmente entre los trabajadores. Los trabajadores jóvenes, con poca educación, y aquellos que laboran en el campo tienen menos probabilidades de estar cubiertos por la seguridad social que los trabajadores mayores, capacitados y urbanos. Si estas desigualdades reflejan las posibilidades de evasión de las empresas, en lugar de las preferencias de los trabajadores por cobertura, estas aumentan las ya considerables desigualdades en las ganancias entre los diferentes trabajadores en la región.

b) Las regulaciones del mercado laboral pueden ser un mecanismo que aumente la desigualdad. Estas tienden a aumentar la desigualdad porque mientras unos trabajadores se benefician otros salen perjudicados. El impacto de la desigualdad es multifacético. Las regulaciones del mercado laboral aumentan las desigualdades porque reducen las perspectivas de empleo de grupos particulares de trabajadores como los trabajadores jóvenes, del sexo femenino, y los no capacitados. Dichas regulaciones también aumentan la desigualdad si, como la evidencia recolectada aquí sugiere, ellas incrementan el tamaño del sector informal.

Por otra parte, debe recordarse que para evaluar la eficiencia de un Programa Social no es válido comparar la situación de las personas que participan en el programa contra la de aquellas que no participaron, sino que se debe comparar el resultado de alguien que participo en un programa contra el resultado que el mismo agentes hubiera obtenido si no hubiera participado.

Por ejemplo, en un programa social de capacitación laboral puede esperarse que las personas participantes en el programa tengan mayor motivación y habilidades (variables no observadas) que las no participantes, y por lo tanto obtengan un mayor salario tomen o no el entrenamiento laboral. Por lo anterior, no es válido comparar los aumentos observados en el ingreso entre los participantes en el programa con el logrado por aquellos que no estuvieron registrados, y decir que este fue el resultado del programa, sino que se debe controlar el efecto en los diferenciales de los ingresos de las variables no observadas. Dependiendo del caso, el empleo de métodos de asignación aleatoria o de modelos de selección en el diseño del programa pueden solucionar este problema.

¿Cuáles son los efectos sobre el bienestar de la población chilena del programa social Chile Solidario? Un análisis teórico y empírico puede considerar dos de los más populares métodos de evaluación de políticas sociales: los modelos de selección y matching, y señalar los supuestos que cada método necesita, estableciendo las situaciones en las que un método es apropiado.

1.4 Econometría Estructural y Econometría de Efectos de Tratamiento

En el enfoque de Heckman se plantea la disyuntiva de buscar estimar parámetros estructurales inicialmente vs. estimar principalmente los efectos de tratamiento. ¿Cuál es la diferencia crucial entre ellos?

En relación a la estimación de efectos de programas, la estimación estructural busca contestar cuál es el efecto probable de un nuevo programa o de un viejo programa aplicado en un nuevo ambiente. Por su lado, la estimación de efectos de tratamiento busca exclusivamente aislar cuál es el efecto de un programa sobre los participantes y no participantes, comparado con la situación en la cual no esta presente el programa, o se encuentra presente un programa alternativo.

La econometría de efectos estructurales es más ambiciosa que la econometría de efectos de tratamiento en el sentido de que los objetivos de la estimación estructural es proveer ingredientes para resolver varios problemas de decisión:

a) Evaluar la eficacia de una política.
b) Proyectar la eficacia de una política a ambientes diferentes de aquel donde se desarrolla.
c) Proyectar los efectos de una nueva política no probada con anterioridad.

Sin embargo, para ciertos problemas de decisión importantes no es necesario el conocimiento de todos (o inclusive de alguno) de los parámetros estructurales de un modelo. La literatura moderna sobre los efectos del tratamiento tiene como principal objetivo la estimación de algunos parámetros de los efectos de tratamiento, y no del rango completo de parámetros buscados por la econometría estructural.

Al concentrarse en un problema particular, la econometría de efectos del tratamiento alcanza sus objetivos bajo condiciones más débiles y más creíbles que las utilizadas en la econometría estructural. Sin embargo, los parámetros así generados son más difíciles de trasladar a ambientes diferentes para estimar los efectos de una política preexistente o de una política nueva.

Para ver las diferencias entre ambos enfoques, utilicemos el siguiente ejemplo acerca de la estimación de los efectos de los impuestos sobre la oferta laboral. Sea **H** la función Marshalliana de oferta laboral (de horas de trabajo), en función de los salarios **W**, y de un vector de otras variables explicativas **X**. Sea **U** un vector de variables no observables por el economista que analiza los datos, el cual puede entrar en forma lineal o no-lineal:

$H = \Phi\ (\mathbf{W, X, U})$
$H = \Phi\ (\mathbf{W, X}) + \mathbf{U}, \quad E(U)=0$

Al introducir parámetros se busca reducir la dimensionalidad del problema de la identificación de una función de infinitas dimensiones a la de identificar un conjunto finito de parámetros.

$H = \Phi(\mathbf{W, X, U,}\ \theta\ ),\ \ \theta$ = vector de parámetros
$H = \Phi(\mathbf{W, X,}\ \theta\ ) + \mathbf{U}$

Por ejemplo, una representación lineal en parámetros muy utilizada en la economía laboral es la log-lineal:

$H = \alpha\ ' \mathbf{X} + \beta\ \mathbf{ln\ W} + \mathbf{U}$

Podemos observar que la econometría estándar se concentra en modelos lineales en parámetros para los diferentes agentes presentes:

$$\mathbf{Y_i = X_i}\ \beta\ + \mathbf{U_i}, \qquad i = 1,2,\ldots,N$$

donde $\mathbf{E(U_i) = 0}$, y analiza los problemas que surgen cuando $\mathbf{E(U_i \mid X_i) \neq 0}$. En la econometría estándar, la heterogeneidad entre los individuos es modelada como heterogeneidad en los interceptos. Sin embargo, toda la experiencia empírica producto de las investigaciones microeconómetricas apoya una versión más general de heterogeneidad de los individuos tanto en pendientes como en interceptos:

$$\mathbf{Y_i = X_i}\ \beta_i + \mathbf{U_i}$$

Sea $\mathbf{b^* = E(b_i)}$ y $\mathbf{v_i = b_i - b^*}$. Gran parte de la investigación contemporánea en microeconometría señala la existencia de correlación entre las $\mathbf{X_i}$ y los $\mathbf{b_i}$, además de que $\mathbf{E(U_i \mid X_i) \neq 0}$ y $\mathbf{E(v_i / X_i) \neq 0}$ (Carneiro, Heckman y Vytlacil 2001, Heckman y Vytlacil 2001). Este modelo de coeficientes aleatorios correlacionados surge de manera natural, por ejemplo, en el análisis microeconómetrico de los retornos económicos a la educación. Si la escolaridad es **X** e **Y** es el logaritmo de las ganancias, el coficiente de **b** asociado con la escolaridad es la tasa de retorno de la educación. Esta tasa de retorno generalmente varía entre los individuos y parece estar fuertemente correlacionada con los diferentes niveles de escolaridad.

Dada la complejidad de estimar modelos estructurales con heterogeneidad en pendientes e interceptos, los microeconometristas han buscado métodos más simples para responder preguntas cuidadosamente focalizadas en los efectos de una política social especifica en una población especifica, en lugar de tratar de solucionar la gama entera de problemas que pueden ser abordados utilizando la estimación estructural.

Continuando con nuestro ejemplo de oferta laboral, supongamos que los agentes perciben correctamente el impuesto, y que no importan los efectos de equilibrio general de la modificación del impuesto. En el lenguaje de efectos de tratamiento, el efecto del tratamiento o "efecto causal" de un cambio impositivo de la oferta laboral definida a un nivel individual es $\mathbf{H = F(W(1-t_j),X,U) - F(W(1-t_k),X,U)}$, para la misma persona sujeta a dos diferentes impuestos $\mathbf{t_j}$ y $\mathbf{t_k}$.

En este contexto, la econometría de efectos de tratamiento busca identificar las diferencias en las horas trabajadas promedio de una población dada **(X,W,U)** que surgen de diferentes políticas impuestas externamente (**t**), sin que se necesite descomponer las horas promedio en **F** o **G** [2] utilizando datos de poblaciones en las cuales **t** no es impuesta en forma exógena.

Los experimentos que cambian **t** pero que no cambian **F** o **G** (o las mantienen invariantes) identifican estos efectos. Para ello, la literatura de efectos de tratamiento utiliza diferentes métodos para controlar las diferencias en los resultados observados o no observados en diferentes regimenes, cuando estas diferencias no están relacionadas a los efectos de la política que se busca evaluar.

Las condiciones requeridas para estimar efectos de tratamiento son generalmente más laxas que las necesarias para identificar **F** o **G**, en el sentido de que se necesitan menos supuestos para identificar los efectos de tratamiento. Sin embargo, a diferencia de los resultados de la estimación estructural de **F**, los efectos de tratamiento no se pueden transportar a nuevos ambientes, ni tampoco se pueden interpretar en términos de efectos causales de todas las variables condicionantes con excepción de **t**. Por otra parte, dado que la literatura de efectos de tratamiento investiga los efectos de políticas de aplicación parcial, donde existen grupos de tratamiento y de control, esta no es útil para evaluar políticas de aplicación general que afecten a todos los individuos, a menos que existan datos de economías separadas experimentando diferentes políticas, y dichas economías estén aisladas una de la otra.

[2] F y G distribuciones asociadas a *f* y *g* en la expresión (•), resp.

2. Sample Selection Models

February 1999 (revised Nov. 2000; Feb. 2002)

We observe data (X, Z) on N individuals or firms. For a subset $N_1 = N - N_0$ of the observations, we also observe a dependent variable of interest, y_1 but this variable is unobserved for the remaining N_0 observations. The following model describes our estimation problem:

$$\begin{aligned} y_{1i} &= X_i\beta + \nu_{1i} \qquad \text{if } y_{2i} > 0 \qquad (1)\\ y_{1i} &= \text{not observed} \qquad \text{if } y_{2i} \leq 0 \\ y_{2i} &= Z_i\delta + \nu_{2i} \qquad (2)\\ D_{2i} &= 1 \qquad \text{if } y_{2i} > 0 \\ D_{2i} &= 0 \qquad \text{if } y_{2i} \leq 0 \end{aligned}$$

The equation for y_{1i} is an ordinary regression equation. However, under some conditions we do not observe the dependent variable for this equation; we denote whether or not we observe its value by a dummy variable D_{2i}. Observation of the dependent variable y_{1i} is a function of the value of another regression equation (the selection equation, which relates a latent variable y_{2i} to some observed characteristics Z_i). The variables in X_i and Z_i may overlap; if they are identical this will create problems for identification in some cases (see the discussion below).

Examples are married women's labor supply (where the first equation is the hours equation and the second equation is an equation for the difference between the market and the unobserved reservation wage) and the firm size and growth relationship (where the first equation is the relation between growth and size and the second equation describes the probability of exit between the first and second periods).

2.1 Bias Analysis

Suppose that we estimate the regression given in equation (1) by ordinary least squares, using only the observed data. We regress y_{1i} on X_i, using $i = N_0 + 1, ..., N$ observations.When are

the estimates of β obtained in this way likely to be biased? We can analyze this question without assuming a specific distribution for the νs. Compute the conditional expectation of y_1 given X and the probability that y_1 is observed:

$$E[y_1|X, y_2 > 0] = X\beta + E[\nu_1|\nu_2 > -Z\delta]$$

From this expression we can see immediately that the estimated β will be unbiased when ν_1 is independent of ν_2 (that is, $E[\nu_1|\nu_2] = 0$), so that the data are missing "randomly," or the selection process is "ignorable." That is the simplest (but least interesting) case.

Now assume that ν_1 and ν_2 are jointly distributed with distribution function $f(\nu_1, \nu_2; \theta)$ where θ is a finite set of parameters (for example, the mean, variance, and correlation of the random variables). Then we can write (by Bayes rule)

$$E[\nu_1|\nu_2 > -Z_i\delta] = \frac{\int_{-\infty}^{\infty}\int_{-Z_i\delta}^{\infty} \nu_1 f(\nu_1, \nu_2; \theta) d\nu_2 d\nu_1}{\int_{-\infty}^{\infty}\int_{-Z_i\delta}^{\infty} f(\nu_1, \nu_2; \theta) d\nu_2 d\nu_1} = \lambda(Z\delta; \theta) \tag{3}$$

$\lambda(Z\delta; \theta)$ is a (possibly) nonlinear function of $Z\delta$ and the parameters θ. That is, in general the conditional expectation of y_1 given X and the probability that y_1 is observed will be equal to the usual regression function $X\beta$ plus a nonlinear function of the selection equation regressors Z that has a non-zero mean.[1] This has two implications for the estimated βs:

1. The estimated intercept will be biased because the mean of the disturbance is not zero. (In fact, it is equal to $E_i[\lambda(Z_i\delta; \theta)]$).

2. If the Xs and the Zs are not completely independently distributed (i.e., they have variables in common, or they are correlated), the estimated slope coefficients will be biased because there is an omitted variable in the regression, namely the $\lambda(Z_i\delta; \theta)$, that is correlated with the included variables X.

Note that even if the Xs and the Zs are independent, the fact that the data is nonrandomly missing will introduce heteroskedasticity into the error term, so ordinary least squares is not fully efficient (*Why?*[2]).

This framework suggests a semi-parametric estimator of the sample selection model, although few researchers have implemented it (see Powell, *Handbook of Econometrics*, Volume IV, for more discussion of this approach). Briefly, the method would have the following steps:

[1] Although I will not supply a proof, only for very special cases will this term be mean zero. For example, in the case of bivariate distributions with unconditional means equal to zero, it is easy to show that $\lambda(.)$ has a nonzero mean unless the two random variables are independent.

[2] Questions in italics throughout these notes are exercises for the interested reader.

1. Estimate the probability of observing the data (equation (**??**)) using a semi-parametric estimator for the binary choice model (*why are these estimates consistent even though they are single equation?*).

2. Compute a fitted value of the index function $\widehat{y}_{2i} = Z_i\widehat{\delta}$.

3. Include powers of $\widehat{y}_{2i}$ in a regression of y_{1i} on X_i to proxy for $\lambda(Z_i\delta;\theta)$. It is not clear how many to include.

Note that it is very important that there be variables in Z_i that are distinct from the variables in X_i for this approach to work, otherwise the regression will be highly collinear. Note also that the propensity score approach of Rubin and others is related to this method: it uses intervals of $\widehat{y}_{2i}$ to proxy for $\lambda(Z_i\delta;\theta)$, interacting them with the X variable of interest (the treatment).

2.3 Heckman Estimator

Semi-parametric estimation can be difficult to do and has very substantial data requirements for identification and for the validity of finite sample results. Therefore most applied researchers continue to estimate sample selection models using a parametric model. The easiest to apply in the case of sample selection is the bivariate normal model, in which case the selection equation becomes the usual Probit model. There are two approaches to estimating the sample selection model under the bivariate normality assumption: the famous two-step procedure of Heckman (1979) and full maximum likelihood. I will discuss each of these in turn. Although ML estimation is generally to be preferred for reasons discussed below, the Heckman approach provides a useful way to explore the problem.

The Heckman method starts with equation (3) and assumes the following joint distribution for the νs:

$$\begin{pmatrix} v_1 \\ \nu_2 \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho\sigma_1 \\ \rho\sigma_1 & 1 \end{pmatrix}\right] \tag{4}$$

where N denotes the normal distribution. Recall that the variance of the distribution in a Probit equation can be normalized to equal one without loss of generality because the scale of the dependent variable is not observed. Using the assumption of normality and the results in the Appendix on the truncated bivariate normal, we can now calculate $E[y_1|y_2 > 0]$.

$$\begin{aligned} E[y_1|y_2 > 0] &= X\beta + E[\nu_1|\nu_2 > -Z\delta] = X\beta + \rho\sigma_1\lambda(\frac{-Z\delta}{1}) \\ &= X\beta + \rho\sigma_1\frac{\phi(-Z\delta)}{1-\Phi(-Z\delta)} = X\beta + \rho\sigma_1\frac{\phi(Z\delta)}{\Phi(Z\delta)} \end{aligned} \tag{5}$$

Let's interpret this equation. It says that the regression line for y on X will be biased upward when ρ is positive and downward when ρ is negative, since the inverse Mills ratio is always positive (see the Appendix). The size of the bias depends on the magnitude of the correlation, the relative variance of the disturbance (σ_1), and the severity of the truncation (the inverse Mills ratio is larger when the cutoff value $Z\delta$ is smaller – see the figure in the Appendix). Note that when ρ is zero there is no bias, as before.[3]

Also note that the simple Tobit model, where y_1 and y_2 conincide and ρ is therefore one, can be analyzed in the same way, yielding

$$E[y_1|y_1 > 0] = X\beta + \sigma_1 \frac{\phi(X\beta)}{\Phi(X\beta)}$$

In this case, because the second term is a monotonic declining function of $X\beta$, it is easy to see that the regression slope will be biased downward (*Why?*).

2.3.1 Estimation using Heckman's Method

Equation (5) suggests a way to estimate the sample selection model using regression methods. As in the semi-parametric case outlined above, we can estimate β consistently by including a measure of $\phi(Z\delta)/\Phi(Z\delta)$ in the equation. Heckman (1979, 1974?) suggests the following method:

1. Estimate δ consistently using a Probit model of the probability of observing the data as a function of the regressors Z.

2. Compute a fitted value of the index function or latent variable $\widehat{y}_{2i} = Z_i\widehat{\delta}$; then compute the inverse Mills ratio $\widehat{\lambda}_i$ as a function of $\widehat{y}_{2i}$.

3. Include $\widehat{\lambda}_i$ in a regression of y_{1i} on X_i to proxy for $\lambda(Z_i\delta)$. The coefficient of $\widehat{\lambda}_i$ will be a measure of $\rho\sigma_1$ and the estimated ρ and σ_1 can be derived from this coefficient and the estimated variance of the disturbance (which is a function of both due to the sample selection; see Heckman for details).

The resultant estimates of β, ρ, and σ_1 are consistent but not asymptotically efficient under the normality assumption. This method has been widely applied in empirical work because of its relative ease of use, as it requires only a Probit estimation followed by least squares, something which is available in many statistical packages. However, it has at least three (related) drawbacks:

[3] In the normal case, $\rho = 0$ is equivalent to the independence result for the general distribution function.

1. The conventional standard error estimates are inconsistent because the regression model in step (3) is intrinsically heteroskedastic due to the selection. (*What is* $Var(\nu_1|\nu_2 > 0)$*?*) One possible solution to this problem is to compute robust (Eicker-White) standard error estimates, which will at least be consistent.

2. The method does not impose the constraint $|\rho| \leq 1$ that is implied by the underlying model (ρ is a correlation coefficient). In practice, this constraint is often violated.

3. The normality assumption is necessary for consistency, so the estimator is no more robust than full maximum likelihood – it requires the same level of restrictive assumptions but is not as efficient.

For these reasons and because full maximum likelihood methods are now readily available, it is usually better to estimate this model using maximum likelihood if you are willing to make the normal distributional assumption. The alternative more robust estimator that does not require the normal assumption is described briefly in Section 2. The ML estimator is described in the next section.

2.4 Maximum Likelihood

Assuming that you have access to software that will maximize a likelihood function with respect to a vector of parameters given some data, the biggest challenge in estimating qualitative dependent variable models is setting up the (log) likelihood function. This section gives a suggested outline of how to proceed, using the sample selection model as an example.[4]

Begin by specifying a complete model as we did in equations (1) and (**??**). Include a complete specification of the distribution of the random variables in the model such as equation (4). Then divide the observations into groups according to the type of data observed. Each group of observations will have a different form for the likelihood. For example, for the sample selection model, there are two types of observation:

1. Those where y_1 is observed and we know that $y_2 > 0$. For these observations, the likelihood function is the probability of the joint event y_1 and $y_2 > 0$. We can write

[4] Several software packages, including TSP, provide the sample selection (generalized Tobit) model as a canned estimation option. However, it is useful to know how to construct this likelihood directly, because often the model you wish to estimate will be different from the simple 2 equation setup of the canned program. Knowing how to construct the likelihood function allows you to specifiy an arbitrary model the incorporates observed and latent variables.

this probability for the ith observation as the following (using Bayes Rule):

$$\begin{aligned}
\Pr(y_{1i}, y_{2i} &> 0|X,Z) = f(y_{1i}) \Pr(y_{2i} > 0|y_{1i}, X, Z) = f(\nu_{1i}) \Pr(\nu_{2i} > -Z_i\delta|\nu_{1i}, X, Z) \\
&= \frac{1}{\sigma_1}\phi\left(\frac{y_{1i} - X_i\beta}{\sigma_1}\right) \cdot \int_{-Z_i\delta}^{\infty} f(\nu_{2i}|\nu_{1i}) d\nu_{2i} \\
&= \frac{1}{\sigma_1}\phi\left(\frac{y_{1i} - X_i\beta}{\sigma_1}\right) \cdot \int_{-Z_i\delta}^{\infty} \phi\left(\frac{\nu_{2i} - \frac{\rho}{\sigma_1}(y_{1i} - X_i\beta)}{\sqrt{1-\rho^2}}\right) d\nu_{2i} \\
&= \frac{1}{\sigma_1}\phi\left(\frac{y_{1i} - X_i\beta}{\sigma_1}\right) \cdot \left[1 - \Phi\left(\frac{-Z_i\delta - \frac{\rho}{\sigma_1}(y_{1i} - X_i\beta)}{\sqrt{1-\rho^2}}\right)\right] \\
&= \frac{1}{\sigma_1}\phi\left(\frac{y_{1i} - X_i\beta}{\sigma_1}\right) \cdot \Phi\left(\frac{Z_i\delta + \frac{\rho}{\sigma_1}(y_{1i} - X_i\beta)}{\sqrt{1-\rho^2}}\right)
\end{aligned}$$

where we have used the conditional distribution function for the normal distribution given in the appendix to go from the second line to the third line. Thus the probability of an observation for which we see the data is the density function at the point y_1 multiplied by the conditional probability distribution for y_2 given the value of y_1 that was observed.

2. Those where y_1 is not observed and we know that $y_2 \leq 0$. For these observations, the likelihood function is just the marginal probability that $y_2 \leq 0$. We have no independent information on y_1. This probability is written as

$$\Pr(y_{2i} \leq 0) = \Pr(\nu_{2i} \leq -Z_i\delta) = \Phi(-Z_i\delta) = 1 - \Phi(Z_i\delta)$$

Therefore the log likelihood for the complete sample of observations is the following:

$$\begin{aligned}
\log L(\beta, \delta, \rho, \sigma; thedata) &= \sum_{i=1}^{N_0} \log\left[1 - \Phi(Z_i\delta)\right] \\
+ \sum_{i=N_0+1}^{N} &\left[-\log\sigma_1 + \log\phi\left(\frac{y_{1i} - X_i\beta}{\sigma_1}\right) + \log\Phi\left(\frac{Z_i\delta + \frac{\rho}{\sigma_1}(y_{1i} - X_i\beta)}{\sqrt{1-\rho^2}}\right)\right]
\end{aligned}$$

where there are N_0 observations where we don't see y_1 and N_1 observations where we do ($N_0 + N_1 = N$). The parameter estimates for the sample selection model can be obtained by maximizing this likelihood function with respect to its arguments. These estimates will be consistent and asymptotically efficient under the assumption of normality and homoskedasticity of the uncensored disturbances. Unfortunately, they will no longer be even consistent if

these assumptions fail. Specification tests of the model are available to check the assumptions (see Hall (1987) and the references therein).

One problem with estimation of the sample selection model should be noted: this likelihood is not necessarily globally concave in ρ, although the likelihhod can be written in a globally concave manner conditional on ρ. The implication is that a gradient maximization method may not find the global maximum in a finite sample. It is therefore sometimes a good idea to estimate the model by searching over $\rho \subset (-1, 1)$ and choosing the global maximum.[5]

2.5 Ditributions

2.5.1 Truncated Normal Distribution

Define the standard normal density and cumulative distribution functions ($y \sim N(0,1)$):

$$\begin{aligned} \phi(y) &= \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}y^2) \\ \Phi(y) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \exp(-\frac{1}{2}u^2)du \end{aligned}$$

Then if a normal random variable y has mean μ and variance σ^2, we can write its distribution in terms of the standard normal distribution in the following way:

$$\begin{aligned} \phi(y;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2}(\frac{y-\mu}{\sigma})^2) = \frac{1}{\sigma}\phi(\frac{y-\mu}{\sigma}) \\ \Phi(y;\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} \frac{1}{\sigma} \exp(-\frac{1}{2}(\frac{u-\mu}{\sigma})^2)du = \Phi(\frac{y-\mu}{\sigma}) \end{aligned}$$

The truncated normal distribution of a random variable y with mean zero is defined as

$$E[y|y \geq c] = \frac{\frac{1}{\sigma}\int_c^\infty u\phi(u/\sigma)du}{\frac{1}{\sigma}\int_c^\infty \phi(u/\sigma)du} = \frac{\phi(c)}{1-\Phi(c)} = \frac{\phi(-c)}{\Phi(-c)}$$

(Can you demonstrate this result?)

[5] TSP 4.5 and later versions perform the estimation of this model by searching on ρ and then choosing the best value.

2.5.2 Truncated bivariate normal

Now assume that the joint distribution of x and y is bivariate normal:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}\right]$$

One of the many advantages of the normal distribution is that the conditional distribution is also normal:

$$f(y|x) = N\left(\mu_y + \frac{\rho\sigma_x\sigma_y}{\sigma_x^2}(x-\mu_x), \sigma_y^2(1-\rho^2)\right) = \phi\left(\frac{y - \mu_y - \frac{\rho\sigma_x\sigma_y}{\sigma_x^2}(x-\mu_x)}{\sigma_y\sqrt{1-\rho^2}}\right)$$

That is, the conditional distribution of y given x is normal with a higher mean when x and y are positively correlated and x is higher than its mean, and lower mean when x and y are negatively correlated and x is higher than its mean. The reverse holds when x is lower than its mean. In general, y given x has a smaller variance than the unconditional distribution of y, regardless of the correlation of x and y.

Using this result, one can show that the conditional expectation of y, conditioned on x greater than a certain value, takes the following form:

$$E[y|x > a] = \mu_y + \rho\sigma_y\lambda(\frac{a-\mu_x}{\sigma_x})$$

where

$$\lambda(u) = \frac{\phi(u)}{1-\Phi(u)} = \frac{\phi(-u)}{\Phi(-u)}$$

The expression $\lambda(u)$ is sometimes known as the inverse Mills' ratio. It is the hazard rate for x evaluated at a. Here is a plot of the hazard rate as a function of $-u$. It is a monotonic function that begins at zero (when the argument is minus infinity) and asymptotes at infinity (when the argument is plus infinity):

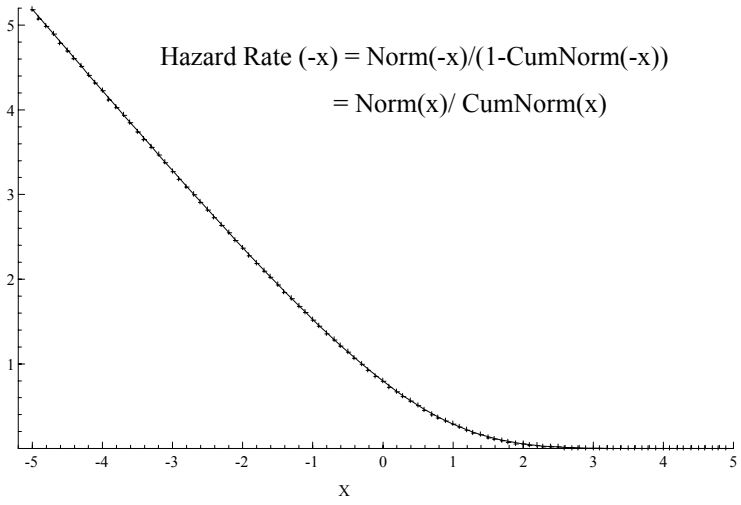
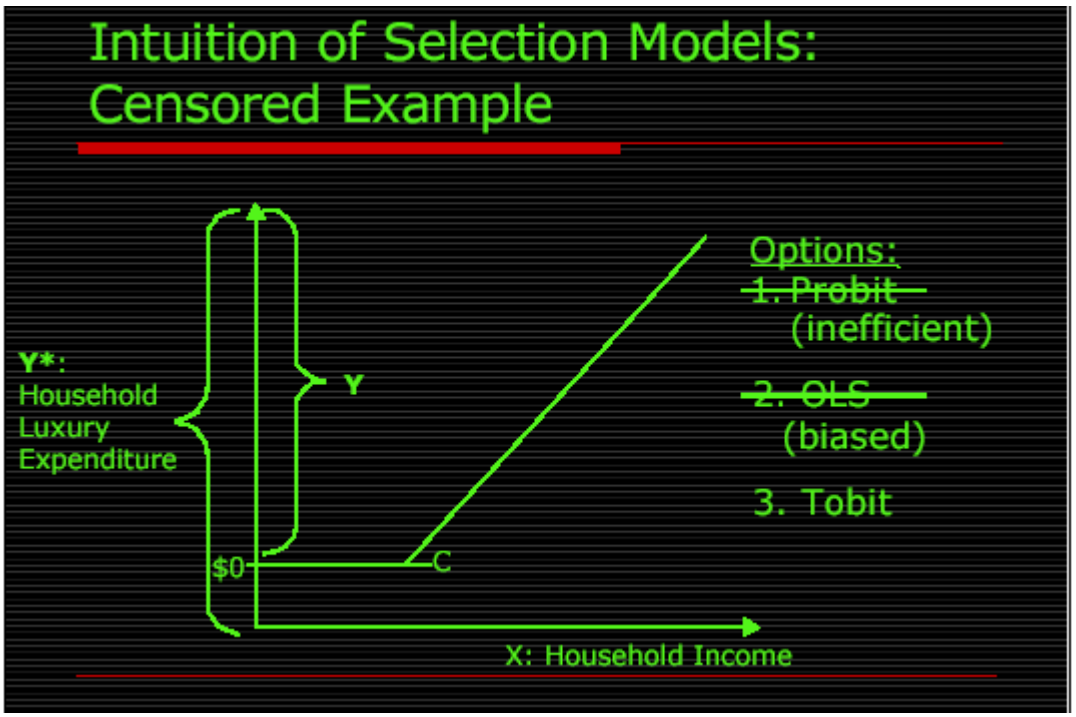


Figure 1: Plot of the Hazard Rate (negative argument)

3. Implementing and Interpreting Sample Selection Models[1]

3.1 The Intuition Behind Selection Models



3.2 Tobit

The Latent Model: $yi^* = xi'b + ui^*$

But, we have censoring at C = 0:

$yi = yi^*$ if $yi^* > C$;
$yi = C$ if $yi^* < C$

So, The Observed Model: $yi = xi'b + ui$ if $yi > 0$
$yi = 0$ otherwise.

[1] By Kevin Sweeney

In the Case of OLS: E(*yi*|*xi*) = *xi'b*

$$L = \left(-\frac{n}{2}\right)\left[\log\left(2\pi\sigma^2\right)\right] - \frac{1}{2}\sum_{i=1}^{n}\left[\frac{(Y_i - \beta' X_i)}{\sigma}\right]^2$$

If we censor y at C=0:

E(*yi*|*xi*) = pr(*yi*>0|*xi*) · E(*yi*|*yi*>0,*xi*)

$$\prod_{0}(1-\Phi_i) \qquad \prod_{1}\Phi_i \qquad \prod_{1}\frac{1}{\sigma}\frac{\phi\left[(y_i - x_i'\beta/\sigma)\right]}{\Phi_i}$$

$$L = \sum_{Y_i>0} -\frac{1}{2}\left[\log\left(2\pi\sigma^2\right)\right] + \left[\frac{(Y_i - \beta' X_i)}{\sigma}\right]^2 + \sum_{Y_i=0}\log\left[1-\Phi\left(\frac{\beta' X_i}{\sigma}\right)\right]$$

Interpreting Coefficients

1. Expected value of the underlying latent variable (Y*)

$$E(Y^*|x_i) = x_i'\beta$$

2. Estimated probability of exceeding C

$$pr(y_i > C) = \Phi\left(\frac{x_i'\beta}{\sigma}\right)$$

3. Expected, unconditional value of the realized variable (Y)

$$E(y_i|x_i) = \Phi_i\left(x_i'\beta + \sigma\frac{\phi_i}{\Phi_i}\right) + (1-\Phi_i)C$$

4. Expected Y, conditional on exceeding C

$$E(y_i|y>C,x_i) = x_i'\beta + \sigma\frac{\phi_i}{\Phi_i} + C$$

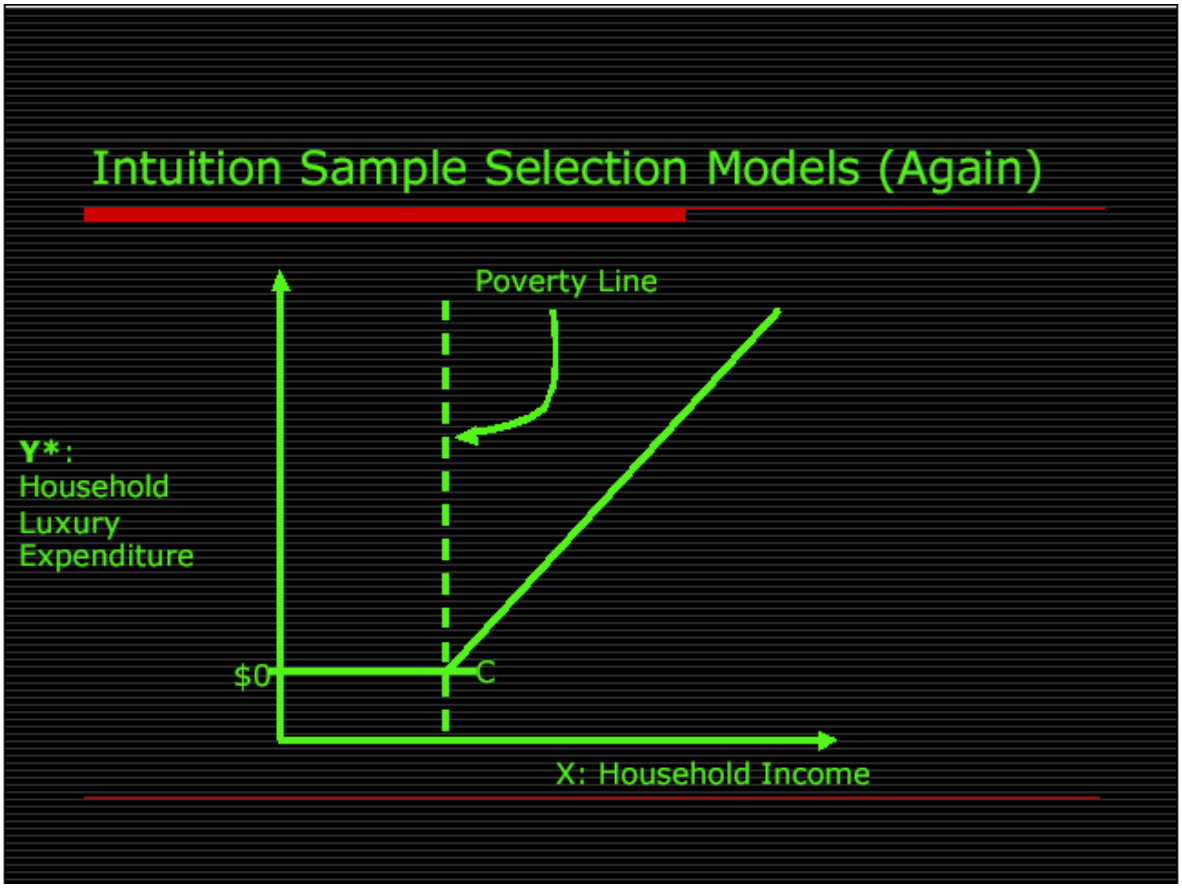
3.3 Sample Selection Models

Tobit Model Limitations:

- Same set of variables, same coefficients determine both P(censored) and the DV (outcome equation)
- Lack of theory as to why observations are censored

Selection Models:

- Different variables and coefficients in censoring (selection equation) and DV (outcome equation)
- Allow theory of censoring, observations are censored by some variable Z
- Allow us to take account of the censoring process because selection and outcome are not independent.



3.4 The Form of Sample Selection Models

- Selection Equation:

$$z_i^* = w_i'\alpha + e_i$$
$$z_i = 0 \ \ if \ z_i^* \le 0;$$
$$z_i = 1 \ \ if \ z_i^* \rangle 0$$

- Outcome equation:

$$y_i^* = x_i'\beta + u_i$$
$$y_i = y_i^* \ \ if \ z_i = 1$$
$$y_i \ \ not \ \ observed \ \ if \ z_i = 0$$

3.5 Where does the Bias Come From?

Step1: To begin, estimate a probit model

$$pr(z_i = 1) = \Phi(w_i'\alpha)$$

Next, estimate the expected value[2] of y, conditional on z=1, and x_i:

$$x_i'\beta + E(u_i \mid e_i \rangle w_i'\alpha) \qquad (1)$$

Evaluate the conditional expectation of u in (1):

$$E(u_i \mid e_i \rangle w_i'\alpha) = \rho\sigma_e\sigma_u \frac{\phi(w_i'\alpha)}{\Phi(w_i'\alpha)} \qquad (2)$$

---

[2] $E(y_i \mid z = 1, x_i) = x_i'\beta + E(u_i \mid z_i = 1)$

Substitute (2) into (1):

$$E(y_i \mid z=1, x_i) = x_i'\beta + \rho\sigma_\varepsilon\sigma_u \frac{\phi(w_i'\alpha)}{\Phi(w_i'\alpha)}$$

Step 2: Use OLS to regress y on x_i and $\hat{\lambda}_i = (\phi_i/\Phi_i)$:

$$E(y_i \mid z=1, x_i) = x_i'\hat{\beta} + \Theta\hat{\lambda}_i$$

3.6 The Likelihood Function:

$$L = \sum_0 \log(1-\Phi_i) + \sum_1 \log\Phi\left[\frac{w_i'\alpha + \rho\left(\frac{y_i - x'\beta}{\sigma_u}\right)}{(1-\rho^2)^{\frac{1}{2}}}\right] + \sum_1 -\frac{1}{2}\left[\log\left(2\pi\sigma_u^2\right)\right] + \left[\frac{(Y_i - \beta' X_i)}{\sigma_u}\right]^2$$

$\rho=0$

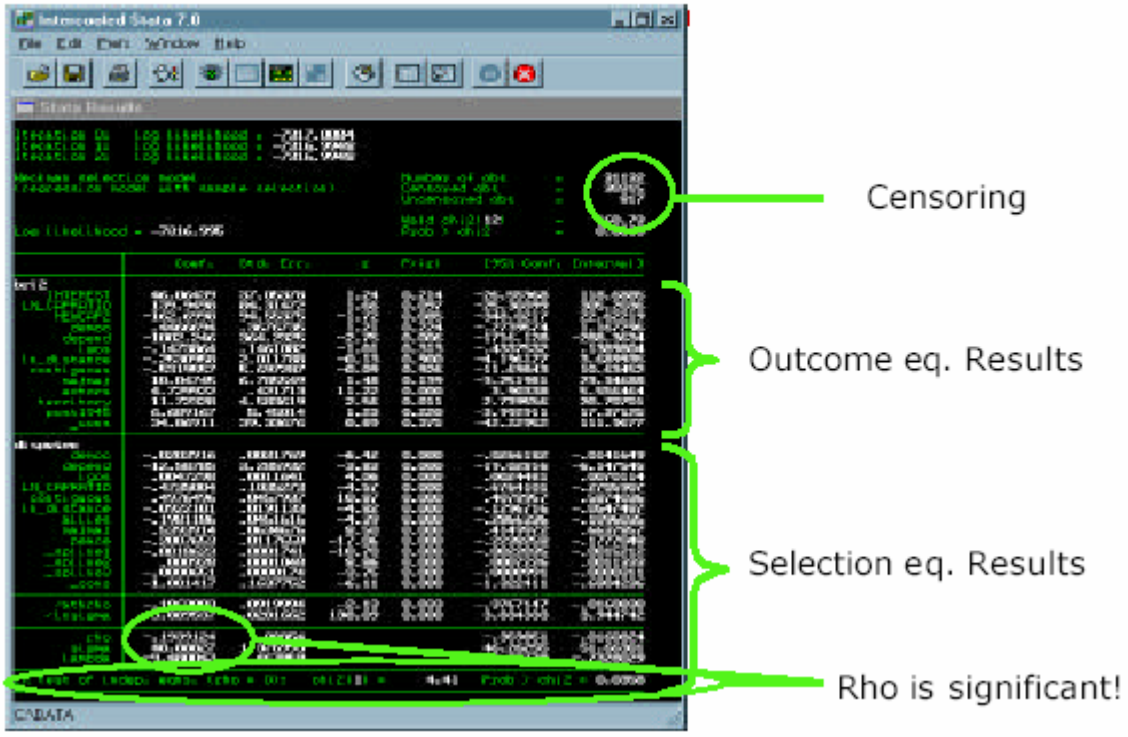
Probit:

$$L = \sum_0 \log(1-\Phi_i) + \sum_1 \log\Phi(w_i'\alpha)$$

OLS:

$$L = \sum_1 -\frac{1}{2}\left[\log\left(2\pi\sigma_u^2\right)\right] + \left[\frac{(Y_i - \beta' X_i)}{\sigma_u}\right]^2$$

3.7 Estimating a Heckman Model in Stata



Remark:

Censored, Sample Selected, and Truncated Variables

Sample	y variable	x variables
Censored	y is known only if some criterion df. in terms of y is met.	x variables are observed for the entire sample.
Sample Selected	y is observed only if a criteria df. in terms of another variable (Z) is met.	x and w are observed for the entire sample.
Truncated	y is known only if some criterion df. In terms of y is met.	x variables are observed only if y is observed.

ANEXOS

A. El inverso del radio de Mill

B. Endogeneidad vs. Sesgo de Selección

Fuente: STATA

Why are there so many formulas for the inverse of Mills' ratio?

What if I have censoring from above/below in my Heckman selection model?

Title Mills' ratios and censoring direction in the Heckman selection model
Author Vince Wiggins, StataCorp
Date May 1999

Someone asked about what Heckman called the "inverse of Mills' ratio" (IMR) and its relation to Heckman's two-step method for estimating selection models.

The definition of the IMR tends to be somewhat inconsistent. In fact, the current help file for **heckman** uses the more intuitive "nonselection hazard" in preference to "inverse mills", primarily because the latter has so many variations in the literature.

Hazard has the unequivocal definition H(x) = f(x) / (1 - F(x)) and Mills' is usually just taken to be 1 / H(x), so the inverse Mills' is just the hazard. Many papers, however, take liberties with this definition of the Mills' and are not always clear about why.

The questioner states

As I understand it, the inverse Mills' ratio (IMR) computed by Stata's **heckman** command, and used in the second stage regression, is lambda=f(x)/F(x) where f(x) is the pdf and F(x) is the cdf (see [R] heckman).

What I do not understand is exactly how this fits in with the definitions of the IMR found in the literature. For example (sorry for any unclear notation):

1. Greene, Econometric Analysis, 3rd Edition, p. 952
   (Note: my A = his alpha; a is the truncation point)
   lambda(A) = f(A) / (1-F(A)) if truncation is x > a
   lambda(A) = -f(A) / F(A) if truncation is x < a

2. Heckman, 1979, p. 156
   lambda = $f(Z_i) / (1 - F(Z_i))$
   where $Z_i = - X_{2i}B_2 / (S_{22})^{0.5}$

3. LimDep 7.0 Manual, p. 639 (Also by W. Greene)
   lambda = f(A'w) / F(A'w)

4. LimDep 7.0 Manual, p. 640
   lambda = f(A) / F(A) if z = 1
   lambda = -f(A) / (1-F(A)) if z = 0

They might also have added that

5. Maddala, 1983, Limited-dependent and qualitative variables in econometrics, p. 231. adds his voice for
   lambda = f(Z) / (1 - F(Z))

Another user noted that,

The key is to remember some basic facts about the standard normal pdf (f) and CDF (F).

1. 1-F(A) = F(-A)
2. f(A) = f(-A)

Using these two facts, and some algebraic manipulation, you can show that all of the different formulas for the IMR are equivalent.

This observation is at the heart of one problem and shows why (5), (4-1), (3), (2), and (1-1) can, in the right context, be used interchangeably.

This first area of confusion results because some authors choose to model selection (e.g., Heckman) while others choose to model nonselection (e.g., Maddala).

If you model nonselection, the natural choice for the nonselection hazard is f(Zg) / (1 - F(Zg)), where typically $Zg = z_1*g_1 + z_2*g_2 + ...$ from the nonselection model. If you model selection, the natural choice for nonselection hazard is f(Zg) / F(Zg) where Zg = z1*g1 + z2*g2 + ... from the selection model. These are the same number because the Gaussian is symmetric.

In both cases, authors are computing the nonselection hazard, they are just beginning in the first case with a model of nonselection — so you get the standard form for the hazard — and in the second case with a model of selection — so the nonselection hazard has a different computation that arrives at the same value.

Authors also tend to write the nonselection hazard in whatever form is convenient. Heckman, for example, models selection and uses f(-Zg) / (1 - F(-Zg)) which is equal to f(Zg) / F(Zg).

A second question is a bit trickier.

Maybe there is something obvious I am missing here, but I'm still missing it. Stata's calculation of the IMR appears to assume x < a ("truncation from above" in Greene's terminology, I think). But in many — perhaps most — cases in econometrics I suspect the truncation is the other way around. Two that come to mind are the classic female labor supply question, and the model I am working on, which is trying to explain the determinants of children's school performance, taking into consideration the selectivity aspect in a country where a large proportion of children do not go to school.

This turns out to be a non-issue for the the Heckman estimator because the direction of the truncation is normalized out in the specification. The shortest story I can think of to show this is somewhat involved.

Let's be very clear about how the two-step estimator works. We have a regression equation of interest

$y_1 = Xb + _{e1}$ where $Xb = x_1*b_1 + x_2*b_2 + ...$

We do not, however, always observe y1; instead, we have a selection equation that determines whether y1 is observed.

$y_2 = Zg + e_2$
where $Zg = z_1*g_1 + z_2*g_2 + ...$
$e_1, e_2 \sim N(0, 0, S_1, 1, rho)$ — bivariate Gaussian
where $S_2 = 1$ is the same normalization used to identify a probit model.

Also, y_2 is not observed. We only know that y1 is observed only when $y_2 > 0$, e.g., when $Zg > e_2$.

The use of 0 as the cutoff for selection is a necessary normalization without which the model is not identified.

With this in hand, we can write the expectation of y_1 conditional on y_1 being observed; that is, y_1 conditional on $y_2 > 0$ or equivalently $e_2 > -Zg$.

So the conditional expectation of y1 is

$$E(y_1 \mid e_2 > -Zg) = Xb + E(e_1 \mid e_2 > -Zg)$$

and from the moments of a censored bivariate Gaussian this is

$$E(y_1 \mid e_2 > -Zg) = Xb + rho*S_1*f(Zg)/F(Zg)$$

Heckman's insight was to formulate this conditional likelihood and then to obtain Zg from a probit estimation on whether y1 is observed. Thus getting consistent estimates of b and $rho*S_1$ when f(Zg)/F(Zg) are included in the regression — $y_1 = Xb + rho*S_1 * f(Zg)/F(Zg)$.

Heckman's f(Zg)/F(Zg) corresponds to Greene's expression (I'm going to change Greene's notation slightly to match the Heckman model.)

f(a-Zg) / (1-F(a-Zg)) if truncation is $y_2 > a$

because, as seen above, we have estimated a selection model and need the nonselection hazard.

Finally, we are ready to answer the question about models where the expected censoring is $y_2 < a$, rather that the $y_2 > a$. This corresponds to the expression Greene quotes as

-f(a-Zg) / F(a-Zg) if truncation is $y_2 < a$

and is the required component of the formula for the conditional expectation of e1 when $y_2 < a$.

Recall that the Heckman model normalized a to be 0 — since a could not be identified separately from the parameter vector g. That means we really have the selection rule $y_2 < 0$ for the case that concerns questioner, and $y_2 > 0$ for the standard formulation of the Heckman model. We also know that $E[e_1] = 0$ from the assumption of the regression model on the unconditional value of y_2.

When centered at [0,0], the bivariate normal is mirror symmetric about the origin. Thus, we can't tell the difference between a model with selection rule $y_2 > 0$ and a positive value for rho and a model with a selection rule $y_2 < 0$ and a negative value of rho.

The most important thing to know is that we will get the same estimates of b from either specification. The data sees to it that the direction of the censoring is accounted for. We would require prior information to differentiate the two models. The data alone cannot distinguish them, it can only identify the the direction of the censoring. One would have to specify the form of the censoring to distinguish the two models and then only the estimate of rho would differ.

What is the difference between 'endogeneity' and 'sample selection bias'?

Title Endogeneity versus sample selection bias
Author Daniel Millimet, Southern Methodist University
Date October 2001

Question:

Many individuals have posted questions using sample selection bias and endogeneity interchangeably or incorrectly. I do not intend to single out one individual, but consider the following:

Consider the case of the effect on wages of workers of being in a trade union. Using a dummy variable to pick up this effect in a pooled sample of union and non-union workers is inappropriate since workers in unions may self select and workers being in a union may not be random.

One approach I have read is to use a probit model to estimate the probability of being in a union (1 being union worker and 0 being non-union worker). Then from the probit equation, obtain predicted probabilities of being a union worker for the entire sample of union and non-union workers. Then use these predicted probabilities in place of a union dummy variable to estimate the effect of being in a union. This approach is supposed to control for sample selection bias.

I am trying to relate this procedure with the standard Heckman's 2-stage procedure that uses the inverse Mills' ratio. Any help will be much appreciated.

Answer:

Sample selection bias and endogeneity bias refer to two distinct concepts, both entailing distinct solutions. In general, sample selection bias refers to problems where the dependent variable is only observed for a restricted, non-random sample. Using the example above, one only observes an individual's wage within a union if the individual has joined a union. Conversely, one only observes an individual's non-union wage if the individual does not belong to a union. Endogeneity refers to the fact that an independent variable included in the model is potentially a choice variable, correlated with unobservables relegated to the error term. The dependent variable, however, is observed for all observations in the data. In the example above, union status may be endogenous if the decision to join or not join a union is correlated with unobservables that affect wages. For instance, if less able workers are more likely to join a union and therefore receive lower wages ceteris paribus, then failure to control for this correlation will yield an estimated union effect on wages that is biased down.

The problem with unions and wages, and a host of other problems, can be treated either as a sample selection problem or as an endogeneity problem. The 'appropriate' model depends on how one believes unions affect wages.

Model I. Endogeneity

If one believes that union status has merely an intercept effect on wages (i.e. results in a parallel shift up or down for various wage profiles), then the appropriate model includes union status as a right-hand side variable, and pools the entire sample of union and non-union workers. Because the entire sample is utilized, there are no sample selection issues (there may be a sample selection issue to the extent that wages are only observed for employed workers; typically this is only a cause for concern in estimating wage equations for females). One can then proceed to estimate a typical wage regression equation via OLS. If you believe union status is endogenous and workers self-select into union/non-union jobs, then one should instrument for union status. One can use either two-step methods, as outlined in the question above, or use the Stata command **treatreg**. Upon estimating the model, the union status coefficient answers the following question: "Conditional on the X's, what is the average effect on wages of belonging to a union?" Note, under this estimation technique, the betas (the coefficients on the X's) are restricted to be the same for union and non-union workers. For example, the return to education is restricted to be the same regardless of whether or not one is in a union.

Model II. Sample Selection

If one believes that union status does not have only an intercept effect, but also a slope effect (i.e., the betas differ according to union status as well), then a sample selection model is called for. To proceed, split the sample into union and non-union workers, and then estimate a wage equation for each sub-sample. If union status is the only potentially endogenous variable in the model, the two separate wage equations may be estimated via OLS, accounting for the fact that each sample is a non-random sample of all workers. This is accomplished via Heckman's selection correction model (utilizing either ML estimation, or two-step estimation where in the first stage a probit model is used to predict the probability of union status and in the second-stage, the inverse Mills' ratio (IMR) is included as a regressor). According to this type of model, the union effect does not show up as a dummy variable, but rather in the fact that the constant term and betas may differ from the union to the non-union sample. The difference in the constants yields the difference in average wages if a union and non-union worker have X=0. The difference in the betas tells one how the returns to different observable attributes vary by union status. Essentially this model allows a full set of interaction terms between union status and the X's. A Chow test could be used to test if the betas differ across by union status. If they do not, Model I is more efficient. This type of model is also known as an endogenous switching regime model.

Other references: Main and Reilly (Economica, 1993) estimate a sample selection model similar to Model II, where they split the sample depending on the size of the firm where the individual works. Thus, their first-stage involves an estimating an ordered probit for three classes of firm size (small, medium, or large), and then estimating three wage equations, each including the appropriate IMR term. Millimet (2000, SMU working paper) estimates the effect of household size on schooling using a similar modeling technique. Maddala (1983) also gives a good introduction to these issues.

Model III. Endogeneity & Sample Selection

One may also confront both types of biases in the same model. For example, say one wants to estimate the effect of union status on wages for women only. Thus, one may choose to include union status as a right-hand side variable (Model I), or wish to split up the sample (Model II). If one opts for Model I, one still has to confront the fact that wages for women are only selectively observed; for those women choosing to participate in the labor force. To estimate this model, one would start by estimating a probit model explaining the decision of women to work or not. One would then generate the IMR and include the IMR and the union dummy in a second-stage wage regression, where one would instrument for union status if it was thought to be endogenous. Finally, if Model II were desired, then one would be confronted with a double selection model. I believe one would estimate a probit for labor force participation first. Upon generating the IMR term, this would be included in a second probit equation explaining union status. The appropriate IMR term from this equation would then be included in the two final wage equations. (This topic is covered in Amemiya 1985.)

Identification:

As in any model, one must be aware from where identification arises. While it is well known that for instrumental variables estimation one requires a variable that is correlated with the endogenous variable, uncorrelated with the error term, and does not affect the outcome of interest conditional on the included regressors, identification in sample selection issues is often not as well grounded. Because the IMR is a non-linear function of the variables included in the first-stage probit model, call these Z, then the second-stage equation is identified — because of this non-linearity — even if Z=X. However, the non-linearity of the IMR arises from the assumption of normality in the probit model. Since most researchers do not test or justify the use of the normality assumption, it is highly questionable whether this assumption should be used as the sole source of identification. Thus, it is advisable, in my opinion, to have a variable in Z that is not also included in X. This makes the source of identification clear (and debatable). In the case of the double selection model discussed above in Model III, two exclusion restrictions would be needed (one for the labor force probit, one for the union probit).

References

Amemiya, T. 1985.
Advanced Econometrics. Cambridge, MA: Harvard University Press.

Maddala, G. S. 1983.
Limited-Dependent and Qualitative Variables in Econometrics. Cambridge, UK: Cambridge University Press.

Main, B. and B. Reilly. 1993.
The Employer Size-Wage Gap: Evidence for Britain. *Economica* 60: 125–142.