Sesión 7 - Modelos de variable censurada y truncada*

Manuel Barrón†

5 de Julio de 2010

1 Mecanismos de censura y truncamiento

1.1 Distribución truncada

Una distribución truncada es una parte de una distribución no truncada que está por encima o por debajo de cierto valor.

La función de densidad de variable aleatoria truncada viene dada por:

$$f(x|x > a) = \frac{f(x)}{Prob(x > a)}$$

Una variable aleatoria truncada es una variable aleatoria cuya distribución está truncada. La gran mayoría de aplicaciones recientes de variables aleatorias truncadas usa la distribución normal truncada. Si $x \sim N(\mu, \sigma^2)$, entonces

$$Prob(x > a) = 1 - \Phi\left(\frac{a-\mu}{\sigma}\right) = 1 - \Phi(\alpha)$$

La distribución normal estándar truncada viene dada por

$$f(x \geq a) = \frac{f(x)}{1-\Phi(\alpha)} = \frac{\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)}{1-\Phi(\alpha)}$$

.

Distribución de Poisson truncada en cero:

$$Prob[Y = y|y > 0] = \frac{f(x)}{Prob[Y > 0]} = \frac{f(x)}{1 - Prob[Y = 0]} = \frac{e^{-\lambda}\lambda^y/y!}{1-e^{-\lambda}},$$

$$\lambda > 0,\ y = 1, 2, ...$$

*Notas de clase para el módulo 4 del curso de Econometría Intermedia, Maestría en Economía, PUCP. Estas notas deben ser tomadas como una introducción al tema, y deben ser complementadas con las secciones correspondientes del libro de texto. La referencia principal para esta sección es Greene capítulo 22. Preguntas, correcciones o comentarios bienvenidos. Contacto: manuel [punto] barron [arroba] pucp.edu.pe o berkeley.edu

†Departamento de Economía, PUCP y Departamento de Economía Agrícola y de Recursos Naturales, UC Berkeley.

1.2 Momentos de una distribución normal truncada

Si $x \sim N(\mu, \sigma^2)$ y a es una constante, entonces

$$E[x|truncación] = \mu + \sigma\lambda(\alpha)$$

$$Var[x|truncación] = \sigma^2[1 - \delta(\alpha)],$$

donde $\alpha = (a - \mu)/\sigma$, $\phi(\alpha)$ es la densidad normal estándar y

$$\lambda(\alpha) = \phi(\alpha)/[1 - \Phi(\alpha)]$$

si la truncación es $x > a$

$$\lambda(\alpha) = -\phi(\alpha)/\Phi(\alpha)$$

si la truncación es $x < a$, y finalmente

$$\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]$$

2 Datos censurados

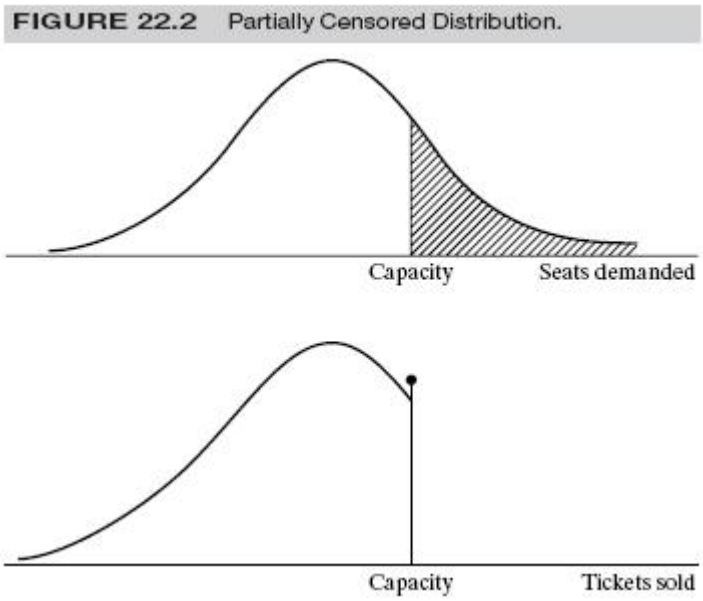
Un problema muy común en datos microeconómicos es la censura en la variable dependiente. Cuando la variable dependiente está censurada, los valores en cierto rango se transforman (o reportan) como un único valor. Por ejemplo, reportar ingresos como "más de \$5000", o "menos de \$100".

La teoría para analizar la distribución normal censurada es similar a la que usamos para estudiar la distribución truncada. Para simplificar la discusión, normalizamos el punto de censura a 0.

Para analizar la distribución, definimos una nueva variable aleatoria y transformada de la original y^*, por:
$y = 0 \quad si\ y^* \leq 0$
$y = y^* \quad si\ y^* > 0$

La distribución que aplica si $y^* \sim N(\mu, \sigma^2)$ es $Prob(y = 0) = Prob(y^* \leq 0) = \Phi(-\mu/\sigma) = 1 - (\mu/\sigma)$, y si $y^* > 0$, y tiene la densidad de y^*.

Figura 2. Distribución parcialmente censurada[1]



2.1 Momentos de una distribución normal censurada

Sea $y^* \sim N(\mu, \sigma^2)$. Si $y = a$ cuando $y^* \leq a$ y $y = y^*$ en otro caso, entonces

$$E[y] = \Phi a + (1 - \Phi)(\mu + \sigma\lambda)$$

y

$$Var[y] = \sigma^2(1 - \Phi)[(1 - \delta) + (\alpha - \lambda)^2\Phi]$$

,

donde

$$\Phi[(a - \mu)/\sigma] = \Phi(\alpha) = Prob(y^* \leq a)$$

$\lambda = \phi/(1 - \Phi)$,
$\delta = \lambda(\lambda - \alpha)$

El siguiente ejemplo fue tomado de Greene (ejemplo 22.3). Queremos analizar el número de tickets demandados para un evento en una ciudad. Lo único que observamos es el número de tickets vendidos, y sabemos que si se vende todas las entradas, el número de tickets demandados es mayor o igual al número de tickets vendidos. En otras palabras, el número de tickets demandados está censurado cuando se transforma para obtener el número de tickets vendidos. Supongan que hay 20,000 asientos y que 25% de los eventos venden todos los tickets. Si el número de entradas vendidas, incluyendo aquellas en las que se venden todas las entradas, es 18,000, ¿cuál es el proemdio y la desviación estándar de la demanda por entradas?

[1] Tomado de Greene, Figura 22.2

$18,000 = 20,000(1-\Phi) + (\mu + \sigma\lambda)\Phi$

Como la censura viene por arriba, $\lambda = -\phi(\alpha)/\Phi(\alpha)$. El argumento de Φ, ϕ, y λ es $\alpha = (20,000 - \mu)/\sigma$. 25% de eventos en los que se agotan las entradas implica $\Phi = 0.75$. Invirtiendo la normal estándar en 0.75 nos da $\alpha = 0.765$. $-\phi(0.675)/0.75 = \lambda = -0.424$. Con ello tenemos:

(i) $18,000 = 0.25(20,000) + 0.75(\mu - 0.424\sigma)$

(ii) $0.675\sigma = 20,000 - \mu$

Las soluciones son $\mu = 18,362$ y $\sigma = 2,426$.

Si nos dicen que la media de 18,000 aplica sólo para los eventos en los que no se vende todas las entradas y que todas las entradas se venden 25% del tiempo, nuestros estimados serían obtenidos de las siguientes ecuaciones:

(i) $18,000 = \mu - 0.424\sigma$

(ii) $0.675\sigma = 20,000 - \mu$

Las soluciones son $\mu = 18,772$ y $\sigma = 1,820$.

3 El modelo de regresión censurada Tobit

La formulación general viene de

$$y_i^* = x_i'\beta + \varepsilon_i \quad ...(1)$$

$$y_i = 0 \quad si\ y_i^* \leq 0$$

$$y_i = 0 \quad si\ y_i > 0$$

Hay tres medias que pueden ser de interés, dependiendo del enfoque del análisis. En primer lugar, para la variable latente,

$$E[y_i^*|x] = x_i'\beta \quad ...(2)$$

En segundo lugar, podemos estar interesados en la esperanza para una observación seleccionada aleatoriamente:

$$E[y_i|x] = \Phi\left(\frac{x_i'\beta}{\sigma}\right)(x_i'\beta + \sigma\lambda_i) \quad ...(3)$$

Donde $\lambda_i = \frac{\phi(x_i'\beta/\sigma)}{\Phi(x_i'\beta/\sigma)}$.

En tercer lugar podemos estar interesados en las observaciones no censuradas, en cuyo caso aplican los resultados del modelo de regresión truncada.

Hay diferencias en los efectos marginales:

$$\frac{\partial E[y_i^*|x_i]}{\partial x_i} = \beta \quad ...(4)$$

$$\frac{\partial E[y_i|x_i]}{\partial x_i} = \beta\Phi\left(\frac{x_i^{'}\beta}{\sigma}\right) \quad ...(5)$$

3.1 Estimación

La función de log-verosimilitud es

$$lnL = \sum_{y_i>0} -\frac{1}{2}\left[log(2\pi) + ln\sigma^2 + \frac{(y_i - x_i^{'}\beta)2}{\sigma^2}\right] + \sum_{y_i=0}\left[1 - \Phi\left(\frac{x_i^{'}\beta}{\sigma}\right)\right]$$

Las dos partes corresponden a la regresión clásica para las observaciones con y positivos y a las probabilidades para las observaciones con $y = 0$, respectivamente. Por lo tanto, esta verosimilitud no es del tipo estándar, dado que es una combinación de distribuciones continuas y dscretas. Sin embargo, Amemiya (1973) muestra que a pesar de estas complicaciones, se puede maximizar lnL de la manera usual.

Se puede reescribir la función de log-verosimilitud de la siguiente manera para simplificar los cálculos, definiendo $\gamma = \beta/\sigma$ y $\theta = 1/\sigma$:

$$lnL = \sum_{y_i>0} -\frac{1}{2}\left[log(2\pi) - ln\theta^2 + (\theta y_i - x_i^{'}\gamma)^2\right] + \sum_{y_i=0}\left[1 - \Phi\left(x_i^{'}\gamma\right)\right]$$

4 Aplicación: Quester & Greene (1982)

En un estudio del número de horas trabajadas por año, Quester & Greene (1982) estudian si las mujeres cuyos matrimonios tienen mayor probabilidad de disolución (desde el punto de vista estadístico) se protegían del riesgo trabajando más horas que las que estaban en un matrimonio estadísticamente más seguro. Los resultados de su modelo están resumidos en la Tabla 1. Los números en la última fila de la tabla implican que una porción importante de las mujeres reporta 0 horas, por lo que una regresión por MCO sería inadecuada.

Los números en paréntesis son los ratios de los coeficientes sobre sus errores estándar asintóticos. La variable dependiente es el número de horas trabajadas durante el año previo a la encuesta. La variable "niños" indica si había niños pequeños en el hogar. Las variables "diferencia educacional" y "ratio de salarios" comparan a la pareja en estas dos dimensiones. La variable "segundo matrimonio" es una dummy que indica si es el segundo matrimonio de la mujer. La probabilidad de divorcio es tomada de un estudio previo (Orcutt, Caldwell & Wertheimer, 1976). Las variables usadas aquí son dummies indicando "promedio" si la probabilidad predicha estaba entre 0.01 y 0.03 y "alta" si era mayor a 0.03. Las "pendientes" son los efectos marginales descritos en la ecuación (5). Noten que la pendiente es el producto del coeficiente multiplicado por $\Phi\left(\frac{x_i^{'}\beta}{\sigma}\right)$ evaluado en $x_i = \bar{x}$, que en este caso es la proporción de mujeres fuera del mercado laboral. Es fácil confirmar que 1324.84*0.29=385.89, o que 25.33*0.46=11.57.

Note la diferencia entre los efectos marginales y los resultados del modelo tobit. Similarmente, el estimador de σ está bastante lejos de ser un buen estimador de la desviación estándar de las horas trabajadas.

Los efectos de las probabilidades de divorcio son significativos, tienen el signo esperado, y son bastante grandes. Una de las principales críticas a este estudio es sobre si la probabilidad de divorcio podria ser

tratada como una variable independiente. Podría ser que, para estas parejas, el número de horas trabajadas fuera un determinante significativo de la probabilidad de divorcio.

Table 1: Restulados Tobit vs MCO

	(1) Blancas - Coef.	(2) Blancas - Pendiente	(3) Negras - Coef.	(4) Negras - Pendiente	(5) MCO
Intercepto	-1803.13		-2753.87		
	(-8.64)		(-9.68)		
Niños pequeños	-1324.84	-385.89	-824.19	-376.53	-352.63
	(-19.78)		(-10.14)		
Diferencia educación	-48.08	-14.00	22.59	10.32	11.47
	(-4.77)		(1.96)		
Ratio de salarios	312.07	90.90	286.39	130.93	123.95
	(5.71)		(3.32)		
Segundo matrimonio	175.85	51.51	25.33	11.57	13.14
	(3.47)		(0.41)		
Probabilidad mediana de divorcio	417.39	121.58	481.02	219.75	219.22
	(6.52)		(5.28)		
Alta probabilidad de divorcio	670.22	195.22	578.66	264.36	244.17
	(8.40		(5.33)		
σ	1559	618	1511	826	
Tamaño de muestra	7459		2798		
% trabajando	0.29		0.46		