Variables Parcialmente Continuas

Walter Sosa Escudero

wsosa@udesa.edu.ar

Universidad de San Andrés

Introduccion

- Variables aleatorias discretas: los valores del soporte pueden ocurrir con probabilidad no nula (bernoulli, poisson, etc.), cara o ceca, numero de hijos.

- Variables aleatorias continuas: si bien cada punto del soporte ocurre con probabilidad cero, intervalos del soporte ocurren con probabilidad no nula (normal, chi^2. Ingresos, llegada a clase, etc.

- Variables aleatorias parcialmente continuas: ciertos valores pueden ocurrir con probabilidad no nula. Gasto en seguro de vida, refinanciamiento de deuda, ingresos de fuentes tributarias.

Porque?

- *Censura:* algun mecanismo altera la informacion (datos tributarios)
- *Soluciones de esquina:* el resultado observable permite 'soluciones de esquina': gasto en durables, restricciones de capacidad.

Un modelo simple

$$\begin{cases} y^* & = & x'\beta + u \\ y & = & \max(0, y^*) \end{cases}$$

,con $E(u|x) = 0$.

Si observasemos una muestra aleatoria (y_i^*, x_i), $i = 1, \ldots, n$, podriamos estimar consistentemente β por MCO, regresando y_i^* en x_i.

Que pasa si solo observamos (y_i, x_i)?

- Regresar y_i en x_i.
- Regresar y_i en x_i, solo para los casos en que $y_i > 0$.

Ambas estrategias conducen a estimaciones inconsistentes.

Ejemplo: consumo, gasto y stocks

(Gourieroux, 2000)

Objetivo: estimar cómo factores observables (ingresos, tamaño de la familia, etc.) afectan el *consumo* de un durable.

c^* =consumo (valor monetario)
s^* = stock disponible del bien (valor monetario)
e^* = variacion deseada en el stock = $c^* - s^*$

g = gasto

$$g = \begin{cases} e^* & \text{si } e^* > 0 \\ 0 & \text{si } e^* \leq 0 \end{cases}$$

Supongamos que: $c^* = x'\beta + \nu$, x= vector de caracteristicas observables, v termino no observable. Definiendo $u \equiv \nu - s^*$:

$$g = \begin{cases} x\beta + u & \text{si } e^* > 0 \\ 0 & \text{si } e^* \leq 0 \end{cases}$$

- Tipicamente, en las encuestas se observa solo g y x.
- Los datos presentan muchas familias con $g = 0$
- El punto clave es que estamos interesados en el *consumo*, el cual no es directamente observable (se observa el gasto): los datos son *censurados*. Es importante notar que las caracteristicas de los hogares se observan para todas las familias.
- Si solo se observase informacion para familias con gasto positivo, los datos serian *truncados*.
- Regresar los gastos en x no funciona. Tampoco funciona restringir la muestra a las familias con gastos positivos.

Distribuciones truncadas

$X \sim f(x)$, $X|X > a$: variable *truncada* en a,

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

Si $X \sim N(\mu, \sigma^2)$, recordando que
$Pr(X < a) = Pr(1/\sigma(X - \mu) < 1/\sigma(a - \mu)) = Pr(z < \alpha)$

$$\begin{aligned} f(x|X > a) &= \frac{\frac{1}{\sigma}\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}{1 - \Phi(\alpha)} \\ &= \frac{(1/\sigma)\phi(z)}{1 - \Phi(\alpha)} \end{aligned}$$

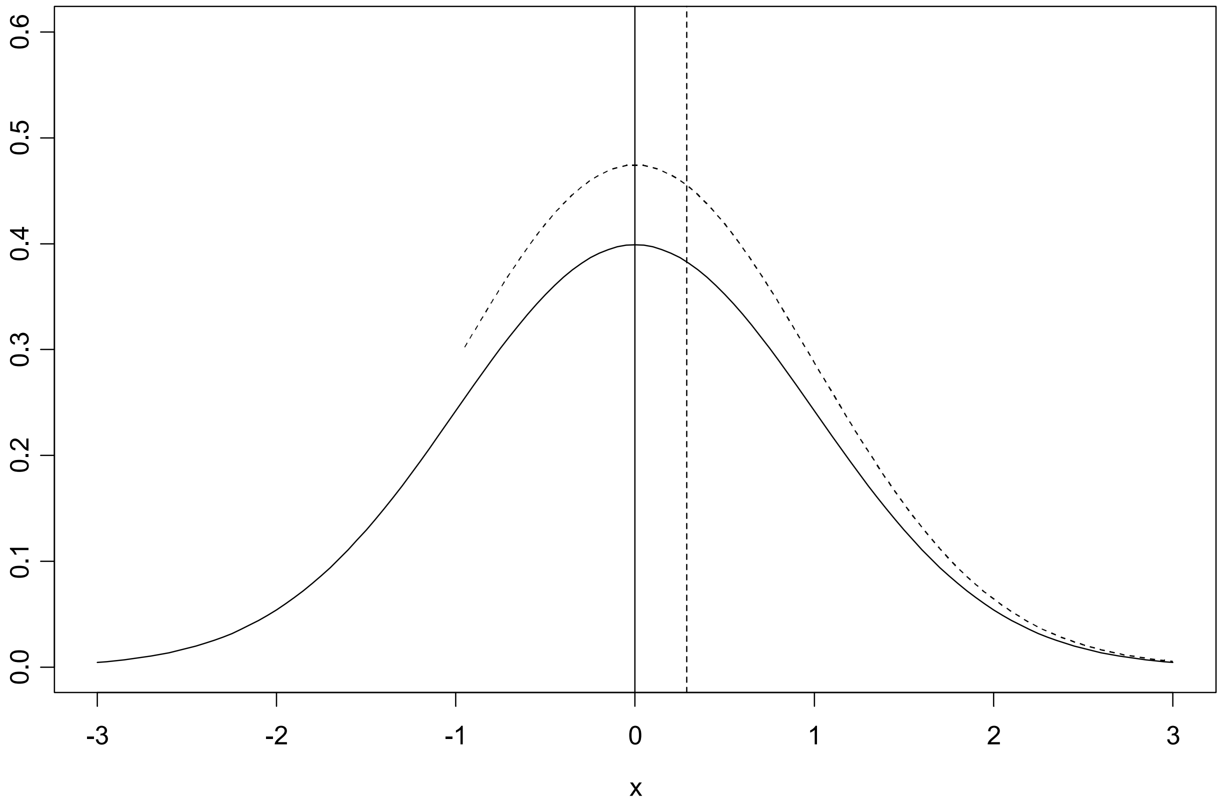
,$\alpha \equiv (a - \mu)/\sigma, \;\; z \equiv (x - \mu)/\sigma$.

Se puede verificar que si $X \sim N(\mu, \sigma^2)$:

$$E(X|X > a) = \mu + \sigma \frac{\phi(\alpha)}{1 - \Phi(\alpha)}$$

- Este resultado es muy importante. En primer lugar, dice que si el truncamiento ocurre a la izquierda, la esperanza se corre a la derecha.
- Tambien informa en *cuanto* se corre a la derecha. Depende del truncamiento α y de la varianza de la variable aleatoria (σ^2)

Distribucion normal truncada



La linea punteada corresponde a la densidad y esperanza de la normal truncada
La linea solida, a la normal estandar

Sesgo en la estimacion de MCO

La insesgadez del estimador de MCO descansa en que si $y = x'\beta + u$, $E(u|x) = 0$, o bien, en que:

$$E(y|x) = x'\beta$$

Consideremos el problema de estimar β utilizando el estimador MCO de regresar y_i en x_i solo para las observaciones con $y_i > 0$ (la muestra *truncada*).

En este caso, deberiamos verificar que sucede con $E(y|x, y > 0)$

$$
\begin{aligned}
y|y>0 &= x'\beta + u \\
E(y|x,y>0) &= x'\beta + E(u \mid x, y>0) \\
&= x'\beta + E(u \mid x, u > -x'\beta) \\
&= x'\beta + \sigma\frac{\phi(-x'\beta/\sigma)}{1-\Phi(-x'\beta/\sigma)} \\
&= x'\beta + \sigma\frac{\phi(x'\beta/\sigma)}{\Phi(x'\beta/\sigma)} \\
&= x'\beta + \sigma\lambda(x'\beta/\sigma) \\
&\neq x'\beta
\end{aligned}
$$

lo que muestra que el estimador de MCO es sesgado si se utiliza la muestra truncada.

$\lambda(z) \equiv \phi(z)/\Phi(z)$ se conoce como la *inversa de la razon de Mills*.

Un argumento similar muestra que MCO es tambien sesgado si se utiliza la muestra *censurada*.

$$E(y|x) = \Phi(x'\beta/\sigma)\Big[x'\beta + \sigma\lambda(x'\beta/\sigma)\Big] \neq x'\beta$$

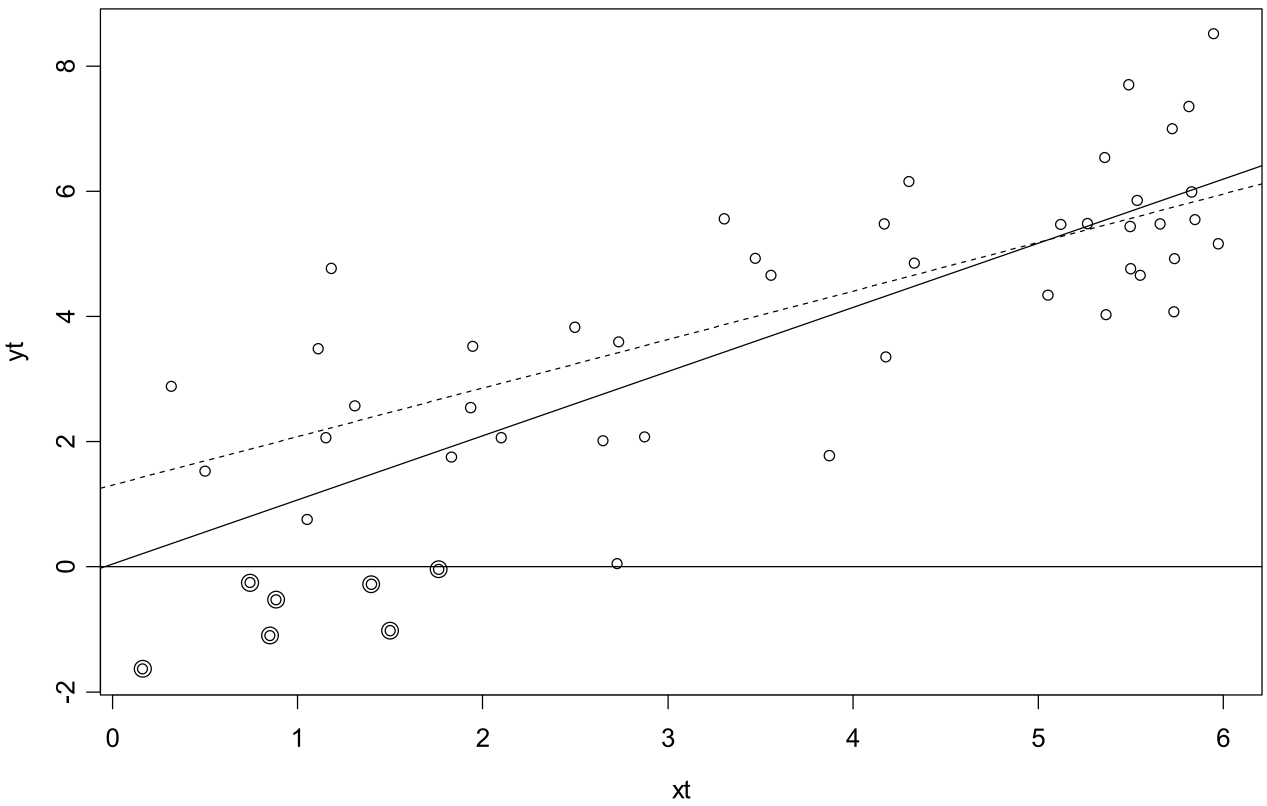
(v/Wooldridge, p. 521). Resultado util:

$$\begin{aligned} E(y) &= E(y|y>0)P(y>0) + E(y \leq 0)(1 - P(y>0)) \\ &= E(y|y>0)P(y>0) \end{aligned}$$

- Tanto para el caso de la muestra truncada como para el de la censurada, la funcion de regresion (la esperanza de y condicional en x y en la informacion relevante) *no es una funcion lineal*

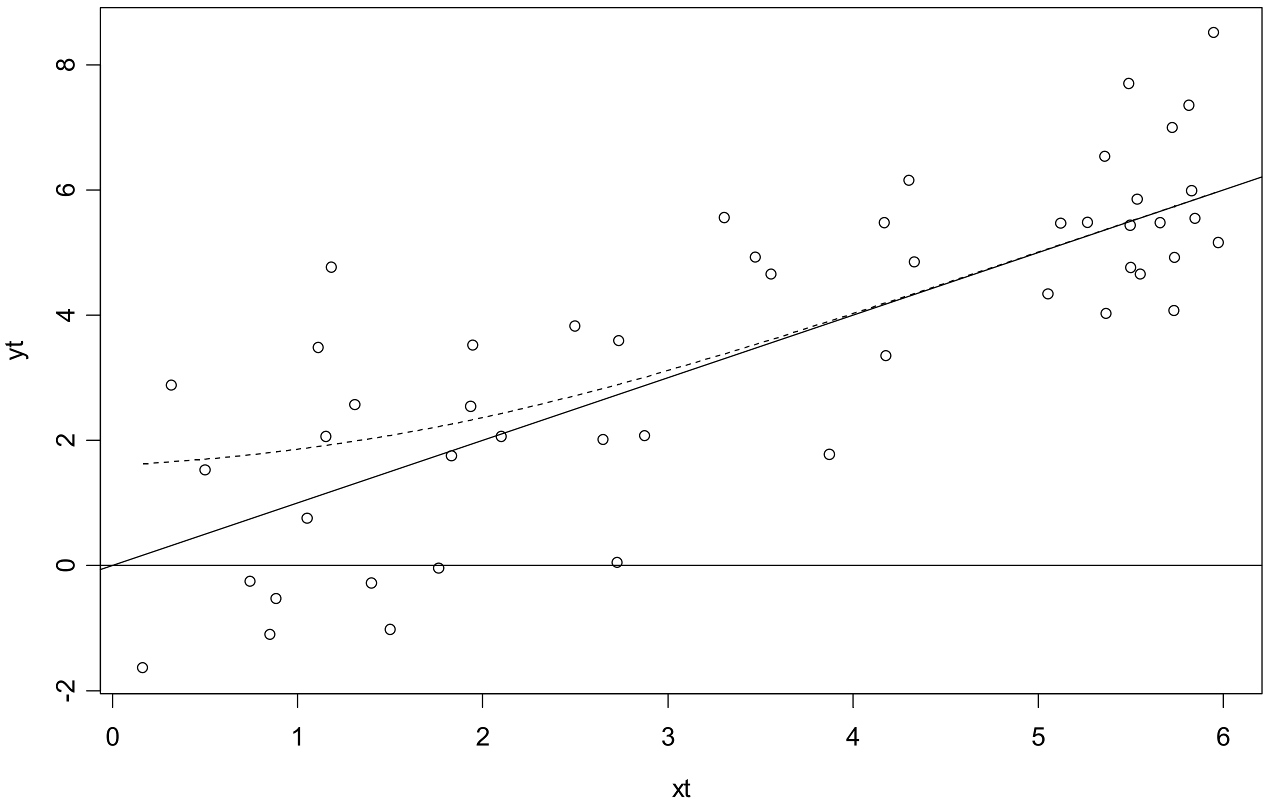
- Los 'sesgos' aparecen por querer estimar la esperanza como si fuese una funcion lineal.

Sesgo en el estimador MCO con datos truncados



- Los puntos con circulos mas grandes son observaciones truncadas
- La linea solida es la estimacion MCO con la muestra original
- La linea punteada es la estimacion con la muestra truncada

Funcion de regresion en el modelo original y truncado



- La linea solida es $E(y|x) = \alpha + \beta x$
- La linea punteada es $E(y|x, y \geq 0) = x'\beta + \sigma\lambda(x'\beta/\sigma)$

Estimacion e inferencia del modelo tobit

Al modelo de datos censurados con errores normales se lo conoce como modelo *tobit*.

Bajo el supuesto de normalidad, es sencillo derivar un estimador maximo verosimil.

Comencemos por dividir la muestra (y_i, x_i) en pares que tienen $y_i = 0$ y pares en que $y_i > 0$. Los primeros ocurren con probabilidad $Pr(y_i = 0|x_i)$ y los segundos ocurren con 'probabilidad' $f(y_i|x_i, y_i > 0)$. Llamemos $w_i \equiv 1[y_i > 0]$.

Entonces:

$$
\begin{aligned}
L(\beta) &= \prod_{i|y_i=0} Pr(y_i=0) \prod_{i|y_i>0} f(y_i|x_i, y_i>0) \\
&= \prod_{i=1}^{n} \left\{ Pr(y_i=0)^{(1-w_i)} \, f(y_i|x_i, y_i>0)^{w_i} \right\} \\
l(\beta) &= \sum_{i=1}^{n} \left\{ (1-w_i) \ln\left[1-\Phi(x_i'\beta/\sigma)\right] + w_i \ln\left[(1/\sigma)\phi[(y_i - x_i'\beta)/\sigma]\right] \right\}
\end{aligned}
$$

que, bajo condiciones estandar, produce un maximo unico utilizando metodos numericos simples.

Interpretacion de resultados

El estimador MV provee estimaciones consistentes de:

$$\beta = \frac{\partial E(y^*|x)}{\partial x}$$

lo cual prove una interpretacion posible: los coeficientes estimados son derivadas parciales de la variable *latente* (ej, consumo).

Pero muchas veces el interes recae sobre otras magintudes.

Efecto sobre la esperanza no-condicional

Recordar que:

$$E(y|x) = P(y > 0|x)\ E(y|x, y > 0)\ =\ \Phi(x'\beta)\big[x'\beta + \sigma\lambda(x'\beta/\sigma)\big]$$

Usando la regla de la cadena:

$$\frac{\partial E(y|x)}{\partial x_k} = \frac{\partial P(y > 0|x)}{\partial x_k} E(y|x, y > 0) + P(y > 0|x)\frac{\partial E(y|x, y > 0)}{\partial x_k}$$

Esta es la *descomposicion de Donald-Moffit.* Es facil mostrar que en este caso:

$$\frac{\partial E(y|x)}{\partial x_k} = \Phi(x'\beta/\sigma)\ \beta_k$$

que es simplemente un reescalamiento de los coeficientes originales.

Ejemplo: efectos sobre el *gasto* (no consumo). Una parte proviene de las familias que pasan de no gastar a hacerlo, y otra parte proviene del aumento en el gasto de las familias que ya lo hacian.

Efecto sobre la esperanza condicional

Tambien de los resultados anteriores

$$E(y|x, y > 0) = x'\beta + \sigma\lambda(x'\beta/\sigma)$$

Tomando derivadas:

$$\frac{\partial E(y|x, y > 0)}{\partial x_k} = \beta_k \Big[1 - \lambda(x'\beta/\sigma) \left[x'\beta/\sigma + \lambda(x'\beta/\sigma) \right] \Big]$$

Ejemplo: efectos sobre el *gasto* para las familias que gastan algo.

The effects of liquidity constraints on consumption: a cross-sectional analysis (Hayashi, 1985)

Test de restricciones de liquidez. c^* = consumo deseado, c = consumo observado. Supongamos que:

$$c^* = x'\beta + \nu \tag{1}$$

x= vector de caracteristicas observables, ν no es observable. No restriccion de liquidez $\Longleftrightarrow c^* = c$.

Con restriccion de liquidez:

$$c = \begin{cases} x'\beta + \nu & \text{si } c^* < K \\ K & \text{si } c^* \geq K \end{cases} \tag{2}$$

Hayashi: $K = 0.85\,(YD + 0.2\,LIQ)$. YD = ingreso disponible, LIQ = activos liquidos.

- En la practica, la idea es censurar *artificialmente* los consumos de todas las familias que estan ahorrando poco, se las trata como potencialmente restringidas.

- Independientemente de que haya restricciones de liquidez, β puede ser consistentemente estimado por *tobit* en (2). OLS produce estimaciones inconsistentes.

- Solo si no hay restricciones de liquidez, OLS es consistente.

- Si no hay restricciones de liquidez, tobit y OLS coinciden: rechazar H_0 en un Test de Hausman implica que hay restricciones de liquidez.

- **Tarea:** leer el paper y mirar los resultados empiricos