ECONOMETRÍA II
Prof.: Begoña Álvarez
2007-2008

Tema 5

MODELOS CENSURADOS, TRUNCADOS Y CON SELECCIÓN MUESTRAL

1. MUESTRAS Y POBLACIÓN

❑ La estimación consistente requiere:

- Disponer de una muestra extraída de forma aleatoria y representativa de la población que se pretende estudiar

- Que los estadísticos (estimadores) converjan a los parámetros poblacionales que estiman.

❑ El problema con las muestras surge cuando se refieren a grupo de la población que no representa a la población que es objeto de estudio.

En ese caso, los estimadores convergerán a las características de esa subpoblación, no a las de la población que se quiere analizar.

1. MUESTRAS Y POBLACIÓN

El objetivo de esta tema es:

- Mostrar la diferencia entre muestras truncadas y censuradas.

- Explicar por qué la estimación por MCO de un modelo lineal es sesgada e inconsistente en tales circunstancias.

- Proponer métodos para estimar muestras en las que la variable dependiente es continua pero limitada (bien por censura o truncamiento).

- También analizaremos el problema del sesgo de selección muestral.

2. MUESTRAS TRUNCADAS Y CENSURADAS

Es posible que no observemos datos de la variable dependiente y de las variables explicativas para toda la población. En este caso, tendremos muestras censuradas o truncadas según cómo sea el tipo de limitación en la información disponible

2. MUESTRAS TRUNCADAS Y CENSURADAS

2.1 MUESTRAS TRUNCADAS

Una muestra está truncada si los datos sólo están disponibles para un subconjunto de la población total.
Los valores de las variables explicativas X sólo se observan cuando se observa Y.

EJEMPLO:
- El gasto médico de una muestra de pacientes entrevistados después de someterse a un tratamiento dental. En este caso, sólo observamos a personas con gasto mayor que cero.

2. MUESTRAS TRUNCADAS Y CENSURADAS

2.2 MUESTRAS CENSURADAS

Una muestra está censurada si los datos se recodifican para un subconjunto de la población.
En una muestra censurada, observo las X de toda la población, pero el valor de la Y se desconoce para un subconjunto de la población.

EJEMPLO:
- Oferta de trabajo: si las personas trabajan, sabemos el número de horas que ofrecen, pero a los que no trabajan les asignamos cero horas…. Sin embargo, podría ser que su oferta de trabajo fuese de 3 horas por semana, pero no encuentra ningún empleo con esas características.

2. MUESTRAS TRUNCADAS Y CENSURADAS

FORMALIZACIÓN

- an observed dependent variable y
- a set of explanatory variables x
- a latent variable y^*

MUESTRAS TRUNCADAS

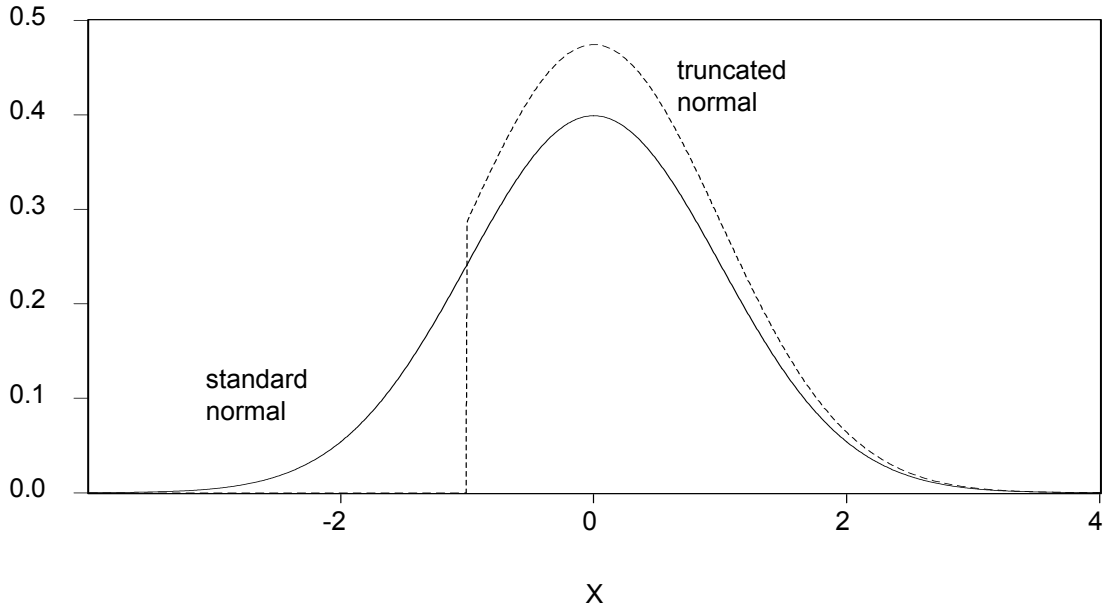
$$\begin{aligned}
&\text{T1: } y = y^* \text{ if } y^* > c \quad \text{not observed otherwise}\\
&\text{T2: } y = y^* \text{ if } y^* < d \quad \text{not observed otherwise}\\
&\text{T3: } y = y^* \text{ if } c < y^* < d \quad \text{not observed otherwise}
\end{aligned}$$

MUESTRAS CENSURADAS

$$\begin{aligned}
&\text{C1: } y = y^*.\mathbf{1}(y^* > c) + c.\mathbf{1}(y^* \leq c)\\
&\text{C2: } y = y^*.\mathbf{1}(y^* < d) + d.\mathbf{1}(y^* \geq d)\\
&\text{C3: } y = y^*.\mathbf{1}(c < y^* < d) + c.\mathbf{1}(y^* \leq c) + d.\mathbf{1}(y^* \geq d)
\end{aligned}$$

2. MUESTRAS TRUNCADAS Y CENSURADAS

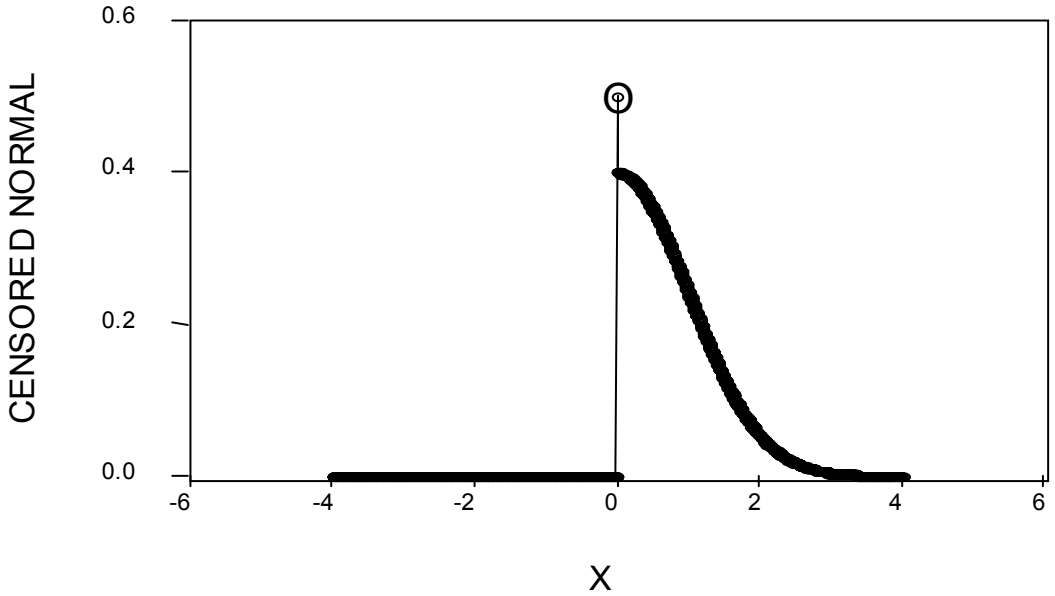
FORMALIZACIÓN



Truncated normal distribution with truncation from below (at x = -1). Source: Heij et al.

2. MUESTRAS TRUNCADAS Y CENSURADAS

FORMALIZACIÓN



Censored normal density with censoring from below (at x = 0) with a point mass p(x = 0) = 0.5. Source: Heij et al.

2. MUESTRAS TRUNCADAS Y CENSURADAS

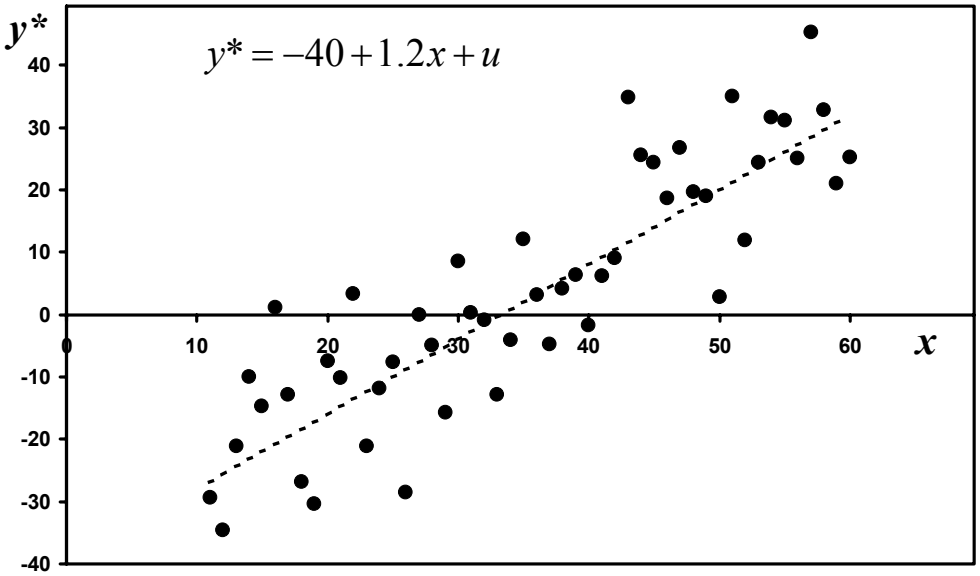
¿Por qué la censura o el truncamiento plantean un problema?

En particular, ¿qué problemas tenemos si especificamos un modelo lineal y estimamos por MCO un modelo en el que la variable está censurada o truncada?

Para ilustrar los problemas, vamos a centrarnos en el caso de una muestra censurada.

2. MUESTRAS TRUNCADAS Y CENSURADAS

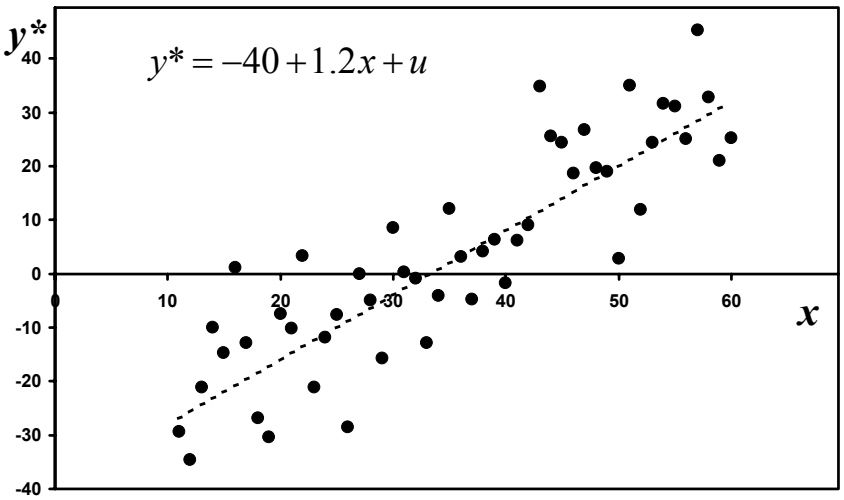
Por ejemplo, supongamos que la verdadera relación que tenemos es la que aparece en el gráfico.



2. MUESTRAS TRUNCADAS Y CENSURADAS

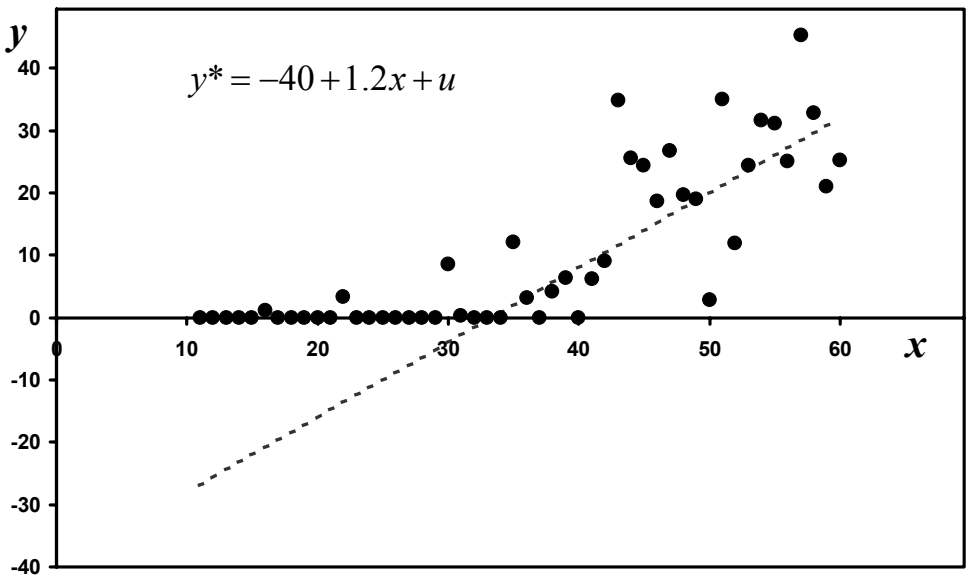
Sin embargo, imaginemos que la variable dependiente está sujeta a una cota inferior fijada en 0. Entonces los valores de variable observada Y serán tales que Y=y* si $y^* > 0$; $Y = 0$ if $y^* <= 0$.

Por ejemplo, supongamos que tenemos modelo de oferta de trabajo en que y son las horas de trabajo semanales. No es posible obtener valores negativos.



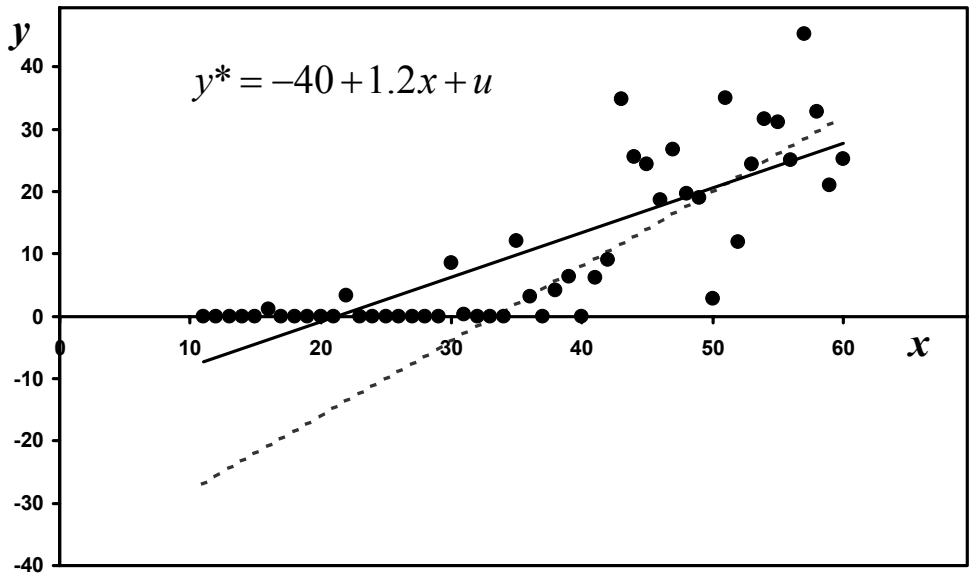
2. MUESTRAS TRUNCADAS Y CENSURADAS

Aquellos individuos con y^* negativa simplemente no trabajan. Para ellos, el valor de Y es 0

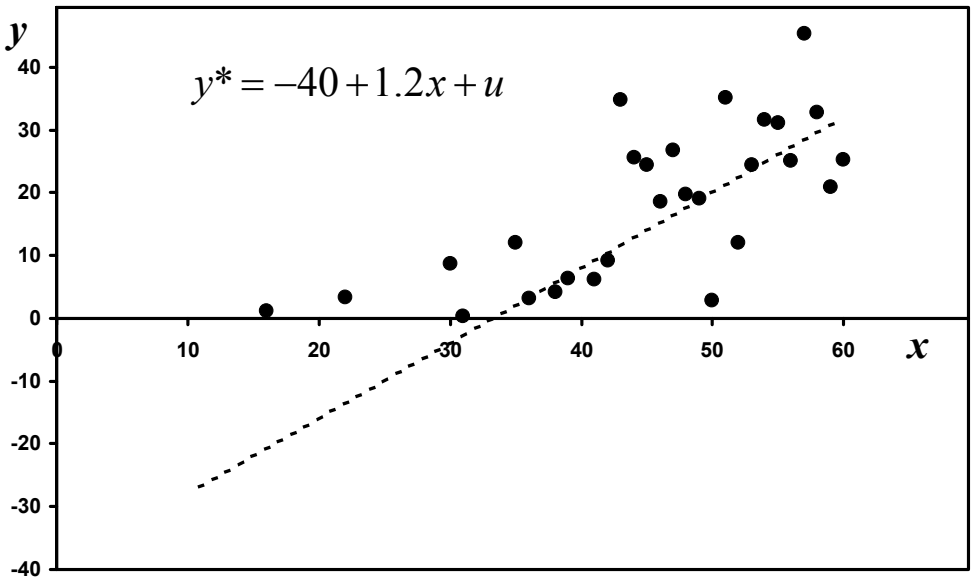


2. MUESTRAS TRUNCADAS Y CENSURADAS

¿Qué ocurriría si ajustásemos un modelo lineal y lo estimásemos por MCO? En este caso, la pendiente estaría sesgada a la baja.



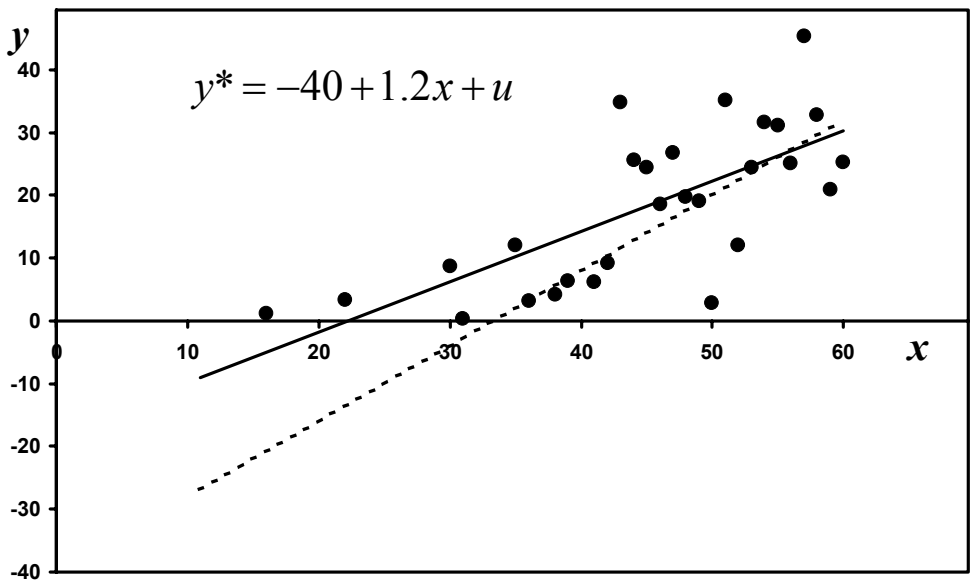
2. MUESTRAS TRUNCADAS Y CENSURADAS



¿Mejorarían las cosas si borrásemos las observaciones restringidas, es decir, las correspondientes a los que no trabajan?

En este caso estaría definiendo una MUESTRA TRUNCADA.

2. MUESTRAS TRUNCADAS Y CENSURADAS



De nuevo tendríamos estimaciones del parámetro de la pendiente sesgadas a la baja.

2. MUESTRAS TRUNCADAS Y CENSURADAS

La solución a este problema es plantear un modelo híbrido que utilice la especificación PROBIT para investigar por qué algunas observaciones toman valor 0 y otras no y, para aquellas observaciones tales que $Y^* > 0$, un modelo de regresión que nos cuantifque la relación.

El **modelo TOBIT** recoge esos dos aspectos.

3. MODELO TOBIT (Modelo censurado)

Supongamos una variable en la cual tenemos una solución esquina. Es decir, esa variable vale cero para una proporción considerable de la población, pero se distribuye de forma aproximadamente continua para los valores positivos.

EJEMPLO: El gasto en alcohol que hace un individuo en un mes determinado.

- **Formalmente, tenemos una variable y que es aproximadamente continua en un rango de valores estrictamente positivos pero que vale cero con probabilidad positiva.**
- **Nada impide que utilicemos un modelo lineal para la variable y, es decir, un modelo lineal para $E(y \mid z_1, x_2, \ldots, x_k)$.**
- **Sin embargo, podríamos obtener predicciones negativas.**
- **Debido a que la distribución de y presenta una acumulación de densidad en cero, no puede tener una distribución condicionada Normal. Por tanto, la inferencia estadística sólo tendrá justificación asintótica.**

3. MODELO TOBIT

ESPECIFICACIÓN

El modelo censurado o modelo Tobit (Tobin, 1958)

- Se dispone de datos para toda la muestra, pero la variable dependiente está censurada en un determinado valor, por ejemplo cero
- Consideremos la siguiente relación latente

$$y_i^* = x_i'\beta + u_i, \quad u_i \sim N(0, \sigma^2)$$

- Donde y es una variable censurada tal que

$$y_i = y_i^*.\mathbf{1}(y_i^* > 0)$$

3. MODELO TOBIT

ESPECIFICACIÓN

- Dada una distribución para u, la probabilidad de observar un dato censurado es:

$$\begin{aligned}\Pr(y_i = 0 \mid x_i) &= \Pr(y_i^* \le 0 \mid x_i) = \Pr(u_i \le -x_i'\beta) \\ &= \Phi(-z_i) = 1 - \Phi(z_i)\end{aligned}$$

- La probabilidad de las observaciones no censuradas es:

$$f(y_i) = \frac{1}{\sigma}.\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)$$

- Por tanto, la función de verosimilitud la escribimos como:

$$L(\beta, \sigma) = \prod_{y_i=0}\left[1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right]\prod_{y_i>0}\frac{1}{\sigma}.\phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)$$

3. MODELO TOBIT

INTERPRETACIÓN

Los β del modelo Tobit miden los efectos marginales de las variables explicativas sobre la variable latente y*. En ocasiones, esta variable tiene una interpretación económica interesante, pero en la mayoría de los casos no es así. La variable que queremos explicar es y, que es la que se puede observar.

¿Qué información podemos obtener del modelo estimado?

- Podemos estar interesados en calcular el efecto marginal de las variables explicativas sobre **E (y | X)**

- O nos pueden interesar los efectos marginales de las variables explicativas sobre **E (y | X, y>0)**

3. MODELO TOBIT

INTERPRETACIÓN

- Podemos obtener **E(y | x)** de forma sencilla. Recordamos que:

$$\Pr(y_i = 0 \mid x_i) = 1 - \Phi(x_i'\beta/\sigma)$$
$$\Pr(y_i = 1 \mid x_i) = \Phi(x_i'\beta/\sigma).$$

Entonces,

$$E(y_i \mid x_i) = \Pr(y_i = 1 \mid x_i).E(y_i \mid x_i, y_i > 0)$$

3. MODELO TOBIT

INTERPRETACIÓN

- La expresión de **E(y | x, y>0)** es

$$E(y_i \mid x_i, y_i > 0) = [x_i'\beta + \sigma.\lambda(x_i'\beta/\sigma)]$$

donde

$$\lambda(z) = \phi(z)/\Phi(z)$$

Ratio inverso de Mills

Esta ecuación nos indica que estimamos un modelo de regresión lineal con las observaciones ***y>0*** no siempre conseguiremos estimaciones consistentes de **β**. El problema que tenemos es el de omisión de variables relevantes; en este caso la variable omitida sería el Ratio inverso de Mills y, generalmente, está correlacionado con los elementos de x.

3. MODELO TOBIT

INTERPRETACIÓN

Efectos marginales

1. Sobre la variable latente

$$\frac{\partial E(y^* \mid x)}{\partial x_k} = \beta_k$$

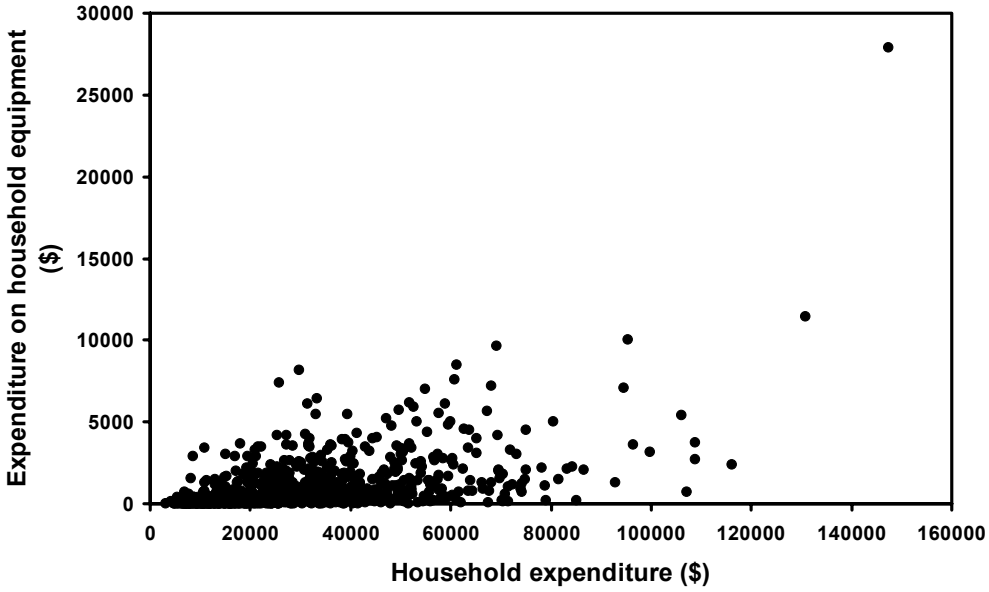
2. Sobre la variable observada sin condicionar a y>0

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

3. Sobre la variable condicionada a y>0

$$\frac{\partial E(y \mid x, y > 0)}{\partial x_k} = \beta_k + \sigma\frac{\partial \lambda}{\partial x_k}$$

EJEMPLO (C. Doughterty, 2002)



We will use the Consumer Expenditure Survey data set to illustrate the use of tobit analysis. The figure plots annual household expenditure on household equipment, *HEQ*, on total household expenditure, *EXP*, both measured in dollars.

EJEMPLO

```
. tab HEQ if HEQ<10
```

HEQ	Freq.	Percent	Cum.
0	86	89.58	89.58
3	1	1.04	90.62
4	2	2.08	92.71
6	1	1.04	93.75
7	1	1.04	94.79
8	5	5.21	100.00
Total	96	100.00	

For 86 households, *HEQ* was 0. (The tabulation has been confined to small values of *HEQ*. We are only interested in finding out how many actually had *HEQ* = 0.)

EJEMPLO

```
. reg HEQ EXP

  Source |       SS       df       MS              Number of obs =     869
---------+------------------------------           F(  1,   867) =  353.91
   Model |   729289164     1   729289164           Prob > F      =  0.0000
Residual |  1.7866e+09   867  2060635.12           R-squared     =  0.2899
---------+------------------------------           Adj R-squared =  0.2891
   Total |  2.5159e+09   868  2898456.01           Root MSE      =  1435.5

------------------------------------------------------------------------------
     HEQ |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     EXP |   .0471546   .0025065     18.813   0.000        .042235    .0520742
   _cons |  -397.2088   89.44449     -4.441   0.000      -572.7619   -221.6558
------------------------------------------------------------------------------
```

Here is a regression using all the observations. We anticipate that the coefficient of *EXP* is biased downwards.

EJEMPLO

```
. reg HEQ EXP if HEQ>0

  Source |       SS       df       MS              Number of obs =     783
---------+------------------------------           F(  1,   781) =  291.04
   Model |   656349265     1   656349265           Prob > F      =  0.0000
Residual |  1.7613e+09   781  2255219.19           R-squared     =  0.2715
---------+------------------------------           Adj R-squared =  0.2705
   Total |  2.4177e+09   782  3091656.59           Root MSE      =  1501.7

------------------------------------------------------------------------------
     HEQ |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     EXP |   .0467672   .0027414     17.060   0.000       .0413859    .0521485
   _cons |  -350.1704   101.8034     -3.440   0.001      -550.0112   -150.3296
------------------------------------------------------------------------------
```

Here is an OLS regression with the constrained observations dropped. The estimate of the slope coefficient is almost the same, just a little lower.

EJEMPLO

```
. tobit HEQ EXP, ll(0)

Tobit Estimates                                   Number of obs =    869
                                                  chi2(1)       = 315.41
                                                  Prob > chi2   = 0.0000
Log Likelihood = -6911.0175                       Pseudo R2     = 0.0223

------------------------------------------------------------------------------
         HEQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         EXP |   .0520828   .0027023    19.273   0.000     .0467789    .0573866
       _cons |  -661.8156   97.95977    -6.756   0.000    -854.0813   -469.5499
-------------+----------------------------------------------------------------
         _se |   1521.896    38.6333              (Ancillary parameter)
------------------------------------------------------------------------------

Obs. summary:         86 left-censored observations at HEQ<=0
                     783 uncensored observations
```

Here is the TOBIT regression.

EJEMPLO

```
. tobit HEQ EXP, ll(0)
------------------------------------------------------------------------------
         HEQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         EXP |   .0520828   .0027023    19.273   0.000     .0467789    .0573866
       _cons |  -661.8156   97.95977    -6.756   0.000    -854.0813   -469.5499
-------------+----------------------------------------------------------------
         _se |   1521.896    38.6333              (Ancillary parameter)
------------------------------------------------------------------------------
. reg HEQ EXP
------------------------------------------------------------------------------
         HEQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         EXP |   .0471546   .0025065    18.813   0.000      .042235    .0520742
       _cons |  -397.2088   89.44449    -4.441   0.000    -572.7619   -221.6558
------------------------------------------------------------------------------
. reg HEQ EXP if HEQ>0
------------------------------------------------------------------------------
         HEQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         EXP |   .0467672   .0027414    17.060   0.000     .0413859    .0521485
       _cons |  -350.1704   101.8034    -3.440   0.001    -550.0112   -150.3296
------------------------------------------------------------------------------
```

We see that the coefficient of *EXP* is indeed larger in the tobit analysis, confirming the downwards bias in the OLS estimates. In this case the difference is not very great. That is because only 10 percent of the observations were constrained.

3. MODELO TOBIT

LIMITACIONES

- El modelo Tobit requiere normalidad y homocedasticidad. Si cualquiera de estos dos supuestos valla, es difícil saber qué estaríamos estimando si utilizásemos MV Tobit.

- No obstante, si estos supuestos no se cumplen, pero no nos alejamos mucho de ellos, el modelo Tobit nos permite obtener buenas estimaciones.

- En un modelo Tobit, suponemos que cada x_j tiene el mismo efecto sobre $P(y>0|x)$ que sobre, que sobre $E(y|y>0,x)$…. Fijaos que sólo se estima un vector de parámetros. Esta restricción es poco realista.

- Este último problema puede resolverse planteando un modelo en dos partes (a two-part model) en el cual $P(y>0|x)$ and $E(y|y>0,x)$ tengan diferentes parámetros.

4. SESGO DE SELECCIÓN MUESTRAL

Ocurre cuando una parte de la población objetivo –con características particulares- es excluida del muestreo.

¿Cuando hay riesgo de sesgo de selección?

- **Cuando seleccionamos de forma intencionada (no aleatoria)** para que confirme nuestras opiniones.

- **Cuando la población objetivo no está bien definida:** cuando se analiza una encuesta de intención de voto, como se define la población objetivo: votantes de las elecciones pasadas que votarán en esta.

- **Cuando no incluimos a toda la población objetivo en el universo muestral.**

- **Cuando sustituimos un número ….When we subsitute a convinient number of a population for a designated member who is not readily available**.

- **Cuando la no- respuesta es relevante y los borramos de la muestra final.**

- **Cuando la muestra está basada en participantes voluntarios.**

4. SESGO DE SELECCIÓN MUESTRAL

EJEMPLO

Informe Hite (1976): *Women and Love: A cultural revolution in progress*

- 84% of women are not satisfied emotionally with their relationship
- 70% of all women married five or more years are having sex outside their marriage
- 95% of women report forms of emotional and psycological harrassment from men with whom they are in love relationship
- 84% of women reports forms of condescension from the men in their love relationship

4. SESGO DE SELECCIÓN MUESTRAL

EJEMPLO

- Aunque fue un "best seller", fue duramente criticado:
  - El error más grave fue generalizar estos resultados a todas las mujeres, hayan o no participado en la encuesta.
  - ¿Por qué no puede utilizarse la información en la que se basa este informe para generalizar?

4. SESGO DE SELECCIÓN MUESTRAL

EJEMPLO

- **Muestra obtenida con autoselección:** los cuestionarios se enviaron por correo y las receptoras decidieron voluntariamente si los cubrían o no, es decir, decidieron voluntariamente estar en la muestra o no -
  - *100.000 entrevistas enviadas; 4500 recibidas*

- Los cuestionarios se remitieron a **asociaciones de mujeres**
  - *Los puntos de vista de mujeres asociadas a un grupo particular pueden ser diferentes de los del resto de mujeres.*

- Los cuestionarios tenían **130 preguntas** y cada una de ellas incluía varios apartados:
  - *Muchas preguntas eran poco precisas por ejemplo en la forma de utilizar la palabra "amor"*
  - *Muchas sugerían claramente lo que la entrevistada debía responder*

4. SESGO DE SELECCIÓN MUESTRAL

EJEMPLO

La justificación de Shere Hite:

"Does research that is **not based on a probability or random sample** give one the right to generalize from the results of the study to the population at large? If a study is large enough and the sample broad enough, and if one generalizes carefully, yes"

Pregunta: Si realizo una encuesta sobre la eutanasia y encuesto a personas en iglesias u otros lugares de culto: ¿Puedo generalizar los resultados de la encuesta a toda la población?