INTRODUCCIÓN

Los *modelos de regresión no lineal* son aquellos no lineales en los parámetros, y que no pueden linealizarse mediante una operación matemática. Un ejemplo de modelo de regresión no lineal pero linealizable es este, basado en una función de producción de tipo Cobb-Douglas, donde y es la producción x_1 es el trabajo y x_2 el capital:

$$y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} e^u$$

que se linealiza fácilmente tomando logaritmos, con lo que tendríamos

$$\ln y = \alpha + \beta_1 \ln x_1 + \beta_2 \ln x_2 + u$$

siendo $\alpha = \ln \beta_0$.

Sin embargo, la función de Cobb-Douglas es un caso particular de la función de elasticidad constante de sustitución (CES), que no se puede linealizar:

$$y = A\left[\delta x_1^{-\beta} + (1-\delta)x_2^{-\beta}\right]^{-\left(\frac{1}{\beta}\right)}$$

donde $0<\delta<1$ y $\beta\geq 1$, y da igual que pongamos el término de error multiplicando o sumando.

Los modelos no lineales, a diferencia de los lineales, no tienen soluciones analíticas por mínimos cuadrados ordinarios. Al minimizar la suma de residuos al cuadrado para obtener las ecuaciones normales, éstas ecuaciones simultáneas se pueden resolver para obtener analíticamente los estimadores de mínimos cuadrados, pero sólo bajo el supuesto de linealidad. Si esto no se da, las incógnitas (los estimadores de los parámetros) no se pueden despejar.

Hay distintas formas de resolver el problema: mediante *prueba y error*, es decir, partiendo de unos valores iniciales de los parámetros ir variando éstos y calculando la suma de errores cuadráticos, lo que puede llevar a un mínimo local y no global; mediante *optimización directa*, también a partir de unos valores iniciales de los parámetros, que se van sustituyendo por otros mediante un proceso iterativo; y mediante *linealización iterativa*, consistente en una linealización en torno a unos valores iniciales de los parámetros (meditante un desarrollo en series de Taylor), cálculo mediante mínimos cuadrados, y vuelta a linealizar.

Los *modelos de elección discreta*, también conocidos como modelos de regresión de respuesta cualitativa, son aquellos en los que la variable dependiente adopta valores discretos, y en el caso más sencillo, binarios, normalmente 0 y 1.

La primera posibilidad es aplicar, simplemente, mínimos cuadrados ordinarios, como haríamos ante cualquier otro caso. Esto se conoce como *modelo de probabilidad lineal*. Si el modelo es

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + u$$

La esperanza condicional de *y* sería

$$E(y/\mathbf{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n = p$$

Ese valor esperado estará situado entre 0 y 1, y señala la probabilidad de que se de el valor 1, o el número de casos en que se da 1 dividido por el número de casos totales. Se puede interpretar por tanto esa esperanza como una probabilidad de *y*, que sigue una distribución de probabilidad de Bernoulli.

Al aplicar MCO sobre este modelo se presentan diversos problemas:

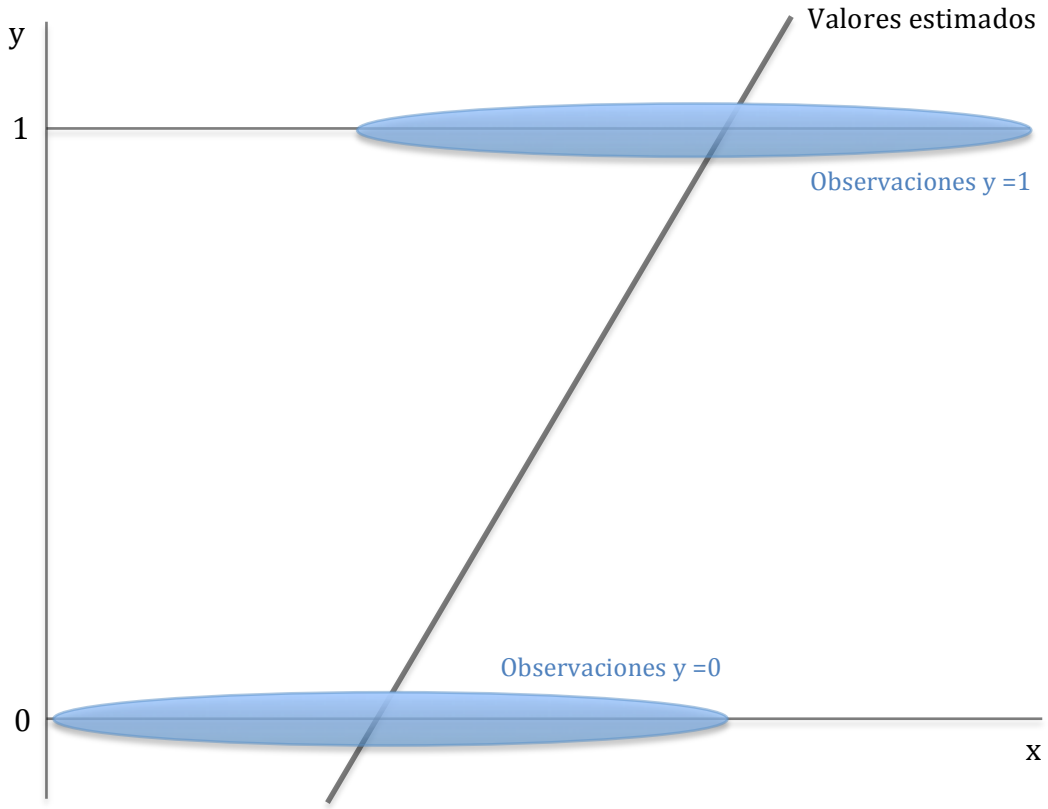
1. Tanto y como u siguen una distribución de Bernouilli, por lo que el supuesto de normalidad se pierde, lo que no es especialmente grave, sobre todo si la muestra es grande.

2. No se cumple tampoco el supuesto de homocedasticidad. En una distribución de Bernouilli la media es p (probabilidad de éxito) y la varianza es p(1-p), y sabemos que p depende de **x**. Este problema se puede tratar dividiendo cada término del modelo por $p(1-p)^{1/2}$. Primero se obtiene una estimación de y, se divide el modelo por las ponderaciones y se repite el proceso.

3. La $E(y/\mathbf{x}) = p$ no necesariamente descansa entre 0 y 1, y este es el problema más serio de este modelo.

4. El R^2 no es una buena medida de la calidad del ajuste.

Estos cuatro problemas pueden intuirse a partir del siguiente gráfico:



Una forma de forzar que el valor estimado de *y* quede comprendido entre 0 y 1 consiste en transformar el modelo original de la siguiente forma

$$p = y = G(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)$$

donde G es una función que toma valores entre 0 y 1. Esta es una función **no lineal**, lo que introduce el problema de la estimación de modelos no lineales, que hemos esbozado antes.

Tradicionalmente se han propuesto diversas funciones: la *función de distribución* de probabilidad (cdf, cumulative distribution function) *logística* en el **modelo Logit**; y la función *normal* en el **modelo normit**, también conocido como **Probit**.

La *función de distribución* de probabilidad logística es[1]

$$G(z) = \exp(z) / [1 + \exp(z)]$$

Por otro lado, en el modelo Probit la función normal es la función de distribución de probabilidad normal, que tiene forma de integral:

$$G(z) = \int_{-\infty}^{z} \phi(v)\,dv$$

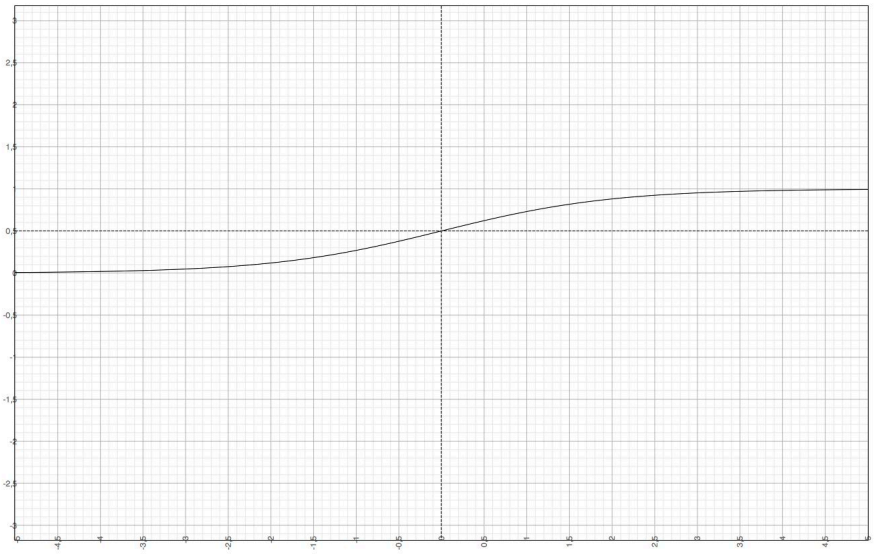
siendo $\phi(z)$ la *función de densidad* de probabilidad normal (con forma de campana de Gauss):

$$\phi(z) = (2\pi)^{-1/2} \exp(-z^2/2)$$

Como puede verse, en el caso normal no hay solución analítica general, lo que hace necesario recurrir a la *simulación* para estimar.

Ambas funciones G(z) nos dan un valor entre 0 y 1 para todo número real *z*. Se trata de una función creciente, que tiende a cero conforme a z tiende a $-\infty$ y a 1 conforme z tiene a ∞. El gráfico sería similar a este:

[1] Sobre la historia del desarrollo del modelo Logit es de gran interés la lección de McFadden (2001) escrita con ocasión de su recogida del premio Nobel.



Como puede observarse

$$p = y = G(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)$$

no contiene un término de error al uso. El ingrediente estocástico lo proporciona la función G.

Una forma de interpretar el modelo *Logit* y el *Probit* de forma más tradicional, con un término de error explícito, es pensar que hay una *variable latente* y* que no observamos y que está relacionada *linealmente* con los regresores:

$$y^* = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + e$$

donde *e* sigue la distribución logística o la normal, ambas simétricas en torno a cero. Esto nos permite aplicar la propiedad según la cual 1 – G(-z) = G(z). El caso es que suponemos que cuando y* toma un valor estrictamente positivo la variable observada, y, toma un valor igual a 1; cuando y* tiene un valor menor o igual a cero y toma un valor igual a 0. La probabilidad de una y otra cosa debe ser la misma, es decir

$$p(y=1/\mathbf{x}) = p(y^*>0/\mathbf{x})$$

donde obviamente p es una *función de distribución de probabilidad* (cdf), normalizada a una varianza igual a 1 en el caso de que se trate de la normal.

Suponemos que el modelo real es el que relaciona la variable latente con los regresores, aunque dicha variable, y*, está claramente conectada con la variable observable *y*. Conociendo el modelo latente real podemos desarrollar la expresión anterior:

$$p(y=1/\mathbf{x}) = p(y^*>0/\mathbf{x})$$

$$= p[e>-(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)/\mathbf{x}]$$

$$=1 - G[-(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)]$$

$$= G[\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n]$$

donde *e*, como sabemos, es una variable que sigue una distribución logística o normal y G(z) es una función de distribución de probabilidad logística o normal, que es aquella que nos da la probabilidad de que z tome un valor menor o igual que z_0, es decir, $G(z=z_0) = p(z \leq z_0)$. También podemos ver que

$$p[e>-(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)/\mathbf{x}]$$

$$= p[-e<(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)/\mathbf{x}]$$

$$=G(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n)$$

donde G es obviamente una *función de distribución de probabilidad* (cdf), que para –*e* coincide con la que correspondería a *e* si la distribución fuera simétrica en torno a cero. Recordemos que estas funciones nos dan la probabilidad de observar un valor menor o igual a *e*. Si *e* sigue una distribución estándar normal, la cdf será la normal estándar; si *e* sigue una distribución logística (también simétrica en torno a cero), la cdf será la propia de una logística.

Cada regresor x_j está relacionado de forma latente con la variable *y**, y no con la variable observada *y* (probabilidad de éxito), lo que resulta complejo debido a la naturaleza no lineal de G(z).

Hay dos formas de conectar la variable latente con el modelo de respuesta binaria. Primero, la variable latente puede considerarse como un índice de la propensión a ocurrir de un evento determinado. Segundo, la variable latente puede ser la diferencia en utilidad que se da si el evento en cuestión ocurre efectivamente (bajo

el supuesto de que el resultado binario es producto de la elección individual)[2]. Será este segundo enfoque el que sigamos para el estudio profundo de los modelos de elección discreta.

La interpretación en términos de una variable latente tiene varias ventajas: se pueden interpretar los resultados en términos de *modelo de utilidad aleatorio*, muy atractivo teóricamente y base de la generalización multinomial y para los modelos de variable censurada; facilita también la comprensión de los modelos Logit y Probit ordenados; facilita la modelización de los problemas de selección muestral; y permite el desarrollo de medidas similares al R^2 de mínimos cuadrados.

Tabla 1. *Una clasificación tradicional de los modelos*

Nº de alternativas	**Tipo de alternativas**	**Tipo de función**	**Los regresores representan:**	
			Características de los individuos	**Atributos de las alternativas**
Modelos de respuesta dicotómica (2 alternativas)	Complementarias	Lineal	Modelo de probabilidad lineal	
		Logística	Logit	
		Normal	Probit	
Modelos de respuesta múltiple (3 alternativas o más)	No ordenadas	Logística	Logit multinomial (anidado o mixto)	Logit condicional (anidado o mixto)
		Normal	Probit multinomial o multivariante	Probit multinomial o multivariante
	Ordenadas	Logística	Logit ordenado	
		Normal	Probit ordenado	

Los efectos marginales en el modelo Logit son fáciles de calcular, ya que $\partial p_i/\partial x_{ij} = p_i(1-p_i)\beta_j$, donde $p_i = G(z_i)$, siendo $z_i = x_i\beta$. Se trata del efecto de un cambio en x_{ij} sobre p. Es muy común la interpretación de los coeficientes en términos de los efectos marginales sobre el *odds ratio*, que es p/(1-p), la probabilidad de que y=1 dividida por la probabilidad de que y=0. En efecto, tenemos

$$p = \exp(z) / [1 + \exp(z)]$$

$$\Rightarrow p/(1-p) = \exp(z)$$

$$\Rightarrow \ln [p/(1-p)] = z = \mathbf{x}\beta$$

[2] Esta segunda opción obliga a diferenciar entre regresores que varían entre individuos (s_n) y aquellos que tienen que ver con las alternativas y los afectan a todos por igual (x_n).

Si un estudio farmacológico tiene como posibles resultados sobrevivir (y=1) o morir (y=0) tras tomar un medicamento, un *odds ratio* de 2 significa que la probabilidad de lo primero duplica la probabilidad de que ocurra lo segundo. Un incremento de 1 en el regresor j-ésimo, para un $\beta_j = 0{,}1$, multiplica el *odds ratio* por $\exp(0{,}1) \approx 1{,}105$, es decir, la probabilidad relativa de supervivencia crece en un 10,5%. Otro ejemplo típico de este tipo de análisis es identificar si un cráneo encontrado en un yacimiento pertenece a un tipo de individuo o a otro, en función de una serie de características de los mismos recogidas en los regresores (tamaños, pesos, etc.).

Los efectos marginales del modelo Probit son más difíciles de calcular. Tenemos que $\partial p_i/\partial x_{ij} = \phi(z)\,\beta_j = \phi(G^{-1}(z))$, donde $p_i = G(z_i)$, siendo $z_i = x_i\beta$.

Tabla 2. *Efectos marginales de cada modelo básico.*

Modelo	Probabilidad de *y*=1 condicionado a los regresores **x**	Efecto marginal, $\partial p/\partial x_j$
Probabilístico lineal	$z = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$	β_j
Logit	$G(z) = \exp(z) / [1 + \exp(z)]$	$G(z)[1-G(z)]\beta_j$
Probit	$G(z) = \int_{-\infty}^{z} \phi(v)\,dv$	$\phi(z)\,\beta_j$

En definitiva, es fácil recordar estos efectos marginales si pensamos que en los tres casos estamos calculando el efecto es:

$$\frac{\partial E(y_i / X_i)}{\partial X_j}$$

En el caso del modelo de probabilidad lineal tenemos que $E(y_i/x_i) = x_i\beta$, mientras que en caso del Logit y el Probit tenemos que $E(y_i/x_i) = G(x_i\beta)$, donde G(.) es la función de distribución logística en el primer caso, y la normal en el segundo. Por tanto, cuando nos preguntamos acerca de los efectos marginales estamos buscando el efecto de cualquier cambio en una de las variables x_i sobre la variable dependiente, esto es, una derivada. Las tres derivadas (modelo probabilístico, Logit y Probit) se calculan directamente:

$$\frac{\partial E(y_i / X_i)}{\partial X_j} = \beta_j$$

$$\frac{\partial E(y_i / x_i)}{\partial x_j} = \frac{\exp(x_i\beta)}{[1+\exp(x_i\beta)]^2}\beta_j$$

$$\frac{\partial E(y_i / X_i)}{\partial X_j} = f(x_i\beta)\beta_j = \phi(z)\beta_j$$

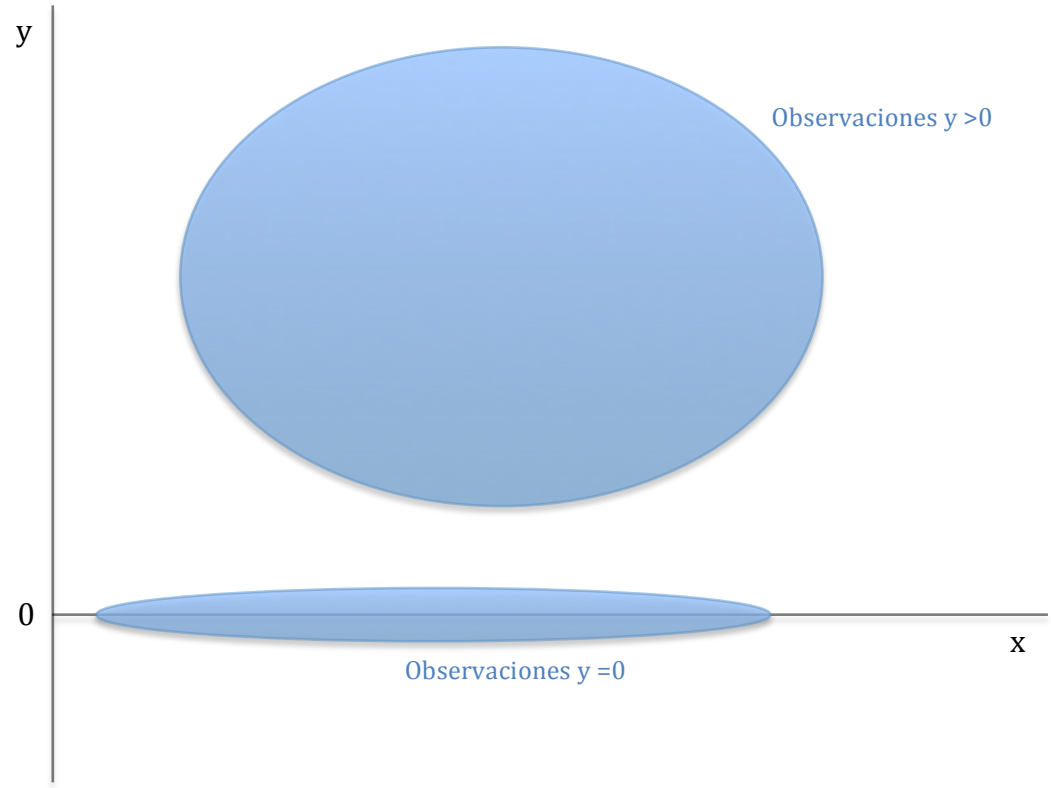
donde, en la última ecuación, f(.) (=ϕ(.)) es la función de densidad de la normal. Obsérvese, en la segunda expresión, que siendo G(z) = exp(z) / [1 + exp(z)], tendremos que G(z)[1-G(z)] = exp(z) / [1 + exp(z)]^2.

Amemiya (1981) propone unas razones de conversión entre los coeficientes de cada uno de los modelos:

- Coef. Probit = Coef. Logit * 0,625
- Coef. Probabilístico = Coef. Logit * 0, 25 [excepto constante]
- Coef. Probabilístico = Coef. Logit * 0, 25 + 0,5 [sólo constante]
- Coef. Probit = Coef. Probabilístico * 2,5 [excepto constante]
- Coef. Probit = Coef. Probabilístico * 2,5 – 1,25 [sólo constante]

No todas las variables dependientes limitadas son binarias, es decir, toman sólo dos valores, cero o uno. También son variables dependientes limitadas aquellas que toman valores dentro de un rango limitado. Los modelos Logit y Probit permiten tratar con variables dependientes binarias, o bien con variables dependientes que toman más de dos valores discretos. Sin embargo, cuando tratamos problemas económicos reales nos encontramos con lo que se conoce en microeconomía como *soluciones de esquina*, es decir, agentes económicos que deciden demandar u ofertar una cantidad nula de un determinado bien o servicio, mientras que otras deciden optar por una cantidad positiva. Cuando la variable dependiente que pretendemos explicar acumula una cantidad importante de ceros. Esto implica que, si usamos mínimos cuadrados ordinarios, podamos obtener predicciones para la variable dependiente negativas; o bien que no podamos emplear logaritmos. Como es evidente, la variable dependiente no puede seguir una distribución normal, simétrica, dado que hay una gran acumulación de valores en el cero. El modelo Tobit está diseñado para tratar con este tipo de variables.

Como puede verse en el siguiente gráfico, podemos encontrarnos con dos grupos de datos. Unos para los que la variable dependiente y alcanza valores positivos, y otros para los que el valor de ésta es siempre cero, para cualquier valor de las variables explicativas. Podemos plantearnos realizar la estimación sólo con el primer grupo de datos, pero es obvio que al no considerar el segundo las estimaciones de los parámetros estarán sesgadas y serán inconsistentes.



Ante un caso como este podemos emplear el método de máxima verosimilitud para desarrollar un procedimiento de estimación (el *modelo Tobit*) que nos permita tratar directamente con este tipo de datos. Tanto el modelo de Heckit como el modelo Tobit dan estimaciones consistentes de los parámetros, si bien el Tobit proporciona estimaciones más consistentes aún.

El modelo Tobit puede entenderse mejor si hacemos depender la variable dependiente (y) de una variable *latente* (y*), de forma que

$$y^* = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + u \qquad u/\mathbf{x} \sim N(0, \sigma^2)$$

Siendo

$$y = \max(0, y^*)$$

La variable latente cumple con todos los supuestos del modelo lineal, incluyendo normalidad, homocedasticidad y linealidad. Nosotros no observamos la variable latente *y**, sino *y*. La variable observada será igual a la latente cuando latente sea $y^* \geq 0$, pero será igual a cero si la variable latente es $y^* < 0$. Recordamos que *y** sigue una distribución normal, por lo que *y* reflejará esa misma distribución, pero sólo para la parte en que los valores son positivos. Tenemos también que la probabilidad de observar un valor para y = 0 dados los valores de las variables explicativas será:

$$P(y=0 \mid \mathbf{x}) = P(y^*<0 \mid \mathbf{x}) = P(u<-\mathbf{x\beta} \mid \mathbf{x}) = P(u/\sigma < -\mathbf{x\beta}/\sigma \mid \mathbf{x})$$

$$= \Phi(-\mathbf{x\beta}/\sigma) = 1 - \Phi(\mathbf{x\beta}/\sigma)$$

donde u/σ es una variable aleatoria con distribución normal estandarizada (media cero, varianza igual a 1). El término constante está incluido dentro de la suma $\mathbf{x\beta}$.

Por tanto, para valores de y > 0 la función de densidad de y dado el vector x es la función de densidad de una normal u/σ, donde $\mathbf{x\beta}$ nos da la media de y, es decir:

$$(1/\sigma)\Phi[(y - x_i\boldsymbol{\beta})/\sigma] = (2\pi\sigma^2)^{-1/2} \exp[-(y - x_i\boldsymbol{\beta})^2/(2\sigma^2)]$$

Ahora necesitamos el logaritmo de la función de verosimilitud para cada observación i, y ello se obtiene sumando para los dos casos posibles

$$\ell_i(\boldsymbol{\beta}, \sigma) = 1(y_i = 0)\log[1 - \Phi(x_i\boldsymbol{\beta}/\sigma)] + 1(y_i > 0)\log\{ (1/\sigma)\Phi[(y_i - x_i\boldsymbol{\beta})/\sigma] \}$$

Esta función se maximiza para obtener los $\boldsymbol{\beta}$ y σ, lo que se hace mediante métodos numéricos, no analíticos. Con programas y ordenadores estos cálculos son muy rápidos.

Se presenta un problema con la interpretación de los parámetros $\boldsymbol{\beta}$, que están ligados a la variable latente y no a la observable. Quiere esto decir que el parámetro β_j mide el efecto parcial de x_j sobre $E(y^*|\mathbf{x})$, donde y^* es la variable latente. Esta variable no siempre tiene un significado económico interpretable. La variable observable y sí. Normalmente estaremos interesados en $E(y|\mathbf{x})$, que no conocemos. Lo que sí tenemos es $E(y \mid y>0, \mathbf{x})$, y a partir de ella podemos obtener $E(y|\mathbf{x})$:

$$E(y|\mathbf{x}) = P(y>0 \mid \boldsymbol{x})\, E(y \mid y>0, \boldsymbol{x}) = \Phi(\boldsymbol{x\beta}/\sigma)\, E(y \mid y>0, \boldsymbol{x})$$

Tenemos $E(y \mid y>0, \mathbf{x})$ porque podemos aplicar la siguiente propiedad. Si $z \sim N(0,1)$ tendremos que $E(z \mid z>c) = \phi(c)/[1-\Phi(c)]$ para cualquier constante c. Además tenemos que $\phi(-c) = \phi(c)$ y $1-\Phi(-c) = \Phi(c)$. Por tanto, considerando que u/σ es una normal estándar e independiente de $\mathbf{x}$,

$$E(y \mid y>0, \boldsymbol{x}) = \boldsymbol{x\beta} + E(u \mid u> -\boldsymbol{x\beta}) = \boldsymbol{x\beta} + \sigma E[(u/\sigma)|(u/\sigma)>-\boldsymbol{x\beta}/\sigma]$$

$$= \boldsymbol{x\beta} + \sigma\phi(\boldsymbol{x\beta}/\sigma)/\Phi(\boldsymbol{x\beta}/\sigma)$$

Que podemos sintetizar bajo la forma:

$$E(y \mid y>0, \boldsymbol{x}) = \boldsymbol{x\beta} + \sigma\lambda(\boldsymbol{x\beta}/\sigma)$$

donde $\lambda(\boldsymbol{x\beta}/\sigma) = \phi(\boldsymbol{x\beta}/\sigma)/\Phi(\boldsymbol{x\beta}/\sigma)$, lo que se conoce como el *cociente inverso de Mills*, es decir, el cociente de la función de densidad y de la función de distribución normal estándar, para el valor determinado por $\boldsymbol{x\beta}/\sigma$. Este cociente se emplea en el *modelo de Heckit*, como hemos apuntado y veremos con detalle más adelante.

La interpretación de la expresión anterior es la siguiente: el valor esperado de *y* condicionado a que y sea positivo es $\boldsymbol{x\beta}$ más un elemento positivo (σ multiplicado por el *cociente inverso de Mills* evaluado en el punto $\boldsymbol{x\beta}/\sigma$.) Queda claro que si usamos mínimos cuadrados sólo para las observaciones para las que y>0 tendremos estimaciones inconsistentes de $\boldsymbol{\beta}$. Falta considerar el *coeficiente inverso de Mills*, que es una especie de variable omitida correlacionada con los elementos de **x**.

Ahora podemos volver al cálculo de $E(y|\mathbf{x})$. Recordamos que

$$E(y|\mathbf{x}) = \Phi(\boldsymbol{x\beta}/\sigma)\, E(y \mid y>0, \boldsymbol{x})$$

y por tanto, sustituyendo $E(y \mid y>0, \boldsymbol{x})$, tenemos:

$$E(y|\mathbf{x}) = \Phi(\boldsymbol{x\beta}/\sigma)\, [\boldsymbol{x\beta} + \sigma\lambda(\boldsymbol{x\beta}/\sigma)] = \Phi(\boldsymbol{x\beta}/\sigma)\, \boldsymbol{x\beta} + \sigma\phi(\boldsymbol{x\beta}/\sigma)$$

El resultado apunta a lo siguiente: cuando *y* tiene las características para las que se ha diseñado el modelo Tobit, $E(y|\mathbf{x})$ depende *no linealmente* de $\boldsymbol{x}$ y de $\boldsymbol{\beta}$. Dado que los dos sumandos de la ecuación anterior son positivos, los $\boldsymbol{\beta}$ estimados nos llevarán a predicciones de *y* –es decir, valores de $E(y|\mathbf{x})$– también positivos, en todo caso. Una estimación lineal (mínimos cuadrados) nos podía llevar a valores de y negativos. Además de la complicación de la estimación no-lineal, reflejada en la expresión para $E(y|\mathbf{x})$, los efectos marginales de las x_j sobre *y* son complicados de interpretar, ya que si bien el signo dependerá del coeficiente β_j, la magnitud del efecto dependerá también del resto de variables explicativas y sus parámetros, y también de σ.

Estos efectos marginales de las variables explicativas sobre la explicada se calculan de la forma usual: derivando. Tenemos por tanto que

$$\partial E(y \mid y>0, \boldsymbol{x})/\partial x_j = \beta_j + \beta_j\, [d\lambda/dc]\, (\boldsymbol{x\beta}/\sigma)$$

siempre que x_j no esté relacionada con las demás explicativas. Ahora hay que aplicar varias reglas para poder pasar al siguiente paso. Primero, recordamos que $\lambda(c) = \phi(c)/\Phi(c)$. Además, sabemos que $d\Phi/dc = \phi(c)$. Por otro lado, tenemos que

$d\phi/dc = - c\phi(c)$. Todo eso nos permite asegurar que $d\lambda/dc = -\lambda(c)[c+\lambda(c)]$. De ahí obtenemos que

$$\partial E(y \mid y>0, \boldsymbol{x})/\partial x_j = \beta_j\{1 - \lambda(\boldsymbol{x\beta}/\sigma)[\boldsymbol{x\beta}/\sigma + \lambda(\boldsymbol{x\beta}/\sigma)]\}$$

Por tanto, el efecto marginal de x_j sobre $E(y \mid y>0, \boldsymbol{x})$ depende de β_j, pero multiplicado por un factor de corrección (entre corchetes) que se mueve entre cero y uno, y en el que está implicado todo el vector $\mathbf{x}$ con los correspondientes parámetros $\boldsymbol{\beta}$. El efecto marginal se puede estimar empleando para ello las estimaciones de máxima verosimilitud de β_j y σ. Obsérvese que la estimación de σ es absolutamente necesaria.

De igual forma que hemos calculado los efectos marginales, podemos calcular también las elasticidades. La elasticidad de y con respecto a x_1, condicionada a que $y>0$, será:

$$\frac{\partial E(y \mid y>0, x)}{\partial x_1} \frac{x_1}{E(y \mid y>0, x)}$$

El efecto marginal puede ser distinto si la variable explicativa no es la típica variable continua sino, por ejemplo, una variable binaria. En ese caso el efecto marginal sería el efecto sobre $E(y \mid y>0, \boldsymbol{x})$ cuando x_1 pasa de valer 0 a valer 1. Lo mismo podría plantearse para variables discretas que adoptan más de dos valores. Para ello se observa simplemente la diferencia entre $E(y \mid y>0, \boldsymbol{x})$ para los casos en los que $x_1 = 0$ y $x_1 = 1$.

Hasta el momento hemos considerado los efectos en $E(y \mid y>0, \boldsymbol{x})$, es decir sobre el valor esperado de la variable dependiente en tanto que esta toma valores positivos. Pero estaremos interesados en el efecto marginal de una variable explicativa continua sobre $E(y|\mathbf{x})$, es decir, considerando las variables explicativas en su conjunto, lo que implica la posibilidad de que al variar la variable explicativa la variable dependiente, en respuesta, "salte" de su valor cero a uno positivo, o al revés. Veamos el cálculo de este efecto marginal.

Volvemos por un momento a la expresión

$$E(y|\mathbf{x}) = \Phi(\boldsymbol{x\beta}/\sigma)\,\boldsymbol{x\beta} + \sigma\phi(\boldsymbol{x\beta}/\sigma)$$

Ahora calculamos la derivada parcial respecto a la variable continua x_j:

$$\partial E(y|\mathbf{x})/\partial x_j = [\partial P(y>0, \boldsymbol{x})/\partial x_j]\, E(y \mid y>0, \boldsymbol{x}) + P(y>0, \boldsymbol{x})\, [\partial E(y \mid y>0, \boldsymbol{x})/\partial x_j]$$

Que se puede leer como la suma del valor esperado de *y* cuando es positiva multiplicado por cómo cambia la probabilidad de que *y* sea positiva, más la probabilidad de que *y* sea positiva multiplicada por cómo cambia el valor esperado cuando es positiva. Ahora, recordando que $P(y>0, \boldsymbol{x}) = \Phi(\boldsymbol{x\beta}/\sigma)$, tendremos que

$$\partial P(y>0, \boldsymbol{x})/\partial x_j = (\beta_j/\sigma)\phi(\boldsymbol{x\beta}/\sigma)$$

Por tanto, todos los términos que aparecen en la expresión $\partial E(y|\mathbf{x})/\partial x_j$ se pueden estimar, con sólo sustituir las estimaciones por máxima verosimilitud de β_j y σ, y dar valores a x_j.

Si tomamos junto a la expresión anterior

$$\partial E(y \mid y>0, \boldsymbol{x})/\partial x_j = \beta_j\{1 - \lambda(\boldsymbol{x\beta}/\sigma)[\, \boldsymbol{x\beta}/\sigma + \lambda(\boldsymbol{x\beta}/\sigma)]\}$$

y sustituimos en $\partial E(y|\mathbf{x})/\partial x_j$, recordando que $\Phi(c)\,\lambda(c) = \phi(c)$, tendremos

$$\partial E(y|\mathbf{x})/\partial x_j = \beta_j\Phi(\boldsymbol{x\beta}/\sigma) = \hat{\gamma}_j$$

Esa expresión se puede obtener también desde $E(y|\mathbf{x}) = \Phi(\boldsymbol{x\beta}/\sigma)\ \boldsymbol{x\beta} + \sigma\phi(\boldsymbol{x\beta}/\sigma)$, recordando que $d\phi(z)/dz = -z\phi(z)$. Los $\hat{\gamma}_j$ son los coeficientes de una regresión por mínimos cuadrados de y_i sobre $x_{i1}, x_{i2},\ldots, x_{ik}$, para i = 1,...,n, usando todos los datos disponibles. Para conectar el coeficiente del modelo Tobit ($\hat{\gamma}_j$) con $\hat{\gamma}_j$ hay que multiplicar el primero por un factor de corrección *estimado*. Ese factor es $\Phi(\bar{x}\hat{\beta}/\hat{\sigma})$, que se encuentra entre cero y uno y que puede considerarse también una estimación (consistente) de $P(y>0|\mathbf{x})$ cuando asignamos a cada variable explicativa la media muestral. Conforme $P(y>0|\mathbf{x}=\bar{x})$ tiende a 1, los coeficientes Tobit y mínimos cuadrados se acercan, y cuando $y_i>0$ para todo i los coeficientes coinciden plenamente.

Surge un problema cuando las variables explicativas son discretas, y no continuas, como variables binarias. En ese caso el cálculo del efecto marginal se complica, y la comparación entre los métodos Tobit y mínimos cuadrados ya no es tan sencilla. En el caso de un Tobit el efecto marginal es la diferencia en la estimación $E(y|\mathbf{x})$

mediante la expresión $E(y|\mathbf{x}) = \Phi(\boldsymbol{x\beta}/\sigma)\,\boldsymbol{x\beta} + \sigma\phi(\boldsymbol{x\beta}/\sigma)$, es decir, si x_1 es la variable binaria, habría que calcular $E(y|\mathbf{x})$ para $x_1=0$ y después para $x_1=1$, quedando el resto de variables en sus medias muestrales o en cualquier otro valor relevante al caso de estudio.

El modelo Tobit es sensible a los problemas de no normalidad y heterocedasticidad, que no afectaban a la insesgadez ni a la consistencia de los estimadores mínimo cuadráticos, al menos en muestras grandes (la heterocedasticidad sí afecta a la inferencia, por lo que deben usarse estimadores robustos). En cambio, las expresiones $E(y|\mathbf{x}) = \Phi(\boldsymbol{x\beta}/\sigma)\,\boldsymbol{x\beta} + \sigma\phi(\boldsymbol{x\beta}/\sigma)$ y $E(y \mid y>0, \boldsymbol{x}) = \boldsymbol{x\beta} + \sigma\lambda(\boldsymbol{x\beta}/\sigma)$ sí dependen de los supuestos de normalidad y homocedasticidad. Si esos supuestos no se cumplen el estimador de máxima verosimilitud no "funciona" según lo esperado. Si los "incumplimientos" no son por un margen muy amplio las estimaciones de los efectos parciales pueden ser válidas.

Hay otro problema relativamente serio con el modelo Tobit. El efecto de una variable explicativa cualquiera x_j sobre $P(y>0|\mathbf{x})$ y sobre $P(y|y>0, \mathbf{x})$ es igualmente proporcional a β_j y a unas funciones que lo multiplican:

$$\partial E(y \mid y>0, \boldsymbol{x})/\partial x_j = \beta_j\{1 - \lambda(\boldsymbol{x\beta}/\sigma)[\,\boldsymbol{x\beta}/\sigma + \lambda(\boldsymbol{x\beta}/\sigma)]\}$$

$$\partial P(y>0, \boldsymbol{x})/\partial x_j = (\beta_j/\sigma)\phi(\boldsymbol{x\beta}/\sigma)$$

Por así decir, un solo mecanismo determina la elección entre y = 0 y y > 0 y la cantidad de y una vez sabemos que y > 0, es decir, $\partial E(y \mid y>0, \boldsymbol{x})/\partial x_j$ y $\partial P(y>0, \boldsymbol{x})/\partial x_j$, que siempre han de tener el mismo signo en un Tobit.

Esto hace imposibles determinados tipos de análisis. Por ejemplo, si analizamos la relación entre la edad de una persona (x_1, variable continua) y la cobertura de un seguro privado de vida (y, variable dependiente). Cuanto mayor es x_1 (edad) mayor es la probabilidad de que y>0 (se tenga un seguro de vida, por una cantidad positiva). Si condicionamos a tener un seguro de vida (damos ya y>0 como cierto), el efecto de x_1 sobre la esperanza matemática de y puede ser negativo, porque aquellos que ya tienen un seguro y son muy mayores pueden tener pocos incentivos a aumentar la cobertura. El modelo Tobit no está diseñado para tener en cuenta casos como este en los que **x** tiene efectos contrarios sobre la probabilidad y la esperanza matemática.

Un truco para ver si el uso de un Tobit es apropiado al caso que nos ocupa es estimar antes un Probit para una variable dependiente alterada, que adopta valor 1 si y>0. El coeficiente de x_j en este Probit sería $\gamma_j = \beta_j/\sigma$. Por tanto, con un Probit podemos estimar ese cociente (aunque no los dos parámetros por separado), y paralelamente hacerlo mediante un Tobit, y comparar. Si el modelo Tobit es

apropiado, las estimaciones del Probit deben ser cercanas a lo que sería una estimación del cociente β_j/σ a partir del Tobit. Nunca serán iguales, pero sí podemos sospechar problemas con el Tobit si hay discrepancias con los signos de los coeficientes significativos, o diferencias importantes en la magnitud de los mismos (si son significativos). Si se encuentra que el Tobit no es apropiado, habría que emplear modelos alternativos, como los *de barrera* (hurdle models) o *en dos partes* (two-tiered models) (Wooldridge, 2002, capítulo 16.7).

Un caso distinto es aquel en el que el Tobit no es aplicable por problemas de truncamientos en la muestra de datos. Un ejemplo permite ver la naturaleza del problema con claridad, tomado de Gronau (1974). Cuando queremos explicar el salario con una serie de variables, como la educación, la edad o la experiencia, nos encontramos con el problema de aquellos que deciden no trabajar. Su salario no es exactamente cero. Simplemente, no los tenemos en la muestra. Para las personas que no trabajan podemos observar las explicativas, pero no la dependiente (el salario). Esto se conoce como truncamiento incidental de la muestra, pues la disponibilidad o no de la variable salario depende de otra variable, que es la participación laboral.

Este problema, que es muy común, se trata mediante la inclusión de una *ecuación de selección* (la segunda):

$$y = \mathbf{x}\boldsymbol{\beta} + u \qquad E(u|\mathbf{x}) = 0$$

$$s = 1[\mathbf{z}\boldsymbol{\gamma} + v \geq 0]$$

donde s = 1 si ocurre que y se observa, pero s = 0 si y no se observa. Las variables explicativas **x** y **z** son todas observables en cualquier caso.

Estamos interesados en estimar la primera ecuación, pero y no se observa para algunos individuos en la muestra. Se supone que en esta ecuación **z** es exógena, es decir $E(u \mid \mathbf{x}, \mathbf{z}) = 0$. Se supone también que **x** es un subconjunto estricto de **z**, es decir, todas las **x** pertenecen a **z**, pero no al revés. Obviamente, en **z** hay más casos, pues se incluyen aquellos para los que y no es observable. Suponemos además que v es independiente de **z** (y por tanto de **x**) y que sigue una distribución normal estándar.

Las variables aleatorias *u* y *v* están correlacionadas. En efecto, si (*u*,*v*) es independiente de **z**, teniendo en cuenta además que **x** es un subconjunto de **z**, tendremos

$$E(y|\mathbf{z},v) = \mathbf{x}\boldsymbol{\beta} + E(u|\mathbf{z}, v) = \mathbf{x}\boldsymbol{\beta} + E(u|v)$$

El supuesto de independencia antedicho explica que $E(u|\mathbf{z}, v) = E(u|v)$. Además, si u y v siguen una distribución normal multivariante (con media cero), tendremos que $E(u|v) = \rho v$ para algún parámetro ρ. Esto nos lleva a

$$E(y|\mathbf{z},v) = \mathbf{x}\boldsymbol{\beta} + \rho v$$

Empleamos esa ecuación para calcular $E(y|\mathbf{z},s)$, para el caso en que s = 1, de forma que

$$E(y|z,s) = \mathbf{x}\boldsymbol{\beta} + \rho\, E(v|\mathbf{z},s)$$

Ahora, considerando la relación que une a *s* y *v*, y teniendo en cuenta que *v* sigue una normal, tenemos que $E(v|\mathbf{z},s)$ es el *cociente inverso de Mills* $\lambda(\mathbf{z}\boldsymbol{\gamma})$ cuando s = 1. Por tanto, podemos escribir,

$$E(y \mid z, s = 1) = \mathbf{x}\boldsymbol{\beta} + \rho\, \lambda(\mathbf{z}\boldsymbol{\gamma})$$

Tenemos unos valores de *y* observados y otros no. Pues bien, dados los valores observados de *y* y **z**, el valor esperado de *y* es $\mathbf{x}\boldsymbol{\beta}$ más un término que depende del *cociente inverso de Mills* para un valor $\mathbf{z}\boldsymbol{\gamma}$**.** De esta forma, podemos estimar $\boldsymbol{\beta}$, y nos basta para ello con la información parcial que tenemos de *y*, más el término $\lambda(\mathbf{z}\boldsymbol{\gamma})$ como regresor.

Cuando *u* y *v* están incorrelacionados ρ = 0 ¿Qué pasa si ρ = 0? En ese caso $\lambda(\mathbf{z}\boldsymbol{\gamma})$ no aparecería, y bastaría con estimar por mínimos cuadrados *y* (la submuestra de los observados) con **x** de regresor. Los estimadores de $\boldsymbol{\beta}$ serían consistentes. Si $\rho \neq 0$ no podemos omitir la variable $\lambda(\mathbf{z}\boldsymbol{\gamma})$, que está además correlacionada con **x**.

Por tanto, el problema de truncamiento en la muestra se reduce a un problema de omisión de variable, que puede superarse introduciendo el término $\lambda(\mathbf{z}\boldsymbol{\gamma})$.

El problema que se presenta es que no conocemos el vector de parámetros $\boldsymbol{\gamma}$, por lo que no podemos calcular $\lambda(\mathbf{z}\boldsymbol{\gamma})$. Bien, pero bajo los supuestos establecidos, *s* dado **z** sigue un modelo Probit:

$$P(s = 1 \mid \mathbf{z}) = \Phi(\mathbf{z}\gamma)$$

Se estima γ mediante un Probit con el que explicamos *s* a partir de $\mathbf{z}$ usando la muestra completa (aquellos con la y observada y los que no). Después se estima β. Es lo que se conoce como *método Heckit*, por Heckman (1976).

En síntesis, el procedimiento tiene dos pasos: primero se emplean todas las observaciones para regresar s sobre z y obtener las estimaciones de los γ mediante un Probit, para después emplearlas en el cálculo de la inversa de Mills, que en este caso sería también un vector $\hat{\lambda} = \lambda(z\gamma)$, del que sólo tomaremos los componentes para los que s = 1; el segundo paso consiste en tomar la submuestra para la que s = 1 y calcular con ella la regresión de *y* sobre $\mathbf{x}$ e $\hat{\lambda}$.

Obtendremos así unas estimaciones de β consistentes y distribuidos aproximadamente como una normal. Además, podemos hacer un test de significatividad de la variable $\hat{\lambda}$ en la segunda regresión, para contrastar la hipótesis nula según la cual $\rho = 0$. Si la hipótesis se cumple no hay problemas de selección muestral. Sin embargo, cuando $\rho \neq 0$ los errores estándar derivados de la segunda regresión no son correctos, pues entre los regresores hay una variable (la estimación inversa del Mills), para cuyo cálculo se han empleado las mismas variables después utilizadas en la regresión, con algunas más (puesto que z contiene a x). Algunos programas econométricos aplican una corrección bastante complicada que permite llegar a unos errores estándar correctos.

En vez de la inversa de Mills se pueden emplear en la segunda regresión los residuos de una estimación Tobit aplicada a los casos en los que $y_i > 0$. Se consiguen estimadores consistentes para las β, y al igual que en el procedimiento anterior, la t de student de la variable formada por los residuos estimados puede servir para un contraste válido de la hipótesis de existencia de un sesgo de selección muestral (para más detalles véase Wooldridge, 2002, capítulo 17).

Existe una alternativa a este procedimiento en dos etapas conocido como *modelo Heckit*. Consiste en la estimación por máxima verosimilitud de las dos ecuaciones, la de regresión y la de selección (véase también Wooldridge, 2002, capítulo 17).