MODELO LOGIT

Cartolin, More, Quispe, Rios, Tamariz, Vargas.

August 22, 2016

Abstract

Es posible diseñar modelos en donde la variable dependiente posea característica cualitativas, ese es el caso que analizaremos en el presente trabajo, enfocándonos únicamente en el modelo LOGIT que nos brinda ciertas ventajas en comparación a un modelo lineal de probabilidad, estimada por mínimos cuadrados ordinarios(MCO) para lo cual resaltaremos dichas diferencias. Los modelos de regresión con respuesta cualitativa son modelos de regresión en los cuales la variable dependiente puede ser de naturaleza cualitativa, mientras que las variables independientes pueden ser cualitativas o cuantitativas, o una mezcla de las dos; por ejemplo, si se está estudiando la relación entre ingresos y el pagar o no impuesto de renta, la respuesta o regresada solo puede tomar dos valores (si paga impuesto de renta o no paga dicho impuesto); otros ejemplos en que la regresada es cualitativa son si la familia posee o no vivienda propia, se aprueba o pierde un curso, padece determinada enfermedad o no la padece. La variable cualitativa en estos tipos de modelos no tiene que restringirse simplemente a respuestas de sí o no, la variable respuesta puede tomar más de dos valores, ser tricotómica o politómica, también se establecen modelos en lo que la variable dependiente es de carácter ordinal o de carácter nominal, en donde no hay preestablecido ningún tipo de orden. En este trabajo se analizara el modelo LOGIT en donde la variable dependiente es de carácter binario o dicotómico (sí o no). (Green 2001) Se trata pues de adoptar una formulación no lineal que obligue a que los valores estimados estén entre 0 y 1 ya que, la regresión con una variable binaria dependiente Y modeliza la probabilidad de que Y = 1. La regresión LOGIT utiliza una función de distribución logística, su función de distribución de probabilidad da lugar a probabilidades ente 0 y 1, y presenta un crecimiento no lineal (con mayores incrementos en la parte central).

1 Introduccion

Antes que nada tenemos que tener en cuenta la diferencia entre modelos, en donde la "Y" (endógena) es cuantitativa o cualitativa. En el primer modelo el objetivo es estimar el valor o medida de "Y". Esta forma del primer modelo se usó en la primera parte del curso "econometría I" centrándonos en el modelo lineal clásico y sus supuestos. En el segundo modelo nuestra variable "Y" es cualitativa así que el objetivo el calcular la probabilidad de que un sujeto tome una determinada decisión de índole discreta, condicionada a ciertas variables explicativas, permitiendo identificar las características del sujeto con variables independientes cualitativas. Esta parte es un tema nuevo para nosotros los

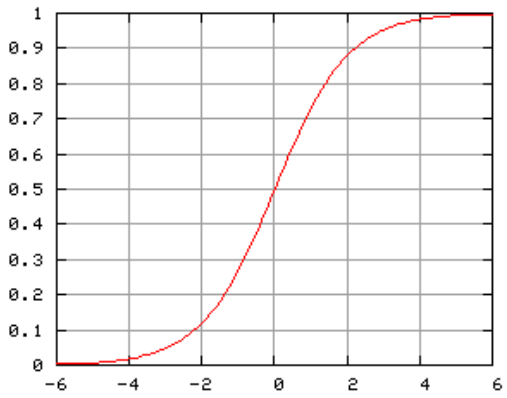

Figure 1: Curva Logistica.

alumnos del curso por lo que la evaluamos con mucho cuidado tanto en la parte conceptual como matemática.

Entre las principales desventajas de estimar un modelo de variable dependiente dicotómica a través de un modelo lineal, se tiene lo siguiente:

- Puesto que la variable dependiente solo toma valores 0 o 1, el supuesto de normalidad de las perturbaciones no se cumple ya que siguen la distribución de Bernoulli.
- Perturbaciones heteroscedásticas.
- Las predicciones de la variable dependiente pueden estar fuera del rango [0,1].
- El modelo lineal de probabilidad implica que el efecto marginal de cada una de las variables explicativas es constante.

Clasificación de los modelos de elección discreta

Número de alternativas	Tipo de alternativas	Tipo de función	El regresor se refiere a:	
			Características (de los individuos)	Atributos (de las alternativas)
Modelo de respuesta dicotómica (2 alternativas)	Complementarias	Lineal	Modelo de probabilidad lineal truncado	
		Logística	Modelo *logit*	
		Normal estándar	Modelo *probit*	
Modelos de respuesta múltiple (más de 2 alternativas)	No ordenadas	Logística	*Logit* multinomial	*Logit* condicional
			Logit anidado	*Logit* anidado
			Logit mixto	*Logit* mixto
		Normal estándar	*Probit* multinomial	*Probit* condicional
			Probit multivariante	*Probit* multivariante
	Ordenadas	Logística	*Logit* ordenado	
		Normal estándar	*Probit* ordenado	

Fuente: tomado de Medina (2003).

Figure 2: Cuadro comparativo.

2 Estimacion del Modelo Logit

2.1 Modelación de elección discreta

Para tener una idea más clara de las ramas de donde se obtiene el modelo logit y de algunos de los elementos que influyen en la especificación de los modelos discretos y sus características mostramos el cuadro de la Figura 2.

2.2 Definición: Modelo Logit

Si tenemos una variable endógena "Y", la cual solo puede tomar valores de Y=1 (presencia de la característica de interés) con una probabilidad de ocurrencia de "p" y Y=0 ausencia de la característica de interés) con portabilidad 1-p. También posee una variable exógena X, la cual puede ser categórica o continua. Si la variable Y es el resultado de un experimento de Bernoulli, es decir la observaciones son independientes, entonces las variable aleatoria tienen un distribución de Bernoulli con:

[Y/X=x]=p(Esperanza condicionada de Y dado X=x)

[Y/X=x]=p(1-p)(Esperanza condicionada de Y dado X=x)

Entonces la probabilidad de que Y=1 es E[Y/X=x]=p y puede ser calculada a partir de una distribución de probabilidad que tiene la forma de la curva sigmoidea, en particular esta curva puede ser logística.

$$E[Y/X = x] = p = \frac{\epsilon^{\alpha+\beta x_i}}{1 + \epsilon^{\alpha+\beta x_i}} \ldots (1.1)$$

Donde es el predictor lineal y la función de enlace canónico es:

$$\theta = \ln\left(\frac{E(Y)}{1 - E(Y)}\right) = \ln\left(\frac{p}{1-p}\right) \ldots (1.2)$$

Su representación como un modelo lineal generalizado:

$$logit(p) = \ln\left(\frac{E(Y/X=x)}{1-E(Y/X=x)}\right) = \ln\left(\frac{p(x)}{1-p(x)}\right) = \alpha + \beta x \ldots (1.3)$$

Con este modelo la variable endógena puede tomar solo dos valores (aprobado o desaprobado) con lo que podemos conocer la probabilidad de que un alumno este aprobado en función de su variables exógenas predictivas.
Estas variables exógenas predictivas pueden tener características cualitativas o cuantitativas, además podremos entender como todas estas variables participan en su conjunto para explicar la probabilidad de respuesta.

2.3 Caracteristicas del Modelo Logit

- A pesar de que el modelo transformado es lineal en las variables, las probabilidades no son lineales.
- El modelo logit supone que el logaritmo de la razón de probabilidades esta linealmente relacionado con las variables explicadoras.
- En el modelo logit los coeficientes de regresión expresan el cambio en el logaritmo de las probabilidades, cuando una de las variables explicadoras cambia en una unidad, permaneciendo constantes las demás (Gujarati 2010).

2.4 Estimación de parámetros modelo Logit

En esta parte la literatura teórica define que existes 2 casos para la estimación de parámetros. Los modelos Logit con observaciones repetidas y no repetidas.

2.4.1 El modelo Logit con observaciones repetidas: Mínimos cuadrados generalizados

El modelo Logit con observaciones repetidas estima los parámetros con el método de mínimos cuadrados generalizados. En este caso la variable endógena queda acotada entre valores de 0 y 1; además es continua por lo que para calcular los parámetros podemos utilizar un método común como el de mínimos cuadrados, pero ya que existe heterocedasticidad se usa el método de mínimos cuadrados generalizados.

No profundizaremos los detalles de este modo de estimación ya que nuestro objetivo en este trabajo es enfocarnos en el método de máxima verosimilitud.

2.4.2 El modelo Logit con observaciones no repetidas: Máxima verosimilitud

El modelo Logit con observaciones no repetidas estima los parámetros con el método de máxima verosimilitud.

Para este método haremos una explicación detallada de los pasos seguidos. El método de máxima verosimilitud nos dice lo siguiente, que teniendo una variable aleatoria, caracterizada por unos paraetros, y dada una muestra, se consideran estimadores de máxima verosimilitud a aquellos parámetros que generarían con mayor probabilidad la muestra observada; es decir, los valores de

máxima verosimilitud son aquellos de los cuales la función de densidad conjunta alcanza un máximo.

Con el supuesto que las observaciones son independientes, la función de densidad conjunta de la variable Yi dicotómica queda:

$$Prob(Y_1, Y_2, Y_3, \ldots, Y_n) = \prod_{1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i} \ldots (2.1)$$

De donde Pi (de la ecuación 1.1) recoge la probabilidad de que Yi=1. Para simplificar se trabajará con la función de densidad en logaritmos:

$$\lambda = lnL = \sum_{i=1}^{n} Y_i ln p_i + \sum_{j=1+i}^{n-j} (1-Y_i) ln(1-p_i) \ldots (2.2)$$

El método de máxima verosimilitud elige el estimador del parámetro que máxima la función de verosimilitud. Para conseguir esto se calculará las derivadas de primer orden con respecto a cada parámetro y las igualamos a 0. Haciendo las modificaciones del caso obtenemos el siguiente sistema de ecuaciones:

$$\frac{\partial \Lambda}{\partial \alpha} = \sum_{i=1}^{n}(Y_i - p_i) = \sum_{i=1}^{n}(Y_i - \frac{\epsilon^{\alpha+\beta x_i}}{1+\epsilon^{\alpha+\beta x_i}}) = 0 \ldots (2.3)$$

$$\frac{\partial \Lambda}{\partial \alpha} = \sum_{i=1}^{n}(Y_i - p_i)X_i = \sum_{i=1}^{n}(Y_i - \frac{\epsilon^{\alpha+\beta x_i}}{1+\epsilon^{\alpha+\beta x_i}})X_i = 0 \ldots (2.4)$$

Nos queda un sistema de ecuaciones no lineales, las cuales serán resueltas por métodos iterativos o algoritmos de optimización que permitan la convergencia de los estimadores.

2.5 Revisión de literatura empírica

Las aplicaciones del modelo logístico son diversas, y en su mayoría estos modelos utilizan la estimación por máxima verosimilitud. En la literatura leída se ha encontrado por ejemplo aplicaciones en el crecimiento poblacional y el crecimiento de bacterias como lo menciona Gujarati (2010) en su libro de econometría. O la permanencia estudiantil por Laura Rosa Llano Díaz y Viardin Mosquera Caicedo (2006), en el cual como un modelo Logit intenta hallar la variable de más influencia en la deserción estudiantil. O como la evolución viral de pacientes de VIH realizado por César Leoncio Calle Hurtado (2004). Que busca las variables de influencia según su condición social que hacen que un paciente de VIH mejore.

2.6 Conclusiones

Según lo leído podemos mencionar algunas ventajas del modelo Logit, como que la variable endógena, al ser dicotómica, no tiene que cumplir el supuesto de normalidad, además sus coeficientes tienen una fácil interpretación. Una desventaja sería la necesidad de tener muchos datos para que la estimación por el método de máxima verosimilitud sea correcta.

Bibliografia

[1] Green, W. (2001), Analisis Econometrico, Prentice Hall.

[2] Gujarati, D. (2010), Econometria, McGraw Hill.