Propiedad de insesgadez de los estimadores de la función logística mediante el Método de Mínimos Cuadrados

Property unbiasedness of the estimates of the logistic function by the method of least squares

Diana Paola Mejía Rojas, Jorge Eduardo Ossa Sánchez, Francisco José Oscar Escobar
Departamento de Matemáticas, Universidad Tecnológica de Pereira, Colombia
dpmejia@utp.edu.co
osgeorge@utp.edu.co
osesco@utp.edu.co

***Resumen*— La función logística es una función con múltiples aplicaciones en las ciencias económicas, naturales y sociales. Esta función contiene dos parámetros los cuales se analizaron para identificar si cumplían con la propiedad de insesgadez o no. La estimación de dichos parámetros se encontró utilizando el Método de Mínimos Cuadrados.**

Palabras clave—Sesgo, Modelos no lineales, Método de Mínimos Cuadrados, Estimadores.

***Abstract*— The logistic function is a function with multiple applications in the economic and natural sciences. This function contains two parameters which were analyzed to identify whether they met the property of unbiasedness or not. The estimation of these parameters was found using the least squares method .**

***Key Word* — Bias , non-linear models , method of least squares estimators .**

I. INTRODUCCIÓN

La función logística plantea que una población puede crecer exponencialmente, en un ambiente en el que no se presente escasez de recursos o amenazas como depredadores. En esta primera etapa, la población se incrementa proporcionalmente frente a su tamaño en un tiempo t. No obstante, esta situación no es muy común, por lo que habría que incluir en el modelo la naturaleza de los recursos, que al ser limitados, pueden interrumpir la tendencia creciente de la población. **[1]**

Las estimaciones realizadas a partir de la función logística han suscitado varios debates entre autores. En este sentido, Ortega M, y Cayuela A **[2]** realizaron una revisión bibliográfica acerca de cuestiones como la determinación del tamaño de muestra más adecuado para este tipo de modelo, y el incremento del sesgo, pues a medida que disminuye el número de eventos de interés, este crece.

Además, Alderete A [3] analiza los fundamentos de la regresión logística en la investigación psicológica. Señala que una de las ventajas de este modelo, es que no requiere cumplir con supuestos como el de normalidad multivariable y homocedasticidad. Una de las principales preocupaciones a la hora de aplicar la regresión logística, es el control de los sesgos, por lo que se han desarrollado procedimientos de bondad y ajuste para estos efectos.

El sesgo en las estimaciones también podría incrementarse por la aparición de intervalos de confianza anormalmente amplios. [4]

En los modelos no lineales aparecen problemas de convergencia en las parametrizaciones de los métodos iterativos afectando las propiedades de sesgo, eficiencia, convergencia y consistencia de los estimadores.

El presente artículo pretende analizar la propiedad de insesgadez de los estimadores de la función logística, a fin de aportar información que permita profundizar en el conocimiento de esta función y sus aplicaciones.

Fecha de Recepción: 30 de noviembre de 2015
Fecha de Aceptación: 08 de junio de 2017

II. ANTECEDENTES

Adolphe Quetelet en 1835 y Pierre Françoise Verhulst en 1838, fueron quienes plantearon unas ideas del comportamiento de la población dando lugar a la función logística que hoy conocemos.

$$p(t) = \frac{a}{1 + be^{-ht}} \qquad (1)$$

La función (1) se puede apreciar gráficamente en la Figura 1.

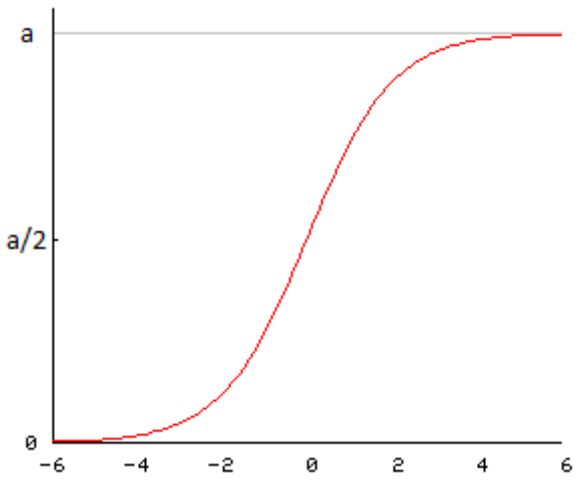

Figura 1: Gráfica de la función logística

Donde a es el límite superior de la curva o capacidad máxima de carga, b son los cambios que afectan la localización de la curva y h es la tasa de crecimiento intrínseco del modelo, un cambio en este parámetro afecta la forma de la curva.

Esta función matemática refleja especialmente la evolución de las poblaciones que experimentan primero un lento crecimiento y luego se acelera hasta llegar al punto de inflexión para a partir de él crecer más lentamente y situarse posterior en términos de crecimiento cero.

La función logística no es lineal, por lo tanto el método de regresión lineal no se puede aplicar directamente.

Linealización de los datos para realizar la estimación por el método de mínimos cuadrados

III LINEALIZACION DE LOS DATOS

A la función logística se le realiza el proceso de linealización para luego ser estimada por el método de mínimos cuadrados.

$$p(t) = \frac{a}{1 + be^{-ht}}\left(\frac{e^{ht}}{e^{ht}}\right) = \frac{ae^{ht}}{e^{ht} + b} \qquad (2)$$

$$p(t)(e^{ht} + b) = ae^{ht}$$

$$p(t)b = ae^{ht} - p(t)e^{ht}$$

$$ht = \ln\left(\frac{p(t)b}{a - p(t)}\right)$$

$$ht = ln\left(\frac{p(t)}{a-p(t)}\right) + lnb$$

$$\ln\left(\frac{p(t)}{a - p(t)}\right) = ht - \ln b \qquad (3)$$

Haciendo un cambio de variable con

$$z = \ln\left(\frac{p(t)}{a - p(t)}\right); \quad \beta = \ln b \qquad (4)$$

El modelo a estimar es: $z = ht - \beta$

En forma general es:

$$z_i = ht_i - \beta + \varepsilon_i \qquad (5)$$

En esta transformación z_i es la predicción de un cociente de probabilidades a lo largo de una recta, t_i variable de predicción, ε_i error aleatorio no observable asociado con z_i, y β, h son los parámetros desconocidos. Cada z_i es una variable aleatoria.

Del modelo se espera que la esperanza de los errores sea cero, la varianza de los errores sea constante y los errores sean mutuamente independientes.

Es decir:

$$E(Z_i) = E(ht_i - \beta + \varepsilon_i) = ht_i - \beta \qquad (6)$$

$$Cov(z_i, z_j) = 0 \qquad i \neq j \qquad (7)$$

$$Var(Z_i) = Var(ht_i - \beta + \varepsilon_i) = Var(\varepsilon_i) = \sigma^2 \qquad (8)$$

IV METODO DE MINIMOS CUADRADOS

El error conocido como residuo viene dado por:

$$\varepsilon_i = z_i + \beta - ht_i \tag{9}$$

Donde B, H serán los estimadores de β, h.

El método busca ajustar una recta a un conjunto de datos, dicho método minimiza los residuos al cuadrado, teniendo como residuo la diferencia entre los datos observados y los valores del modelo.

$$S_r = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (z_i - Ht_i + B)^2 \tag{10}$$

$$\begin{cases} \dfrac{\partial S_r}{\partial H} = -2\sum_{i=1}^{n} (z_i - Ht_i + B)t_i \\ \dfrac{\partial S_r}{\partial B} = 2\sum_{i=1}^{n} (z_i - Ht_i + B) \end{cases}$$

$$\begin{cases} -2\sum_{i=1}^{n} (z_i - Ht_i + B)t_i = 0 \\ 2\sum_{i=1}^{n} (z_i + Ht_i + B) = 0 \end{cases}$$

$$\begin{cases} \sum_{i=1}^{n} z_i t_i = \sum_{i=1}^{n} Ht_i^2 - \sum_{i=1}^{n} Bt_i \\ \sum_{i=1}^{n} z_i = \sum_{i=1}^{n} Ht_i - \sum_{i=1}^{n} B \end{cases}$$

Las ecuaciones normales encontradas son:

$$\begin{cases} \sum_{i=1}^{n} z_i t_i = H\sum_{i=1}^{n} t_i^2 - B\sum_{i=1}^{n} t_i \\ \sum_{i=1}^{n} z_i = H\sum_{i=1}^{n} t_i - nB \end{cases} \tag{11}$$

Al dividir la segunda ecuación entre n

$$\frac{1}{n}\sum_{i=1}^{n} z_i = \frac{H}{n}\sum_{i=1}^{n} t_i - \frac{nB}{n}$$

$$B = \frac{H}{n}\sum_{i=1}^{n} t_i - \frac{1}{n}\sum_{i=1}^{n} z_i = H\bar{t} - \bar{z} \tag{12}$$

Sustituyendo en la primera

$$\sum_{i=1}^{n} z_i t_i = H\sum_{i=1}^{n} (t_i)^2 - \left(\frac{H}{n}\sum_{i=1}^{n} t_i - \frac{1}{n}\sum_{i=1}^{n} z_i\right)\sum_{i=1}^{n} t_i$$

$$H = \frac{\sum_{i=1}^{n} z_i t_i - \frac{1}{n}\sum_{i=1}^{n} z_i \sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} t_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} t_i\right)^2} = \frac{\sum_{i=1}^{n} (t_i - \bar{t})(z_i - \bar{z})}{\sum_{i=1}^{n} (t_i - \bar{t})^2} \tag{13}$$

$$\sum_{i=1}^{n} z_i t_i = H\left(\sum_{i=1}^{n} (t_i)^2 - \frac{1}{n}\left(\sum_{i=1}^{n} t_i\right)^2\right) + \frac{1}{n}\sum_{i=1}^{n} z_i \sum_{i=1}^{n} t_i$$

Despejando H

$$H = \frac{\sum_{i=1}^{n} (t_i - \bar{t})(z_i - \bar{z})}{\sum_{i=1}^{n} (t_i - \bar{t})^2}$$

Entonces

$$B = H\bar{t} - \bar{z} \tag{14}$$

$$H = \frac{\sum_{i=1}^{n} (t_i - \bar{t})(z_i - \bar{z})}{\sum_{i=1}^{n} (t_i - \bar{t})^2} \tag{15}$$

Como B es el estimador de β y $\beta = \ln b$ por lo tanto un estimador de b es

$$B^* = e^B \tag{16}$$

Entonces

$$B^* = e^{H\bar{t} - \bar{z}} \tag{17}$$

Donde B^* y H son los estimadores por el método de mínimos cuadrados.

V PROPIEDAD DE INSESGADEZ DE LOS ESTIMADORES

- $E(H) = h$

$$E\left(\frac{\sum_{i=1}^{n} (t_i - \bar{t})(z_i - \bar{z})}{\sum_{i=1}^{n} (t_i - \bar{t})^2}\right) = \frac{1}{\sum_{i=1}^{n} (t_i - \bar{t})^2} E\left(\sum_{i=1}^{n} (t_i - \bar{t})(z_i - \bar{z})\right)$$

$$= \frac{1}{\sum_{i=1}^{n}(t_i-\bar{t})^2}\left(\sum_{i=1}^{n}(t_i-\bar{t})E(z_i-\bar{z})\right)$$

Ahora:

$$E(z_i-\bar{z}) = E(z_i) - E(\bar{z}) \quad (19)$$

$$= ht_i - \beta - \left(\frac{1}{n}\sum_{i=1}^{n}E(z_i)\right)$$
$$= h(t_i-\bar{t})$$

Por lo tanto:

$$E(z_i-\bar{z}) = h(t_i-\bar{t}) \quad (20)$$

Entonces

$$= \frac{1}{\sum_{i=1}^{n}(t_i-\bar{t})^2}\left(\sum_{i=1}^{n}(t_i-\bar{t})E(z_i-\bar{z})\right)$$

$$= \frac{1}{\sum_{i=1}^{n}(t_i-\bar{t})^2}\left(\sum_{i=1}^{n}(t_i-\bar{t})h(t_i-\bar{t})\right)$$

$$E(H) = h$$

El estimador H es insesgado

- $E(B^*) = E(e^B)$

$$\approx E\left(\frac{\sum_{n=0}^{\infty}(H\bar{t}-\bar{z})^n}{n!}\right)$$

Tomando los dos primeros términos de la serie

$$\approx E\left(1+\frac{H\bar{t}-\bar{z}}{1!}\right)$$
$$\approx 1 + E(H\bar{t}-\bar{z})$$
$$\approx 1 + H\bar{t} - E\left(\frac{\sum_{i=1}^{n}(ht_i-\beta+\varepsilon_i)}{n}\right)$$
$$\approx 1 + H\bar{t} - \frac{1}{n}E(\sum_{i=1}^{n}ht_i-\beta+\varepsilon_i)$$
$$\approx 1 + H\bar{t} - \frac{1}{n}\left(\sum_{i=1}^{n}ht_i-\beta\right)$$
$$\approx 1 + \beta$$

Se observa que el estimador B^*es sesgado para los dos primeros términos, esto se da porque proviene de una función no lineal.

Al estudiar la varianza de H:

$$H = \frac{\sum_{i=1}^{n}(t_i-\bar{t})(z_i-\bar{z})}{\sum_{i=1}^{n}(t_i-\bar{t})^2} = \frac{S_{xy}}{S_{xx}}$$

$$Var(H) = Var\left(\frac{S_{xy}}{S_{xx}}\right) = Var\left(\sum_{i=1}^{n}c_iz_i\right)$$

Donde $c_i = \frac{t_i-\bar{t}}{S_{xx}}$

$$Var(H) = \sum_{i=1}^{n}(c_i)^2Var(z_i)$$
$$= \sum_{i=1}^{n}(c_i)^2\sigma^2$$
$$= \frac{\sigma^2}{S_{xx}}$$

Para la varianza de B

$$Var(B) = Var(\bar{z}-H\bar{t})$$

$$Var(B) = Var(\bar{z}) + (\bar{t}^2)Var(H) - 2\bar{t}cov(\bar{z},H)$$

$$Var(\bar{z}) = Var(\bar{e}) = \frac{\sigma^2}{n}$$

$$Cov(\bar{z},H) = Cov\left(\frac{\sum_{i=1}^{n}z_i}{n}, \sum_{i=1}^{n}c_iz_i\right)$$

$$= \frac{\sigma^2}{n}\sum_{i=1}^{n}c_i = 0$$

Para el estimador B^*

$$Var(B^*) = Var(\bar{z}) + \bar{t}^2Var(H)$$

$$= \sigma^2\left(\frac{1}{n} + \frac{\bar{t}^2}{S_{xx}}\right)$$

VI CONCLUSIONES

La insesgadez del estimador H es una propiedad de poco interés para el estimador, lo que verdaderamente interesa es su proximidad al parámetro, sea este su esperanza matemática o no.

Para garantizar el resultado de insesgadez del estimador H se tendrá que valorar la dispersión cuadrática con el error cuadrático medio.

Las ecuaciones normales resultantes de aplicar el Método de Mínimos cuadrados para encontrar los estimadores de la función o de otros modelos no lineales, no siempre tienen solución cerrada y se debe utilizar los métodos numéricos apropiados.

El método natural de estimación es el de máxima verosimilitud, por las complicaciones que presenta en estos casos no lineales los mínimos cuadrados.

REFERENCIAS

[1]. Stewart J. (2010). "Ecuaciones diferenciales en Cálculo de una variable" pp. 567-568. México D.F: CENCAGE Learning.
[2]. Ortega M, Cayuela A (2002). "Regresión logística no condicionada y tamaño de muestra: una revisión bibliográfica". Rev. Esp. Salud Publica, 76, pp. 85-93. 23 de noviembre de 2015, De Scielo Base de datos.
[3]. Alderete A. (2006). "Fundamentos del Análisis de Regresión Logística en la Investigación Psicológica", pp. 52-67. 23 de noviembre de 2015, De Google académico Base de datos.
[4]. Irala J, Fernández R, Serrano A. (1997). "Intervalos de confianza anormalmente amplios en regresión logística: interpretación de resultados de programas estadísticos". Rev Panam Salud Publica/Pan Am J Public Health, 1, pp. 230- 234. 23 de noviembre de 2015 , De Scielo Base de datos.
[5]. Mendenhall, William, Wackerly, Dennis D., Scheaffer, Richard L. "Estadística matemática con aplicaciones", Grupo Editorial Iberoamericana, 1994. México.
[6]. Parra Rodriguez, Francisco. "Métodos de estimación no lineales". [Online].Available: http://econometria.files.wordpress.com/2008/05/metodos-de-estimacion-no-lineales1.pdf