REGRESIÓN LINEAL SIMPLE CON TÉRMINOS DE PERTURBACIÓN NO NORMALES

M.Sc. Rivero Suguiura, Fernando Oday

✉ *fors2004@yahoo.com.ar*

RESUMEN

Existen muchas funciones de distribución de probabilidad discreta o continua que no son consideradas en el análisis del comportamiento aleatorio de un conjunto de datos y se asume normalidad en la mayoría de los casos, aunque este conjunto de datos a veces no presente cercanía a una distribución normal. El presente artículo describe el método de regresión lineal simple considerando valores de perturbación basados en otro tipo de distribución como es el de la función secante hiperbólica generalizada (SHG), que complica su tratamiento matemático al momento en que se aplica el método de estimación por máxima verosimilitud.

PALABRAS CLAVE

Distribución secante hiperbólica generalizada (SHG), Verosimilitud modificada.

1. DISTRIBUCIÓN SECANTE HIPERBÓLICA GENERALIZADA (SHG)

La distribución secante hiperbólica debida a Vaughan en el año 2002, es una distribución de probabilidad continua, cuya función de densidad y función característica son proporcionales a la función secante hiperbólica. Es simétrica de media cero y varianza , enfocada en este caso a la variable aleatoria no observable de perturbación . Es una distribución leptocúrtica simétrica, con pico agudo y coeficiente de curtosis entre 3 y 9, muy similar a la normal estándar y t-student.

La función de densidad, se define como:

$$f(u;t)=\frac{1}{\sigma_u}\left(\frac{ae^{bu/\sigma_u}}{e^{2bu/\sigma_u}+2ce^{bu/\sigma_u}+1}\right) \qquad t:\text{parámetro y} \qquad -\infty<u<\infty$$

Si: $-\pi<t<0$

$$a=\frac{sen\,t}{t}b;\ b=\sqrt{\frac{\pi^2-t^2}{3}};\ c=\cos t$$

Si: $t=0$

$$a=b=\frac{\pi}{\sqrt{3}};\ c=1$$

Si: $t>0$

$$a=\frac{senh\,t}{t}b;\ b=\sqrt{\frac{\pi^2+t^2}{3}};\ c=\cosh t$$

2. MODELO DE REGRESIÓN LINEAL SIMPLE

Un modelo de regresión lineal simple es de la forma:

$$Y=\beta_0+\beta_1X+u$$

A menudo se asume que los valores de perturbación u_i, $i = 1, \ldots, N$ son independientes e idénticamente distribuidos con distribución normal, media cero y varianza σ^2_u. Sin embargo, no siempre puede ocurrir esto, tal el ejemplo de que las respuestas sean dicotómicas, es decir, 0 o 1, entonces tiene distribución Bernoulli, y en otras oportunidades, puede darse una distribución continua sesgada.

En este artículo se considera que los valores de perturbación del modelo siguen distribución Secante Hiperbólica Generalizada (SHG), es decir, $u \sim SHG\ (0,\sigma^2_u; t)$.

Se toma una muestra aleatoria de tamaño n de una población de tamaño N, considerándose el modelo de estimación

$$Y = \hat{Y} + e = \hat{\beta}_0 + \hat{\beta}_1 X + e$$

luego

$$e = Y - \hat{Y} = Y - \hat{\beta}_0 - \hat{\beta}_1 X$$

con

$$E(e) = 0 \text{ y } V(e) = \sigma^2_u$$

3. FUNCIÓN DE VEROSIMILITUD

Sea

$$z_i = \frac{u_i}{\sigma_u} = \frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma_u} \qquad i = 1, \ldots, n$$

la función de verosimilitud, es

$$L = \frac{a^n}{\sigma_u^n} \prod_{i=1}^{n} \left(\frac{a e^{bz_i}}{e^{2bz_i} + 2ce^{bz_i} + 1} \right) = \frac{a^n}{\sigma_u^n} \prod_{i=1}^{n} \left[\frac{a e^{b\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma_u}\right)}}{e^{2b\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma_u}\right)} + 2ce^{b\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma_u}\right)} + 1} \right]$$

Tomando logaritmo de L, derivando respecto a β_0, β_1 y σ_u e igualando a cero, se tiene:

$$\frac{\partial lnL}{\partial \beta_0} = -\frac{bn}{\sigma_u} + \frac{2b}{\sigma_u} \sum_{i=1}^{n} g(z_i) = 0 \qquad (1)$$

$$\frac{\partial lnL}{\partial \beta_1} = -\frac{b}{\sigma_u} \sum_{i=1}^{n} X_i + \frac{2b}{\sigma_u} \sum_{i=1}^{n} X_i g(z_i) = 0 \qquad (2)$$

$$\frac{\partial lnL}{\partial \sigma_u} = -\frac{n}{\sigma_u} - \frac{b}{\sigma_u} \sum_{i=1}^{n} z_i + \frac{2b}{\sigma_u} \sum_{i=1}^{n} z_i g(z_i) = 0 \qquad (3)$$

donde

$$g(z_i) = \frac{e^{2bz_i} + ce^{bz_i}}{e^{2bz_i} + 2ce^{bz_i} + 1} \qquad i = 1, \ldots, n \qquad (4)$$

Las expresiones de (1), (2) y (3) no admiten soluciones explícitas dado a los términos relacionados con la función no lineal de (4). Para establecer soluciones en las ecuaciones anteriores, se emplean expresiones de verosimilitud modificadas mediante el uso de los dos primeros términos de la serie de Taylor alrededor de $t_{(i)}$ dado por Tiku y Suresh.

$$g[z_{(i)}] \cong g[t_{(i)}] + g'[t_{(i)}][z_{(i)} - t_{(i)}] = \theta_i + \alpha_i z_{(i)} \qquad 1 \le i \le n \qquad (5)$$

Donde $\theta_i = g\ [t_{(i)}] - \alpha_i t_{(i)}$ y $\alpha_i = g'\ [t_{(i)}]$ Incluyendo (5) en (1), (2) y (3), se obtienen las ecuaciones de verosimilitud modificada L^*, es decir:

$$\frac{\partial lnL}{\partial \beta_0} \cong \frac{\partial lnL^*}{\partial \beta_0} = -\frac{bn}{\sigma_u} + \frac{2b}{\sigma_u}\sum_{i=1}^{n}[\theta_i + \alpha_i z_{(i)}] = 0 \qquad (6)$$

$$\frac{\partial lnL}{\partial \beta_1} \cong \frac{\partial lnL^*}{\partial \beta_1} = -\frac{b}{\sigma_u}\sum_{i=1}^{n} X_{[i]} + \frac{2b}{\sigma_u}\sum_{i=1}^{n} X_{[i]}[\theta_i + \alpha_i z_{(i)}] = 0 \qquad (7)$$

$$\frac{\partial lnL}{\partial \sigma_u} \cong \frac{\partial lnL^*}{\partial \sigma_u} = -\frac{n}{\sigma_u} - \frac{b}{\sigma_u}\sum_{i=1}^{n} z_{(i)} + \frac{2b}{\sigma_u}\sum_{i=1}^{n} z_{(i)}[\theta_i + \alpha_i z_{(i)}] = 0 \qquad (8)$$

4. SOLUCIONES DE ESTIMACIÓN

De (6), (7) y (8), se logran las soluciones de estimación de los parámetros

$$\hat{\beta}_0 = \bar{Y}_{[.]} - \hat{\beta}_1 \bar{X}_{[.]} \qquad \hat{\beta}_1 = A - \hat{\sigma}_u B$$

con

$$\hat{\sigma}_u = \left(-C + \sqrt{C^2 + \frac{4nD}{b}}\right) \Big/ \frac{2n}{b}$$

y

$$\bar{X}_{[.]} = \frac{\sum_{i=1}^{n} \alpha_i X_{[i]}}{\sum_{i=1}^{n} \alpha_i} \qquad \bar{Y}_{[.]} = \frac{\sum_{i=1}^{n} \alpha_i Y_{[i]}}{\sum_{i=1}^{n} \alpha_i}$$

$$A = \frac{\sum_{i=1}^{n} \alpha_i (X_{[i]} - \bar{X}_{[.]}) Y_{[i]}}{\sum_{i=1}^{n} \alpha_i (X_{[i]} - \bar{X}_{[.]})^2} \qquad B = \frac{1/2 \sum_{i=1}^{n} X_{[i]} - \sum_{i=1}^{n} \theta_i X_{[i]}}{\sum_{i=1}^{n} \alpha_i (X_{[i]} - \bar{X}_{[.]})^2}$$

$$C = \sum_{i=1}^{n} Y_{[i]} - A\sum_{i=1}^{n} X_{[i]} - 2\sum_{i=1}^{n} \theta_i Y_{[i]} + 2A\sum_{i=1}^{n} \theta_i X_{[i]}$$

$$D = 2\left[\sum_{i=1}^{n} \alpha_i (Y_{[i]} - \bar{Y}_{[.]})^2 - A\sum_{i=1}^{n} \alpha_i (X_{[i]} - \bar{X}_{[.]}) Y_{[i]}\right]$$

BIBLIOGRAFÍA

Bent Jorgensen (1993). The Theory of Linear Models. New York, London.

G. A. F. Seber (1977). Linear Regression Analysis. New York.

S. R. Searle (1971). Linear Models. New York.

David W. Hosmer. Applied Logistic Regression. New York.