**PESQUIMAT**
Revista del Instituto de
Investigación de la Facultad
de Ciencias Matemáticas de la UNMSM
Vol I, N°1,pp 91-102 Lima - Perú Julio -1998

# ESTIMADORES NORMA $L_P$ EN REGRESION LINEAL

**Mg. Ysela Agüero P.**

**LABORATORIO DE SERIES DE TIEMPO-FCM**

## 1. INTRODUCCION

El uso de modelos estocásticos tiene una larga historia. Por ejemplo, sabemos que los científicos durante el siglo XVIII construían modelos estocásticos para describir el movimiento de los cuerpos celestes. Ordinariamente, los modelos estocásticos contienen parámetros desconocidos que deben ser estimados, por lo tanto, el problema de estimación de parámetros también tiene una larga historia. La primeras técnicas de estimación se basaban en diferentes tipos de promedios o funciones de observaciones empíricas.

Boscovich en 1757, fue el primero que propuso una metodología completamente nueva para estimar los parámetros desconocidos en modelos estocásticos. Él propuso estimar los parámetros minimizando una función de los errores de medición. La función que Boscovich propuso era la suma de los valores absolutos de los errores de medición. Este método aún se emplea y es conocido como estimador de mínimo error absoluto (MDA) o estimador norma $L_1$ . La escasa popularidad de este estimador se debió posiblemente a que su cálculo era complicado y sólo podía ser aplicado a modelos relativamente sencillos (con pocos parámetros).

Legendre en 1805, y posteriormente, Gauss en 1809, propusieron otra función a ser minimizada, esta era la suma de los cuadrados de los errores de medición. Este estimador es conocido como estimador de mínimos cuadrados o estimador norma $L_2$ . A partir de la propuesta de Legendre y Gauss, el método de mínimos cuadrados alcanzó una gran popularidad, tal vez por su facilidad de cálculo y por el hecho que cuando los residuos son independientes e idénticamente distribuidos, con distribución normal el estimador de mínimos cuadrados de un modelo de regresión lineal es MELI (mejor estimador lineal

insesgado). Además, es equivalente al estimador de máxima verosimilitud, por lo que se facilita la inferencia estadística.

Entonces, surge de manera natural la pregunta ¿por qué buscar otros estimadores si este es el mejor?. La respuesta es que el estimador de mínimos cuadrados es muy sensible a pequeñas desviaciones de la suposición de que los residuos se distribuyen normalmente. Existe una gran variedad de aplicaciones en las cuales hay evidencias tanto teóricas como empíricas de que los residuos presentan características que difieren de lo que podría esperarse de una distribución normal. Por lo tanto es importante estudiar y desarrollar estimadores alternativos.

El propósito de este artículo es estudiar la familia de estimadores Norma $L_p$ aplicados a los modelos de regresión lineal.

## 2. MODELO DE REGRESION LINEAL

Una clase importante de modelos que es muy aplicado en las diferentes áreas de investigación es la familia de modelos de regresión lineal la cual puede ser escrita como

$$y = X\beta + \varepsilon \qquad (2.1)$$

donde $y$ es un vector n dimensional con observaciones de la variable respuesta (v. endógena ), $X$ es una matriz (de observaciones) de rango completo y orden nxk (v. Exógenas), $\beta$ es un vector k dimensional de parámetros desconocidos y $\varepsilon$ es un vector n dimensional de residuos no observables. Las columnas de $X$ denominadas variables predictor, son denotadas por $x_1, x_2, \ldots, x_k$, donde $x_j = (x_{1j}, x_{2j}, \ldots, x_{nj})$; j=1,2,..., k. Los residuos son causados por variaciones estocásticas en la población de referencia, errores de medición, efecto de idealizaciones en la forma funcional del modelo, predictores erroneamente excluidos, etc.

Nuestro problema es estimar el vector de parámetros, $\beta$ del modelo (2.1). Boscovich introdujo la idea de, en algún sentido, minimizar el vector de residuos,

$$\varepsilon = y - X\beta .$$

Esta misma idea es utilizada en el método de mínimos cuadrados propuesto por Legendre y Gauss. Pero, minimizar los residuos no es otra cosa que; tratar de encontrar el hiperplano $X\beta$ que pase *lo más cerca posible* del vector $y$. Es decir, se trata de minimizar la *distancia* del vector de observaciones $y \in \Re^n$ a un hiperplano generado por la combinación lineal de los vectores $x_1, x_2, \ldots, x_k$ ; $x_j \in \Re^n$ . Esto nos lleva a pensar que el problema de elección de un estimador; no es otra cosa que la elección de una métrica particular, la cual podría ser por ejemplo; la norma del vector de residuos. Así, dada una norma adecuada ( $\| \cdot \|$ ), las estimaciones serán elegidas de modo que se minimice $\| \varepsilon \|$ .

Como la clase de las normas vectoriales es ilimitada, podemos elegir una interesante subclase denominada norma $L_p$ . Así el estimador de regresión será aquel que,

$$\underset{\beta}{\text{Min}} \ \| \varepsilon \|_p \qquad \forall \ \beta \in \Re^k ; \quad 1 \le p < \infty \tag{2.2}$$

donde,

$$\| \varepsilon \|_p = \begin{cases} \left\{ \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{k} x_{ij}\beta_j \right| \right\}^{1/p} & 1 \le p < \infty \\ \underset{1 \le i \le n}{Max} \left| y_i - \sum_{j=1}^{k} x_{ij}\beta_j \right| & p = \infty \end{cases}$$

Luego, un estimador que minimice la norma $L_p$ del vector de residuos observados, será denominado un estimador de regresión norma $L_p$ . Así, el estimador propuesto por Boscovich y el de mínimos cuadrados se constituyen en casos particulares de la familia de estimadores norma $L_p$ .

(1) El estimador norma $L_1$ o estimador mínimo valor absoluto, $\hat{\beta}$, es obtenido resolviendo el problema de

$$\underset{\beta \in \Re^k}{Min} \left\{ \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{k} x_{ij}\beta_j \right| \right\}$$

(2) El estimador norma $L_2$ o de mínimos cuadrados, $\hat{\beta}$, es obtenido resolviendo el problema de

$$\underset{\beta \in \mathbb{R}^k}{Min}\left\{\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right)^2\right\}$$

(3) El estimador norma $L_\infty$ , también conocido como estimador minimax o estimador de Tchebyshev. , $\hat{\beta}$, se obtiene resolviendo el problema de

$$\underset{\beta \in \mathbb{R}^k}{Min}\left\{\underset{1 \le i \le n}{Max}\sum_{i=1}^{n}\left|y_i - \sum_{j=1}^{k} x_{ij}\beta_j\right|\right\}$$

**Observación.-** En un artículo publicado en 1857 Tchebyshev propuso que los parámetros de un modelo podían ser estimados minimizando el más grande valor absoluto de las diferencias entre la función dada y la función estimada.

Un caso simple del modelo (2.1) es el modelo de posición en el cual hay solamente una variable predictor el cual sólo toma valor 1. Entonces el modelo se reduce a

$$y = \beta + \varepsilon \qquad (2.3)$$

donde $\beta$ es el parámetro de posición en una distribución de probabilidades de y . La estimación norma $L_1$ de $\beta$ es igual a la mediana muestral, la estimación norma $L_2$ de $\beta$ es igual a la media muestral y la estimación norma $L_\infty$ es igual al rango medio muestral.

## 3. EXISTENCIA Y UNICIDAD DE LOS ESTIMADORES NORMA $L_P$

Para cada vector $\hat{\beta}$ que minimiza la función objetivo dada por (2.2) existe un vector de estimaciones $\hat{y} = X\hat{\beta}$. Todos los vectores así definidos forman un conjunto que denotaremos por $P_G$ ( y ). Dado que, la matriz X es de rango completo; el vector $\beta$ está únicamente determinado para cada vector $y \in P_G$ ( y ). Los vectores $x_1$ , $x_2$ , ... , $x_k$ , y $\hat{y}$ son elementos del espacio vectorial norma $L_p$.

Sea G , el conjunto de todos los vectores que son combinaciones lineales de los vectores $x_1$ , $x_2$ , ... , $x_k$ , esto es,

$$G = \{ g \in L_p : g = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k ; \beta_j \in \Re , j=1,2, \ldots , k \}$$

es un subespacio lineal del espacio lineal $L_p$ generado por los vectores $x_1, x_2, \ldots, x_k$. Luego, los vectores $y$ serán aquellos que satisfacen,

$$\underset{g \in G}{Min} \; \| y - g \|_p = \| y - \hat{y} \|_p$$

definiendo formalmente $P_G ( y )$ tenemos

$$P_G ( y ) = \left\{ \hat{y} \in G : \| y - \hat{y} \|_p = \underset{g \in G}{Inf} \| y - g \|_p \right\}$$

donde el subespacio lineal G es de dimensión finita.

Ahora necesitamos garantizar que el conjunto $P_G ( y )$ es no vacío, esto es, que exista como mínimo una solución para el problema de optimización. Cheney (1966) demostró que dado un espacio lineal normado, existe por lo menos un punto con distancia mínima a partir de un punto fijado. Basándonos en este resultado; podemos decir que en el espacio normado $L_p$ existe como mínimo un vector $\hat{y} \in P_G ( y )$, esto es, existe como mínimo una estimación norma $L_p$ para el conjunto de datos. El teorema siguiente presenta los valores de p para los cuales $P_G ( y )$ tiene un único elemento.

**Teorema 2.1** Sea G un subespacio lineal del espacio norma $L_p$ ; $1 < p < \infty$. Entonces, existe exactamente un elemento $\hat{y}$ en $P_G ( y )$.

**Prueba.-** La demostración se basa en el hecho que los espacios lineales norma $L_p$ con $1 < p < \infty$ son estrictamente convexos.

**Observación.-** El teorema 2.1 no es aplicable para $p = 1$ y $p = \infty$, pues los espacios lineales $L_1$ y $L_\infty$ no son estrictamente convexos, luego en estos casos no se garantiza solución única..

## 3. REPRESENTACION GEOMETRICA DE LOS ESTIMADORES NORMA $L_p$

Para facilitar las interpretaciones usaremos el modelo lineal con dos regresores y tres observaciones.

$$G = \{ g \in L_p : g = \beta_1 x_1 + \beta_2 x_2 ; \beta_j \in \Re , j=1,2 \} .$$

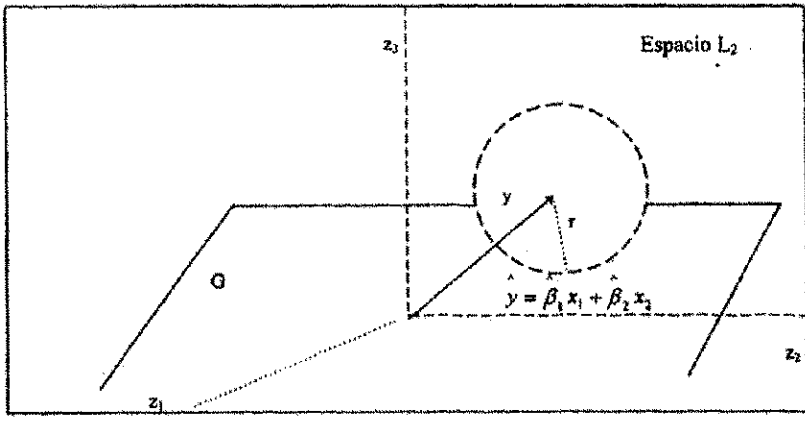
Representaremos las bolas asociadas a los estimadores de mínimos cuadrados, mínimo valor absoluto y Tchebyshev, respectivamente.

**a) Estimador norma $L_2$ .-** consideremos todos los vectores que están a una cierta distancia r del vector $y \in \Re^3$, es decir, en el espacio $L_2$ consideremos el conjunto

$$\left\{ s \in L_2 \ : \left( \sum_{i=1}^{3} (y_i - s_i)^2 \right)^{1/2} = r \right\},$$

el cual forma una esfera con centro en $y \in \Re^3$ y radio r. Para un radio suficientemente grande, la esfera toca al plano $g \in G$ . Cuando $n > 3$, el subespacio G estará formado por hiperplanos y el conjunto de vectores con distancia r del vector y será una hiperesfera. Para un rayo suficientemente grande, la hiperesfera toca al hiperplano $g \in G$ más próximo en un sólo punto. Este punto tangente corresponde a $\hat{y} = X \hat{\beta}$. Luego siempre existe una única estimación norma $L_2$ .

**Figura 3.1** El espacio $L_2$ y el conjunto de todos los vectores con distancia r partiendo del vector y.

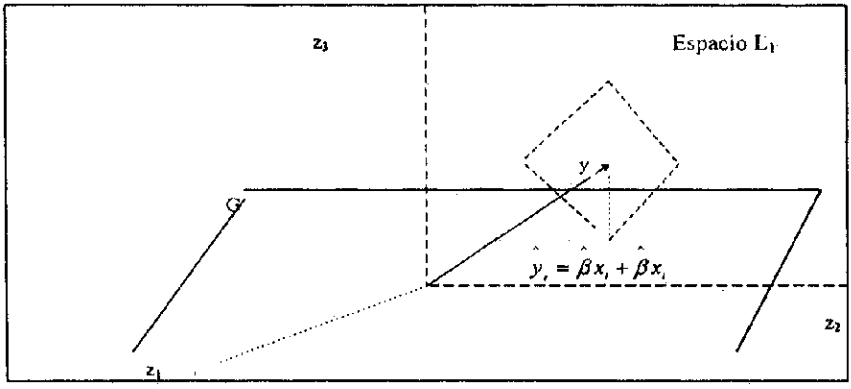



**b) Estimador norma $L_1$ .-** El conjunto de puntos que están a una cierta distancia r a partir del vector y según la métrica $L_1$ es dado por

$$\left\{ s \in L_1 \ : \left( \sum_{j=1}^{1} |y_i - s_i| \right) = r \right\}.$$

Este conjunto tiene la forma de un romboide con centro en $y \in \Re^3$ y con diagonales de longitud 2r paralelas a los ejes de coordenadas (figura 3.2). Para r suficientemente grande, el romboide toca al plano $g \in G$. En consecuencia, está garantizada la existencia de, como mínimo, una estimación norma $L_1$. El plano pueden interceptarse en un único punto, en una arista o una fase del romboide, luego las estimaciones no necesariamente son únicas.

**Figura 3.2** El espacio $L_1$ y el conjunto de todos los vectores con centro distancia r partiendo del vector y.




c) **Estimador norma $L_\infty$**.- La figura 3.3 muestra las características de la bola asociada al estimador norma $L_\infty$ o estimador de Tchebyshev. El conjunto de los vectores que están a una distancia r del vector $y \in \Re^3$, según la métrica $L_\infty$, será dado por,

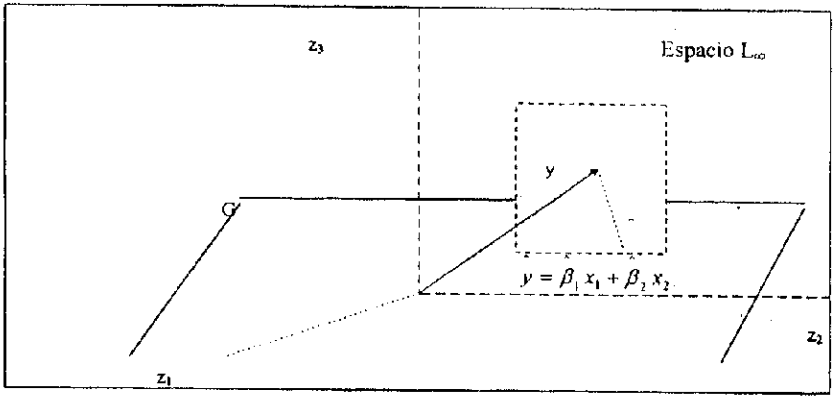
$$\left\{ s \in L_\infty : \underset{1 \le i \le n}{Max} |y_i - s_i| = r \right\}$$

los mismos forman un cubo con centro en $y \in \Re^3$ y lados de longitud 2r, paralelos a los ejes de coordenadas. Como en el caso de los estimadores norma $L_1$ la existencia de, como mínimo una solución, está garantizada, pero, no se garantiza la unicidad.

## 4. ALGORITMOS PARA EL CALCULO DE LOS ESTIMADORES NORMA $L_P$

El método de mínimos cuadrados (estimador norma $L_2$ ) es ampliamente conocido y muy utilizado por la mayoría de los estadísticos. Sus propiedades han sido bastante estudiadas.

**Figura 3.3** El espacio $L_\infty$ y el conjunto de todos los vectores con centro distancia r partiendo del vector y.




Las suposiciones básicas que se establecen para utilizar este método son, $E(\varepsilon \cdot \varepsilon) = \sigma^2 I$ , donde I es una matriz identidad de orden n x n y $\sigma^2$ es la varianza de la distribución de los residuos, además $E(x_j \varepsilon)=0$, para $j=1,2,\ldots, k$.

Una de las ventajas del método de método de mínimos cuadrados es que es computacionalmente simple. El cálculo es equivalente a la solución de un sistema de

ecuaciones lineales; que es conocido como *ecuaciones normales*, cuya solución es dada por

$\beta = [X' X]^{-1} X'y$.

El estimador es una función lineal del vector y , lo cual es muy importante cuando se deduce otras características del estimador, por ejemplo, el valor esperado y la varianza, los cuales son dados por: $E(\hat{\beta}) = \beta$ y $V(\hat{\beta}) = \sigma^2 (X'X)^{-1}$. La linealidad del estimador también implica que sea más fácil deducir su distribución de probabilidad, cuando la distribución de los residuos es conocida, en consecuencia, se simplifica la inferencia estadística.

Las estimaciones norma $L_1$ y $L_\infty$ son obtenidas usando métodos recursivos, pero su cálculo es mucho más fácil actualmente por que las computadoras son cada vez más veloces.

El cálculo de la estimación norma $L_1$ de un modelo de regresión lineal con k parámetros es equivalente a obtener la solución del problema de optimización

$$\min_{\beta} |\varepsilon| \qquad \forall \ \beta \in \Re^k \tag{2.2}$$

Sujeto a las restricciones :

$$y_i = \sum_{j=1}^{k} \hat{\beta}_j x_{ij} + \hat{\varepsilon}_i \quad ; i=1,2,...,n \ ;$$

donde $\hat{\varepsilon}_i$ ; i=1,2, ... , n son los residuos estimados.

El cálculo de la estimación norma $L_1$ es equivalente a resolver un problema de programación lineal. Esta formulación fue primero desarrollada por Barrodale y Young (1966). Posteriormente, aparecieron varias sugerencias para mejorar este algoritmo.

Cuando se calcula la estimación $L_\infty$, se formula el problema de programación lineal,

$$Min\left\{ \underset{1 \le i \le n}{Max} \left| y_i - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right| \right\} \quad ; \quad \beta \in \Re^k .$$

Sujeto a las restricciones

$$\sum_{j=1}^{k} \left( \hat{\beta}_i - \hat{\beta}_j \right) x_{ij} - y_i \le d \quad ; \quad i=1,2,...,n$$

$$-\sum_{j=1}^{k} \left( \hat{\beta}_i - \hat{\beta}_j \right) x_{ij} + y_i \le d \quad ; i=1,2,...,n$$

Del mismo modo que con el estimador norma $L_1$ , este problema de optimización se resolverá usando métodos iterativos. Uno de los primeros algoritmos fue desarrollado por Barrodale y Roberts (1972).

## 5. EQUIVALENCIA DE LOS ESTIMADORES NORMA $L_P$ Y DE MAXIMA VEROSIMILITUD.

Turner (1960) presentó una función densidad de probabilidad dada por

$$f(y_i) = \frac{1}{2\delta\Gamma\left(\frac{1}{\gamma}\right)} \exp\left\{ -\frac{|y_i - \beta|^{\gamma}}{\delta^{\gamma}} \right\} \quad -\infty < y_i < \infty \; ; \quad -\infty < \beta < \infty \; ; \delta > 0 \; ; \gamma > 0 . \qquad (5.1)$$

el símbolo $\Gamma$ se refiere a la función Gamma. Esta función de densidad es simétrica al rededor de $\beta$.

La función (5.1) constituye una familia de distribuciones de probabilidad, cuyos miembros son, por ejemplo, la distribución normal ($\gamma$=2), la distribución de Laplace ($\gamma$=1), Uniforme (límite cuando $\gamma$ tiende a $\infty$), etc.

Asumamos ahora que $\gamma > 0$ conocido y se extrae n observaciones en la variable aleatoria y . El logaritmo de la función de verosimilitud es dado por,

$$l(\beta,\delta,\gamma) = \ln\gamma^* - \ln 2\delta\Gamma\left(\frac{1}{\gamma}\right)n - \frac{1}{\delta^\gamma}\sum_{i=1}^{n}|y_i - \beta|^\gamma \qquad (5.2)$$

Claramente, se observa que el último término de la función de verosimilitud (5.1) tiene signo negativo y es el único que depende de β. Luego el estimador máximo verosímil de β es aquel $\beta$ que minimiza $\sum_{i=1}^{n}|y_i - \beta|^\gamma$

En el caso de la distribución uniforme (γ=∞), el estimador máximo verosímil de β es aquel $\beta$ que minimiza $\lim_{\gamma\to\infty}|y_i - \beta|^\gamma$, es decir, aquel $\beta$ que minimiza $\max_{1\le i\le n}|y_i - \beta|$

En conclusión el estimador máximo verosímil de β es igual al estimador norma $L_\rho$ cuando p=γ. Además, el estimador máximo verosímil de δ cuando γ=1 es igual a la media de las desviaciones absolutas $\hat{\delta} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{\beta}|$; para γ=2 es la raíz cuadrada de dos veces la desviación estándar muestral $\hat{\delta} = \left\{2\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}\right)^2\right\}^{1/2}$, y para γ=∞, el estimador máximo verosímil de δ es igual al rango medio muestral, $\hat{\delta} = \left(y_{Max} - y_{Min}\right)/2$.

Es fácil generalizar las observaciones de Turner a los modelos de regresión lineal Así, el estimador máximo verosímil y el estimador norma $L_\rho$ son equivalentes para p=γ, cuando los residuos $\varepsilon_i$ son independientes e idénticamente distribuidos con función de densidad dada en (5.2).

## 6. CONCLUSIONES

Algunas conclusiones que podemos establecer son:

a) Los estimadores de regresión norma $L_\rho$ constituyen una clase de estimadores, uno de cuyos miembros es el popular estimador de mínimos cuadrados el cual es óptimo cuando los residuos se distribuyen normalmente.

b) Cuando los residuos tienen distribuciones con "colas más pesadas" que la normal o provienen de distribuciones contaminadas, podemos encontrar estimadores norma $L_p$ que tienen una mejor performance y son más recomendables.

c) Cuando $1 < p < \infty$ existe una estimación única para el conjunto de datos. Para $p=1$ y $p = \infty$ sólo se garantiza la existencia pero no la unicidad de la estimación.

d) Las dificultades para el cálculo de algunos miembros de la familia de estimadores norma $L_p$ limitó su uso, pero, en los últimos años las facilidades computacionales están permitiendo una mayor popularización de estos métodos.

e) Los estimadores norma $L_p$ son equivalentes a los estimadores máximo verosímiles para una cierta familia de distribuciones de probabilidad.

## 7. BIBLIGRAFIA


AGUERO, P.Y. (1994) "Estimadores de regressao com alto ponto de Ruptura e detecçao de Múltiplas observaçôes Discrepantes". Biblioteca de la Facultad de Ciencias Matemáticas. UNMSM.

BARRODALE , I. & ROBERTS, F.D.K. (1972). "Solution of an overdetermined system of equations in the $L_1$ norm", Matematical Dept. Report N° 69. University of Victoria.

BARRODALE, I. & Young, A. (1966). "Algorithms for best $L_1$ and $L_\infty$ linear approximations on a discrete set" Numer. Math. 8, pp 295-306.

CHENEY, E.W. (1966) "Introduction of Aproximation Theory". McGraw-Hill, New York.

TURNER, M. (1960). " On Heuristic Estimation Methods". Biometrics, 16, pp 299-301.