Ciencia e Ingeniería Neogranadina, Vol. 20-1, pp. 133-157. Bogotá, Junio de 2010. ISSN 0124-8170

MODELOS DE OPTIMIZACIÓN POR METAS PARA EL CÁLCULO DE ESTIMADORES EN REGRESIÓN MÚLTIPLE

GOAL OPTIMIZATION MODELS FOR THE ESTIMATORS CALCULUS IN MULTIPLE REGRESSION PROBLEMS

Héctor Andrés López Ospina
Estudiante de Doctorado en Sistemas Complejos de Ingeniería. Universidad de Chile Msc. Matemática Aplicada. Matemático. Universidad Nacional de Colombia, hlopezospina@ing.uchile.cl

Rafael David López Ospina
Estadístico. Universidad Nacional de Colombia. Estadístico Banco Agrario, rdlopezo@unal.edu.co

Fecha de recepción: 3 de enero de 2010
Fecha de aceptación: 24 de mayo de 2010

RESUMEN

Este trabajo introductorio presenta y describe diversos modelos de regresión múltiple y su respectiva formulación como un problema de optimización por metas. Se describen los modelos de regresión mediana, regresión mediana ponderada, regresión cuantílica, regresión cuantílica ponderada y formulación minimax. Además, se describe la formulación dual de estos modelos y se presentan algunos ejemplos sencillos se presentan para explicar los conceptos desarrollados y las aplicaciones de dichos modelos en ingeniería y ciencias.

Palabras clave: modelos de regresión múltiple, programación por metas, regresión cuantílica, optimización minimax, regresión restringida.

ABSTRACT

This introductory work shows several multiple regression models and their relevant development as a problem of goal programming (eliminar…optimization by goals). It describes the median regression, weighted median regression, quantile regression, weighted quantile regression, and minimax formulation models. Furthermore, describes

their dual formulation. We describe some simple examples to explain the concepts developed and applications of such models on engineering and sciences.

Key words: multiple regression models, goal programming, quantile regression, minimax optimization, constrained regression.

INTRODUCCIÓN

Uno de los grandes problemas de los modelos de regresión lineal múltiple es el cumplimiento de los supuestos básicos como homocedasticidad, valor esperado y normalidad de los errores, etc. El presente documento tiene por objetivo relacionar algunos modelos poco usados en la aplicación econométrica y útiles para el análisis de la información. Además, se integra la modelación estadística con los modelos de optimización con múltiples objetivos.

Los modelos de regresión múltiple son algunas de las técnicas estadísticas más usadas para analizar datos. El objetivo de los modelos de regresión múltiple es encontrar la relación entre variables. De manera más formal, dados n vectores en $\Re^m$ $\left(x^1, x^2, ..., x^n \in \Re^m\right)$, que representan las variables explicativas y n valores reales $\left(y_1, y_2, ..., y_n \in \Re\right)$, que representan la variable explicada. En estos modelos, se tiene por objetivo encontrar un vector de estimadores $\beta = (\beta_1, \beta_2, ..., \beta_m) \in \Re^m$ que minimicen un problema determinado de optimización que depende de los valores de $x^i, y_i, i = 1, ..., n$. La regresión por mínimos cuadrados busca resolver este tipo de problemas de optimización [18]:

$$\min f(\beta) = \sum_{i=1}^{n} (y_i - \beta^T x^i)^2$$

Y Además, supone lo siguiente:

$$y_i - \beta^T x^i = u_i, \; i = 1, ..., m$$

$$E(u_i | x^i) = 0$$

Siendo u_i una sucesión de variables aleatorias e idénticamente distribuidas (generalmente, se asume la distribución normal).

En muchos problemas prácticos, los supuestos descritos anteriormente no se satisfacen con facilidad y es necesario, hacer algún tipo de transformaciones sobre las variables o utilizar otro tipo de técnicas matemáticas. Para dar solución a este tipo de problemas, surgen los conceptos de regresión mediana [28], y regresión cuantílica [2], [8], ya que se ha demostrado que estos tipos de modelos de regresión son más eficientes que el estimador máximo verosímil de muchos modelos paramétricos convencionales.

Existen diversas aplicaciones de dichas técnicas en áreas tales como: ecología [1] y [2], economía [3], [5], [6], retribución salarial [19], [29] y [12], predicción de demanda [16], calidad de la educación [4] y [38], desnutrición infantil [39], entre otros [7], [10], [11], [15], [17], [20], [21], [25] y [28]. En la sección siguiente, se presenta una breve introducción a la programación por metas. En la sección 2, se hace la formulación de diversos modelos de regresión, por medio de esta técnica de programación lineal multiobjetivo.

1. PROGRAMACIÓN POR METAS

La programación u optimización por metas (goal programming), [24], [27], [36] y [37] tiene por objetivo alcanzar unas metas o niveles de logro para determinados objetivos y fue desarrollada por Charnes y Cooper en 1955. De acuerdo con Güneş [33], la programación por metas (GP), es una técnica o herramienta muy útil para los tomadores de decisiones de tal forma que sea factible discutir y encontrar un conjunto de soluciones apropiadas y aceptables en problemas de decisión con múltiples objetivos o criterios. Por otra parte, determinar con precisión el valor real de cada objetivo es muy difícil por que se obtiene sólo información parcial.

Matemáticamente, los problemas de programación por metas se definen de la siguiente manera: Dadas n funciones objetivo $Z(x) = (Z_1(x), Z_2(x),...,Z_n(x))$, donde $x \in \Re^m$ define el vector de variables de decisión y un vector de metas $z_{meta} = (z_{1,meta}, z_{2,meta}, z_{3,meta},..., z_{n,meta})$, donde $z_{i,meta}$ define la meta del objetivo i, el método ofrece como solución el punto factible más cercano a dicho punto. Así pues, el modelo matemático que representa dicha situación es:

$$\min \sum_{i=1}^{n} \left| z_{i,meta} - Z_i(x) \right| \qquad (1)$$

$$sujeto\ a \quad x \in X \subseteq \Re^m$$

Donde X simboliza la región factible del problema de optimización.

La función objetivo del problema (1) es no lineal y no diferenciable. Sin embargo, es posible transformarla en una función lineal, introduciendo dos variables auxiliares positivas para cada objetivo, que son:

- d_i^+ corresponde a la desviación positiva o variable de exceso en cuanto al cumplimiento de la meta $i, i = 1,...,n$.
- d_i^- correspondiente a la desviación negativa o faltante sobre el nivel requerido para la meta $i, i = 1,...,n$.

Además, se cumple que $d_i^+ d_i^- = 0$ con $i, i = 1,...,n$. Es decir, no es posible que una variable de desviación positiva y una variable de desviación negativa pertenezcan de

forma simultánea a la solución básica, ya que en ningún caso se puede exceder y ser inferior a la meta. Por otra parte:

$$d_i^+ + d_i^- = \left| z_{i,meta} - Z_i(x) \right|, \; i = 1,....,n. \quad (2)$$

$$d_i^+ - d_i^- = Z_i(x) - z_{i,meta}, \quad i = 1,...,n \quad (3)$$

$$d_i^+, d_i^- \geq 0, \; i = 1,...,n \quad (4)$$

Utilizando (2), (3) y (4), el problema de optimización (1) se puede escribir de la siguiente manera:

$$\begin{aligned} &\min \sum_{i=1}^{n} d_i^+ + d_i^- \; sujeto \; a \\ &Z_i(x) + d_i^- - d_i^+ = Z_{i,meta}, \; i = 1,....,n \\ &x \in X, d_i^+, d_i^- \geq 0, \; i = 1,...,n \end{aligned} \quad (5)$$

En algunos casos, los problemas de programación por metas pueden tener diferentes niveles de ponderación para cada objetivo. En este caso, la función objetivo se escribe de la siguiente forma:

$$\min \sum_{i=1}^{n} w_i (d_i^+ + d_i^-)$$

Donde w_i es la ponderación de la meta i. Existen algunas variantes de la formulación descrita anteriormente desarrolladas en Smith R., et al [24].

Otro tipo de formulaciones que se usan en programación por metas, es la basada en la minimización de la distancia o norma *Lp* del valor obtenido a la meta. La formulación general del problema de optimización es [41], [42], y [43]:

$$\begin{aligned} &\min \left[\sum_{i=1}^{n} \left(d_i^+ + d_i^- \right)^p \right]^{1/p} \; sujeto \; a \\ &Z_i(x) + d_i^- - d_i^+ = Z_{i,meta}, \; i = 1,....,n \\ &x \in X, d_i^+, d_i^- \geq 0, \; i = 1,...,n \end{aligned} \quad (6)$$

Donde $p \in [1,+\infty]$. La formulación anterior implica tener una función objetivo no diferenciable; para evitarla se resuelve el siguiente problema equivalente de optimización:

$$\begin{aligned} &\min \sum_{i=1}^{n} \left(d_i^+ + d_i^-\right)^p \; sujeto \quad a \\ &Z_i(x) + d_i^- - d_i^+ = Z_{i,meta}, \; i = 1,....,n \\ &x \in X, d_i^+, d_i^- \geq 0, i = 1,...,n \end{aligned} \tag{7}$$

Cuando *p=1*, se tiene el problema lineal de programación por metas descrito en (5). Cuando se tiene la norma del máximo ($p = \infty$), la formulación de la función objetivo se presenta de la siguiente forma *minimax* [44]:

$$\min \max\{(d_i^+ + d_i^-), \; i = 1,...,n\}$$

El modelo (7) para la norma del máximo, tiene la siguiente formulación basada en la norma de Chevyshev [44], [45]:

$$\begin{aligned} &\min D, sujeto \; a \\ &Z_i(x) + d_i^- - d_i^+ = Z_{i,meta}, \; i = 1,....,n \\ &d_i^+ + d_i^- \leq D, \quad i = 1,...,n \\ &x \in X, d_i^+, d_i^- \geq 0, i = 1,...,n, D \geq 0 \end{aligned} \tag{8}$$

Donde *D* es una cota superior de las suma de las desviaciones con respecto de la meta en cada uno de los *n* objetivos. La formulación minimax es típica de los problemas aplicados en teoría de juegos o teoría de la decisión multicriterio [46], [47] y [48].

Vale la pena anotar que existen muchas aplicaciones de la programación por metas, en las cuales se utiliza la formulación original combinada con otras estrategias de modelación como programación lineal difusa aplicada a diversos sectores y temas específicos [30], [37]. Análisis envolvente de datos [31], modelación en cadenas de abastecimiento, Planeación de la producción y Localización de entes físicos [32], [49], [50], Planeación de recursos humanos [34] y Gestión financiera [35], [37].

2. REGRESIÓN MEDIANA

La regresión mediana o L_1 descrita en Toshiyuki S. et al [26] y Li, Y. et al [40] tiene por objetivo minimizar la suma de errores absolutos. Es decir:

$$\min \sum_{i=1}^{n} \left| y_i - \beta^T x^i \right| \tag{9}$$

Haciendo $\varepsilon_i = y_i - \beta^T x^i$, donde ε_i representa el error o desviación del valor estimado sobre el valor real. Si $\varepsilon_i > 0$, la estimación $\beta^T x^i < y_i$ (el valor real (y_i), es menor que el valor estimado $\beta^T x^i$). En caso contrario, el valor real observado es menor que el valor estimado. Defínase:

$$\varepsilon_i^+ = \max\{0, \varepsilon_i\}, \ \varepsilon_i^- = \max\{0, -\varepsilon_i\} \qquad (10)$$

que representan variables de desviación o errores en las estimaciones. Por lo desarrollado en la sección 1, el modelo de optimización (9) formulado por medio de programación por metas es [26]:

$$\min \ f(\varepsilon_i, i = 1,...,n, \beta) = \sum_{i=1}^{n} \varepsilon_i^+ + \varepsilon_i^-, Sujeto \ a \qquad (11)$$

$$\varepsilon_i^+ - \varepsilon_i^- = y_i - \beta^T x^i, \ i = 1,...,n$$

$$\beta \in \Re^m, \varepsilon_i^+, \ \varepsilon_i^- \geq 0$$

El problema de programación lineal (11) tiene $m + 2n$ variables y n restricciones. Las restricciones del modelo de optimización (11) representan la desviación (positiva o negativa), de la estimación con respecto del valor real y la función objetivo es equivalente a la minimización de los errores absolutos.

Ejemplo 1. Tomado de Montgomery, et al [18]. La tabla 1 presenta algunos datos de pruebas de energía solar térmica, donde:
y : flujo total del calor (Kwatts)
x_1 : insolación (watts/m^2)
x_2 : posición del foco en dirección este (pulgadas)
x_3 : posición del foco en dirección sur (pulgadas)
x_4 : posición del foco en dirección norte (pulgadas)
x_5 : hora del día

Tabla 1. Datos del ejemplo para regresión mediana

K	y_k	x_1^k	x_2^k	x_3^k	x_4^k	x_5^k
1	271.8	783.35	33.53	40.55	16.66	13.20
2	264.0	748.45	36.50	36.19	16.46	14.11
3	238.8	684.45	34.66	37.31	17.66	15.68
4	230.7	827.80	33.13	32.52	17.50	10.53
5	251.6	860.45	35.75	33.71	16.40	11.00
6	257.9	875.15	34.46	34.14	16.28	11.31
7	263.9	905.55	34.60	34.85	16.06	11.96
8	266.5	909.45	35.38	35.89	15.93	12.58
9	229.1	756.00	35.85	33.53	16.60	10.66

Por mínimos cuadrados el modelo de optimización es:

$$\min \sum_{k=1}^{1}\left(y_k - \sum_{j=1}^{5} x_j^k \beta_j \right)^2$$

Obteniendo la solución de la tabla 2:

Tabla 2. Estimadores por mínimos cuadrados del problema de regresión

β_1	β_2	β_3	β_4	β_5
0.131	1.986	4.545	-6.607	2.043

Por otro lado, utilizando regresión mediana se obtiene el siguiente modelo:

$$\min \sum_{k=1}^{1}\left| y_k - \sum_{j=1}^{5} x_j^k \beta_j \right|$$

Y su formulación por medio de programación por metas:

$$\min \sum_{k=1}^{10} (\varepsilon_k^+ + \varepsilon_k^-), s.a.$$
$$y_k - \sum_{j=1}^{5} x_j^k \beta_j = \varepsilon_k^+ - \varepsilon_k^-,\ k = 1,...,10$$
$$\varepsilon_k^+, \varepsilon_k^+ \geq 0$$

El conjunto de restricciones del problema anterior, define el error (desviación positiva o negativa), del valor real y_k con respecto de la estimación $\sum_{j=1}^{5} x_j^k \beta_j = x^k \beta$. Por ejemplo: para la primera observación ($k = 1$), la restricción obtenida es:

$$271.8 - 783.35\beta_1 - 33.53\beta_2 - 40.55\beta_3 -$$
$$16.66\beta_4 - 13.20\beta_5 = \varepsilon_1^+ - \varepsilon_1^-,$$

De este modelo de regresión L_1 los estimadores obtenidos son López H. [13].

Tabla 3. Estimadores por regresión mediana del problema de regresión

β_1	β_2	β_3	β_4	β_5
0.134	1.543	5.089	-5.575	0.292

Además, se puede hacer las siguientes comparaciones para los dos modelos:

Tabla 4. Comparaciones de resultados para los errores con mínimos cuadrados y regresión mediana

	Regresión mediana	**Regresión mínimos cuadrados**
$\sum(\varepsilon_i)^2$	252.93	201.052
$\sum\lvert\varepsilon_i\rvert$	28.262	34.527
$\lvert\sum\varepsilon_i\rvert$	4,58	0,035

Aunque los resultados obtenidos por mínimos cuadrados parecen mejores que los resultados obtenidos con regresión mediana, es importante anotar que sus objetivos son distintos ya que no es necesario hacer los mismos supuestos sobre los errores.

Ejemplo 2. Polinomio de regresión. Supongamos que se tienen los datos de la tabla 5 de la demanda anual de un producto para ser ajustados por una función polinomial de orden 2.

Tabla 5. Datos de demanda de producto

Período x	1	2	3	4	5	6	7	8	9	10
Demanda d	23	50	100	140	230	287	340	520	650	709

La ecuación del polinomio de regresión está dada por: $d = \beta_0 + \beta_1 x + \beta_2 x^2$**.** Luego el modelo por programación por metas es:

$$\min \sum_{k=1}^{10} (\varepsilon_k^+ + \varepsilon_k^-), s.a.$$

$$d_k - \beta_0 - \beta_1 x_k - \beta_2 x_k^2 = \varepsilon_k^+ - \varepsilon_k^-, \; k = 1,...,10$$

$$\varepsilon_k^+, \varepsilon_k^+ \geq 0$$

Por ejemplo: para $k = 5$, la restricción sería:

$$230 - \beta_0 - 5\beta_1 - 25\beta_2 = \varepsilon_5^+ - \varepsilon_5^-$$

Obteniendo la solución por medio de mínimos cuadrados y por programación por metas que se presenta en la tabla 6:

Tabla 6. Estimadores por mínimos cuadrados y programación por metas

	β_0	β_1	β_2
Mínimos cuadrados	5.86	10.23	6.303
Programación por metas	5.33	11.81	5.856

Gráficamente el polinomio solución se representa en la siguiente figura 1:

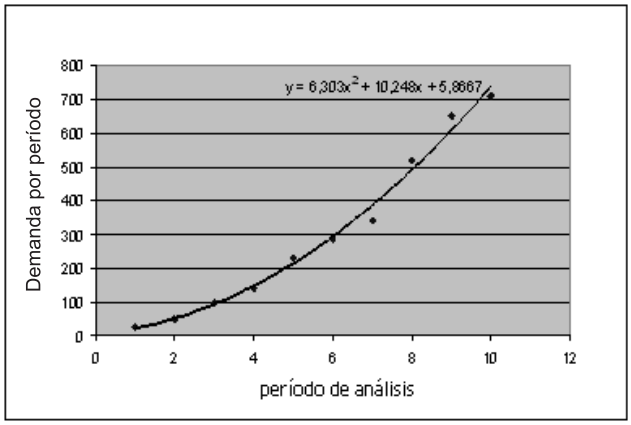

Figura 1. Ajuste polinomial de orden 2 por medio de mínimos cuadrados

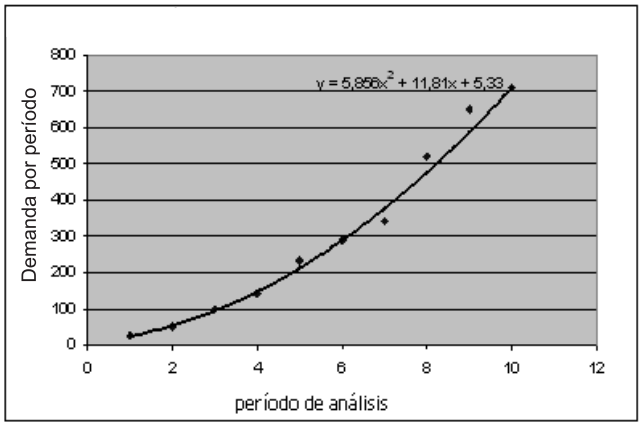

Figura 2. Ajuste polinomial de orden 2 por medio de programación por metas

Ejemplo 3. Influencia de datos atípicos. Una de las grandes desventajas de la regresión por mínimos cuadrados, es la influencia de datos atípicos. Por ejemplo: los datos descritos en la tabla 7, representan la demanda de un artículo en determinados períodos.

Tabla 7. Demanda de un artículo en determinados periodos.

Período x	**Demanda** d
1	15,90
2	18,00
3	21,43
4	24,76
5	27,86
6	30,39
7	33,47
8	36,18
9	39,64
10	42,30
11	45,73
12	12,00

En la figura 3, se presenta un diagrama de dispersión de la información anteriormente descrita.

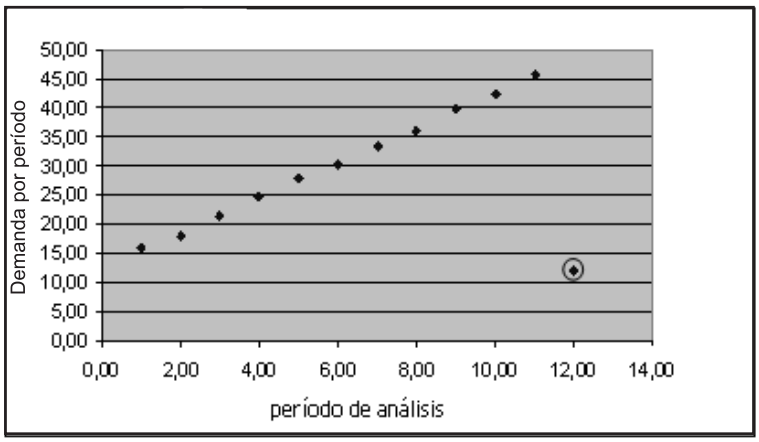

Figura 3. Diagrama de dispersión para los datos del ejemplo 3

Se podría asociar dicho comportamiento con una recta de regresión, notando que el último dato no sigue la tendencia esperada. Utilizando mínimos cuadrados, la ecuación obtenida es:

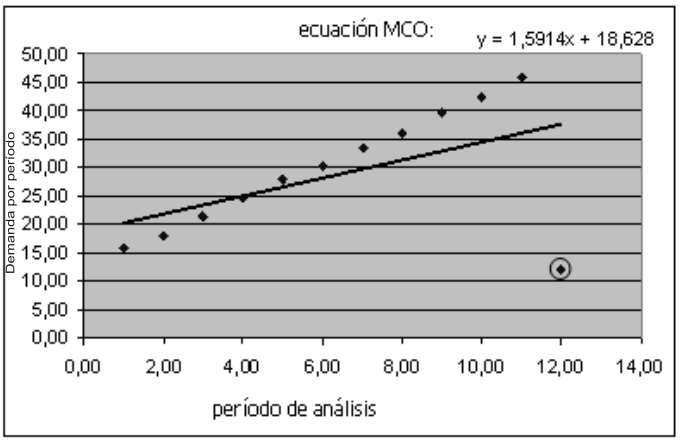

Figura 4. Recta de regresión por mínimos cuadrados

Utilizando regresión mediana, se obtienen la siguiente ecuación y recta de regresión:

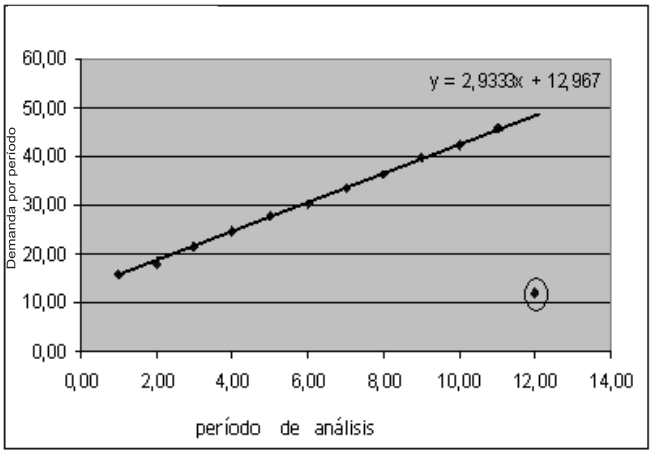

Figura 5. Recta de regresión mediana

Es posible darse cuenta de que en regresión mediana o con percentil 50, la influencia de datos atípicos es menor que en los modelos por mínimos cuadrados [8], debido a que dicha medida de tendencia es más robusta que el promedio.

Según lo indica [56], existen antecedentes teóricos, tales como lo demostrado en [57], donde se afirma que en los casos en los cuales existen datos atípicos o anómalos la regresión mediana es más consistente que por mínimos cuadrados. De esta forma, utilizando dicha técnica es posible no eliminar estos datos de la muestra. Por otra parte, Dielman T. [58], demuestra que en el caso de no tener residuos normales en mínimos cuadrados, esta situación no genera problemas en la formulación presentada en este trabajo, debido a que es posible extenderlo a otro tipo de distribuciones.

3. REGRESIÓN CUANTÍLICA O POR PERCENTILES

En los métodos de regresión lineal clásicos, el objetivo es minimizar la suma de los residuales al cuadrado y utilizar la media como estimador. Por otra parte, la regresión cuantílica o por percentiles introducida por Koenker [8], y Koenker et al [9], donde se busca minimizar una suma de errores con pesos asimétricos y utilizar los cuantíles como estimadores. La motivación para utilizar cuantíles en vez de la media, es debido a la relación estocástica que puede existir en las variables aleatorias trabajadas en los modelos, siendo posible reflejar una mejor relación y encontrar menor impresión en las inferencias realizadas. La estimación por mínimos cuadrados puede llegar a ser muy deficiente cuando los errores no distribuyen normal. En el caso de la regresión cuantílica, las estimaciones generan modelos más robustos y por consiguiente más confiables. El problema de la regresión por mínimos, que busca estimar funciones mediante el promedio, es la poca robustez de dicha medida y además, dada la heterogeneidad de la variable explicada en algunos casos, es posible que el promedio sea una medida muy pobre para realizar el análisis de la información.

Los modelos de regresión cuantílica generalizan el modelo de regresión mediana (11). En este caso, no es necesario que $E(u_i|x^i) = 0$, pero el $\tau - ésimo$ cuantil o percentil del error con respecto de las variables regresoras debe ser cero.

Cabe anotar que las condiciones para aplicar los modelos de regresión cuantílica son menos fuertes que los modelos por mínimos cuadrados. En este caso, se realiza la estimación del $\tau - ésimo$ cuantil o percentil de las variables explicadas y_i con respecto de las variables regresoras, que se anota de la siguiente forma:

$$Q_\tau(y_i|x^i) = \beta_\tau^T x^i$$

La estimación de β_τ se encuentra por medio de:

$$\min_{\beta \in \Re^m} \left\{ \begin{array}{l} \sum_{y_i \geq \beta_\tau x^i} \tau \left| y_i - \beta_\tau{}^T x^i \right| \\ + \sum_{y_i < \beta_\tau x^i} (1-\tau) \left| y_i - \beta_\tau{}^T x^i \right| \end{array} \right\} \quad (12)$$

Es decir, los errores positivos se ponderan con un valor de τ y los errores negativos se ponderan con un valor de $(\tau - 1)$. Una manera más compacta de escribir el problema de optimización (12) es por medio de la *función de chequeo* definida de la siguiente manera:

$$\rho_\tau(r) = r \times (\tau - I(r < 0)), \ 0 < \tau < 1$$

Donde: $I(r<0)=\begin{cases}1, & r<0\\ 0, & r\geq 0\end{cases}$.

De este modo, el modelo matemático (12) se puede escribir como:

$$\min\ f(\beta)=\sum_{i=1}^{n}\rho_{\tau}(y_i-\beta^T x^i) \qquad (13)$$

Cuando $\tau=0.5$, el problema de regresión cuantílica es equivalente al problema de regresión mediana.

La técnica más usada para solucionar el problema de regresión cuantílica (13), es por medio de su representación como un problema de programación lineal por metas [14]. La función de chequeo se puede escribir como la suma de dos funciones positivas:

$$\rho_{\tau}(r)=\tau p^{+}(r)+(1-\tau)p^{-}(r),$$

Donde:
$p^{+}(r)=\max\{0,r\}\ y\ p^{-}(r)=\max\{0,-r\}$.

Sean $\varepsilon_i^{+}=p^{+}(y_i-\beta^T x^i), \varepsilon_i^{-}=p^{-}(y_i-\beta^T x^i),\ \varepsilon^{+}=(\varepsilon_1^{+},...,\varepsilon_n^{+}), \varepsilon^{-}=(\varepsilon_1^{-},...,\varepsilon_n^{-})$.

De esta forma, la formulación del problema de regresión cuantílica (13), como un problema de programación lineal por metas, está dada por Koenker, R., [8]:

$$\min \sum_{i=1}^{n}\tau\varepsilon_i^{+}+(1-\tau)\,\varepsilon_i^{-}$$

Sujeto a (14)

$$\varepsilon_i^{+}-\varepsilon_i^{-}=y_i-\beta^T x^i,\ \ i=1,...,n$$
$$\beta\in\Re^m,\ \varepsilon_i^{+},\ \varepsilon_i^{-}\geq 0$$

En este caso, las restricciones son equivalentes a las del problema de regresión mediana.

La función objetivo cambia por las ponderaciones asimétricas de las variables de desviación. A continuación, se muestra un caso de aplicación de la regresión cuantílica con una sola variable explicativa.

Ejemplo 4. Los datos presentados a continuación son simulados. Sea $x=$ años invertidos en educación y la variable explicada $y=$ retribución salarial por hora en unidades monetarias. El diagrama de dispersión para los datos es el siguiente:

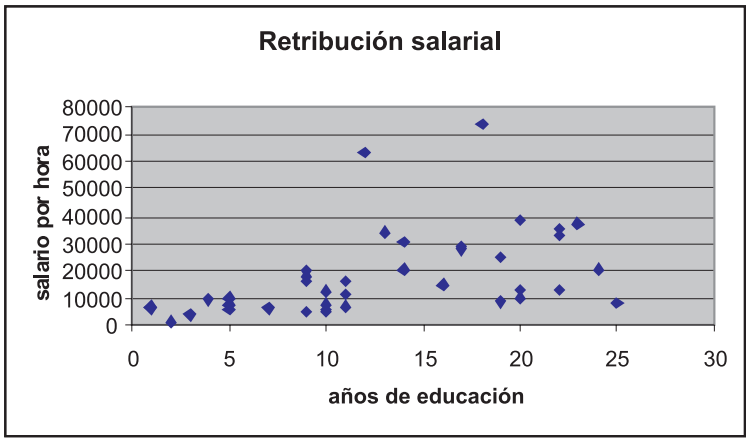

Figura 6. Diagrama de dispersión de los datos del ejemplo 4

Obteniendo las curvas de regresión por percentiles con el paquete estadístico R [22], descritas en la figura 7, con los valores de τ =0.10, 0.25, 0.5, 0.75 y 0.9.

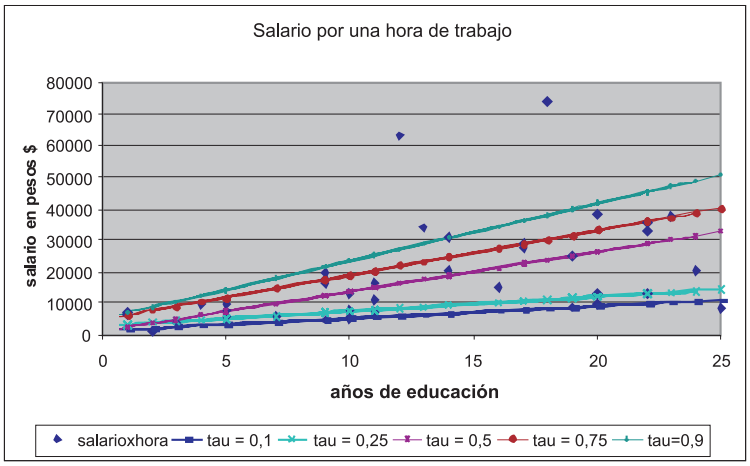

Figura 7. Rectas de regresión cuantílica para los datos del ejemplo 4

Análisis de resultados. Por ejemplo: la curva obtenida para $\tau = 0.1$ es $y = 1.3 + 384x$, implica que para los elementos de la población con menor salario (inferior al 10% o por debajo del percentil 10), se tendrá que por cada año adicional invertido en educación, su aumento promedio salarial por hora es de 384 unidades monetarias. Otra aplicación importante en este ejemplo, es la siguiente: para una persona que ha estudiando 16 años, se obtienen las siguientes estimaciones en los percentiles 10 y 90:

$y_{10}(16)$ = 7472 unidades monetarias
$y_{90}(16)$ = 34306 unidades monetarias

Como en ese rango se encuentra el 80% de la población, se espera que con dicho nivel de estudio (16 años), la retribución salarial por hora de los individuos se encuentre entre esas dos estimaciones. En la tabla 8, se presenta el valor $\beta_{0,\tau}$ y $\beta_{1,\tau}$ para los distintos modelos de regresión cuantílica de la gráfica 7, recordando que la recta de regresión para el percentil τ se nota: $y_\tau = \beta_{0,\tau} + \beta_{1,\tau} x$.

Tabla 8. Valores de los estimadores para los modelos de regresión cuantílica

τ =	0.1	0.25	0.5	0.75	0.9
$\beta_{0,\tau}$	1332	2807	1067	4806,75	5122
$\beta_{1,\tau}$	384	458	1262	1411,25	1824

4. REGRESIÓN CUANTÍLICA PONDERADA

En regresión, uno de los objetivos de los modelos por mínimos cuadrados ponderados es corregir problemas tales como varianza no constante para los errores o simplemente dar una jerarquía o nivel de importancia a cada una de las observaciones. En este tipo de modelos, es necesario introducir un nuevo parámetro definido de la siguiente forma:

$$w = (w_1, w_2, ..., w_n)$$

Cada w_i representa el factor de ponderación de la observación $i, i = 1,...,n.$

El problema de optimización para mínimos cuadrados ponderados se formula de la siguiente manera Montgomery, et al [18]:

$$\min \sum_{i=1}^{n} w_i (y_i - \beta^T x^i)^2$$

Por otra parte, el modelo de regresión cuantílica ponderada se escribe como:

$$\min_{\beta \in \Re^m} \left\{ \begin{array}{l} \sum_{y_i \geq \beta_\tau x^i} \tau \; w_i \left| y_i - \beta_\tau^T x^i \right| \\ + \sum_{y_i < \beta_\tau x^i} (1-\tau) w_i \left| y_i - \beta_\tau^T x^i \right| \end{array} \right\} \qquad (15)$$

De forma equivalente (15), se expresa como un modelo de programación por metas de la siguiente manera:

$$\min \sum_{i=1}^{n} \tau\, w_i \varepsilon_i^+ + (1-\tau)\, w_i \varepsilon_i^-$$

Sujeto a (16)

$$\varepsilon_i^+ - \varepsilon_i^- = y_i - \beta^T x^i,\ i = 1,...,n$$
$$\beta \in \Re^m, \varepsilon_i^+,\ \varepsilon_i^- \geq 0$$

En el caso que $w_i = 1,\ i = 1,...,n$ el problema (16) es equivalente a (14). Cuando $w_i = 1,\ i = 1,...,n$ y $\tau = \frac{1}{2}$, el problema (16) es equivalente al problema de regresión mediana (11). Matricialmente (16), se expresa de la siguiente manera:

$$\min z(\beta, \varepsilon^+, \varepsilon^-) = \left[\tau w^T, (1-\tau) w^T, \vec{0}_m\right] \begin{bmatrix} \varepsilon^+ \\ \varepsilon^- \\ \beta \end{bmatrix}$$

$s.\ a.$

$$\left[X \quad I_n \quad -I_n\right] \begin{bmatrix} \beta \\ \varepsilon^+ \\ \varepsilon^- \end{bmatrix} = y \qquad (17)$$

$$\varepsilon_i^+, \varepsilon_i^+ \geq 0,\ i = 1,...,n$$
$$\beta \in \Re^m$$

Nótese que una ventaja de esta formulación por metas, es que permite incluir restricciones sobre los parámetros [56], por conocimientos a priori, tales como intervalos de pertenencia (cotas superiores e inferiores), relaciones entre parámetros, signos, etc., extendiendo la formulación clásica de mínimos cuadrados ordinarios. Estas restricciones sobre los parámetros controlan valores no factibles de los estimadores y problemas de mal ajuste [59]. A continuación, se presenta un ejemplo para mostrar una de las utilidades de la regresión cuantílica ponderada.

Ejemplo 5. Con los datos del ejemplo 4, se obtiene la curva de regresión cuantílica con $\tau = 0.9$ y dada por: $y_{90} = 5122 + 1824x$. Una de las aplicaciones de los modelos de regresión ponderados es en la influencia de datos atípicos. En este caso, se encuentran dos datos fuera del comportamiento promedio de la población.

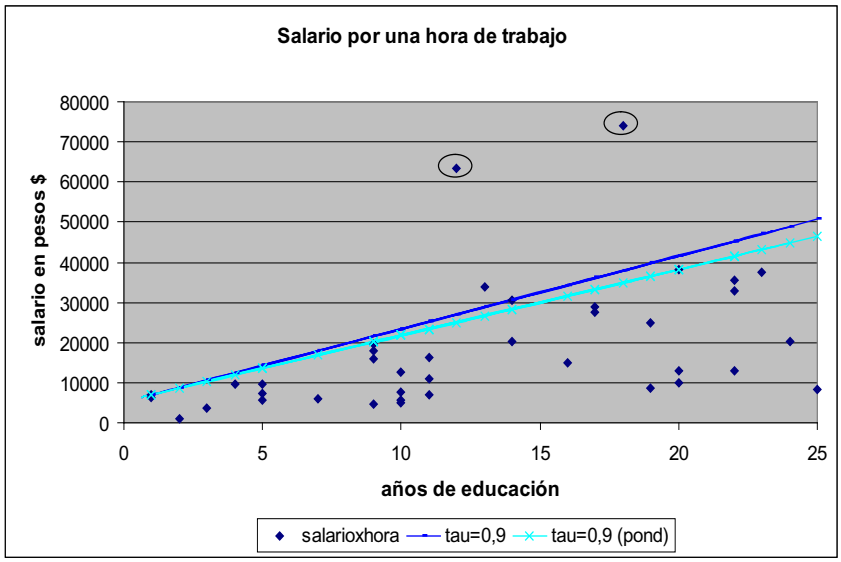

Figura 8. Rectas de regresión cuantílica y cuantílica ponderada

Para evitar la influencia de dichos datos, se hace una ponderación dándole menor importancia. Se observa que en el modelo de regresión cuantílica con $\tau = 0.9$ ponderado, la estimación promedio es menor. Es importante anotar que los datos atípicos sólo influyen en los percentiles extremos. Otra aplicación de los modelos ponderados de mayor uso, es cuando se tienen factores de expansión por muestreo [23].

5. MODELO DUAL DE OPTIMIZACIÓN

Es posible asociar a (17), un problema de programación lineal dual [3]. Sea $d = [d_1,...,d_n]^T$, el vector de variables duales correspondiente al problema (14). Así, el dual asociado al problema anterior es:

$$\begin{aligned} &\max\ y^T d, \textit{ sujeto a} \\ &X^T d = 0 \\ &(\tau - 1)w \le d \le \tau w \end{aligned} \qquad (18)$$

El problema dual es un problema de programación lineal con variables acotadas y tiene m+2n restricciones y n variables. Es decir, son menos variables que en problema primal (17). En la práctica, es más fácil resolver el problema dual para regresión cuantílica ponderada que el primal. Algebraicamente, el modelo matricial (18) se escribe de la siguiente manera:

$$\begin{aligned} &\max \sum_{j=1}^{n} y_j d_j, \; sujeto \quad a \\ &\sum_{j=1}^{n} x_i^j d_j = 0, \quad i = 1,...,m \\ &(\tau - 1) w_j \le d_j \le \tau w_j, \quad j = 1,...,n. \end{aligned} \tag{19}$$

6. FORMULACIÓN DEL PROBLEMA DE OPTIMIZACIÓN POR METAS CUADRÁTICO Y MINI-MAX

La norma L_p aplicada a los modelos de regresión lineal múltiple tiene por objetivo minimizar la siguiente función objetivo:

$$\min \left(\sum_{i=1}^{n} \left| y_i - \beta^T x^i \right|^p \right) \tag{20}$$

Haciendo $\varepsilon_i = y_i - \beta^T x^i$, donde ε_i representa el error o desviación del valor estimado sobre el valor real. Si se define $\varepsilon_i^+ = \max\{0, \varepsilon_i\}$ y $\varepsilon_i^- = \max\{0, -\varepsilon_i\}$ son las variables de desviación o errores en las estimaciones. De acuerdo con lo desarrollado en la sección 1, el modelo de optimización (20) formulado por medio de la programación por metas no lineal es:

$$\min \; f(\varepsilon_i, i = 1,...,n, \beta) = \sum_{i=1}^{n} \left(\varepsilon_i^+ + \varepsilon_i^- \right)^p$$

Sujeto a

$$\begin{aligned} &\varepsilon_i^+ - \varepsilon_i^- = y_i - \beta^T x^i, \; i = 1,...,n \\ &\beta \in \Re^m, \varepsilon_i^+, \; \varepsilon_i^- \ge 0 \end{aligned} \tag{21}$$

Claramente cuando *p=2*, se tiene el problema típico de regresión por mínimos cuadrados, y se obtiene una formulación de optimización por metas equivalente a la formulación clásica de econometría [51], [52] y cuando p=1, se tiene el problema de regresión mediana. Un caso interesante es cuando p=∞, debido a que se tiene el siguiente problema de regresión múltiple min y max:

$$\min \max \left\{ \left| y_i - \beta^T x^i \right|, i = 1,...,n \right\}$$

Que de forma equivalente, se puede escribir como un problema de programación por metas de la siguiente forma:

$$\min f(\varepsilon_i, i=1,...,n,\beta) = \max\left\{\varepsilon_i^+ + \varepsilon_i^- : i=1,...,n\right\}$$

Sujeto a (22)

$$\varepsilon_i^+ - \varepsilon_i^- = y_i - \beta^T x^i,\ i=1,...,n$$
$$\beta \in \Re^m, \varepsilon_i^+,\ \varepsilon_i^- \geq 0$$

Como el problema anterior no es diferenciable en su función objetivo, entonces al añadir una cota superior a todas las sumas de las desviaciones [27], el problema anterior es equivalente a: $\min \quad D$

Sujeto a (23)

$$\varepsilon_i^+ - \varepsilon_i^- = y_i - \beta^T x^i,\ i=1,...,n$$
$$\varepsilon_i^+ + \varepsilon_i^- \leq D,\quad i=1,...,n$$
$$\beta \in \Re^m, \varepsilon_i^+,\ \varepsilon_i^- \geq 0$$

Nótese que en este caso, se minimiza la máxima desviación de los residuos (*D*), y se encuentra un modelo más equilibrado que los modelos clásicos [60]. Debido a que esta formulación se basa en el máximo de una muestra, esta medida no es robusta ante la presencia de datos atípicos, por lo cual se sugiere su uso en el caso de una muestra homogénea [56].

Aznar y et al [56], proponen cambiar el conjunto de restricciones $\varepsilon_i^+ + \varepsilon_i^- \leq D,\quad i=1,...,n$ por restricciones de la forma $\varepsilon_i^+ + \varepsilon_i^- \leq Dy_i,\quad i=1,...,n$ cuando se están estimando precios, viajes o cualquier variable positiva.

La formulación *minimax* para modelos de regresión múltiple tiene aplicaciones en varias ramas de la estadística e ingeniería, tal como regresión robusta [53] y [54], regresión local [55], Estadística no paramétrica, etc.

Algunos autores proponen el uso de un modelo biobjetivo [56], de tal forma que se pueda incluir la máxima desviación de los errores y la regresión mediana (siendo esta una técnica excelente en el caso de datos atípicos en la muestra), y llegar a soluciones intermedias por medio de la generación de la frontera de Pareto. Matemáticamente, el problema biobjetivo se escribe como:

$$\min \quad \left(D, \sum_{i=1}^{n}\left(\varepsilon_i^+ + \varepsilon_i^-\right)\right)$$

Sujeto a (24)

$$\varepsilon_i^+ - \varepsilon_i^- = y_i - \beta^T x^i, \ i = 1,...,n$$
$$\varepsilon_i^+ + \varepsilon_i^- \leq D, \quad i = 1,...,n$$
$$\beta \in \Re^m, \varepsilon_i^+, \ \varepsilon_i^- \geq 0$$

Esta formulación es interesante, debido a que es posible utilizar técnicas clásicas de optimización multiobjetivo [24] para analizar el conjunto de soluciones y su implicación en la estimación y predicción.

7. CONCLUSIONES

Los modelos de programación multiobjetivo por metas, son una herramienta útil para el cálculo de estimadores en modelos de regresión múltiple debido a su formulación equivalente.

Es interesante aplicar estos modelos de regresión cuantílica y mediana cuando los supuestos sobre los errores en mínimos cuadrados no se cumplen de forma satisfactoria o de forma completa. Además, las formulaciones como un problema de programación por metas, permite incluir restricciones sobre los parámetros por estimar y de esta forma, controlar a priori condiciones no factibles o conocimiento dado sobre los estimadores.

Los modelos de regresión mediana permiten menor influencia de datos atípicos con respecto de los modelos de regresión por mínimos cuadrados. De esta forma, utilizando dicha técnica, es posible no eliminar estos datos de la muestra.

La regresión cuantílica permite analizar la distribución de la variable respuesta en distintos puntos y determinar rangos en la predicción para un individiduo determinado.

La estimación por mínimos cuadrados puede llegar a ser muy deficiente cuando los errores no distribuyen normal. En el caso de la regresión cuantílica, las estimaciones generan modelos más robustos y por consiguiente, más confiables. El problema de la regresión por mínimos cuadrados, que busca estimar funciones con el promedio, es la poca robustez de dicha medida y además, dada la heterogeneidad de la variable explicada en algunos casos, es posible que el promedio sea una medida muy pobre para realizar el análisis de la información.

La programación por metas permite formular modelos de optimización que involucren el valor absoluto en la función objetivo (optimización no diferenciable), como un problema de programación lineal, introduciendo variables de desviación.

AGRADECIMIENTOS

Los autores agradecen los valiosos comentarios, sugerencias y observaciones de los evaluadores anónimos.

REFERENCIAS BIBLIOGRÁFICAS

[1] Cade B., Noon B., (2003) "A Gentle Introduction to Quantile Regression for Ecologists", *Frontiers in Ecology and the Environment* **1**(8), 412-420.

[2] Cade B., Terrell J., y Schroeder R.; (1999) Estimating effects of limiting factors with regression quantiles, Ecology, 80, 311-323.

[3] Chernozhukov, V. and L. Umantsev. 2001. "Conditional Value-at-Risk: Aspects of Modeling and Estimation." *Empirical Economics.* March, 26:1, pp. 271–92.

[4] Eide E., y Showalter M., (1998) The effect of school quality on student performance: a quantile regression approach, Economics Letters, 58, 345-50.

[5] Engle Robert and Simone Manganelli.; (1999) "*CaViaR: Conditional Autoregressive Value at Risk by Regression Quantiles*." University of California, San Diego, Department of Economics Working Paper.

[6] Fitzenberger B., Koenker R., Machado J., A.F. (Eds.); (2002) Economic Applications of Quantile Regression. Series: Studies in Empirical Economics . VI, 324 p. 74 illus., Hardcover.

[7] Knight K., Bassett G., y Mo-Yin S. Tam.; (2002) "Comparing Quantile Estimators for the Linear Model." Preprint.

[8] Koenker R., (2005) *Quantile Regression*, Econometric Society Monographs, Cambridge University Press.

[9] Koenker R., Basset G.; (1978) *Regression Quantiles.* Econometrica 46, 33-50.

[10] Koenker R., Geiling O.; (2001) *'Reappraising medfly longetivity: A quantile regression approach',* Journal of American Statistic Association 96, 458–468.

[11] Koenker R., Machado J.; (1999) *'Goodness of fit and related inference processes for quantile regression'*, Journal of the American Statistical Association 94, 1296–1310.

[12] López R., (2007) La brecha de la distribución salarial en Colombia, un efecto de discriminación?. Monografía de pregrado. Carrera de Estadística. Universidad Nacional de Colombia.

[13] López H.; (2006) *Introducción a GAMS y su aplicación en la solución de modelos matemáticos de optimización, in* 'Memorias del XXII Coloquio distrital de Matemáticas y Estadística', Universidad Nacional de Colombia, Bogotá.

[14] López H., Mora H.; (2007) Cálculo de los estimadores de regresión cuantílica lineal por medio del método ACCPM. *Revista Colombiana de Estadística 30 (1).* 53 a 68.

[15] Ma L., Koenker R.; (2003) Quantile regression methods for recursive structural models, Technical report.

[16] Manning W., Blumberg L., y Moulton L.; (1995) The demand for alcohol: the differential response to price, Journal of Health Economics, 14, 123-148.

[17] McKeague I., Subramanian S., y Sun Y. Q.; (2002) Median regression and the missing information principle, Journal of Nonparametric Statistics 13, 709-727.

[18] Montgomery Peck, y Vining; (2004) *Al análisis de regresión lineal.* 3ª edición. Editorial CECSA.

[19] Poterba James, and Kim Rueben; (1995) "*The Distribution of Public Sector Wage Premia: New Evidence Using Quantile Regression Methods*." NBER Working Paper No. 4734.

[20] Powell J.L.; (1986) *'Censored regression quantiles'*, Journal of Econometrics 32, 143– 55.

[21] Powell J.; (2002) Notes On Median and Quantile Regression. Department of Economics.University of California, Berkeley.

[22] R. Development Core Team (2006), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. *http://www.R-project.org.

[23] Särndal C. Swensson B., Wretman J., Model Assisted Survey Sampling. Springer Series in Statistics. Segunda Edición. 2003.

[24] Smith R., Mesa O., Dyner I., Jaramillo P., Poveda G., Valencia D.; (2000) *Decisiones con múltiples objetivos e incertidumbre.* 2a edición. Facultad de Minas. Universidad Nacional de Colombia. Sede Medellín.

[25] Sosa Escudero W.; (2006) *Perspectivas y Avances Recientes en Regresion por Cuantiles*, en Marchionni, M. (editora), Ahumada, H., Jorrat, J., Navarro, M. y Sosa Escudero, W., Progresos en Econometría, Temas Grupo Editorial, Buenos Aires.

[26] Toshiyuki S., Yih-Long C.; (1989) Goal Programming Approach for Regression Median. Decision Sciences 20 (4), 700–713.

[27] Winston W., *Investigación de operaciones: aplicaciones y algoritmos.* Séptima edición. Editorial Thomson. 2005.

[28] Wooldridge, J.M.; (2002) *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge, Massachusetts.

[29] Zarate, H.; (2002) *Cambios en la estructura salarial: una historia desde la regresión cuantílica.* CEMLA. Colombia.

[30] Dutta D., y Murthy S.; (2010) MULTI-CHOICE GOAL PROGRAMMING APPROACH FOR A FUZZY TRANSPORTATION PROBLEM. *International Journal of Research and Reviews in Applied Sciences. Volume 2*.

[31] Makui A., Alinezhad A., Kiani Mavi R., Zohrehbandian M.; (2008) A Goal Programming Method for Finding Common Weights in DEA with an Improved Discriminating Power for Efficiency. *Journal of Industrial and Systems Engineering.* Vol. 1, No. 4, pp 293-303.

[32] Mezghani M., Rebai A., Dammak A., Loukil T.; (2009) A Goal Programming model for Aggregate Production Planning problem. *International Journal of Operational Research. Vol. 4, No.1 pp. 23 – 3.*

[33] Güneş y Umarosman.; (2005) FUZZY GOAL PROGRAMMING APPROACH ON COMPUTATION OF THE FUZZY ARITHMETIC MEAN**.** *Mathematical and Computational Applications*, Vol. 10, No. 2, pp. 211-220.

[34] Glynn J., College C.; (2005) A Goal Programming Approach To Human Resource Planning With A Concentration On Promotion Policy. Journal of Business & Economics Research . Volumne 3, No. 3.

[35] Lin T., y O'Leary D., (1993) Goal programming applications in financial management. Advances in mathematical programming and financial planning*,* Volume 3, pages 211-229.

[36] Tamiz M., Jones D., y El-Darzi E., (1995) A review of Goal Programming and its applications. *Annals of Operations Research***.** Volume 58, Number 1.

[37] Dylan, Tamiz, Mehrdad, Ries, Jana (Eds.); (2010) New Developments in Multiple Objective and Goal Programming. Series: Lecture Notes in Economics and Mathematical Systems, Vol. 638.

[38] Marcenaro O., Navarro L., (2007) El éxito en la universidad: Una aproximación cuantílica. Revista de Economía Aplicada Número 44 (vol. XV), págs. 5 a 39.

[39] Block S., Masters W., y Bhagowalia P.; (2010) Child Undernutrition, Household Poverty and National Income in Developing Countries: Quantile Regression Results. Selected Paper prepared for presentation at the Agricultural & Applied Economics Association.

[40] Li Y., y Zhu, J.; (2008) L1-norm quantile regressions, *Journal of Computational and Graphical Statistics* 17: 163–185.

[41] González-Pachón y Romero. (2001) Aggregation of partial ordinal rankings: an interval goal programming approach. *Computers & Operations Research*, Volume 28, Issue 8. Pág. 827-834.

[42] Charnes A., Collomb B.; (1972) Optimal economic stabilization policy: Linear goal-programming models. *Socieconomic Planning Sciences*, Vol. 6. 431-5.

[43] Ignizio J.P.; (1974) Interval goal programming and applications. Documento de trabajo. Pennsylvania State University, USA.

[44] Du D., y Pardalos P.M.; (1995) Minimax And Applications, Kluwer, Wiley: Dordrecht.

[45] Demyanov V.F., y Molozemov N.V.; (1974) Introduction to Minimax. Wiley: New York.

[46] Rao S.S.; (1987) Game Theory approach for multiobjetive structural optimization. Computers & Structures. Volume 25. Issue 1. Pag. 119-127.

[47] Aumann R., y Hart S.; (1994) Handbook of game theory with theory applications. Libro. Elsevier, 814 páginas.

[48] Simon D. A.; (2006) Game Theory Approach to constrained Minimax State Estimation. IEEE Transactions on signal Processing.

[49] Schniederjans M., Kwak N.K., y Helmer M.; (1982) An Application of Goal Programming to Resolve a site Location Problem. *Interfaces*. Vol 12. No. 3.

[50] Badri M.; (1999) Combining the analiytic hierchachy process and goal programming for global facility location-allocation problem. International Journal of Production Economics. Volume 62, Issue 3, pág: 237-248.

[51] Greene W.H.; (1998) *Análisis econométrico.* Prentice Hall.

[52] Novales A.; (1996) *Econometría. McGraw Hill.*

[53] Wiens D.; (1990) Robust minimax designs for multiple linear regression. *Linear Algebra and its Applications*. Volume 127. Páginas 327-340.

[54] Huber P., y Ronchetti E.; (2009) Robust Statics. John Wiley and Sons. 354 páginas.

[55] Fan J.; (1993) Local Linear regression Smoothers and their minimax efficiencies. The Annals of Statistics. Vol, 21, No 1, 196-216.

[56] Aznar J., y Guijarro F.; (2004) Métodos de valoración basados en la programación por metas: modelo de valoración restringida. *Estudios Agrosociales* y *Pesqueros*. Numero 204. Pp 29-45.

[57] Basset G., y Koenker R.; (1978) Asymptotic theory of least absolute Errors. *The American Statistician*, 51 (2): pp. 99-105.

[58] Dielman T. A.; (1986) Comparison of forecasts from least absolute value and least squares regression. *Journal of Forecasting*, 5 (3): pp. 189-195.

[59] Charnes A., Cooper W. W., y Sueyoshi T.; (1986) Least squares/ridge regression and goal programming/constrained regression alternatives. European Journal of Operational Research, 27 (2). Pp. 146-157.

[60] Romero C.; (2001) Extended lexicographic goal programming: a unifying approach. *Omega*, 29 (1): pp. 63-71.