Revista Colombiana de Estadística
Volumen 30 No. 1. pp. 53 a 68. Junio 2007

Cálculo de los estimadores de regresión cuantílica lineal por medio del método ACCPM

Calculus of the Estimators of Linear Quantile Regression by the Method ACCPM

Héctor Andrés López[1,a], Héctor Manuel Mora[2,b]

[1]Universidad de la Sabana, Facultad de Ingeniería, Área de Matemáticas Aplicadas y Estadística, Chía, Colombia

[2]Universidad Nacional de Colombia, Facultad de Ciencias, Departamento de Matemáticas, Bogotá

Resumen

Se muestra cómo calcular los estimadores en regresión cuantílica por medio del método de optimización no diferenciable ACCPM (Analytic Center Cutting Plane Method). El cálculo de dichos estimadores usualmente se encuentra por medio de programación lineal y sus respectivas técnicas de solución (método simplex, métodos de punto interior, etc.). La primera parte presenta las generalidades de la regresión cuantílica y su formulación como un problema de programación lineal. Además, se realiza una breve descripción del método ACCPM. Por último, se muestra la aplicación del método ACCPM para el cálculo de estimadores por cuantiles y los resultados numéricos y comparaciones del método ACCPM con el paquete estadístico R y el paquete de optimización GAMS.

***Palabras clave*:** optimización, estimador de regresión, programación lineal, estimación cuantílica.

Abstract

The present work shows how to calculate the estimators in quantile regression by nondifferentiable optimization method ACCPM (Analytic Center Cutting Plane Method). The calculus of the estimators is usually found by linear programming and its respective techniques of solution (Simplex method, interior point methods, etc.). The first part presents some generalities of quantile regression and its formulation as a linear programming problem. Also, a brief description of the ACCPM method is made. Finally, it is shown the application of the ACCPM method for the calculation of the estimators by quantiles and the numerical results and comparisons of the ACCPM with the statistic package R and the optimization package GAMS.

***Key words*:** Optimization, Regression estimator, Linear programming, Quantile estimation.

[a]Profesor. E-mail: hector.lopez1@unisabana.edu.co
[b]Profesor titular. E-mail: hmmorae@unal.edu.co

1. Introducción a la regresión cuantílica

En los modelos de regresión, los errores se asumen como una sucesión u_n de variables aleatorias independientes e idénticamente distribuidas con media cero ($E(u_n) = 0$). Generalmente la distribución que se asume es la normal. Sin embargo, no siempre se cumple el supuesto de normalidad ya que la distribución puede ser asimétrica. Koenker & Bassett (1978) introducen el concepto de regresión cuantílica (RC) como una solución a dichos problemas y demuestran que los estimadores por cuantiles son más eficientes que el estimador máximo verosímil de muchos modelos paramétricos convencionales.

En los métodos de regresión clásicos el objetivo es minimizar la suma de los residuales al cuadrado y utilizar la media como estimador. La regresión cuantílica busca minimizar una suma de errores absolutos ponderados con pesos asimétricos y utiliza los cuantiles como estimadores.

1.1. Definición de cuantil

La RC utiliza la noción clásica de cuantil para el cálculo de las estimaciones.

Dado un $\tau \in (0, 1)$ y una variable aleatoria Y (continua o discreta), el $\tau-$ésimo cuantil es definido como:

$$Q(\tau) = \inf \{y : F(y) \geq \tau\}$$

donde F es la función de distribución de Y.

Por otro lado, si se tiene $\{Y_1, Y_2, \ldots, Y_n\}$, una muestra con observaciones independientes, es posible encontrar una estimación de la función de distribución por medio de la distribución empírica de la muestra definida como el cociente entre el número de las observaciones inferiores o iguales al valor de interés y el número total de las observaciones:

$$\widehat{F}(y) = \frac{\#(Y_i \leq y)}{n} \tag{1}$$

Análogamente, es posible definir una estimación de los cuantiles por medio de la distribución empírica así:

$$\widehat{Q}(\tau) = \inf \left\{y : \widehat{F}(y) \geq \tau\right\} \tag{2}$$

El problema (2) es equivalente a:

$$\widehat{Q}(\tau) = \underset{\varepsilon_\tau \in \mathbb{R}}{\operatorname{argmin}} \left\{ \sum_{y_i \geq \varepsilon_\tau} \tau \left|y_i - \varepsilon_\tau\right| + \sum_{y_i < \varepsilon_\tau} (1 - \tau) \left|y_i - \varepsilon_\tau\right| \right\}$$

Otra manera de encontrar $\widehat{Q}(\tau)$ es a través de una *función de chequeo* definida de la siguiente manera:

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

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

De este modo el problema (3) correspondiente al cálculo del $\tau-$ésimo cuantil queda reformulado así:

$$\widehat{Q}(\tau)=\underset{\varepsilon_\tau\in\mathbb{R}}{\operatorname{argmin}}\sum_i \rho_\tau(y_i-\varepsilon_\tau) \tag{3}$$

1.2. Regresión cuantílica

Dados m vectores $x^1,\ldots,x^m\in\mathbb{R}^n$, que representan las variables explicativas y m valores reales $y_1,y_2,\ldots,y_m$, que representan la variable explicada[1], en los problemas de regresión por mínimos cuadrados se busca un vector $\beta=(\beta_1,\ldots,\beta_{n-1},\beta_n)^T\in\mathbb{R}^n$, solución del siguiente problema de optimización:

$$\min f(\beta)=\sum_{i=1}^{m}\left(y_i-\beta^T x^i\right)^2 \tag{4}$$

Si asumimos que $y_i-\beta^T x^i=u_i$, $i=1,2,\ldots,n$ y que el valor esperado condicional de u_i con respecto a las observaciones es cero $\left(E(u_i\mid x^i)=0\right)$, entonces la media condicional de y_i con respecto a x^i es

$$E(y_i\mid x^i)=\beta^T x^i$$

La solución del problema de optimización (4) está dada por

$$\beta=\left(X^TX\right)^{-1}X^Ty$$

donde $X=\left[\begin{array}{cccc}x^1 & x^2 & \cdots & x^m\end{array}\right]^T$ y $y=[y_1,y_2,\ldots,y_m]$.

Ahora, si se supone que $y_i=\beta_\tau^T x^i+u_{i,\tau}$ y además que el valor esperado condicional no necesariamente es cero, pero el $\tau-$ésimo cuantil del error con respecto a las variables regresoras es cero $\left(Q_\tau(u_{i,\tau}\mid x^i)=0\right)$, entonces el $\tau-$ésimo cuantil de y_i con respecto a las variables regresoras se puede escribir

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

La estimación de β_τ se encuentra por medio de

$$\widehat{\beta}_\tau=\arg\min_{\beta_\tau\in\mathbb{R}^n}\left\{\sum_{y_i\geq\beta_\tau^T x^i}\tau\left|y_i-\beta_\tau^T x^i\right|+\sum_{y_i<\beta_\tau^T x^i}(1-\tau)\left|y_i-\beta_\tau^T x^i\right|\right\} \tag{5}$$

que es equivalente al siguiente problema de optimización:

$$\widehat{\beta}_\tau=\underset{\beta_\tau\in\mathbb{R}^n}{\operatorname{argmin}}\sum_{i=1}^{m}\rho_\tau\left(y_i-\beta^T x^i\right) \tag{6}$$

donde ρ_τ es la *función de chequeo* y τ es un valor en $(0,1)$.

El problema (6) resulta ser un problema de optimización convexa.

[1]Es decir, se tienen n variables explicativas y el tamaño de la muestra es m.

1.3. Cálculo de $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\tau}}$ por medio de programación lineal

La técnica más usada para solucionar el problema de regresión cuantílica (6) es por medio de su representación como un problema de programación lineal (Koenker 2005). La *función de chequeo* ρ_τ 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 $u_i = p^+(y_i - \beta^T x^i)$, $v_i = p^-(y_i - \beta^T x^i)$, $u = (u_1, \ldots, u_m)$, $v = (v_1, \ldots, v_m)$ y $\mathbf{1} = [1, 1, 1, \ldots, 1]$ un vector de unos de dimensión adecuada.

La formulación del problema de regresión cuantílica como un problema de programación lineal está dada por:

$$\min\{\tau \mathbf{1}^T u + (1-\tau)\mathbf{1}^T v : y = X\beta + u - v, (u, v) \in \mathbb{R}_+^{2m}\} \tag{7}$$

El problema de programación lineal (7) tiene $n + 2m$ variables, m restricciones y $2m$ variables no negativas. La formulación dual del problema de regresión cuantílica es

$$\max\left\{y^T d : X^T d = 0, d \in [\tau - 1, \tau]^m\right\} \tag{8}$$

donde $d = [d_1, d_2, \ldots, d_m]^T$ es el vector de variables duales. Dicho problema tiene $n+2m$ restricciones y m variables. Es decir, son menos variables que en el problema primal. Por lo tanto, en la práctica es más fácil resolver el problema dual para regresión cuantílica que el problema primal.

La formulación del problema dual para regresión cuantílica es equivalente a la usada en la formulación estándar de los métodos de punto interior para programación lineal con variables acotadas. Dicho algoritmo se encuentra descrito en Koenker (2005) e implementado en el paquete `quantreg` del software estadístico R. Este paquete es el más usado por las personas que trabajan regresión cuantílica.

2. Método ACCPM

El método ACCPM (Analytic Center Cutting Plane Method) fue creado por Goffin, Haurie & Vial (1992). El método ACCPM hace parte de los métodos de planos de corte. Se presentan los conceptos básicos del método y algunas observaciones sobre su implementación desarrollada en Petón & Vial (2001).

2.1. Métodos de planos de corte

La mayoría de algoritmos de planos de corte resuelven problemas como el siguiente:

$$\begin{aligned} &\min c^T x \\ &\text{s.a. } x \in X \end{aligned}$$

donde $X \subset \mathbb{R}^n$ es un conjunto convexo y acotado. Los problemas de la forma

$$\begin{aligned} &\min f(y) \\ &\text{s.a. } y \in Y \end{aligned}$$

donde $Y \subset \mathbb{R}^{n-1}$ es un conjunto convexo y f es convexa, se pueden convertir a problemas con la formulación del problema inicial de la siguiente manera:

$$\begin{aligned} &\min z \\ &\text{s.a. } f(y) - z \leq 0 \\ &y \in Y \end{aligned}$$

tomando $x = (z, y)$ y $X = \{(z, y) : f(y) - z \leq 0, y \in Y\}$.

Estos métodos construyen una aproximación lineal de la región factible X "mejorándola" en cada iteración.

Sea P_0 una aproximación poliédrica de X ($X \subset P_0$) y x^0 el punto óptimo de $c^T x$ en P_0. La formulación general de un algoritmo de planos de corte para resolver el problema anterior es:

Método de planos de corte
Inicialización
$k := 0$
Definir $P_0 \supset X$
Encontrar $x^0 = \arg\min \{c^T x : x \in P_0\}$
Mientras $x^k \notin X$ hacer
Definir un hiperplano $H_k : \{x : a_k^T x = b_k\}$ que separe x^k de X
$P_{k+1} = P_k \cap \{x : a_k^T x \leq b_k\}$
$x^{k+1} = \arg\min \{c^T x : x \in P_{k+1}\}$
$k = k + 1$
Fin mientras

Los diversos algoritmos de planos de corte difieren en la manera de seleccionar el nuevo punto x^{k+1}. Este es el aspecto de mayor importancia ya que cuanto mejor sea el corte definido por x^{k+1}, más rápido convergerá el algoritmo.

Entre los métodos de planos de corte, se encuentran los métodos basados en centros. Estos métodos definen x^{k+1} por medio del cálculo del centro de un conjunto convexo y compacto llamado conjunto de localización $\mathcal{L}$.

El conjunto de localización $\mathcal{L}$ está formado por la intersección de los semiespacios generados por la aproximación lineal de la región factible y por una cota superior de la función objetivo

$$\mathcal{L} = \{x : Ax \leq b, \quad c^T x \leq \overline{z}\}$$

Los métodos basados en centros difieren en la manera de definir dicho punto del conjunto de localización. Entre los métodos más conocidos se encuentran: el método del centro de gravedad, el método volumétrico y el ACCPM.

2.2. Fundamentos matemáticos del método ACCPM

El método ACCPM se aplica a los problemas de optimización que pueden ser representados de la siguiente manera:

$$\min\{f(x) : x \in X \subseteq X_0\} \tag{9}$$

donde el conjunto $X \subset \mathbb{R}^n$ es convexo, la función $f : \mathbb{R}^n \to \mathbb{R}$ es convexa y X_0 es un poliedro acotado.

Los métodos de planos de corte se basan en la interacción de dos procedimientos: el oráculo y el programa principal.

El programa principal trabaja sobre la relajación lineal de la región factible del problema de optimización (9), calculando en cada iteración del método un nuevo punto central. Además, controla la convergencia del proceso.

El oráculo toma el punto central como entrada y retorna uno o varios planos de corte al programa principal. Estos planos son de dos tipos: cortes de optimalidad o cortes de factibilidad, dependiendo de la naturaleza del punto.

2.2.1. El oráculo

Dado el punto $\overline{x} \in X_0$, la salida del oráculo está dada así:

- **Cortes de factibilidad:** si $\overline{x} \notin X$ ($\overline{x}$ no es factible), el oráculo retorna el vector $(\gamma_0, \gamma) \in \mathbb{R} \times \mathbb{R}^n$ y el corte de factibilidad:

$$\langle \gamma, x - \overline{x} \rangle + \gamma_0 \leq 0, \ \forall x \in X_t \tag{10}$$

- **Cortes de optimalidad**: el punto es factible ($\overline{x} \in X$); el oráculo retorna $f(\overline{x})$ y un subgradiente $\gamma \in \partial f(\overline{x})$, que definen la desigualdad conocida como corte de optimalidad:

$$f(x) \geq f(\overline{x}) + \langle \gamma, x - \overline{x} \rangle, \ \forall x \in X \tag{11}$$

2.2.2. Conjunto de localización

Sea $(x^1, \ldots, x^K)$ una sucesión de puntos centrales. El conjunto de índices K puede ser expresado como la unión de dos conjuntos disyuntos I_K y J_K donde

$$I_k = \{k : x^k \text{ es no factible (corte de factibilidad)}\}$$
$$J_k = \{k : x^k \text{ es factible (corte de optimalidad)}\}$$

Si $J_K \neq \varnothing$ se define la cota superior de la solución del problema (9) como $\overline{z}_K = \min\{f(x^k) \mid k \in J_K\}$. Tomando la unión de todos los cortes y desigualdades obtenidos anteriormente, se define un subconjunto del epígrafo de f. Este conjunto contiene la solución óptima y se conoce como el *conjunto de localización* y se

denota $\mathcal{L}_K \subseteq \mathbb{R}^{n+1}$. El conjunto de localización está constituido por las siguientes desigualdades:

$$z \geq f(x^k) + \langle \gamma^k, x - x^k \rangle, \quad k \in J_K \tag{12}$$

$$0 \geq \langle \gamma^k, x - x^k \rangle + \gamma_0^k, \quad \forall k \in I_K \tag{13}$$

$$\overline{z}_K \geq z \tag{14}$$

$$b \geq \langle B, x \rangle \tag{15}$$

El primer conjunto de restricciones corresponde a los cortes de optimalidad (12), el segundo conjunto recibe el nombre de cortes de factibilidad, el tercer conjunto de restricciones (14) define la cota superior del problema de optimización, el último conjunto de restricciones son fijas. Usualmente se utilizan restricciones de caja para las variables de decisión con el objetivo de definir esta última desigualdad (15).

Por otro lado, es posible asociar con (12), (13), (14) y (15) las variables duales $\alpha_{jk} \geq 0$, $\mu_k \geq 0$, $\nu \geq 0$ y $\rho \in \mathbb{R}$ que satisfacen la desigualdad

$$z \geq \sum_{k \in J_K} \alpha_k \big(f(x^k) - \langle \gamma^k, x^k \rangle \big) + \sum_{k \in I_K} \mu_k \big(\gamma_0^k - \langle \gamma^k, x^k \rangle \big) + \langle b, \rho \rangle \tag{16}$$

para todo z tal que $(z, x) \in \mathcal{L}_K$. La expresión del lado derecho en (16) es una cota inferior del problema (9). Dicha cota se notará $\underline{z}_K$. Dadas las cotas superior e inferior es posible definir una brecha o salto de dualidad: $d_{g,k} = \overline{z}_K - \underline{z}_K$. En la implementación del método es usual trabajar con la brecha de dualidad *relativa:*

$$d_{g,k} = \frac{\overline{z}_K - \underline{z}_K}{\max\{1, \overline{z}_K\}}$$

Dicho valor es muy importante debido a que con él se construye el criterio de parada. El método se detiene cuando $d_{g,k} \leq \epsilon$.

2.2.3. Método genérico de planos de corte

A continuación se presentan los pasos básicos de los métodos de planos de corte:

Método genérico de planos de corte
1. Prueba de terminación del método
2. Escoger un punto central $(\overline{x}, \overline{z}) \in \mathcal{L}_K$
3. Calcular una cota inferior para $z \in \mathcal{L}_K$
4. Llamar al oráculo para $\overline{x}$. El oráculo retorna
(a) Cortes de factibilidad (si $\overline{x}$ es no factible)
(b) Cortes de optimalidad (si $\overline{x}$ es factible)
5. Calcular la cota superior para $z \in \mathcal{L}_K$
6. Agregar el nuevo corte al conjunto de localización $\mathcal{L}_K$

Los diversos métodos de planos de corte difieren en la forma de escoger el punto central en $\mathcal{L}_K$. El método ACCPM encuentra el centro analítico del conjunto

de localización $\mathcal{L}_K$. En la biblioteca desarrollada por Petón & Vial (2001) se encuentran implementados los pasos 2, 3 y 5 en `Visual C++`. Los pasos 1 y 4 debe implementarlos el usuario. Dicha implementación depende del problema a resolver.

2.2.4. Cálculo del centro analítico

Como se mencionó anteriormente, ACCPM calcula el centro analítico del conjunto de localización. De forma compacta el conjunto de localización se escribe de la siguiente manera:

$$\mathcal{L}_K = \{\widetilde{x} : A^T\widetilde{x} \leq c\}$$

El centro analítico del poliedro acotado $\mathcal{L}_K$ es la única solución (en caso de existir) del siguiente problema de optimización[2]:

$$\operatorname{argmin}\left\{ -\sum_{i=1}^{K} \log(s_i) : s = c - A^T\widetilde{x} \right\}$$

En la ecuación anterior, se penalizan los puntos cercanos a la frontera, es decir, las variables de holgura (s_i) que tiendan a cero.

Para el cálculo del centro analítico, los métodos se basan en algoritmos de punto interior para programación no lineal, tales como método primal, método dual, método primal-dual y método primal proyectivo (Vial 1998). La implementación usada del método ACCPM obtiene el centro analítico por medio del método primal proyectivo desarrollado en Du Merle (1995).

2.2.5. Restricciones de caja

El conjunto de localización es acotado si X_0 es acotado. En la mayoría de aplicaciones de optimización es posible asumir que cada una de las variables de decisión se encuentra restringida por unos valores máximos y mínimos (restricciones de caja), es decir,

$$x_{\text{mín}} \leq x_i \leq x_{\text{máx}}, \quad i = 1, \ldots, n$$

La implementación de ACCPM supone la existencia de restricciones de caja.

3. Cálculo de los estimadores de regresión cuantílica por medio del método ACCPM

En esta sección se presenta la forma de aplicar el método ACCPM en el cálculo de los estimadores de regresión cuantílica. Además, se presentan resultados nu-

[2]Para que sea válido el cálculo del centro analítico se supone que $\{\widetilde{x} : A^T\widetilde{x} \leq c\}$ es acotado y tiene interior no vacío.

méricos y comparaciones con el paquete R (R Development Core Team 2006) y GAMS[3] (López 2006*b*).

3.1. Subgradiente y oráculo para el problema de regresión cuantílica

El modelo de optimización para regresión cuantílica se escribe de la forma

$$\widehat{\beta}_\tau = \text{argmin}_{\beta\in\mathbb{R}^n} f(\beta) = \sum_{i=1}^{m} \rho_\tau \left(y_i - \beta^T x^i\right) \tag{17}$$

donde $y_i \in \mathbb{R}$, $x^i \in \mathbb{R}^n$, $i = 1, \ldots, m$. Es decir, m es el número de datos y n el número de variables explicativas, $\rho_\tau(u)$ es la función de chequeo con $\tau \in (0, 1)$.

Para el anterior problema un subgradiente (Mora 2005) está dado por

$$\gamma = \gamma(f, \beta) = -\tau \sum_{\substack{i=1 \\ r_i>0}}^{m} x^i - (\tau - 1) \sum_{\substack{i=1 \\ r_i<0}}^{m} x^i$$

donde $r_i = y_i - \beta^T x^i$.

El subgradiente descrito anteriormente se puede escribir en forma desagregada como

$$\gamma = \sum_{i=1}^{m} \gamma_i$$

donde

$$\gamma_i = \begin{cases} -\tau x^i, & \text{si} \quad y_i - \beta^T x^i > 0; \\ \mathbf{0}, & \text{si} \quad y_i - \beta^T x^i = 0; \\ -(\tau - 1)\, x^i, & \text{si} \quad y_i - \beta^T x^i < 0. \end{cases}$$

Para efectos de programación se da el valor vectorial $\mathbf{0}$ al subgradiente γ_i cuando $r_i \in (-\epsilon, \epsilon)$, con ϵ positivo y pequeño, es decir, cuando el residuo es casi 0. En este caso γ_i queda reformulado de la siguiente manera:

$$\gamma_i = \begin{cases} -\tau x^i, & \text{si} \quad y_i - \beta^T x^i \geq \epsilon_s; \\ \mathbf{0}, & \text{si} \quad y_i - \beta^T x^i \in (-\epsilon_s, \epsilon_s); \\ -(\tau - 1)\, x^i, & \text{si} \quad y_i - \beta^T x^i \leq \epsilon_s. \end{cases}$$

Por otro lado, como el problema de regresión cuantílica es un problema de optimización sin restricciones, solo se generan cortes de optimalidad en el método ACCPM. Dichos cortes se expresan de la siguiente forma:

$$f(\beta) \geq f(\beta^{k+1}) + \left\langle \gamma^{k+1}, \beta - \beta^{k+1} \right\rangle$$

[3]GAMS (General Algebraic Modeling System) es un lenguaje de programación que tiene por objetivo encontrar solución a diversos problemas de optimización a pequeña y gran escala. Es posible obtener más información, manuales, ayuda y una versión demo en la página `www.gams.com`

donde β^{k+1} y γ^{k+1} son, respectivamente, el centro analítico y el subgradiente generados en la $k-$ésima iteración.

A continuación se presenta la descripción del algoritmo del oráculo para el problema de regresión cuantílica.

Oráculo para el problema de regresión cuantílica

Datos del problema: x^i, y_i $i = 1, \ldots, m$. ε, τ $\widehat{\beta}$
  $\gamma := \overrightarrow{0}$ **(Inicialización del subgradiente)**
  $f = f(\widehat{\beta}) = 0$ **(Inicialización de la función)**
para $i = 1, \ldots, m$
  $r_i = y_i - \widehat{\beta}^T x^i$
  si $r_i \geq \varepsilon_s$
    $\gamma = \gamma - \tau x^i$
    $f = f + \tau r_i$
  fin si
  si $r_i \leq -\varepsilon_s$
    $\gamma = \gamma - (\tau - 1)x^i$
    $f = f + (\tau - 1)r_i$
  fin si
fin para
corte de optimalidad: $f(\beta) \geq f + \gamma^T(\beta - \widehat{\beta})$

En el caso del método ACCPM, $\widehat{\beta}$ se obtiene por medio del cálculo del centro analítico del conjunto de localización. Otro factor de gran importancia es el valor de ε_s. Dicho valor se llamará épsilon del subgradiente.

Otros aspectos importantes del método ACCPM son la brecha de dualidad y las restricciones de caja. La brecha de dualidad utilizada es

$$duality_{gap} = \frac{|\overline{\beta}_k - \underline{\beta}_k|}{\max\{1, |\overline{\beta}_k|\}}$$

donde $\overline{\beta}_k$ y $\underline{\beta}_k$ son las cotas superior e inferior del valor óptimo obtenidas en la $k-$ésima iteración. Las restricciones de caja utilizadas son de la forma

$$-c \leq \beta_i \leq c$$

donde $i = 1, \ldots, n$ y $c > 0$.

3.2. Resultados numéricos y comparaciones

Los resultados numéricos presentados a continuación corresponden al tiempo de ejecución de los paquetes utilizados para hallar los estimadores del problema de regresión cuantílica (ACCPM, R y GAMS[4]). El equipo utilizado para el desarrollo de las pruebas fue un computador con sistema operacional Windows XP, procesador Pentium 4 con 2.4 GHz y 512 Mb de RAM.

[4]El problema resuelto por GAMS fue el problema dual para regresión cuantílica debido a que es de menor tamaño que el problema primal.

El *solver* de optimización utilizado por GAMS es el BDMLP 1.3.

Los tiempos dados son aproximados e incluyen el tiempo de algunas tareas propias del sistema operativo. No se hacen comparaciones de requerimientos de memoria debido a que no se tiene esta información para GAMS ni para R. El criterio de parada utilizado en el método ACCPM es la obtención de una brecha de dualidad menor que un valor dado (θ):

$$d_{g,k} \leq \theta$$

El valor de τ es 0.8. No es necesario presentar resultados con otros valores de τ debido a que cambios en dicho valor no generan cambios en el tiempo de ejecución de los algoritmos.

Todos los archivos de prueba utilizados fueron generados por medio de números aleatorios. Los valores de n se encuentran entre 5 y 20. Se supone que la matriz de datos no tiene variables redundantes y es de naturaleza densa. En todas las pruebas se tomó el épsilon del subgradiente como $\varepsilon_s = 10^{-5}$.

La tabla 1 muestra la solución obtenida con R, GAMS y ACCPM para una base de datos con $m = 10000$ y $n = 13$. Para el método ACCPM se hicieron dos pruebas, con $\theta = 10^{-3}$ y $\theta = 10^{-6}$. Además, para las restricciones de caja $c = 1000$.

Tabla 1: Comparación de resultados para $n = 13$ y $m = 10000$.

estimador	ACCPM ($\theta = 10^{-3}$)	ACCPM ($\theta = 10^{-6}$)	R	GAMS
β_1	9.001168	9.001113	9.00034857	9.000
β_2	2.000805	2.000924	1.99998761	2.000
β_3	3.000632	3.000590	2.99983450	3.000
β_4	8.000662	8.000673	7.99980355	8.000
β_5	2.000971	2.000980	2.00017629	2.000
β_6	2.000971	2.001014	2.00006537	2.000
β_7	14.000552	14.000556	13.99974277	14.000
β_8	2.001135	2.000774	2.00023659	2.000
β_9	7.000982	7.000984	7.00010026	7.000
β_{10}	6.000836	6.000583	5.99967501	6.000
β_{11}	8.000817	8.001068	8.00030165	8.000
β_{12}	1.000432	1.000470	0.99968552	0.999
β_{13}	4.001463	4.001559	4.00046326	4.000

Es importante notar que con tres cifras decimales, la solución obtenida con los tres paquetes es la misma.

La tabla 2 muestra las diferencias de cálculo del método ACCPM cambiando el valor de θ. Los diferentes valores son 10^{-3}, 10^{-4} , 10^{-5} y 10^{-6} con $m = 25000$ y $n = 10$ y las restricciones de caja : $-1000 \leq \beta_i \leq 1000$, con $i = 1, \ldots, 15$.

Para una aproximación de 10^{-6} es necesario generar 57 cortes más que en el caso de 10^{-3}. Es decir un 62% más de cortes. Además, el tiempo de ejecución con $\theta = 10^{-6}$ fue 2.3 segundos mayor que con $\theta = 10^{-3}$. Por lo tanto, utilizó el 61% más de tiempo.

Tabla 2: Tiempos y cortes dependiendo del valor de θ.

θ	Tiempo (s)	Cortes
10^{-3}	3.8	91
10^{-4}	4.7	111
10^{-5}	5.4	131
10^{-6}	6.1	148

La tabla 3 muestra la solución de un problema con $m = 15000$ y $n = 10$, variando los valores de las restricciones de caja: $-c \leq \beta_i \leq c$, con $i = 1, \ldots, 10$. Se tomó $\theta = 10^{-4}$.

Tabla 3: Tiempos y cortes para varios valores de c (restricciones de caja).

c	Tiempo (s)	Cortes
100	2.5	72
200	2.7	77
500	2.6	74
1000	2.7	76
5000	2.5	74
10000	2.5	73
20000	2.6	74
100000	2.6	74

El tamaño de la caja no tiene mayor influencia en el tiempo de ejecución y número de iteraciones (cortes) del método ACCPM.

La tabla 4 muestra la solución del método ACCPM con $n = 12$, $m = 18000$ y $\theta = 10^{-3}$ para cuatro archivos distintos. El primer archivo es generado por números aleatorios en el intervalo $(0, 1)$, el segundo archivo es generado por números aleatorios en $(0, 100)$, el tercer archivo con números aleatorios en $(100, 1000)$ y el cuarto archivo con números aleatorios en $(1000, 100000)$. El objetivo de realizar dichas comparaciones es revisar las diferencias de ejecución del método para problemas de igual tamaño y datos diferentes.

Tabla 4: Solución de problemas con $n = 12$, $m = 18000$ para datos distintos.

Archivo	Valores entre	Tiempo (s)	Cortes
Archivo 1	$(0, 1)$	2.8	73
Archivo 2	$(0, 100)$	3.4	73
Archivo 3	$(100, 1000)$	3.3	69
Archivo 4	$(1000, 100000)$	3.4	70

Según la tabla 4, no existe mucha diferencia entre el número de iteraciones y el tiempo de ejecución para problemas con datos distintos y el mismo tamaño.

Las siguientes tablas (5, 6 y 7) presentan el tiempo de ejecución de cada uno de los paquetes, variando el valor de n y el valor de m. El símbolo **X** indica que no se dispone de ese valor de tiempo porque el problema resultó demasiado grande y no pudo ser resuelto por el paquete indicado.

La tabla 5 muestra los resultados para varios valores de $m = 100$, 300, 1000, 5000, 10000, 30000, 50000, 80000, 100000, 200000 300000 y 400000 con $n = 5$. Se toma θ con un valor de 10^{-3} y $c = 1000$.

Tabla 5: Comparación de tiempos ACCPM, GAMS y R para $n = 5$.

$n = 5$	ACCPM		R	GAMS
m	Cortes	Tiempo (s)	Tiempo (s)	Tiempo (s)
100	29	0.8	0.0	0.0
300	31	0.9	0.0	1.0
1000	30	0.9	0.0	1.0
5000	32	1.0	0.0	14.1
10000	28	1.0	0.8	50.9
30000	32	1.2	3.1	167.2
50000	30	1.3	8.1	278.5
80000	31	1.4	16.9	438.5
100000	29	1.6	26.6	542.6
200000	31	2.2	105.5	**X**
300000	29	2.6	237.9	**X**
400000	30	3.1	492.5	**X**

Análogamente, la tabla 6 muestra los resultados para los mismos valores de m y $n = 10$, $\theta = 10^{-3}$ y $c = 1000$.

Tabla 6: Comparación de tiempos ACCPM, GAMS y R para $n = 10$.

$n = 10$	ACCPM		R	GAMS
m	Cortes	Tiempo (s)	Tiempo (s)	Tiempo (s)
100	52	1.5	0.0	0.0
300	55	1.6	0.0	1.0
1000	59	1.7	0.0	1.0
5000	59	1.9	0.5	20.3
10000	61	2.2	1.0	62.8
30000	62	2.6	3.7	204.1
50000	61	2.8	8.3	323.1
80000	62	3.5	18.5	505.1
100000	58	3.7	28.5	625.2
200000	61	5.5	110.1	**X**
300000	57	6.7	250.9	**X**
400000	63	9.1	**X**	**X**

La tabla 7 muestra los resultados para los mismos valores de m de las tablas anteriores y $n = 20$, $\theta = 10^{-3}$, $c = 1000$.

De las tablas 5, 6 y 7 es posible notar lo siguiente:

Para $m \leq 10000$, el paquete R es muy eficiente debido a que en todas las pruebas realizadas encuentra los estimadores en menos de 2 segundos y para $m \geq 30000$ el tiempo del método ACCPM es menor que el tiempo de R y el de la solución obtenida por GAMS para el problema dual de regresión cuantílica. Además, el tiempo de ACCPM aumenta de forma más o menos lineal. Con los otros dos paquetes el tiempo de ejecución crece de forma más acelerada y en algunos casos no es posible encontrar la solución. Por ejemplo, para $n = 5$ los tiempos con

Tabla 7: Comparación de tiempos ACCPM, GAMS y R para $n = 20$.

$n = 20$	ACCPM		R	GAMS
m	Cortes	Tiempo (s)	Tiempo (s)	Tiempo (s)
100	106	2.7	0.0	0.0
300	116	3.1	0.0	1.0
1000	120	3.3	0.0	1.0
5000	120	3.6	0.9	32.9
10000	121	4.3	1.3	80.5
30000	127	5.9	4.9	243.2
50000	121	7.3	10.1	405.2
80000	119	8.8	22.9	629.1
100000	115	9.9	33.5	778.1
200000	117	16.4	130.2	**X**
300000	114	21.3	**X**	**X**
400000	118	27.9	**X**	**X**

ACCPM se encuentran entre 0.8 y 3.1 segundos (0.8 segundos para $m = 100$ y 3.1 segundos para $m = 400000$). Los tiempos de R se encuentran entre 0 y 492 segundos. Gams entre 0 y 542 segundos. Este último no logra encontrar la solución cuando $m \geq 200000$.

De forma análoga, para $n = 20$, los tiempos de ACCPM se encuentran entre 2.7 y 27.9 segundos. Además, con el método ACCPM en todos los casos fue posible encontrar la solución. Con R los tiempos de ejecución se encuentran entre 0 y 130 segundos. En este caso no fue posible hallar la solución cuando $m \geq 300000$. Para GAMS los tiempos oscilan entre 0 y 778 segundos. Además, no fue posible encontrar la solución cuando $m \geq 200000$.

Para un valor fijo de n y variando el valor de m, el número de cortes de optimalidad (iteraciones) no varía de forma significativa. Por ejemplo, cuando $n = 10$, el mínimo número de cortes generados es 52 con $m = 100$ y la mayor cantidad es 63 con $m = 400000$.

Con $n = 20$, el número de cortes (iteraciones) varía entre 106 y 127; se nota que para este caso hay menor cantidad de cortes con $m = 400000$ que cuando $m = 30000$, $m = 50000$, $m = 1000$, $m = 5000$. Es decir, el número de iteraciones del método depende exclusivamente del número de variables explicativas y no de la cantidad de datos. La diferencia del tiempo de ejecución depende del cálculo del subgradiente debido a que en problemas de mayor tamaño es necesario hacer más operaciones para su obtención.

Para problemas con $m \leq 1000$, la solución del problema dual por medio de GAMS se encuentra más rápido que con ACCPM. Para $m \geq 1000$, el método ACCPM es más rápido.

3.3. Dificultades del método ACCPM

El método ACCPM tiene restricciones para su ejecución y depende del valor de θ. En la tabla 8 se presentan diferentes tamaños máximos para la matriz de datos. Varias de las dificultades se generan por la capacidad de cálculo del computador

y del método. En algunos casos fue posible resolver problemas de mayor tamaño que el indicado en la tabla 8 con valores de restricciones de caja de la forma

$$d_i \leq \beta_i \leq g_i$$

con los valores de d_i y g_i cercanos a los estimadores de los parámetros de regresión cuantílica y además no simétricos, es decir $g_i \neq -d_i$.

Tabla 8: Restricciones del método ACCPM.

θ	$n \leq$	$m \leq$
10^{-6}	26	1000
10^{-6}	20	400000
10^{-6}	18	500000
10^{-4}	29	1000
10^{-4}	23	400000
10^{-4}	20	500000
10^{-3}	33	1000
10^{-3}	30	200000
10^{-2}	38	200000

4. Conclusiones

La implementación del método de punto interior primal-dual de programación lineal para la solución del problema de regresión cuantílica desarrollada en el software R por medio del paquete `quantreg` es la más eficiente cuando el número de datos es menor que 10000. Por otro lado, el método ACCPM resultó ser el más eficiente cuando el número de datos es mayor que 30000. Además, es bastante estable tanto en el tiempo de ejecución como en el número de iteraciones generadas. Es decir, es recomendable usar el método ACCPM cuando se tiene un número grande de datos.

Según los resultados, para $m \geq 1000$ el paquete de optimización GAMS resulta ser el menos eficiente. Esto se debe en parte a las restricciones de ejecución con respecto al tamaño del problema de programación lineal y al método utilizado para encontrar la solución (simplex). En el caso de $m < 1000$, GAMS resulta ser el más eficiente.

El número de cortes de optimalidad (iteraciones) generados en el método ACCPM depende del número de variables explicativas. En este caso, no es un factor influyente el número de datos, ni la naturaleza de los mismos.

El tamaño de las restricciones de caja no influye en el cálculo de los estimadores. La solución generada para el problema de regresión cuantílica por medio de ACCPM no presenta cambios significativos con respecto a cambios sobre el valor de la cota para la brecha de dualidad (θ) y del épsilon del subgradiente (ϵ_s).

Agradecimientos

El presente trabajo se deriva de la tesis de maestría en Matemática Aplicada del primer autor (López 2006*a*).

Agradecemos al profesor Edilberto Ruiz, del Departamento de Estadística de la Universidad Nacional de Colombia, por su asesoría y lectura del documento en el área de regresión cuantílica y al estadístico Rafael López por su apoyo e indicaciones en el manejo del paquete estadístico R. También, las sugerencias y comentarios hechos por los evaluadores de este artículo.

Recibido: diciembre de 2006
Aceptado: febrero de 2007

Referencias

Du Merle, O. (1995), Points intérieurs et plans coupants: mise en uvre et développement d'une méthode pour l'optimisation convexe et la programmation linéaire structurée de grand taille, Tesis de doctorado, Universidad de Ginebra.

Goffin, J., Haurie, A. & Vial, J. (1992), 'Decomposition and Nondiferentiable Optimization with the Projective Algorithm', *Management Science* (2), 284–302.

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

Koenker, R. & Bassett, G. W. (1978), 'Regression Quantiles', *Econometrica* **46**, 33–50.

López, H. (2006*a*), Cálculo de la regresión cuantílica por medio del método ACCPM, Tesis para optar al título de Maestría en Matemática Aplicada, Universidad Nacional de Colombia, Departamento de Matemáticas, Bogotá.

López, H. (2006*b*), 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.

Mora, H. (2005), 'Métodos numéricos para la estimación de parámetros en regresión cuantílica', *Revista Colombiana de Estadística* **28**(2), 221–231.

Petón, O. & Vial, J. (2001), A tutorial of ACCPM. Version 2.01, Technical report, HEC/logilab. Universidad de Ginebra.

R Development Core Team (2006), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Vial, J. P. (1998), 'Analytic center of polytope', Universidad de Ginebra. Manuscrito.