Introducción a la Regresión Cuantil. Estimación y extensión a modelos no paramétricos.

Ana Martín Escura

Trabajo de fin de grado en Matemáticas
Universidad de Zaragoza

Director del trabajo: Dr. José Tomás Alcalá Nalvaiz
Febrero de 2019

Prólogo

La regresión cuantil fue introducida por Roger Koenker y Gib Basset (1978) buscando extender las ideas de estimación de función cuantil condicional. Estos modelos constan de una distribución condicional de la variable respuesta expresada en función de las covariables observadas.

Los métodos de regresión cuantil son competitivos con el método tradicional de mínimos cuadrados en lo que se refiere al esfuerzo computacional gracias al descubrimiento del método simplex y al desarrollo en la programación lineal. La localización de los cuantiles asegura un tipo de robustez carente en muchos procedimientos estadísticos habituales, como, por ejemplo, los basados en minimizar una suma de residuos al cuadrado.

La regresión cuantil está llegando a ser cada vez más útil en áreas como la Econometría, Finanzas, Biomedicina, búsqueda de patrones y en Estudios Ambientales.

Abstract

Quantile regression was introduced by Roger Koenker and Gib Basset in 1978. This regression estimates the conditional median or other quantiles of the response variable Y in contrast to least squares that estimates the conditional mean of the response variable given some values of the predictor variables X.

Quantile regression is becoming increasingly useful in areas such as econometrics, finance, biomedicine, data mining and environmental studies.

In chapter 1, certain basic definitions for quantile regression are introduced such as the τ-th population quantile, denoted by c_τ, the τ-th sample quantile, denoted by $\hat{c}_\tau$, or the quantile function.

The quantile loss function is also defined as $\rho_\tau(u) = u(\tau - \mathbb{1}(u < 0))$ such that it is a piecewise linear function where $\mathbb{1}$ represents the indicator function. Then, the τ-th sample quantile could be saw as the solution of a linear programming problem. In addition, it is shown, for an absolutely continuous random variable, that the estimator $\hat{c}_\tau$ converges in distribution to a normal distribution with zero mean and variance $\frac{\tau(1-\tau)}{f(c_\tau)}$. Then less density of data there is around the population quantile c_τ, there will be more doubt of the asymptotic distribution of the sample quantile $\hat{c}_\tau$.

When there are p explanatory variables we want to model the dependence of the quantiles of the conditional distribution of the response variable Y given $\mathbf{X} \in \mathbb{R}^p$. The basic model of quantile regression is $c_\tau(Y|\mathbf{X}) = \mathbf{x}_i^t \mathbf{b}(\tau)$, where you can estimate $\mathbf{b}(\tau)$ as a solution to a minimization problem and now it is satisfied that the τ-th quantile of the residuals is zero. Similarly to the previous case without covariates, it is shown that, under certain specific conditions, the estimator $\hat{\mathbf{b}}(\tau)$ converges to a normal distribution with $\mathbf{b}(\tau)$ mean.

In this chapter, also different estimation methods are compared, such as the interior point method based on the Karmarkar algorithm or the Barrodale and Roberts algorithm based on simplex with the well-known method of the least squares.

In chapter 2, an introduction to the nonparametric quantile regression is made. In this chapter we focus on local linear smoothing techniques. It is shown the local linear estimator depends on kernel function K and the bandwidth parameter h_τ. In addition, the asymptotic expression for the mean square error (MSE) of the conditional τ-th quantile estimator is given. It is shown how practical calculation selection of the bandwidth parameter could be made by plug-in and cross-validation techniques.

Finally, in chapter 3 and the last chapter, all the previous models are applied in two data sets. Several functions of R software package `quantreg` is used to obtain quantile regression, summary or anova tables.

The first data set, was made up by our own data, is used to model the dependency of the box office in two consecutive weekends in a Spanish exhibition network. Parametric lineal models allow us a good

representation of this relationship.

In the second data set, related to the relative mineral density in spinal bone for teenagers, nonparametric modelling and automatic selection procedure are implemented. Moreover, in this case a cross-over phenomenon is described and it is resolved by applying an isotonic regression algorithm.

In the appendix all code and script used to develop the applied data analysis are commented.

Índice general

Capítulo 1

Estimación e inferencia en Regresión Cuantil

En esta primera sección se realiza una introducción de los conceptos básicos de la regresión cuantil, como el término cuantil y la función de pérdida de un cuantil. Por otro lado, se ve cómo el cuantil es solución de un problema de optimización y se da la distribución asintótica del cuantil muestral.

1.1. Estadísticos ordenados. Primeras propiedades

Sea una variable aleatoria X, se entenderá, salvo que se diga lo contrario, que $F(x) = P(X \leq x)$ es su función de distribución y $f(x)$ su función de densidad. Si X es discreta o absolutamente continua se tiene respectivamente la siguiente expresión para la función de distribución:

$$F(x) = \sum_{x_i \leq x} P(X = x_i) \quad \text{o} \quad F(x) = \int_{-\infty}^{x} f(u)du.$$

A continuación, se ve el equivalente a la función de distribución cuando se dispone de una muestra aleatoria $x_1, ..., x_n$. Para ello, se introduce el concepto de muestra ordenada.

Definición. Dada una muestra aleatoria $x_1, ..., x_n$ se define la **muestra ordenada** como $x_{(1)}, ..., x_{(n)}$ donde $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$.

Definición. Dada una muestra aleatoria $x_1, ..., x_n$ se define la **función de distribución empírica** como sigue:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, x]}(x_i) = \begin{cases} 0 & \text{si} \quad x < x_{(1)} \\ \frac{k}{n} & \text{si} \quad x_{(k)} \leq x < x_{(k+1)} \\ 1 & \text{si} \quad x \geq x_{(n)}, \end{cases} \tag{1.1}$$

donde $k = 1, ..., n-1$ y se cumple que $F_n(x) \xrightarrow{c.s.} F(x)$ (véase Corolario 4.1 del capítulo 3 en [1]).

Definición. Dada una variable aleatoria X para cada $0 < \tau < 1$ se define el **cuantil poblacional de orden** τ, y se denota c_τ, al valor tal que:

$$P(X \leq c_\tau) \geq \tau \quad \text{y} \quad P(X \geq c_\tau) \geq 1 - \tau. \tag{1.2}$$

En el caso de que $\tau = 0{,}5$ se denomina mediana (Q_2), $\tau = 0{,}25$ es el primer cuartil (Q_1) y $\tau = 0{,}75$ se trata del tercer cuartil (Q_3).

Dada una muestra aleatoria $x_1, ..., x_n$, para cada $0 < \tau < 1$ se define el **cuantil muestral de orden** τ, y se denota $\hat{c}_\tau$, al estadístico ordenado $x_{(n\tau)}$ si $n\tau$ es entero, y al estadístico $x_{([n\tau+1])}$ si $n\tau$ no es entero, donde $[x]$ es el menor entero mayor o igual que x.

Si $n\tau$ es entero, se podría tomar cualquier valor entre $x_{(n\tau)}$ y $x_{(n\tau+1)}$ como cuantil muestral de orden τ. Si $\tau = \frac{1}{2}$ y n es par, se podría tomar cualquier valor entre $x_{(\frac{n}{2})}$ y $x_{(\frac{n}{2}+1)}$, en particular se toma el promedio de estos valores. De esta manera, la mediana muestral se define como:

$$\hat{c}_{0,5} = \begin{cases} x_{(\frac{n+1}{2})} & \text{si} \quad \text{n es impar} \\ x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)} & \text{si} \quad \text{n es par.} \end{cases} \tag{1.3}$$

Definición. La **función cuantil de una variable aleatoria** X, si existe, se define como la inversa de su función de distribución $F(x)$. Si F es estrictamente creciente existe una única antiimagen y se tiene que $c_\tau = F^{-1}(\tau)$; en otro caso, se obtiene:

$$F^{-1}(\tau) = inf\{x \in \mathbb{R} \mid F(x) \geq \tau\}, \tag{1.4}$$

con $\tau \in (0,1)$.

Sea la muestra $x_1, ..., x_n$ se define la **función cuantil empírica de la muestra** como la función:

$$\hat{F}_n^{-1}(\tau) = inf\{x \in \mathbb{R} \mid F_n(x) \geq \tau\}, \tag{1.5}$$

para todo $\tau \in (0,1)$. Para $\tau = 0{,}5$ se tiene que $\hat{F}_n^{-1}(0{,}5) = \hat{Q}_2$ es la mediana muestral, para $\tau = 0{,}25$ y $\tau = 0{,}75$ se tienen los cuartiles $\hat{Q}_1$ y $\hat{Q}_3$.

1.2. Estimación

En esta sección se estima con ayuda de la función de pérdida de un cuantil. Para mayor claridad, se realiza en el contexto de muestras aleatorias sin covariables y, posteriormente, se pasa a hacerlo en general con p covariables.

Definición. Se define la **función de pérdida de un cuantil** como una función lineal a trozos:

$$\rho_\tau(u) = u(\tau - \mathbb{1}(u < 0)) = \begin{cases} u\tau & \text{si} \quad u \geq 0 \\ u(\tau - 1) & \text{si} \quad u < 0. \end{cases} \tag{1.6}$$

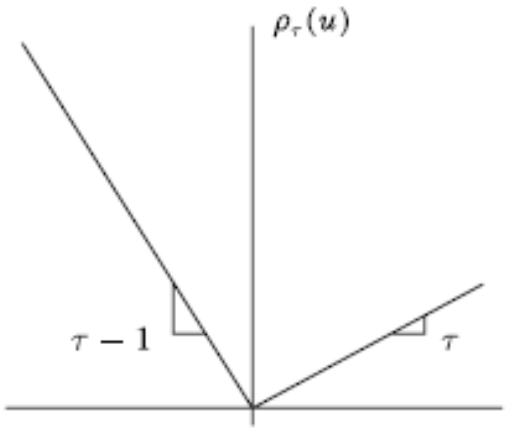

Figura 1.1: Función de pérdida de un cuantil τ (Fuente: Figura 2 de Koenker y Hallock [14]).

Se pueden calcular los cuantiles muestrales como solución de un problema de optimización.

Teorema 1.1. *Sea $\tau \in (0,1)$ se tiene que $\hat{c}_\tau$ es solución del siguiente problema:*

$$\min_{\beta \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(x_i - \beta). \tag{1.7}$$

Demostración. Véase primero para la mediana, es decir, con $\tau = 0{,}5$ donde se cumple que $\tau = 1 - \tau$:

$$\sum_{i=1}^{n} \rho_\tau(x_i - \beta) = \tau \left[\sum_{x_i > \beta} (x_i - \beta) + \sum_{x_i \leq \beta} (\beta - x_i) \right]. \tag{1.8}$$

Ya que se busca el mínimo, a continuación, se deriva (1.8) respecto β y se iguala a 0, obteniéndose que $\hat{\beta}$ es la mediana muestral:

$$\#\{x_i \leq \beta\} - \#\{x_i > \beta\} = 0 \Leftrightarrow \#\{x_i \leq \beta\} = \#\{x_i > \beta\} \Leftrightarrow \hat{\beta} = \hat{Q}_2.$$

En general, si $\tau \neq 0{,}5$, se tiene:

$$\sum_{i=1}^{n} \rho_\tau(x_i - \beta) = \sum_{x_i > \beta} \tau(x_i - \beta) + \sum_{x_i \leq \beta} (\tau - 1)(x_i - \beta).$$

Del mismo modo, se deriva e iguala a 0 obteniendo que $\hat{\beta}$ es el cuantil de orden τ:

$$-\tau \#\{x_i > \beta\} - (\tau - 1)\#\{x_i \leq \beta\} = 0 \Leftrightarrow \tau(\#\{x_i > \beta\} + \#\{x_i \leq \beta\}) = \#\{x_i \leq \beta\} \Leftrightarrow \tau n = \#\{x_i \leq \beta\}.$$

Es decir, β cumple $\frac{\#\{x_i \leq \beta\}}{n} = \tau$. Por lo tanto, $\hat{\beta} = \hat{c}_\tau$. □

Se han visto los cuantiles muestrales como solución de un problema de optimización. Este problema se puede reformular introduciendo variables artificiales $u = (u_1, ..., u_n) \in \mathbb{R}^n$ y $v = (v_1, ..., v_n) \in \mathbb{R}^n$ tales que $u_i = (x_i - \beta)^+$ y $v_i = (x_i - \beta)^-$ para $i = 1, ..., n$, es decir, son la parte positiva y negativa de $(x_i - \beta)$ respectivamente. Entonces, el problema de minimizar (1.7) queda:

$$\min_{(\beta, u, v) \in \mathbb{R} x \mathbb{R}^{2n}_+} \{\tau \mathbb{1}_n^t u + (1 - \tau)\mathbb{1}_n^t v : \mathbb{1}_n \beta + u - v = X\}, \tag{1.9}$$

donde $\mathbb{1}_n$ denota el vector n-dimensional de unos y $X = (x_1, ..., x_n)$. Se minimiza una función lineal en un conjunto poliédrico de restricciones formado por la intersección de hiperplanos de dimensión $2n+1$.

A continuación, se va a ver cuál es la distribución asintótica del estimador que se ha calculado en el modelo (1.7).

Teorema 1.2. *Dada una variable aleatoria absolutamente continua X tal que $f(c_\tau) > 0$ en el entorno del cuantil de orden τ. Entonces, la distribución asintótica del cuantil muestral, $\hat{c}_\tau$ viene dada por:*

$$\sqrt{n}(\hat{c}_\tau - c_\tau) \xrightarrow{D} N\left(0, \frac{\sqrt{\tau(1-\tau)}}{f(c_\tau)}\right). \tag{1.10}$$

Demostración. La variable aleatoria X tiene función densidad continua y positiva en c_τ, pues $f(c_\tau) > 0$, donde c_τ es el cuantil de orden τ con $0 < \tau < 1$. Sea $\hat{c}_\tau$ el cuantil muestral de orden τ, es decir, es $x_{(k)}$ con $k = [n\tau] + 1$. Ahora, se define $T_\tau = \sqrt{n}(\hat{c}_\tau - c_\tau)$ cuya función de densidad es:

$$f_{T_\tau}(t) = \frac{1}{\sqrt{n}} f\left(c_\tau + \frac{t}{\sqrt{n}}\right) =$$

$$\frac{1}{\sqrt{n}} \frac{n!}{(k-1)!(n-k)!} \left[F\left(c_\tau + \frac{t}{\sqrt{n}}\right)\right]^{k-1} \left[1 - F\left(c_\tau + \frac{t}{\sqrt{n}}\right)\right]^{n-k} f\left(c_\tau + \frac{t}{\sqrt{n}}\right), \tag{1.11}$$

adaptada de la expresión de la función de densidad de un estadístico ordenado.

Tomando límites cuando $n \to \infty$ en (1.11):

$$\lim_{n\to\infty} \underbrace{\frac{1}{\sqrt{n}} \frac{n!}{(k-1)!(n-k)!}}_{(1)} \underbrace{\left[F\left(c_\tau + \frac{t}{\sqrt{n}}\right)\right]^{k-1} \left[1 - F\left(c_\tau + \frac{t}{\sqrt{n}}\right)\right]^{n-k} f\left(c_\tau + \frac{t}{\sqrt{n}}\right)}_{(2)}. \tag{1.12}$$

En (1), al aplicar la fórmula de Stirling $n! \simeq n^n e^{-n}\sqrt{2\pi n}$, se tiene que:

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \frac{n^n e^{-n}\sqrt{2\pi n}(n-n\tau)}{(n\tau)^{n\tau} e^{-n\tau}\sqrt{2\pi n\tau}(n-n\tau)^{n-n\tau} e^{-n-+n\tau}\sqrt{2\pi(n-n\tau)}} = \lim_{n\to\infty} \frac{1-\tau}{\tau^{n\tau}(1-\tau)^{n(1-\tau)}\sqrt{2\pi(1-\tau)}}. \tag{1.13}$$

Ahora en (2) desarrollando por Taylor en $x = c_\tau$ y tomando $c_\tau^* \in \left(c_\tau, c_\tau + \frac{t}{\sqrt{n}}\right)$ tal que $f(c_\tau^*) = f(c_\tau) + o(1)$, por ser f continua:

$$F\left(c_\tau + \frac{t}{\sqrt{n}}\right) = F(c_\tau) + \frac{t}{\sqrt{n}} f(c_\tau) + R_n = \tau + \frac{t}{\sqrt{n}} f(c_\tau) + R_n, \tag{1.14}$$

donde $R_n = o\left(\frac{1}{\sqrt{n}}\right)$, y:

$$1 - F\left(c_\tau + \frac{t}{\sqrt{n}}\right) = 1 - \tau - \frac{t}{\sqrt{n}} f(c_\tau) - R_n. \tag{1.15}$$

A continuación, uniendo (1.14) y (1.15) con parte del límite de (1.13) donde se toma logaritmo y exponencial y se utiliza el desarrollo de McLaurin de $\log(1+x)$, con $|x| < 1$, se tiene:

$$\lim_{n\to\infty} \frac{\left[F\left(c_\tau + \frac{t}{\sqrt{n}}\right)\right]^{k-1} \left[1 - F\left(c_\tau + \frac{t}{\sqrt{n}}\right)\right]^{n-k+1}}{\tau^{n\tau}(1-\tau)^{n(1-\tau)}} =$$

$$\lim_{n\to\infty} \frac{\left[\tau + \frac{t}{\sqrt{n}} f(c_\tau) + R_n\right]^{k-1} \left[1 - \tau - \frac{t}{\sqrt{n}} f(c_\tau) - R_n\right]^{n-k+1}}{\tau^{k-1}(1-\tau)^{n-k+1}} =$$

$$\lim_{n\to\infty} \left[1 + \frac{t}{\tau\sqrt{n}} f(c_\tau) + \frac{R_n}{\tau}\right]^{k-1} \left[1 - \frac{t}{(1-\tau)\sqrt{n}} f(c_\tau) - \frac{R_n}{1-\tau}\right]^{n-k+1} =$$

$$exp\left\{\lim_{n\to\infty} (k-1)\log\left[1 + \frac{t f(c_\tau)}{\tau\sqrt{n}} + \frac{R_n}{\tau}\right] + \lim_{n\to\infty} (n-k+1)\log\left[1 - \frac{t f(c_\tau)}{(1-\tau)\sqrt{n}} - \frac{R_n}{1-\tau}\right]\right\} =$$

$$exp\left\{\lim_{n\to\infty} (n\tau)\left[\frac{t}{\tau\sqrt{n}} f(c_\tau) + \frac{R_n}{\tau} - \frac{t^2}{2n\tau^2} f^2(c_\tau) + o\left(\frac{1}{n}\right)\right] +\right.$$

$$\left.\lim_{n\to\infty} (n-n\tau)\left[-\frac{t}{(1-\tau)\sqrt{n}} f(c_\tau) - \frac{R_n}{1-\tau} - \frac{t^2}{2n(1-\tau)^2} f^2(c_\tau) + o\left(\frac{1}{n}\right)\right]\right\} =$$

$$exp\left\{\lim_{n\to\infty}\left[t\sqrt{n} f(c_\tau) + nR_n - \frac{t^2 f^2(c_\tau)}{2\tau} - t\sqrt{n} f(c_\tau) - nR_n - \frac{t^2 f^2(c_\tau)}{2(1-\tau)}\right]\right\} =$$

$$exp\left\{-\frac{t^2}{2} f^2(c_\tau)\left[\frac{1}{\tau} + \frac{1}{1-\tau}\right]\right\} = exp\left\{-\frac{t^2}{2} f^2(c_\tau)\frac{1}{\tau(1-\tau)}\right\}.$$

Llevando los límites anteriores a (1.12) se obtiene:

$$\lim_{n\to\infty} f_{T_\tau}(t) = \frac{1-\tau}{\sqrt{2\pi\tau(1-\tau)}} \cdot exp\left\{-\frac{t^2}{2} f^2(c_\tau)\frac{1}{\tau(1-\tau)}\right\} \cdot \lim_{n\to\infty} \frac{f\left(c_\tau + \frac{t}{\sqrt{n}}\right)}{1-F\left(c_\tau + \frac{t}{\sqrt{n}}\right)} =$$

$$\frac{exp\left\{-\frac{t^2 f^2(c_\tau)}{2\tau(1-\tau)}\right\}}{\sqrt{2\pi\tau(1-\tau)}} \cdot \lim_{n\to\infty} \frac{f\left(c_\tau + \frac{t}{\sqrt{n}}\right)(1-\tau)}{1-\tau-\frac{t}{\sqrt{n}}f(c_\tau)-R_n} = \frac{exp\left\{-\frac{t^2 f^2(c_\tau)}{2\tau(1-\tau)}\right\}}{\sqrt{2\pi\tau(1-\tau)}} \cdot \lim_{n\to\infty} \frac{f\left(c_\tau + \frac{t}{\sqrt{n}}\right)}{1-\frac{t}{\sqrt{n}(1-\tau)}f(c_\tau)-\frac{R_n}{1-\tau}} =$$

$$\frac{f(c_\tau)}{\sqrt{2\pi\tau(1-\tau)}} \cdot exp\left\{-\frac{t^2}{2}\frac{f^2(c_\tau)}{\tau(1-\tau)}\right\}.$$

Como $f_{T_\tau}(t)$ es uniformemente acotada para todo intervalo $(t_1,t_2]$, el límite $\lim_{n\to\infty} F_{T_\tau}(t)$ será la función de distribución de una normal $N\left(0, \frac{\sqrt{\tau(1-\tau)}}{f(c_\tau)}\right)$, por lo que $\hat{c}_\tau \sim N\left(c_\tau, \frac{\sqrt{\tau(1-\tau)}}{\sqrt{n}f(c_\tau)}\right)$.

Luego, $\sqrt{n}(\hat{c}_\tau - c_\tau) \xrightarrow{D} N\left(0, \frac{\sqrt{\tau(1-\tau)}}{f(c_\tau)}\right)$. Como se quería probar. □

Para más detalles de esta demostración véase Teorema 3.1 del capítulo 3 de [1]. Notar que, a menor densidad de datos alrededor de c_τ, habrá mayor incertidumbre de la distribución asintótica de $\hat{c}_\tau$.

El recíproco de la densidad en c_τ se denomina **función 'sparsity'** y viene dada por:

$$s(\tau) = \frac{1}{f(F^{-1}(\tau))}. \tag{1.16}$$

Esta función es la derivada de la función cuantil, es decir, $\frac{\partial F^{-1}(t)}{\partial t} = s(t)$ y representa la pendiente de la tangente a la función cuantil en el punto t.

1.3. Regresión Cuantil

Dada una muestra aleatoria conocida $\{(X_1,Y_1),...,(X_n,Y_n)\}$ tal que $X_i = (X_{1,i},...,X_{p,i}) \in \mathbb{R}^p$ y $Y_i \in \mathbb{R}$. Se define $\mathbf{x_i}^t = (1, X_i^t) \in \mathbb{R}^{p+1}$. Sea ahora Y variable respuesta y $X \in \mathbb{R}^p$ vector de valores formado de p variables regresoras o predictoras de la respuesta Y.

En regresión múltiple, se modela $\mathbb{E}(Y|\mathbf{X}=\mathbf{x_i}) = \mathbf{x}_i^t\beta$ donde $\beta \in \mathbb{R}^{p+1}$ son los coeficientes de regresión.

Ahora se va a modelar la dependencia de los cuantiles de la distribución condicional de la variable Y dada $\mathbf{X}$. Se modela $c_\tau(Y|\mathbf{X}=\mathbf{x_i})$, donde $c_\tau(.|.)$ denota la función cuantil para el cuantil de orden τ de Y condicionada a los valores $\mathbf{X}=\mathbf{x}_i$ de las p variables predictoras. El modelo básico en regresión cuantil es:

$$c_\tau(Y|\mathbf{X}=\mathbf{x_i}) = \mathbf{x}_i^t\mathbf{b}(\tau). \tag{1.17}$$

Se va a ver cómo estimar $\mathbf{b}(\tau)$. Se sigue utilizando la misma función de pérdida de un cuantil (1.6). En la práctica, se va a tener que utilizar una modificación, ya que hasta ahora en (1.7) no dependía de las covariables. De nuevo, se plantea el problema (1.7) en este contexto con las p covariables.

Se tiene que $\hat{\mathbf{b}}(\tau) \in \mathbb{R}^{p+1}$ es solución del siguiente problema:

$$\min_{b\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{x}_i^t\mathbf{b}), \tag{1.18}$$

Es decir, se tiene el modelo lineal:

$$Y_i = \mathbf{x}_i^t \mathbf{b} + \varepsilon_i, \quad i = 1, ..., n, \tag{1.19}$$

donde ε_i es el término del error para $i = 1, ..., n$. Ahora se cumple que el cuantil de orden τ del error es cero, esto es, que $\mathbb{P}(\varepsilon_i \leq 0|X) = \tau$.

En el siguiente resultado se da la distribución asintótica del estimador que se ha calculado en el modelo básico en regresión cuantil (1.17) que extiende el resultado del Teorema 1.2 cuando se tienen p covariables.

Teorema 1.3. *Sean F_i las funciones de distribución condicionales (Y_i condicionada a X_i) que son absolutamente continuas y f_i las respectivas funciones de densidad uniformemente acotadas en los cuantiles condicionados c_τ. Supuesto que existen las respectivas matrices definidas positivas $\mathbf{D}_0$ y $\mathbf{D}_1(\tau)$ que cumplen:*

$$\mathbf{D}_0 = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^t,$$

$$\mathbf{D}_1(\tau) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_i(c_\tau) \mathbf{x}_i \mathbf{x}_i^t$$

y

$$\max_{i=1,...,n} \frac{\| \mathbf{x}_i \|}{\sqrt{n}} \to 0.$$

Se tiene que la distribución asintótica del estadístico muestral $\hat{\mathbf{b}}(\tau)$ es:

$$\sqrt{n}(\hat{\mathbf{b}}(\tau) - \mathbf{b}(\tau)) \xrightarrow{D} \mathbf{N}(\mathbf{0}, \tau(1-\tau)\mathbf{D}_1^{-1}(\tau)\mathbf{D}_0\mathbf{D}_1^{-1}(\tau)), \tag{1.20}$$

donde $\mathbf{x}_i^t = (1, X_i^t)$ y las dimensiones de las matrices $\mathbf{D}_0$ y $\mathbf{D}_1(\tau)$ son $(p+1)x(p+1)$.

En el caso particular de que ε_i sean independientes e idénticamente distribuidas (i.i.d.) para todo $i = 1, ..., n$ se tiene que $\mathbf{D}_1(\tau) = f(c_\tau)\mathbf{D}_0$, y el resultado anterior (1.20) *se reduce a:*

$$\sqrt{n}(\hat{\mathbf{b}}(\tau) - \mathbf{b}(\tau)) \xrightarrow{D} \mathbf{N}\left(\mathbf{0}, \frac{\tau(1-\tau)}{f^2(c_\tau)}\mathbf{D}_0^{-1}\right). \tag{1.21}$$

El resultado de la distribución asintótica del estadístico muestral $\hat{\mathbf{b}}(\tau)$ se da en el Teorema 4.1 del capítulo 4 de Koenker [10], donde se puede consultar la demostración laboriosa, con un número importante de detalles técnicos, pero, en esencia, sigue los pasos de la demostración del Teorema 1.2.

De la misma manera que se ha obtenido (1.9), se puede reformular este problema de regresión cuantil como un problema de programación lineal. Análogamente, se introducen las variables artificiales u y v en $\mathbb{R}^n$ tal que $u_i = (Y_i - \mathbf{x}_i^t\mathbf{b})^+$ y $v_i = (Y_i - \mathbf{x}_i^t\mathbf{b})^-$ para $i = 1, ..., n$, es decir, son la parte positiva y negativa de $(Y_i - \mathbf{x}_i^t\mathbf{b})$. Entonces, el problema de minimizar (1.18) queda:

$$\min_{(b,u,v)\in\mathbb{R}^{p+1}x\mathbb{R}_+^{2n}} \{\tau \mathbb{1}_n^t u + (1-\tau)\mathbb{1}_n^t v | \ \mathbb{X}\mathbf{b} + u - v = Y\}. \tag{1.22}$$

Pero en este modelo básico de regresión cuantil no se puede obtener de forma explícita la solución para estos coeficientes ya que la función de pérdida de un cuantil (1.6) no es diferenciable en el origen. Luego, se necesitan algoritmos numéricos para encontrar solución a este problema. Existen distintos algoritmos para resolver estos problemas de programación lineal anteriores con alguna forma del algoritmo simplex.

Alternativamente, existe un método en programación lineal denominado de punto interior. Este método está basado en el algoritmo de Karmarkar [16] que adapta los problemas reformulando el problema de optimización como un problema de frontera. Consiste en atravesar el interior de la región factible hasta llegar a una frontera. Karmarkar [16] demostró que es posible crear este algoritmo de programación lineal caracterizado por la dificultad polinómica y que, además, es competitivo con el método simplex. Transformando el problema de programación lineal (1.22) al problema dual, se obtiene:

$$\max_{\mathbf{d}} \{Y^t\mathbf{d} : \mathbb{X}^t\mathbf{d} = 0 |\ \mathbf{d} \in [\tau-1, \tau]^n\}. \tag{1.23}$$

Ahora se toma $\mathbf{a} = \mathbf{d} + (1-\tau)\mathbb{1}_n$ y el problema dual anterior queda:

$$\max_{\mathbf{a}} \{Y^t\mathbf{a} |\ \mathbb{X}^t\mathbf{a} = (1-\tau)\mathbb{X}^t\mathbb{1}_n, \mathbf{a} \in [0,1]^n\}. \tag{1.24}$$

La formulación dual del problema de regresión cuantil se ajusta bien a la formulación estándar del método de punto interior para programación lineal con variables acotadas. Añadiendo variables de holgura $\mathbf{s}$ tal que $\mathbf{a} + \mathbf{s} = \mathbb{1}_n$, se define la función barrera como sigue:

$$B(\mathbf{a}, \mathbf{s}, \mu) = Y^t\mathbf{a} + \mu \sum_{i=1}^{n} (\log \mathbf{a}_i + \log \mathbf{s}_i), \tag{1.25}$$

donde $\mu > 0$ real y el problema de programación lineal se formula como:

$$\max_{(\mathbf{s},\mu)} \{B(\mathbf{a}, \mathbf{s}, \mu) | \mathbb{X}^t\mathbf{a} = (1-\tau)\mathbb{X}^t\mathbb{1}_n, \mathbf{a} + \mathbf{s} = \mathbb{1}_n\}. \tag{1.26}$$

Se puede prescindir de la restricción de que $\mathbf{a} \geq 0$ debido a la presencia del término logarítmico en la función barrera.

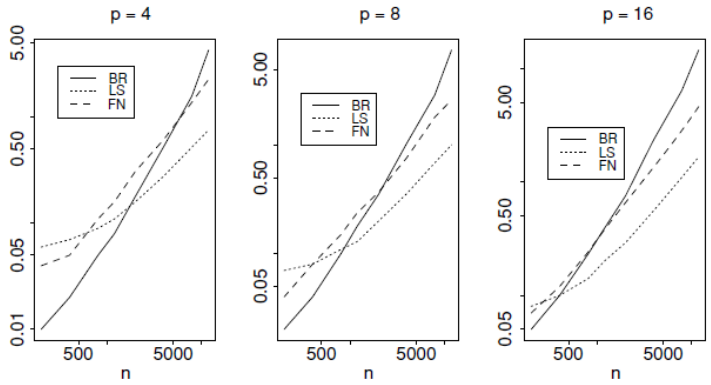

Figura 1.2: Comparación de tiempos de los algoritmos punto interior y exterior en regresión mediana con mínimos cuadrados. El tiempo está en segundos para 5 réplicas de datos Gaussianos i.i.d., la dimensión del modelo es $p+1$, y los valores de n son 200, 400, ..., 12000 (Fuente: Figura 6.3 de Koenker [10]).

Según las conclusiones de un estudio de simulación realizado por Koenker [10], hay tres algoritmos diferentes principales que se pueden elegir según el tamaño de muestra y otras características:

- **BR**: algoritmo de Barrodale y Roberts basado en simplex (línea continua en la Figura 1.2).
- **LS**: algoritmo de mínimos cuadrados basado en la función de S $lm(y \sim x)$ (línea de puntos en la Figura 1.2).
- **FN**: algoritmo de Frisch Newton de punto interior (línea discontinua en la Figura 1.2).

Para problemas de tamaño n modesto, los algoritmos de FM y BR son competitivos con los mínimos cuadrados en términos de velocidad computacional. A medida que el tamaño de la muestra aumenta, tales comparaciones están inevitablemente ligadas al estilo de programación, sobrecarga del sistema y otros factores. Esto se debe a que los mínimos cuadrados cuentan con décadas de refinamiento mientras que los procedimientos de punto interior todavía se encuentran en sus comienzos.

Para problemas de dimensión p pequeños, se tiene que los algoritmos BR y FN funcionan más rápidos que LS. En particular, el mejor método es el de BR con crecimiento aproximadamente cuadrático. Por otro lado, si el tamaño de la muestra es más grande, el algoritmo FN resulta bastante mejor que el algortimo simplex, con un crecimiento lineal. Estos detalles históricos y otros detalles técnicos de los algortimos se pueden consultar en [10].

Todas las ideas anteriores se pueden extender al contexto no lineal. Sea el modelo de regresión:

$$Y_i = q_\tau(X_i, \theta_\tau) + \varepsilon_i, \tag{1.27}$$

donde q_τ es conocida salvo por el parámetro θ_τ y $\{(X_1,Y_1),...,(X_n,Y_n)\}$ es una muestra aleatoria de las variables $(X,Y) \in \mathbb{R}^{p+1}$. Así mismo, el cuantil condicional de orden τ de los errores es nulo. En estas condiciones, se puede considerar el siguiente estimador de θ:

$$\hat{\theta}_\tau = arg \min_{\theta \in \mathbb{R}^q} \sum_{i=1}^{n} \rho_\tau(Y_i - q_\tau(X_i, \theta)). \tag{1.28}$$

Koenker en el capítulo 4 [10] presenta el comportamiento asintótico del estimador (1.28) como una extensión del Teorema 1.3.

Elegir esta función q_τ no siempre es fácil en ausencia de otra información sobre el modelo. Por ello, en el siguiente capítulo se introducirán algunas técnicas para regresión cuantil no paramétrica que son convenientes para conseguir enfoques más flexibles.

Capítulo 2

Regresión cuantil no paramétrica

2.1. Introducción a la regresión no paramétrica

En primer lugar, se va a ver una de las ideas básicas en la que se basan la mayoría de las técnicas de estimación no paramétrica de curvas.

Sea $m(x)$ la función de regresión de Y condicional a un valor x_0 de X, no necesariamente lineal y desconocida. Si se asume que la función $m(x)$ es suficientemente suave, admite para valores x próximos a x_0 un desarrollo o aproximación lineal en torno a dicho punto,

$$m(x) \approx m(x_0) + m'(x_0)(x - x_0) + O((x - x_0)^2) \approx a_0 + a_1(x - x_0).$$

La estimación de $m(x_0)$ (y de $m'(x_0)$) se puede realizar modificando el problema de mínimos cuadrados de regresión lineal a un problema similar introduciendo pesos locales. Sea K una **función kernel o núcleo**, normalmente una función de densidad simétrica de soporte compacto, por ejemplo $[-1,1]$, con $\sigma_K^2 = \int u^2 K(u)du$, y sea $K_h(u) = h^{-1}K(\frac{u}{h})$ con $h > 0$.

Dada una muestra aleatoria simple de $\{(X_1,Y_1),...,(X_n,Y_n)\}$ del vector aleatorio bidimensional $(X,Y) \in \mathbb{R}^2$, la suma de cuadrados local se expresa como:

$$\sum_{i=1}^{n} \{Y_i - a_0 - a_1(X_i - x_0)\}^2 K_h(X_i - x_0). \tag{2.1}$$

El estimador (lineal local) de la función de regresión es $\hat{m}(x_0) = \hat{a}_0$, con $\hat{a}_0$ el valor que minimiza (2.1). La estimación lineal local se puede extender al modelo polinómico local, tal y como se presenta en el libro de Fan y Gijbels [6].

El valor h se denomina **parámetro de suavizado o de ventana** y juega un papel importante en las propiedades asintóticas del estimador lineal local. El Teorema 1 de Fan [4] proporciona la siguiente expresión del error cuadrático medio (MSE) del estimador local lineal $\hat{m}(x_0)$ cuando $n \to \infty$, $h \to 0$ y $nh \to \infty$:

$$E((\hat{m}(x_0) - m(x_0))^2) = \frac{1}{4}h^4\sigma_K^2 m''^2(x_0) + \frac{1}{nh}\frac{\sigma^2(x_0)}{f_X(x_0)}R(K) + o\left(h^4 + \frac{1}{nh}\right) \tag{2.2}$$

donde $R(K) = \int K^2(u)du$, $\sigma^2(x_0) = Var(Y|X = x_0)$ y $f_X(x_0) > 0$ es la función de densidad de X en x_0.

Se puede seleccionar el parámetro h de forma que minimice el término principal de (2.2), y su expresión es:

$$h_{AMSE} = \left(\frac{\sigma^2(x_0)R(K)}{\sigma_K^2 m''^2(x_0)f_X(x_0)}\right)^{1/5} n^{-1/5}. \tag{2.3}$$

La selección de h_{AMSE} no es fácil de llevar a la práctica ya que, como se ve en (2.3), depende de elementos desconocidos. En [6] se pueden consultar diversas estrategias de selección del parámetro h_{AMSE} basadas en la estimación piloto de las cantidades desconocidas, mediante nuevos estimadores de tipo núcleo basados en otros parámetros de suavizado, de forma que se obtiene un estimador consistente de este parámetro óptimo de suavizado (método plug-in).

Alternativamente, es posible estimar este parámetro por el método de validación cruzada. La idea es predecir el rendimiento de un modelo en datos no disponibles mediante el cálculo numérico en lugar del análisis teórico.

2.2. Regresión cuantil no paramétrica

2.2.1. Estimador lineal local cuantil

La teoría desarrollada en el capítulo anterior se puede extender a regresión cuantil no paramétrica. Dados $X, Y \in \mathbb{R}$, se considera el siguiente modelo:

$$Y = q_\tau(X) + \varepsilon,$$

donde el cuantil de orden τ condicional del error dada la covariable es cero. Se va a adaptar a la regresión cuantil la suma (2.1). Dada una muestra aleatoria de observaciones independientes $\{(X_1,Y_1),...,(X_n,Y_n)\}$, un estimador no paramétrico del cuantil condicional se puede definir como $\hat{q}_{\tau,h_\tau} = \hat{a}_0$, donde $\hat{a}_0$ y $\hat{a}_1$ minimizan:

$$\sum_{i=1}^{n} \rho_\tau(Y_i - a_0 - a_1(X_i - x_0))K\left(\frac{X_i - x_0}{h_\tau}\right),$$

donde h_τ es un parámetro de suavizado. Notar que es el estimador lineal local de la función de regresión por cuantiles. El parámetro de suavizado h_τ muestra una fuerte influencia en la estimación resultante. Como se va a ver, existen varios enfoques para la selección de este parámetro de suavizado h_τ.

La expresión del error cuadrático medio (MSE) del estimador puntual de la función cuantil condicional de orden τ se da en el Teorema 3 de [5] y es:

Teorema 2.1. *El MSE de $\hat{q}_\tau(x_0)$ cuando $n \to \infty$, $h_\tau \to 0$ y $nh_\tau \to \infty$ es:*

$$\hat{q}_\tau(x_0) - q_\tau(x_0) \sim N\left(\frac{1}{2}q_\tau^{(2)}(x_0)\sigma_K^2 h_\tau^2, \frac{\tau(1-\tau)R(K)}{nh_\tau f_X(x_0) f(q_\tau(x_0)|X = x_0)}\right). \tag{2.4}$$

En este mismo artículo [5] se puede consultar la expresión del error cuadrático medio integrado (MISE).

2.2.2. Métodos de selección del parámetro de suavizado h

En [2] se describen diversos selectores de $h_{AMISE,\tau}$ para este problema. Algunos de ellos son el método plug-in o el de validación cruzada, como se va a ver a continuación.

- **Método plug-in**

  La técnica plug-in consiste en minimizar los términos dominantes del Error Cuadrático Medio Integrado (MISE) del estimador. Dada g función densidad de X, $f(q_\tau(x)|X = x)$ la densidad condicional de Y dado $X = x$ en $q_\tau(x)$, $q_\tau^{(i)}(x) = \frac{\partial^i q_\tau(x)}{\partial x^i}$ y $\mu_i(K) = \int u^i K(u)du$. Un parámetro de suavizado asintóticamente óptimo se obtiene como:

$$h_{AMISE,\tau} = \left[\frac{R(K)\tau(1-\tau)}{n\mu_2(K)^2 \int q_\tau^{(2)}(x)^2 g(x)dx} \int \frac{1}{f(q_\tau(x)|X = x)^2}dx\right]^{\frac{1}{5}}, \tag{2.5}$$

donde las integrales son desconocidas por lo que tienen que ser estimadas, y $\mu_2(K)$ y $R(K)$ se obtienen de la función kernel. La expresión (2.5) es bastante similar al método plug-in para la regresión media pero la función 'sparsity' jugará un papel importante.

- **Validación cruzada**

  La adaptación del método de validación cruzada consiste en sustituir el criterio de pérdida cuadrática por la función de pérdida de cuantiles. Entonces, se puede aplicar este procedimiento para la selección del parámetro de suavizado asociado con una regresión cuantil de tipo kernel, de tal manera que:

$$\hat{h}_{\tau,CV} = arg\min_{h} \sum_{i=1}^{n} \rho_\tau \left(Y_i - \hat{q}^{-i}_{\tau,h}(X_i) \right), \tag{2.6}$$

  donde $\hat{q}^{-i}_{\tau,h}(X_i)$ es el estimador de la función del cuantil de orden τ obtenido de una muestra que omite el dato del individuo i-ésimo.

En [2] se hace una revisión detallada de diferentes alternativas en la estimación del parámetro de suavizado h_τ. En concreto, en [3] se presenta un método de tipo plug-in para la estimación lineal local de la regresión cuantil no paramétrica. Muestran el selector sin imponer restricciones en la variabilidad condicional y la distribución del término del error. En su lugar, se propone la estimación no paramétrica de la curvatura en el cuantil dado τ, así como la estimación no paramétrica de la dispersión. Además, se muestra la convergencia de su estimador del parámetro de suavizado al óptimo y que la tasa de convergencia es la misma que en el contexto de regresión no paramétrica clásica.

2.3. Cruce de cuantiles

Cuando se quiere estimar dos o más cuantiles condicionales simultáneamente, las funciones cuantiles condicionales se pueden cruzar o superponerse. En [18] abordan este problema y muestran como corregirlo con determinadas restricciones. Este fenómeno desconcertante, denominado cruce de cuantiles, es causado porque la función cuantil condicional se estima independientemente para cada cuantil. En los intervalos donde el fenómeno sucede, se aborda el problema introduciendo unas restricciones sin cruces.

Capítulo 3

Aplicación de la Regresión Cuantil

3.1. Software para Regresión Cuantil en R

R es un lenguaje de programación para realizar análisis estadísticos que cuenta con grandes posibilidades gráficas. Dispone de un almacenamiento efectivo de datos en diferentes clases de objetos. Por otra parte, se pueden combinar métodos de análisis estándar con análisis desarrollados de forma adecuada para una situación específica. Véase [17].

R es totalmente gratuito bajo los términos de la GNU (General Public License). Además, R dispone de varias librerías creadas a partir de una amplia comunidad de usuarios que contribuyen a implementar sus capacidades básicas y extender su funcionalidad.

En este caso, se va a utilizar el paquete `quantreg` creado por Roger Koenker. En él, se describen métodos de estimación e inferencia para modelos de regresión de cuantiles condicionales: modelos lineales y no lineales, paramétricos y no paramétricos. Ver [13].

La función más importante que se va a usar es `rq` y representa un ajuste de regresión cuantil. Además de la fórmula objeto, en los argumentos de la función `rq` se incluye el cuantil o cuantiles τ a ser estimados. En el caso de tomar un cuantil $\tau \notin (0,1)$, devuelve un objeto de la clase `rq.process`, es decir, el proceso completo. Del mismo modo, se puede tomar τ como un vector de valores entre 0 y 1, en cuyo caso se devuelve un objeto de la clase `rqs`, donde se tiene una matriz de estimaciones de coeficientes en los cuantiles especificados. Otro argumento que se utiliza es `na.action`, en particular, `na.omit` será útil para eliminar observaciones con datos faltantes. En el caso de que se quiera hacer a continuación un `summary`, hay que indicar `model=TRUE`. En cuanto al método de estimación se recomienda la siguiente estrategia:

- `method=`"`br`": el método de Barrodale y Roberts para problemas de tamaño moderado (predeterminado).
- `method=`"`fn`": para problemas grandes resulta mejor el algoritmo de Frisch-Newton.
- `method=`"`pfn`": para problemas muy grandes, se puede usar el enfoque de Frisch-Newton con preprocesamiento.

Esta función `rq` solamente produce estimaciones de los coeficientes. Para evaluar la precisión de estas estimaciones es necesario un `summary` que determina si las covariables son significativas en cuantiles particulares. Para especificar el método de cálculo de error estándar hay hasta seis métodos disponibles. En este caso, se utiliza `se=`"`rank`" que devuelve intervalos de confianza de los parámetros estimados (predeterminado), o bien `se=`"`nid`" para obtener el p-valor.

Con la función `anova` se puede calcular un test estadístico para dos o más ajustes de regresión cuantil. En ésta hay que especificar la función objeto que suele ser `rq` o un objeto de tipo `rqs` con varios cuantiles τ. Para más detalles véase la viñeta disponible desde R en [12] y [10].

En este capítulo se van a realizar dos partes de estudio. Una primera en la que se considera un fichero de datos de los ingresos en las dos primeras semanas de estreno de películas, donde podrá verse el efecto de lo recaudado un fin de semana en el estreno de una película sobre el siguiente fin de semana y, también, se incorporará información adicional sobre el número de cines donde se proyecta. Con estos datos se llevará a cabo un estudio en escala logarítmica en la que se verá cómo en ciertos modelos se da paralelismo con mínimos cuadrados de regresión simple. Con esto se ve que hay modelos paramétricos que describen muy bien el conjunto de datos. Por otro lado, en el segundo conjunto de datos los modelos paramétricos no son satisfactorios, por ello se van a utilizar unos modelos de tipo no paramétrico diferenciando hombres y mujeres.

3.2. Regresión Cuantil paramétrica

En esta sección se va a utilizar el paquete `quantreg`. En el apéndice A se encuentra el script de lo que se realiza.

3.2.1. Datos

Se va a trabajar con un fichero de datos extraídos de la página [9] que consta de 268 filas y 11 columnas. El conjunto de datos *boxoffice* está formado por las recaudaciones de estrenos de películas en el cine durante los años 2015 a 2017 en las dos primeras semanas, el número de salas en las que se ha proyectado y la posición en el ranking TOP10 en la primera semana.

Las variables que se tienen son las siguientes:

- *Nweek*: fecha de su estreno (mes, fin de semana, año).
- *TW1*: posición en el ranking TOP10 de recaudación (total) tras el primer fin de semana.
- *Movie*: nombre de la película (nombre original).
- *Studio*: estudio o distribuidora de la película.
- *Weekend.Gross.1w*: recaudación en euros en el primer fin de semana de proyección.
- *Theaters.1w*: número de salas en las que se ha proyectado en su primer fin de semana.
- *Weekend.Gross.2w*: recaudación en euros en el segundo fin de semana de proyección.
- *Theaters.2w*: número de salas en las que se ha proyectado en su segundo fin de semana.
- *TW2*: posición en el ranking de recaudación (total) tras el segundo fin de semana.

3.2.2. Estudio de los datos

Se comienza realizando un primer modelo en el que se va a predecir la recaudación bruta de la segunda semana conociendo la recaudación bruta de la primera semana de estreno. Luego, se toma como X la variable *Weekend.Gross.1w* y como Y se toma la variable *Weekend.Gross.2w*. Para ello, se utiliza la función `rq` y se hace el ajuste de la fórmula:

$$Weekend.Gross.2w \sim Weekend.Gross.1w \tag{3.1}$$

para los cuantiles de orden 0,25, 0,50 y 0,75. Con los resultados obtenidos, el ajuste del modelo de regresión cuantil para los tres cuantiles especificados es:

$$\begin{cases} Y = -2274{,}41 + 0{,}47X & \text{si} \quad \tau = 0{,}25 \\ Y = -2965{,}41 + 0{,}57X & \text{si} \quad \tau = 0{,}50 \\ Y = -12267{,}98 + 0{,}71X & \text{si} \quad \tau = 0{,}75. \end{cases} \tag{3.2}$$

Con esto se tiene la estimación de los coeficientes para los tres cuantiles especificados. Observar que para cada cuantil τ se tiene un ajuste diferente con distintas pendientes, así la variable *Weekend.Gross.1w* influye de diferente forma en cada cuantil de la variable respuesta *Weekend.Gross.2w*.

A continuación, se realiza un gráfico de un diagrama de dispersión con las rectas de regresión ajustadas del modelo propuesto (3.1) para los cuantiles desde $\tau = 0{,}1$ hasta $\tau = 0{,}9$:

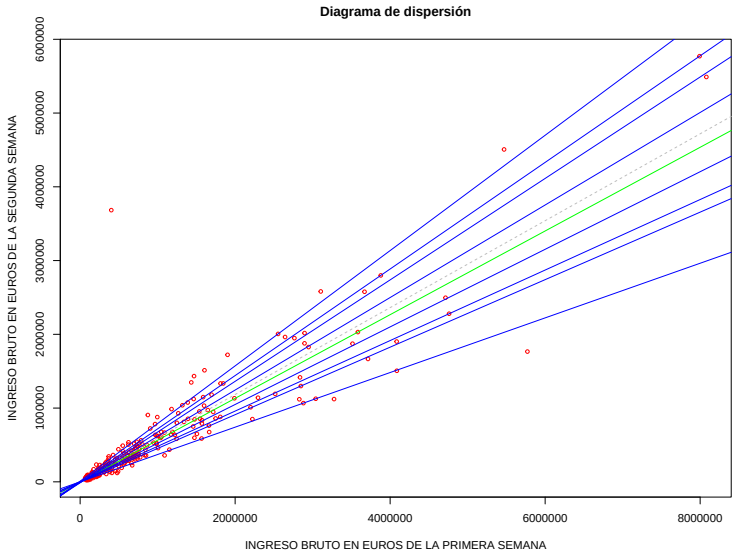

Figura 3.1: Diagrama de dispersión del ajuste regresión cuantil de la recaudación bruta de la segunda semana conocida la recaudación bruta de la primera semana.

Las líneas azules en la Figura 3.1 son los ajustes `rq` para los cuantiles especificados, la línea verde es el ajuste del cuantil 0,5 y la línea discontinua rosa es el ajuste por mínimos cuadrados de la mediana condicional.

Ahora, se va a evaluar la precisión de estas estimaciones del modelo (3.1) con `summary` en el que se obtiene para cada coeficiente estimado un intervalo de confianza que, por defecto, se calcula con el método predeterminado para hallar el error residual `se="rank"`. Veamos estos IC en la siguiente tabla:

Quantiles	(Intercept)	Weekend.Gross.1w
0,25	−2274,409	0,466
	(−9389,718, 8338,827)	(0,438, 0,497)
0,50	−2965,411	0,567
	(−13033,665,5025,653)	(0,528, 0,631)
0,75	−12267,978	0,709
	(−21521,088, −690,166)	(0,679, 0,726)

Cuadro 3.1: Intervalos de confianza de los coeficientes estimados en modelo (3.1).

En el Cuadro 3.1, se observa que el intervalo de confianza del coeficiente estimado de la variable predictora *Weekend.Gross.1w* en los tres casos no contiene el 0, de manera que hay evidencias de que el predictor influye significativamente en el modelo propuesto (3.1). Notar que para cada coeficiente estimado se tienen distintos IC, disjuntos entre sí dos a dos para los tres cuantiles especificados. Sin embargo, el intervalo de confianza del intercepto si que contiene el 0, lo que implica que el coeficiente es nulo.

Para mayor claridad, se realiza un `plot` de los coeficientes estimados con sus intervalos de confianza correspondientes para el intercepto y la variable predictora. Éste es un tipo de gráfico que se va a usar para resumir la significancia de las variables en los diferentes cuantiles τ que se eligen.

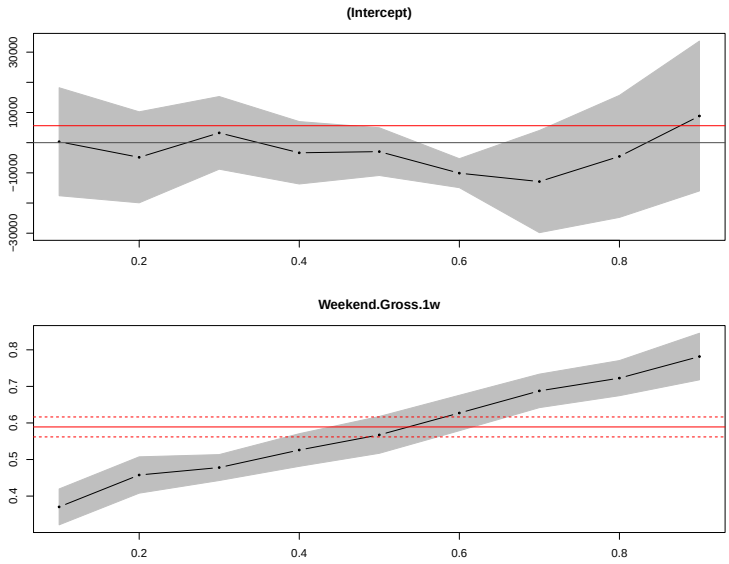

Figura 3.2: Gráfico de los coeficientes estimados en el modelo *Weekend.Gross.2w* $\sim$ *Weekend.Gross.1w* desde $\tau = 0{,}1$ hasta $\tau = 0{,}9$.

En la Figura 3.2 se muestran como varían los coeficientes de regresión cuantil estimados para los cuantiles $\tau = 0{,}1$ hasta $\tau = 0{,}9$. Además, se dibuja el coeficiente de regresión estimado por mínimos cuadrados (línea continua roja en la Figura 3.2) y su intervalo de confianza al 95 % se indica con la línea discontinua. En el caso del intercepto, se ve como el 0 es el valor que predomina para muchos de los percentiles (la línea negra está denro de la banda), luego se admite el 0 como valor de los coeficientes estimados. Por otro lado, en el caso de los coeficientes estimados de *Weekend.Gross.1w* se observa la discrepancia de los coeficientes obtenidos con regresión cuantil y mínimos cuadrados. Por lo tanto, se confirma que la variable *Weekend.Gross.1w* es significativa para el modelo.

Además para cada cuantil τ se tiene una pendiente distinta, pues al hacer `anova` se obtiene que si que hay diferencia significativa entre los coeficientes de regresión de los cuantiles 0,25, 0,5 y 0,75.

Si se quieren obtener los p-valores es necesario especificar el método `se=`"`nid`" para hallar el error residual en la función `summary`. Se obtiene para los tres cuantiles indicados en el modelo (3.1), que el p-valor de la variable predictora *Weekend.Gross.1w* es $0{,}00 < 0{,}05$ lo cual implica que se acepta que el coeficiente estimado es no nulo, es decir, la variable predictora es significativa en el modelo propuesto. Por otro lado, el p-valor del intercepto para los tres ajustes es mayor que 0,05, luego el coeficiente del intercepto es nulo.

Ahora se va a añadir alguna otra variable adicional al modelo (3.1) para ver si mejora. Se plantean los siguientes modelos:

(I) *Weekend.Gross.2w* $\sim$ *Weekend.Gross.1w*

(II) *Weekend.Gross.2w* $\sim$ *Weekend.Gross.1w + Theaters.1w*

(III) *Weekend.Gross.2w* $\sim$ *Weekend.Gross.1w + Theaters.2w*

(IV) *Weekend.Gross.2w* $\sim$ *Weekend.Gross.1w + Theaters.1w + Theaters.2w*

(V) *Weekend.Gross.2w* $\sim$ *Weekend.Gross.1w + I(Theaters.1w - Theaters.2w)*

Para comparar los modelos, se enfrentan dos a dos con la función `anova` y se analiza el p-valor obtenido para distintos valores de τ. En la siguiente tabla se muestran los p-valores que se obtienen:

Comparativa de los modelos propuestos para distintos τ					
Modelos a comparar	$\tau = \mathbf{0{,}2}$	$\tau = \mathbf{0{,}4}$	$\tau = \mathbf{0{,}5}$	$\tau = \mathbf{0{,}6}$	$\tau = \mathbf{0{,}8}$
(I) vs (II)	0,213	0,115	0,022$*$	0,012$*$	0,118
(I) vs (III)	0,614	0,401	0,084	0,039$*$	0,673
(II) vs (IV)	0,170$*$	0,000$**$	0,002$**$	0,012$*$	0,040$*$
(III) vs (IV)	0,001$**$	0,000$**$	0,000$**$	0,000$**$	0,008$**$
(I) vs (IV)	0,007$**$	0,000$**$	0,000$**$	0,000$**$	0,021$*$
(I) vs (V)	0,000$**$	0,000$**$	0,020$*$	0,056	0,010$*$

Cuadro 3.2: Tabla comparativa de los modelos propuestos para datos boxoffice.

En el Cuadro 3.2 se observa que, añadiendo por separado al modelo (3.1), las variables *Theaters.1w* y *Theaters.2w* no son influyentes en el modelo. Cuando se comparan los modelos (II) y (IV) comienza a notarse una influencia importante, pues en este caso la suma de las variables *Theaters.1w* y *Theaters.2w* es significativa de forma aditiva en el modelo. La comparación entre los modelos (III) y (IV) es similar a la comparativa anterior. Así, se concluye que las variables *Theaters.1w* y *Theaters.2w* añadidas por separado al modelo no influyen, pero si se predice el ingreso teniendo estas dos variables en cuenta entonces sí que existen diferencias significativas. Las dos últimas comparaciones son las más importantes. Únicamente se obtiene que no es significativa en el caso de un cuantil al comparar el primer y último modelo.

Una medida local de ajuste para modelos de regresión cuantil es el coeficiente $R^1(\tau)$. Se trata de una medida pseudo-R^2 sugerida por Koenker y Machado (1999) que mide la bondad de ajuste comparando la suma de las desviaciones ponderadas para el modelo de interés con la misma suma de un modelo en el que solo aparece el intercepto. Se calcula como:

$$R^1(\tau) = 1 - \frac{\widehat{V}(\tau)}{\tilde{V}(\tau)}, \tag{3.3}$$

donde $\tilde{V}(\tau)$ es el modelo nulo, es decir, sin covariables y $\widehat{V}(\tau)$ es el modelo nuevo propuesto con términos de las $\mathbf{x}$:

$$\tilde{V}(\tau) = \min_{\beta\in\mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \beta) \quad \text{y} \quad \widehat{V}(\tau) = \min_{\mathbf{b}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i^t\mathbf{b}).$$

Por definición, como $\widehat{V}(\tau)$ es más pequeño que $\tilde{V}(\tau)$, se tiene que el coeficiente $R^1(\tau)$ debe estar en el intervalo $[0, 1]$, donde valores de $R^1(\tau)$ próximos a 1 corresponderían a un ajuste perfecto, ya que de esta manera la función de pérdida con los nuevos términos $\widehat{V}(\tau)$ es mucho más pequeña que la función de pérdida sin esos términos de las $\mathbf{x}$, $\tilde{V}(\tau)$. Consiguiendo así un mínimo más pequeño en cuanto al valor en el modelo $\widehat{V}(\tau)$ respecto del modelo $\tilde{V}(\tau)$. Por lo tanto, cuanto más pequeña sea $\widehat{V}(\tau)$, mejor

se estimará el cuantil τ gracias a la información del modelo.

Si se calcula el coeficiente de bondad de ajuste $R^1(\tau)$ en los modelos que se han propuesto para los cuantiles $\tau = 0{,}2$, $0{,}4$, $0{,}5$, $0{,}6$ y $0{,}8$ se obtienen los siguientes valores, respectivamente, para cada modelo:

```
[1] 0.5801361 0.6535443 0.6788192 0.7069694 0.7673077
[1] 0.5864874 0.6590203 0.6836757 0.7180489 0.7716879
[1] 0.5979031 0.6649507 0.6888416 0.7224491 0.7751113
[1] 0.6094620 0.6758955 0.6964613 0.7296407 0.7817980
[1] 0.6092097 0.6758614 0.6964268 0.7259136 0.7813825
```

Pero son valores no próximos a 1, solamente se observan valores más cercanos a 1 en los coeficientes de $R^1(\tau)$ en los modelos (IV) y (V). Notar que esta medida de pseudo-R^2 no tiene en cuenta el número de parámetros, simplemente el grado de ajuste. Recordar que el modelo (IV) tiene 3 parámetros mas la pendiente, sin embargo el modelo (V) tiene 2 parámetros mas la pendiente, por lo que son modelos igual de buenos.

Para ver como influyen estas variables en el modelo (V) se va a hacer un gráfico `plot` en el que se plasma la significancia de las variables para los cuantiles τ desde $0{,}1$ hasta $0{,}9$:

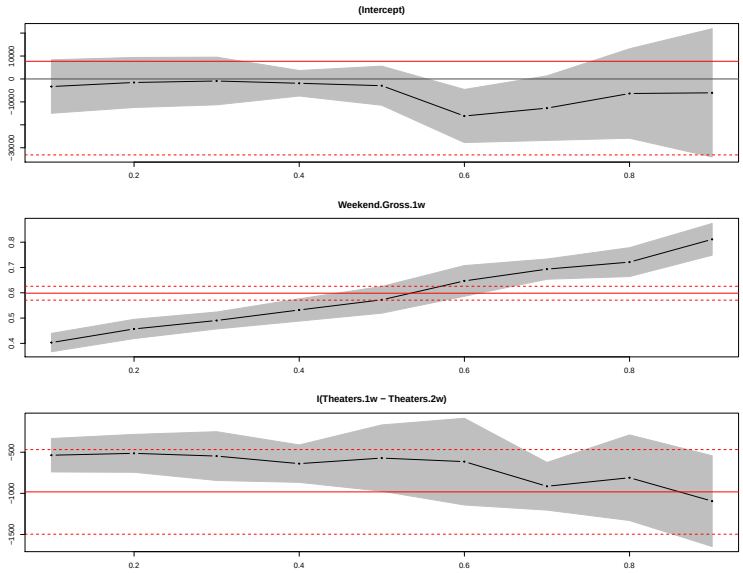

Figura 3.3: Gráfico de los coeficientes estimados en el modelo *Weekend.Gross.2w* $\sim$ *Weekend.Gross.1w* + *I(Theaters.1w - Theaters.2w)* desde $\tau = 0{,}1$ hasta $\tau = 0{,}9$.

En la Figura 3.3 se observa como en el caso del intercepto el 0 es un valor admisible para los coeficientes estimados, al contrario de lo que ocurre para las covariables, pues se ve como la línea negra no entra en la banda de confianza roja.

3.2.3. Modelo en escala logarítmica

A continuación, se va a presentar un modelo en escala logarítmica. Debido a que en el modelo (V) se tiene la resta *Theaters.1w-Theaters.2w*, desde el punto de vista económico lo que se va a hacer es pasar a la recaudación por sala y se estudiará el siguiente modelo:

$$\log\left(\frac{Weekend.Gross.2w}{Theaters.2w}\right) \sim \log\left(\frac{Weekend.Gross.1w}{Theaters.1w}\right). \tag{3.4}$$

Antes de comenzar, se realiza un gráfico de dispersión y se observa que los datos 16 y 131 son atípicos, por lo que se quitan del modelo y entonces se pasa a escala logarítmica. Se hace el ajuste `rq` de la fórmula (3.4) para los cuantiles $\tau = 0{,}1,\ 0{,}2,\ 0{,}3,\ 0{,}4,\ 0{,}5,\ 0{,}6,\ 0{,}7,\ 0{,}8$ y $0{,}9$ y se representa en el siguiente diagrama de dispersión:

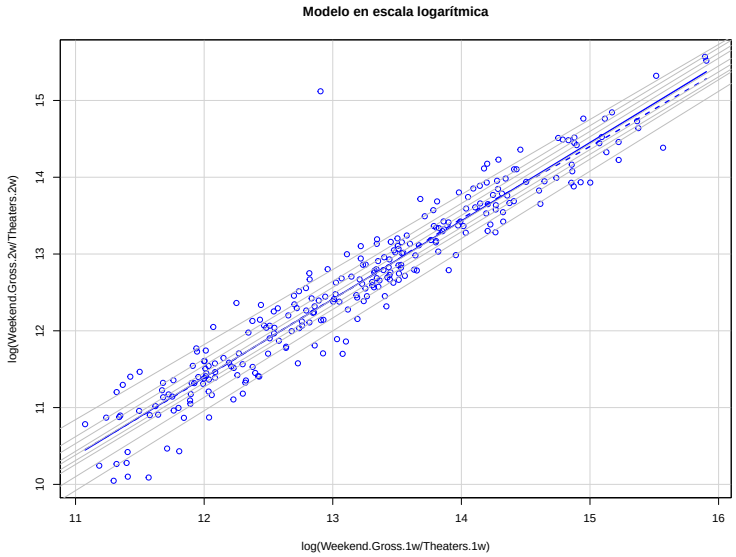

Figura 3.4: Gráfico del ajuste de regresión cuantil en escala logarítimica de recaudación por sala desde $\tau = 0{,}1$ hasta $\tau = 0{,}9$.

En la Figura 3.4 se observa que todas las pendientes son muy similares para todos los cuantiles especificados.

El coeficiente de bondad de ajuste $R^1(\tau)$, en este caso, debe de estar muy próximo a 1 al tratarse de un ajuste casi perfecto. Se obtienen los siguientes coeficientes:

```
[1] 0.9999994 0.9999995 0.9999995 0.9999995 0.9999995
0.9999988 0.9999994 0.9999996 0.9999998
```

Por lo tanto, se confirma que el ajuste es perfecto y que el parámetro se puede interpretar como un parámetro de elasticidad, es decir, se tiene lo que se denomina elasticidad perfecta en cuantiles. Esto es, cada 10 % que aumenta un cuantil en la primera semana de estreno, también ese mismo cuantil aumenta un 10 % en la segunda semana.

Se realiza la tabla `anova` para comparar las pendientes de los ajustes con los distintos cuantiles τ elegidos y se obtiene:

```
Quantile Regression Analysis of Deviance Table

Model: log(y) ~ log(x)
Joint Test of Equality of Slopes: tau in {  0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9  }

  Df Resid Df F value Pr(>F)
1  8     2386   0.826 0.5796
```

Luego, se acepta la hipótesis de igualdad de pendientes de una forma muy clara. Así, se concluye que todas las pendientes son iguales.

3.3. Modelos no paramétricos y de tipo spline

A lo largo de esta sección se utilizará el paquete `quantreg` descrito al comienzo del capítulo. Véase los script en apéndice B.

En esta sección se va a trabajar con un nuevo conjunto de datos que muestra un comportamiento diferente al fichero de datos anterior. El fichero de datos *bone* se puede encontrar en el paquete de R [7]. Este fichero reúne la medición de la densidad mineral ósea de 261 adolescentes norteamericanos. El fichero consta de 485 observaciones de las siguientes variables:

- *idnum*: identifica al niño debido a que hay varias mediciones en un mismo individuo.
- *age*: edad media del individuo en las dos visitas consecutivas.
- *gender*: indica el género (femenino o masculino).
- *spnbmd*: medición de la densidad mineral relativa en columna vertebral, medida como la diferencia entre dos registros consecutivos dividida entre la semisuma de ambos registros.

De acuerdo con la definición de la Organización Mundial de la Salud (OMS), la densidad ósea en un adulto joven es normal si la medición se encuentra en el intervalo $(-1,\ 1)$. Si la densidad ósea es menor de $-2{,}5$ se tiene osteoporosis. En este conjunto de datos se tienen todos los valores en el intervalo $(-1, 1)$, es lógico ya que son mediciones en adolescentes con edades comprendidas entre 9 y 25 años.

3.3.1. Exploración de datos y modelos paramétricos

Se va a comenzar realizando un gráfico de violín, que resulta útil para estudiar una variable continua en las diferentes categorías de una variable categórica. Este gráfico es similar a generar una curva de densidad para cada categoría, dispuestos de otra manera para facilitar la comparación. Al igual que en un boxplot, en el gráfico de violín se visualizan los cuartiles, los extremos y la presencia de posibles datos atípicos.

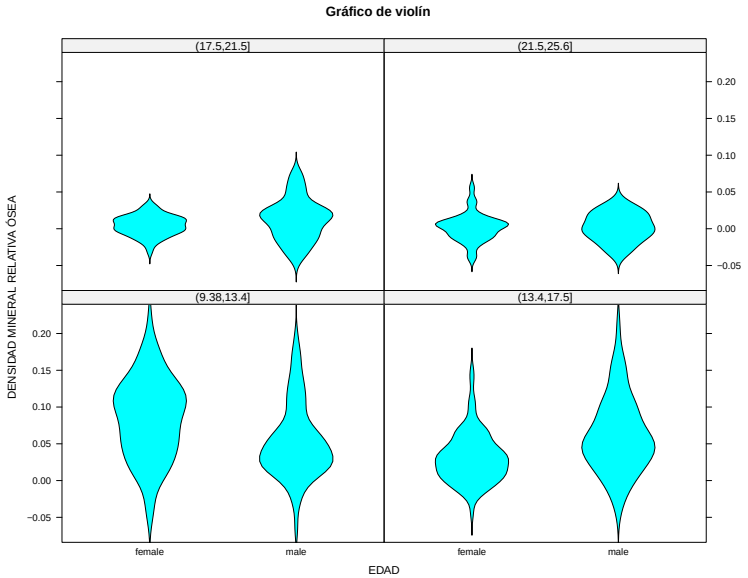

Figura 3.5: Gráfico de violín de la densidad mineral relativa ósea respecto la edad en cuatro intervalos por grupos según el género.

En la Figura 3.5 se muestra como los cuartiles difieren con la edad de forma no lineal y de manera distinta en chicos y chicas. En el intervalo de edad $(9{,}38,\ 13{,}4]$ se ve como los cuartiles son más

amplios y el rango intercuartílico es mayor. Se observa como para el género masculino se concentran más los datos, mientras en el género femenino se aprecian dos modas y los datos están más dispersos. En el intervalo $(13{,}4,\ 17{,}5]$, en el caso de las chicas se aglomeran más los datos, y se tiene un rango intercuartílico menor, mientras que en los chicos se dispersan más los datos y sigue teniendo un rango intercuartílico similar al intervalo anterior.

En el intervalo de edad $(17{,}5,\ 21{,}5]$ se ve como se concentran más los datos en ambos géneros, incluso se puede apreciar como las chicas tienen menor densidad mineral relativa ósea que los chicos debido a que el rango intercuartílico es menor que en los chicos. Por último, en el intervalo $(21{,}5,\ 25{,}5]$ se observa como se condensan aún más en el caso del género femenino, mientras que en el caso del género masculino se ve como el rango intercuartílico es mayor.

En el apéndice B, la Figura B.1 se complementa a este gráfico con un gráfico de cajas más usual donde se ve con mayor precisión la posición de los cuartiles en cada intervalo de edad, tanto para chicos como para chicas, así como la presencia de datos atípicos.

En la exploración de datos se ha visto que la variable género influye de alguna forma en la densidad mineral relativa ósea, por lo que se propone el estudio de los siguientes modelos: un primer modelo sin covariables, otro con efecto solo de la edad sobre la densidad mineral relativa ósea, un tercer modelo con efecto solo de la variable género, otro con efecto de las variables edad y género de forma aditiva y un último modelo con interacción de las variables género y edad:

(I) *spnbmd* $\sim$ *1*

(II) *spnbmd* $\sim$ *age*

(III) *spnbmd* $\sim$ *gender*

(IV) *spnbmd* $\sim$ *age* + *gender*

(V) *spnbmd* $\sim$ *age*gender*.

Se comparan dos a dos estos cinco modelos con la función `anova`. En la siguiente tabla comparativa se recogen los p-valores que se han obtenido:

Comparativa de los modelos propuestos para distintos τ

Modelos a comparar	$\tau = \mathbf{0{,}2}$	$\tau = \mathbf{0{,}4}$	$\tau = \mathbf{0{,}5}$	$\tau = \mathbf{0{,}6}$	$\tau = \mathbf{0{,}8}$
(I) vs (II)	0,000**	0,000**	0,000**	0,000**	0,000**
(I) vs (III)	0,722	0,016*	0,162	0,226	0,283
(II) vs (IV)	0,514	0,882	0,718	0,193	0,351
(III) vs (IV)	0,000**	0,000**	0,000**	0,000**	0,000**
(II) vs (V)	0,514	0,882	0,718	0,193	0,351
(III) vs (V)	0,649	0,049*	0,000**	0,000**	0,191

Cuadro 3.3: Tabla comparativa de los modelos propuestos para datos bone.

En el Cuadro 3.3, al comparar los modelos (II) y (IV) se obtiene que no hay diferencia entre ellos, luego la variable género parece que no es significativa de forma aditiva en el modelo. A pesar de que estos dos modelos son similares, se tiene que el modelo (V) es mucho mejor que (IV), pues (V) mejora significativamente respecto a (IV) y respecto a (II).

Por lo tanto, se tiene que el género influye en las curvas cuantiles de forma significativa aunque no de forma aditiva ya que no desplaza la curva cuantil, sino que la modifica de una forma más completa.

Luego la interacción hace que el género sea significativo.

Se puede ver como el género es significativo en el modelo (V) en el siguiente `plot` donde se representa el ajuste del modelo con efecto de las variables edad y género con interacción sobre la densidad mineral relativa ósea con las diferentes rectas en hombres y mujeres para el cuantil de orden $\tau = 0{,}5$:

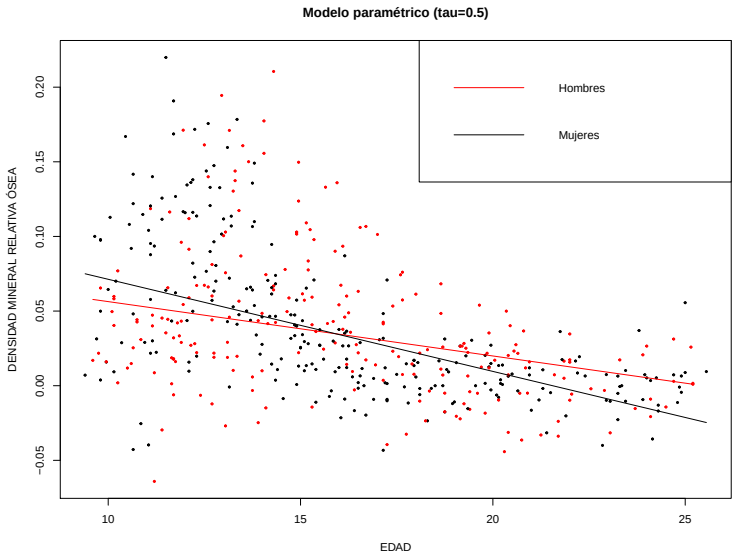

Figura 3.6: Gráfico del ajuste regresión cuantil del modelo con interacción con género y edad para $\tau = 0{,}5$ diferenciado para hombres y mujeres.

En la Figura 3.6 se observa que el efecto edad es diferente entre hombres (línea roja en Figura 3.6) y mujeres (línea negra en Figura 3.6), la influencia de la edad es diferente ya que se tienen dos rectas con distinta pendiente.

Debido a que la regresión cuantil paramétrica no puede captar las tendencias de la nube de puntos, se necesita la utilización de suavización que se llevará a cabo en la siguiente sección.

3.3.2. Modelos no paramétricos

Para captar los comportamientos de los datos y poder ver el efecto de unas variables sobre otras, se necesitan técnicas de suavización como ya se ha comentado anteriormente.

En esta sección se utilizará la función `lprq` del paquete `quantreg`, que proporciona el cálculo del estimador local lineal de regresión cuantil. Para el uso de esta función se debe especificar la covariable condicional X, la variable respuesta Y, el parámetro de suavizado h, los cuantiles τ a ser estimados y el número m de puntos en los que la función es estimada.

En este caso, se considera $h = 0{,}75$ como valor único y se elige $m = 200$. Para obtener una mejor visualización de los gráficos, se ordena el conjunto de datos por la edad en el fichero de datos.

A continuación, se representa el ajuste no paramétrico de la densidad mineral relativa ósea respecto de la edad diferenciado para hombres y mujeres de los cuantiles elegidos $\tau = 0{,}2$, $0{,}4$, $0{,}5$, $0{,}6$ y $0{,}8$.

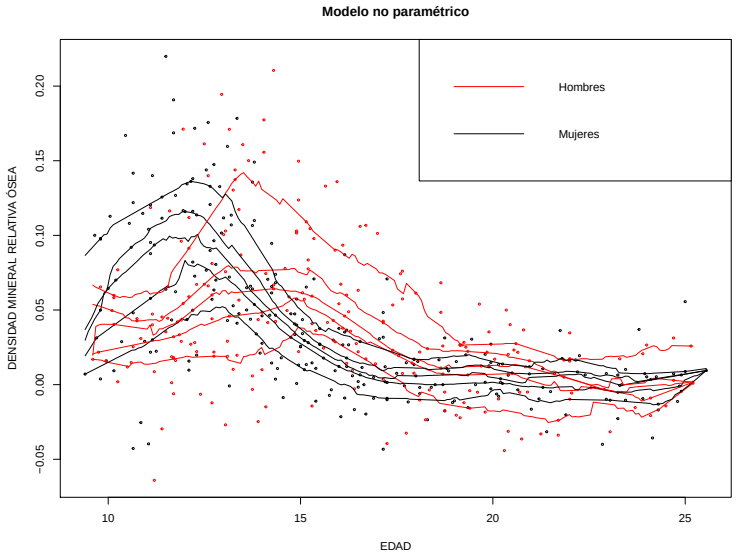

Figura 3.7: Gráfico del ajuste regresión cuantil no paramétrica de la densidad mineral relativa ósea respecto edad diferenciado para hombres y mujeres de los cuantiles $\tau = 0{,}2$, 0,4, 0,5, 0,6 y 0,8.

En la Figura 3.7 se observa que hay curvaturas diferentes en hombres y mujeres para los distintos cuantiles especificados. En el intervalo de edad de 9 a 15 años se observa un crecimiento más rápido de los cuantiles, más acelerado para las mujeres. A partir de los 15 años, se percibe un decrecimiento más lento de la concentración relativa. Notar que, en estas edades, en el caso de las chicas los cuantiles se concentran más que en los chicos.

3.3.3. Modelo con splines y cruce de cuantiles

En esta sección, se van a utilizar las bases de splines para conseguir una modelización flexible de las curvas cuantiles y, a la vez, que nos permita comparar y seleccionar el mejor modelo.

Para todos los cuantiles considerados $\tau = 0{,}2$, 0,4, 0,5, 0,6 y 0,8, se selecciona en base al mínimo AIC el número óptimo de grados de libertad. En realidad, en los modelos de regresión cuantil, el cálculo del AIC está realizado asumiendo un modelo de errores normales. Esto permite un alto grado de compatibilidad con otras funciones de *R*, pero se trata de una pseudo-verosimilitud ya que dicha normalidad no suele ser una hipótesis asumible. Este valor del AIC sigue siendo útil para la comparación y selección de modelos anidados, como es este caso.

A continuación, se ve el ajuste que se consigue introduciendo un criterio de selección, que consiste en minimizar el AIC. Es el mejor modelo que se podría conseguir con los splines. Se obtiene el siguiente gráfico diferenciado para hombres y mujeres:

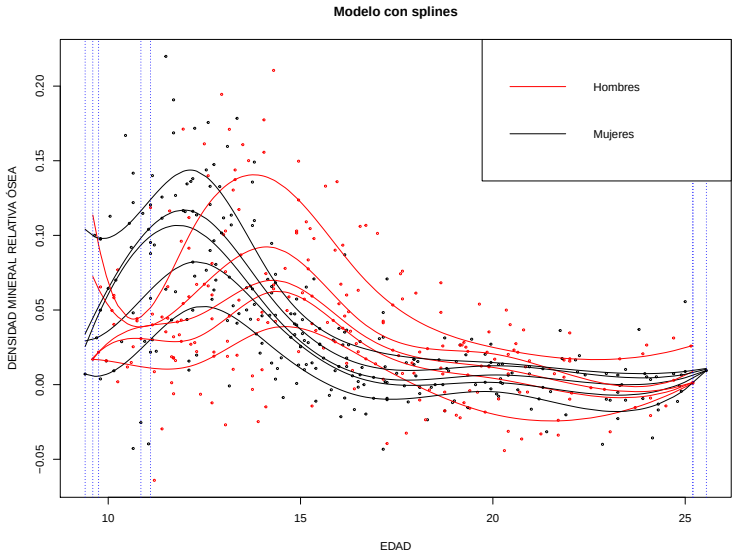

Figura 3.8: Modelo con splines diferenciado para hombres y mujeres de los cuantiles $\tau = 0{,}2$, 0,4, 0,5, 0,6 y 0,8.

En la Figura 3.8 se observa como las curvas son más flexibles que en la Figura 3.7. En este gráfico se puede ver con más claridad las diferencias entre hombres y mujeres para las curvas de los distintos cuantiles τ elegidos.

Si se calcula el coeficiente de bondad de ajuste $R^1(\tau)$ en el modelo para los cuantiles $\tau = 0{,}2$, 0,4, 0,5, 0,6 y 0,8 se obtienen los siguientes valores:

```
[1] 0.1685667
[1] 0.2297703
[1] 0.2700062
[1] 0.3143658
[1] 0.3963982
```

no son muy grandes, pues hay mucha dispersión. Entonces, es razonable que el ajuste no sea tan bueno como podría imaginarse. El problema es un problema lineal y siempre es más complicado dar un buen ajuste.

En la Figura 3.8 se puede observar, especialmente en los extremos del eje horizontal de la edad, que las curvas de regresión cuantil pueden tocarse e incluso llegar a cruzarse. Como se ha comentado en el capítulo anterior, este fenómeno de 'crossing' no es deseable y para evitarlo se propone un procedimiento de isotonización. Este método se basa en la regresión isotónica de los cuantiles condicionales en cada uno de los puntos donde se estiman las curvas.

Esta técnica para evitar el 'crossing' es básicamente visual y se podría incorporar en el proceso de estimación de las curvas cuantil imponiendo la monotonía. Para resolver este problema se va a aplicar un procedimiento basado en la regresión isotónica [15].

Sean $a_1, ..., a_n$ valores que deberían estar ordenados, entonces se calculan los valores $y_1, ..., y_n$ tales que $y_1 \leq y_2 \leq ... \leq y_n$ y además minimizan la suma de cuadrados siguiente:

$$\sum_{i=1}^{n} (a_i - y_i)^2. \tag{3.5}$$

Para proceder con esta técnica, se comienza detectando para que valores de la variable edad se producen cruces:

```
[1] 25.55 11.10 10.85 25.20 25.20  9.40  9.75  9.6
```

Se tienen 8 valores de la variable edad donde se producen cruces en las curvas estimadas de los distintos cuantiles. Estos valores se corresponden con las rectas verticales en la Figura 3.8. Notar que en el gráfico solo hay 7 rectas verticales, ya que hay un valor de x (para $x = 25{,}20$) en el que se producen dos cruces. A continuación, se muestra la matriz de los cuantiles condicionales en la que se ve que no hay monotonía para los valores de x obtenidos.

```
             [,1]        [,2]        [,3]        [,4]       [,5]
[1,] 0.0094674560 0.009467456 0.009467456 0.010352912 0.01084100
[2,] 0.0111181420 0.030284139 0.039815170 0.039639219 0.05154655
[3,] 0.0120379689 0.030897210 0.038701395 0.038601144 0.04536399
[4,] 0.0009803922 0.001607200 0.001607200 0.008942388 0.02625506
[6,] 0.0070721360 0.029255901 0.025528047 0.034048811 0.10402326
[7,] 0.0166732012 0.021748590 0.021748590 0.064080878 0.09365411
[8,] 0.0169779300 0.017858519 0.016977930 0.072454567 0.11361501
```

Ahora, se aplica la función `isoreg` para isotonizar las curvas en las que hay cruces, obteniéndose la siguiente matriz de cuantiles condicionados ya isotonizados:

```
             [,1]        [,2]        [,3]        [,4]       [,5]
[1,] 0.0094674560 0.009467456 0.009467456 0.010352912 0.01084100
[2,] 0.0111181420 0.030284139 0.039727194 0.039727194 0.05154655
[3,] 0.0120379689 0.030897210 0.038651269 0.038651269 0.04536399
[4,] 0.0009803922 0.001607200 0.001607200 0.008942388 0.02625506
[6,] 0.0070721360 0.027391974 0.027391974 0.034048811 0.10402326
[7,] 0.0166732012 0.021748590 0.021748590 0.064080878 0.09365411
[8,] 0.0169779300 0.017418224 0.017418224 0.072454567 0.11361501
```

Estos valores de la matriz son monótonos crecientes, pero no estrictamente crecientes ya que se observa, en alguna fila, que hay valores duplicados. Estos valores corresponden a los casos en los que no se cumplía la condición de monotonía.

En conclusión, ante un conjunto de datos con una estructura compleja y distribuciones condicionales muy diferentes, se han utilizado técnicas de regresión cuantil de tipo no paramétrico para evidenciar una relación cuantil no lineal. Además, el uso de modelos sobre bases de tipo spline permite seleccionar de forma automática los modelos flexibles óptimos separados para hombres y mujeres para cada cuantil τ elegido. Finalmente, se ha usado una técnica de isotonización para corregir cruces inadecuados en las curvas cuantil.

Bibliografía

[1] J.A. CRISTÓBAL CRISTÓBAL, *Inferencia Estadística, Editado por Prensas Universitarias de Zaragoza, 2ª edición, 2000.*

[2] M. CONDE-AMBOAGE, W. GONZÁLEZ-MANTEIGA Y C. SÁNCHEZ-SELLERO, *Quantile regression: estimation and lack-of-fit tests, Boletín de Estadística e Investigación Operativa,* ***34****, (2) (2018), 97–117.*

[3] M. CONDE-AMBOAGE Y C. SÁNCHEZ-SELLERO, *A plug-in bandwidth selector for nonparametric quatile regression. TEST.(2018)*, https://doi.org/10.1007/s11749-018-0582-6.

[4] FAN, *Local Linear Regression Smoothers and Their Minimax Efficiencies, The Annals of Statistics,* ***21****, (1), (1993), 196–216*, https://projecteuclid.org/download/pdf_1/euclid.aos/1176349022.

[5] J. FAN, T.C. HU, AND Y.K. TRUONG, *Robust Non-parametric Function Estimation, Scandinavian Journal of Statistics,* ***21****, (4), (1994), 433-466*, https://www.jstor.org/stable/4616328?seq=1#metadata_info_tab_contents.

[6] J. FAN AND I. GIJBELS, *Local Polynomial Modelling and Its Applications, Chapman & Hall, 1996.*

[7] K. B. HALVORSEN, *ElemStatLearn*, https://CRAN.R-project.org/package=ElemStatLearn.

[8] L. HAO Y D.Q. NAIMAN, *Quantile Regression, SAGE Publications, 1949.*

[9] IMDB, *Box Office Mojo*, "*Information courtesy of Box Office Mojo. Used with permission.*", https://www.boxofficemojo.com/intl/spain/, disponible en https://www.boxofficemojo.com.

[10] R. KOENKER, *Quantile Regression, Cambridge University Press, 2005.*

[11] R. KOENKER, *Quantile Regression: 40 Years On, Annual Review of Economics,* ***9****, (2017), 155–176*, https://doi.org/10.1146/annurev-economics-063016-103651.

[12] R. KOENKER, *Quantile Regression in R: a vignette, (2006)*, http://www.et.bs.ehu.es/cran/web/packages/quantreg/vignettes/rq.pdf.

[13] R.KOENKER, *quantreg: Quantile Regression*, https://CRAN.R-project.org/package=quantreg.

[14] R. KOENKER AND KEVIN F. HALLOCK, *Quantile Regression, Journal of Economic Perspectives,* ***15****, (4), (2001), 143–156*, http://www.econ.uiuc.edu/~roger/research/rq/QRJEP.pdf.

[15] J. DE LEEUW, K. HORNIK AND P. MAIR, *Isotone Optimization in R: Pool-Adjacent-Violators Algorithm (PAVA) and Active Set Methods*, https://pdfs.semanticscholar.org/6eb1/652e19db26a8356b6457431da0cf109707e9.pdf.

[16] S. PORTNOY AND R. KOENKER, *The Gaussian Hare and the Laplacian Tortoise: Computability of Squared-Error versus Absolute-Error Estimators, Statistical Science,* ***12****, (4), (1997), 279–300.*

[17] *The R Project for Statistical Computing*, `https://www.r-project.org/`.

[18] I. TAKEUCHI, Q.V. LE, T. SEARS AND A.J. SMOLA, *Nonparametric Quantile Regression, Journal of Machine Learning Research,* ***7****, (2006), 1231–1264.*

Apéndice A

Script R datos cine

```
#Se carga el paquete quantreg y el conjunto de datos
library(quantreg)

load("C:/Users/ANA MARTÍN/Desktop/TFG/Rstudio/boxofficeRQ.RData")
datos<- boxofficeRQ

#Resumen datos
summary(datos)
```

Modelo inicial

```
modelo0 <- rq(Weekend.Gross.2w ~Weekend.Gross.1w, tau=c(0.25,0.5 ,0.75), data=datos)

#IC de los coeficientes obtenidos con rq, por defecto con método se ="rank"
summary(modelo0)

#Se especifica el método se="nid" para obtener el p-valor
summary(modelo0, se ="nid")

modeloq<-rq(Weekend.Gross.2w ~Weekend.Gross.1w, tau=c(0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9),
data=datos)

#Gráfico de diagrama de dispersión con ajuste rq
x<-datos$Weekend.Gross.1w
y<-datos$Weekend.Gross.2w

plot(x, y, cex = 0.25, type = "n", xlab = "INGRESO BRUTO EN EUROS DE LA PRIMERA SEMANA",
ylab= "INGRESO BRUTO EN EUROS DE LA SEGUNDA SEMANA", main="Diagrama de dispersión")

points(x, y, cex=0.7, col = "red")

abline(rq(y~x, tau=0.5), col = "green")
abline(lm(y~x), lty = 2, col = "grey")

taus <- c(0.1,0.2,0.3,0.4,0.6,0.7,0.8,0.9)
for (i in 1:length(taus)){
  abline(rq(y~x, tau=taus[i]), col = "blue")
}

plot(summary(modeloq, se="nid"))

#Para ver si los coeficientes de dos o más ajustes son distintos
anova(modelo0)
```

Comparativa modelos

```
modelo1<- rq(Weekend.Gross.2w ~Weekend.Gross.1w, tau=0.2, data=datos)
modelo2<- rq(Weekend.Gross.2w ~Weekend.Gross.1w+ Theaters.1w, tau=0.2, data=datos)
modelo3<- rq(Weekend.Gross.2w ~Weekend.Gross.1w+ Theaters.2w, tau=0.2, data=datos)
modelo4<- rq(Weekend.Gross.2w ~Weekend.Gross.1w+ Theaters.1w+ Theaters.2w, tau=0.2,
data=datos)
modelo5<- rq(Weekend.Gross.2w ~Weekend.Gross.1w+I(Theaters.1w-Theaters.2w), tau=0.2,
data=datos)

anova(modelo1,modelo2)
anova(modelo1,modelo3)
anova(modelo2,modelo4)
anova(modelo3,modelo4)
anova(modelo1,modelo4)
anova(modelo1,modelo5)

#Medida bondad de ajuste

fitnull<-rq(Weekend.Gross.2w~1, tau=c(0.2, 0.4, 0.5, 0.6, 0.8), data=datos)

rho <- function(u,tau=.5)u*(tau - (u < 0))

fit1<-rq(Weekend.Gross.2w~Weekend.Gross.1w, tau=c(0.2, 0.4, 0.5, 0.6, 0.8), data=datos)
fit2<-rq(Weekend.Gross.2w~Weekend.Gross.1w+Theaters.1w, tau=c(0.2, 0.4, 0.5, 0.6, 0.8),
data=datos)
fit3<-rq(Weekend.Gross.2w~Weekend.Gross.1w+Theaters.2w, tau=c(0.2, 0.4, 0.5, 0.6, 0.8),
data=datos)
fit4<-rq(Weekend.Gross.2w~Weekend.Gross.1w+Theaters.1w+Theaters.2w,
tau=c(0.2, 0.4, 0.5, 0.6, 0.8), data=datos)
fit5<-rq(Weekend.Gross.2w~Weekend.Gross.1w+I(Theaters.1w-Theaters.2w),
tau=c(0.2, 0.4, 0.5, 0.6, 0.8), data=datos)

1-fit1$rho/fitnull$rho
1-fit2$rho/fitnull$rho
1-fit3$rho/fitnull$rho
1-fit4$rho/fitnull$rho
1-fit5$rho/fitnull$rho
```

Modelo en escala logarítmica

```
x<-Weekend.Gross.1w/Theaters.1w
y<-Weekend.Gross.2w/Theaters.2w

fitlog<-rq(log(y)~log(x), tau=seq(0.1, 0.9, by=0.1), data=datos)

Call:
rq(formula = log(y) ~ log(x), tau = seq(0.1, 0.9, by = 0.1),
    data = datos)

Coefficients:
             tau= 0.1  tau= 0.2   tau= 0.3   tau= 0.4   tau= 0.5
(Intercept) -1.520278 -1.235151 -0.8613901 -0.8016799 -0.7638842
log(x)       1.039942  1.031323  1.0109131  1.0107867  1.0132940
              tau= 0.6   tau= 0.7   tau= 0.8   tau= 0.9
(Intercept) -0.7961359 -0.6486331 -0.4751874 0.02131932
log(x)       1.0213459  1.0164143  1.0094060 0.98215686

Degrees of freedom: 268 total; 266 residual
```

```
#Diagrama de dispersión con las rectas de regresión cuantil ajustadas
scatterplot(log(y)~log(x), regLine=TRUE, smooth=list(span=0.5, spread=FALSE),
xlab="log(Weekend.Gross.1w/Theaters.1w)", ylab="log(Weekend.Gross.2w/Theaters.2w)",
main="Modelo en escala logarítmica", boxplots=FALSE, data=datos, subset=-c(16,131))

taus<-seq(0.1,0.9,by=0.1)
z<-rq(log(y)~log(x), tau= taus, data=datos, subset= -c(16, 131))

for( i in 1:length(taus)){
  abline(coef(z)[,i],col="gray")
}

#Coeficiente bondad de ajuste
1-fitlog$rho/fitnull$rho

anova(z)
```

Apéndice B

Script R datos bone

```
#Se cargan los paquetes y el fichero de datos
library(ElemStatLearn)
library(quantreg)

data("bone")

#Se ordena la edad para visualizar mejor los gráficos
sortbone<-bone[order(bone$age),]

hombres<-(sortbone$gender=="male")
mujeres<-(sortbone$gender=="female")  # mujeres<- !hombres
```

Gráfico de cajas

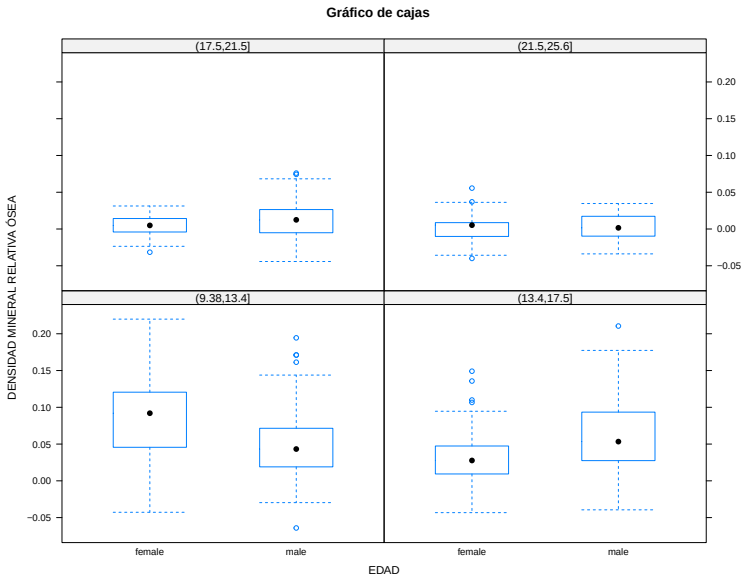

Figura B.1: Gráfico de cajas de la densidad mineral relativa ósea respecto la edad en cuatro intervalos por grupos según el género.

Gráfico de violín

```
require(lattice)

bwplot(spnbmd~gender|cut(age,4), panel=panel.violin,
       xlab="Edad", ylab="Densidad mineral relativa ósea",
       main="Gráfico de violín ", data=bone)
```

Comparación de modelos

```
modelo1<-rq(spnbmd~1, tau=0.2, data = bone)
modelo2<-rq(spnbmd~age, tau=0.2, data = bone)
modelo3<-rq(spnbmd~gender, tau=0.2, data = bone)
modelo4<-rq(spnbmd~age + gender, tau=0.2, data = bone) #aditivo
modelo5<-rq(spnbmd~age*gender, tau=0.2, data = bone) #interacción

anova(modelo1, modelo2)
anova(modelo1, modelo3)
anova(modelo2, modelo4)
anova(modelo3, modelo4)
anova(modelo2, modelo4)
anova(modelo2, modelo5)
```

Plot modelo paramétrico

```
plot(bone$age,bone$spnbmd,pch=10,col=as.numeric(bone$gender), cex=.4, xlab="Edad",
ylab="Densidad mineral relativa ósea", main="Modelo paramétrico (tau=0.5)")

lines(sortbone$age[hombres], fitted(rqpbone3)[order(bone$age)][hombres],
col=sortbone$gender[hombres])

lines(sortbone$age[mujeres], fitted(rqpbone3)[order(bone$age)][mujeres],
col=sortbone$gender[mujeres])

legend(x = "topright", legend = c("Hombres", "Mujeres"),lty =c(1,1),
col = c("red", "black"))
```

Plot modelo no paramétrico

```
taus=c(0.2, 0.4, 0.5, 0.6, 0.8)

plot(bone$spnbmd~bone$age, type="n", xlab="Edad", ylab="Densidad mineral relativa ósea",
main="Modelo no paramétrico")

points(bone$age, bone$spnbmd, col=bone$gender, data=bone, cex=0.4)

for (tauh in taus){
  fit.h<- lprq(sortbone$age[hombres], sortbone$spnbmd[hombres], h = 0.75, tau=tauh, m=200)
  fit.m<- lprq(sortbone$age[mujeres], sortbone$spnbmd[mujeres], h = 0.75, tau=tauh, m=200)
  lines(fit.h$xx, fit.h$fv, col=2)
  lines(fit.m$xx, fit.m$fv, col=1)
}

legend(x = "topright", legend = c("Hombres", "Mujeres"),lty =c(1,1),
col = c("red", "black"))
```

Modelo con splines

```
bsmax<-12 #Trabajamos con splines de aproximación
vAIC<-matrix(0, nrow=bsmax,ncol=2)

#Función V(\tau), es el valor de función de pérdida
rho <- function(u,tau=.5)u*(tau - (u < 0))

library(splines)
X.model<- model.matrix(bone$spnbmd~bs(bone$age, df=bsmax)*bone$gender)

taus=c(.2,.4,.5,.6,.8)
```

```
cq<-matrix(0,nrow=nrow(X.model), ncol=length(taus)) #matriz de cuantiles condicionales

plot(bone$spnbmd~bone$age, type="n", xlab="EDAD", ylab="DENSIDAD MINERAL RELATIVA ÓSEA",
main="Modelo con splines")

points(bone$age, bone$spnbmd, col=bone$gender, cex=0.4)

#Se construye el AIC para cada modelos para cada tau, con grado de libertad óptimo
for (tauh in taus ) {
  fit0<-rq(bone$spnbmd~1,tau=tauh, data=bone) #ajuste sin ningún tipo de variable
  for (i in 1:bsmax) { #bucle para cada grado de libertad (valor mínimo de i es 3)
    vAIC[i,]<-AIC(rq(bone$spnbmd~bs(bone$age, df=i)*bone$gender, tau=.4, data= bone))
  }
  dfaic<-which.min(vAIC[,1])#grados de libertad óptimo
  fit<- rq(bone$spnbmd~bs(bone$age, df=dfaic)*bone$gender, tau=tauh,
  data= bone) #modelo con interacción

  cq[, match(tauh,taus) ]<- fitted(fit) #cuantiles condicionales para cada edad
  X.model<- model.matrix(bone$spnbmd~bs(bone$age, df=dfaic)*bone$gender)
  spn.fit<- X.model[order(bone$age),]%*%fit$coef
  print(1-fit$rho/fit0$rho)  #R1 (es como el R^2)

  lines(sortbone$age[hombres], spn.fit[hombres], col=sortbone$gender[hombres])
  lines(sortbone$age[!hombres], spn.fit[!hombres], col=sortbone$gender[!hombres])

}
```

Isotonización

```
pcross<-which(apply(cq, 1, function (x) {any(order(x)!=1:length(as.vector(x))) }))

bone$age[pcross] #x en los que se producen los cortes

#Matriz de los cuantiles condicionales sin monotonía
cq[pcross,]

isoqr<- function(x,cq)
{
  n<-length(x) #longitud de x
  ncq<-0
  tcq<-0
  ncq<-dim(cq)[1]
  tcq<-dim(cq)[2]
  if(n!=ncq) {
    stop("dimensiones incompatibles") }
  else
  {
    cqiso<-cq
    for (i in 1:n)
    {
      cqx<-as.vector(cq[i, ]) #cogemos cada fila como un vector
      if(any(order(cqx) != 1:n)) #si el orden del vector no coincide
      con el orden natural de la permutacion 1 a n
        cqiso[i, ]<-isoreg(cqx)$yf   #se isotoniza el vector
    }
  }
  return(cqiso)
}
```

```
#Se aplica la función isoqr a los cuantiles condicionales donde hay cruces
cqiso<-isoqr(bone$age,cq)

#Matriz de cuantiles condicionales isotonizados
cqiso[pcross,]
```