Regresión de cuantiles
Una aplicación a la estimación de umbrales extremos móviles

Guillermo Anglés Fernando

Trabajo de fin de grado en Matemáticas
Universidad de Zaragoza

Directora del trabajo: Ana Carmen Cebrián Guajardo

28 de junio de 2017

Resumen

Regression analysis is a statistical process that allows us to estimate the relationship between two or more variables. It has many applications and is used in many fields, including climatology, medicine, economy or engineering. Regression helps us to understand how the dependent variable varies when changing the value of an independent variable, while the rest of independent varaibles are fixed. The most common regression model focuses on the mean of the dependent variable when the independent variables are fixed. This model can be formulated as follows:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{1}$$

where $\mathbf{Y}_{(n\times 1)}$ is the vector of responses, $\mathbf{X}_{(n\times(p+1))}$ is the regressor matrix, $\beta_{((p+1)\times 1)}$ is the vector of unknown parameters and $\varepsilon_{(n\times 1)}$ the vector of unknown errors, being n the number of observations and p the number of regressor variables. The β_i coefficient can be interpreted as the rate of change of the mean of the dependent variable distribution per unit change in the value of the i-th regressor while the rest of independent varaibles are fixed. This model works under some assumptions:

1. Linearity
2. Errors identically distributed
3. Incorrelated errors
4. Normality

The unknown parameters can be obtained by the least square method, based on minimazing the sum of squared errors. This paramater is linear, unbiased and of minimun variance due to the Gauss-Markov Theorem, under the assumptions mentioned before, except normality.

Such a function can be more or less complex, but it restricts exclusively on a specific location of the Y conditional distribution. Quantile regression, introduced by [7] extends this approach, allowing one to study the conditional distribution of $\mathbf{Y}$ on $\mathbf{X}$ at different quantiles in order to offer a global view of the relationship between the $\mathbf{Y}$ distribution and the $\mathbf{X}$ distribution.

Quantile regression has been used in a broad range of application settings. In pediatric medicine, quantile regression methods are used to estimate upper anfd lower quantile reference curves as function of age, sex and other variables being height or weight the response variable. This regression methods are also used in economy to study wages or discrimination effects.

The quantile regression model, for a given conditional quantile $\theta \in (0,1)$ can be formulated as follows:

$$Q_y(\theta|\mathbf{X}) = \mathbf{X}\beta(\theta), \tag{2}$$

where $0 < \theta < 1$, $Q_y(.|.)$ denotes the conditional quantile function for the θth quantile, $\mathbf{X}_{(n\times(p+1))}$ is the regressor matrix and $\beta_{((p+1)\times 1)}(\theta)$ the vector of unknown parameters for the generic conditional quantil θ, being n the number of observations and p the number of regressor variables. The β_i coefficient can be interpreted as the rate of change of the θ-th quantile of the dependent variable distribution per unit change in the value of the i-th regressor while the rest of independent varaibles are fixed.

This unknown parameters are obtained as the solution of the following minimization problem:

$$\min_{\beta} \sum_{i=1}^{n} \rho_\theta(y_i - x_i^T \beta(\theta)) \tag{3}$$

where $\rho_\theta(y) = [(1-\theta)I(y \leq 0) + \theta I(y > 0)]|y|$.

Under an empiric study made by [3] will see how quantile regression works in different situations, including those where least square method is not applicable.

When errors are independent and identically distributed the quantile regression estimator $\hat{\beta}(\theta)$ is asymptotically distributed as:

$$\sqrt{n}[\hat{\beta}(\theta) - \beta(\theta)] \to N(0, \omega^2(\theta)\mathbf{D}^{-1}), \tag{4}$$

being

$$\omega^2 = \frac{\theta(1-\theta)}{f(F^{-1}(\theta))^2} \qquad \mathbf{D} = \lim_{n\to\infty} \frac{1}{n} \sum_i \mathbf{x}_i^T \mathbf{x}_i \tag{5}$$

positive definite matrix.

Once the quantile regression model is obtained it is useful to study the influence that a specific variable has in the model. The posible influence that the variable X_i might have would dissapear if the coefficient β_i is zero. The statistical relevance is assessed through the usual Student-t statistic (Koenker 2005).

It is possible to study the influence of two or more variables at the same time. In this case the Likelihood Ratio test is used. It is asymptotically distributed as a χ^2 with degrees of freedom equal to the number of coefficients under test.

Resampling methods are simulation techniques used in statistic inference. From the observed data, they obtain new samples in order to estimate the distribution of an estimator or just do some type of inference. Bootstrap is a resampling method which consists on randomly sampling with replacement from the original dataset to produce new samples. The *xy*-pair bootstrap method will be introduced.

When errors are independent and no identically distributed the quantile regression estimator $\hat{\beta}(\theta)$ is asymptotically distributed as:

$$\sqrt{n}[\hat{\beta}(\theta) - \beta(\theta)] \to N(0, \theta(1-\theta)\mathbf{D}_1(\theta)^{-1}\mathbf{D}\mathbf{D}_1(\theta)^{-1}), \tag{6}$$

being $\mathbf{D}_1 = \lim_{n\to\infty} \frac{1}{n} \sum_i f_i(F^{-1}(\theta))\mathbf{x}_i^T \mathbf{x}_i$ positive definite matrix and $\mathbf{D} = \lim_{n\to\infty} \frac{1}{n} \sum_i \mathbf{x}_i^T \mathbf{x}_i$ positive definite matrix.

In the case of errors being identically distributed but not independent, defined by $e_i = ae_{i-1} + a_i$, with $|a| < 1$, and a_i variables independient and identically distributed, after purging serial correlation with the Prais-Winsten (PW) or Cochrane-Orcutt (OC) procedures, the quantile regression estimator $\hat{\beta}(\theta)$ is asymptotically distributed as:

$$\sqrt{n}[\hat{\beta}(\theta) - \beta(\theta)] \to N(0, \omega^2(\theta)\mathbf{A}^{-1}), \tag{7}$$

being $\mathbf{D} = \lim_{n\to\infty} \frac{1}{n} \sum_i \mathbf{x}_i^T \mathbf{x}_i$ positive definite matrix and $\mathbf{A} = \lim_{n\to\infty} \mathbf{D} + \frac{1}{n} \sum_i \varphi(e_i)\varphi(e_{i-1})(\mathbf{x}_i^T \mathbf{x}_{i-1} + \mathbf{x}_{i-1}^T \mathbf{x}_i)$ positive definite matrix. PW and CO procedures transform the model with dependent error terms into a model with erros independent and identically distributed in order to use the inference techniques for models with i.i.d errors.

Goodness of fit measures describe how the model fits to the original sample. This measure, callled *pseudo*R^2, is defined as:

$$pR^2 = 1 - \frac{RASW_\theta}{TASW_\theta}, \tag{8}$$

with

$$RASW_\theta = \sum_{y_i \geq \mathbf{x}_i^T \hat{\beta}} \theta|y_i - \mathbf{x}_i^T \hat{\beta}| + \sum_{y_i < \mathbf{x}_i^T \hat{\beta}} \theta|y_i - \mathbf{x}_i^T \hat{\beta}| \qquad TASW_\theta = \sum_{y_i \geq \theta} \theta|y_i - \hat{\theta}| + \sum_{y_i < \theta} (1-\theta)|y_i - \hat{\theta}| \tag{9}$$

The pR^2 ranges between 0 and 1. If pR^2 is close to 1 it is said that the fit for for a given quantile it is a good fit. It is worth noting that the index cannot be considered a measure of the goodness of fit of the whole model because it is related to a given quantile.

Finally, a practical aplicattion will be introduced related to heat waves since the analysis of heat waves is an increasingly important issue due to the serious impact of this phenomenon on ecosystems an human health. AEMET define heat waves when temperatutures register maximum values over the 95th percentile in a referenced period. This is why a regression model of the 95th percentile will be introduced in order to study it's evolution.

Índice general

Capítulo 1

Introducción a la regresión de cuantiles

En este capítulo se hará una introducción a la regresión lineal, destacando las hipótesis que este modelo necesita para poder llevarse a cabo. Posteriormente, se introducirá la regresión de cuantiles que requiere menos hipótesis y, usando conjuntos de datos con distintas características, se estudiarán las propiedades de la regresión de cuantiles.

1.1. Regresión

El análisis de regresión es un análisis estadístico que permite estimar la relación entre variables. A menudo resulta de interés predecir, en mayor o menor grado, valores de una variable a partir de otras e incluso cuantificar el efecto que una o varias variables pueden causar sobre otra. Son numerosas las aplicaciones de la regresión, y las hay en casi cualquier campo, incluyendo la climatología, cambios medio-ambientales, la ingeniería, la medicina, la economía y las ciencias sociales, entre otros.

El término regresión fue introducido por Francis Galton en el siglo XIX con el estudio de variables antropométricas al comparar la estatura de padres e hijos. Resultó que los hijos cuyos padres tenían una altura superior a la media tendían a igualarse a ésta y lo mismo ocurría con los hijos cuyos padres tenían una estatura inferior a la media, tendían a reducir esta distancia, es decir, *regresaban* al valor medio. La primera forma de estimación de la regresión lineal fue la de mínimos cuadrados. En 1801, Gauss utilizó el método de mínimos cuadrados para resolver un problema de astronomía, pero no fue hasta 1805 cuando Legendre lo publicó en su trabajo y posteriormente Gauss en 1809.

Los modelos de regresión son el conjunto de técnicas usadas para explorar y cuantificar la relación de dependencia entre una variable llamada variable dependiente o respuesta $\mathbf{Y}$ y una o más variables independientes llamadas variables predictoras o regresoras $\mathbf{X}_i$. Nos ayuda a entender como varía la variable dependiente al cambiar el valor de una de las variables independientes, manteniendo el resto de las variables independientes fijas. El modelo de regresión más habitual se centra en la estimación de la media condicional de la variable dependiente. Otros tipos de modelos de regresión se centran en otras medidas de localización como pueden ser los cuantiles.

1.2. Modelo de regresión lineal para la media

La regresión lineal múltiple se utiliza para establecer una relación lineal entre una variable dependiente $\mathbf{Y}$ y una o más variables independientes $\mathbf{X}_i$. Este modelo se expresa de la siguiente forma:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \tag{1.1}$$

donde $\mathbf{Y}_{(n\times 1)}$ es el vector respuesta, $\mathbf{X}_{(n\times(p+1))}$ es la matriz regresora, siendo la primer columna todo unos, $\beta_{((p+1)\times 1)}$ el vector de los coeficientes y $\varepsilon_{(n\times 1)}$ un vector de variables que representa los errores, siendo n el número de observaciones y p el número de variables independientes.

Los parámetros β_i con $i = 0, ..., p$ representan el cambio en media en $\mathbf{Y}$ al aumentar una unidad la variable $\mathbf{X}_i$ manteniendo fijas el resto de variables.

El modelo de regresión planteado impone las siguientes hipótesis:

1. **Linealidad.** La variable respuesta depende linealmente de las variables regresoras.
2. **Idéntica distribución de los errores ε_i.** Por lo tanto, la varianza de los errores ε_i es igual para todos ellos, es decir, presentan **homocedasticidad**.
3. **Incorrelación de los errores.**
4. **Normalidad de los errores.**

   Es decir, los errores ε_i son variables aleatorias independientes con $\varepsilon_i \sim N(0, \sigma)$.

A partir del modelo expresado en (1.1) se puede obtener que

$$E[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta$$

Por ejemplo, supongamos que se quiere estudiar si las variables peso, altura, edad, sexo, fumador y bebedor, que son las variables regresoras, influyen en el pulso de un individuo tras haber realizado una actividad física (variable dependiente). La relación entre la variable dependiente y el resto de variables se puede expresar como:

$$Pulso_i = \beta_0 + \beta_1 Peso_i + \beta_2 Altura_i + \beta_3 Edad_i + \beta_4 Sexo_i + \beta_5 Fumador_i + \beta_6 Bebedor_i + \varepsilon_i$$

1.2.1. Estimación del modelo

Tanto el método de Máxima Verosimilitud como el de mínimos cuadrados son los más utilizados a la hora de estimar el parámetro β.

Método de Máxima Verosimilitud

Se considera el modelo con una única variable regresora:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma) \qquad i = 1, ..., n \tag{1.2}$$

Como cada y_i se distribuye como $N(\beta_0 + \beta_1 x_i, \sigma)$ su función de densidad es:

$$f_i = \frac{exp\{\frac{-(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\}}{\sqrt{2\pi\sigma^2}}$$

Dado que las y_i son independientes la función de densidad conjunta de $y_1, ..., y_n$ será el producto de las funciones de densidad para cada y_i. A esta función se le llama función de verosimilitud y depende de los parámetros β_j y σ:

$$L(\beta_0, \beta_1, \sigma, y_1, ..., y_n) = \prod_{i=1}^{n} \frac{exp\{\frac{-(y_i - (\beta_0 + \beta_1 x_i))^2}{2\sigma^2}\}}{\sqrt{2\pi\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{n/2}} exp\{-\frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{\sigma^2}\} \tag{1.3}$$

El método de Máxima Verosimilitud consiste en maximizar la función de verosimilitud L. Como la función logaritmo es creciente es equivalente maximizar la siguiente función:

$$l(\beta_0, ..., \beta_p, \sigma, y_1, ..., y_n) = logL = -\frac{n}{2} log(2\pi) - nlog\sigma - \frac{1}{2}\sum_{i=1}^{n} \frac{(y_i - (\beta_0 + \beta_1 x_i))^2}{\sigma^2} \tag{1.4}$$

Considerando p variables regresoras:

$$l(\beta_0,...,\beta_p,\sigma,y_1,...,y_n) = -\frac{n}{2}log(2\pi) - nlog\sigma - \frac{1}{2}(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta) \tag{1.5}$$

Maximizar esta función con respecto a β es equivalente a minimizar $(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)$.
Derivando respecto β e igualando a 0

$$\frac{\partial l(\beta)}{\partial \beta} = 2(\mathbf{X}^T\mathbf{X})\beta - 2\mathbf{X}^T\mathbf{Y} = 0 \tag{1.6}$$

se obtiene

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Los estimadores Máximo Verosímiles son invaraintes, consistentes, asintóticamente insesgados y asintóticamente normales.

Método de Mínimos cuadrados

Otro procedimiento utilizado para estimar el vector β es el método de mínimos cuadrados (MCO). El método de mínimos cuadrados tiene por objetivo minimizar la suma de los errores al cuadrado. Encontrar un vector β que minimice la siguiente expresión:

$$(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta) \tag{1.7}$$

Se observa que hay que minimizar la misma expresión que para el método de Máxima Verosimilitud. Luego el estimador de β obtenido por mínimos cuadrados coincide con el estimador máximo verosímil bajo la hipótesis de normalidad.

El teorema de Gauss-Markov asegura que el estimador de β obtenido por el método de mínimos cuadrados es lineal, insesgado y de mínima varianza bajo las hipótesis mencionadas anteriormente sin imponer la hipótesis de normalidad.

La regresión por mínimos cuadrados es uno de los métodos más utilizados debido a su facilidad de cálculo. Sin embargo, las hipótesis necesarias para su aplicación se incumplen en algunas ocasiones. La presencia de heterocedasticidad, errores dependientes y presencia de datos atípicos dan lugar a estos incumplimientos.

Medida de bondad de ajuste

Para medir la bondad del ajuste se utiliza el coeficiente estadístico R^2 que es la proporción de la varianza explicada (VE) entre la varianza total (VT):

$$R^2 = \frac{VE}{VT} \tag{1.8}$$

siendo

$$VE = \sum_{i=1}^{n}(\hat{y}_i - \overline{\mathbf{Y}})^2 \qquad\qquad VT = \sum_{i=1}^{n}(y_i - \overline{\mathbf{Y}})^2$$

Este coeficiente determina qué porcentaje de variación de la variable dependiente es explicado por el modelo de regresión. Se cumple $0 < R^2 < 1$ y cuanto más próximo a 1 mejor es el ajuste aunque no implica que se verifiquen las hipótesis del modelo.

1.3. Regresión de cuantiles

La regresión de cuantiles fue introducida por [7] como una extensión de la estimación de la media condicional a través del método de mínimos cuadrados a una estimación de los cuantiles condicionales de la variable respuesta.

La regresión de cuantiles se utiliza en muchos campos. Las curvas de crecimiento de altura y peso de niños tienen una gran importancia en pediatría; este método se usa para estudiar cuantiles grandes y pequeños como función del sexo, la edad y otras variables. También se utiliza mucho en economía para estudiar salarios, efectos discriminantes y tendencias de desigualdad en ingresos. Para modelizar los resultados de alumnos de colegios públicos en exámenes estándar en función de características socio-económicas (ingresos de los padres, nivel educativo, tamaño de la clase, profesorado cualificado, gastos efectuados en enseñanza...) se ha utilizado la regresión de cuantiles.

La regresión por mínimos cuadrados se restringe exclusivamente a una ubicación específica de la distribución condicional $\mathbf{Y}$, la media condicional. La regresión de cuantiles extiende este enfoque permitiendo estudiar la distribución condicional $\mathbf{Y}$ en diferentes localizaciones, diferentes cuantiles, como puede ser, por ejemplo, la mediana y ofrece, de esta forma, una visión global de las relación de las distribuciones de $\mathbf{Y}$ y de $\mathbf{X}$. Este tipo de regresión permite estimar cualquier cuantil, pudiendo valorar así lo que ocurre con valores extremos de la población y detectar de esta forma los distintos efectos de la distribución y ver lo que es importante a estudiar y por qué.

Dada Y una variable aleatoria y $\theta \in (0,1)$ el θ-ésimo cuantil es el valor y tal que $P(Y \leq y) = \theta$. Dada la función de distribución:

$$F_y(y) = P(Y \leq y), \tag{1.9}$$

la función cuantílica se define como su inversa:

$$Q_y(\theta) = F_y^{-1}(\theta) = \text{ínf}[y : F_y(y) > \theta] \tag{1.10}$$

El modelo de regresión de cuantiles se puede formular de la siguiente forma:

$$Q_y(\theta|\mathbf{X}) = \mathbf{X}\beta(\theta), \tag{1.11}$$

donde $0 < \theta < 1$, $Q_y(.|.)$ es la función cuantil condicionada, $\mathbf{X}_{(n\times(p+1))}$ es la matriz regresora y $\beta_{((p+1)\times 1)}(\theta)$ es el vector de los coeficientes desconocidos para el cuantil θ, siendo n el número de observaciones y p el número de variables independientes.

1.3.1. Estimación por mínimos errores absolutos

Es posible definir un método de estimación para los cuantiles, denominado de mínimos errores absolutos. Este método se reduce a resolver un problema de optimización:

$$\hat{q}_\theta = \min_c E[\rho_\theta(Y-c)] \tag{1.12}$$

donde $\rho_\theta(.)$ denota la siguiente función de peso:

$$\rho_\theta(y) = \begin{cases} \theta y & si \quad y \geq 0 \\ (\theta-1)y & si \quad y < 0 \end{cases} \tag{1.13}$$

La función $\rho_\theta(.)$ es continua, lineal por segmentos, asimétrica si $\theta \neq 0,5$ y no diferenciable en $y = 0$.

En el caso de una variable discreta Y con función de probabilidad $p(y) = P(Y = y)$, la expresión (1.12) queda:

$$\hat{q}_\theta = \text{mín}_c E[\rho_\theta(Y-c)] = \text{mín}_c \left\{ (1-\theta)\sum_{y\leq c} |y-c|p(y) + \theta \sum_{y>c} |y-c|p(y) \right\}$$

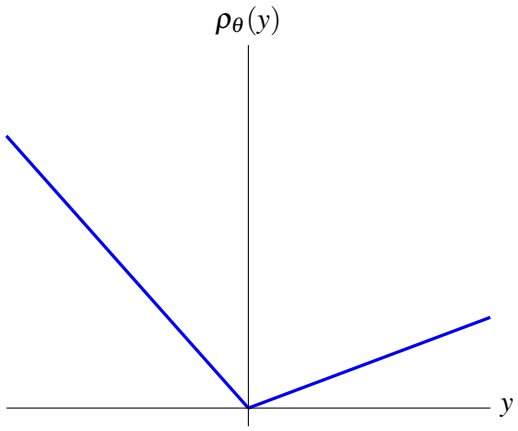

Figura 1.1: Gráfica de la función $\rho_\theta(y)$ para $\theta = 0{,}25$

Si se trata de una variable aleatoria continua con función de densidad f(y):

$$\hat{q}_\theta = \text{mín}_c\, E[\rho_\theta(Y-c)] = \text{mín}_c \left\{ (1-\theta) \int_{-\infty}^{c} |y-c| f(y)\, dy + \theta \int_{c}^{+\infty} |y-c| f(y)\, dy \right\}$$

La estimación $\hat{\beta}(\theta)$ de $\beta(\theta)$ no requiere condiciones y se obtiene como la solución del siguiente problema de minimización:

$$\min_{\beta} \sum_{i=1}^{n} \rho_\theta(y_i - x_i^T \beta(\theta)) \tag{1.14}$$

Los parámetros $\beta_i(\theta)$ con $i = 0, ..., p$ se interpretan como el cambio en el θ-ésimo cuantil de la variable dependiente $\mathbf{Y}$ al cambiar en una unidad la variable regresora $\mathbf{X}_i$, con $i = 0, ..., p$, manteniendo las otras fijas.

1.3.2. Cómo funciona la regresión de cuantiles

Una de las ventajas de la regresión de cuantiles es que requiere muy pocas condiciones para que se pueda aplicar. Mediante la realización de un estudio empírico basado en simulación realizado por [3] se va a ver cómo funciona el modelo de regresión de cuantiles en distintas situaciones, incluidas aquellas donde la regresión por mínimos cuadrados no se puede aplicar. Se van a utilizar los siguientes modelos para simular conjuntos de datos con distintas características y a los que se les aplicará tanto la regresión de cuantiles como la de mínimos cuadrados:

1. **Modelo normal:** $y = 1 + 2x + e_N$ con $e_N \sim N(0,1)$ incorrelados

   Se caracteriza por cumplir las hipótesis para poder utilizar el método de mínimos cuadrados.

2. **Modelo heterocedástico:** $y = 1 + 2x + (1+x)e_N$ con $e_N \sim N(0,1)$ incorrelados

   No cumple la hipótesis de homocedasticidad; la varianza no es constante a lo largo de la distribución.

3. **Modelo no normal:** $y = 1 + 2x + e_{LN}$ con $e_{LN} \sim LN(0,0{,}5)$ incorrelados siendo LN la distribución logNormal.

   En este caso no se cumple la hipótesis de normalidad y además la distribución del error es asimétrica.

4. **Modelo dependiente:** $y = 1 + 2x + e_D$

   Este modelo no cumple la hipótesis de incorrelación; los términos del error dependen del anterior, es decir, $e_i = -0{,}2e_i + a_i$ siendo $a_i \sim N(0,1)$ incorrelados.

En primer lugar, para poder apreciar los beneficios que ofrece la regresión de cuantiles, se van a comparar los coeficientes estimados para el *modelo normal* y para el *modelo heterocedástico*. La representación de la recta de mínimos cuadrados así como las rectas de la regresión de cuantiles para los valores de $\theta = \{0{,}05,\ 0{,}1,\ 0{,}25,\ 0{,}5,\ 0{,}75,\ 0{,}9,\ 0{,}95\}$ se representan en la *Figura 1.2*.

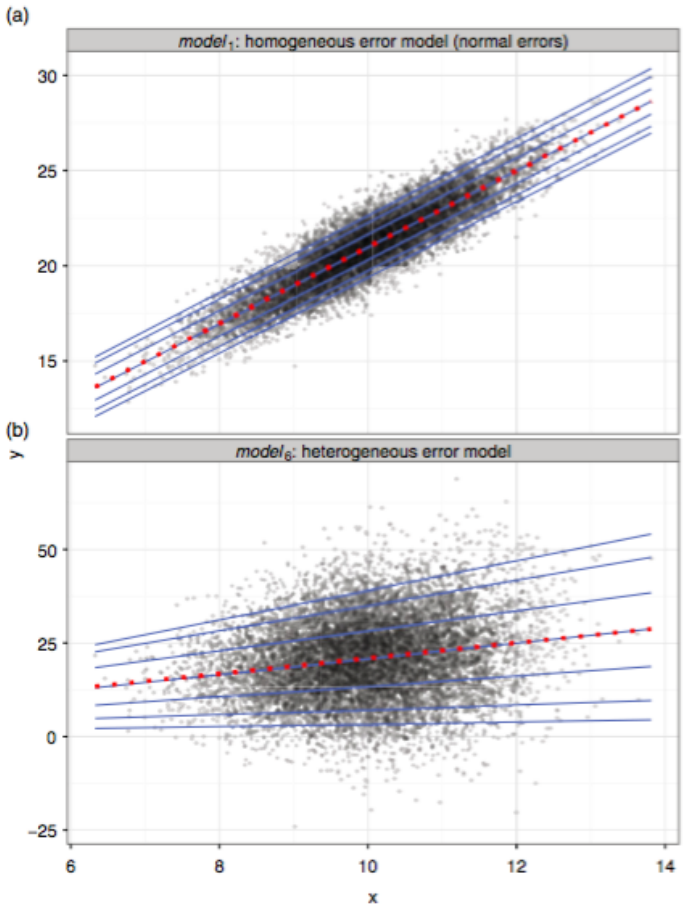

Figura 1.2: Representación de la nube de puntos obtenida mediante el *modelo normal* (figura a) y el *modelo heterocedástico* (figura b). La línea de puntos corresponde a la recta obtenida mediante el MCO y las líneas sólidas corresponden a las rectas de la regresión de cuantiles para los valores de $\theta = \{0{,}05,\ 0{,}1,\ 0{,}25,\ 0{,}5,\ 0{,}75,\ 0{,}9,\ 0{,}95\}$. Imagen obtenida de [3]

En el *modelo normal* se puede apreciar, al igual que ocurre en una distribución normal estándar donde la media coincide con la mediana, la estimación de la media coincide con la estimación de la mediana. Los coeficientes estimados para las pendientes de las rectas de regresión de cuantiles son todos iguales, se diferencian en los estimadores de β_0. La pendiente estimada para la media en la regresión por mínimos cuadrados es igual a la pendiente estimada para los cuantiles en la regresion de cuantiles.

Si se produce un cambio en media y en varianza en la distribución de $\mathbf{Y}$, la pendiente estimada para los cuantiles en la regresion de cuantiles no es iguale para todos ellos como ocurre en el *modelo heterocedástico* donde se aprecia el no paralelismo entre las rectas obtenidas. En este caso el modelo por mínimos cuadrados proporciona una relación incompleta debido a que sólo se centra en cambios de la media condicional.

Hay que resaltar la importancia de la regresión de cuantiles a la hora de describir toda la distribución condicional de la variable dependiente ofreciendo una visión más completa de ésta. Permite modelar cambios en la variable respuesta en múltiples puntos de la distribución.

Intervalos de predicción para los valores de la respuesta

La regresión de cuantiles permite obtener intervalos de predicción de los valores de la variable respuesta.

El intervalo proporcionado por dos cuantiles, $q_Y(\theta_1|X=x)$ y $q_Y(\theta_2|X=x)$ para un valor de x, es el intervalo donde se encuentran el $(\theta_2-\theta_1)\,\%$ de los valores de la variable respuesta.

[3], mediante una simulación, compararon los intervalos obtenidos a partir de la regresión de cuantiles con los intervalos reales .

En la simulación se utilizaron los cuatro modelos introducidos anteriormente. Para cada modelo se simularon 1000 conjuntos de datos distintos y para cada conjunto de datos se obtuvo el intervalo correspondiente al 80 % de los datos centrales usando tanto el método de mínimos cuadrados como el método de regresión de cuantiles. Los intervalos han sido calculados para dos valores de x, $x=8$ y $x=11$. En la Tabla (1.1) se muestra el porcentaje de veces que los intervalos de predicción simulados cubren el intervalo real de la población. Este valor debería ser próximo al nivel de confianza de los intervalos calculados.

		$x=8$	$x=11$
modelo normal	MCO	81.6	78.7
	RC	80.3	78.6
modelo no normal	MCO	56.2	54.1
	RC	80.2	79.3
modelo heterocedástico	MCO	26.5	95.3
	RC	80.0	79.4
modelo dependiente	MCO	80.1	76.2
	RC	78.5	79.2

Tabla (1.1): Cada celda de la tabla muestra el porcentaje de veces que el intervalo estimado cubre al intervalo real. Véase [3].

Se puede observar la consistencia del nivel de cubrimiento de los intervalos obtenidos a partir de la regresión de cuantiles y en cuanto a los obtenidos por la regresión por mínimos cuadrados la ausencia de consitencia es más que notable en el momento en el que no se tiene normalidad en los errores.

Cómo seleccionar los cuantiles a estimar

Es importante elegir bien qué cuantiles se van a estimar para poder caracterizar de la forma más completa posible la distribución de la variable respuesta.

Una opción es escoger una gran cantidad de cuantiles igualmente espaciados en el intervalo $(0,1)$ para obtener una aproximación lo más precisa posible. En el caso de que se aprecie un problema en una zona específica se realizará un estudio más preciso de la zona en cuestión introduciendo más cuantiles para obtener una mejor visión de la distribución.

El número de cuantiles que se pueden calcular está relacionado con en el tamaño de la muestra. A medida que aumenta el tamaño de la muestra el número de cuantiles que se pueden calcular aumenta, ver [3].

Capítulo 2

Inferencia sobre los coeficientes del modelo de regresión de cuantiles

En este capítulo se presenta la distribución del estimador $\hat{\beta}(\theta)$ de $\beta(\theta)$ obtenido mediante mínimos errores absolutos en el caso de que los los errores sean independientes e idénticamente distribuidos y en el caso de que sean independientes y no idénticamente distribuidos. También se comentarán distintos test para poder realizar contrastes de hipótesis en relación a los coeficientes $\beta_i(\theta)$. Se introducirán métodos de remuestreo (bootstrap) y se estudiará como realizar inferencia en los modelo con errores dependientes. Por último, se presentarán medidas para estudiar la bondad del ajuste.

2.1. Modelos con errores independientes e idénticamente distribuidos

Cuando los errores son independientes e idénticamente distribuidos el estimador $\hat{\beta}(\theta)$ de la regresión de cuantiles se distribuye asintóticamente de la siguiente froma:

$$\sqrt{n}[\hat{\beta}(\theta) - \beta(\theta)] \to N(0, \omega^2(\theta)\mathbf{D}^{-1}), \tag{2.1}$$

siendo

$$\omega^2 = \frac{\theta(1-\theta)}{f(F^{-1}(\theta))^2} \qquad \mathbf{D} = \lim_{n\to\infty} \frac{1}{n} \sum_i \mathbf{x}_i^T \mathbf{x}_i \tag{2.2}$$

una matriz definida positiva, $\mathbf{x}_i$ es el vector fila de dimensión $1 \times p$ compuesto por las i-ésimas observaciones de cada una de las p variables. La demostración de este resultado se encuentra en [7].

El término $f(F^{-1}(\theta))$ es desconocido y debe ser estimado. [9] propuso un estimador para este término.

2.1.1. Contraste sobre el valor de $\beta_i(\theta)$

En esta sección se van a considerar los contrastes de hipótesis necesarios para estudiar la influencia de cada una de las variables regresoras $\mathbf{X}_i$. Se plantea la siguiente pregunta: ¿Se dispone de suficiente evidencia muestral para afirmar que la variable $\mathbf{X}_i$ tiene una influencia significativa sobre el cuantil θ de la variable dependiente $\mathbf{Y}$? La posible influencia de $\mathbf{X}_i$ desaparece si su coeficiente $\beta_i(\theta)$ se anula. Para ello se considera el siguiente contraste de hipótesis:

$$H_0 : \beta_i(\theta) = 0$$
$$H_1 : \beta_i(\theta) \neq 0$$

La relevancia estadísitica de estos coeficientes se estudia mediante el test t al igual que se hace con la regresión por mínimos cuadrados, véase [5]. Como el estimador $\hat{\beta}_i(\theta)$ es asintóticamente normal estandarizarlo con su error estándar estimado en lugar del verdadero error estándar, que se desconoce, da lugar a un a una distribución t-Student. Por lo tanto, el estadístico t es el siguiente:

$$t = \frac{\hat{\beta}_i(\theta)}{se(\hat{\beta}_i(\theta))}$$

que bajo la hipótesis nula tiene una distribución t_{n-p-1} Basándose en dicho estadístico, dado un nivel de significación α, la región crítica del test es $RC = \{\mathbf{X}||t| > t_{n-p-1,\alpha/2}\}$. El p-valor $= P(t_{n-p-1} > |t| \cup t_{n-p-1} < -|t|)$ corresponde al nivel de significación más pequeño posible para el que se rechaza la hipótesis nula.

2.1.2. Contraste para un coeficiente en distintos cuantiles

Es de gran interés también comparar un coeficiente para los modelos de diferentes cuantiles. Se va a considerar el modelo más simple, con una variable regresora, que es extensible a modelos más complejos, con más de una variable regresora. La diferencia entre los cuantiles de orden θ_1 y θ_2 es:

$$Q_y(\theta_2|\mathbf{X}) - Q_y(\theta_1|\mathbf{X}) = [\beta_0(\theta_2) - \beta_0(\theta_1)] + [\beta_1(\theta_2) - \beta_1(\theta_1)]X = \gamma - \delta X \tag{2.3}$$

El coeficiente $\gamma = \beta_0(\theta_2) - \beta_0(\theta_1)$ mide la diferencia de los coeficientes β_0 de las rectas calculadas y el coeficiente $\delta = \beta_1(\theta_2) - \beta_1(\theta_1)$ mide la diferencia entre las pendientes de las dos rectas de los cuantiles seleccionados. En cuanto a la interpretación de γ, un cambio en los coeficientes β_0 no es de extrañar debido a que al cambiar de cuantil la recta estimada se mueve a lo largo del eje vertical. La inferencia sobre el coeficiente δ es de mayor utilidad ya que permite estudiar el paralelismo entre las rectas. La relevancia estadística de la diferencia entre las pendientes estimadas en las distintas rectas, estimada por $\hat{\delta}$, se puede cuantificar con la técnica de inferencia presentada anteriormente, el test-t. Para ello se considera el siguiente contraste de hipótesis:

$$H_0 : \delta = 0$$
$$H_1 : \delta \neq 0$$

Cuando la hipótesis nula $H_0 : \delta = 0$ es rechazada la diferencia entre las pendientes estimadas es estadísticamente diferente de 0, las dos rectas estimadas no son paralelas. Por otro lado, si la hipótesis nula es aceptada esta diferencia es estadísticamente 0 y las rectas estimadas se pueden considerar paralelas.

2.1.3. Test de Razón de Verosimilitudes

Una hipótesis de interés en un modelo de regresión es estudiar si son necesarias todas las variables regresoras o, si por el contrario, hay algunas de ellas que se pueden excluir dando lugar a un modelo más simple. Con el test t ya se ha estudiado la exclusión de una única variable mientras que ahora se va a proceder a la exclusión simultánea de dos o más variables regresoras. Para ello se plantea el siguiente contraste de hipótesis:

$$H_0 : \beta_i(\theta) = 0 \text{ para algunos } i \in (0, ..., p)$$
$$H_1 : \beta_i(\theta) \neq 0 \text{ para algún } i \in (0, ..., p)$$

Para estudiar este contraste de hipótesis se utilizará el test de Razón de Verosimilitudes (RV).
En primer lugar, se define la función objetivo como:

$$V(\theta) = \sum_{i=1}^{n} \rho_\theta(y_i - x_i^T \hat{\beta}(\theta)) \tag{2.4}$$

El estadístico para el test RV es:

$$L = 2\omega^{-1}(\tilde{V}(\theta) - \hat{V}(\theta)) \tag{2.5}$$

siendo $\tilde{V}(\theta)$ la función objetivo del modelo suponiendo la hipótesis nula cierta, conocido como el modelo restringido , y $\hat{V}(\theta)$ la función objetivo del modelo bajo H_1, el modelo no restringido y

$\omega^2 = \frac{\theta(1-\theta)}{f(F^{-1}(\theta))^2}$. Bajo la hipótesis nula este estadístico se distribuye asintóticamnete como χ^2_n siendo n la diferencia entre el número de parámetros en el modelo no restringido y el número de parámetros en el modelo no restringido, véase [5].

A partir de este estadístico, la región crítica dado un nivel de significación α, queda definida como $RC = \{\mathbf{X}|L \geq \chi^2_{n,\alpha}\}$. Si los coeficiente β_i no son nulos, la diferencia $(\tilde{V}(\theta) - \hat{V}(\theta))$ es estadísticamente relevante. Entonces, dado un nivel de significación α, el correspondiente p-valor $= P(\chi^2_n > L)$ será igual o menor que α y la hipótesis nula será rechazada. En caso contrario, si los coeficientes son nulos, las funciones objetivo son próximas la una a la otra, el p-valor será mayor que α y la hipótesis nula no será rechazada.

Otros test que se pueden utilizar para llevar a cabo este tipo de hipótesis múltiple son el test Wald y el test del Multiplicador de Lagrange, véase [9].

2.1.4. Métodos de remuestreo

Los métodos basados en el remuestreo son técnicas de simulación empleadas en inferencia estadística. A partir de los datos observados, se simulan nuevas muestras con el fin de estimar la distribución de un estimador o realizar algún tipo de inferencia.

Las técnicas de bootstrap consisten en seleccionar muestras con reemplazamiento a partir de la original. Para cada muestra se calcula el estimador de interés y se obtiene una muestra de los estimadores. A partir de esta muestra se puede calcular la varianza muestral, construir intervalos de confianza y realizar contrastes de hipótesis.

Método del par xy

En esta sección se introduce el método bootstrap del par *xy* aunque se pueden utilizar otros como el método pwy, véase [9]. Se considera el modelo de regresión de cuantiles más simple, con una única variable regresora y n observaciones:

$$Q_y(\theta|\mathbf{X}) = \hat{\beta}_0(\theta) + \hat{\beta}_1(\theta)\mathbf{x}$$

El método del par xy consiste en la construcción de un número de muestras (B) del mismo tamaño que la muestra inicial, donde cada muestra se obtiene tomando aleatoriamente y con remplazamiento elementos (x_i, y_i) de la muestra original. Para cada cuantil de interés se estiman B modelos de regresión de cuantiles, uno con cada muestra y se obtiene un vector con los coeficientes estimados $\hat{\beta}_b(\theta)$ con $b = 1, ..., B$. Cualquier estadístico de interés se puede calcular a partir de la muestra de los estimadores de los coeficientes obtenidos mediante bootstrap, como la media:

$$\overline{\hat{\beta}}(\theta) = \frac{1}{B}\sum_{b=1}^{\mathbf{B}} \hat{\beta}_b(\theta) \tag{2.6}$$

Por ejemplo, la desviación típica de la muestra de estimadores de los coeficientes $\hat{\beta}_b(\theta)$ permite estimar el error estándar de $\hat{\beta}(\theta)$, útil para obtener intervalos de confianza y realizar contrastes de hipótesis.

Si se consideran k cuantiles lo que se obtiene es una matriz de dimensión $B \times k$:

$$\begin{pmatrix} \hat{\beta}_{1j}(\theta_1) & \dots & \hat{\beta}_{1j}(\theta_k) \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \hat{\beta}_{Bj}(\theta_1) & \dots & \hat{\beta}_{Bj}(\theta_k) \end{pmatrix}$$

donde, para un j fijo, la q-ésima columna representa la muestra de los estimadores del coeficiente j-ésimo para el cuantiles θ_q con $q = 1,...,k$.

A partir de esta matriz se puede obtener la matriz de varianzas y covarianzas de $\hat{\beta}(\theta_1)$,...,$\hat{\beta}(\theta_q)$ donde los elementos (i,i) de la matriz son $Var(\hat{\beta}(\theta_i))$, la varianza muestral de los estimadores de los coeficientes obtenidos mediante bootstrap y los elementos (i,j) con $i \neq j$ son $Cov(\hat{\beta}(\theta_i), \hat{\beta}(\theta_j))$, la covarianza muestral de los estimadores de los coeficientes obtenidos mediante bootstrap.

En el caso de regresión de cuantiles múltiple con p variables regresoras, la varianza muestral de los estimadores de los coeficientes obtenidos mediante bootstrap para el coeficiente j-ésimo y cuantil θ_q es:

$$\hat{V}(\hat{\beta}_j(\theta_q)) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\beta}_{bj}(\theta_q) - \overline{\hat{\beta}_j}(\theta_q))(\hat{\beta}_{bj}(\theta_q) - \overline{\hat{\beta}_j}(\theta_q))^T, \tag{2.7}$$

donde $j = 1,...,p$, $q = 1,...,k$ y $\overline{\hat{\beta}_j}(\theta_q) = \frac{1}{\mathbf{B}}\sum_{b=1}^{B}\hat{\beta}_{bj}(\theta_q)$.

Para cada coeficiente de la regresión de cuantiles se puede obtener su intervalo de confianza. El intervalo de confianza para el j-ésimo coeficiente y el q-ésimo cuantil es:

$$\overline{\hat{\beta}_j}(\theta_q) \pm z_{\alpha/2} SD(\hat{\beta}_i(\theta_q)), \tag{2.8}$$

donde $\overline{\hat{\beta}}(\theta_q)$ es la media muestral de los estimadores del coeficiente obtenido mediante bootstrap y $SD(\hat{\beta}_j(\theta_q))$ es la raíz cuadrada de la varianza $\hat{V}(\hat{\beta}_j(\theta_q))$ vista en (2.7).

2.2. Modelo con errores independientes y no idénticamente distribuidos

Los errores independientes y no idénticamente distribuidos implican una función de densidad f_i del error diferente para cada ε_i. En este caso, el estimador obtenido a partir de la regresión de cuantiles se distribuye asintóticamente de la siguiente manera:

$$\sqrt{n}[\hat{\beta}(\theta) - \beta(\theta)] \to N(0, \theta(1-\theta)\mathbf{D}_1(\theta)^{-1}\mathbf{D}\mathbf{D}_1(\theta)^{-1}), \tag{2.9}$$

siendo $\mathbf{D}_1 = \lim_{n\to\infty}\frac{1}{n}\sum_i f_i(F^{-1}(\theta))\mathbf{x}_i^T\mathbf{x}_i$ matriz definida positiva, $\mathbf{D} = \lim_{n\to\infty}\frac{1}{n}\sum_i \mathbf{x}_i^T\mathbf{x}_i$ matriz definida positiva y $\mathbf{x}_i$ es el vector fila de dimensión $1 \times p$ compuesto por las i-ésimas observaciones de cada una de las p variables, ver [8].

Nótese que cuando los errores son independientes e idénticamente distribuidos esta expresión se reduce a la expresión en (2.1) ya que la densidad f_i es igual para todos los errores, así $f_i = f$. Al ser $f(F^{-1}(\theta))$ constante $\mathbf{D}_1 = f(F^{-1}(\theta))\mathbf{D}$. De esta forma la matriz de covarianza queda:

$$\theta(1-\theta)\mathbf{D}_1(\theta)^{-1}\mathbf{D}\mathbf{D}_1(\theta)^{-1} = \frac{\theta(1-\theta)}{f(F^{-1}(\theta))^2}\mathbf{D}(\theta)^{-1}\mathbf{D}\mathbf{D}(\theta)^{-1} = \omega^2\mathbf{D}$$

2.3. Modelo con errores dependientes

En esta sección se presenta la inferencia en modelos con errores dependientes. En particular, se asume que los errores están correlacionados porque el valor de la variable dependiente está influenciado por sus valores pasados. Para simplificar notación, aunque es extensible a modelos más complejos, se considera el modelo más simple con una única variable regresora, $y_i = \beta_0(\theta) + \beta_1(\theta)x_i + e_i$ con los términos del error dependientes e idénticamente distribuidos definidos como $e_i = ae_{i-1} + a_i$, siendo $|a| < 1$, y a_i variables aleatorias independientes e idénticamente distribuidas. La condición $|a| < 1$ implica que los errores pasados influyen en y_i pero su influencia disminuye con el paso del tiempo.

El estimador de los coeficientes de la regresión de cuantiles de la mediana con $e_i = ae_{i-1} + a_i$ se distribuye asintóticamente de la siguiente manera:

$$\sqrt{n}[\hat{\beta}(\theta)-\beta(\theta)] \to N(0,\omega^2(\theta)\mathbf{D}^{-1}\mathbf{A}\mathbf{D}), \tag{2.10}$$

siendo $\mathbf{D}=\lim_{n\to\infty}\frac{1}{n}\sum_i \mathbf{x}_i^T\mathbf{x}_i$ matriz definida positiva, $\mathbf{A}=\lim_{n\to\infty}\mathbf{D}+\frac{1}{n}\sum_i \varphi(e_i)\varphi(e_{i-1})(\mathbf{x}_i^T\mathbf{x}_{i-1}+\mathbf{x}_{i-1}^T\mathbf{x}_i)$ matriz definida positiva, $\mathbf{x}_i$ es el vector fila de dimensión $1\times p$ compuesto por las i-ésimas observaciones de cada una de las p variables y la función $\varphi(.)$ es la derivada de la función objetivo. La demostración de este resultado se encuentra en [10].

Ignorar la correlación serial implica una incorrecta estimación del término $\omega^2(\theta)\mathbf{D}^{-1}\mathbf{A}\mathbf{D}$ por el término $\omega^2(\theta)\mathbf{D}^{-1}$.

Los siguientes resultados se aplicarán para un θ fijo para comodidad de la notación y se seguirá suponiendo $e_i = ae_{i-1}+a_i$. A continuación, se presentarán dos formas de corrección del modelo para eliminar la autocorrelación.

La primera consiste en sustituir en el modelo $y_i=\beta_0(\theta)+\beta_1(\theta)x_i+ae_{i-1}+a_i$, el término e_{i-1} por su definición $e_{i-1}=y_{i-1}-\beta_0(\theta)-\beta_1(\theta)x_{i-1}$ quedando:

$$\begin{aligned}
y_i &= \beta_0(\theta)+\beta_1(\theta)x_i+a(y_{i-1}-\beta_0(\theta)-\beta_1(\theta)x_{i-1})+a_i\\
y_i &= (1-a)\beta_0(\theta)+ay_{i-1}+(x_i-ax_{i-1})\beta_1(\theta)+a_i\\
y_i &= b_0+ay_{i-1}+\beta_1(\theta)x_i-b_1x_{i-1}+a_i
\end{aligned}$$

siendo $b_1=a\beta_1$ y $b_0=(1-a)\beta_0$. Llegando así a un modelo con errores a_i independientes e idénticamente distribuidos. No se puede aplicar cuando se trata de obtener proyecciones a largo plazo ya que como covariable se tiene el retardo y_{i-1} de la variable respuesta.

Otra forma de corrección del modelo para eliminar la autocorrelaciónes es mediante los procedimientos de Prais-Winsten (PW) y Cochrane-Orcutt (CO) que transforman los datos usando el coeficiente de autoccorrelación, a. Primero se verá su funcionamiento en el modelo más simple y luego se extenderá a modelos con p variables regresoras.

$$\begin{aligned}
y_i &= \beta_0(\theta)+\beta_1(\theta)x_i+a(y_{i-1}-\beta_0(\theta)-\beta_1(\theta)x_{i-1})+a_i\\
y_i &= (1-a)\beta_0(\theta)+ay_{i-1}+(x_i-ax_{i-1})\beta_1(\theta)+a_i\\
y_i-ay_{i-1} &= b_0+(x_i-ax_{i-1})\beta_1(\theta)+a_i\\
y_i^* &= \beta_0(\theta)+\beta_1(\theta)x_i^*+a_i
\end{aligned}$$

siendo $y_i^*=y_i-ay_{i-1}$ y $x_i^*=x_i-ax_{i-1}$. Obteniendo así, un modelo con errores a_i independientes e idénticamente distribuidos.

Se considera el siguiente modelo con p variables regresoras para un cuantil θ fijo:

$$\mathbf{Y}=\mathbf{X}\beta+\varepsilon \qquad \varepsilon_i=a\varepsilon_{i-1}+\eta_i \tag{2.11}$$

La matriz $(n\times n)$ de transformación para PW es la siguiente:

$$M_1=\begin{pmatrix}
\sqrt{1-a} & 0 & \dots & 0 & 0\\
-a & 1 & \dots & 0 & 0\\
\vdots & \vdots & \dots & \vdots & \vdots\\
0 & 0 & \dots & -a & 1
\end{pmatrix}$$

Multiplicando el modelo (2.11) por M_1 se tiene:

$$M_1\mathbf{Y}=M_1\mathbf{X}\beta+M_1\varepsilon \tag{2.12}$$

dando lugar al siguiente modelo:

$$\mathbf{Y}^*=\mathbf{X}^*\beta+\eta \tag{2.13}$$

siendo $\mathbf{Y}^*$ el vector dependiente transformado, $\mathbf{X}^*$ la matriz regresora transformada y η el vector de errores incorrelados.

De esta forma, el estimador $\hat{\beta}(\theta)$ de $\beta(\theta)$ se distribuye asintóticamente como sigue, ver [10]:

$$\sqrt{n}[\hat{\beta}(\theta) - \beta(\theta)] \to N(0, \omega^2(\theta)\mathbf{A}^{-1}), \tag{2.14}$$

Estos dos procedimientos se diferencian en cuanto al trato de la primera observación. La matriz de transformación para el procedimiento CO es la matriz $(n-1) \times n$ obtenida a partir de la matriz M_1 a la que se le quita la primera fila. [10] estudió las propiedades asintóticas de CO y demostró que los estimadores resultantes eran consistentes y asintóticamente normales.

En el procedimiento CO se usan $(n-1)$ observaciones en vez de n para estimar el modelo. Asintóticamente, la pérdida de una solo observación es de escasa influencia. En el caso de tamaños de muestra muy pequeños, este efecto puede ser más notable, y sería conveniente usar el procedimiento PW, véase [4].

Una vez que se tiene el modelo con errores independientes e idénticamente distribuidos se pueden aplicar los procedimientos de inferencia descritos en la Sección (2.1)

2.4. Bondad de ajuste

Las medidas de bondad de ajuste de un modelo estadístico describen como el modelo se ajusta a la muestra. Una medida de bondad de ajuste para la regresión de cuantiles se basa en en la idea del coeficiente R^2 utilizado para estudiar la bondad de ajuste para la regresión por mínimos cuadrados. Teniendo en cuenta que la regresión de cuantiles se diferencia de la regresión por mínimos cuadrados en que minimiza la suma ponderada de valores absolutos de los errores y no la suma de errores cuadráticos, se define el coeficiente $pseudoR^2$:

$$pR^2 = 1 - \frac{RASW_\theta}{TASW_\theta}, \tag{2.15}$$

siendo

$$RASW_\theta = \sum_{y_i \geq \mathbf{x}_i^T \hat{\beta}} \theta |y_i - \mathbf{x}_i^T \hat{\beta}| + \sum_{y_i < \mathbf{x}_i^T \hat{\beta}} \theta |y_i - \mathbf{x}_i^T \hat{\beta}| \tag{2.16}$$

y

$$TASW_\theta = \sum_{y_i \geq \theta} \theta |y_i - \hat{\theta}| + \sum_{y_i < \theta} (1-\theta) |y_i - \hat{\theta}| \tag{2.17}$$

con $\hat{\theta}$ el cuantil estimado.

Como $RASW_\theta$ es siempre más pequeño que $TASW_\theta$ se tiene que el valor del coeficiente pR^2 pertenece al intervalo $(0,1)$. Este coeficiente no se puede considerar como una medida para estudiar la bondad de ajuste de todo el modelo sino para un determinado cuantil.

Si el coeficiente pR^2 es próximo a 1 se podrá decir que el ajuste para un determinado cuantil θ es bueno. Que sea bueno para un cuantil no quiere decir que el ajuste sea bueno para otros cuantiles.

R^2relativo

Dado un modelo de regresión, se considera a partir de él otro más simple, es decir, un modelo con menos variables regresoras. El estudio de cuál de los dos modelos es mejor se puede realizar mediante el coeficiente $R^2 relativo$, (R_R^2). Este coeficiente se define de la siguiente manera:

$$R_R^2 = 1 - \frac{\hat{V}(\theta)}{\tilde{V}(\theta)}, \tag{2.18}$$

siendo $\tilde{V}(\theta)$ la función objetivo del modelo más simple y $\hat{V}(\theta)$ la función objetivo del modelo más complejo.

El coeficiente R_R^2 se utiliza como una medida descriptiva. Cuando el coeficiente es próximo a 0, es decir, las funciones objetivo de ambos modelos son similares, se elegirá el modelo más simple. En caso contrario, si la función objetivo del modelo simple es mayor, el coeficiente R_R^2 no será próximo a 0 y se elegirá el modelo más complejo.

Capítulo 3

Aplicación de la regresión de cuantiles

3.1. Introducción

La importancia del estudio de olas de calor es cada día mayor debido al impacto negativo que este fenómeno causa en el ecosistema y en la salud humana. Debido al calentamiento global, inducido por la creciente concentración de gases de efecto invernadero, cabe esperar que las olas de calor serán más frecuentes.

Una importante herramienta para prevenir el calentamiento global es la caracterización y obtención de futuras proyecciones de olas de calor incluyendo información de las temperaturas máximas y mínimas diarias.

Para el estudio del impacto del cambio climático, las proyecciones diarias y locales de temperatura son necesarias. Hoy en día, los Modelos de Circulación General, son las mejores técnicas para obtener futuras proyecciones de diferentes variables atmosféricas a nivel mensual y sobre un área muy amplia. Sin embargo, no son capaces de ofrecer proyecciones fiables diarias y locales de la temperatura y no pueden ser usadas directamente para proyectar el comportamiento extremo de las temperaturas a estas escalas.

Este trabajo tiene como principal objetivo estudiar la evolución de umbrales extremos, en particular, el percentil 95, de la temperatura máxima diaria de Zaragoza entre los años 1951 y 2005 durante los meses de mayo, junio, julio, agosto y septiembre. Además, el modelo obtenido se podría utilizar para obtener proyecciones de la evolución futura del percentil 95. También se realizará una comparación con otro umbral extremo, en este caso, el percentil 5 y con la mediana.

En este trabajo se utiliza el programa estadístico R, en particular, el paquete *quantreg*, véase [6], que nos permite estimar los coeficientes en un modelo de regresión de cuantiles. A continuación sólo se mostrarán los resultados más relevantes; el código R entero se encuentra en el Apéndice A.

3.2. Análisis descriptivo

Los datos que se van a analizar son las series de la temperatura de Zaragoza proporcionados por AEMET. Zaragoza está situada en el valle del Ebro a 240 m sobre el nivel del mar y se caracteriza por tener un clima semiárido frío; los inviernos son ligeramente fríos y los veranos son cálidos. El cierzo sopla frecuentemente en invierno y a principios de la primavera. Se tienen los datos correspondientes a las temperaturas máximas diarias de Zaragoza entre los años 1951 y 2005 durante los meses de mayo a septiembre medidas en décimas de grado.

Evolución de las temperaturas máximas diarias

Según European Climate Assessment (ECA), ver [1], la media de la temperatura máxima diaria durante 1951-2005 en la cuenca del Ebro fue estable en otoño y aumentó en el resto de estaciones. Este incremento fue significativo en verano y más aún en primavera.

Para mostrar la evolución la temperatura en Zaragoza durante el periodo 1951-2005 de mayo a septiembre se representa en la Figura (3.1) un suavizado tipo lowess, ver [2], con una ventana del 30 %.

Cabe destacar que el periodo hasta el año 1980, a pesar de sufrir algunas oscilaciones, se caracteriza por ser un periodo estable mientras que es a partir de 1980 es cuando se produce un aumento en la media de la temperatura máxima diaria.

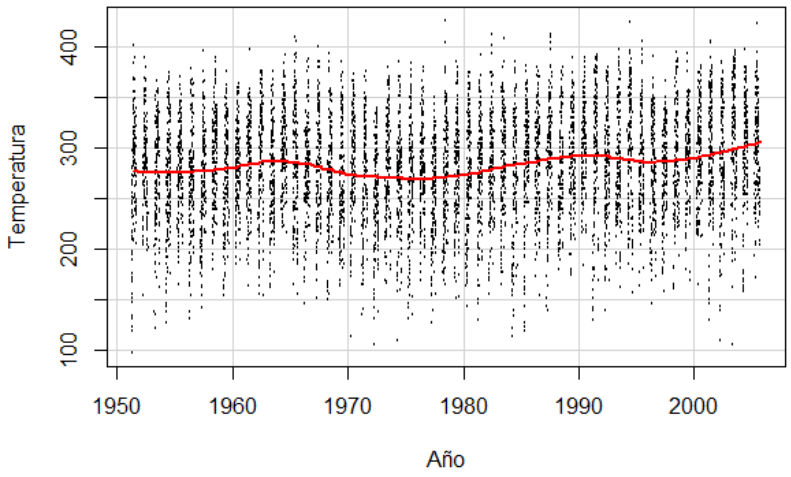

Figura 3.1: Representación de las temperaturas máximas diarias (Tx) entre los años 1951-2005 para los meses de mayo a septiembre. La línea roja corresponde a un suavizado tipo lowess con una ventana del 30 %

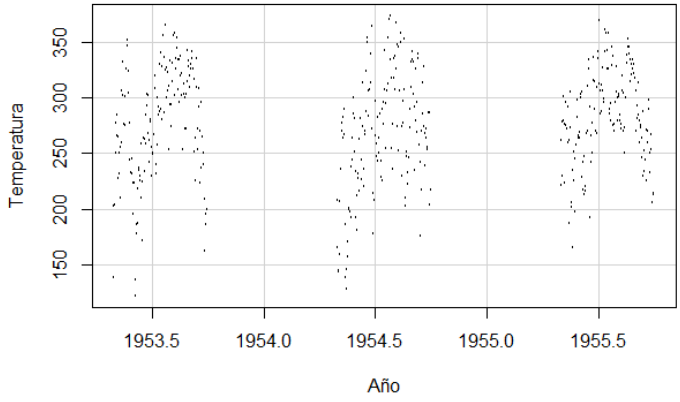

Figura 3.2: Representación de las temperaturas máximas diarias (Tx) entre los años 1951-1956 para los meses de mayo a septiembre.

Ampliando la Figura (3.1), se puede apreciar en la Figura (3.2) la estacionalidad de la temperatura máxima diaria a lo largo de los años. En este caso se ha observado entre 1953 y 1956 pero es extensible para todos los años.

3.3. Modelo para percentil alto

Según AEMET, se define ola de calor cuando la temperatura máxima diaria se encuentra por encima del percentil 95 en un periodo de referencia. En Zaragoza, por ejemplo, es de 37ºC. Luego, el siguiente objetivo será estudiar la evolución del percentil 95 durante los años entre 1951-2005 de los meses de mayo a septiembre. Por lo tanto, se planteará un modelo de regresión de cuantiles para el cuantil $\theta = 0{,}95$, siendo Tx la variable respuesta. Como variables regresoras, se van a considerar dos tipos diferentes de variables:

1. Términos de temperatura. . Se tienen dos variables de temperatura, ambas medidas en décimas de grado:

   - Una señal de la evolución de la temperatura a largo plazo, TTx, que representa la evolución decadal.
   - Una señal de la evolución de la temperatura a corto plazo, Tx_{m31}.

   Como el modelo que se va a obtener se podría utilizar para obtener proyecciones futuras del percentil 95, las proyecciones diarias y locales de la temperatura serían necesarias. Como ya se ha mencionado, las técnicas actuales no son capaces de ofrecer proyecciones fiables diarias y locales. Por lo tanto, en vez de proyecciones diarias, se utilizarán proyecciones a corto y largo plazo.

2. Términos estacionales. La estacionalidad de la temperatura se tiene en cuenta considerando como variables la restricción a los meses de mayo a septiembre del armónico de primer orden que describe el ciclo anual:

   $$cos(\tfrac{2\pi i}{365}) \qquad sen(\tfrac{2\pi i}{365})$$

   con $i = 121, ..., 273$ correspondiente al día del año y que se denotan *Acos* y *Asen* respectivamente. En este caso, sólo las variables relacionadas con la temperatura máxima han sido consideradas. Sin embargo, también se podrían incluir variables relacionadas con la temperatura mínima.

Modelo inicial

En primer lugar, se considera el modelo incluyendo todas las variables regresoras:

$$Q_{Tx}(\theta = 0{,}95|\mathbf{X}) = \beta_0(0{,}95) + \beta_1(0{,}95)Acos + \beta_2(0{,}95)Asen + \beta_3(0{,}95)Tx_{m31} + \beta_4(0{,}95)TTx \quad (3.1)$$

```
## Call:
## rq(formula = Tx ~ Acos + Asen + Txm31 + TTx, tau = 0.95, data = Datos)
##
## Coefficients:
## (Intercept)    Acos        Asen      Txm31      TTx
## 100.1391315 -6.4816237 5.2831103 0.9973749 -0.1600211
##
## Degrees of freedom: 8415 total; 8410 residual
```

$$Q_{Tx}(0{,}95|\mathbf{X}) = 100{,}13 - 6{,}48Acos + 5{,}28Asen + 0{,}99Tx_{m31} - 0{,}16TTx \quad (3.2)$$

Claramente la temperatura máxima diaria es una serie con correlación serial. Los errores, se espera que sean correlados. Calculando el coeficiente de autocorrelación $\rho = 0{,}5836638$ se puede afirmar la presencia de correlación en los errores. En series correladas, la estimación es válida pero no la estimación de la variabilidad, luego no se pueden aplicar las herramientas de inferencia para errores i.i.d. Ver [3].

Se transformarán las variables con el fin de obtener un modelo con errores independientes e idénticamente distribuidos y poder aplicar las técnicas de inferencia presentadas en el capítulo anterior. Se utilizará el procedimiento de Cochrane-Orcutt, ver Sección (2.3), siendo el i-ésimo termino de la variable respuesta modificada $y_i - \rho y_{i-1}$. El procedimiento es análogo para el resto de variables.

Modelo con las variables transformadas

Con las variables transformadas se ha estimado el siguiente modelo:

```
## Call:
## rq(formula = Txmod ~ Acosmod + Asenmod + Txm31mod + TTxmod, tau = 0.95,
## data = Datos)
##
## Coefficients:
## (Intercept)    Acosmod      Asenmod    Txm31mod     TTxmod
## 45.36667111 -26.51416241 -0.01468816 0.82310385 0.07440750
##
## Degrees of freedom: 8415 total; 8410 residual
```

$$Q_{Tx}(0{,}95|\mathbf{X}) = 45{,}36 - 26{,}51Acosmod - 0{,}01Asenmod + 0{,}82Tx_{m31}mod - 0{,}074TTxmod \qquad (3.3)$$

Al obtener el coeficiente de autocorrelación, $\rho = 0{,}0061$, se está ante un modelo en el que se podrán aplicar las técnicas de inferencia que se han estudiado con anterioridad.

Se va a estudiar si las variables introducidas en el modelo tienen una influencia significativa sobre el percentil 95 de la variable respuesta $Txmod$. En primer lugar, se estudia la influencia que tiene el termino armónico de primer orden. Para ello se realizará el siguiente contraste de hipótesis:

$$H_0 : \beta_1(0{,}95) = \beta_2(0{,}95) = 0$$
$$H_1 : \beta_1(0{,}95) \neq 0 \text{ ó } \beta_2(0{,}95) \neq 0$$

Utilizando el test de Razón de Verosimilitud (RV) se obtienen los siguientes resultados:

```
## Likelihood ratio test
##
## Model 1: Txmod ~ Txm31mod + TTxmod
## Model 2: Txmod ~ Acosmod + Asenmod + Txm31mod + TTxmod
## #Df LogLik Df Chisq Pr(>Chisq)
## 1 3 -41539
## 2 5 -41473 2 131.38 < 2.2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor es pequeño, luego la hipótesis nula es rechazada a un nivel de significación $\alpha = 0{,}05$. Por lo tanto, el término armónico de primer orden no se excluye del modelo.

La siguiente variable a la que se le va a estudiar su influencia en el modelo es Tx_{m31}. El contraste para estudiar su influencia es el siguiente:

$$H_0 : \beta_3(0{,}95) = 0$$
$$H_1 : \beta_3(0{,}95) \neq 0$$

Se obtienen los siguientes resultados del test RV:

```
## Likelihood ratio test
##
## Model 1: Txmod ~ Acosmod + Asenmod + TTxmod
## Model 2: Txmod ~ Acosmod + Asenmod + Txm31mod + TTxmod
## #Df LogLik Df Chisq Pr(>Chisq)
## 1 4 -41853
## 2 5 -41473 1 759.47 < 2.2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El p-valor es muy pequeño (<0,05) luego la hipótesis nula es rechazada y no se excluye la variable del modelo.

Por último, se estudia la influencia de TTx a través del siguiente contraste:

$$H_0 : \beta_3(0{,}95) = 0$$
$$H_1 : \beta_3(0{,}95) \neq 0$$

```
## Likelihood ratio test
##
## Model 1: Txmod ~ Acosmod + Asenmod + Txm31mod
## Model 2: Txmod ~ Acosmod + Asenmod + Txm31mod + TTxmod
## #Df LogLik Df Chisq Pr(>Chisq)
## 1 4 -41474
## 2 5 -41473 1 1.0177 0.3131
```

Se tiene un p-valor=0.3131 grande ($> 0{,}05$) luego se acepta la hipótesis nula y la variable TTx se excluye del modelo.

Por lo tanto el modelo queda de la siguiente forma:

$$Q_{Txmod}(\theta = 0{,}95|\mathbf{X}) = \beta_0(0{,}95) + \beta_1(0{,}95)Acosmod + \beta_2(0{,}95)Asenmod + \beta_3(0{,}95)Tx_{m31}mod \quad (3.4)$$

Modelo final

Tras el estudio de inferencia realizado, para estudiar la evolución del umbral que define las olas de calor en Zaragoza correspondientes al periodo entre 1951-2005 de los meses de mayo a septiembre, se considera el siguiente modelo:

$$Q_{Tx}(\theta = 0{,}95|\mathbf{X}) = 61{,}21 - 11{,}001Acos + 3{,}56Asen + 0{,}96Tx_{m31} \quad (3.5)$$

A continuación, en la Figura (3.2) se representa mediante una línea de suavizado la evolución del valor ajustado del percentil 95 de la temperatura máxima diaria desde el año 1951 hasta 2005.

Se puede apreciar como la evolución del valor ajustado del percentil 95 de la temperatura máxima diaria en Zaragoza entre los años 1951 y 1980 se han mantenido estable, a pesar de algunas oscilaciones que se pueden observar. A partir de 1980 se observa como el percentil 95 aumenta, varía desde los 34ºC hasta los 35.5ºC.

3.4. Comparación de la evolución de distintos percentiles

Por último se va a comparar la evolución para el percentil 50, la mediana, y la evolución para otro umbral extremo, como es el percentil 5, con la evolución del percentil 95 que se acaba de estudiar. Al igual que se ha hecho para el modelo del percentil 95, se debería hacer una búsqueda para el modelo de la mediana y para el modelo del percentil 5. En este caso, se van a utilizar las mismas variables que se han usado para el modelo del percentil 95.

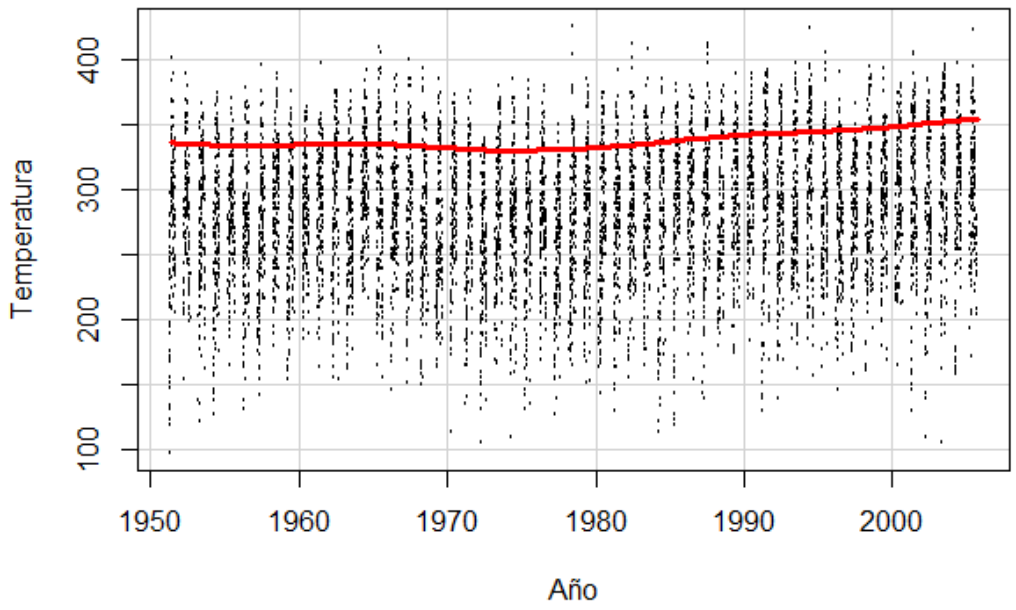

Figura 3.3: Representación de la temperatura máxima diaria (Tx). La línea roja es una línea de suavizado correspondiente a la evolución del percentil 95 de Tx.

Modelo para la mediana

Estimación de los coeficientes del modelo de regresión para la mediana:

```
## Call:
## rq(formula = Tx ~ Acos + Asen+ Txm31, tau = 0.5, data = Datos)
##
## Coefficients:
## (Intercept)   Acos     Asen     Txm31
## -21.093144 11.042932 2.544040 1.113657
## Degrees of freedom: 8415 total; 8411 residual
```

$$Q_{Tx}(\theta = 0{,}95|\mathbf{X}) = -21{,}07 + 11{,}04Acos + 2{,}54Asen + 1{,}11Tx_{m31} \tag{3.6}$$

Modelo para el percentil 5

Para el modelo de regresión del percentil 5 se estiman los siguiente coeficientes:

```
## Call:
## rq(formula = Tx ~ Acos + Asen + Txm31, tau = 0.05, data = Datos)
##
## Coefficients:
## (Intercept)    Acos     Asen     Txm31
## -40.5674561 4.8495542 -4.5238026 0.9373135
## Degrees of freedom: 8415 total; 8411 residual
```

$$Q_{Tx}(\theta = 0{,}95|\mathbf{X}) = -40{,}56 + 4{,}84Acos - 4{,}52Asen + 0{,}93Tx_{m31} \tag{3.7}$$

3.5. Conclusiones

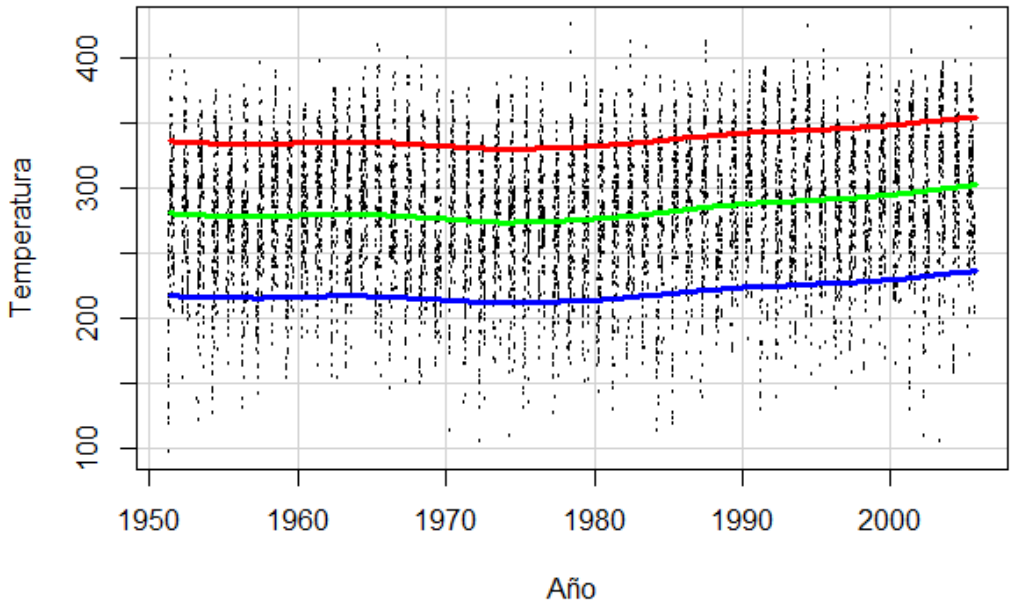

Figura 3.4: Representación de la nube de puntos de la temperatura máxima diaria (Tx). Las líneas de suavizado corresponden a la evolución del percentil 95 de Tx (línea roja), a la evolución de la mediana de Tx (línea verde) y a la evolución del percentil 5 de Tx (línea azul).

En la Figura (3.3) se representan las líneas de suavizado correspondientes al percentil 95 (línea roja), a la mediana (línea verde) y al percentil 5 (línea azul). Cabe destacar que la evolución tanto para la mediana como para el percentil 5 presenta algunas oscilaciones pero se mantiene estable hasta el año 1980 y a partir de este año, aumenta. Lo mismo ocurría con el percentil 95.

Cabe destacar que el aumento que se produce a partir del año 1980 es más significativo en el percentil 5, aumenta aproximadamente 3^oC, mientras que para la mediana y para el percentil 95, los incrementos son aproximadamente de 1.5^oC y 1^oC, respectivamente.

Bibliografía

[1] J. ABAURREA, J. ASÍN, O. ERDOZAIN Y E. FERNÁNDEZ, Climate variability analysis of temperature series in the medium Ebro river basin, *Springer-Verlag* (2001), 109–118.

[2] W.S. CLEVELAND, Robust locally weighted regression and smoothing scatterplots *J. American Statistical Association* **74** (1979), 829–836.

[3] C. DAVINO, M. FURNO Y D. VISTOCCO, *Quantile Regression: Theory and Application*, 1.ª ed., John Wiley & Sons, Ltd, 2014.

[4] T.E. DIELMAN Y E.L. ROSE, Estimation in least absolute value regression with autocorrelated errors: A small sample study, *International Journal of Forecasting* **10** (1994), 539–547.

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

[6] R. KOENKER, *quantreg: Quantile Regression*, `http://CRAN.R-project.org/package=quantreg`..

[7] R. KOENKER Y G. BASSET, Regression quantiles, *Econometrica* **46** (1) (1978), 33–50.

[8] R. KOENKER Y G. BASSET, Robust tests for heteroskedasticity based on regression quantiles, *Econometrica* **50** (1982a), 43–61.

[9] M. SIDDIQUI, Distributions of quantiles from a bivariate population, *Journal of Research of the National Bureau of Standards.* **64** (1960), 145–150.

[10] A. WEISS, Least absolute error estimation in the presence of serial correlation, *Econometrica* **44** (1990), 127–159.