FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Modelos lineais de regresión cuantil

Lucía Gil Rial

2018/2019

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Modelos lineais de regresión cuantil

Lucía Gil Rial

Xullo 2019

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Traballo proposto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Modelos lineais de regresión cuantil
Breve descrición do contido
Un modelo de regresión permítenos establecer a relación entre unha variable resposta e unha ou varias variables explicativas. Aínda que a forma máis coñecida de estudar esta relación é mediante a regresión en media, unha formulación baseada na regresión en mediana, ou en xeral a regresión cuantil, adquiriu unha gran relevancia nos últimos anos debido a que permite unha descrición máis detallada do comportamento da variable resposta, adáptase a situacións baixo condicións máis xerais da distribución do erro e goza de propiedades de robustez. Neste traballo imos completar o estudo dos modelos lineais de regresión, non so dende un punto de vista dos mínimos cadrados, senón a través dun enfoque máis robusto. A título orientativo, o traballo podería organizarse nas seguintes seccións: ▪ Presentación do modelo de regresión cuantil ▪ Cálculo dos estimadores mediante programación lineal ▪ Propiedades dun modelo de regresión cuantil ▪ Inferencia para a regresión cuantil Presentaremos diferentes modelos de regresión cuantil aplicados a conxuntos de datos tanto reais como simulados. Para elo empregaremos o software libre R (`https://www.r-project.org/`).

Índice xeral

Resumo

O obxectivo desta memoria é presentar os modelos de regresión cuantil. Para isto defínese previamente o concepto de variable aleatoria así como as súas principais características. Préstase especial atención á estimación de cuantís mostrais, que pode ser vista como un problema de optimización, e resultará de gran utilidade para a estimación do modelo cuantil. A continuación preséntase o modelo de regresión en media, que ten como propósito estudar posibles relacións entre distintas variables aleatorias a través do método de mínimos cadrados. Paralelamente, defínese o modelo lineal de regresión cuantil así como métodos inferenciais asociados á estimación dos parámetros da regresión. Preséntanse tamén as propiedades máis salientables de dito modelo como a robustez (que se verá a través da función de influencia) que presenta fronte a datos atípicos e tamén unha das súas debilidades como o cruce entre cuantís. Por outra banda, realízanse dous estudos de simulación a través do programa R. O primeiro ten como obxectivo mostrar, mediante os erros cadráticos medios, que canto maior sexa a densidade da variable resposta avaliada no cuantil de interese, menor será a variabilidade dos estimadores da regresión. O segundo estudo mostra a través de representacións gráficas a robustez que presenta a regresión en mediana fronte a datos atípicos mentres que a regresión en media vese moi afectada por ditas observacións. Despois realízase unha aplicación a datos reais coa base de datos *Engel* que proporciona o paquete *quantreg* de R. Para finalizar preséntase as principais conclusións derivadas deste traballo, así como un anexo no que se recollen todos os códigos de R necesarios para levar a cabo os estudos feitos ao longo do traballo.

Abstract

The objective of this research is to introduce the quantile regression model. For this purpose, we will first define a random variable and its main characteristics. Then, we will focus on the sample quantile estimation, which can be seen as an optimization problem, and it would be useful for the quantile regression model estimation. Subsequently, we will introduce the regression toward the mean model of which purpose is to study the possible relationships among different random variables through the least-square method. Concurrently, we will define the linear quantile regression model as well as inference methods

associated with the estimation of the regression parameters. We will show the most important properties of the aforementioned model such as robustness (which will be seen through the influence function) that the model shows in the presence of outliers and we will also show one of its weaknesses like the quantile crossing. Furthermore, we will conduct two simulation studies through the well-known software R. The first one has the purpose of showing, by the mean squared error, that a higher density of the target variable evaluate in a particular quantile implicates a minor variability of the regression estimators. The second one presents (through graphic representations) the robustness that median regression has in front of outliers while the regression to the mean is really sensitive to this observations. Lastly, we will show a real data application with the data basis Engel provided by the R package *quantreg*. Finally, we will summarize the main ideas derived from this work, and we will add an annex which contains the programming code that we use during this project.

Resumen

El objetivo de esta memoria es presentar los modelos de regresión cuantil. Para esto se define previamente el concepto de variable aleatoria así como sus principales características. Se presta especial atención a la estimación de cuantiles mostrales, que puede ser vista como un problema de optimización, y será de gran utilidad para la estimación del modelo cuantil. A continuación se presenta el modelo de regresión en media, cuyo propósito es estudiar posibles relaciones entre distintas variables aleatorias a través del método de mínimos cuadrados. Paralelamente, se define el modelo lineal de regresión cuantil así como métodos inferenciales asociados a la estimación de los parámetros de la regresión. Se presentan también las propiedades más importantes de dicho modelo como la robustez (que se verá a través de la función de influencia) que presenta frente a datos atípicos y también una de sus debilidades como el cruce entre cuantiles. Por otra banda, se realizan dos estudios de simulación a través del programa R. El primero tiene como objetivo mostrar, mediante los errores cuadráticos medios, que cuanto mayor sea la densidad de la variable respuesta evaluada en el cuantil de interés, menor será la variabilidad de los estimadores de la regresión. El segundo estudio muestra a través de representaciones gráficas la robustez que presenta la regresión en mediana frente a datos atípicos mientras que la regresión en media se ve muy perjudicada por dichas observaciones. Después se realiza una aplicación

a datos reales con la base de datos *Engel* que proporciona el paquete *quantreg* de R. Para finalizar se presentan las principales conclusiones derivadas de este trabajo, así como un anexo en el que se recogen todos los códigos de R necesarios para llevar a cabo los estudios hechos a lo largo de este proyecto.

Introdución

Aínda que a regresión en media tivese unha gran repercusión neste último século na área de Estatística, é curioso que o concepto de regresión cuantil xurdiu moito antes da idea de regresión en media axustada grazas ao método de mínimos cadrados. O citado método xurde no ano 1805 grazas ás publicacións de Legendre especializadas neste tema, mentres que a regresión cuantil emerxe medio século antes da man de Boscovich.

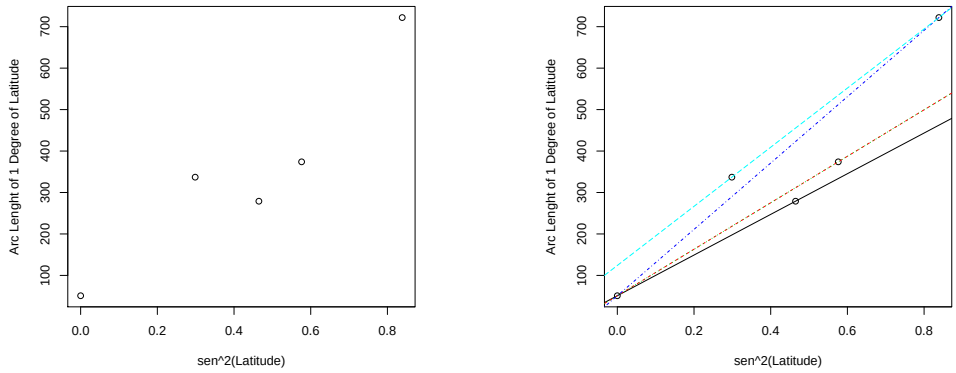
O obxectivo de Boscovich era estudar a forma elíptica da Terra. Para isto mediuse a lonxitude de arco no desplazamento dun grao de latitude cara o norte dende distintos puntos da Terra. É evidente que se o desplazamento se fai dende áreas con maior latitude, maior será ao lonxitude de arco, o que proba a conxectura de Newton. A aproximación á que chegou foi a seguinte:

$$y = a + b\,\mathrm{sen}^2\,\lambda$$

onde y é a lonxitude de arco e λ a latitude. O parámetro a representa a lonxitude do arco asociado a un grao de incremento na latitude, considerando que o ángulo se mide a partir do ecuador. Debemos ter en conta que se a Terra fose esférica entón dita lonxitude de arco sería constante. Ademais, o parámetro b representa o exceso de lonxitude de arco cando o ángulo dun grao se toma no polo Norte. Así obtense a seguinte estimación da elipticidade da Terra:

$$\frac{1}{\text{elipticidade}} = \frac{3a}{b}.$$

Boscovich non axustou a recta por mínimos cadrados, pero aplicou un método moi semellante ao da regresión cuantil: estimou os parámetros a e b de maneira que se faga mínima a suma dos valores absolutos suxeitos á restrición de que a suma dos erro sexa cero. Na Figura 1 (a) pode verse o diagrama de dispersión dos datos empregados por Boscovich e que se atopan na base de datos *Bosco*, dispoñible no paquete *quantreg* de R. E na Figura 1 (b) pode verse o diagrama de dispersión coas rectas axustadas para os cuantís: $0'1$, $0'3$, $0'4$, $0'5$ e $0'75$.



(a) Diagrama de dispersión.

(b) Rectas axustadas.

Figura 1: Diagrama de dispersión da base de datos *Bosco* do paquete *quantreg* (figura (a)), xunto coas rectas axustadas de regresión cuantil para os cuantís $0'1$, $0'3$, $0'4$, $0'5$ e $0'75$ (figura (b)).

A día de hoxe séguese investigando na regresión cuantil posto que o modelo conta con grandes propiedades como a flexibilidade a escenarios onde non se esixen hipóteses como a homocedasticidade ou a normalidade dos erros de regresión (como se pide na regresión en media axustada por mínimos cadrados) e a robustez fronte a datos atípicos.

Neste traballo recóllense as principais características da regresión cuantil. No Capítulo 1 introdúcense os conceptos básicos necesarios para definir a regresión cuantil e tamén se aporta unha definición teórica deste citado modelo. No Capítulo 2 móstrase o procedemento de estimación dos parámetros e preséntanse as vantaxes deste modelo como tamén unha das súas debilidades. No Capítulo 3 realízanse dous estudos de simulación: o primeiro mostra por medio de erros cadráticos medios que canto maior sexa a densidade da variable resposta avaliada no cuantil de interese, menor será a variabilidade que presenten os estimadores da regresión; e o segundo mostra a robustez da regresión en mediana comparado coa regresión en media fronte á presenza de datos atípicos. No Capítulo 4 faise unha aplicación a datos reais para comprobar que na práctica a regresión cuantil funciona de forma axeitada. Posteriormente no Capítulo 5 faise un pequeno resumo das principais ideas que podemos extraer sobre os modelos de regresión cuantil. Por último, no Anexo A achéganse os códigos de R necesarios para levar a cabo os estudos de simulación deseñados ao longo do Capítulo 3 e para analizar a base de datos estudada no Capítulo 4.

Capítulo 1

Preliminares

Neste capítulo abordaranse os conceptos previos necesarios para poder definir unha variable aleatoria e, de seguido, o modelo de regresión, que será a ferramenta estatística fundamental ao longo deste traballo.

Definición 1.1. Unha poboación é un conxunto de elementos de estudo que comparten certas características. Cada elemento da poboación denomínase individuo.

Xeralmente o número de individuos dunha poboación é moi elevado e, como consecuencia, é inviable traballar con todos eles. Polo tanto, tomaremos un subconxunto representativo da poboación, que chamaremos **mostra**, coa que será posible realizar o estudo estatístico.

Agora introducimos uns conceptos novos coa finalidade de caracterizar a idea de probabilidade. Para isto seguiremos a notación de [3] (páxinas 17-23).

Definición 1.2. Un **experimento aleatorio** é unha proba na que os posibles resultados son coñecidos de antemán pero non se pode predecir o resultado antes de realizala.

Definición 1.3. Sexa Ω un conxunto de puntos ω. En teoría de probabilidade este conxunto está composto por tódolos posibles resultados ω dun experimento aleatorio. Este conxunto Ω denomínase **espazo de mostras**. Cada un dos posibles resultados denomínase **suceso elemental**, $\omega \in \Omega$. E un **suceso** é un subconxunto de Ω.

Definición 1.4. Consideramos un espazo arbitrario non baleiro Ω. Unha clase $\mathcal{F}$ de subconxuntos de Ω, $\mathcal{F} \subset P(\Omega)$[1] é unha **$\sigma$-álxebra** se:

[1] Sexa X un conxunto. Notaremos por $P(X)$ ao conxunto das partes de X.

1. $\Omega \in \mathcal{F}$.
2. $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$.[2]
3. $A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$.
4. $A_1, ..., A_n \in \mathcal{F} \Rightarrow A_1 \cup ... \cup A_n \in \mathcal{F}$.[3]

Definición 1.5. Unha función real, $\mathbb{P}$, definida nunha clase de subconxuntos de Ω, que denotaremos por $\mathcal{F}$, nunha σ-álxebra é unha **medida de probabilidade** se verifica as seguintes condicións:

1. $0 \leq \mathbb{P}(A) \leq 1$ se $A \in \mathcal{F}$.
2. $\mathbb{P}(\emptyset) = 0$, $\mathbb{P}(\Omega) = 1$.
3. Se $A_1, A_2, ...$ conforman unha sucesión disxunta de $\mathcal{F}$-conxuntos e se $\bigcup_{k=1}^{\infty} A_k \in \mathcal{F}$, entón:
$$\mathbb{P}\left(\bigcup_{k=1}^{\infty} A_k\right) = \sum_{k=1}^{\infty} \mathbb{P}(A_k)$$

Se $\mathcal{F}$ é unha σ-álxebra en Ω e $\mathbb{P}$ é unha medida de probabilidade en $\mathcal{F}$, a tripla $(\Omega, \mathcal{F}, \mathbb{P})$ coñécese como **espazo de probabilidade**.

Agora xa estamos nas condicións de definir unha variable aleatoria.

Definición 1.6. Unha variable aleatoria é unha aplicación do tipo:

$$\begin{array}{rccc} X: & \Omega & \longrightarrow & \mathbb{R} \\ & \omega & \longmapsto & X(\omega) \end{array}$$

É dicir, a cada resultado do experimento asóciaselle un número real. Na Sección 1.1 definiremos as variables aleatorias discretas e continuas.

Definición 1.7 (Función de distribución)**.** A función de distribución asociada a unha variable X é unha función que describe a probabilidade de que X teña un valor menor ou igual que cada valor de $x \in \mathbb{R}$, é dicir:

$$F(x) = \mathbb{P}(X \leq x), x \in \mathbb{R}.$$

[2]Sexa X un conxunto. Denotaremos por X^c ao complementario de X.
[3]Sexan $X_1, ..., X_n$ conxuntos arbitrarios. Denotaremos por $X_1 \cup ... \cup X_n$ á unión de tales conxuntos.

1.1. Variables aleatorias

1.1.1. Variable aleatoria discreta

Diremos que unha variable aleatoria X é discreta se toma un número finito ou infinito numerable de valores $\{x_i\}_{i\in\mathbb{N}}$. Neste caso definimos a **función de masa de probabilidade** como unha aplicación que asocia a cada valor x_i a probabilidade de que a variable X tome tal valor, é dicir,

$$p_i = \mathbb{P}(X = x_i)$$

con $i = 1, ..., n$. Verificándose que:

$$\sum_{i=1}^{n} p_i = 1$$

É inmediato, entón, que neste caso se pode calcular a función de distribución da seguinte maneira:

$$F(x) = \mathbb{P}(X \leq x) = \sum_{x_i \leq x} p_i.$$

Agora centrarémonos en como se definen as medidas características[4] asociadas a unha variable aleatoria discreta:

1. Medidas de centralización

   a) Esperanza ou media

   Definición 1.8. Sexa X unha variable aleatoria discreta que toma os valores $\{x_1, x_2, ..., x_n\}$, con función de masa de probabilidade $p_1, p_2, ..., p_n$. Definimos a media, que denotaremos por μ, como segue:

   $$\mu = \mathbb{E}[X] = \sum_{i=1}^{n} x_i p_i$$

   As propiedades máis importantes da esperanza preséntanse a continuación:

   1) $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
   2) $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
   3) $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \Leftrightarrow$ X e Y son variables independentes[5].

[4] As medidas características dunha variable aleatoria resumen a información máis importante da variable aleatoria en cuestión. As medidas características máis coñecidas son as medidas de posición, que indican por onde se moven os datos, e as medidas de dispersión, que miden a súa variabilidade.

[5] Sexan X e Y dúas variables aleatorias con funcións de distribución F_X e F_Y e sexa F a función de distribución conxunta. Diremos que estas variables son independentes se verifican: $F(x, y) = F_X(x)F_Y(y)$.

b) Cuantís

Definición 1.9. O cuantil de orden $p \in (0,1)$, que denotaremos por c_p, é aquel punto que verifica:

$$\mathbb{P}(X \leq c_p) \geq p$$
$$\mathbb{P}(X \geq c_p) \geq 1-p$$

É dicir, o cuantil de orde p, c_p, é aquel valor de X tal que deixa por debaixo unha proporción p de valores de X e deixa por riba unha proporción $1-p$ de valores de X.

O cuantil máis utilizado é a mediana, $p = \frac{1}{2}$. Na Sección 1.2 veremos con máis detalle o concepto de cuantil.

2. Medidas de dispersión

a) Varianza

Definición 1.10. A varianza, que denotaremos por σ^2, mide a dispersión que presentan os datos con respecto á media.

$$\sigma^2 = Var[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

En caso de que a variable sexa discreta a varianza defínese como segue:

$$Var[X] = \sum_{i=1}^{n} (x_i - \mathbb{E}[X])^2 p_i$$

Algunhas das propiedades máis importantes da varianza mostránse de seguido:

1) $Var[aX + b] = a^2 Var[X]$
2) $Var[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$

b) Desviación típica

Definición 1.11. A desviación típica, que denotaremos por σ, é a raíz cadrada da varianza.

$$\sigma = \sqrt{Var[X]}$$

1.1.2. Variable aleatoria continua

Unha variable aleatoria X é continua se pode tomar calquer valor nun intervalo, unión de intervalos ou toda a recta real $\mathbb{R}$. Este tipo de variable caracterízase mediante a **función**

de densidade f, que se define como segue:

$$F(x) = \int_{-\infty}^{x} f(u)du$$

A continuación enuméranse algunhas das propiedades das que consta a función de densidade:

1. A función de densidade é continua.
2. $f(x) \geq 0$
3. $\mathbb{P}(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$
4. $\int_{-\infty}^{\infty} f(x)dx = 1$

Tras as definicións de función de densidade e función de distribución, podemos observar a seguinte relación entre elas:

1. $f(x) = \frac{dF(x)}{dx}$
2. $F(x_0) = \int_{-\infty}^{x_0} f(x)dx = \mathbb{P}(X \leq x_0)$

Seguindo co xa visto para variables aleatorias discretas, definimos agora as medidas características das variables aleatorias continuas.

1. Medidas de centralización

   a) Esperanza ou media

   Definición 1.12. Sexa X unha variable aleatoria continua con función de densidade f. Defínese a súa esperanza da seguinte maneira:

   $$\mathbb{E}[X] = \int_{-\infty}^{\infty} xf(x)dx$$

   cando

   $$\int_{-\infty}^{\infty} |x|f(x)dx < \infty$$

   b) Cuantís

   Definición 1.13. O cuantil p-ésimo, c_p, sendo $p \in (0,1)$, é o punto que cumpre: $\mathbb{P}(X \leq c_p) \geq p$ e $\mathbb{P}(X \geq c_p) \geq 1 - p$.

2. Medidas de dispersión

   a) Varianza

   Definición 1.14. Se X é unha variable aleatoria continua con función de densidade f, a súa varianza é:

$$\sigma^2 = Var[X] = \int_{-\infty}^{\infty} (x - \mathbb{E}[X])^2 f(x) dx$$

   b) Desviación típica

   Definición 1.15. A desviación típica dunha variable aleatoria continua é a raíz cadrada da varianza.

$$\sigma = \sqrt{Var[X]}$$

1.2. Os cuantís e a súa estimación

Na Sección 1.1 vimos a definición teórica de cuantil. Agora veremos como se define este mesmo na práctica. Se a función de distribución F dunha variable aleatoria Y é continua e estrictamente monótona, a inversa da función de distribución redúcese a unha inversa 'clásica' e devolve un único valor c_τ tal que $\mathbb{P}(Y \leq c_\tau) = \tau$. Denotamos por c_τ ao cuantil de orde $\tau \in (0, 1)$ da variable Y. Na Figura 1.1 podemos ver un exemplo desta situación.

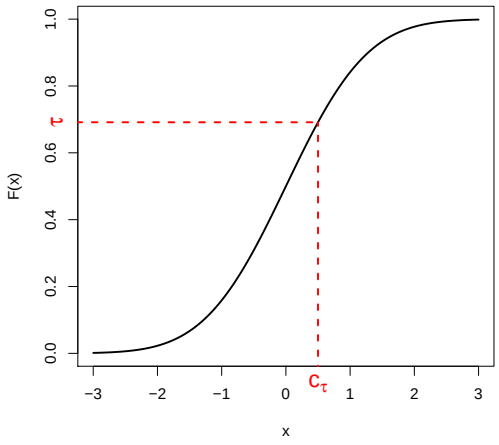


Figura 1.1: Función de distribución continua e estrictamente crecente.

Pero existen outras situacións nas que a definición do cuantil non é tan sinxela. Debemos diferenciar os seguintes casos:

- A función de distribución F podería ter saltos no grafo da mesma. Neste caso, a inversa non está ben definida polo que necesitariamos introducir unha caracterización máis xeral. Un exemplo desta situación pódeser ver na Figura 1.2.

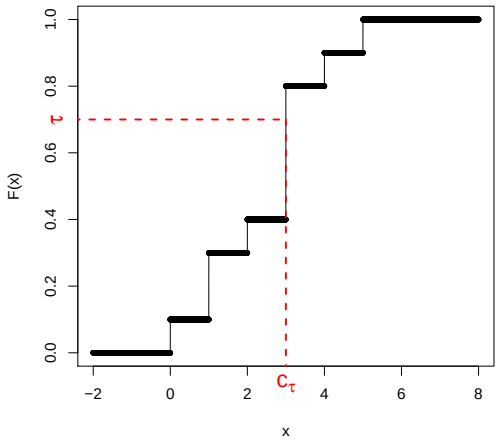


Figura 1.2: Función de distribución con saltos.

- Se a función de distribución é igual a τ nalgún rango, entón temos un intervalo de cuantís de orde τ. Podemos observar un exemplo desta situación na Figura 1.3.

Á vista dos exemplos anteriores podemos establecer a seguinte definición formal de cuantil:

Definición 1.16. Sexa Y unha variable aleatoria con función de distribución $F(y) = \mathbb{P}(Y \leq y)$. Entón definimos o **cuantil** de orde τ da variable Y, que denotaremos por c_τ, como segue:

$$F^{-1}(\tau) = \text{ínf}\{y : F(y) \geq \tau\}$$

con $\tau \in (0, 1)$ e onde ínf denota o ínfimo.

De igual xeito que a media é o valor que minimiza unha función de perda cadrática,

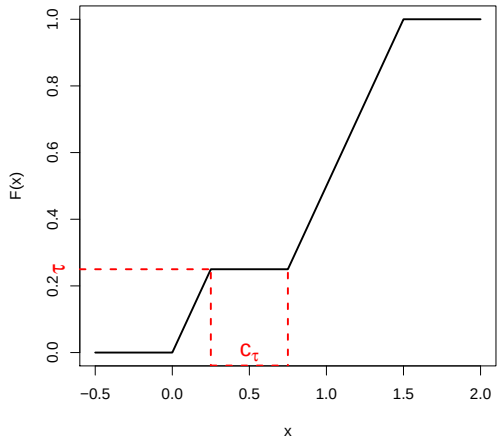


Figura 1.3: Función de distribución con tramos 'planos'.

podemos expresar o cuantil de orde τ da variable Y como segue:

$$Q_Y(\tau) = \arg\min_y \mathbb{E}(\rho_\tau(Y - y))$$

onde ρ_τ se coñece como a **función de perda cuantílica**:

$$\rho_\tau(u) = u(\tau - \mathbb{I}(u < 0)) = \begin{cases} \tau u & se \quad u \geq 0 \\ u(\tau - 1) & se \quad u < 0 \end{cases}$$

sendo $\mathbb{I}$ a función característica. Na Figura 1.4 está representada a función de perda cuantílica para cuantís de distintas ordes τ.

Podemos ver como na Figura 1.4 asociada ao cuantil $0'25$ os valores menores que o cero están máis penalizados que os maiores que cero. En cambio, para o cuantil $0'75$ penalízanse máis os valores maiores que cero. Para a mediana, os datos menores e maiores que cero teñen a mesma penalización, pois a función de perda cuantílica coincide neste caso co valor absoluto.

Dada $\{Y_1, ..., Y_n\}$ unha mostra aleatoria simple da variable Y, podemos calcular o cuantil mostral como:

$$\arg\min_{\xi \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(Y_i - \xi) \tag{1.1}$$

Na ecuación (1.1) temos expresado a busca do cuantil como un problema de optimización. En contraste coa busca da media mostral, neste caso a función de perda non é

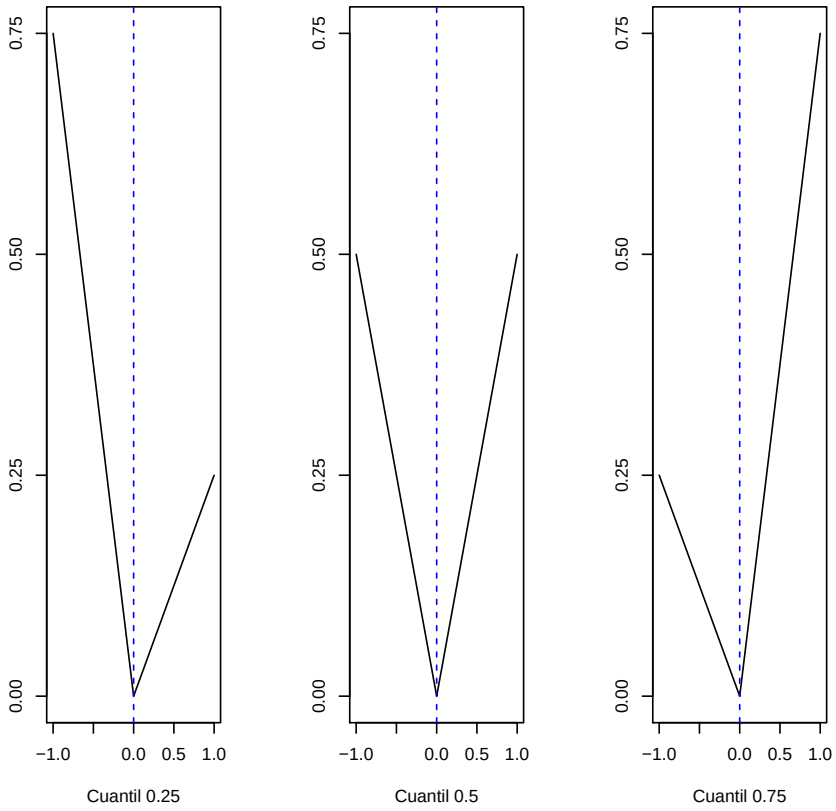


Figura 1.4: Función de pérdida cuantílica para os cuantís $0'25$, $0'5$, $0'75$.

derivable e en consecuencia non é posible aplicar o mesmo método de optimización. Aínda así, o problema (1.1) pode reformularse como un problema de programación lineal engadindo $2n$ variables, u_i e v_i sendo $i = 1, ..., n$, que fan referencia á parte positiva e negativa do vector $Y - \xi$, respectivamente, sendo ξ a cantidade a optimizar. Así obtemos un novo problema,

$$\min_{(\xi,u,v)\in\mathbb{R}\times\mathbb{R}_+^{2n}} \{\tau 1_n^T u + (1-\tau)1_n^T v \mid 1_n\xi + u - v = y\}$$

onde 1_n denota un vector n-dimensional con todas as compoñentes iguais a 1.

Esta formulación do problema da busca do cuantil permitiranos coñecer algoritmos para o seu cálculo como o método do simplex, explicado en [2], ou algoritmos acelerados do mesmo como o *'algoritmo de Barrodale e Roberts'*, explicado en [1].

1.3. Modelos de regresión en media

Supoñamos que temos dúas variables aleatorias distintas, X e Y, que son características dunha poboación, da cal coñecemos unha mostra. É lóxico preguntarse se existe algunha relación entre ámbalas dúas variables. Para este obxectivo son utilizados os **modelos de regresión**. Grazas a ditos modelos podemos coñecer a dependencia da variable Y, que chamaremos variable resposta ou dependente, en base á variable X, que coñeceremos como variable explicativa ou independente. Ademais podemos efectuar predicións do valor de Y sempre e cando coñezamos un novo valor de X.

A regresión en media ven dada pola seguinte función: $m(x) = \mathbb{E}(Y|X = x)$, onde x é un valor de X. Podemos escribir entón a variable resposta como:

$$Y = m(X) + \epsilon$$

sendo ϵ o erro, que verifica $\mathbb{E}(\epsilon|X = x) = 0$ para todo x valor de X.

Imos centrarnos no modelo de regresión lineal simple. En dito caso tense que

$$Y = \beta_0 + \beta_1 X + \epsilon$$

sendo β_0 a ordenada na orixe, β_1 a pendente e ϵ o erro. Dada unha mostra $(x_1, Y_1), ..., (x_n, Y_n)$ baixo deseño fixo será fundamental estimar os parámetros β_0 e β_1. Denotaremos por $\widehat{\beta_0}$ e $\widehat{\beta_1}$ a tales estimadores. Así para un valor observado x_i obteriamos a predición $\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$ da variable resposta. Notemos que Y_i é o valor observado e $\widehat{Y_i}$ é a predición. Logo, chamaremos aos erros da predición,

$$\widehat{\epsilon_i} = Y_i - \widehat{Y_i} = Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i$$

$\forall i \in \{1, ..., n\}$, **residuos** da regresión. A citada estimación dos parámetros farase por mínimos cadrados, pois buscamos os estimadores de β_0 e β_1 que produzan os residuos mínimos.

Na Figura 1.5 podemos ver un diagrama de dispersión dunha certa mostra coa recta axustada por mínimos cadrados. Os segmentos verticais representan os residuos da regresión.

Seguindo o método de mínimos cadrados, os estimadores serán aqueles que minimicen os segmentos discontinuos da Figura 1.5. É dicir, os estimadores de mínimos cadrados defínense como segue:

$$\sum_{i=1}^{n} (Y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i)^2 = \min_{\beta_0, \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 \tag{1.2}$$

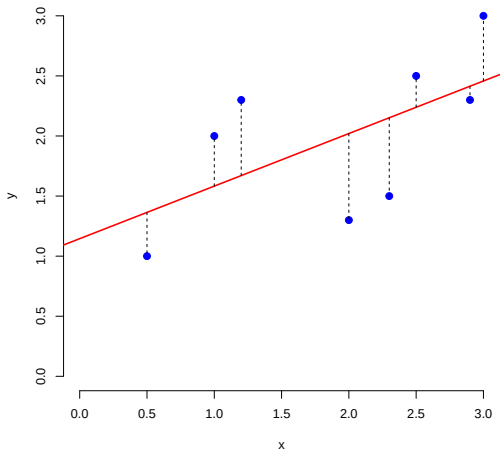


Figura 1.5: Diagrama de dispersión coa recta axustada por mínimos cadrados.

Para calcular o mínimo desta ecuación débese derivar respecto de β_0 e β_1, igualar a cero e despexar os seus valores, que serán os posibles puntos críticos. Coa segunda derivada compróbase que ditos puntos son os mínimos da ecuacion (1.2). A solución que se obten é a seguinte:

$$\widehat{\beta_0} = \bar{Y} - \frac{S_{xy}}{S_x^2}\bar{x}$$
$$\widehat{\beta_1} = \frac{S_{xy}}{S_x^2}$$

onde:

- $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ é a media da mostra Y.
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ é a media da mostra X.
- $S_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})$ é a covarianza[6] entre X e Y.
- $S_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ é a varianza da mostra X.

Esta estimación da recta de regresión pasa polo centro de gravidade $(\bar{x}, \bar{Y})$ e ten como pendente $\widehat{\beta_1} = \frac{S_{xy}}{S_x^2}$.

[6] A covarianza é o valor que indica a variación de dúas variables aleatorias respecto das súas medias. Serve para medir a relación lineal entre dúas variables X e Y.

Dita estimación está sustentada nas seguintes hipóteses:

1. **Linealidade**: A función de regresión é unha recta, é dicir, $m(x) = \beta_0 + \beta_1 x$.
2. **Homocedasticidade**: A varianza do erro é constante, é dicir, $Var(\epsilon|X = x) = \sigma^2$.
3. **Normalidade**: A distribución do erro é gaussiana, isto é, $\epsilon \in N(0, \sigma^2)$[7].
4. **Independencia**: Os erros $\epsilon_1, ..., \epsilon_n$ son mutuamente independentes.

A continuación mencionamos as propiedades máis salientables dos estimadores.

- $\mathbb{E}[\widehat{\beta_0}] = \beta_0$. Como consecuencia, $\widehat{\beta_0}$ é un estimador insesgado[8] de β_0.
- $Var[\widehat{\beta_0}] = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}\right)$.
- O estimador $\widehat{\beta_0}$ segue unha distribución normal: $\widehat{\beta_0} \in N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{nS_x^2}\right)\right)$.
- $\mathbb{E}[\widehat{\beta_1}] = \beta_1$, polo que $\widehat{\beta_1}$ é un estimador insesgado de β_1.
- $Var[\widehat{\beta_1}] = \frac{\sigma^2}{nS_x^2}$.
- $\widehat{\beta_1}$ segue tamén unha distribución normal: $\widehat{\beta_1} \in N\left(\beta_1, \frac{\sigma^2}{nS_x^2}\right)$.

Vimos de ver un modelo moi sinxelo, que só conta cunha variable explicativa. Pero en moitas ocasións precísanse de varias variables explicativas para poder explicar a variable dependente. O modelo que formula esta situación é o **modelo lineal múltiple**. Neste modelo de regresión contamos coa variable resposta Y e $p-1$ variables explicativas $X_1, ..., X_{p-1}$ e temos que

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + \epsilon$$

onde $\beta_0, ..., \beta_{p-1}$ son o intercepto e os parámetros asociados á variable explicativa correspondente e ϵ o erro.

[7] O modelo de distribución de probabilidade para variables continuas máis importante é a distribución normal. A súa función de densidade é $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp[-\frac{1}{2\sigma^2}(x-\mu)^2]$ cuxa gráfica é unha campana de Gauss. f depende dos parámetros μ, que é a media, e σ, que é a desviación típica. Diremos logo que unha variable é $N(\mu, \sigma)$ cando segue a citada función de densidade.

[8] Un estimador insesgado é aquel que o seu sesgo, a diferenza entre a esperanza do estimador e o parámetro a estimar, é nulo pois a súa esperanza é igual ao estimador que se desexa estimar.

Se consideramos unha mostra $(x_{1,1},...,x_{1,p-1},Y_1),...,(x_{n,1},...,x_{n,p-1},Y_n)$ podemos expresar Y_i, a variable resposta do individuo i-ésimo, como segue:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_{p-1} x_{i,p-1} + \epsilon_i$$

sendo $x_{i,1},...,x_{i,p-1}$ as variables explicativas de dito individuo e ϵ_i o seu erro asociado con $\mathbb{E}(\epsilon|X) = 0$.

Denotando por $\mathbf{x_i} = (1, x_{i,1}, \dots, x_{i,p-1})$ ao vector asociado ás observacións das variables explicativas do individuo i-ésimo e $\beta = (\beta_0, \dots, \beta_{p-1})$ ao vector de coeficientes, temos a seguinte expresión da función de regresión:

$$m(x_i) = \mathbf{x_i}^T \beta = \beta_0 + \beta_1 x_{i,1} + ... + \beta_{p-1} x_{i,p-1}.$$

Notemos que os erros seguen unha distribución normal $\epsilon_1,...,\epsilon_n \in N(0,\sigma^2)$ e son mutuamente independentes, é dicir, verifican as hipóteses de normalidade, homocedasticidade e independencia. Coa notación que acabamos de ver podemos expresar o modelo de regresión múltiple de forma matricial:

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} = \mathbb{X}\beta + \epsilon$$

e denominaremos á matriz $\mathbb{X}$ como a matriz de deseño do modelo.

O vector de parámetros β pode estimarse mediante mínimos cadrados. É dicir, o noso estimador $\widehat{\beta}$ debe verificar:

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - \mathbf{x_i}^T \beta)^2$$

ou equivalentemente en notación matricial, $\widehat{\beta}$ debe ser aquel onde se alcance:

$$\min_{\beta} (Y - \mathbb{X}\beta)^T (Y - \mathbb{X}\beta)$$

Derivando a función $(Y - \mathbb{X}\beta)^T (Y - \mathbb{X}\beta)$ respeto β e igualando a cero obtemos o que coñeceremos como ecuacións normais da regresión: $\mathbb{X}^T\mathbb{X}\beta = \mathbb{X}^T Y$. E desto obtemos o estimador $\widehat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$.

Unha vez coñecido o estimador dos parámetros podemos facer as predicións para novos individuos: $\widehat{Y_i} = \mathbf{x_0}^T \widehat{\beta}$ onde $\mathbf{x_0} = (1, x_{0,1}, ..., x_{0,p-1})$. Para profundizar máis na clásica regresión en media pode verse [9].

1.4. Pequena introdución á regresión cuantil

Introducida xa a regresión en media (visto na Sección 1.3) e a estimación do cuantil (visto na Sección 1.2) estamos nas condicións de facer unha pequena presentación da regresión cuantil e das súas propiedades, que constitúen a base deste traballo.

Xa vimos que a análise de regresión é utilizada cando queremos concluir sobre a existencia de relacións entre unha variable resposta, Y, e un conxunto de variables explicativas, $X \in \mathbb{R}^{p-1}$. Como vimos na Sección 1.3, para este fin adóitase usar a regresión lineal en media. Este procedemento emprega o método de mínimos cadrados para a estimación da recta de regresión. O problema xurde cando no modelo existen observacións atípicas[9] (pois a recta regresión en media vese afectada por este tipo de datos) ou distribucións de erro non gaussianas ou escenarios non homocedásticos.

É ben sabido que os principais estimadores clásicos, como a media mostral ou os estimadores de regresión por mínimos cadrados, son poucos robustos. Isto debése a que poden sufrir grandes alteracións con un cantidade moi pequena de datos atípicos. Para solventar dito problema, pois os datos atípicos son moi comúns en situacións reais, desenvolvéronse novos principios e métodos, dentro do ámbito coñecido como Estatística Robusta. A regresión cuantil ten unha próxima relación con esta área posto que comparte algunhas propiedades e métodos coa mesma.

Outra gran vantaxe que alberga a regresión cuantil é que esta fíxase en toda a distribución condicional da variable resposta, mentres que a regresión en media só se centra no comportamento central dos datos.

1.4.1. Modelo lineal de regresión cuantil

Dada unha variable resposta Y, un conxunto de variables explicativas que denotaremos por $X = (X_1, ..., X_{p-1}) \in \mathbb{R}^{p-1}$, e fixado un cuantil de orde $\tau \in (0, 1)$, o modelo de regresión lineal múltiple cuantil defínese da seguinte forma:

$$Y = \beta_0^\tau + \beta_1^\tau X_1 + ... + \beta_{p-1}^\tau X_{p-1} + \epsilon$$

sendo $\beta^\tau = (\beta_0^\tau, \beta_1^\tau, ..., \beta_{p-1}^\tau)$ un vector de $\mathbb{R}^p$ e ϵ o erro que verifica $\mathbb{P}(\epsilon \leq 0 \mid X) = \tau$. Esta condición é análoga a hipótese de que o erro da regresión en media tivese esperanza condicionada nula, como vimos na Sección 1.3. Neste caso, pedimos que o cuantil condicional de ϵ de orde τ sexa 0.

[9] Un valor atípico é un dato que se afasta moito do comportamento esperado baixo o modelo. Profundizaremos máis neste concpeto na Sección 3.3.

A partir dagora consideraremos unha mostra aleatoria simple $\{(\mathbf{X_1}, Y_1), ..., (\mathbf{X_n}, Y_n)\}$ de tamaño n por comodidade, sendo $\mathbf{X}_i$ o vector que contén todas as variables explicativas, $\mathbf{X}_i = (1, X_{i,1}, ..., X_{i,p-1})$ con $i = 1, ..., n.$. Como vimos na Sección 1.2, os cuantís mostrais poden ser expresados como solución dun problema de optimización. De forma análaga, un estimador de β^τ pode obterse mediante a resolución do problema de optimización:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau (Y_i - \mathbf{X}_i^T \beta^\tau) \tag{1.3}$$

Esta idea foi introducida por Koenker e Basset, pode verse máis información en [8]. Daquí en diante denotaremos a dito estimador como $\widehat{\beta}^\tau$.

Capítulo 2

Propiedades da regresión cuantil

2.1. Cálculo dos estimadores mediante programación lineal

Vimos de ver que o modelo de regresión cuantil se pode estimar buscando a mínima perda cuantílica mediante a expresión (1.3). Esta estimación formúlase como un problema de programación lineal. Para isto engadimos $2n$ variables, $\{u_i, v_i : i = 1, ..., n\}$, que representan a parte positiva e a parte negativa do vector de residuos da regresión, $r_i = Y_i - \mathbf{X}_i^T \beta^\tau$. Entón definimos o novo problema como segue:

$$\min_{(\beta,u,v)\in\mathbb{R}^p\times\mathbb{R}_+^{2n}} \{\tau 1_n^T u + (1-\tau)1_n^T v \mid \mathbb{X}\beta^\tau + u - v = y\}$$

onde $\mathbb{X}$ denota a matriz de deseño da regresión que neste caso é unha matriz de dimensión $n \times p$.

É dicir, temos o problema de optimización seguinte:

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

suxeito a

$$\begin{aligned} & Y - \mathbb{X}\beta^\tau = u - v \\ & \beta^\tau \in \mathbb{R}^p \\ & u_i \geq 0, \;\; v_i \geq 0 \;\; \forall i \in \{1, ..., n\} \end{aligned}$$

Este problema de programación lineal[1] consiste en minimizar a función obxectivo,

[1] Debemos lembrar que o termo solución factible fai referencia a aquela solución que cumpre todas as restricións plantexadas no problema e que unha solución óptima é unha solución factible que optimiza a función obxectivo do problema. Esta última non ten porque ser única.

$\sum_{i=1}^{n}(\tau u_i + (1-\tau)v_i)$ co requisito de que as variables de decisión (β, u_i e v_i) satisfagan as tres restricións lineais propostas. Neste problema contamos con $p+2n$ variables, p coeficientes do vector β^τ e $2n$ variables referentes á u_i e v_i, e con n restricións.

Dado que o problema só ten n restricións, se existe solución, a solución óptima terá n variables non nulas: os p coeficientes do modelo e $n-p$ variables que corresponden aos residuos, u_i e v_i. Esto implica que dos n individuos hai p que teñen residuo cero. Por tanto, o modelo de regresión lineal cuantil pasa por p datos da mostra. Por exemplo, no caso particular de regresión lineal simple, a recta pasa por dous puntos da mostra como pode verse na Figura 2.1 para diferentes valores do cuantil τ de interese.

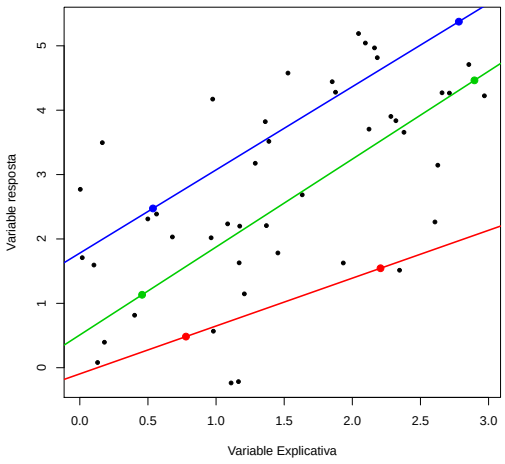


Figura 2.1: Rectas de regresión cuantil asociadas a diversos cuantís de interese que vemos que pasar por $p=2$ puntos. En vermello representamos a recta asociada a $\tau = 0'1$, en verde a asociada a $\tau = 0'4$ e en azul a asociada a $\tau = 0'8$.

De xeito análogo ao visto na Sección 1.3, onde a regresión en media constaba da propiedade de que a esperanza do residuo é cero, podemos enunciar o seguinte corolario (pódese ver en [7], páxina 37) para o caso da regresión cuantílica:

Corolario 2.1. *Sexan P e N as proporcións positivas e negativas do vector de residuos $y - X\widehat{\beta}^\tau$, respectivamente. Sexa Z o número de residuos cero, é dicir, $Z = p$. A proporción de residuos negativos é aproximadamente τ*

$$\frac{N}{n} \leq \tau \leq \frac{N+p}{n}$$

e a proporción de residuos positivos é aproximadamente $(1-\tau)$

$$\frac{P}{n} \leq 1-\tau \leq \frac{P+p}{n}$$

Disto obtemos as seguintes conclusións:

- O número de residuos menores estrictamente que cero é aproximadamente τ.
- O número de residuos maiores estrictamente que cero é aproximadamente $(1-\tau)$.
- O número de residuos iguais a cero é exactamente p.

2.2. Inferencia sobre os parámetros

Esta sección está adicada a coñecer a distribución dos estimadores dos parámetros da regresión cuantil. Para esto seguiremos [7] (páxina 120).

Consideremos unha mostra do modelo de regresión lineal cuantil $(\mathbf{X_1}, Y_1), ..., (\mathbf{X_n}, Y_n)$ onde recordemos que $\mathbf{X_i} = (1, X_{i,1}, ..., X_{i,p-1})$ con $i = 1, ..., n$. As variables $Y_1, ..., Y_n$ son independentes con funcións de distribución asociadas $F_1, ..., F_n$, respectivamente. Notemos que non é necesario que as funcións de densidade $f_1, ..., f_n$ sexan iguais (escenario heterocedástico). Fixemos o cuantil de orde $\tau \in (0,1)$. Supoñamos que o cuantil condicional,

$$Q_{Y_i}(\tau|X=x) = x^T\beta^\tau \ \ \forall i \in \{1, ...n\}$$

segue o modelo lineal. As funcións de distribución condicionais de cada Y_i denotaranse como $\mathbb{P}(Y_i < y|\mathbf{X_i}) = F_{Y_i}(y|\mathbf{X_i}) = F_i(y)$ e denotaremos ao cuantil de Y_i como $\xi_i(\tau) = Q_{Y_i}(\tau|\mathbf{X_i}) = F^{-1}_{Y_i}(\tau|\mathbf{X_i})$.

Consideraremos agora dúas condicións, **A1** e **A2**, necesarias para determinar o comportamento asintótico do estimador do vector de parámetros β^τ,

$$\widehat{\beta}^\tau = arg\min_{b\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X_i}^T b)$$

Condición A1. As funcións de distribución, F_i, son absolutamente continuas, con densidades continuas, $f_i(\xi)$, uniformemente acotadas (coa excepción do cero e do infinito) nos puntos $\xi_i(\tau)$, $\forall i \in \{1, ..., n\}$.

Condición A2. Existen dúas matrices, D_0 e $D_1(\tau)$, simétricas e definidas positivas verificando:

1. $\lim_{n\to\infty} n^{-1} \sum \mathbf{X_i}\mathbf{X_i}^T = D_0$
2. $\lim_{n\to\infty} n^{-1} \sum f_i(\xi_i(\tau))\mathbf{X_i}\mathbf{X_i}^T = D_1(\tau)$
3. $\max_{i=1,\ldots,n} \frac{\|\mathbf{X_i}\|}{\sqrt{n}} \to 0$

Introducimos agora unha propiedade moi importante da regresión cuantil baseada nas condicións anteriores.

Teorema 2.2. *Baixo as condicións **A1** e **A2**,*

$$\sqrt{n}(\widehat{\beta}^\tau - \beta^\tau) \longrightarrow N(0, \tau(1-\tau)D_1^{-1}D_0D_1^{-1})$$

No caso de que os erros sexan independentes e identicamente distribuídos tense que:

$$\sqrt{n}(\widehat{\beta}^\tau - \beta^\tau) \longrightarrow N(0, \omega^2 D_0^{-1})$$

sendo $\omega^2 = \frac{\tau(1-\tau)}{f^2(\xi_i(\tau))}$.

O papel que xoga a matriz D_0 é análogo ao da matriz de covarianzas da variable X, S_x^2, nos estimadores da regresión en media, visto na Sección 1.3. A citada matriz describe a dispersión da variable X.

Por outra banda, a matriz $D_1(\tau)$ ten un relevante rol ante unha situación na que os erros non son identicamente distribuidos. Cabe mencionar que mentres que na regresión en media pediamos a hipótese de homocedasticidade dos erros para a estimación dos parámetros, a regresión cuantil pode utilizarse cando o carácter dos erros é heterocedástico.

No caso de que os erros sexan independentes e teñan a mesma distribución, a estimación de β^τ depende inversamente da densidade da resposta avaliada no cuantil de interese, que se coñece como **función sparsity**. Esta cantidade describe a densidade de observacións que hai preto do cuantil de interese. Se os datos están moi dispersos nesa rexión a estimación será difícil e, polo tanto a varianza asintótica elevada. No caso contrario, cando hai unha cantidade inxente de observacións próximas ao cuantil fixado, a estimación do cuantil será máis exacta. É dicir, cando a *sparsity* é baixa obtemos unha menor varianza asintótica do estimador e como consecuencia unha mellor aproximación de $\widehat{\beta}^\tau$. A *función sparsity* xoga un papel análogo ao da desviación típica dos erros na estimación por mínimos cadrados.

2.3. Robustez

Para o estudo da robustez dos metodos cuantís centrarémonos na **función de influencia** dos estimadores. Seguindo a notación de [7] (páxina 42), imos definir tal función.

Definición 2.3. Sexa $\widehat{\theta}(F)$ un estimador dun parámetro θ que caracteriza unha variable con función de distribución F. Se modificamos a distribución F do seguinte xeito:

$$F_\alpha = \alpha\delta_y + (1-\alpha)F$$

onde δ_y denota a función característa do punto y, unha pequena probabilidade α de datos pasan a tomar o valor y. Isto é o que se coñece no ámbito da Estatística Robusta como **función contaminada**.

Agora estamos nas condicións para poder definir a función de influencia de $\widehat{\theta}(F)$ como segue:

$$IF_{\widehat{\theta}}(y,F) = \lim_{\alpha\to 0} \frac{\widehat{\theta}(F_\alpha) - \widehat{\theta}(F)}{\alpha}$$

Por exemplo, para a media temos que:

$$\bar{\theta}(F_\alpha) = \int y dF_\alpha = \int y d(\alpha\delta_y + (1-\alpha)F(y)) = \alpha y + (1-\alpha)\bar{\theta}(F)$$

e polo tanto a súa función de influencia será:

$$IF_{\bar{\theta}}(y,F) = y - \bar{\theta}(F)$$

Mentres que para a mediana o estimador ven dado por:

$$\tilde{\theta}(F_\alpha) = F_\alpha^{-1}\left(\frac{1}{2}\right)$$

e a función de influencia será:

$$IF_{\tilde{\theta}}(y,F) = \operatorname{sgn}\frac{(y-\tilde{\theta}(F))}{f(F^{-1}\left(\frac{1}{2}\right))} \tag{2.1}$$

sendo f a función de densidade asociada á función de distribución F e sgn a función signo.

Comparando as anteriores expresións podemos concluír que no caso da media a influencia da contaminación de F en y é proporcional a y. É dicir, se alonxamos o punto y poderiamos cambiar a estimación da media tanto como queiramos. En cambio, a influencia da contaminación en y na mediana está delimitado pola constante $s(\frac{1}{2}) = \frac{1}{f(F^{-1}(\frac{1}{2}))}$, que como dixemos anteriormente denomínase *sparsity*.

Na Figura 2.2 podemos observar a función de influencia para a media e a mediana. Nesta móstrase a robustez desta última fronte á 'contaminación' das observacións. Isto xustifica a influencia de datos atípicos na media fronte á robustez da mediana.

A función de influencia para o cuantil de orde τ obténse substituíndo $\frac{1}{2}$ na ecuación (2.1) por τ. Pódese profundizar máis na función de influencia para un cuantil de orde τ en [7] (páxina 44).

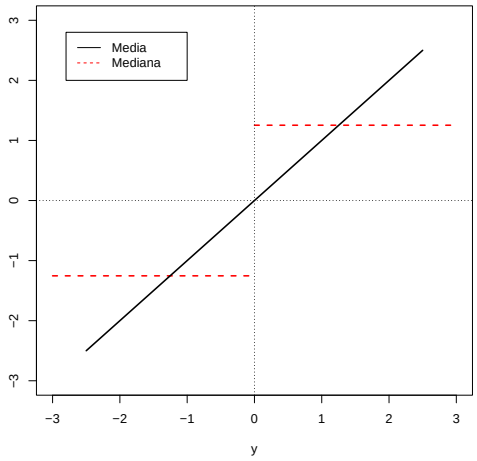


Figura 2.2: Funcións de influencia para a media e a mediana

2.3.1. A función de influencia dos estimadores de regresión

A idea de función de influencia pode estenderse ao contexto da regresión. Aínda que non se trata do obxectivo deste traballo, introduciremos brevemente as funcións de influencia asociadas a modelos de regresión en media e mediana. En [5], páxina 106, podemos atopar o cálculo da función de influencia do estimador de mínimos cadrados que ven dado por:

$$IF((x,y),\widehat{\beta}_{MC},F) = \mathbb{E}(\mathbb{X}\mathbb{X}')^{-1}x(y - x'\widehat{\beta}_{MC}(F))$$

onde F representaría a distribución de probabilidade do vector aleatorio (X,Y). Á vista da ecuación anterior, e na liña do vista para a media mostral, destacamos a influencia debida ao residuo, é dicir, a ocasionada pola desviación da resposta y respecto do que lle correspondería segundo o modelo axustado.

Por outra banda, a función de influencia da regresión cuantil ven dada por

$$IF_{\widehat{\beta}^\tau}((x,y),F) = Q^{-1}\ x\ \mathrm{sgn}(y - x^T\widehat{\beta}^\tau)$$

onde

$$Q = \int xx^T\ f(x^T\widehat{\beta}^\tau|x)\ dG(x)$$

sendo f a función de densidade condicional da resposta dada a covariable e asumindo que $dF = dG(x)f(y|x)dy$. Podemos entón observar que a regresión cuantil non se ve influenciada pola magnitude dos residuos como a regresión en media.

En calquera caso, ambos estimadores sofren da influencia da posición de x no espazo das variables explicativas. Así, se x está moi alonxado dos valores previsibles, o par (x, y) influirá máis no axuste. Isto está moi relacionado co concepto do **efecto panca** en regresión. Para solventar dito problema debemos recurrir a **M-estimadores generalizados** que se atopan completamente fóra do alcance deste traballo.

2.4. Cruce entre cuantís

Nesta sección exporemos unha pequena 'debilidade' que sofre a regresión cuantil. Como vimos na conclusión da Sección 2.2 só a información local é relevante en regresión cuantil, é dicir, só nos centramos nos datos próximos ao cuantil de estudo.

Dado que as curvas de regresión cuantil son estimadas de forma individual, as curvas cuantílicas poden cruzarse, o que nos leva a unha distribución non válida da resposta. O feito de que ditas curvas se crucen contradí o principio básico de que as funcións de distribución e as súas respectivas inversas deben ser monótonas crecentes. Na Figura 2.3 podemos ver un exemplo dunha situación onde observamos o cruce das funcións cuantís.

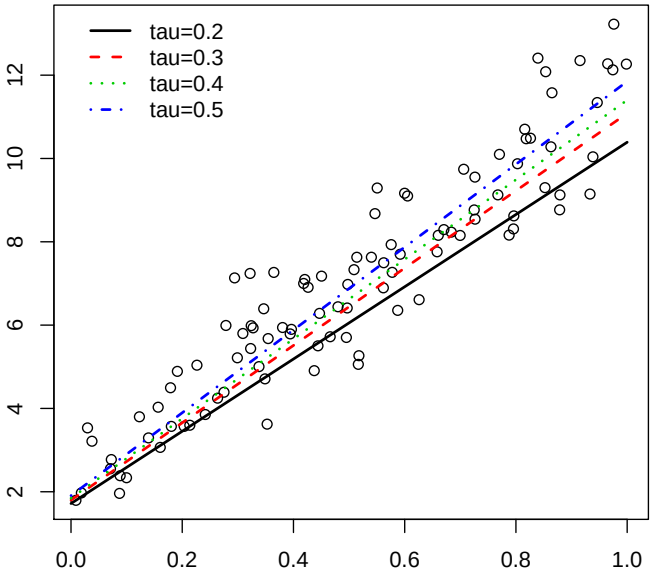


Figura 2.3: Rectas de regresión cuantil asociadas a diversos cuantís de interese.

O certo é que o mencionado cruce prodúcese nas colas da distribución. Se denotamos por $\bar{X} = (\bar{X}_1, \dots, \bar{X}_{p-1})$ ao centroide do deseño onde

$$\bar{X}_j = \frac{1}{n} \sum_{i=1}^{n} X_{i,j}$$

representa a media mostral asociada a cada variable X_j. O teorema que introducimos a continuación mostra que no centroide do diseño, a estimación da función cuantil condicional

$$\widehat{Q}_Y(\tau|\bar{X}) = (1, \bar{X})^T \widehat{\beta}^\tau$$

é monótona en τ.

Teorema 2.4. *A función estimada $\widehat{Q}_Y(\tau|\bar{X})$ é non decrecente en $\tau \in (0, 1)$.*

Demostración. A demostración do teorema pode verse en [7] (páxina 56). □

A monotonía en $X = \bar{X}$ non garantiza que $\widehat{Q}_Y(\tau|\bar{X})$ sexa monótona en τ para outros valores de X. Se $\widehat{Q}_Y$ é lineal, o cruce darase o suficientemente lonxe de $\bar{X}$. No caso de que o cruce cuantil se produza será fora da envolvente convexa das observacións X e o modelo estimado non será adecuado nesa rexión. Nótese como na Figura 2.3 observamos o cruce cerca da fronteira dos datos.

Capítulo 3

Estudo de simulación con R

Este capítulo está adicado a analizar os modelos de regresión cuantil dende un punto de vista práctico, para o que empregaremos un estudo de simulación. Na Sección 3.1 explicamos como se realizou dito estudo mediante o software estatístico libre R. Na Sección 3.2 ilustraremos o Teorema 2.2 mediante un exemplo levado a cabo con dito programa. E na seguinte Sección 3.3 mostraremos cando é mellor o uso da regresión mediana fronte á regresión en media e tamén poderemos observar, mediante a axuda de gráficas, a robustez que presenta a regresión cuantil fronte a aqueles datos que non esperaríamos no noso modelo. É dicir, ilustremos as ideas teóricas derivadas da análise da función de influencia no contexto da regresión.

3.1. Realización das simulacións con R

Os estudos de simulación que presentaremos a continuación nas Seccións 3.2 e 3.3 son deseñados usando a linguaxe R, que é un programa de software libre orientado á análise de datos estatísticos. Pode descargarse ou consultar máis información do software R na seguinte dirección `https://www.r-project.org`. Unha introducción a dito programa pode atoparse en [6] e [11].

Ademais, será fundamental para o desenrolo dos estudos de simulación, o uso do paquete *quantreg* orientado a métodos de estimación e inferencia para modelos de regresión cuantil e deseñado por Koenker. Toda a información sobre dito paquete pode atoparse na páxina `cran.r-project.org/web/packages/quantreg`.

3.2. Os estimadores da regresión cuantil

Como vimos no Teorema 2.2 da Sección 2.2, canto maior sexa a densidade da variable resposta Y avaliada no cuantil de interese, mellor será a aproximación de $\widehat{\beta}^{\tau}$ a β^{τ} en termos de varianza asintótica. Vexamos agora este resultado dunha forma práctica.

Dada unha mostra aleatoria simple $\{(X_i, Y_i)\}_{i=1}^{n}$ do modelo lineal

$$Y = 3 + 5X + (\epsilon - c_{\tau}) \tag{3.1}$$

onde X segue unha distribución uniforme no intervalo $[0, 1]$, ϵ é o erro do modelo que seguirá distintas distribucións e c_{τ} o cuantil condicional de orde τ de ϵ dado un valor da variable X. Neste caso, dito erro é independente da variable explicativa X, é dicir, o modelo é homocedástico. As distribucións de erro que usaremos nesta práctica serán as seguintes:

- Distribución normal con media 0 e varianza 1, o que se coñece como distribución normal estándar, que denotaremos por $\epsilon \sim N(0, 1)$.
- Distribución uniforme continua no intervalo $[-1, 1]$, que denotaremos por $\epsilon \sim U[-1, 1]$.
- Distribución *chi*-cadrado con 2 graos de liberdade, que denotaremos por $\epsilon \sim \chi^2_2$.
- Distribución *log*-normal[1] caracterizada polos parámetros $\tilde{\mu} = 0$ e $\tilde{\sigma} = 1$, que denotaremos por $\epsilon \sim logN(0, 1)$.

Na Figura 3.1 representamos as distintas distribucións de erro consideradas para poder visualizar como varía a densidade do erro avaliada no cuantil de interese nos diferentes escenarios que imos considerar. Notemos que se mantén a mesma escala en todas as gráficas para facilitar a comparación entre as distintas densidades de erro citadas anteriormente.

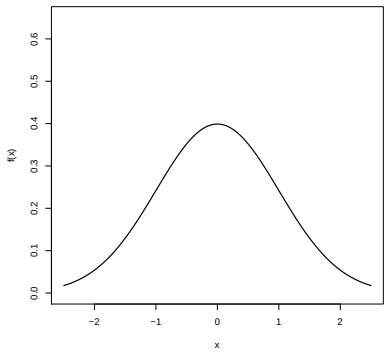
O obxectivo principal deste estudo de simulación será mostrar a calidade de distintas estimacións dos parámetros β_0^{τ} e β_1^{τ} en función do cuantil de interese e para distintos tamaños de mostra ($n = 50$ e $n = 100$). Para medir dita calidade calcularemos os erros cadráticos medios (ECM) dos estimadores obtidos. Recordemos que:

Definición 3.1. Sexa θ un certo parámetro e denotemos por $\hat{\theta}$ un estimador do mesmo. Definimos o erro cadrático medio (ECM) do estimador $\hat{\theta}$ respecto de θ como segue:

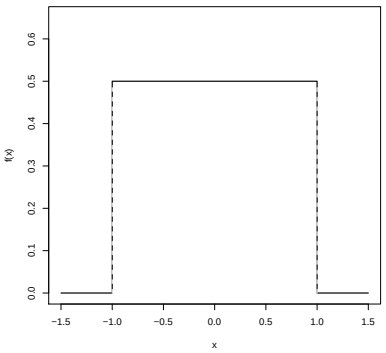
$$\mathrm{ECM}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathrm{Var}(\hat{\theta}) + \mathrm{Sesgo}^2(\hat{\theta})$$

onde $\mathrm{Sesgo}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$.

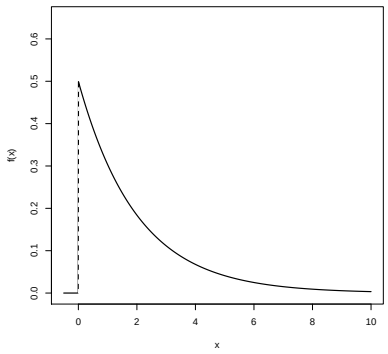
[1] Dada unha variable Z diremos segue unha distribución *log*-normal de parámetros $\tilde{\mu}$ e $\tilde{\sigma}$, se $log(Z)$ segue unha distribución normal de media $\tilde{\mu}$ e desviación típica $\tilde{\sigma}$.



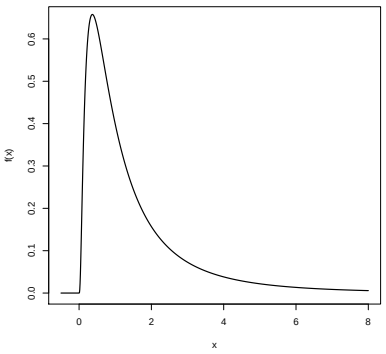
(a) Distribución normal estándar



(b) Distribución uniforme no intervalo $(-1, 1)$



(c) Distribución χ^2



(d) Distribución log-normal estándar

Figura 3.1: Gráficas das funcións de densidade asociadas ás distintas distribucións do erro utilizadas neste estudo: $N(0,1)$, $U[-1,1]$, χ^2_2 e $logN(0,1)$.

Para calcular os ECM dos estimadores $\widehat{\beta_0}^{\tau}$ e $\widehat{\beta_1}^{\tau}$ utilízase o **método de Monte Carlo**, que nos permite aproximar estes erros cadráticos medios. Detallamos a continuación o procedemento a seguir:

1. Xeramos unha mostra do modelo (3.1) de tamaño n.
2. Calculamos os estimadores $\widehat{\beta_0}^{\tau}$ e $\widehat{\beta_1}^{\tau}$ para a mostra obtida.
3. Repetimos os pasos 1 e 2 M veces, $M = 1000$ por exemplo, e isto permitiramos estimar o sesgo, a varianza e o ECM dos M estimadores que xeramos comparándoos cos valores reais β_0^{τ} e β_1^{τ} (que coñecemos por tratarse dun estudo de simulación).

Na Sección A.1 pode verse o código de R necesario para levar a cabo este estudo de

simulación. Ademais, as Táboas 3.1 e 3.2 móstrase o sesgo, a varianza e o erro cadrático medio (ECM) dos estimadores $\widehat{\beta_0}^{\tau}$ e $\widehat{\beta_1}^{\tau}$ para distintos valores do cuantil de interese, de tamaños de mostra e da distribución de erro.

			$\epsilon \sim N(0,1)$			$\epsilon \sim U[-1,1]$		
τ	$\widehat{\beta}^{\tau}$	n	Sesgo	Varianza	ECM	Sesgo	Varianza	ECM
0.10	$\widehat{\beta_0}^{\tau}$	50	$1'583 \times 10^{-2}$	$2'230 \times 10^{-1}$	$2'232 \times 10^{-1}$	$3'377 \times 10^{-2}$	$3'099 \times 10^{-2}$	$3'213 \times 10^{-2}$
		100	$8'689 \times 10^{-3}$	$1'175 \times 10^{-1}$	$1'175 \times 10^{-1}$	$1'227 \times 10^{-2}$	$1'477 \times 10^{-2}$	$1'493 \times 10^{-2}$
	$\widehat{\beta_1}^{\tau}$	50	$5'165 \times 10^{-3}$	$7'319 \times 10^{-1}$	$7'319 \times 10^{-1}$	$-6'848 \times 10^{-3}$	$9'047 \times 10^{-2}$	$9'052 \times 10^{-2}$
		100	$-6'203 \times 10^{-3}$	$3'655 \times 10^{-1}$	$3'656 \times 10^{-1}$	$2'918 \times 10^{-3}$	$4'374 \times 10^{-2}$	$4'375 \times 10^{-2}$
0.25	$\widehat{\beta_0}^{\tau}$	50	$-4'771 \times 10^{-3}$	$1'488 \times 10^{-1}$	$1'488 \times 10^{-1}$	$1'643 \times 10^{-2}$	$5'617 \times 10^{-2}$	$5'644 \times 10^{-2}$
		100	$-3'403 \times 10^{-3}$	$7'998 \times 10^{-2}$	$7'999 \times 10^{-2}$	$1'380 \times 10^{-2}$	$3'061 \times 10^{-2}$	$3'080 \times 10^{-2}$
	$\widehat{\beta_1}^{\tau}$	50	$2'084 \times 10^{-2}$	$4'628 \times 10^{-1}$	$4'632 \times 10^{-1}$	$2'561 \times 10^{-3}$	$1'717 \times 10^{-1}$	$1'717 \times 10^{-1}$
		100	$7'057 \times 10^{-3}$	$2'423 \times 10^{-1}$	$2'424 \times 10^{-1}$	$-8'754 \times 10^{-3}$	$9'443 \times 10^{-2}$	$9'450 \times 10^{-2}$
0.50	$\widehat{\beta_0}^{\tau}$	50	$-1'219 \times 10^{-3}$	$1'323 \times 10^{-1}$	$1'323 \times 10^{-1}$	$-1'481 \times 10^{-2}$	$7'768 \times 10^{-2}$	$7'790 \times 10^{-2}$
		100	$7'801 \times 10^{-3}$	$6'556 \times 10^{-2}$	$6'562 \times 10^{-2}$	$4'231 \times 10^{-3}$	$4'124 \times 10^{-2}$	$4'125 \times 10^{-2}$
	$\widehat{\beta_1}^{\tau}$	50	$6'860 \times 10^{-3}$	$4'154 \times 10^{-1}$	$4'155 \times 10^{-1}$	$2'560 \times 10^{-2}$	$2'424 \times 10^{-1}$	$2'430 \times 10^{-1}$
		100	$-6'101 \times 10^{-3}$	$2'008 \times 10^{-1}$	$2'008 \times 10^{-1}$	$-1'119 \times 10^{-2}$	$1'172 \times 10^{-1}$	$1'173 \times 10^{-1}$
0.75	$\widehat{\beta_0}^{\tau}$	50	$-8'155 \times 10^{-3}$	$1'466 \times 10^{-1}$	$1'467 \times 10^{-1}$	$-2'642 \times 10^{-2}$	$5'942 \times 10^{-2}$	$6'012 \times 10^{-2}$
		100	$-6'123 \times 10^{-5}$	$7'670 \times 10^{-2}$	$7'670 \times 10^{-2}$	$-7'164 \times 10^{-3}$	$3'102 \times 10^{-2}$	$3'107 \times 10^{-2}$
	$\widehat{\beta_1}^{\tau}$	50	$7'730 \times 10^{-3}$	$4'407 \times 10^{-1}$	$4'408 \times 10^{-1}$	$1'939 \times 10^{-2}$	$1'823 \times 10^{-1}$	$1'826 \times 10^{-1}$
		100	$-3'957 \times 10^{-3}$	$2'334 \times 10^{-1}$	$2'334 \times 10^{-1}$	$-7'782 \times 10^{-3}$	$9'134 \times 10^{-2}$	$9'140 \times 10^{-2}$
0.90	$\widehat{\beta_0}^{\tau}$	50	$-9'095 \times 10^{-3}$	$2'514 \times 10^{-1}$	$2'515 \times 10^{-1}$	$-2'780 \times 10^{-2}$	$3'267 \times 10^{-2}$	$3'344 \times 10^{-2}$
		100	$-8'128 \times 10^{-3}$	$1'195 \times 10^{-1}$	$1'196 \times 10^{-1}$	$-1'774 \times 10^{-2}$	$1'510 \times 10^{-2}$	$1'541 \times 10^{-2}$
	$\widehat{\beta_1}^{\tau}$	50	$1'170 \times 10^{-2}$	$7'515 \times 10^{-1}$	$7'516 \times 10^{-1}$	$-8'439 \times 10^{-3}$	$9'843 \times 10^{-2}$	$9'850 \times 10^{-2}$
		100	$3'252 \times 10^{-3}$	$3'582 \times 10^{-1}$	$3'582 \times 10^{-1}$	$-3'293 \times 10^{-3}$	$4'347 \times 10^{-2}$	$4'348 \times 10^{-2}$

Táboa 3.1: Sesgo, varianza e erro cadrático medio dos estimadores asociados ao modelo de regresión cuantil (3.1) para distintas distribucións do erro, distintos tamaños de mostra e distintos ordes do cuantil de interese.

Así, fixándonos nos ECM nas Táboas 3.1 e 3.2, podemos ver o efecto da densidade

			$\epsilon \sim \chi^2_2$			$\epsilon \sim logN(0,1)$		
τ	$\widehat{\beta}^\tau$	n	Sesgo	Varianza	ECM	Sesgo	Varianza	ECM
0.10	$\widehat{\beta}_0{}^\tau$	50	$5'367 \times 10^{-2}$	$4'622 \times 10^{-2}$	$4'911 \times 10^{-2}$	$1'867 \times 10^{-2}$	$1'974 \times 10^{-2}$	$2'009 \times 10^{-2}$
		100	$2'200 \times 10^{-2}$	$2'239 \times 10^{-2}$	$2'287 \times 10^{-2}$	$1'033 \times 10^{-2}$	$9'694 \times 10^{-3}$	$9'801 \times 10^{-3}$
	$\widehat{\beta}_1{}^\tau$	50	$-2'263 \times 10^{-2}$	$1'306 \times 10^{-1}$	$1'311 \times 10^{-1}$	$4'266 \times 10^{-3}$	$6'539 \times 10^{-2}$	$6'540 \times 10^{-2}$
		100	$-3'205 \times 10^{-3}$	$6'460 \times 10^{-2}$	$6'461 \times 10^{-2}$	$-1'185 \times 10^{-3}$	$2'964 \times 10^{-2}$	$2'964 \times 10^{-2}$
0.25	$\widehat{\beta}_0{}^\tau$	50	$6'434 \times 10^{-2}$	$1'175 \times 10^{-1}$	$1'217 \times 10^{-1}$	$1'482 \times 10^{-2}$	$3'972 \times 10^{-2}$	$3'994 \times 10^{-2}$
		100	$2'994 \times 10^{-2}$	$5'921 \times 10^{-2}$	$6'010 \times 10^{-2}$	$8'011 \times 10^{-3}$	$2'087 \times 10^{-2}$	$2'093 \times 10^{-2}$
	$\widehat{\beta}_1{}^\tau$	50	$-4'594 \times 10^{-2}$	$3'356 \times 10^{-1}$	$3'377 \times^{-1}$	$1'258 \times 10^{-2}$	$1'237 \times 10^{-1}$	$1'238 \times 10^{-1}$
		100	$-2'555 \times 10^{-2}$	$1'752 \times 10^{-1}$	$1'759 \times 10^{-1}$	$4'078 \times 10^{-3}$	$6'462 \times 10^{-2}$	$6'464 \times 10^{-2}$
0.50	$\widehat{\beta}_0{}^\tau$	50	$9'096 \times 10^{-2}$	$3'275 \times 10^{-1}$	$3'358 \times 10^{-1}$	$3'040 \times 10^{-2}$	$1'416 \times 10^{-1}$	$1'426 \times 10^{-1}$
		100	$3'789 \times 10^{-2}$	$1'697 \times 10^{-1}$	$1'711 \times 10^{-1}$	$2'460 \times 10^{-2}$	$6'721 \times 10^{-2}$	$6'782 \times 10^{-2}$
	$\widehat{\beta}_1{}^\tau$	50	$-9'764 \times 10^{-2}$	$9'635 \times 10^{-1}$	$9'730 \times 10^{-1}$	$6'611 \times 10^{-3}$	$4'356 \times 10^{-1}$	$4'357 \times 10^{-1}$
		100	$-2'820 \times 10^{-2}$	$5'294 \times 10^{-1}$	$5'302 \times 10^{-1}$	$-8'565 \times 10^{-3}$	$2'052 \times 10^{-1}$	$2'053 \times 10^{-1}$
0.75	$\widehat{\beta}_0{}^\tau$	50	$5'858 \times 10^{-2}$	$9'518 \times 10^{-1}$	$9'552 \times 10^{-1}$	$5'790 \times 10^{-2}$	$5'761 \times 10^{-1}$	$5'795 \times 10^{-1}$
		100	$2'126 \times 10^{-2}$	$5'080 \times 10^{-1}$	$5'085 \times 10^{-1}$	$3'848 \times 10^{-2}$	$3'001 \times 10^{-1}$	$3'016 \times 10^{-1}$
	$\widehat{\beta}_1{}^\tau$	50	$-4'897 \times 10^{-2}$	$2'903$	$2'905$	$9'800 \times 10^{-3}$	$1'730$	$1'730$
		100	$3'061 \times 10^{-2}$	$1'480$	$1'481$	$-1'095 \times 10^{-2}$	$9'029 \times 10^{-1}$	$9'030 \times 10^{-1}$
0.90	$\widehat{\beta}_0{}^\tau$	50	$8'602 \times 10^{-2}$	$2'866$	$2'873$	$1'942 \times 10^{-1}$	$3'665$	$3'703$
		100	$7'280 \times 10^{-3}$	$1'568$	$1'568$	$8'332 \times 10^{-2}$	$1'646$	$1'653$
	$\widehat{\beta}_1{}^\tau$	50	$-1'161 \times 10^{-1}$	$8'506$	$8'519$	$1'145 \times 10^{-2}$	$1'028 \times 10^{1}$	$1'028 \times 10^{1}$
		100	$9'031 \times 10^{-2}$	$4'547$	$4'556$	$-4'841 \times 10^{-3}$	$4'850$	$4'850$

Táboa 3.2: Sesgo, varianza e erro cadrático medio dos estimadores asociados ao modelo de regresión cuantil (3.1) para distintas distribucións do erro, distintos tamaños de mostra e distintos ordes do cuantil de interese.

avaliada no cuantil de interese[2]. Por exemplo, á vista da Táboa 3.1, observamos como para $\tau = 0'5$ os ECM son menores cando $\epsilon \sim U[-1,1]$ que cando $\epsilon \sim N(0,1)$ dado que a densidade da $U[-1,1]$ no cuantil $0'5$ é $0'5$ mentres que no caso da normal é aproximadamente $0'4$. Situacións similares obsérvanse cando comparamos o comportamento do cuantil $0'1$

[2]Recordemos que vimos que no caso de que os erros sexan independentes e identicamente distribuídos (como é este caso) tense que:

$$\sqrt{n}(\widehat{\beta}_n^\tau - \beta^\tau) \longrightarrow N(0, \omega^2 D_0^{-1})$$

sendo $\omega^2 = \frac{\tau(1-\tau)}{f^2(c_\tau)}$ e $D_0 = \lim_{n\to\infty} n^{-1} \sum \mathbf{X}_i \mathbf{X}_i^T$.

	$\tau = 0'10$	$\tau = 0'25$	$\tau = 0'50$	$\tau = 0'75$	$\tau = 0'90$
$\epsilon \sim N(0,1)$	0'175	0'318	0'399	0'318	0'175
$\epsilon \sim U[-1,1]$	0'500	0'500	0'500	0'500	0'500
$\epsilon \sim \chi^2_2$	0'450	0'375	0'250	0'125	0'050
$\epsilon \sim logN(0,1)$	0'632	0'624	0'399	0'162	0'049

Táboa 3.3: Densidades das distintas distribucións de erro avaliadas nos correspondentes cuantís de orde τ, para diferentes valores do cuantil de interese.

cando $\epsilon \sim N(0,1)$ ou $\epsilon \sim \chi^2_2$ ou do cuantil 0'75 cando $\epsilon \sim N(0,1)$ ou $\epsilon \sim logN(0,1)$. De cara a clarificar estas situacións, mostramos na Táboa 3.3 os valores das densidades do erro avaliadas nos correspondentes cuantís de orde τ.

Por outra banda, fixándonos no caso de $\epsilon \sim U[-1,1]$ vemos que hai un claro comportamento de 'simetría' respecto da mediana dado que nese caso á densidade é constante e o factor $\tau(1-\tau)$ é o que marca a diferenza entre os distintos cuantís. Ademais, non sempre a regresión en mediana ten asociados os menores erros cadráticos, véxase por exemplo o caso $\epsilon \sim \chi^2_2$ onde a mellor estimación dos parámetros dase na regresión cuantil con $\tau = 0'1$. Finalmente, debemos destacar que a medida que aumenta o tamaño de mostra, os ECM diminúen en todos os escenarios considerados.

3.3. Regresión en media versus regresión en mediana

A continuación presentamos un novo estudo de simulación que nos permitará comparar a regresión en media e a regresión en mediana. Ao longo deste estudo será fundamental a análise de datos atípicos co que poderemos ver o impacto que teñen os mesmos sobre a recta de regresión en media fronte á robustez da recta de regresión en mediana. Para este fin introduciremos agora certos conceptos necesarios para este estudo. Unha **observación atípica** é un dato que se afasta moito do comportamento esperado baixo o modelo. Falaremos agora do efecto panca polo cal uns datos teñen máis influencia que outros no axuste do modelo. Recordemos que o axuste do modelo en media obtense da seguinte forma $\widehat{Y} = HY$ sendo $H = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$ a '*matriz hat*' do modelo e onde $\mathbb{X}$ é a matriz de deseño. Aos elementos correspondentes á diagonal de H, h_{ii}, coñécense como '*leverages*'. Canto máis grande sexa esta cantidade, máis pequena será a varianza do residuo posto que $Var(\widehat{\epsilon}) = \sigma^2(1 - h_{ii})$, onde σ^2 é a desviación típica do erro. Isto significa que o axuste da

regresión en media está obrigado a aproximarse á observación Y_i para que así o residuo sexa o máis pequeno posible.

Como vimos de ver, as cantidades h_{ii} conforman a diagonal da matriz H e tal matriz obtense da matriz de deseño $\mathbb{X}$. Así pois, os individuos con valores $\mathbf{X}_i = (1, X_{i,1}, ..., X_{i,p-1})$, con $i = 1, ..., n$, moi alonxados dos datos, é dicir, os valores extremos de X, terán '*leverages*' grandes e como consecuencia posúen unha forte influencia no axuste pois obrígano a pasar o máis preto posible do valor da variable resposta. Os residuos estandarizados, $r_i = \frac{\widehat{\epsilon_i}}{\widehat{\sigma}\sqrt{(1-h_{ii})}}$, serán útiles para a búsqueda de datos atípicos posto que un dato considérase atípico se o seu residuo estandarizado é maior, en valor absoluto, a 2. Tendo definidos estes conceptos realizamos agora o estudo da influencia de datos atípicos e o efecto panca.

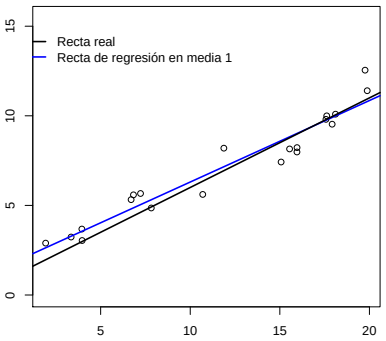
Dada unha mostra aleatoria simple $\{(X_1, Y_1), ..., (X_{20}, Y_{20})\}$ do modelo lineal

$$Y = 1 + 0'5X + \epsilon \tag{3.2}$$

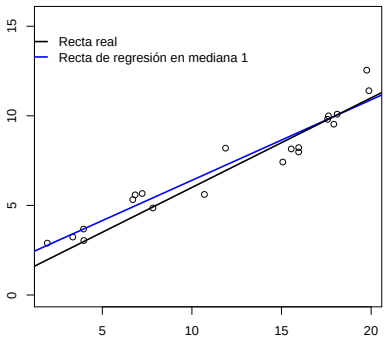
onde X segue unha distribución uniforme no intervalo $(0, 20)$ e ϵ representa o erro do modelo que segue unha distribución normal estándar e é independente da variable explicativa X. Notemos que este escenario é óptimo para a regresión en media. Na Figura 3.2 (a), podemos ver a recta real en negro e a recta axustada de regresión en media en azul para este conxunto de datos. Na Figura 3.2 (b) móstranse a recta real en negro e a recta axustada de regresión en mediana en azul. As tres rectas son moi semellantes, como cabería esperar, pois o $\epsilon \sim N(0, 1)$.

Xeramos agora un novo dato con compoñente ordenada $x_0 = 30$ e a con abscisa, $Y_0 - 8$, sendo Y_0 a predición do modelo para dito punto. Tal dato pódese ver nas Figuras 3.3 (a) e (b) en cor vermello. Nestas mesmas gráficas podemos ver en (a) o diagrama de dispersión do modelo (3.2) coa recta real en negro, a recta de regresión en media para o modelo orixinal en azul e a recta de regresión en media para o modelo con este novo dato en vermello e na figura (b) a recta real en negro, a recta de regresión en mediana do modelo orixinal en azul e a recta de regresión en mediana para o modelo engadindo este novo dato en vermello. Aínda que o dato $(x_0, Y_0 - 8)$ é atípico, con residuo estandarizado $-3'87$, a recta de regresión en media sofre un desprazamento en contraste coa recta de regresión en mediana.

Consideramos agora outro dato con ordenada $x_0 = 30$ e con abscisa $Y_0 - 28$. Este dato tamén é atípico, con residuo estandarizado $-4'31$. Na Figura 3.3 (c) móstranse o novo punto en vermello, a recta real en negro, a recta de regresión en media do modelo orixinal en azul e a recta de regresión en media do modelo co dato $(x_0, Y_0 - 28)$ en vermello. Da



(a) Regresión en media.



(b) Regresión en mediana.

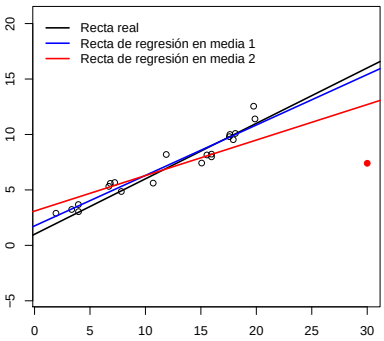
Figura 3.2: Diagramas de dispersión do modelo (3.2) de 20 datos coas rectas reais en negro, e rectas estimadas grazas a modelos de regresión en media (figura (a)) e de regresión en mediana en azul (figura (b)).

mesma forma, na Figura 3.3 (d) podemos ver o novo punto en vermello, a recta real en negro, a recta de regresión en mediana do modelo orixinal en azul e a recta de regresión en mediana para o modelo co novo dato en vermello. Comparando ambas gráficas, (c) e (d), comprobamos que o novo dato produce un efecto panca moi pronunciado na recta de regresión en media, mentres a recta de regresión en mediana permanece indemne.

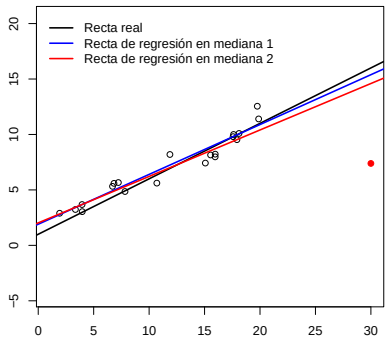
Debemos notar que canto maior sexa o número de datos da mostra será menor o efecto panca que se produce na recta de regresión en media. Por exemplo, se tomamos o mesmo modelo (3.2) cun tamaño de mostra $n = 50$ e realizamos o mesmo análise exposto anteriormente obtemos as gráficas que podemos ver na Figura 3.4.

Cando engadimos o dato $(x_0, Y_0 - 8)$, a recta de regresión en media sofre unha menor desviación, como podemos ver na Figura 3.4 (a), a pesar de que dito dato é atípico cun residuo estandarizado $-5'13$. En cambio, cando engadimos o dato $(x_0, Y_0 - 28)$, que tamén é atípico con residuo estandarizado $-6'77$, a recta de regresión en media, que se pode ver na Figura 3.4 (c), sofre unha gran desviación debido o efecto panca. Aínda así, se comparamos as Figuras 3.3 e 3.4 vemos claramente como cando o tamaño de mostra é menor o efecto panca que sofre a recta de regresión en media do modelo (3.2) é máis notable.

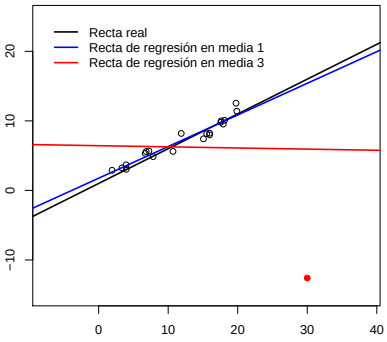
Agora faremos un estudo de simulación para comparar a regresión en media fronte



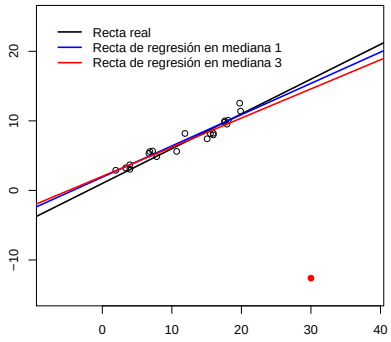
(a) Diagrama de dispersión co punto $(x_0, Y_0 - 8)$ e a regresión en media.



(b) Diagrama de dispersión co punto $(x_0, Y_0 - 8)$ coa recta de regresión en mediana.



(c) Diagrama de dispersión co punto $(x_0, Y_0 - 28)$ coa recta de regresión en media.



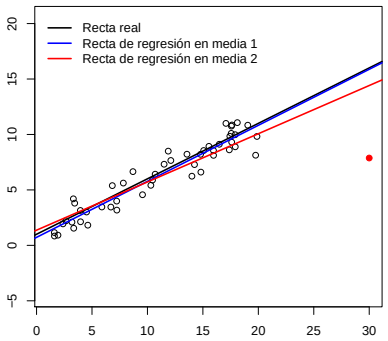
(d) Diagrama de dispersión co punto $(x_0, Y_0 - 28)$ coa recta de regresión en mediana.

Figura 3.3: Diagrama de dispersión do modelo (3.2), engadindo un novo punto, que aparecerá en vermello, e as rectas real en negro, de regresión en media do modelo orixinal e de regresión en mediana do modelo orixinal en azul e a recta de regresión de media e regresión en mediana en vermello.

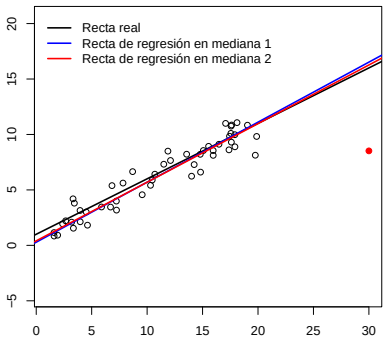
a regresión en mediana. Tal comparación farase a través dos erros de predición, aquelas cantidades que miden a diferenza entre o valor real e a súa estimación. Canto menor sexa esta cantidade mellor será a estimación do dato.

Dada unha mostra aleatoria simple $\{(X_i, Y_i)\}_{i=1}^n$, usaremos dous criterios:

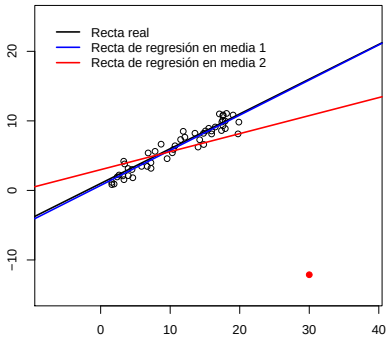
- Erro cadrático medio (ECM): $\sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$



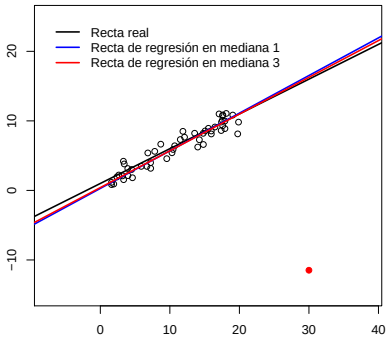
(a) Diagrama de dispersión co punto $(x_0, Y_0 - 8)$ e a regresión en media.



(b) Diagrama de dispersión co punto $(x_0, Y_0 - 8)$ coa recta de regresión en mediana.



(c) Diagrama de dispersión co punto $(x_0, Y_0 - 28)$ coa recta de regresión en media.



(d) Diagrama de dispersión co punto $(x_0, Y_0 - 28)$ coa recta de regresión en mediana.

Figura 3.4: Diagrama de dispersión do modelo (3.3) con 50 datos, engadindo un novo punto, que aparecerá en vermello, e as rectas real en negro, de regresión en media do modelo orixinal e de regresión en mediana do modelo orixinal en azul e a recta de regresión de media e regresión en mediana en vermello.

- Erro absoluto medio (EAM): $\sum_{i=1}^{n} |Y_i - \widehat{Y}_i|$

Neste estudo imos xerar datos do modelo

$$Y = 1 + 0'5X + (\epsilon - c_\tau) \tag{3.3}$$

onde a variable explicativa X segue unha distribución uniforme no intervalo $(0, 20)$, $X \sim$

$U(0, 20)$, o erro ϵ seguirá distintas distribucións e c_τ é o cuantil condicional de orde τ de ϵ dado un valor da variable X. As distribucións de erro que consideraremos son as seguintes:

1. O erro, ϵ, segue unha distribución normal estándar, con media 0 e varianza 1 ($\epsilon \sim N(0, 1)$).
2. O erro segue unha distribución T de Student con 1 grao de liberdade ($\epsilon \sim T_1$).
3. O erro segue unha distribución T de Student con 4 graos de liberdade ($\epsilon \sim T_4$).
4. O erro segue unha distribución de erro normal estándar con datos atípicos nas colas, o que se coñece como unha distribución *normal contaminada*. Neste caso, o 90 % dos datos seguen a distribución normal estándar e os 10 % dos datos restantes toman os valores $+6$ ou -6 con igual probabilidade. Denotaremos a esta distribución como D_1 ($\epsilon \sim D_1$).
5. O erro segue unha distribución na que o 75 % dos datos proveñen dunha $N(0, 5)$ e o 25 % dos datos proveñen dunha $N(1, 2)$. É dicir, o que se coñece como unha mistura de normais. A esta nova distribución denotarémola por D_2 ($\epsilon \sim D_2$).

Na Figura 3.5 representamos as distintas funcións de distribución dos erros citados anteriormente para poder ver como, a medida que a distribución do erro se alonxa da distribución normal, a regresión en mediana é máis recomendable que a regresión en media (que é o modelo óptimo cando o erro segue unha distribución normal).

Na Táboa 3.4 móstranse os distintos erros cadráticos medios (ECM) e os erros absolutos medios (EAM) para diferentes tamaños de mostra (n) e para as distintas distribucións de erro. Notemos que para cada modelo axustado computaremos só dez predicións. Na Táboa 3.4 podemos ver que no caso de que o erro siga unha distribución gaussiana entón os erros de predición da regresión en media son menores que os da regresión en mediana como cabería esperar dado que a regresión en media é óptima neste contexto. Notemos que canto maior sexa o tamaño da mostra menor serán os erros tanto para a regresión en media como para a regresión en mediana e ademais, as diferenzas entre ambas redúcense.

No caso de que $\epsilon \sim T_4$ os erros de predición da regresión en mediana son mellores, aínda así non hai unha gran diferenza cos erros de predición da regresión en media. Esto débese a que a distribución T_4 é semellante á $N(0, 1)$. En contraste, cando $\epsilon \sim T_1$ existe unha grandísima diferenza entre os erros de predición da regresión mediana e da regresión media, pois a distribución T_1 afástase notablemente de $N(0, 1)$, especialmente nas colas. Así vemos como ante un modelo no que o erro siga unha distribución normal estándar a

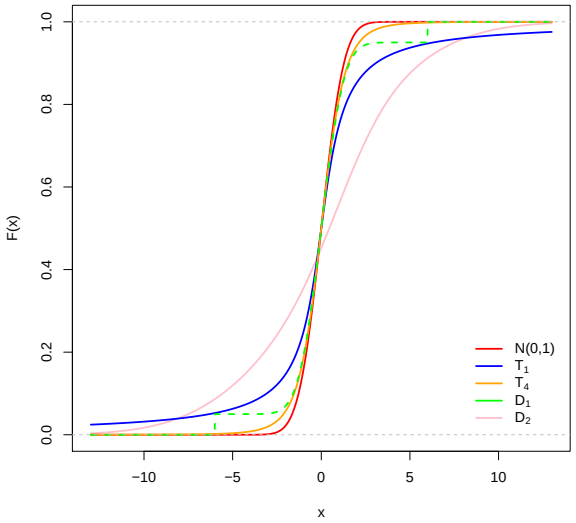


Figura 3.5: Representación gráfica das diferentes distribucións dos erros que consideramos neste estudo de simulación. En vermello móstrase a función de distribución de $N(0,1)$; en azul móstrase a función de distribución de T_1; en amarelo móstrase a función de distribución de T_4; en verde móstrase a función de distribución de D_1 e en rosa móstrase a función de distribución de D_2.

regresión en media é mellor que a regresión en mediana, mentres que noutras distribucións como as T de student con distintos graos de liberdade a regresión en mediana é unha mellor opción.

Se o erro segue a distribución D_1 os erros de predición son menores, aínda que a diferenza non é maiúscula, na regresión en mediana, polo que podemos concluír que no caso de que o erro siga unha distribución normal estándar con observacións atípicas nas colas é preferible a regresión en mediana.

Por outra banda, se o erro segue a distribución D_2 os erros cadráticos medios da regresión en media son máis pequenos que os da regresión en mediana mentres que os erros absolutos medios son maiores na regresión en media. Posto que a regresión en media se realiza minimizando o erro cadrático medio, polo tanto é normal que os ECM da regresión en media sexan menores que os da regresión en mediana. Polo tanto, guiándonos polos EAM, a elección da regresión en mediana sería máis acertada.

Na Sección A.2 móstranse os scripts dos que fixemos uso para realizar os estudos de

		Regresión en media		Regresión en mediana	
		ECM	EAM	ECM	EAM
$\epsilon \sim N(0,1)$	$n = 50$	1′029	0′812	1′055	0′822
	$n = 100$	1′046	0′815	1′054	0′818
	$n = 500$	0′999	0′798	1′001	0′799
$\epsilon \sim T_1$	$n = 50$	108356′48	18′978	85454′82	9′811
	$n = 100$	15066′061	13′020	4813′439	5′647
	$n = 500$	4254′273	9′379	3600′122	5′460
$\epsilon \sim T_4$	$n = 50$	1′909	1′008	1′898	1′005
	$n = 100$	2′044	1′020	2′036	1′018
	$n = 500$	1′929	0′988	1′930	0′990
$\epsilon \sim D_1$	$n = 50$	4′603	1′368	4′510	1′333
	$n = 100$	4′637	1′353	4′585	1′336
	$n = 500$	4′502	1′322	4′490	1′318
$\epsilon \sim D_2$	$n = 50$	6′691	1′996	7′251	1′838
	$n = 100$	6′541	1′983	7′121	1′813
	$n = 500$	6′453	1′964	7′122	1′801

Táboa 3.4: Táboa na que se mostran os distintos erros cadráticos medios e erros absolutos medios de predición asociados ás estimacións do modelo (3.3) obtidas grazas a modelos de regresión en media e en mediana, para distintos tamaños de mostra, n, e para distintas distribucións do erro, ϵ.

simulación aquí expostos.

Capítulo 4

Aplicación a datos reais

Ao longo deste capítulo presentaremos a utilidade da regresión cuantil grazas a unha aplicación a datos reais, e para isto faremos uso da base de datos *Engel* que foi empregada por primeira vez por Koenker e Basset en [8], que ademais está dispoñible no paquete *quantreg* de R.

Como vimos no Capítulo 2, a regresión cuantil consta dunhas boas propiedades onde cabe destacar a robustez fronte a datos atípicos e a súa adaptación a condicións máis xerais que a clásica regresión en media como pode ser a falta de normalidade do erro ou escenarios non homocedásticos. Tamén vimos de ver no Capítulo 3 que o modelo de regresión cuantil funciona dunha maneira apropiada nos diversos escenarios de simulación que se formularon. Agora veremos como tamén funciona adecuadamente en casos reais.

4.1. Presentación da base de datos reais

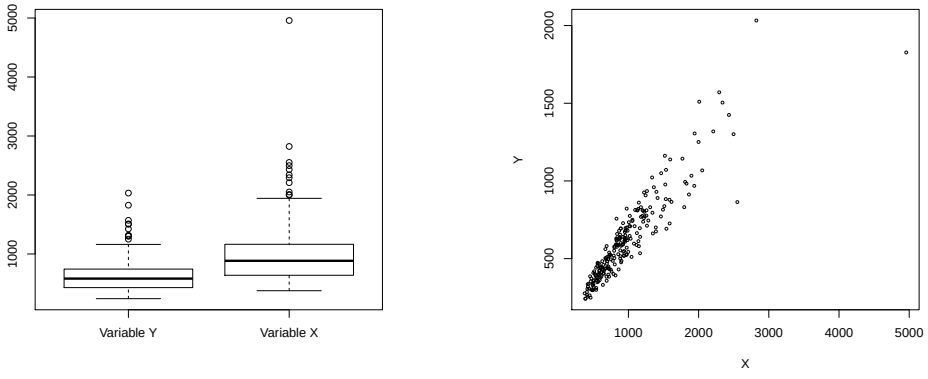
Con este obxectivo faremos unha aplicación aos datos reais que proporciona a base de datos *Engel*. Esta base de datos atópase no paquete '*quantreg*' do programa R e conta con 235 observacións dun estudo sobre o gasto en comida fronte aos ingresos da clase traballadora belga. Nesta base de datos dispoñemos de dúas variables:

- como variable resposta, que denotaremos por Y, consideraremos o gasto anual en comida de cada fogar belga, medidos en francos belgas.
- como variable explicativa, que denotaremos por X, consideramos os ingresos que recibe cada fogar belga, medidos en francos belgas.

Na Táboa 4.1 atopamos un resumo das principais medidas característica de ambas variables e amais represéntanse grazas a diagramas de caixa na parte (a) da Figura 4.1. Logo, na Figura 4.1 (b) representamos o diagrama de dispersión que nos permite ver a relación entre estas dúas variables. Ademais, se calculamos a covarianza entre ambas variables, danos un valor de $0'911$, polo que a correlación é positiva, é dicir, a medida que aumenta Y aumenta X, como observamos no diagrama de dispersión.

	Variable X	Variable Y
Mínimo	$377'1$	$242'3$
Cuantil $0'25$	$638'9$	$429'7$
Mediana	884	$582'5$
Media	$982'5$	$624'2$
Cuantil $0'75$	1164	$743'9$
Máximo	$4957'8$	$2032'7$
Desviación típica	$519'2$	$276'4$

Táboa 4.1: Principais medidas características das variables X e Y.



(a) Diagrama de caixas de X e Y.

(b) Diagrama de dispersión.

Figura 4.1: Boxplot para as variables Y e X (figura (a)) e diagrama de dispersión da base de datos *Engel* (figura (b)).

4.2. Axuste de modelos de regresión cuantil

Á vista da sección anterior, podemos considerar o seguinte modelo lineal de regresión cuantil que nos permita explicar o comportamento da variable resposta

$$Y = \beta_0^\tau + \beta_1^\tau X + \epsilon \tag{4.1}$$

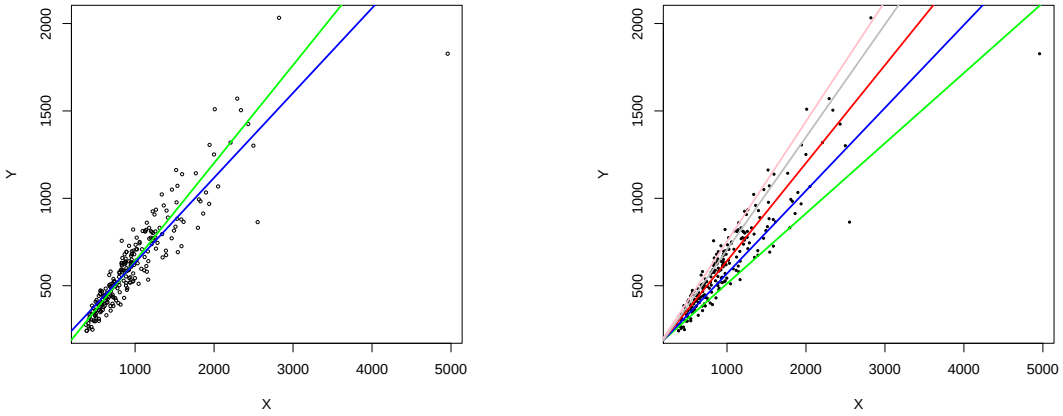
onde ϵ representa o erro do modelo, que debe verificar que o seu cuantil condicional de orde τ sexa igual a cero. Á vista da Figura 4.1, cabe mencionar que nos atopamos nun escenario heterocedástico dado que a variabilidade do erro parece que aumenta a medida que aumenta a variable X. Esta hipótese pódese comprobar grazas ao test de Breusch-Pagan (para máis información ver [4]) que podemos aplicar en R grazas á función *bptest* dispoñible no paquete *lmtest*. Para os nosos datos obteriamos un *p-valor*[1] menor a $2'2 \times 10^{-16}$, polo que podemos concluír que existen evidencias estatisticamente significativas de que os datos proveñen dun modelo heterocedástico.

Ademais intúese a presenza de observacións atípicas que provocarían un efecto panca na recta de regresión en media ("tirando da recta cara o datos atípicos"), mentres que a recta de regresión en mediana non debería verse moi afectada. Á vista da Figura 4.2 (a), podemos observar as diferenzas entre os modelos de regresión en media e mediana poñendo en evidencia a non simetría da distribución do erro. Así podemos comprobar a través dun exemplo real a robustez que presenta a regresión cuantil fronte a datos atípicos.

Presentamos a continuación na Táboa 4.2 as estimacións dos parámetros β_0^τ e β_1^τ do modelo de regresión cuantil (4.1) para distintos cuantís. Recordemos que como mencionamos no Capítulo 2, o método de estimación empregado para esta estimación é unha variante do algoritmo de Barrodale e Roberts (máis detalles en [1]). Se nos fixamos agora na Táboa 4.2, vemos como todos os coeficientes β_1^τ son positivos, o que quere dicir que, a medida que aumenta a variable X tamén aumenta a variable Y. Como cabería esperar, a medida que aumenta o ingresos nos fogares, tamén aumenta o gasto en comida. Nótese ademais que a medida que aumentamos o valor de τ, o valor da pendente tamén aumenta.

Finalmente, na Figura 4.2 (b) móstranse as rectas axustadas de regresión lineal cuantil para os distintos cuantís propostos. Isto proporciona unha idea máis ampla do efecto que a variable explicativa X, os ingresos, produce sobre a variable resposta Y, os gastos en

[1] Dado un estatístico de contraste, hai un valor de α (nivel de significación) a partir do cal xa non podemos rexeitar a hipótese nula. A dito valor chamarémolo *p-valor* do contraste. Entón se $\alpha > p-valor$ podemos rexeitar a hipótese nula a nivel α.



(a) Regresión en media fronte a regresión en mediana

(b) Regresión cuantil para diversos valores de τ

Figura 4.2: Diagrama de dispersión da base de datos *Engel* xunto con distintas rectas de regresión. Na parte (a) representamos en azul a recta de regresión en media e en verde a recta de regresión en mediana. Mentres que na parte (b) representamos en verde a recta de regresión cuantil para o $\tau = 0'1$, en azul o modelo para $\tau = 0{,}25$, en vermello o modelo para $\tau = 0'5$, en gris o modelo para $\tau = 0'75$ e en rosa o modelo para $\tau = 0'9$.

	$\tau = 0'10$	$\tau = 0'25$	$\tau = 0'50$	$\tau = 0'75$	$\tau = 0'90$
$\widehat{\beta_0}^{\tau}$	110'141	95'483	81'482	62'396	67'351
$\widehat{\beta_1}^{\tau}$	0'402	0'474	0'560	0'644	0'686

Táboa 4.2: Estimadores dos parámetros β_0^{τ} e β_1^{τ} asociados ao modelo (4.1) para diferentes valores do cuantil de interese.

comida, non só nos no centro da distribución (aportada pola mediana), senón tamén nos valores máis pequenos e máis grandes de Y. Ademais, vemos como as rectas asociadas aos modelos cuantís non son paralelas polo que pon de manifesto o carácter heterocedásticos dos datos que estamos tratando.

Capítulo 5

Conclusións

Este capítulo está adicado a expoñer as conclusións máis importantes derivadas da realización deste traballo. Como vimos ao longo do Capítulo 1, este traballo enmárcase no estudo de modelos de regresión cuantil, que é un tema de gran actualidade no ámbito da Estatística debido ás súas boas propiedades. Recordemos que dada unha variable resposta Y e un conxunto de variables explicativas $X = (X_1, ..., X_{p-1})$ definimos o modelo de regresión cuantil linear asociado a un certo cuantil, $\tau \in (0, 1)$, como segue:

$$Y = \beta_0^\tau + \beta_1^\tau X_1 + ... + \beta_n^\tau X_{p-1} + \epsilon$$

sendo $\beta^\tau = (\beta_0^\tau, \beta_1^\tau, ..., \beta_{p-1}^\tau) \in \mathbb{R}^p$ o vector de parámetros e ϵ o erro que verifica que o seu cuantil condicional de orde τ é cero, é dicir, $\mathbb{P}(\epsilon \leq 0 \mid X) = \tau$ (condición análoga a pedir no contexto da regresión en media que a esperanza condicional do erro sexa 0). Neste contexto, dada unha mostra aleatoria simple $\{(\mathbf{X_1}, Y_1), ..., (\mathbf{X_n}, Y_n)\}$ con $\mathbf{X_i} = (1, X_{i,1}, ..., X_{i,p-1})$, podemos estimar o parámetro β^τ como

$$\widehat{\beta^\tau} = \arg \min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau(Y_i - \mathbf{X_i}^T b),$$

onde $\rho_\tau(u) = u(\tau - \mathbb{I}(u < 0))$ coñécese como función de perda cuantílica. Débese destacar que neste punto xurde unha das principais diferencias entre a regresión cuantil e a regresión en media dado que a función de perda cuantílica non é derivable, e polo tanto, non podemos usar os mesmos métodos de optimización que no contexto dos mínimos cadrados.

Posto que os cuantís mostrais poden ser vistos como a solución dun problema de optimización (Sección 1.2), podemos expresar o estimador $\widehat{\beta^\tau}$ tamén como un problema de programación lineal:

$$\min_{(\beta,u,v) \in \mathbb{R}^p \times \mathbb{R}_+^{2n}} \left\{ \tau 1_n^T u + (1-\tau) 1_n^T v \mid \mathbb{X}\beta^\tau + u - v = y \right\}$$

onde $\mathbb{X}$ denota a matriz de deseño $\{u_i, v_i : i = 1, ..., n\}$ representan a parte positiva e negativa do vector de residuos da regresión, respectivamente.

O interesante é que a solución óptima deste problema, de existir, terá n variables non nulas (os p coeficientes do modelo e $n-p$ variables que se corresponden aos residuos), o que significa que dos n individuos hai p que teñen residuo cero. E en consecuencia, o modelo de regresión cuantil pasará por p datos da mostra. Ademais, como vimos no Corolario 2.1, temos que a proporción de residuos negativos é aproximadamente τ e a de residuos positivos é aproximadamente $(1-\tau)$ e o número de residuos iguais a cero é exactamente p, como mencionabamos anteriormente.

Por outra banda, o Teorema 2.2 mostra a converxencia asintótica a unha distribución gaussiana do estimador da regresión cuantil,

$$\widehat{\beta}^{\tau} = \arg \min_{b \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_{\tau}(Y_i - \mathbf{X_i}^T b),$$

sendo ρ_τ a función de perda cuantílica. De dito teorema dedúcese que canto maior sexa a densidade da variable resposta Y avaliada no cuantil de interese, menor será a varianza asintótica do estimador do parámetro de regresión cuantil, β^τ. Dito resultado ilústrase grazas a un estudo de simulación que se presenta na Sección 3.2, onde observamos nun escenario real como se comporta o erro cadrático medio do estimador $\widehat{\beta}^{\tau}$ para diversos valores do cuantil τ e diversas distribucións do erro.

Tamén é interesante mencionar que a regresión cuantil funciona en escenarios onde os erros teñen un carácter heterocedástico, mentres que na regresión en media se esixe a homocedasticidade como hipótese. Ademais, para presentar un modelo de regresión cuantil non precisamos que a distribución do erro siga unha distribución normal, como no caso da regresión por mínimos cadrados.

Outra das propiedades máis importantes que ten a regresión cuantil é a robustez que presenta fronte a observacións atípicas, tal e como podemos ver na Sección 2.3 a través da función de influencia. Ademais, na Sección 3.3 móstrase a través dun exemplo práctico como a regresión en mediana é capaz de resistir o efecto deste tipo de datos mentres que a regresión en media sofre unha gran modificación, que coñeceremos como efecto panca.

Ademais, tamén debemos falar dunha das 'debilidades' que presenta o modelo de regresión cuantil. Posto que as curvas de regresión cuantil se estiman de maneira individual, estas curvas poderían cruzarse nas colas da distribución. É dicir, o que se coñece como cruce entre cuantís, o que daría lugar a estimacións inconsistentes para certos cuantís de interese. En todo caso, débese destacar que este cruce entre cuantís está asociado á fronteira

do conxunto de datos.

Finalmente, ademais de mostrar as boas propiedades da regresión cuantil grazas aos estudos de simulación mencionados anteriormente (ver Capítulo 3), presentouse a utilidade da regresión cuantil a través dunha aplicación a datos reais que se desenrola ao longo do Capítulo 4. Nesta liña débese destacar que, mentres a regresión en media permite modelar o comportamento central da variable resposta en función das variables explicativas, a regresión cuantil permítemos estimar completamente a distribución condicional da variable resposta. É dicir, a regresión cuantil permite describir completamente o comportamento da variable resposta dado un conxunto de variables explicativas.

Apéndice A

Scripts

A.1. Cálculo dos erros cadráticos medios dos estimadores de β_0^τ e β_1^τ.

Na Sección 3.2 presentasenos o modelo 3.1 para o que calcularemos os erros cadráticos medios, ECM, dos estimadores β_0^τ e β_1^τ para corroborar o Teorema 2.2. A continuación mostramos un dos scripts usados para facer dito estudo de simulación para distintos tamaños de mostra ($n = 50$, $n = 100$), distintas distribucións de erro, ϵ ($N(0,1)$, $U[-1,1]$, χ^2_2 e $logN(0,1)$) e distintos cuantís, τ ($0'1$,$0'25$,$0'5$,$0'75$ e $0'9$).

```
set.seed(123456) #Fixamos a semilla
library(quantreg) #Cargamos a libraría de interese para a regresión cuantil

# Para calcular o ECM dos estimadores calculamos
# moitas veces os estimadores para distintas mostras (o número de veces será M).
# Logo calculamos sesgo e varianza deses M estimadores calculados.

M=1000 # Mostras de Monte Carlo
n=50 # Tamaño de mostra
beta=c(3,5) # Parámetros
estimadores=matrix(NA,nrow=M,ncol=2)
tau=0.9 #Cuantil

for(i in 1:M){
```

```
x.axuste=runif(n,min=0,max=1)
error.axuste=rlnorm(n, meanlog = 0, sdlog = 1)
-qlnorm(tau, meanlog = 0, sdlog = 1)
# O erro segue unha distribución normal
y.axuste=beta[1]+beta[2]*x.axuste+error.axuste
mod2=rq(y.axuste~x.axuste,tau=tau) # regresion cuantil 0.9
estimadores[i,]=coef(mod2)
}

##  beta_0
sesgo.beta0=mean(estimadores[,1])-beta[1];sesgo.beta0
var.beta0= var(estimadores[,1]);var.beta0
ECM.beta0=var.beta0+sesgo.beta0^2;ECM.beta0

##  beta_1
sesgo.beta1=mean(estimadores[,2])-beta[2];sesgo.beta1
var.beta1= var(estimadores[,2]);var.beta1
ECM.beta1=var.beta1+sesgo.beta1^2;ECM.beta1

# Resumimos todos os datos nunha táboa
Resultados=matrix(NA,ncol=3,nrow=2)
colnames(Resultados)=c("Sesgo","Varianza","ECM")
rownames(Resultados)=c("beta0","beta1")
Resultados[1,]=c(sesgo.beta0,var.beta0,ECM.beta0)
Resultados[2,]=c(sesgo.beta1,var.beta1,ECM.beta1)
Resultados

# Notación científica
formatC(Resultados, format = "e", digits = 3)
```

A.2. Regresión en media versus regresión en mediana

A.2.1. Gráficas da recta de regresión en media e da recta de regresión en mediana

A continuación móstrase o script co que fixemos o primeiro estudo de simulación da Sección 3.3, no que se pode ver a robustez da regresión en mediana.

```
set.seed(123456) #Fixamos a semilla
library(quantreg) #Cargamos o paquete

# Xeración dos datos do modelo:
n=20
x=1:n
x=runif(n,min=0,max=20)
y=1+0.5*x+rnorm(n)

#####-----> MODELO 1

# Representación dos datos orixinais para a media:
mod_media_1=lm(y~x);summary(mod_media_1)
plot(y~x,xlab="",ylab="",asp=1)
abline(mod_media_1,col="blue",lwd=2)
abline(1,0.5,lwd=2)
legend(0,15,
c("Recta real","Recta de regresión en media 1"),lwd=2,col=c("1","4"),box.lty=0)

# Representación dos datos orixinais para a mediana:
mod_mediana_1=rq(y~x);summary(mod_mediana_1)
plot(y~x,xlab="",ylab="",asp=1)
abline(mod_mediana_1,col="blue",lwd=2)
abline(1,0.5,lwd=2)
legend(0,15,
c("Recta real","Recta de regresión en mediana 1"),lwd=2,col=c("1","4"),box.lty=0)

#####-----> MODELO 2
```

```
# Representación dos datos orixinais para a media cun dato máis
# CON efecto panca:
xmarco=c(1,30)
ymarco=c(0,16)
x0=30
y0p=predict(mod_media_1,data.frame(x=c(x0)))
names(y0p)=c()
y0=y0p-8;y0
x_ampliado=c(x,x0)
y_ampliado=c(y,y0)

mod_media_2=lm(y_ampliado~x_ampliado);summary(mod_media_2)

residuals(mod_media_2) # Residuos brutos
rstandard(mod_media_2) # Residuos estandarizados

plot(xmarco,ymarco,type="n",xlab="",ylab="",asp=1)
points(x,y)
abline(1,0.5,lwd=2)
abline(mod_media_1,col="blue",lwd=2)
points(x0,y0,col="red",pch=19)
abline(mod_media_2,col="red",lwd=2)
legend(0,21,c("Recta real","Recta de regresión en media 1",
"Recta de regresión en media 2"),lwd=2,col=c("1","4","2"),box.lty=0)

# Representación dos datos orixinais para a mediana cun dato máis
# SEN efecto panca
xmarco=c(1,30)
ymarco=c(0,16)
x0=30
y0p=predict(mod_mediana_1,data.frame(x=c(x0)))
names(y0p)=c()
y0=y0p-8
x_ampliado=c(x,x0)
y_ampliado=c(y,y0)
```

```
mod_mediana_2=rq(y_ampliado~x_ampliado);summary(mod_mediana_2)

plot(xmarco,ymarco,type="n",xlab="",ylab="",asp=1)
points(x,y)
abline(1,0.5,lwd=2)
abline(mod_mediana_1,col="blue",lwd=2)
points(x0,y0,col="red",pch=19)
abline(mod_mediana_2,col="red",lwd=2)
legend(0,21,c("Recta real","Recta de regresión en mediana 1",
"Recta de regresión en mediana 2"),lwd=2,col=c("1","4","2"),box.lty=0)

#####-----> MODELO 3

# Representación dos datos orixinais para a media cun dato máis
# CON efecto panca:
xmarco=c(1,30)
ymarco=c(-15,25)
x0=30
y0p=predict(mod_media_1,data.frame(x=c(x0)))
names(y0p)=c()
y0=y0p-28
x_ampliado=c(x,x0)
y_ampliado=c(y,y0)

mod_media_3=lm(y_ampliado~x_ampliado);summary(mod_media_3)
residuals(mod_media_3) # Residuos brutos
rstandard(mod_media_3) # Residuos estandarizados

plot(xmarco,ymarco,type="n",xlab="",ylab="",asp=1)
points(x,y)
abline(1,0.5,lwd=2)
abline(mod_media_1,col="blue",lwd=2)
points(x0,y0,col="red",pch=19)
abline(mod_media_3,col="red",lwd=2)
legend(-8,25,c("Recta real","Recta de regresión en media 1",
"Recta de regresión en media 3"),lwd=2,col=c("1","4","2"),box.lty=0)
```

```
# Representación dos datos orixinais para a mediana cun dato máis
# SEN efecto panca
xmarco=c(1,30)
ymarco=c(-15,25)
x0=30
y0p=predict(mod_mediana_1,data.frame(x=c(x0)))
names(y0p)=c()
y0=y0p-28
x_ampliado=c(x,x0)
y_ampliado=c(y,y0)

mod_mediana_3=rq(y_ampliado~x_ampliado);summary(mod_mediana_3)

plot(xmarco,ymarco,type="n",xlab="",ylab="",asp=1)
points(x,y)
abline(1,0.5,lwd=2)
abline(mod_mediana_1,col="blue",lwd=2)
points(x0,y0,col="red",pch=19)
abline(mod_mediana_3,col="red",lwd=2)
legend(-8,25,c("Recta real","Recta de regresión en mediana 1",
"Recta de regresión en mediana 3"),lwd=2,col=c("1","4","2"),box.lty=0)
```

A.2.2. Cálculo do erro cadrático medio e do erro absoluto medio

Para o segundo estudo de simulación que realizamos na Sección 3.3, no que se presentan os ECM e os EAM para distintas distribucións de erro e para distintos tamaños de mostra utilizamos os seguintes scripts:

```
set.seed(123456) # Fixamos unha semilla
library(quantreg) # Libraría necesaria para axustar
# modelos de regresión cuantil

n=50 # Tamaño de mostra
beta=c(1,0.5) # Parámetros
M=1000 # Réplicas Monte Carlo
```

```
EAM.media=numeric(M);ECM.media=numeric(M)
EAM.mediana=numeric(M);ECM.mediana=numeric(M)

for(i in 1:M){
x.axuste=runif(n,min=0,max=20); x.predicion=runif(10,min=0,max=20)
error.axuste=rnorm(n,mean=0,sd=1)-qnorm(0.5,mean=0,sd=1)
error.predicion=rnorm(10,mean=0,sd=1)-qnorm(0.5,mean=0,sd=1)
y.axuste=beta[1]+beta[2]*x.axuste+error.axuste
y.predicion=beta[1]+beta[2]*x.predicion+error.predicion
#plot(x.axuste,y.axuste)

# Axustamos os modelos de regresión en media e mediana
modelo.media=lm(y.axuste~x.axuste)
modelo.mediana=rq(y.axuste~x.axuste,tau=0.5)

# Facemos prediccións empregando os modelos axustados
novos.datos=data.frame("x.axuste"=x.predicion)
predicion.media=predict(modelo.media,newdata=novos.datos)
predicion.mediana=predict(modelo.mediana,newdata=novos.datos)

# Calculamos os erros de prediccion
EAM.media[i]=mean(abs(predicion.media-y.predicion))
ECM.media[i]=mean((predicion.media-y.predicion)^2)

EAM.mediana[i]=mean(abs(predicion.mediana-y.predicion))
ECM.mediana[i]=mean((predicion.mediana-y.predicion)^2)
}

Resultados=matrix(NA,ncol=2,nrow=2)
colnames(Resultados)=c("ECM","EAM")
rownames(Resultados)=c("Reg.media","Reg.mediana")
Resultados[1,1]=mean(ECM.media)
Resultados[1,2]=mean(EAM.media)
Resultados[2,1]=mean(ECM.mediana)
Resultados[2,2]=mean(EAM.mediana)
round(Resultados,3)
```

Para a distribución, D_1, de erro normal estándar con datos atípicos nas colas, cun 90 % dos datos con distribución normal estándar e os 10 % dos datos restantes teñen os valores $+6$ ou -6 utilizamos o seguinte script:

```
set.seed(123456) # Fixamos unha semilla
library(quantreg) # Libraría necesaria para axustar
# modelos de regresion cuantil

n=100 # Tamaño de mostra
beta=c(1,0.5) # Parámetros
M=1000 # Réplicas Monte Carlo

EAM.media=numeric(M);ECM.media=numeric(M)
EAM.mediana=numeric(M);ECM.mediana=numeric(M)

x.axuste=runif(n,min=0,max=20);x.predicion=runif(10,min=0,max=20)
atipicos=function(n,p,media,desv,ati){
X=numeric(n)
for(j in 1:n){
u=runif(2)
if(u[1]<p){X[j]=rnorm(1,mean=media,sd=desv)}else{
if(u[2]<=0.5){X[j]=ati}else{X[j]=-ati}}}
return(X)}
datos.mostra.grande=atipicos(100000,0.9,0,1,6)
ctau.atipicos=quantile(datos.mostra.grande,0.5)
# Aproximamos a mediana

for(i in 1:M){
error.axuste=atipicos(n,0.9,0,1,6)-ctau.atipicos
error.predicion=atipicos(n,0.9,0,1,6)-ctau.atipicos
y.axuste=beta[1]+beta[2]*x.axuste+error.axuste
y.predicion=beta[1]+beta[2]*x.predicion+error.predicion
#plot(x.axuste,y.axuste)

# Axustamos os modelos de regresión en media e mediana
modelo.media=lm(y.axuste~x.axuste)
modelo.mediana=rq(y.axuste~x.axuste,tau=0.5)
```

```
# Facemos prediccións empregando os modelos axustados
novos.datos=data.frame("x.axuste"=x.predicion)
predicion.media=predict(modelo.media,newdata=novos.datos)
predicion.mediana=predict(modelo.mediana,newdata=novos.datos)

# Calculamos os erros de predicción
EAM.media[i]=mean(abs(predicion.media-y.predicion))
ECM.media[i]=mean((predicion.media-y.predicion)^2)

EAM.mediana[i]=mean(abs(predicion.mediana-y.predicion))
ECM.mediana[i]=mean((predicion.mediana-y.predicion)^2)
}

Resultados=matrix(NA,ncol=2,nrow=2)
colnames(Resultados)=c("ECM","EAM")
rownames(Resultados)=c("Reg.media","Reg.mediana")
Resultados[1,1]=mean(ECM.media)
Resultados[1,2]=mean(EAM.media)
Resultados[2,1]=mean(ECM.mediana)
Resultados[2,2]=mean(EAM.mediana)
round(Resultados,3)
```

E para finalizar mostramos o script co que nos apoiamos para facer os ECM e EAM da distribución, D_2, de erro na que o 75 % dos datos proveñen dunha $N(0,5)$ e o 25 % dos datos proveñen dunha $N(1,2)$.

```
set.seed(123456) # Fixamos unha semilla
library(quantreg) # Libraría necesaria para axustar
# modelos de regresión cuantil

n=100 # Tamaño de mostra
beta=c(1,0.5) # Parámetros
M=1000 # Réplicas Monte Carlo

EAM.media=numeric(M);ECM.media=numeric(M)
EAM.mediana=numeric(M);ECM.mediana=numeric(M)
```

```
x.axuste=runif(n,min=0,max=20);x.predicion=runif(10,min=0,max=20)
mixtura = function(n,p,mu1,mu2,sigma1,sigma2){
X=rnorm(n,mu1,sigma1)
U=runif(n)
Ind=U>=p
nI=sum(Ind)
X[Ind]=rnorm(nI,mu2,sigma2)
return(X)}
datos.mostra.grande=mixtura(100000,0.75,0,5,1,2)
ctau.mixtura=quantile(datos.mostra.grande,0.5)
# Aproximamos a mediana

for(i in 1:M){
error.axuste=mixtura(n,0.75,0,5,1,2)-ctau.mixtura
error.predicion=mixtura(n,0.75,0,5,1,2)-ctau.mixtura
y.axuste=beta[1]+beta[2]*x.axuste+error.axuste
y.predicion=beta[1]+beta[2]*x.predicion+error.predicion
#plot(x.axuste,y.axuste)

# Axustamos os modelos de regresión en media e mediana
modelo.media=lm(y.axuste~x.axuste)
modelo.mediana=rq(y.axuste~x.axuste,tau=0.5)

# Facemos predicciónes empregando os modelos axustados
novos.datos=data.frame("x.axuste"=x.predicion)
predicion.media=predict(modelo.media,newdata=novos.datos)
predicion.mediana=predict(modelo.mediana,newdata=novos.datos)

# Calculamos os erros de predicción
EAM.media[i]=mean(abs(predicion.media-y.predicion))
ECM.media[i]=mean((predicion.media-y.predicion)^2)

EAM.mediana[i]=mean(abs(predicion.mediana-y.predicion))
ECM.mediana[i]=mean((predicion.mediana-y.predicion)^2)
}
```

```
Resultados=matrix(NA,ncol=2,nrow=2)
colnames(Resultados)=c("ECM","EAM")
rownames(Resultados)=c("Reg.media","Reg.mediana")
Resultados[1,1]=mean(ECM.media)
Resultados[1,2]=mean(EAM.media)
Resultados[2,1]=mean(ECM.mediana)
Resultados[2,2]=mean(EAM.mediana)
round(Resultados,3)
```

A.3. Aplicación a datos reais

Finalmente, para o Capítulo 4 utilizamos o seguinte script:

```
set.seed(123456)
library(quantreg)
data(engel) # cargamos a base de datos Engel
attach(engel)

##--> Boxplot para as variables Y e X
boxplot(cbind(foodexp,income),names=c("Variable Y","Variable X"))

##--> Principais medidas características das variables
summary(foodexp) # Variable Y
sd(foodexp)
summary(income) # Variable X
sd(income)
cor(income,foodexp) #Covarianza

##--> Test de heterocedasticidade
library(lmtest)
bptest(foodexp~income)

##--> Representamos o diagrama de dispersión dos datos xunto
# coas rectas de regresión en media e en mediana
plot(engel, cex=0.5, xlab="X",ylab="Y")
```

```
abline(rq(foodexp~income,tau=0.5),col="green",lwd=2)
abline(lm(foodexp~income), col="blue",lwd=2)

##--> Xeramos os modelos lineais de regresión cuantil
# para distintos cuantís
mod1=rq(foodexp~income,tau=0.1); summary(mod1) #cuantil 0.1
mod2=rq(foodexp~income,tau=0.25); summary(mod2) #cuantil 0.25
mod3=rq(foodexp~income,tau=0.5); summary(mod3) #cuantil 0.5
mod4=rq(foodexp~income,tau=0.75); summary(mod4) #cuantil 0.75
mod5=rq(foodexp~income,tau=0.9); summary(mod5) #cuantil 0.9

##--> Representación gráfica das distintas rectas de regresión cuantil
# para distintos cuantís
plot(engel, cex=0.5, xlab="X",ylab="Y",pch=20)
abline(mod1,col="green",lwd=2)
abline(mod2,col="blue",lwd=2)
abline(mod3,col="red",lwd=2)
abline(mod4,col="gray",lwd=2)
abline(mod5,col="pink",lwd=2)
```

Bibliografía

[1] Barrodale, I. e Roberts, F. D. K. *An improved algorithm for discrete l_1 linear approximation.*, SIAM J. Numer. Anal. 10 (1973), 839-848.

[2] Bhatti, M. A. *Practical optimization methods*, Primeira Edición, Springer-Verlag, 2000.

[3] Billingsley, P., *Probability and Measure*, Terceira Edición, Wiley-Interscience, 1995.

[4] Breusch, T. S. e Pagan, A. R., *A Simple Test for Heteroscedasticity and Random Coefficient Variation.*, Econometrica 47 (1979), 1287–1294.

[5] Cook, R. D. e Weisberg, S. *Residuals and influence in regression.*, Primeira Edición, Chapman and Hall, 1982.

[6] García Pérez, A., *Estadística básica con R.*, UNED, 2010.

[7] Koenker, R., *Quantile Regression*, Primeira Edición, Cambridge University Press, 2005.

[8] Koenker, R. e Bassett, G. *Regression quantiles.*, Econometrica 46 (1978), 33-50.

[9] Peña, D., *Regresión y diseño de experimentos.*, Primeira Edición, Alianza Editorial, 2002.

[10] Peña, D., *Fundamentos de estadística*, Primeira Edición, Alianza Editorial, 2005.

[11] Verzani, J., *Using R for introductory Statistics.*, Primeira Edición, Taylor & Francis, 2005.