UNIVERSIDADE DA CORUÑA

Universida$_{\text{de}}$Vigo

Traballo Fin de Máster

Regresión cuantil con datos censurados por intervalos

Díaz Candán, Santiago

Máster en Técnicas Estatísticas

Curso 2019-2020

Índice general

Resumo

Resumo en español

La regresión cuantil estudia el cuantil de la variable respuesta condicionado a ciertas variables explicativas. Incluye como caso particular la regresión en mediana, que es una alternativa muy interesante a la regresión en media. La regresión cuantil goza de propiedades de robustez y flexibilidad frente a la distribución del error de regresión, que puede adoptar formas muy diferentes a la normal, y permite representar diversas posiciones de la variable respuesta, no sólo la posición central. Además, es especialmente útil para tratar situaciones con información incompleta, como los datos censurados por intervalos, en los cuales para cada individuo sólo se conoce que la variable respuesta se encuentra en cierto intervalo.
En este trabajo se hará una breve revisión de la regresión cuantil y del problema de estimación con datos censurados por intervalos, para más adelante proponer nuevos estimadores de la regresión cuantil con datos censurados por intervalos, y analizar sus propiedades.

Resumo en galego

A regresión cuantil estuda o cuantil da variable resposta condicionado a certas variables explicativas. Inclúe como caso particular a regresión na mediana, que é unha alternativa moi interesante á regresión en media. A regresión cuantil goza de propiedades de robustez e flexibilidade fronte á distribución do erro de regresión, que pode adoptar formas moi diferentes á normal, e permite representar diversas posicións da variable resposta, non só a posición central. Ademais, é especialmente útil para tratar situacións con información incompleta, como os datos censurados por intervalos, nos cales para cada individuo só se coñece que a variable resposta se atopa en certo intervalo.
Neste traballo farase unha breve revisión da regresión cuantil e do problema de estimación con datos censurados por intervalos, para máis adiante propoñer un novo estimador da regresión cuantil con datos censurados por intervalos, e analizar as súas propiedades.

English abstract

Quantum regression studies the quantile of the conditioned response variable a certain explanatory variables. Includes as a particular case the regression in median, which is a very interesting alternative to mean regression. Quantile regression enjoys properties of robustness and flexibility over the regression error distribution, which can take very different forms to normal, and allows to represent diverse positions of the variable answer, not just the central position. In addition, it is especially useful for treating situations with incomplete information, such as interval-censored data, in which for each individual the variable answer is only known is in a certain range.
In this work, a brief review of quantile regression and estimation problem with interval-censored data will be done, to later propose new quantum regression estimator with interval-censored data, and analyze their properties.

Prefacio

O estudo da censura estatística proporciona unha serie de ferramentas útiles á hora de abordar problemas relacionados con diferentes campos: dende estudos epidemiolóxicos relacionados co VIH ata os tempos de faio de produtos dentro do campo industrial. Co fin de establecer en que contextos existe censura resulta lóxico fixar un marco teórico, pero non parece doado abordar esta cuestión sen explicar previamente en que consiste a censura. A tal efecto, responderemos a continuación unha serie de preguntas fundamentais para entender os aspectos básicos da censura estatística.

- **Que é censura estatística?**

  O fenómeno da censura en estatística consiste na ausencia de certa información dos datos que son obxecto de estudo. En lugar de coñecer o valor exacto do dato, proporciónase un rango de valores ou un conxunto que o delimita.

- **Como se relaciona cos datos incompletos?**

  Os datos censurados son un tipo de datos incompletos, pero existen outros tipos de datos incompletos que non son censurados. Un exemplo de datos incompletos que poden non ser censurados son os datos truncados. A diferenza dos datos censurados, neste caso non existe constancia da existencia do dato a menos que pertenza a un conxunto de datos determinado.

- **Que tipos de censura existen?**

  Atendendo á forma de obtención dos datos, a censura divídese en informativa, se a forma de obtención dos datos inflúe na verosimilitude, ou censura non informativa. Outra clasificación atende á forma da información que proporcione a censura. Por exemplo, nun caso de censura en datos temporais, o estado actual dos datos proporciona como única información se un suceso determinado aconteceu ou non nun instante de tempo, fronte o caso xeral, onde pode vir a información restrinxida entre dous instantes temporais.

- **Cales son os métodos con censura?**

  Igual que con datos non censurados, un dos principais obxectivos dos datos censurados consiste en analizar como se comportan os datos. Para iso constrúense modelos matemáticos que proporcionan información útil de como se distribúen os datos, como se relacionan con outras variables, etc.

- **Que funcións están presentes nun estudo con datos censurados?**

  Igual que en estatística non censurada, a función de distribución e a función de densidade son básicas para analizar o comportamento dos datos. Ademais, dada a situación excepcional dos datos censurados e a súa relación cas estimacións temporais, estará presente a función de supervivencia, que se define como a complementaria da función de distribución e que mide a probabilidade de que un individuo continúe vivo nun determinado momento.

- **Existe software dispoñible para censura?**

Se ben é certo que existen numerosos algoritmos relativos á censura que aínda non teñen software implementado, existe en R o paquete Icens que pode resultar moi útil á hora de abordar a censura por intervalos. Inclúe rutinas relativas ao algoritmo de estimación e maximización (Algoritmo EM), e tamén permite estimar a función de supervivencia.

- **Que é a regresión cuantil?**

  A regresión consiste nun conxunto de modelos estatísticos cuxo fin é determinar o efecto que unha ou máis variables, coñecidas como variables explicativas, producen sobre outra variable, coñecida como variable resposta. Ditos modelos axustaranse mediante diferentes criterios, sendo dito criterio no caso da regresión cuantil a función de perda cuantílica. Deste xeito a regresión cuantil coincide con certo cuantil da variable resposta condicionado ao valor das variables explicativas.

- **Cantas seccións incorporará o documento e que incluirán?**

  O traballo dividirase en catro seccións básicas. A primeira inclúe todo o necesario sobre a estimación da función de distribución e de densidade con datos censurados por intervalos, abordando os desenvolvementos teóricos presentados por Dempster [5] e Turnbull [23] e exemplos prácticos.

  A sección segunda está centrada na regresión cuantil, en que consiste e cales son as súas vantaxes respecto da regresión baseada en mínimos cadrados.

  A sección terceira desenvolve unha proposta de método de estimación da regresión cuantil con datos censurados por intervalos, incluíndo o código en R do procedemento e algún exemplo.

  A cuarta e última sección incorpora un procedemento de simulación de mostras censuradas por intervalos a partir das cales é posible obter aproximacións do sesgo, varianza e erros cadráticos medios dos estimadores dos parámetros de regresión. Tamén se aborda unha breve comparación do modelo de regresión cuantil con datos censurados por intervalos e o modelo paramétrico exponencial para datos censurados.

Capítulo 1

Datos censurados por intervalos

1.1. Definición do problema

Procederemos a continuación a definir formalmente a censura estatistica. Para isto, seguiremos a notación introducida por Turnbull [23]. Considérase X unha variable aleatoria con valores en $\mathbb{R}$, da que consideramos n observacións independentes $X_1=x_1,....,X_n=x_n$. Ditos datos son censurados cando non se coñecen ditos valores a priori, pero si se coñecen n conxuntos en $\mathbb{R}$, A_1, A_2,...., A_n, tales que $X_1 \epsilon A_1$, $X_2 \epsilon A_2$,...., $X_n \epsilon A_n$. No caso en que os conxuntos A_i sexan disxuntos falaremos de datos agrupados.
O anterior non determina a forma en que a censura se produce. Imaxinemos que a censura se produce a partires dos resultados de persoas que asisten ao médico. Naturalmente, non garda a mesma relación ca enfermidade programar unha cita en función dunha doenza puntual dun paciente que unha programación periódica. No caso de que garde relación o mecanismo de censura co acontecemento dun evento, denominarase censura informativa; pola contra, se o mecanismo de censura é independente do evento, entón estaremos ante un caso de censura non informativa. No resto do documento, a menos que se especifique o contrario, supoñeremos este último tipo.
Volvendo ao noso obxecto de estudo, consideramos censura por intervalos cando os conxuntos A_i antes descritos teñen forma de intervalo de datos reais.
Imos clasificar en tres categorías os datos censurados por intervalos. Existe un exemplo descrito en Sun [21], que analiza os resultados obtidos en dous grupos de control sobre a retracción mamaria. Someteuse aos dous grupos a tratamentos distintos: ao primeiro a radioterapia e ao segundo a radioterapia e a quimioterapia. Independentemente do tratamento as pacientes experimentaron retracción mamaria, e o interrogante consistía en descubrir se o tratamento influíu na retracción. Unha vez rematou o experimento, presentáronse tres situacións:

- Cando se someteu a control por primeira vez a certas pacientes, estas xa experimentaran retracción mamaria. Estamos ante un caso de censura pola esquerda, posto que non hai ningún momento de referencia previo.
- Outras pacientes experimentaron a retracción mamaria entre varias sesións de control, dando o que se denomina de maneira sinxela como censura por intervalos.
- Por último, determinadas pacientes unha vez rematado o estudo non experimentaron retracción mamaria, o que significa que dito proceso pode producirse fora da liña de tempo marcada no estudo. Esta última coñécese como censura pola dereita. Retomaremos este exemplo en seccións posteriores.

Debemos definir con linguaxe alxébrica axeitada os intervalos onde se asenta o anterior. Para cada dato i censurado en intervalos, con i=1,...,n, denotaremos por L_i o extremo esquerdo do intervalo censurado

pola dereita, e por R_i o extremo dereito do intervalo. Por conseguinte, para a observación X_i obtemos un intervalo $[L_i, R_i]$. Situando un valor de L_i=0, estamos ante un caso de censura pola esquerda, e con $R_i = \infty$ ante censura pola dereita.

Como é obvio, calquera análise rigoroso dos datos censurados pasa por estimar a correspondente función de distribución. Para iso, existen diferentes procedementos e técnicas de gran utilidade. Non obstante, dado que algunhas desas técnicas son perfectamente extensibles a casos de truncamento e de datos agrupados, abordaremos unha xeneralización do anterior a casos de datos incompletos. Usaremos unha notación similar á empregada por Dempster [5].

1.2. Estimación con datos incompletos: o algoritmo EM

A incompletitude parte da base dunha relación entre dúas variables, que chamaremos X e Y. Dunha delas, a variable X, descoñecemos os seus valores exactos pero coñecemos a súa relación ca variable Y, que si é observable. A variable X coñecerase como variable completa e a variable Y como variable incompleta. No caso da censura por intervalos, a variable X será o tempo exacto, mentres que o intervalo $[L, R]$ é a variable Y.

Dun modo xenérico, a situación de datos incompletos pódese pensar como unha aplicación de X en Y, na cal parte da información de X queda reducida ao valor observable de Y. Nun contexto paramétrico, no cal a distribución de X dependa dun parámetro θ, a distribución de Y tamén dependerá de θ, pero pode non ser tan sinxelo estimar θ a partir dunha mostra da variable incompleta Y como se tiveramos unha mostra de X. Escrito formalmente, se f(x| θ) denota a función de densidade mostral da variable aleatoria X e g(y| θ) a función de densidade mostral da variable aleatoria Y, tense a relación seguinte:

$$g\left(y|\theta\right) = \int_{X(y)} f\left(x|\theta\right) dx.$$

sendo X(y) o conxunto de valores da variable X que dan lugar ao valor y da variable Y.

Para simplificar, supoñamos que a distribución de X segue a forma da familia exponencial, familia moi empregada en numerosos contextos matemáticos e que resolve unha gran cantidade de problemas. Máis adiante, xeneralizarase para calquera función de distribución. Entón a función de densidade mostral para a variable X condicionada ao parámetro θ ten a forma:

$$f\left(x|\theta\right) = \frac{b\left(x\right) e^{\theta t(x)^T}}{a\left(\theta\right)}$$

sendo $t(x)^T$ a trasposta do estatístico suficiente da variable aleatoria X.

O método iterativo proposto por Dempster [5] en 1977, coñecido como algoritmo EM, consta de dúas fases, E e M. No contexto da familia exponencial, na fase E do algoritmo calcúlase a esperanza do estatístico suficiente $t(x)$ condicionada ao valor da variable observada y e a un valor do parámetro θ. Como algoritmo iterativo que é, pártese dun valor inicial de θ. Na fase M do algoritmo calcúlase un novo valor de θ empregando o valor aproximado na fase E do estatístico suficiente $t(x)$. De forma abreviada, as fases son as seguintes, onde se parte dun valor $\theta^{(p)}$ e obténse un novo valor $\theta^{(p+1)}$ do parámetro θ:

- Fase E: Calcular $t^{(p)} = \mathbb{E}(t(x) \mid y, \theta^{(p)})$
- Fase M: Calcular $\theta^{(p+1)}$ como a solución de $IE(t(x)|\ \theta\ ) = t^{(p)}$

Omitindo a suposición de que a distribución segue a familia exponencial, o algoritmo EM exprésase directamente a través da verosimilitude. Así, na fase calcúlase a esperanza da log-verosimilitude cos datos dispoñibles, e na fase M maximízase a función calculada no paso E. En concreto, definimos

$$Q\left(\theta^{'} \mid \theta\right) = E\left(logf\left(x|\theta'\right)|y,\theta\right)$$

Nótese que estamos traballando con dous valores do parámetro, θ e θ'. No algoritmo sustituiremos θ polo valor $\theta^{(p)}$ resultante da iteración anterior, e θ' polo valor $\theta^{(p+1)}$ que queremos obter despois da iteración $(p+1)$. Deste xeito, o proceso EM resultante constará de novo de dúas fases para unha iteración $(p+1)$ determinada:

Fase E: Calculamos a esperanza Q($\theta \mid \theta^{(p)}$)

Fase M: Obténse $\theta^{(p+1)}$ como o valor que maximiza Q($\theta \mid \theta^{(p)}$) en θ.

Tal como ilustra Dempster [5], o algoritmo EM é monótono, no sentido de que a log-verosimilitude medra en cada iteración, e converxente.

Daremos paso a continuación a un exemplo que facilitará a comprensión do algoritmo EM.

1.3. Exemplo práctico do algoritmo EM

Imaxinemos un experimento aleatorio definindo unha variable aleatoria X, que consistirá no número de caras no lanzamento de dúas veces unha moeda ao aire. O comportamento segue unha distribución binomial de parámetros n=2, e p descoñecido. Respecto da sección anterior, consideraremos a X como os datos completos e a θ =p.

Temos tres posibles valores de X en función do número de caras obtidas: 0, 1 e 2. Podemos calcular a probabilidade de cada caso:

$$P\left(X=0\right) = \binom{2}{0} p^0 \left(1-p\right)^2 = \left(1-p\right)^2$$

$$P\left(X=1\right) = \binom{2}{1} p^1 \left(1-p\right)^1 = 2p\left(1-p\right)^2$$

$$P\left(X=2\right) = \binom{2}{2} p^2 \left(1-p\right)^0 = p^2$$

Consideremos unha variable dependendo da variable aleatoria X. Definimos Y unha variable cuxos valores serán 0 no caso de que non saia ningunha cara, e 1 no caso de que haxa polo menos unha cara. Polo tanto, se X respondía o número de caras, Y responde se hai caras. Igual que antes, a probabilidade de cada caso será

$$P\left(Y=0\right) = \left(1-p\right)^2$$

$$P\left(Y=1\right) = 2p\left(1-p\right) + p^2 = 2p - p^2$$

Consideremos a repetición do experimento un número de N=100 veces, obtendo as seguintes frecuencias absolutas:

X	0	1	2
FRECUENCIA ABSOLUTA	20	55	25

de onde podemos deducir os correspondentes valores de Y.

Y	0	1
FRECUENCIA ABSOLUTA	20	80

Para calcular neste suposto a esperanza da variable X, cun tamaño N=100, sabemos que IE(X)=Np. Con esta información, sería doado establecer un valor aproximado de p:

$\overline{X} = \frac{20 \cdot 0 + 55 \cdot 1 + 25 \cdot 2}{100} = 1,05$

$\overline{X} = n\hat{p} \rightarrow \hat{p} = \frac{\overline{X}}{n} = \frac{1,05}{2} = 0,525$

Imaxinemos que no experimento de N=100 non temos o reparto de 0, 1 e 2 caras,pero si o valor de Y, e polo tanto o reparto de 20 casos sen caras e 80 casos de caras. Se tivésemos o número de caras exacto, poderíamos obter p, e viceversa. Como sabemos que X segue unha distribución binomial, se queremos aproximar p podemos iniciar dándolle un valor p=0,5 e, mediante o algoritmo EM, aproximariámonos ao resultado.

$P(X = 1) = \frac{2p(1-p)}{2p(1-p)+p^2} = \frac{2 \cdot 0,5 \cdot 0,5}{2 \cdot 0,5 \cdot (0,5)+0,25} = \frac{2}{3}$

$P(X = 2) = \frac{p^2}{2p(1-p)+p^2} = \frac{0,25}{2 \cdot 0,25 \cdot (0,5)+0,25} = \frac{1}{3}$

Do valor aproximado da probabilidade anterior, podemos calcular o reparto como segue:

X	0	1	2
FRECUENCIA ABSOLUTA	20	$\frac{2}{3} \cdot 80$	$\frac{1}{3} \cdot 80$

Co reparto estimado, podemos calcular unha estimación de p:

$\overline{X} = \frac{20 \cdot 0 + \frac{2}{3} \cdot 80 \cdot 1 + \frac{1}{3} \cdot 80}{100} = 1,07$

$\hat{p} = \frac{\overline{X}}{n} = \frac{1,07}{2} = 0,535$

Se supoñemos $\hat{p} = 0,535$, podemos volver a facer o reparto e mellorar a aproximación de p.

1.4. Algoritmo EM con datos censurados por intervalos

Presentaremos agora un algoritmo para a estimación da función de distribución con datos censurados por intervalos. Aínda que é un algoritmo de tipo EM, os elementos para a súa construcción foron dados por Turnbull [23] nun artigo simultáneo ao de Dempster [5]. Además, no artigo de Turnbull [23] dase a estimación con datos non só censurados senón tamén truncados ou agrupados. Aquí ímonos ceñir aos datos censurados por intervalos.

Buscamos un estimador de máxima verosimilitude de F, sendo F a función de distribución da variable X, a partir dunha mostra de datos censurados por intervalos, que denotamos como $[L_1, R_1], \ldots, [L_n, R_n]$ sendo n o tamaño mostral. Entón a función de versomilitude ven dada por:

$$L^*(F) = \prod_{i=1}^{n} \left[F\left(R_i^+\right) - F\left(L_i^-\right) \right]$$

Unha das claves para entender o funcionamento do algoritmo consiste na definición dos intervalos de Turnbull. Coloquialmente, son todos os trozos que resultan facendo intersecciónss dos intervalos observados $[L_i, R_i]$. Dito doutro modo, cada intervalo de Turnbull ten extremo esquerdo en algún L_i, extremo dereito en algún R_i, e non contén outros puntos L_j ou R_j no seu interior. Suponse que existen m intervalos que verifican as condicións anteriores, que denotamos $[q_1,p_1]$, $[q_2,p_2]$,..., $[q_m,p_m]$, con $q_1 \leq p_1$ ¡$q_2 \leq p_2$ ¡$q_m \leq p_m$.

A clave para o funcionamento do algoritmo resida na idea de que para a verosimilitude non importa os valores intermedios dos intervalos, e ademais acadarase nos intervalos de Turnbull. O propio Turnbull [23] proporciona dous resultados teóricos ao respecto: No lema 1, afirma que unha estimación da función de distribución que non estea definida sobre o conxunto de intervalos de Turnbull non pode proporcionar un estimador de máxima verosimilitude, e o lema 2, que afirma que o comportamento da función de distribución nos interiores dos intervalos de Turnbull non afecta á verosimilitude.

Polo tanto, podemos reconstruír a verosimilitude $L^*(F)$ para que empregue unicamente os extremos de Turnbull nos cálculos, obtendo resultados máis eficientes. Definiremos $s_j = F(P_j^+) - F(q_j^-)$, para j=1,...,m. Loxicamente, a suma de todos os s_j teñen que valer 1 e curiosamente cada s_j será positivo. Isto último débese á monotonía, demostrada por Dempster [5]. Tamén se empregarán indicadores $\alpha_{ij} = I([q_j, p_j] \subset [L_i, R_i])$ que servirán para construír a verosimilitude, pois

$$L^*(F) = L^*(s_1, ..., s_m) = \prod_{i=1}^{n} \sum_{j=1}^{m} \alpha_{ij} s_j$$

Turnbull[23] presenta o seu estimador desde un enfoque de autoconsistencia, aínda que obtén como resultado un algoritmo de tipo EM. A informacióm completa neste contexto consistiría en coñecer en qué intervalo de Turnbull, $[q_j, p_j]$ se atopa a observación X_i, sabendo que pode estar en calquera dos que intersecan co seu intervalo observado $[L_i, R_i]$. Se tiveramos valores $s = (s_1, \ldots, s_m)$ para as probabilidades dos intervalos de Turnbull, a probabilidade condicionada de que o dato X_i pertenza ao intervalo $[q_j, p_j]$ sería

$$\mu_{ik}(s) = \frac{\alpha_{ik} s_k}{\sum_{j=1}^{m} \alpha_{ij} s_j}$$

Ao mesmo tempo, se tiveramos os valores μ_{ik}, sería moi sinxelo estimar $s = (s_1, \ldots, s_m)$ da seguinte maneira:

$$s_k = \frac{1}{n} \sum_{i=1}^{n} \mu_{ik}$$

Deste xeito, podemos construir un algoritmo iterativo que alterne as dúas expresións anteriores, que realmente serían os pasos E e M dun algoritmo de tipo EM:

Paso inicial. Partirase dun valor inicial para $s^0 = (s_1{}^0, ..., s_j{}^0, ..., s_m{}^0)$ Normalmente carecerase deste valor, pero dado que o método é iterativo, o lóxico é que con calquera valor inicial que cumpra certos requisitos conduciranos co algoritmo ao valor aproximado. A condición fundamental que debe verificar $(s_1{}^0, ..., s_j{}^0, ..., s_m{}^0)$ é que $\sum_{j=1}^{m} s_j^0 = 1$. O máis cómodo é coller todos os $s_j{}^0 = \frac{1}{m}$, para j=1,...,m, para que a suma dea 1.

Paso E. Obter os valores de

$$\mu_{ik}\left(s^{(p)}\right) = \frac{\alpha_{ik} s_k^{(p)}}{\sum_{j=1}^{m} \alpha_{ij} s_j^{(p)}}$$

Paso M. Calcular un novo valor de s como

$$s_k^{(p+1)} = \frac{1}{n}\sum_{i=1}^{n}\mu_{ik}(s^{(p)})$$

Repetimos os pasos E e M ata converxencia.

Proseguiremos o seguinte apartado cun exemplo do algoritmo de Turnbull con datos censurados en intervalos. Dito exemplo mostrará como é posible concretar a notación de Dempster ao caso do algoritmo de Turnbull. Usaremos ese exemplo de referencia para enlazar os algoritmos de Turnbull e Dempster.

1.5. Exemplo práctico do algoritmo EM

Nesta sección incluirase un exemplo sinxelo de datos censurados por intervalos co fin de ilustrar o funcionamento tanto do algoritmo EM aplicado a datos censurado como do algoritmo de Turnbull. Como veremos, non existen diferenzas entre ámbolos dous algoritmos salvo polo valor de partida da función de probabilidade, o que conduce a unha converxencia máis rápida de Turnbull. Combinaremos desenvolvementos xerais ca correspondente adaptación ao caso concreto do exemplo.

Tal como foi presentado en seccións previas, considerarase unha variable aleatoria censurada por intervalos, e tomarase unha mostra aleatoria de tamaño n X_i, con i=1,...,n. Para cada valor X_i existirá un intervalo $(L_i,R_i]$ verificando que $X_i \in (L_i,R_i]$ $\forall$ i=1,...,n.

No exemplo presentaranse 3 observacións X_1, X_2 e X_3 censuradas por intervalos. Para cada observación X_i, con $1 \leq i \leq 3$ presentaremos un intervalo $(L_i,R_i]$ no que está contida. Os valores serán os seguintes:

$$X_1 \in (L_1,R_1]=(0,2]$$

$$X_2 \in (L_2,R_2]=(1,4]$$

$$X_3 \in (L_3,R_3]=(3,Inf)$$

Denotaremos por Inf o valor. De forma gráfica, representamos a continuación para cada individuo o intervalo censurado no que está contido, a partición que empregaría o algoritmo EM e tamén os intervalos que se empregarán no algoritmo de Turnbull.

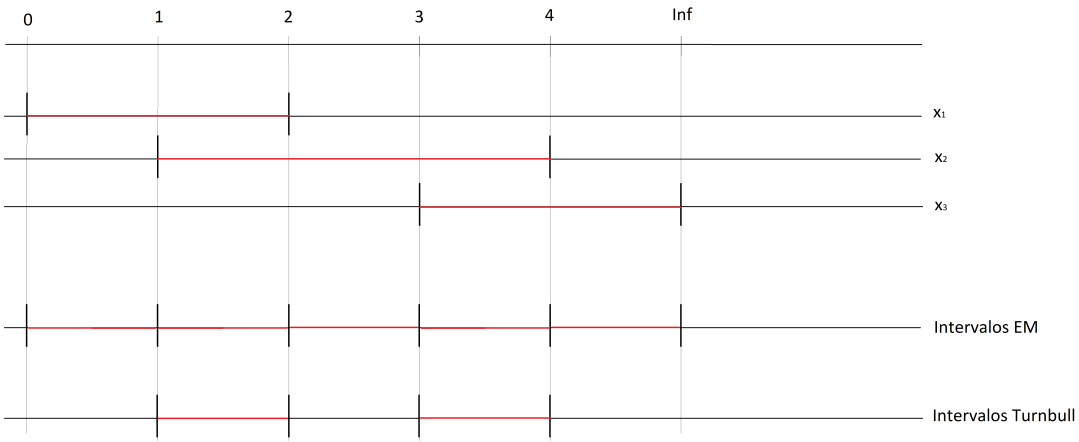


Figura 1.5.1. Intervalos censurados

A continuación, definiremos unha serie de valores t_j, con $0 \leq j \leq m+1$, que consistirán no conxunto de valores dos intervalos censurados onde pode variar a distribución. Tratándose de intervalos censurados unicamente nos extremos dos intervalos pode cambiar a función de distribución, polo que os valores t_j consistirán simplemente nos valores extremos distintos (tanto inferiores como superiores) que forman os intervalos. Como se pode apreciar no gráfico, os valores que forman a partición máis fina no noso exemplo serán os m+1=5 valores seguintes: t_0=0, t_1=1, t_2=2, t_3=3,t_4=4 e t_5=Inf. Polo tanto, os intervalos serán (0,1], (1,2], (2,3], (3,4] e (4,Inf].

A simple vista pode comprobarse que os intervalos de Turnbull non coinciden cos intervalos definidos para o algoritmo EM. Esta será a principal diferenza entre a aplicación de EM en bruto respecto da aplicación de Turnbull. Turnbull [23] presenta unha serie de resultados teóricos que garanten que a converxencia do algoritmo se produce no que se denomina intervalos de Turnbull. Tal e como os definimos no apartado anterior, estes intervalos restrinxen a partición máis fina de tal forma que unicamente se toman aqueles extremos esquerdo e dereito que non teñen no seu interior outro extremo. Polo tanto, no noso exemplo os intervalos de Turnbull serán (1,2] e (3,4].

A continuación desenvolveremos o algoritmo EM no exemplo. Tras finalizar o desenvolvemento retomaremos a mellora da eficiencia dada polos intervalos de Turnbull, que como veremos conduce a unha converxencia do algoritmo máis rápida.

Cando unha variable aleatoria está censurada nun intervalo prodúcese unha perda de información da distribución, pero a forma dos intervalos censurados pode axudarnos a recuperar parte desa información perdida.

Tendo en conta que $i \in [1,n]=[1,3]$, $j \in [0,m+1]=[0,5]$, se renomearemos cada un dos intervalos do exemplo incorporando os correspondentes t_j a cada extremo obtense $(L_1,R_1]=(t_0,t_2]$, $(L_2,R_2]=(t_1,t_4]$ e $(L_3,R_3]=(t_3,t_5]$.

A función de probabilidade estará definida en función de cada intervalo censurado e de cada extremo da partición. Así, definimos para cada observación censurada $X_i \in (L_i, R_i]$ a función de probabilidade condicionada como

$$p_i(t_j) = P(t_{j-1} < X_i \leq t_j | X_i \in (L_i, R_i])$$

para todo i=1,...,n e j=1,...,m+1. A probabilidade de que cada intervalo $(t_{j-1},t_j]$ con $1\leq j\leq m+1$ conteña algunha observación censurada vira dada por

$$p\left(t_j\right)=P\left(t_{j-1}<X_i\leq t_j\right)$$

Neste punto estamos en condicións de aplicar o algoritmo EM [5]. Recordemos que consta dunha fase de esperanza (paso E) e dunha fase de maximización (paso M). Resumiremos brevemente en que consiste estes dous pasos de maneira xeral para datos censurados en intervalos. Empregaremos no algoritmo a seguinte función indicadora:

$$I_{\{t_j\in(L_i,R_i]\}}=\begin{cases}1\ se\ t_j\in(L_i,R_i]\\0\ se\ t_j\notin(L_i,R_i]\end{cases}$$

Partiremos dunha probabilidade de base asociada a cada valor t_j, con $1\leq j\leq m+1$, que pode tomar calquera valor que cumpra unha única condición. Ao tratarse dunha función de probabilidade, a suma das m+1 probabilidades valerá 1. Normalmente, tomarase o seguinte vector m+1-dimensional

$$\hat{p}\quad=\quad\left[p\left(t_1\right),\ldots,p\left(t_{m+1}\right)\right]\quad=\quad\left[\tfrac{1}{m+1},\ldots,\tfrac{1}{m+1}\right].$$

Paso E. Para cada $1\leq i\leq n$ obtemos $\hat{p}_i\left(t_j\right)=\frac{\hat{p}(t_j)I_{\{t_j\in(L_i,R_i]\}}}{\sum_{t_k\in(L_i,R_i]}\hat{p}(t_k)}$, con $1\leq j\leq m+1$.

Polo tanto, trátase de obter

$$\hat{p}_1=\left[\hat{p}_1\left(t_1\right),\ldots,\hat{p}_1\left(t_{m+1}\right)\right]$$

$$\hat{p}_2=\left[\hat{p}_2\left(t_1\right),\ldots,\hat{p}_2\left(t_{m+1}\right)\right]$$

...

$$\hat{p}_n=\left[\hat{p}_n\left(t_1\right),\ldots,\hat{p}_n\left(t_{m+1}\right)\right]$$

Paso M: Para cada $1\leq j\leq m+1$ obteremos $\hat{p}\left(t_j\right)=\frac{1}{n}\sum_{i=1}^{n}\hat{p}_i\left(t_j\right)$, con $1\leq i\leq n$.

Como resultado, deducimos un novo valor de $\hat{p}=\left[p\left(t_1\right),\ldots,p\left(t_{m+1}\right)\right]$, que será a función de probabilidade da cal partiremos na seguinte iteración. Tratándose dun algoritmo iterativo, alternaremos pasos E e M en cada iteración ata acadar a converxencia.

Procedamos pois ca execución do algoritmo ao caso concreto do noso exemplo. Aproximaremos os cálculos a partir da cuarta cifra decimal.

Iteración 1

- Paso E. Partiremos do seguinte vector de probabilidades co fin de calcular os valores de $\hat{p}_1$, $\hat{p}_2$ e $\hat{p}_3$.

$$\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = \left[\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}\right]$$

Procederemos ca estimación de cada un dos valores de $\hat{p}_1$, $\hat{p}_2$ e $\hat{p}_3$

$$\hat{p}_1(t_1) = \frac{\hat{p}(t_1)}{\hat{p}(t_1)+\hat{p}(t_2)} = 0,5 \; ; \; \hat{p}_1(t_2) = \frac{\hat{p}(t_2)}{\hat{p}(t_1)+\hat{p}(t_2)} = 0,5 \; , \; \hat{p}_1(t_3) = 0 \; , \; \hat{p}_1(t_4) = 0 \; , \; \hat{p}_1(t_5) = 0$$

$$\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0{,}5, 0{,}5, 0, 0, 0]$$

$\hat{p}_2(t_1) = 0 \; ; \; \hat{p}_2(t_2) = \frac{\hat{p}(t_2)}{\hat{p}(t_2)+\hat{p}(t_3)+\hat{p}(t_4)} = 0,3333 \; , \; \hat{p}_2(t_3) = \frac{\hat{p}(t_3)}{\hat{p}(t_2)+\hat{p}(t_3)+\hat{p}(t_4)} = 0,3333 \; , \; \hat{p}_2(t_4) = \frac{\hat{p}(t_4)}{\hat{p}(t_2)+\hat{p}(t_3)+\hat{p}(t_4)} = 0{,}3333 \; , \; \hat{p}_2(t_5) = 0$

$$\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0{,}3333, 0{,}3333, 0{,}3333, 0]$$

$\hat{p}_3(t_1) = 0 \; ; \; \hat{p}_3(t_2) = 0 \; , \; \hat{p}_3(t_3) = 0 \; , \; \hat{p}_3(t_4) = \frac{\hat{p}(t_4)}{\hat{p}(t_4)+\hat{p}(t_5)} = 0{,}5 \; , \; \hat{p}_3(t_5) = \frac{\hat{p}(t_5)}{\hat{p}(t_4)+\hat{p}(t_5)} = 0{,}5$

$$\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0, 0, 0{,}5, 0{,}5]$$

- Paso M. Renovaremos o valor de $\hat{p}$. Para iso calcularemos una nova aproximación dos valores $\hat{p}(t_j)$ con 1≤ j≤ 5.

$$\hat{p}(t_1) = \frac{1}{3}[\hat{p}_1(t_1) + \hat{p}_2(t_1) + \hat{p}_3(t_1)] = \frac{1}{3}[0{,}5 + 0 + 0] = 0,1667$$

$$\hat{p}(t_2) = \frac{1}{3}[\hat{p}_1(t_2) + \hat{p}_2(t_2) + \hat{p}_3(t_2)] = \frac{1}{3}[0{,}5 + 0{,}3333 + 0] = 0{,}2778$$

$$\hat{p}(t_3) = \tfrac{1}{3}[\hat{p}_1(t_3) + \hat{p}_2(t_3) + \hat{p}_3(t_3)] = \tfrac{1}{3}[0 + 0{,}3333 + 0] = 0{,}1111$$

$$\hat{p}(t_4) = \frac{1}{3}[\hat{p}_1(t_4) + \hat{p}_2(t_4) + \hat{p}_3(t_4)] = \frac{1}{3}[0 + 0{,}3333 + 0{,}5] = 0{,}2778$$

$$\hat{p}(t_5) = \frac{1}{3}[\hat{p}_1(t_5) + \hat{p}_2(t_5) + \hat{p}_3(t_5)] = \frac{1}{3}[0 + 0 + 0{,}5] = 0,1667$$

$$\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [\,0{,}1667,\, 0.2778,\, 0,1111,\, 0.2778,\, 0{,}1667]$$

Iteración 2

- Paso E. Partindo do valor $\hat{p} = [0,1667,\, 0.2778,\, 0{,}1111,,\, 0.2778,\, 0{,}1667]$ obtido no paso anterior, realizaremos a estimación de $\hat{p}_1$, $\hat{p}_1$ e $\hat{p}_3$ dun modo análogo ao do paso E da iteración anterior.

$$\hat{p}_1(t_1) = \frac{\hat{p}(t_1)}{\hat{p}(t_1)+\hat{p}(t_2)} = 0.3750; \; \hat{p}_1(t_2) = \frac{\hat{p}(t_2)}{\hat{p}(t_1)+\hat{p}(t_2)} = 0.6250,$$

$$\hat{p}_1(t_3) = 0 \; , \; \hat{p}_1(t_4) = 0 \; , \; \hat{p}_1(t_5) = 0 \; ,$$

$$\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0{,}3750,\ 0.6250,\ 0,\ 0,\ 0]$$

$$\hat{p}_2(t_1) = 0\ ;\ \hat{p}_2(t_2) = \frac{\hat{p}(t_2)}{\hat{p}(t_2)+\hat{p}(t_3)+\hat{p}(t_4)} = 0.4167,\ \ \hat{p}_2(t_3) = \frac{\hat{p}(t_3)}{\hat{p}(t_2)+\hat{p}(t_3)+\hat{p}(t_4)} = 0.1667,$$

$$\hat{p}_2(t_4) = \frac{\hat{p}(t_4)}{\hat{p}(t_2)+\hat{p}(t_3)+\hat{p}(t_4)} = 0.4167,\ \ \hat{p}_2(t_5) = 0$$

$$\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0{,}4167, 0{,}1667,\ 0{,}4167,\ 0]$$

$$\hat{p}_3(t_1) = 0\ ;\ \hat{p}_3(t_2) = 0\ ,\ \ \hat{p}_3(t_3) = 0\ ,\ \ \hat{p}_3(t_4) = \frac{\hat{p}(t_4)}{\hat{p}(t_4)+\hat{p}(t_5)} = 0.6250,$$

$$\hat{p}_3(t_5) = \frac{\hat{p}(t_5)}{\hat{p}(t_4)+\hat{p}(t_5)} = 0.3750$$

$$\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0, 0, 0{,}6250,\ 0{,}3750]$$

- Paso M. Obteremos cada valor $\hat{p}(t_j)$ con $1 \leq j \leq 5$.

$$\hat{p}(t_1) = \tfrac{1}{3}[\hat{p}_1(t_1) + \hat{p}_2(t_1) + \hat{p}_3(t_1)] = \tfrac{1}{3}[0{,}3750 + 0 + 0] = 0{,}1250$$

$$\hat{p}(t_2) = \tfrac{1}{3}[\hat{p}_1(t_2) + \hat{p}_2(t_2) + \hat{p}_3(t_2)] = \tfrac{1}{3}[0{,}6250 + 0{,}4167 + 0] = 0{,}3472$$

$$\hat{p}(t_3) = \tfrac{1}{3}[\hat{p}_1(t_3) + \hat{p}_2(t_3) + \hat{p}_3(t_3)] = \tfrac{1}{3}[0 + 0{,}1667 + 0] = 0{,}0556$$

$$\hat{p}(t_4) = \tfrac{1}{3}[\hat{p}_1(t_4) + \hat{p}_2(t_4) + \hat{p}_3(t_4)] = \tfrac{1}{3}[0 + 0{,}4167 + 0{,}0{,}6250] = 0{,}3472$$

$$\hat{p}(t_5) = \tfrac{1}{3}[\hat{p}_1(t_5) + \hat{p}_2(t_5) + \hat{p}_3(t_5)] = \tfrac{1}{3}[0 + 0 + 0{,}3750] = 0{,}125$$

$$\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [0,1250,\ 0,3472,\ 0,0556,\ 0,3472,\ 0,1250]$$

Na iteración 3 partirase da aproximación $\hat{p}$ anterior, e repetiranse os pasos E e M. Este proceso iterarase ata converxencia. Quedando ilustrado o seu funcionamento, presentase un resumo dos resultados das iteracións 1-25 no apéndice A. Para a iteración 25, cunha aproximación de 4 cifras decimais, obtivéronse os seguintes resultados:

Iteración 25

- Paso E:

$$\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0, 1, 0, 0, 0]$$

$$\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0{,}5, 0, 0, 0]$$

$$\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0, 0, 1, 0]$$

- Paso M:

$$\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [0,\ 0.5,0,0.5,0]$$

Con isto conclúe a aplicación do algoritmo EM. Agora abordaremos o papel dos intervalos de Turnbull. Recordemos que a partición máis ideada por Turnbull non coincide necesariamente ca partición máis fina que se aplica no algoritmo EM en bruto. No noso exemplo, reducíamos de 5 a 2 o número de intervalos da partición máis fina a ter en conta. A única diferencia é que cos intervalos de Turnbull partimos dunha distribución de probabilidades mellor escollida para iniciar o algoritmo EM.

Recordemos que no contexto do exemplo os intervalos de Turnbull toman valores (t_1, t_2] =(1,2] e (t_3, t_4] =(3,4]. A probabilidade de que algunha observación censurada estea en algún dos intervalos non seleccionados por Turnbull será nula. Repartindo de maneira equitativa a probabilidade entre os intervalos de Turnbull seleccionados, obterase que:

$$\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [0,\ 0.5,0,0.5,0]$$

que, curiosamente, coincide co valor ao que converxe o método. Outros casos non se resolverán de maneira tan inmediata, e precisarán iterar o algoritmo de Turnbull ata a converxencia dun modo análogo ao presentado antes. En calquera caso, é doado comprobar como se reduce considerablemente o número de iteracións ata converxencia.

Capítulo 2

Regresión cuantil

2.1. As limitacións da regresión en media

Neste segundo capítulo dedicarase a un recurso tan importante para a estatística como pode ser a regresión, sendo o seu uso a análise da relación entre variables. En concreto, se a regresión é simple estúdase como varía unha variable Y (variable resposta) en función doutra variable X (variable explicativa). No caso de existir varias variables explicativas, a regresión é múltiple.

Centrándonos no caso simple as condicións que verifican as variables e a forma de relacionarse entre elas determina os distintos tipos de regresión, sendo a forma máis coñecida a regresión lineal. Represéntase a relación por Y=Xβ +ε , onde β =(β_0, β_1) no caso simple, e β =(β_0,..., $\beta_{\text{p-1}}$) no caso de p-1 variables explicativas. Ademais, ε correspondese cos erros da regresión.

A forma de obter o vector paramétrico β é a partir da redución do valor dos residuos ao valor mínimo, e para iso o método lineal emprega o método de mínimos cadrados. En dito método considéranse dous conceptos importantes: os valores aproximados da variable resposta, cuxo valor está determinado pola recta de regresión; e os residuos da regresión, que son a diferenza entre os valores reais e aproximados da variable resposta.

O método consiste en minimizar a suma dos residuos da regresión elevados ao cadrado, para o cal é necesario atopar os valores de β de tal xeito que dita suma sexa a mínima posible. Formalmente, se tomamos unha mostra de tamaño n das variables (X,Y) da forma (x_i, y_i), con i=1,...,n, onde X pode representar unha única variable explicativa (caso simple, $x_i \in X$) ou p-1 variables explicativas (caso múltiple $x_i=(x_{i1},..., x_{ip-1}) \in X$). Para estes casos, o método de mínimos cadrados ten por obxectivo minimizar:

Caso simple:

$$\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^{'} \beta \right)^2$$

Caso múltiple:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

Este método permite establecer a estimación da recta de regresión lineal que mellor se axusta aos datos tendo en conta as desviacións cadráticas respecto dos erros. Dito axuste garante que a recta de regresión pase pola media dos datos.

Consideremos por un momento unicamente unha das variables a nivel poboacional, concretamente a variable resposta Y. En ausencia de outras variables, o valor esperado da variable Y, que proporciona as menores desviacións cadráticas, será a media. Isto é:

$$IE\left(Y\right) = arg\min_{y} IE\ \left[\left(Y-y\right)^2\right]$$

Trasladando o anterior ao terreo mostral, tomarase unha mostra da variable Y de tamaño n, cuxos valores denotaremos por Y_1,...,Y_n, O valor esperado de Y tendo en conta a mostra anterior é:

$$\overline{Y} = arg\min_{y} \frac{1}{n}\sum_{i=1}^{n}\left(Y-y\right)^2$$

Para que a regresión lineal simple funcione, son necesarios catro supostos: a relación entre as variables ten que ser lineal, homocedasticidade da varianza, normalidade dos erros e independencia dos erros. Non obstante, estes supostos non sempre se verifican, polo que se fai patente a necesidade de buscar métodos alternativos de regresión simple onde a rixidez destes supostos sexa menor. Dito cambio é posible grazas a cambiar o método de mínimos cadrados por outro método que, como veremos, empregará o concepto de cuantil.

2.2. O cuantil como instrumento na regresión

Recordemos o cuantil corresponde cun instrumento estatístico que calcula o valor dunha determinada variable en función da posición que ocupa na mostra ordenada. Tras a ordenación da mostra de menor a maior, dividirase a mesma en tantas partes como indica o cuantil, e buscarase os valores que ocupan os cortes (no caso de que o corte este situado entre dous valores, calcúlase a media dos dous valores).

Formalmente, dada unha variable aleatoria X cunha función de distribución asociada F, e un cuantil de orde p comprendida entre 0 e 1, defínese o cuantil como o valor x da variable que fai que F(x)=P(X$\leq$ x)=p.

Aínda que comprender o concepto de cuantil é importante, dito concepto non se aplica directamente á regresión, senón que a aplicación é a través da función de perda cuantílica. Para unha variable determinada U con u ϵ U, e un cuantil de orde τ , defínese a función de perda cuantílica como

$$\rho_\tau\left(u\right) = \left\{\begin{array}{cc} \tau & se\ u \geq 0 \\ -\left(1-\tau\right)u & se\ u < 0 \end{array}\right\}$$

Conseguintemente, os valores resultantes da aplicación da función serán positivos, e en función do valor da orde do cuantil os valores negativos estarán máis penalizados que os positivos (τ entre 0 e 0.5), igualmente penalizados (τ =0.5) ou menos penalizados (τ entre 0.5 e 1). Vexamos na figura 2.1 como transforma o cuantil unha variable cuxos valores están comprendidos entre -1 e 1:

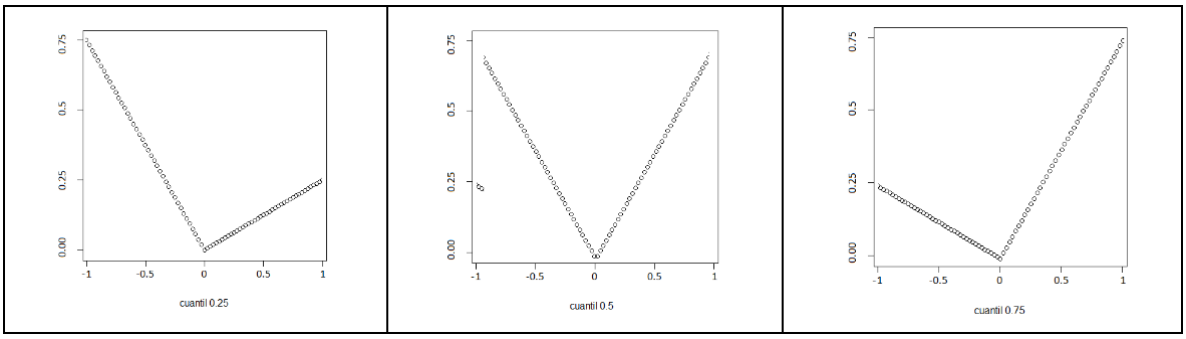


Figura 2.1: Función de perda cuantílica

Esta función será a base da regresión cuantil, posto que permitirá intercambiar as desviacións cadráticas dos residuos por desviacións absolutas ponderadas polo cuantil.

Recordemos que no caso dos mínimos cadrados, se considerábamos unha única variable Y, o valor esperado de dita variable viña determinado polo valor que proporciona unha mínima perda cadrática, sendo dito valor a media. Imaxinemos que en lugar de considerar perda cadrática empregásemos desviacións absolutas. Loxicamente, a media non sería o valor que minimizaría as desviacións. No seu lugar sería a mediana a que proporcionaría un argumento mínimo, isto é:

$$Mediana\,(Y) = arg\,\min_{y} IE|Y - y|$$

Considerando o nivel mostral, para unha mostra Y_1,...,Y_n tense:

$$Mediana\ \ mostral\,(Y) = arg\,\min_{y}\ \frac{1}{n}\sum_{i=1}^{n}|Y - y|$$

Cun razoamento análogo ao anterior, podemos considerar como alternativa a perda cuantílica. Sexa $Q_Y(\ \tau\ )$ o cuantil de orde τ respecto da variable Y, e sexa Y_1,...,Y_n unha mostra aleatoria de tamaño n. A nivel poboacional, verifícase que

$$Q_Y\,(\tau) = arg\,\min_{y} IE\,(\rho_\tau\,(Y - y))$$

e a nivel mostral

$$\hat{Q}_Y\,(\tau) = arg\,\min_{y}\,\frac{1}{n}\sum_{i=1}^{n}(\rho_\tau)\,(Y - y)$$

2.3. A regresión cuantil

A regresión cuantil mantén o mesmo obxectivo que a regresión lineal, que consiste en atopar unha recta, e polo tanto uns parámetros β_0 e β_1 (caso simple) ou un vector de parámetros β p dimensional (caso múltiple) que reduzan a un valor mínimo os residuos. Porén, diferéncianse na forma en que se conseguen que a suma residual sexa mínima.

Na regresión simple, cada residuo era elevado ao cadrado, e facíase a suma global. Non obstante, na regresión cuantil cada residuo non se eleva ao cadrado senón que se lle aplica unha función de perda cuantílica (recordemos que, igual que os cadrados, a función de perda cuantílica deixa

valores positivos). Do mesmo xeito que se presentou previamente, dependendo da orde do cuantil que se lle aplique os residuos positivos ou negativos terán un peso distinto na valoración global.

Formalmente, tomando unha mostra de tamaño n (x_i,y_i) das variables (X,Y) (recordemos, caso simple $x_i \epsilon$ X, ou caso múltiple x_i?=(x_{i1},..., x_{ip-1})ϵ X con p-1 variables explicativas) o obxectivo será minimizar

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i^{'} \beta \right) \text{(caso múltiple)}$$

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} \rho_\tau \left(y_i - \beta_0 - \beta_1 x_i \right) \text{(caso simple)}$$

Tal e como adiantamos na primeira sección deste apartado, o enfoque cuantil da regresión proporciona unha relaxacións das restricións do modelo. En concreto, a regresión cuantil pode aplicarse a contextos con ausencia de normalidade ou incluso con ausencia de homocedasticidade. Por último, outra vantaxe que presenta este modelo respecto ao modelo lineal é mellor axuste fronte a datos atípicos.

A forma de realizar esta estimación correspóndese a métodos de programación lineal. Esbozaremos a continuación as principais liñas de actuación.

O primeiro paso consistirá en renomear os residuos, de tal xeito que en lugar de considerar residuos con valores positivos e negativos, pasarase a ter a diferenza entre dous vectores positivos. Para iso, partindo do caso xeral, en lugar de considerar $y_i - x_i^{'}\beta$ considerarase $u_i - v_i$, sendo u_i=0 se o residuo é negativo e v_i=0 se é positivo. Co anterior introducirase dúas variables u=(u_1,..., u_n)' e v=(v_1,..., v_n)', e o problema pasa a ser

$$\min_{\beta} \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i^{'} \beta \right) = \min_{\beta} \sum_{i=1}^{n} \left[\tau_i u_i - (1 - \tau) v_i \right]$$

Polo tanto, o problema de programación lineal consistirá en

$$\min_{\beta} \sum_{i=1}^{n} \left[\tau_i u_i - (1 - \tau) v_i \right]$$

suxeito a

$$Y - X\beta = u - v$$

$$\beta \epsilon I\!R^p, u_i \geq 0,\ v_i \geq 0,\ 1 \leq i \leq n\}$$

Unha forma de resolver o anterior corresponde co uso do método simplex, aínda que existe unha versión máis eficiente, coñecida como o algoritmo de Barrodale e Roberts [2]

2.4. Inferencia sobre os parámetros

Outra das cuestións básicas que presenta a regresión cuantil consiste na realización dunha inferencia estatística para estimar os parámetros. Considerando unha mostra $\{(x_i, y_i)\}_{i=1}^{n}$, supoñeremos, como fixemos con anterioridade, que o comportamento da mostra é lineal.

Se denotaremos por $Q_Y(\tau)$ o cuantil de orde τ da variable Y, entón a suposición da linealidade da regresión cuantil equivale a que o cuantil condicional verifique $Q_{y_i}(\tau|x_i) = x_i^{'}\beta$. Para a inferencia consideraremos F_i como a función de distribución asociada a y_i.

Imos supoñer para a inferencia que se verifican os seguintes supostos (para todo $1 \leq i \leq n$):

- As función de distribución F_i son absolutamente continuas.
- As funcións de densidade f_i son absolutamente continuas.
- $f_i\left(Q_{y_i}(\tau|x_i)\right) \epsilon\, (0, \infty)$
- Existen as matrices simétricas e definidas positivas D_0 e $D_1(\tau)$ verificando
- $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} x_i x_i^{'} = D_0$
- $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f_i\left(Q_{y_i}(\tau|x_i)\right) x_i x_i^{'} = D_1$
- $\max_{1\leq i\leq n} \frac{||x_i||}{n} \to 0$

Baixo estas condicións, aproximaremos o parámetro β mediante $\hat{\beta}$ do seguinte xeito:

$$\hat{\beta} = arg \min_{\beta \epsilon\, I\!R^p} \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i^{'}\beta\right)$$

Veremos agora a converxencia asintótica do parámetro $\hat{\beta}$. Baixo as condicións anteriores, tense que

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \to N\left(0, \tau\,(1-\tau)\, D_i^{-1} D_0\; D_i^{-1}\right)$$

No caso de que os erros teñan a mesma distribución, e polo tanto F_1=...=F_n tense que

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \to N\left(0, \frac{\tau\,(1-\tau)\, D_0^{-1}}{f_i^2\left(Q_{y_i}(\tau|x_i)\right)}\right)$$

O teorema e condicións anteriores están recollido na obra de Koenker [9]

2.5. Librería *quantreg* de R

Afortunadamente, o software estatístico R inclúe rutinas baseadas en regresión cuantil cuxo cometido será o axuste da recta de regresión. A librería quantreg inclúe a función rq, que obtén a recta de regresión introducindo os valores da variable explicativa e a variable resposta tendo en conta a orde do cuantil. Por defecto, ese cuantil será a mediana a menos que se indique o contrario. Aínda que nesta sección non se detalla, en seccións posteriores veremos que esta función inclúe a posibilidade de incorporar un vector de pesos que ponderará cada observación da variable resposta da regresión cuantil.

Co fin de ilustrar o funcionamento da regresión cuantil, mostraremos un exemplo en R sen censura formado por dúas covariables. Os datos empregados, que proveñen do Instituto Nacional

de Estatística, correspóndense co consumo anual de familias durante o período comprendido entre o ano 2004 e o ano 2018. Dunha banda, incluirase para cada ano o gasto en consumo total dos fogares e doutra, o gasto en hoteis, cafés e restaurantes [7]

Tras introducir os paquetes e datos correspondentes, ca función rq axustarase a recta de regresión cun determinado cuantil. Na figura 2.2. ofreceremos diferentes cuantís e observaremos cal é o efecto de cada recta de regresión. Representarase en gris as rectas de regresión cuantil para os valores de 0.1 a 0.9, con saltos 0.1. En vermello, representaremos a recta de regresión cuantil para un cuantil de orde 0.5, en verde para cuantil 0.1 e en azul para cuantil 0.9.

Como podemos ver na figura 1, a función de perda cuantíliza axusta as rectas de regresión de tal forma que, en función do cuantil pertinente, a recta pase por determinados valores e teña unha pendente determinada. Por exemplo, para o cuantil 0.1 a recta axustada ten un intercepto de -258.96724, e unha pendente de 8.46783, pero para o cuantil 0.9 o intercepto é 245.22131 e a pendente 4.17891. Como pode apreciarse, existen grandes diferenzas para o axuste da recta en función do cuantil empregado.

O código correspondente a este exemplo está contido no apéndice B. Graficamente, apreciase que a homocedasticidade non supón un impedimento para o axuste do modelo. Mentres que un modelo de regresión lineal baseado en mínimos cadrados non sería posible sen realizar un axuste dos datos ou sen realizar unha eliminación de datos atípicos e influentes, no caso da regresión cuantil a falta de simetría dos erros non supón un problema.

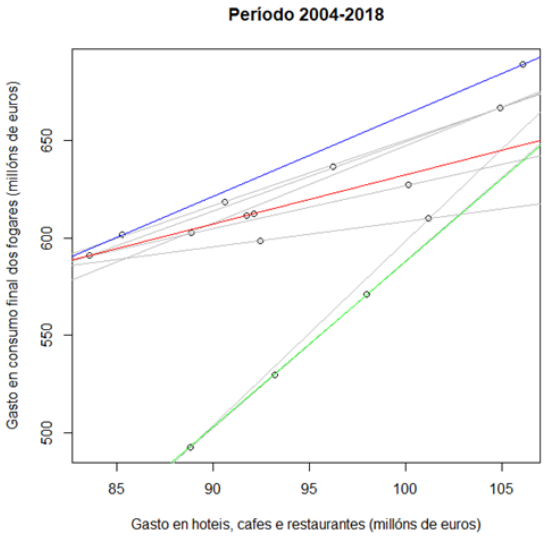


Figura 2.2: Axuste da regresión cuantil

Capítulo 3

Estimación da regresión cuantil con datos censurados por intervalos

3.1. Estimación da regresión con datos censurados por intervalos.

A problemática dos modelos de regresión con datos censurados foi estudada por diversos autores, variando a metodoloxía en función da forma da censura ou dos criterios de eficiencia. Neste apartado abordarase o traballo realizado por outros autores ao respecto, e veremos en que punto se sitúa a nosa proposta.

Unha das características que dividen aos modelos é a forma en que se produce a censura. Recordemos que aínda que existe a censura informativa, o noso traballo está centrado no caso da censura non informativa. Neste último tipo, son moitos os autores [26] que dividen en dous os casos principais de censura, sendo o primeiro un caso particular do segundo. O caso I de datos censurados, coñecido comunmente como estado actual dos datos, preséntase cando para cada dato só se ten unha única referencia temporal que indique se un determinado suceso asociado a dito caso aconteceu ou non. Pola contra, o caso II de datos censurados corresponde cos datos censurados por intervalos, onde o suxeito pode presentar censura pola dereita, pola esquerda ou en intervalos.

Dada a diferenza de información que aporta un caso ou outro, os estimadores da función de distribución serán diferentes, tendo o caso I nalgunhas ocasións forma pechada, fronte o caso II que non ten unha forma pechada. O modelo que será proposto na sección seguinte pode aplicarse ao caso II, que é máis xenérico e que aporta menos información, pero tamén ao caso I.

A relación entre regresión e datos censurados foi presentada por Tobin [22]. En relación ao problema da censura e regresión, abordase a posibilidade de que a relación entre unha variable aleatoria explicativa non censurada e unha variable resposta censurada estea determinada por unha relación lineal directa. Denotando por T a variable resposta censurada e Z a variable explicativa non censurada, verificarase a relación

$$\mathrm{T}= \beta\text{'}\mathrm{Z}+ \varepsilon$$

sendo β un vector de parámetros e ε independente e identicamente distribuído. Para unha mostra $(\mathrm{T_i}, \mathrm{Z_i})$ para i=1,...,n, obterase que

$$T_i = Z_i^T \beta + \varepsilon_i.$$

Como é lóxico, existen multitude de modelos que tratan a cuestión da regresión e os datos censurados por intervalos cando o modelo non presenta unha forma lineal tan directa como a anterior. Zhang e Sun [26] abordan os modelos semiparamétricos máis importantes, co inconvinte de que ditos modelos están baseados no método de mínimos cadrados para os axustes. Non obstante, incluiremos a continuación algúns deses métodos dado a relación que garda ca nosa proposta.

O modelo de tempo de vida acelerado.

Cando se presenta unha variable aleatoria censurada T, e conxuntamente unha variable non censurada Z, o modelo de tempo de vida acelerado toma a forma

$$\log(\mathrm{T}) = \beta\text{'}\mathrm{Z} + \varepsilon$$

sendo β o vector de parámetros e ε un erro cuxa distribución é descoñecida. Para iso, partindo da estimación da función de distribución do erro, Rabionwitz e Betensky [4] proporcionaron os seus respectivos métodos. Para salvar o problema da estimación do erro, Li e Pu [14] aportaron un método que salvaba o problema da distribución do erro empregando unha función de rango. Aínda que por vías diferentes, o método de regresión cuantil proposto neste traballo relaxa a restrición da distribución de erros que habitualmente se presenta na regresión baseada en métodos cadráticos.

O modelo lineal parcial.

Consiste nunha ampliación do anterior. Neste caso considerarase unha variable aleatoria censurada T e dúas variables non censuradas Z_1 e Z_2. Respecto do modelo, a variable Z_2 relacionase cas demais variables mediante unha función de suavizado g,

$$\log(\mathrm{T}) = \beta\text{'}\mathrm{Z}_1 + \mathrm{g}(\mathrm{Z}_2) + \varepsilon$$

sendo a función de suavizado descoñecida e a distribución dos erros ε coñecida. Autores como Shiboski [19] presentaron modelos nesta liña.

Tanto este modelo como o anterior están orientados á censura coñecida como tempos de faio, incluído no caso I de datos censurados.

O modelo de transformación lineal.

Nas condicións anteriores, sendo T a variable aleatoria censurada, Z unha variable aleatoria non censurada, o modelo de transformación lineal supón que se produce unha relación lineal entre as variables Z e T mediante unha función de variable real crecente h, a priori descoñecida. Por conseguinte, o modelo tomará a forma

$$\mathrm{h}(\mathrm{T}) = \beta\text{'}\mathrm{Z} + \varepsilon$$

representado ε os erros cuxa distribución é coñecida. Na literatura, pode atoparse desenvolvementos inferenciais na obra de Sun e Sun [21] Younes e Lachin [24] e Zhang [25]

Existen outros moitos exemplos de modelos de regresión que empregan datos censurados: dende modelos semiparamétricos, como o modelo de riscos proporcionais, ata modelos paramétricos como o das familias exponenciais. Todos estes modelos presentan como gran dificultade a estimación dunha determinada función de base.

En xeral, os modelos antes resumidos inclúen diferentes formas de linealidade na relación entre as variables. O modelo proposto neste documento comparte esta relación, polo tanto será posible nalgúns casos empregar o noso modelo como alternativa, mais non en tódolos casos.

A diferenza máis importante dos modelos anteriores respecto da nosa proposta é a forma de axuste da recta de regresión: mentres que todas as propostas anteriores están baseadas na estimación da recta de regresión mediante o método de mínimos cadrados, a nosa aposta será mediante a función de perda cuantílica. Ao longo desta sección, vimos a importancia que ten admitir certos supostos sobre os erros, como a súa distribución e por isto, tal como se abordou no capítulo anterior, ten especial importancia o uso da regresión cuantílica ao non presupoñer un determinado tipo de forma dos erros.

Existe múltiple bibliografía sobre os modelos de regresión que incorporan a función de perda cuantílica, e para abordar o noso problema será preciso ver en que consisten as principais contribucións e en que se diferencian da nosa proposta.

Na obra de Koenker [11] inclúense varios dos métodos máis destacados que incorporan censura pola dereita e regresión cuantil. Toma como punto de partida o caso de Tobin [22], presentado con anterioridade, e inclúe en primeiro lugar uns valores C_i con i=1,...,n, (en xeral non observables, aínda que en Powell [17] sí o son) en segundo lugar define $Y_i = max\{C_i, T_i\}$, tras o cal proporciona unha estimación para o valor do vector de parámetros

$$\hat{\beta} = arg \min_{b \epsilon \, I\!R^p} \sum_{i=n}^{n} \rho_\tau \left(Y_i - max\{C_i, x_i^T b\} \right).$$

Continuando co anterior, Powell [17] traslada á regresión cuantil o modelo de vida acelerado para datos censurados. Para iso, e tendo en conta a notación antes incluída, considera os valores C_i e Y_i antes definidos e unha mostra (T_i, Z_i) para i=1,...,n verificando

$$\log\left(T_i\right) = Z_i^T \beta + \varepsilon_i$$

.

O modelo de regresión cuantil ven dado pola función:

$$Q_{log(Y_i)|Z_i}\left(\tau | Z_i\right) = Z_i^T \beta\left(\tau\right)$$

. O estimador de Powell minimiza:

$$\sum_{i=n}^{n} \rho_\tau \left(Y_i - max\{C_i, x_i^T b\} \right)$$

Naturalmente, reláxase a restrición de que os erros eran independentes e identicamente distribuídos.

Existe software dispoñible con regresión cuantil que permite implementar o anterior. Por unha banda, Koenker [10] inclúe no paquete quantreg de R a función crq complementada coa función survival para o caso do estimador de Powell para o tempo de supervivencia acelerado. Dita función tamén inclúe o método de Portnoy [16]. Resumiremos este método.

Na obra de Portnoy presentase o modelo de tempo de vida acelerado dun xeito moi similar a como acabamos de definilo. A través do estimador de Kaplan-Meier da función de supervivencia, e da reordeación dos tempos do evento $Y_{(i)}$ con i=1,...,n cos correspondentes indicadores de censura $\delta_{(i)}$, Koenker [12] constrúe pesos asociados a cada variable a partir da función

$$\hat{\xi}(\tau) = arg\min_{\xi} \sum_{i=n}^{n} \rho_\tau (Y_i - \xi)$$

con i$\in\left(\frac{i-1}{n}, \frac{i}{n}\right)$. Se τ_i denota o valor para o cal $\hat{\xi}(\tau_i) = Y_{(i)}$ con $\delta_{(i)}$=0, os pesos serán , $w_i(\tau) = \frac{\tau - \tau_i}{1-\tau_i}$.

Unha das contribucións máis interesantes de Koenker é a librería quantreg de R [10], que inclúe unha serie de rutinas centradas en datos censurados. Nos apartados vindeiros, empregaremos algúns dos algoritmos para a execución do modelo de regresión cuantil con datos censurados proposto neste traballo.

Uns autores que abordan a cuestión da censura por intervalos e a regresión cuantil son Zhou e outros [27]. Resumiremos en que consiste a súa proposta e para iso, dada a forma singular en que se desenvolve este método, empregaremos a mesma notación que usa Zhou. Consideremos unha variable aleatoria y_i e unha covariable x_i. Para $\tau \in (0,1)$, expresa o cuantil condicional para un vector de parámetros $\theta(\tau) \in IR^m$ como

$$Q_{Y_i}(\tau|x_i) = \theta(\tau)^T x_i, i = 1, \ldots, n.$$

Como é lóxico, ao haber censura suporase que y_i está censurada, polo que non se coñecerá o valor directamente, senón dous valores t_{1i} e t_{2i} que verificarán para todo i=1,...,n que P($t_{1i} \leq y_i \leq t_{2i}$)=1. Pode comprobarse que o caso ata este punto resulta análogo ao proposto por Powell e, de feito, o autor emprega parte da literatura presentada por Powell no desenvolvemento. Non obstante, presenta dun xeito interesante a relación da función de distribución aproximada ao respecto da variable censurada e o intervalo de censura. A partires da función de perda cuantílica obtén para cada i=1,...,n

$$\widetilde{F}_i(\tau, \theta) = \begin{cases} \tau|y_i - \theta^T x_i| & se\ y_i \geq \theta^T x_i \\ (1-\tau)|y_i - \theta^T x_i| & se\ y_i < \theta^T x_i \end{cases}$$

Porén, compre aclarar que cada dato está situado nun intervalo censurado. Para incorporar este efecto na función de distribución, modificamos o anterior incorporando para cada i os dous valores xa citados t_{1i} e t_{2i} chegando a conclusión de que

$$\widetilde{F}_i(\tau, \theta) = \begin{cases} \tau|y_i - \theta^T x_i| & se\ t_{1i} \geq \theta^T x_i \\ \Psi_i(\tau, \theta) & se\ t_{1i} < \theta^T x_i \leq t_{2i} \\ (1-\tau)|y_i - \theta^T x_i| & se\ t_{2i} < \theta^T x_i \end{cases}$$

con $\Psi_i(\tau, \theta)$ descoñecida, polo que non entraremos en detalles. Polo tanto, a estimación dos parámetros virá dada polo proceso de optimización

$$arg\min_{\theta \in \Theta} \{\sum_{i=1}^{n} \widetilde{F}_i(\tau, \theta)\}$$

No manuscrito definitivo, Zhou e outros [27] proporciona unha serie de desenvolvementos teóricos para a minimización e sobre a converxencia normal. Tamén demostra a converxencia de $\hat{\theta}_n(\tau) \to \theta_0(\tau)$. Non nos deteremos nisto, pero si na idea que presenta Zhou sobre a corrección do sesgo. Presenta dous métodos para corrixilo, que resumiremos a continuación.

O método bootstrap empregado por Zhou consiste, como é lóxico, nun método iterativo onde para un número B de iteracións, seguindo en cada iteración catro pasos aborda o problema do sesgo.

Isto inclúe a revisión sobre a literatura máis próxima á nosa proposta, tendo presente que, como veremos a continuación, ningún dos métodos coincidirá ca nosa proposta. Abordaremos a construción do noso modelo partindo dun axuste lineal entre unha variable explicativa e unha variable resposta censurada tal como foi introducido por Tobin [22].

3.2. O método proposto de estimación da regresión cuantil con datos censurados.

Consideramos Y unha variable resposta e x unha covariable. Para unha mostra aleatoria independente (Y_i, x_i) de tamaño n, existirán un intercepto β_0 e unha pendente β_1 que verifican

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

sendo ε_i os erros.

Se pola contra, tivésemos interese en aportar un caso múltiple, bastaría con ampliar o caso anterior a un vector paramétrico p-dimensional β e transformar o valor x_i nun vector p dimensional, verificando:

$$Y_i = x_i^{'}\beta + \varepsilon_i$$

A variable resposta será unha variable censurada por intervalos polo que consideraremos, dun xeito similar a Rabinowitz [4], para cada valor de Y_i un vector $(t_{i1}, ..., t_{in_i})$ de tempos de observación ordenados de xeito que existirá un valor k coñecido que cumpra $t_{ik} \leq Y_i \leq t_{ik+1}$. Renomearemos o anterior como $L_i \leq Y_i \leq R_i$ para todo i=1,..,n. Igual que noutras propostas, podemos asumir que no caso L_i=0 existiría censura pola esquerda e no caso R_i=∞ censura pola dereita.

Na primeira sección do noso modelo abordamos a situación de censura por intervalos e presentamos dous métodos para a estimación da función de distribución para datos censurados, sendo o método de Turnbull un método adecuado para o caso da censura por intervalos. Denotamos os intervalos de dita partición como [q_1,p_1], [q_2,p_2],..., [q_m,p_m]. Ademais, calcularemos o vector (c_1,...,c_m) asociado a estes intervalos de tal forma que

$$c_j = \begin{cases} \frac{q_j+p_j}{2} & \text{se } p_j \neq \infty \\ q_j & \text{se } p_j = \infty \end{cases}$$

con 1$\leq$ t$\leq$ m.

A execución do método proporciona para cada observación a probabilidade de que pertenza a cada un dos intervalos de Turnbull. Nalgúns casos, a probabilidade será nula, pero noutros tomará unha certa probabilidade obtida a partir da función de distribución estimada. Non repetiremos todo o razoamento esgrimido na sección primeira de como se obteñen os intervalos de Turnbull ou de cal é o desenvolvemento teórico do algoritmo. Simplemente, executaremos o método ata converxencia [23].

Como resultado da execución do algoritmo, para cada dato censurado Y_i obtense un vector (μ_{i1},...,μ_{im}) con 1$\leq$ i$\leq$ n, sendo μ_{ij}=P($Y_i \in$ [q_i,p_j]) para todo 1$\leq$ j$\leq$ m. Aquí retómase a idea presentada no capítulo 1 de como calcular os pesos de Turnbull en cada intervalo de Turnbull mediante

$$\mu_{ik}(s) = \frac{\alpha_{ik} s_k}{\sum_{j=1}^{m} \alpha_{ij} s_j}$$

para logo calcular os pesos condicionais en cada intervalo $[L_i, R_i]$

Chegados a este punto faremos unha breve paréntese no noso modelo. O seguinte aspecto a tratar do método está relacionado cunha idea importante presentada por Stute [20], onde se estima un modelo paramétrico de regresión con variable resposta censurada pola dereita. Polo tanto, antes de continuar resumiremos a proposta de Stute e veremos como se relaciona co noso método.

Sobre uns supostos similares aos da nosa proposta, parte dunha mostra aleatoria independente e identicamente distribuída de tamaño n (X_i, Y_i) nun espazo Euclídeo d+1 dimensional definido nun espazo probabilístico (Ω ,A,$\mathbb{P}$). Considerando as correspondentes covariables aleatorias X e Y asociadas, θ_0 un vector de parámetros p dimensional descoñecido, a relación entre as dúas variables virá dada por unha función descoñecida f que verificará

$$Y = f\left(X, \theta_0\right) + \varepsilon$$

con

$$\mathbb{E}\left(\varepsilon | X\right) = 0.$$

Ademais, contempla escenarios onde a mostra sexa heterocedástica.

Sobre estes supostos Stute incorpora o concepto de censura mediante a introdución, por un lado da variable censura C, e por outro de $Z = \text{mín}\left(Y, C\right)$ e $\delta = 1_{\{Y \leq C\}}$. A idea subxacente será a de realizar unha estimación non paramétrica da función de distribución, partindo do estimador de Kaplan-Meier [8]. Co anterior, aborda unha cuestión importante, a da estimación do vector de pesos, integrado por:

$$W_{in} = \frac{\delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_{[j:n]}}$$

e sendo os valores $\delta_{[i:n]}$ os indicadores asociados a $Z_{i:n}$, e a súa vez $Z_{1:n} \leq Z_{i:n} \ldots \leq Z_{n:n}$ son os valores de Z ordenados. Aquí, a idea máis importante, que que influirá na nosa proposta, será a do estimador de mínimos cadrados ponderados, que será aquel que minimice

$$S_n\left(\theta\right) = \sum_{i=1}^{n} W_{in} \left[Z_{i:n} - f\left(X_{[i:n],}\theta\right)\right]^2.$$

Unha mención especial respecto da comparación dos modelos é que, mentres Stute realiza a estimación do modelo mediante mínimos cadrados, no noso caso será mediante a función de perda cuantílica. A pesar das diferenzas, como a forma dos pesos ou da función de distribución para variable censurada, o feito de incorporar os pesos e a función de distribución ao modelo de regresión serán dúas cuestións similares, que non idénticas, en ambos modelos.

Tras este paréntese, continuaremos co último paso da construción do modelo. Seguindo un razoamento análogo ao emprendido por Stute, concluímos que o método de estimación da regresión cuantil con datos censurados por intervalos emprega un estatístico baseado nos pesos e na función de perda cuantílica. Ese estimador, para o caso da regresión lineal simple, será o seguinte:

$$\min_{\beta} \sum_{i=1}^{n} \sum_{j=1}^{m} \rho_\tau \left(c_{ij} - \beta_0 - \beta_1 x_i\right) \mu_{ij}$$

Se pola contra estivésemos interesados no caso múltiple, bastaría con transformar o estimador anterior en:

$$\min_{\beta} \sum_{i=1}^{n} \sum_{j=1}^{m} \rho_\tau \left(c_{ij} - x_i^{'} \beta \right) \mu_{ij}$$

Naturalmente, estamos nun caso de regresión múltiple, e aínda que existen un pesos asociados e un sumatorio dobre, o desenvolvemento teórico deriva da regresión cuantil desenvolta na sección segunda.

Co anterior finaliza o desenvolvemento teórico da regresión cuantil con datos censurados.

3.3. Método de estimación da regresión cuantil con datos censurados en R.

A terceira e última cuestión que se abordará neste capítulo será a obtención mediante R dun algoritmo que execute o método proposto na sección anterior. Tratándose dunha proposta novidosa é lóxico que non existise software dispoñible para a implementación do método, polo que veremos como podemos combinar o software dispoñible para acadar o estimador .

Diferentes paquetes de R proporcionan suficientes recursos para a construción do algoritmo. Recordemos que o método incorpora por un lado o algoritmo de Turnbull para a obtención dun estimador non paramétrico da función de distribución dunha variable censurada por intervalos, e por outro lado a regresión lineal cuantil ponderada para a estimación da recta de regresión respecto da variable censurada e dunha covariable. Pois ben, de modo independente, existe software dispoñible que pode ser empregado.

En primeiro lugar, o algoritmo de Turnbull está dispoñible na librería interval de R. O autor Fay [6] parte de tres librerías de R: a librería Survival, a librería perm e a librería Icens. Esta última, incorpora o estimador non paramétrico de máxima verosimilitude para variables aleatorias censuradas; estimador a partir do cal obtemos os pesos con que se ponderará a regresión cuantil.

En segundo lugar, a librería quantreg, creada entre outros autores polo xa citado Roger Koenker [10], representa a mellor librería de R para abordar a cuestión da regresión cuantil, incluíndo tanto modelos non lineais como lineais, incluíndo a estimación ponderada mediante pesos. Loxicamente, estaremos neste último caso.

Cunha combinación dos dous paquetes anteriores incluíndo unha serie de adaptacións, é posible executar o noso método. Para iso, incluiremos a continuación dous exemplos que ilustrarán o funcionamento do código. Ademais, observarase con que facilidade pode adaptarse o código a problemas diferentes, xa que as modificacións serán mínimas e afectarán exclusivamente aos datos introducidos e sobre os eixos dos gráficos.

Exemplo I. Estudo sobre VIH

Ao longo do exemplo, e co fin de expresar o funcionamento do código, combinaremos explicacións sobre os datos ou o modelo co código oportuno que realiza as accións descritas nas explicacións.

O primeiro exemplo corresponde cunha mostra de observación de 297 pacientes daneses con VIH positivo, pero que xa non manifestan síntomas da enfermidade. Tras un período determinado de tempo que varía segundo os pacientes, algúns deles experimentan o retorno da enfermidade. [3] [15]

O estudo comeza en 1983 e esténdese ata 1989, incluíndo información sobre a data de entrada no estudo, a última data observada onde o paciente non manifesta síntomas, a primeira data onde o paciente manifesta síntomas, e outros indicadores.

O noso obxectivo será, empregando regresión cuantil, establecer se existe algunha relación entre o regreso da enfermidade e a idade do paciente. Intuitivamente, é de supoñer que pacientes

con maior idade poidan experimentar unha maior sensibilidade ao retorno da enfermidade, pero compre facer as comprobacións oportunas.

O código completo asociado está incluído no apéndice C. En primeiro lugar, será necesario introducir os datos do estudo, incluídos na librería Epi de R. Para iso, introducimos tanto a librería interval como a librería Epi:

```
library(interval)
library(Epi)
```

A continuación, será necesario chamar aos datos do estudo e editalos de tal forma que se presenten na forma buscada de datos censurados. Para iso, o primeiro que se realizará será fixar os datos asociados ao conxunto de datos *hivDK*, tras o cal fixaremos os extremos superiores e inferiores dos intervalos censurados.

Na primeira fase, por un lado fixarase os extremos inferiores como a diferenza entre a última observación do paciente onde non manifestaba síntomas e a primeira observación realizada do paciente. Doutra banda, establecerase os extremos superiores como a diferenza entre a primeira observación do paciente con síntomas do retorno da enfermidade e a primeira observación. Loxicamente, a diferenza de datas exprésase en días.

Modificarase en ámbolos dous casos anteriores os datos que falten. No caso dos extremos inferiores, que falte un dato significa que existe censura pola esquerda, e será renomeado como 0. Nos extremos superiores, que falte un dato implica censura pola dereita, e renomearase ese datos como Inf, en referencia ao infinito.

```
data(hivDK)
data(hivDK)
d<-hivDK
left<-as.numeric(d$well-d$entry)
left.na<-left
left[is.na(left)]<-0
right<- as.numeric(d$ill - d$entry)
right.na<-right
right[is.na(right)]<-Inf
```

En segundo lugar, construirase a matriz de datos incorporando á información xa obtida a idade de cada paciente participante no estudo. A variable bth dentro de hivDK calcula a desviación de cada paciente respecto dos 30 anos, co que resulta doado modificala para obter a idade de cada paciente. Tamén calcularemos os puntos medios de cada intervalo censurado que, se ben é certo que non ten ningunha utilidade nos cálculos, serviranos para representar os intervalos graficamente.

```
d<-data.frame(left,left.na,right,right.na,us=d$us,year.of.birth=d$bth+1950,age.at.entry
=d$bth+30,partners.per.year=d$pyr) #matriz de datos
y<-as.numeric((d$right+d$left)/2)
y.na<-y
y[is.infinite(y)]<-3057
y #puntos medios dos intervalos censurados
x<-d$age.at.entry #Variable explicativa
```

A partir da matriz de datos, resulta doado calcular os valores extremos dos intervalos de confianza.

```
L<-d[,"left"] #extremos inferiores dos intervalos censurados
R<-d[,"right"] #extremos superiores dos intervalos censurados
```

A segunda fase consistirá en calcular os intervalos de Turnbull e a probabilidade asociada a cada un deles. A partires destas probabilidades obterase posteriormente a probabilidade de que cada observación estea nun intervalo. A tal efecto, empregarase a función icfit sobre os extremos inferiores e superiores dos intervalos censurados.

Incorporaremos a continuación os vectores que representen os extremos superiores, inferiores e as probabilidades de cada intervalo. A figura 3.1. consiste na representación gráfica da función de distribución da variable censurada.

```
est=icfit(L,R)
u=est$intmap[1,]; u #extremos inferiores dos intervalos de Turnbull
v=est$intmap[2,]; v #extremos superiores dos intervalos de Turnbull
pf=est$pf; pf #probabilidade de cada intervalo de Turnbull
plot(icfit(L,R))
```

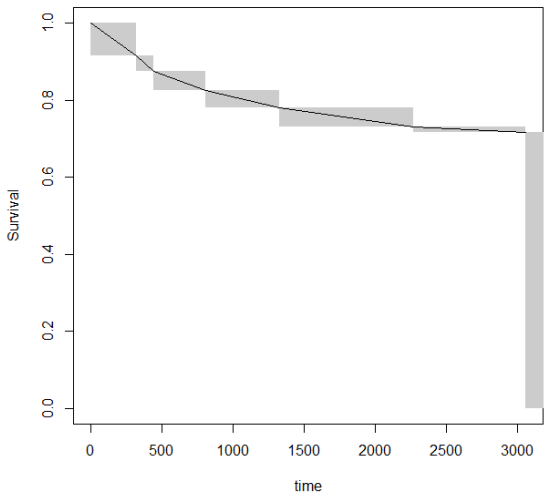


Figura 3.1: Función de distribución

A terceira fase consistirá na asociación dos intervalos de Turnbull a cada intervalo censurado. Para iso fixarase o número de intervalos de Turnbull e de individuos da mostra e obtense, por un lado, o índice do extremo inferior máis próximo pola esquerda ao extremo inferior do intervalo censurado, e por outro, o índice do extremo superior máis próximo pola dereita ao extremo superior do intervalo censurado.

```
p=length(u)  # Número de intervalos de Turnbull
n=nrow(d) # Número de individuos da mostra
# Extremos inferiores
z=c(L,u)
zord=sort(z,index.return=T)
```

```
ind=zord$ix
a=c()
a0=1 # Indice nos intervalos de Turnbull
for (iz in 1:(n+p)){if (ind[iz]<=n){a[ind[iz]]=a0}else{a0=a0+1}}
# Extremos superiores
z=c(v,R)
zord=sort(z,index.return=T)
ind=zord$ix
b=c()
b0=p # Indice nos intervalos de Turnbull
for (iz in (n+p):1){if (ind[iz]>p){b[ind[iz]]=b0}else{b0=b0-1}}
b=b[(p+1):(p+n)]
```

Para a cuarta fase será preciso obter tres vectores de datos. O primeiro vector incluirá o conxunto de puntos medios dos intervalos de Turnbull asociados a cada intervalo censurado, tendo en conta que cada intervalo censurado inclúe varios intervalos de Turnbull. O segundo vector presentará, para cada elemento do vector anterior, o valor da idade asociada. Terceiro e último, para cada un dos valores do primeiro vector inclúese o pesos que terá asociado. Chamaranse ynew, xnew e pesos respectivamente. Respecto dos puntos medios do primeiro vector, recordemos que para valores censurados pola dereita asociaremos valor do extremo inferior asociado.

```
inew=0
xnew=c()
ynew=c()
pesos=c()
for (i in 1:n){
pt=sum(pf[a[i]:b[i]])
for (ip in a[i]:b[i]){
inew=inew+1
xnew[inew]=x[i]
ynew[inew]=(v[ip]+u[ip])/2
pesos[inew]=pf[ip]/pt }}
ynew<-as.numeric(ynew)
ynew.na<-ynew
ynew[is.infinite(ynew)]<-3057
```

A quinta e última fase incorpora a regresión lineal cuantil ao conxunto de datos. O tratamento dos datos, tanto da variable non censurada como da variable censurada foi realizado na fase cuarta. Para proceder ca regresión, será necesario cargar a librería quantreg previamente. O vector *xnew* será tomado como variable explicativa, o vector *ynew* como variable resposta, e o vector *pesos* como as ponderacións de cada observación. A función rq será encargada de realizar a regresión, sendo *tau* o valor da orde do cuantil.

```
library(quantreg)
m=rq(ynew~xnew,tau=0.2,weights=pesos)
```

Finalmente, representaremos graficamente os intervalos censurados fronte a variable explicativa, e a recta de regresión. Para facilitar a interpretación, a recta de regresión representase en negro, os intervalos que presenten censura pola esquerda en azul, os intervalos censurados pola dereita en amarelo, e o resto de intervalos en vermello.

A Figura 3.2 representa o gráfico cando o axuste da regresión correspóndese a tau=0.3. Para valores de tau comprendidos entre 0.3 e 0.8 o gráfico non presenta variacións perceptibles.

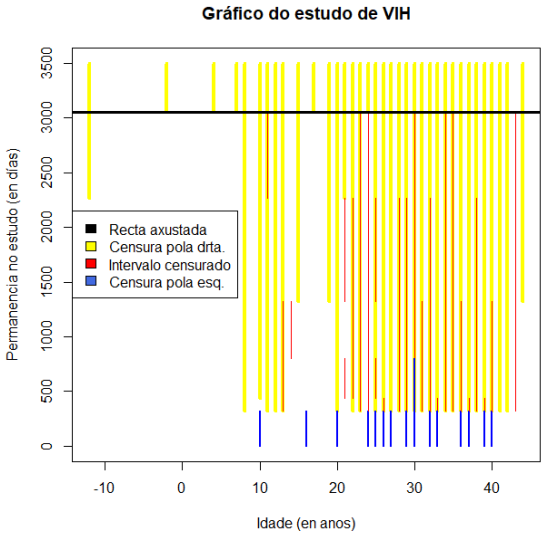


Figura 3.2: Intervalos censurados (permanencia no estudo) fronte á idade. Axuste da recta de regresión (en negro) para τ de 0.3 a 0.9

O gran cambio prodúcese para valores inferiores a tau=0.3. Na figura 5 e 6 incluímos o cambio producido cando tau pasa a ter un valor de 0.2 e 0.1 respectivamente.

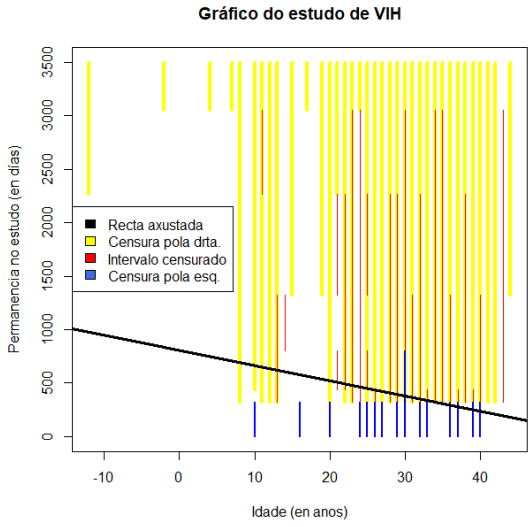


Figura 3.3: Intervalos censurados (permanencia no estudo) fronte á idade. Axuste da recta de regresión (en negro) para τ=0.1.

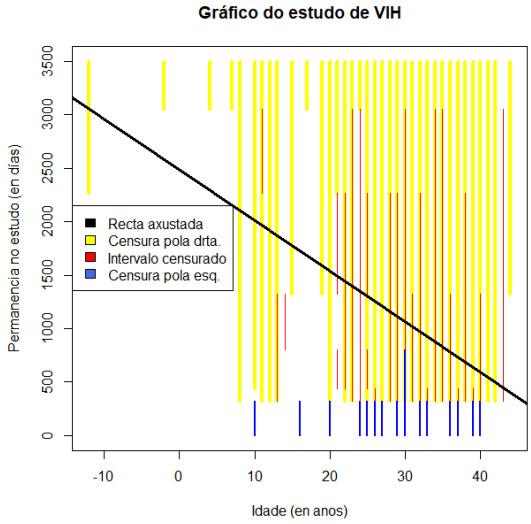


Figura 3.4: Intervalos censurados (permanencia no estudo) fronte á idade. Axuste da recta de regresión (en negro) para τ=0.2.

A pregunta que xorde inmediatamente é a que se debe esa estabilidade para valores de tau comprendidos entre 0.3 e 0.9, e por que para valores inferiores a 0.3 a pendente varía tanto.

Recordemos que os valores de tau penalizan determinado tipo de observacións. Cuantis de orde superior a 0.5 penalizanse a través da función de perda cuantílica máis ás observacións á dereita da posición central . Pola contra, para cuantis de orde baixa esa penalización invértese e pasa a penalizar máis ás observacións á esquerda da posición central. Por iso, ao existir un gran número de persoas que sobreviven ao estudo, será preciso que o cuantil tome un valor baixo para obter conclusións sobre os pacientes que non sobreviven ao estudo.

A conclusión que podemos obter do estudo é que, aínda que un gran número de pacientes sobreviven ao estudo, os que non sobreviven vense afectados pola idade. Á vista da recta de regresión, a medida que aumenta a idade diminúe a probabilidade de sobrevivir ao estudo sen enfermidade.

Capítulo 4

Simulación

Tras proporcionar unha serie de resultados teóricos e de implementar un novo método de estimación da regresión con datos censurados por intervalos, o novo obxectivo consistirá no estudo das propiedades do estimador, para o cal será necesario empregar técnicas de simulación para a xeración de mostras aleatorias.

Consideraremos catro factores que provocan variacións no estimador. En primeiro lugar, o valor da orde do cuantil, comprendida entre 0 e 1, inflúe no axuste da recta de regresión do modelo, e polo tanto no valor dos coeficientes de cada recta de regresión asociada a cada mostra. En segundo lugar, o tamaño de cada mostra simulada inflúe tamén na estimación da recta de regresión. En terceiro lugar, o mecanismo de censura. En cuarto lugar, o número de simulacións condicionará dun xeito directo a calidade da aproximación por Montecarlo.

O código que será presentado nesta sección ten en conta estes catro factores para construír a simulación. Permite fixar o valor do cuantil, o tamaño da mostra, o número de observacións e o número de simulacións, e realizar as simulacións tendo en conta os valores asignados.

Dividiremos en tres os apartados deste capítulo. No primeiro apartado describiremos o funcionamento do código, explicando que funcións se realizan para cada segmento do código. Para o segundo apartado variaremos o tamaño mostral, o número de simulacións, a densidade de observacións, o valor do cuantil, e veremos como incide no sesgo, na varianza e no erro cadrático medio do estimador. No terceiro e último apartado, faremos unha comparación con outro método de regresión con datos censurados, co fin de valorar a calidade do método proposto.

4.1. Funcionamento do algoritmo da simulación

O proceso de simulación do modelo consistirá na construción de mostras aleatorias dunhas determinadas características que permitan corroborar a boa aproximación do método de regresión cuantil con datos censurados por intervalos definido neste traballo. Loxicamente, empregarase o código relativo ao método proposto que xa foi debidamente explicado no capítulo anterior. Centrarémonos en explicar como se xeran as mostras aleatorias con regresión e censura por intervalos, para o cal combinaremos teoría e implementación en R.

Ante todo, para explicar o proceso de simulación consideraremos uns tamaños fixos de mostra, de número de simulacións ou de orde do cuantil. Na sección posterior veremos como evoluciona o erro cadrático medio a medida que varían estes factores. Procedamos pois ca descrición do proceso de simulación.

Inclúese o código completo da simulación no apéndice D. Comezarase cargando as librerías interval e quantreg necesarias para a execución do método de regresión cuantil en intervalos censurados.

```
library(interval)
library(quantreg)
```

En primeiro lugar, introduciremos os valores numéricos que sirvan de referencia para a xeración de mostras simuladas. En concreto, "n" representará o tamaño mostral, "$nobs$" consistirá no número de observacións asociadas a cada individuo que se empregarán para xerar os extremos dos intervalos censurados, e "ns" o número de mostras simuladas. Tamén se fixará unha semente para iniciar o xerador dos números aleatorios en 123456. A continuación podemos ver no código uns exemplos concretos de posibles valores de "n" ," $nobs$" e " ns" .

```
set.seed(123456)
n=100
nobs=10
ns=1000
```

A continuación, iniciaremos o bucle de mostras simuladas que se reproducirá tantas veces como indicamos antes. Fixaremos previamente dous vectores, que se completarán con cada execución do algoritmo, e que inclúen o valor de β_0 (no código representado por vbeta0) e de β_1 (representado no código por vbeta1). Cada simulación gardará os valores de β_0 e de β_1 estimados para a mostra correspondente. Ademais, incluímos unha matriz nula con tantas filas como o tamaño da mostra, e tantas columnas como o número de observacións que empregaremos para a censura.

```
vbeta0=c()
vbeta1=c()
obs=matrix(0,nrow=n,ncol=nobs)
for (is in 1:ns){  # Inicio do bucle das mostras simuladas
```

Describiremos como se xerará cada unha das mostras dentro dese bucle. O primeiro que debemos ter en conta e que a censura establecerase a partir de valores coñecidos e obtidos mediante simulación, polo que coñecemos cales son os datos reais que están censurados por intervalos. Noutras palabras, pola forma da construír a simulación sabemos cales serán os valores reais da variable censurada.

Construirase unha mostra de tamaño n seguindo unha distribución U(0,1) para definir a variable aleatoria explicativa, isto é

$$\{x_i\}_{i=1}^{n} \sim U\left(0,1\right).$$

Por outro lado, a variable aleatoria resposta Y seguirá unha relación lineal cos valores da variable X con intercepto 1 e pendente 3 pero considerando un erro en cada observación, isto é

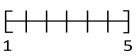
$$y_i = 1 + 3 \cdot x_i + \varepsilon_i$$

con $\varepsilon_i \sim U\left(0,1\right)$ e $1 \leq$ i$\leq$ n. Dito erro, como foi presentado en seccións previas, pode non ser normal, e nesta simulación de feito non o será.

```
x=runif(n)
y=1+3*x+runif(n)
```

Os valores anteriores "x" e "y" representan as variables explicativa e resposta antes mencionadas. Como adiantamos anteriormente, a partires dos valores da variable Y construiremos intervalos censurados para cada dato. Para iso, será necesario xerar un determinado número de valores aleatorios asociados á variable Y, de tal xeito que o valor da variable aleatoria estea limitado por algún deses valores. O número de valores aleatorios asociados a cada variable é un valor fixo que introducimos ao comezo da sección, que recibía o nome de "*nobs*" .

Para cada valor y_i obteremos unha serie de valores crecentes $(C_{i1}, C_{i2}, \ldots, C_{inobs})$. A notación anterior resulta similar á presentada por Rabinowitz [18] e Lawless [13] para definir a censura. Cada un destes valores xerados obtéñense a partir dunha función lineal crecente con pendente constante, intercepto 1, e que incorpora un erro aleatorio distribuído mediante unha uniforme pero ponderada mediante un coeficiente comprendido entre 0 e 1.



Trátase de dividir o intervalo [1,5] en *"nobs"* trozos e tomamos un número ao azar de maneira uniforme en cada subintervalo.

En concreto,

$$C_{ij} = 1 + \frac{4}{nobs} \cdot (j-1) + \frac{4}{nobs} \cdot u_{ij}$$

con $u_{ij} \sim U(0,1)$ e 1≤ j≤ nobs. A forma de xerar estes elementos é a seguinte:

```
for (iobs in 1:nobs){obs[,iobs]=1+4*(iobs-1)/nobs+4*runif(n)/nobs}
```

A medida que aumenta o valor do índice j, aumenta o valor de C_{ij}, o cal é coherente co obxectivo proposto. Pola forma en que están definidos e acoutados os valores de C_{ij}, o valor real de y_i estará limitado por algún dos valores xerados por este procedemento. L_i e R_i serán dous valores consecutivos de C_{ij} tales que y_i está condido entre eles.

Se o valor de "*pobs*" é cero, entón defínese o intervalo como censurado pola esquerda; se o valor de "*pobs*" é igual ao valor de "*nobs*" , entón o intervalo será censurado pola dereita; e finalmente, se o valor de "*pobs*" é distinto dos dous casos anteriores, tratarase dun intervalo censurado cuxos extremos inferior e superior serán respectivamente os C_{ij} con valores máis próximos inferior e superior ao valor y_i.

```
L=c()
R=c()
for (i in 1:n){
pobs=sum(obs[i,]<y[i])
if (pobs==0){L[i]=0;R[i]=obs[i,1]}else{
if (pobs==nobs){L[i]=obs[i,nobs];R[i]=5} else{
L[i]=obs[i,pobs];R[i]=obs[i,pobs+1] }
}}
```

Con isto remata o proceso de construción de dúas variables aleatorias X e Y, e de intervalos censurados para cada valor y_i da variable Y. A continuación prodúcese a implementación do método de regresión cuantil con datos censurados por intervalos, cuxo funcionamento foi descrito detalladamente no capítulo anterior.

```
est=icfit(L,R)
u=est$intmap[1,]
v=est$intmap[2,]
pf=est$pf
p=length(u)
z=c(L,u)
zord=sort(z,index.return=T)
ind=zord$ix
a=c()
a0=1 # Indice nos intervalos de Turnbull
for (iz in 1:(n+p)){
if (ind[iz]<=n){a[ind[iz]]=a0}else{
a0=a0+1}}
z=c(v,R)
zord=sort(z,index.return=T)
ind=zord$ix
b=c()
b0=p # Indice nos intervalos de Turnbull
for (iz in (n+p):1){if (ind[iz]>p){b[ind[iz]]=b0}else{b0=b0-1}}
b=b[(p+1):(p+n)]
inew=0
xnew=c()
ynew=c()
pesos=c()
for (i in 1:n){
pt=sum(pf[a[i]:b[i]])
for (ip in a[i]:b[i]){
inew=inew+1
xnew[inew]=x[i]
ynew[inew]=(v[ip]+u[ip])/2
pesos[inew]=pf[ip]/pt }}
m=rq(ynew~xnew,weights=pesos)
```

Ata aquí, executouse o método proposto neste traballo. A modo de resumo, a partires dos valores dos intervalos censurados construíronse os intervalos de Turnbull, e obtívose unha probabilidade asociada a cada observación de pertencer a cada intervalo de Turnbull. Esas probabilidades empregáronse como pesos para ponderar unha regresión cuantil (cun orde de cuantil de 0.5) tomando como valor da variable resposta o punto medio de cada intervalo de Turnbull, a excepción da censura pola dereita cuxos valores consistiron nos extremos inferiores.

Como resultado do proceso anterior, obtense unhas estimacións do intercepto e pendente, que gardaremos no vector vbeta0 e vbeta1 xa mencionados. Ademais, daremos a orde de que mostre en pantalla o valor concreto de β_0 e de β_1 para cada mostra simulada.

Todo o proceso anterior repetirase tantas veces como indique "*ns*" . En cada repetición, simulará unha mostra nova, a partir da cal realizará todo o proceso descrito, que culminará na obtención de dous valores β_0 e β_1 que gardaremos nos correspondentes vectores.

```
beta=coef(m)
vbeta0[is]=beta[1]
vbeta1[is]=beta[2]
cat("Mostra", is, "Beta0", vbeta0[is], "Beta1", vbeta1[is], "\n")
}#Fin do bucle
```

Polo tanto, se vbeta0 inclúe todos os valores para cada mostra simulada do intercepto da recta de regresión, entón podemos calcular a media e a varianza deses valores.

```
mean(vbeta0); var(vbeta0)
```

Para o vector vbeta1, que inclúe as pendentes da recta de regresión asociadas a cada mostra simulada, tamén se pode calcular a media e a varianza.

```
mean(vbeta1); var(vbeta1);
```

Para rematar, o último que realizaremos será o cálculo do sesgo, varianza e erro cadrático medio dos estimadores, tanto do intercepto como da pendente. Tal e como construímos a variable explicativa, verificarase que, para $1 \leq i \leq n$ e orde do cuantil 0.5, $\mathbb{E}(y_i|\ x_i=0)=1+\mathbb{E}(\varepsilon_i)=1.5$, xa que como sabemos, $IE(\varepsilon_i)=0.5$ para cuantil 0.5 ao seguir os erros unha distribución U(0,1). Polo tanto, o erro cadrático medio para o intercepto coincide co cadrado do sesgo máis a varianza.

```
ecm0=(mean(vbeta0)-1.5)^2+var(vbeta0); ecm0
```

Cun razoamento similar, séguese que a pendente teórica tal e como se definen as mostras será 3. O resto do razoamento resulta análogo ao do intercepto.

```
ecm1=(mean(vbeta1)-3)^2+var(vbeta1); ecm1
```

4.2. Resultados da simulación.

Habendo abordado na sección anterior a descrición do funcionamento do algoritmo de simulación, nesta sección describirase os resultados da simulación para diferentes valores dos parámetros. O obxectivo será analizar como varía o sesgo, varianza e erro cadrático medio dos estimadores en función da orde do cuantil, do número de simulacións, do número de observacións e do tamaño mostral.

Efecto da orde do cuantil

A continuación figura de forma tabulada os resultados de implementar en R a simulación tendo en conta os diferentes cuantís. Consideremos un caso onde o número de observacións para os intervalos censurados son 10, o tamaño da mostra 100 e o número de simulacións 1000. Vexamos como varían os valores aproximados ao variar a orde do cuantil.

Orde	Valor real intercepto	Sesgo intercepto	Sesgo pendente	Varianza intercepto	Varianza pendente	ECM intercepto	ECM pendente
0.1	1.1	0.0186	-0.1166	0.00818	0.0257	0.1535	0.0393
0.2	1.2	0.0379	-0.0878	0.00765	0.0234	0.0763	0.0311
0.3	1.3	0.0389	-0.0739	0.00791	0.0242	0.0338	0.0296
0.4	1.4	0.0340	-0.0639	0.00827	0.0243	0.0126	0.0284
0.5	1.5	0.0288	-0.0586	0.00845	0.0240	0.0092	0.0275
0.6	1.6	0.0293	-0.0651	0.00877	0.0243	0.0255	0.0285
0.7	1.7	0.03475	-0.0759	0.00828	0.0236	0.0633	0.0294
0.8	1.8	0.05411	-0.0976	0.00764	0.0223	0.1330	0.0318
0.9	1.9	0.10372	-0.1282	0.00792	0.0227	0.2616	0.0392

Táboa 1. Sesgo, varianza e erro cadrático medio (ECM) dos estimadores do intercepto e da pendente para distintas ordes do cuantil

Parece lóxico que, para este número de observacións e de tamaños mostrais, o cuantil de orde 0.5 sexa o que mellores resultados aporte. Presenta un menor erro cadrático tanto para o intercepto como para a pendente, os sesgos máis próximos a cero e a menor varianza.

Efecto do número de simulacións

Tomando a orde do cuantil 0.5 como referencia, e mantendo o mesmo número de observacións dos intervalos censurados (10) e o mesmo tamaño mostral (100), veremos como varían os resultados variando o número de simulacións:

Número de simulacións	Sesgo intercepto	Sesgo pendente	Varianza intercepto	Varianza pendente	ECM intercepto	ECM pendente
100	0.0179	-0.0459	0.00871	0.0238	0.00903	0.0259
1000	0.0288	-0.0586	0.00845	0.0240	0.00928	0.0275
10000	0.0318	-0.0632	0.00821	0.0244	0.00821	0.0244
100000	0.0308	-0.0620	0,00805	0,0245	0.00905	0.0284

Táboa 2. Sesgo, varianza e erro cadrático medio (ECM) dos estimadores do intercepto e da pendente para distinto número de simulacións.

Veremos que as aproximacións do sesgo, varianza e erro cadrático medio son semellantes cando se realiza alomenos 1000 mostras. Ao pasar de 100 a 1000 simulacións prodúcese unha variación de aproximadamente unha centésima na pendente e no intercepto (0.010861 e 0.012723 respectivamente). Pola contra, pasar de 10000 a 100000 simulacións a variación é de aproximadamente unha milésima (0.000984 para a intercepto e 0.001277 para a pendente).

Efecto da densidade de observacións

A continuación analizaremos como varían os sesgos, as varianzas e os erros cadráticos medios para diferentes valores da densidade de observacións. A mostra terá tamaño 100, o número de simulacións será 1000, a orde do cuantil 0.5, e variaremos a densidade de observacións (variable *"nobs"*).

Tal como foi deseñada a estimación, a medida que aumenta a densidade de observacións o tamaño dos intervalos censurados diminuirá, co que a perda de información será menor e o axuste dos datos será mellor. A táboa 3 confirma as nosas sospeitas oscilando en torno a cero os sesgos para o intercepto e a pendente a medida que aumenta a densidade de observacións, e unha estabilización dos erros cadráticos medios entorno a valores inferiores a unha centésima (para o intercepto) e tres centésimas (para a pendente).

Obs.	Sesgo do intercepto	Sesgo da pendente	Varianza intercepto	Varianza pendente	ECM intercepto	ECM pendente
10	0.0288	-0.058665	0.00845	0.0240	0.00928	0.0275
20	0.0105	-0.020390	0.00898	0.0274	0.00909	0.0278
30	0.0048	-0.008290	0.00942	0.0276	0.00944	0.0276
40	0.0015	-0.004888	0.00944	0.0289	0.00944	0.0290
50	-0.0016	-0.002178	0.00998	0.0295	0.00999	0.0295
60	-0.00003	0.001512	0.00908	0.0273	0.00908	0.0273
70	-0.0026	0.005801	0.01007	0.0304	0.01008	0.0305
80	-0.0018	0.002306	0.01034	0.0297	0.01035	0.0297
90	0.0064	-0.006188	0.01026	0.0315	0.01030	0.0315
100	-0.0054	0.011552	0.00946	0.0297	0.00949	0.0299

Táboa 3. Sesgo, varianza e erro cadrático medio (ECM) dos estimadores do intercepto e da pendente para distintas densidades de observacións.

Efecto do tamaño mostral

Prosigamos agora co estudo do efecto do tamaño mostral. Deixaremos fixos os valores para a orde do cuantil (0.5), do número de simulacións (1000), da densidade de observacións (10), e variaremos o tamaño mostral. Os resultados están recollidos na táboa 4.

A medida que aumenta o tamaño mostral redúcese o valor absoluto do sesgo do intercepto e da pendente, achegándose cada vez máis a cero. Tamén, aumentando o valor do tamaño mostral redúcese os valores das varianzas e dos erros cadráticos medios. Trátase polo tanto dun estimador consistente dos coeficientes de regresión.

Tamaño mostral	Sesgo do intercepto	Sesgo da pendente	Varianza intercepto	Varianza pendente	ECM intercepto	ECM pendente
50	0.0366	-0.0753	0.0149	0.0459	0.0163	0.0516
100	0.0288	-0.0586	0.00845	0.0240	0.0092	0.0275
150	0.0282	-0.0571	0.00610	0.0199	0.0061	0.0199
200	0.0284	-0.0551	0.00389	0.0117	0.0046	0.0147
250	0.0259	-0.0517	0.00335	0.0100	0.0040	0.0127
300	0.0256	-0.0511	0.00242	0.0076	0.0030	0.0102

Táboa 4. Sesgo, varianza e erro cadrático medio (ECM) dos estimadores do intercepto e da pendente para distintas variacións do tamaño mostral

A conclusión é, polo tanto, que a simulación corrobora o bo funcionamento do método á vista dos valores dos sesgos, das varianzas e dos erros cadráticos medios. A súa aplicación pode resultar de gran utilidade en contextos onde se desexa aplicar unha regresión lineal, pero unha das variables presenta censura por intervalos. Ademais, esa utilidade ven reforzada polo feito de empregar a función de perda cuantílica que amplía a súa aplicación a casos onde os erros non presentan normalidade.

4.3. Comparación cun estimador paramétrico

Nesta última sección procederemos a realizar unha comparativa do modelo de regresión cuantil con datos censurados por intervalos proposto neste traballo respecto doutro modelo. Na obra de Shang e Sun [25] cítase un dos modelos máis empregados para axustar regresión con datos censurados por intervalos, o modelo paramétrico exponencial, onde se supón que os erros de regresión seguen unha distribución exponencial. Compre facer un breve apunte teórico [1] sobre como se estima a log-verosimilitude no contexto de modelos paramétricos. Tomaremos unha mostra aleatoria de tamaño n formada polos valores y_i con $1 \leq i \leq n$ que se atopa nunha situación de censura por intervalos, co que para cada un destes valores asociaremos dous valores L_i e R_i, respectivamente, extremos inferior e superior de cada intervalo censurado. Se representamos S_0 a función de supervivencia base, x_i cada valor mostral da variable explicativa, a log-verosimilitude consistirá en

$$\sum_{i=1}^{n} \log \left(S_0 \left(L_{i-} \right)^{e^{x_i \beta}} - S_0 \left(R_i \right)^{e^{x_i \beta}} \right).$$

No artigo de Anderson-Bergman [1] resúmense algúns dos paquetes e comandos máis empregados en R nos diferentes modelos con datos censurados. Para modelos paramétricos, o paquete icenReg proporciona o comando ic_par *(interval censored parametric regression)*, que permite executar a regresión con datos censurados para o modelo exponencial, cun axuste mediante un modelo AFT (tempo de vida acelerado).

Considerarase o mecanismo de simulación da sección anterior e incorporarase o modelo paramétrico exponencial co obxectivo de realizar unha comparativa co modelo proposto neste documento. O apéndice E inclúe o código empregado para a comparativa.

Fixaremos un valor 10 de densidade de observacións para a construción dos intervalos censurados, 1000 simulacións e unha orde de cuantil 0.5. Aumentaremos o tamaño mostral, "n" , e compararemos os valores do sesgo, varianza e erro cadrático medio nos dous modelos.

	Modelo regresión cuantil			Modelo paramétrico exponencial		
n	Sesgo	Varianza	ECM	Sesgo	Varianza	ECM
100	0.05866	0.02406	0.02750	0.06720	0.03232	0.03684
200	0.05517	0.01174	0.01478	0,06575	0.01445	0.01877
300	0.05115	0.00765	0.01026	0.06334	0.01040	0.01442

Táboa 5. Comparación do sesgo, varianza e erro cadrático medio para a estimación de β_1

Se analizamos os resultados obtidos ca simulación, que están incluídos na táboa 5, vemos que os resultados do modelo de regresión cuantil con datos censurados por intervalos son mellores que o modelo da familia exponencial.

En primeiro lugar, a medida que aumentamos o tamaño mostral, o sesgo, a varianza e o erro cadrático medio diminúen tanto no modelo paramétrico exponencial como no modelo de regresión cuantil. En segundo lugar, os sesgos que obteríamos no modelo de regresión cuantil para tamaños mostrais 100, 200 e 300 sería 0.05866, 0.05517 e 0.05115 respectivamente. Pola contra, no modelo paramétrico exponencial eses valores serían os valores 0.06720, 0.06575 e 0.06334 respectivamente. Polo tanto, comparando sesgo a sesgo para cada tamaño obtense valores inferiores co modelo de regresión cuantil.

En terceiro lugar comparemos a varianza. No modelo de regresión cuantil obtense valores 0.02750, 0.01174 e 0.00765 para tamaños mostrais 100, 200 e 300. Para o modelo paramétrico exponencial os valores da varianza serían 0.03232, 0.01445 e 0.01040 para os tamaños mostrais 100, 200 e 300. Igual que co sesgo, o modelo de regresión cuantil presenta mellor axuste que o modelo de familia exponencial comparando dato a dato.

En cuarto lugar, veremos como varía o erro cadrático medio. Nunha man, para o modelo de regresión cuantil obterase uns valores 0.02750, 0.01478 e 0.01026 para tamaños mostrais 100, 200 e 300 respectivamente. Na outra man, o modelo paramétrico exponencial presenta uns valores 0.03684, 0.01877 e 0.01442. De novo, os valores do erro cadrático medio son menores co modelo de regresión cuantil.

Apéndice A

Resultados do exemplo práctico do algoritmo EM

$\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0.2, 0.2, 0.2, 0.2, 0.2]
Iteración 1

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0.5, 0.5, 0, 0, 0]$
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0.3333, 0.3333, 0.3333, 0]$
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0, 0, 0.5, 0.5]$
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0.1667, 0.2778, 0.1111, 0.2778, 0.1667]
  Iteración 2
- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0.3750, 0.625, 0, 0, 0]$
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0.4167, 0{,}1667, 0.4167, 0, 0]$
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0.625, 0.3750, 0, 0]$
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [0.1250, 0.3472, 0.0556, 0.3472, 0.1250]$
  Iteración 3
- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [\ 0{,}2647, 0.7353, 0, 0, 0, 0\ ]$
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0.463, 0.0741, 0.4630, 0]$
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0, 0, 0.7353, 0.2647\ ]$
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [0.0882, 0.3994, 0.0247, 0.39940, 0.0882]$
  Iteración 4
- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0.1809, 0.8191, 0, 0, 0\ ]$
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [\ 0, 0.4850, 0.0300, 0.4850, 0\ ]$
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = 0, 0, 0, 0.8191, 0.1809]$
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [0.0603, 0.4347, 0.01, 0.4347, 0.0603]$
  Iteración 5

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0.1218, 0{,}8782, 0, 0, 0]$
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0.4943, 0.0114, 0.4943, 0]$
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0, 0, 0.8782, 0.1218]$

- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [0.0406, 0.4575, 0.0038, 0.4575, 0.0406]$
  Iteración 6

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0.0815, 0.9185, 0, 0, 0]$
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0.4979, 0.0041, 0.4979, 0]$
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0, 0, 0.9185, 0.0815]$

- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [\ 0.0272, 0.4721, 0.0014, 0.4721, 0.0272]$
  Iteración 7

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0.0544, 0{,}9456, 0, 0, 0]$
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0.4993, 0.0015, 0.4993, 0]$
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0, 0, 0.9456, 0.0544]$

- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [\ 0.0181, 0.4816, 0.0005, 0.4816, 0.0181]$
  Iteración 8

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0.0363, 0.9637, 0, 0, 0]$
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0.4997, 0.0005, 0.4997]$
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0, 0, 0.9637, 0.0363\ ]$

- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [0.0121, 0.4878, 0.0002, 0.4878, 0.0121]$
  Iteración 9

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0.0242, 0.9758, 0, 0\ ,0]$
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0.4999, 0.0002, 0.4999, 0]$
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0, 0, 0.9758, 0.0242]$

- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [0, 0.0081, 0.4919, 0.0001, 0{,}4919, 0.0081]$
  Iteración 10

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0.0161, 0.9839, 0, 0, 0]$
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0.5000, 0.0001, 0.5, 0]$
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0, 0, 0.9839, 0.0161]$

- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [0.0054, 0.4946, 0, 0.4946, 0.0054]$
  Iteración 11

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0.0108, 0.9892, 0, 0, 0]$
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0.5, 0, 0.5, 0]$
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0, 0, 0.9892, 0.0108]$

- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0.0036, 0.4964, 0.0000, 0.4964, 0.0036]
  Iteración 12
- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] =$ [0.0072, 0.9928, 0, 0, 0]
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] =$ [0, 0.5, 0, 0.5, 0]
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] =$ [0, 0, 0, 0.9928, 0.0072]
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0.0024, 0.4976, 0.0000, 0.4976, 0.0024]
  Iteración 13
- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] =$ [0.0048, 0.9952, 0, 0, 0]
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] =$ [0, 0.5, 0, 0.5, 0]
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] =$ [0, 0, 0, 0.9952, 0.0048]
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0.0016, 0.4984, 0.0000, 0.4984, 0.0016]
  Iteración 14
- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] =$ [0.0032, 0.9968, 0, 0, 0]
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] =$ [0, 0.5, 0, 0.5, 0]
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] =$ [0, 0, 0, 0.9968, 0.0032]
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0.0011, 0.4989, 0.0000, 0.4989, 0.0011]
  Iteración 15
- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] =$ [0.0021, 0.9979, 0, 0, 0]
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] =$ [0, 0.5, 0, 0.5, 0]
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] =$ [0, 0, 0, 0.9979, 0.0021]
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0.0007, 0.4993, 0.0000, 0.4993, 0.0007]
  Iteración 16
- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] =$ [0.0014, 0.9986, 0, 0, 0]
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] =$ [0, 0.5, 0, 0.5, 0]
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] =$ [0, 0, 0, 0.9986, 0.0014]
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0.0005, 0.4995, 0.0000, 0.4995, 0.0005]
  Iteración 17
- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] =$ [0.0009, 0.9991, 0, 0, 0]
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] =$ [0, 0.5, 0, 0.5, 0]
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] =$ [0, 0, 0, 0.9991, 0.0009]
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0.0003, 0.4997, 0.0000, 0.4997, 0.0003]
  Iteración 18
- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] =$ [0.0006, 0.9994, 0, 0, 0]
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] =$ [0, 0.5, 0, 0.5, 0]

$\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] =$ [0, 0, 0, 0.9994, 0.0006]

- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0.0002, 0.4998, 0, 0.4998, 0.0002]

Iteración 19

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] =$ [0.0004, 0.9996, 0, 0, 0]
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] =$ [0, 0.5, 0, 0.5, 0]
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] =$ [0, 0, 0, 0.9996, 0.0004]
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0.0001, 0.4999, 0.0000, 0.4999, 0.0001]

Iteración 20

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] =$ []
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] =$ [0, 0.5, 0, 0.5, 0]
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] =$ []
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ []

Iteración 21

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] =$ [0.0003, 0.9997, 0, 0, 0]
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] =$ [0, 0.5, 0, 0.5, 0]
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] =$ [0, 0, 0, 0.9997, 0.0003]
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0.0001, 0.4999, 0.0000, 0.4999, 0.0001]

Iteración 22

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] =$ [0.0001, 0.9999, 0, 0, 0]
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] =$ [0, 0.5, 0, 0.5, 0]
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] =$ [0, 0, 0, 0.9999, 0.0001]
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0.0000, 0.5000, 0.0000, 0.5000, 0.0000]

Iteración 23

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] =$ [0.0001, 0.9999, 0, 0, 0]
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] =$ [0, 0.5, 0, 0.5, 0]
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] =$ [0, 0, 0, 0.9999, 0.0001]
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0, 0. 5, 0, 0,5, 0]

Iteración 24

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] =$ [0.0001, 0.9999, 0, 0, 0]
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] =$ [0, 0.5, 0, 0.5, 0]
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] =$ [0, 0, 0, 0.9999, 0.0001]
- Paso M:
  $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] =$ [0, 0. 5, 0, 0,5, 0]

Iteración 25

- Paso E:
  $\hat{p}_1 = [\hat{p}_1(t_1), \hat{p}_1(t_2), \hat{p}_1(t_3), \hat{p}_1(t_4), \hat{p}_1(t_5)] = [0, 1, 0, 0, 0]$
  $\hat{p}_2 = [\hat{p}_2(t_1), \hat{p}_2(t_2), \hat{p}_2(t_3), \hat{p}_2(t_4), \hat{p}_2(t_5)] = [0, 0.5, 0, 0.5, 0]$
  $\hat{p}_3 = [\hat{p}_3(t_1), \hat{p}_3(t_2), \hat{p}_3(t_3), \hat{p}_3(t_4), \hat{p}_3(t_5)] = [0, 0, 0, 1, 0]$
- Paso M: $\hat{p} = [\hat{p}(t_1), \hat{p}(t_2), \hat{p}(t_3), \hat{p}(t_4), \hat{p}(t_5)] = [0, 0.5, 0, 0.5, 0]$

Apéndice B

Regresión cuantil

A continuación, presentamos o código relativo á gráfica amosada na Figura 2.2.

```
library(quantreg)\par
ct<-c(492.934,530.024,570.855,609.744,627.013,598.490,612.349,611.386,602.781,590.837,
601.586,618.514,636.323,666.407,688.641)\par
cp<-c(88.795,93.211,97.997,101.211,100.128,92.471,92.114,91.719,88.866,83.581,85.277,
90.594,96.217,104.914,106.109)\par
plot(ct~cp)\par
mod<-rq(ct~cp)\par
taus <- c(0.1, 0.25,0.5, 0.75, 0.9)\par
plot(ct~cp,ylab="Gasto en consumo final dos fogares (millóns de euros)",xlab="Gasto
en hoteis, cafes e restaurantes (millóns de euros)",main="Período 2004-2018")
taus <- seq(0,1,by=0.1)\par
for (i in 1:length(taus)) {abline(rq(ct~cp, tau = taus[i]), col = "gray")}
abline(rq(ct~cp, tau = 0.5), col = "red")
```

Apéndice C

Regresión cuantil con datos censurados por intervalos

Recollemos neste apéndice o código completo en R do método proposto neste documento para a estimación da regresión cuantil con datos censurados por intervalos. Digo código está implementado para o conxunto de pacientes con VIH positivo previamente citado.

```
library(interval)
library(Epi)
data(hivDK)
d<-hivDK
left<-as.numeric(d$well-d$entry)
left.na<-left
left[is.na(left)]<-0
right<- as.numeric(d$ill - d$entry)
right.na<-right
right[is.na(right)]<-Inf
d<-data.frame(left,left.na,right,right.na,us=d$us,year.of.birth=d$bth+1950,age.at.entry=
d$bth+30,partners.per.year=d$pyr)
y<-as.numeric((d$right+d$left)/2)
y.na<-y
y[is.infinite(y)]<-3057
x<-d$age.at.entry
L<-d[,"left"]
R<-d[,"right"]
est=icfit(L,R)
u=est$intmap[1,]; u
v=est$intmap[2,]; v
pf=est$pf; pf
plot(icfit(L,R))
p=length(u)  # Número de intervalos de Turnbull
n=nrow(d) # Número de individuos da mostra
#-- Atopamos os intervalos de Turnbull
#-- correspondentes a cada individuo da mostra
# Extremos inferiores
```

```
z=c(L,u)
zord=sort(z,index.return=T)
ind=zord$ix
a=c()
a0=1 # Indice nos intervalos de Turnbull
for (iz in 1:(n+p)){
if (ind[iz]<=n){a[ind[iz]]=a0}else{
a0=a0+1}
}
# Extremos superiores
z=c(v,R)
zord=sort(z,index.return=T)
ind=zord$ix
b=c()
b0=p # Indice nos intervalos de Turnbull
for (iz in (n+p):1){
if (ind[iz]>p){b[ind[iz]]=b0}else{
b0=b0-1}
}
b=b[(p+1):(p+n)]
inew=0
xnew=c()
ynew=c()
pesos=c()
for (i in 1:n){
pt=sum(pf[a[i]:b[i]])
for (ip in a[i]:b[i]){
inew=inew+1
xnew[inew]=x[i]
ynew[inew]=(v[ip]+u[ip])/2
pesos[inew]=pf[ip]/pt }
}
ynew<-as.numeric(ynew)
ynew.na<-ynew
ynew[is.infinite(ynew)]<-3057
library(quantreg)
m=rq(ynew~xnew,tau=0.1,weights=pesos)
m
plot(y~x,type="n", xlab="Idade (en anos)", ylab="Permanencia no estudo (en días)",ylim=c(0,3500)
,main="Gráfico do estudo de VIH",axes=T)
legend(x = "left",y = "right", legend = c("Recta axustada","Censura pola drta.","Intervalo
censurado","Censura pola esq."), fill = c("black", "yellow","red","royalblue"))
for(i in 1:n){if(R[i]==Inf){segments(x[i],L[i],x[i],3500, col ="yellow",lwd=4)}}
for(i in 1:n){if(L[i]==0){segments(x[i],L[i],x[i],R[i], col ="blue",lwd=2)}
else{segments(x[i],L[i],x[i],R[i], col ="red")}}
abline(m,col="black", lwd=3)
```

Apéndice D

Simulación do método de regresión cuantil con datos censurados por intervalos

Neste apéndice incluímos o código en R relativo á simulación do método. O axuste da simulación presenta un valor para a orde do cuantil de 0.5, un tamaño mostral 100, unha densidade de observacións de 10 e 1000 simulacións.

```
library(interval)
set.seed(123456)
n=100
nobs=10
ns=1000
vbeta0=c()
vbeta1=c()
obs=matrix(0,nrow=n,ncol=nobs)
for (is in 1:ns){  # Bucle das mostras simuladas
x=runif(n)
y=1+3*x+runif(n)
for (iobs in 1:nobs){
obs[,iobs]=1+4*(iobs-1)/nobs+4*runif(n)/nobs}
L=c()
R=c()
for (i in 1:n){
pobs=sum(obs[i,]<y[i])
if (pobs==0){L[i]=0;R[i]=obs[i,1]}else{
if (pobs==nobs){L[i]=obs[i,nobs];R[i]=5} else{
L[i]=obs[i,pobs];R[i]=obs[i,pobs+1] }
}
}
est=icfit(L,R)
u=est$intmap[1,]
v=est$intmap[2,]
pf=est$pf
```

```
p=length(u)  # Número de intervalos de Turnbull
#-- Atopamos os intervalos de Turnbull
#-- correspondentes a cada individuo da mostra
# Extremos inferiores
z=c(L,u)
zord=sort(z,index.return=T)
ind=zord$ix
a=c()
a0=1 # Indice nos intervalos de Turnbull
for (iz in 1:(n+p)){
if (ind[iz]<=n){a[ind[iz]]=a0}else{
a0=a0+1}
}
# Extremos superiores
z=c(v,R)
zord=sort(z,index.return=T)
ind=zord$ix
b=c()
b0=p # Indice nos intervalos de Turnbull
for (iz in (n+p):1){
if (ind[iz]>p){b[ind[iz]]=b0}else{
b0=b0-1}
}
b=b[(p+1):(p+n)]
inew=0
xnew=c()
ynew=c()
pesos=c()
for (i in 1:n){
pt=sum(pf[a[i]:b[i]])
for (ip in a[i]:b[i]){
inew=inew+1
xnew[inew]=x[i]
ynew[inew]=(v[ip]+u[ip])/2
pesos[inew]=pf[ip]/pt }
}
library(quantreg)
m=rq(ynew~xnew,tau=0.5, weights=pesos)
# m
#abline(m)
beta=coef(m)
vbeta0[is]=beta[1]
vbeta1[is]=beta[2]
cat("Mostra", is, "Beta0", vbeta0[is], "Beta1", vbeta1[is], "\n")
}
mean(vbeta0); mean(vbeta1)
var(vbeta0); var(vbeta1)
ecm0=(mean(vbeta0)-1.5)^2+var(vbeta0); ecm0
ecm1=(mean(vbeta1)-3)^2+var(vbeta1); ecm1
```

Apéndice E

Comparación entre os métodos

Neste apéndice está incluído o código da comparativa entre o modelo paramétrico exponencial e o modelo de regresión cuantil con datos censurados. O número de simulacións aquí fixadas será 1000, o tamaño mostral 100, a orde do cuantil 0.5 e a densidade de observacións 10.

```
library(icenReg)
library(interval)
set.seed(123456)
n=100
nobs=10
ns=100
vbeta0=c()
vbeta1=c()
vbeta1_par=c()
vbeta0_par=c()
obs=matrix(0,nrow=n,ncol=nobs)
for (is in 1:ns){  # Bucle das mostras simuladas
x=runif(n)
y=1+3*x+runif(n)
for (iobs in 1:nobs){
obs[,iobs]=1+4*(iobs-1)/nobs+4*runif(n)/nobs}
L=c()
R=c()
for (i in 1:n){
pobs=sum(obs[i,]<y[i])
if (pobs==0){L[i]=0;R[i]=obs[i,1]}else{
if (pobs==nobs){L[i]=obs[i,nobs];R[i]=5} else{
L[i]=obs[i,pobs];R[i]=obs[i,pobs+1] }
}
}
#--- Método paramétrico de regresión en media
datos=data.frame(L,R,x)
aft_exp_fit <- ic_par(Surv(L,R,type="interval2") ~ x, data=datos,
dist = "exponential", model = "aft")
summary(aft_exp_fit)
```

```
vbeta1_par[is]=exp(coef(aft_exp_fit))[2]
vbeta0_par[is]=exp(coef(aft_exp_fit))[1]
#--- Método proposto no TFM
est=icfit(L,R)
#plot(icfit(L,R))
u=est$intmap[1,]
v=est$intmap[2,]
pf=est$pf
p=length(u)  # Número de intervalos de Turnbull
#-- Atopamos os intervalos de Turnbull
#-- correspondentes a cada individuo da mostra
# Extremos inferiores
z=c(L,u)
zord=sort(z,index.return=T)
ind=zord$ix
a=c()
a0=1 # Indice nos intervalos de Turnbull
for (iz in 1:(n+p)){
if (ind[iz]<=n){a[ind[iz]]=a0}else{
a0=a0+1}
}
# Extremos superiores
z=c(v,R)
zord=sort(z,index.return=T)
ind=zord$ix
b=c()
b0=p # Indice nos intervalos de Turnbull
for (iz in (n+p):1){
if (ind[iz]>p){b[ind[iz]]=b0}else{
b0=b0-1}
}
b=b[(p+1):(p+n)]
inew=0
xnew=c()
ynew=c()
pesos=c()
for (i in 1:n){
pt=sum(pf[a[i]:b[i]])
for (ip in a[i]:b[i]){
inew=inew+1
xnew[inew]=x[i]
ynew[inew]=(v[ip]+u[ip])/2
pesos[inew]=pf[ip]/pt }
}
library(quantreg)
m=rq(ynew~xnew,weights=pesos)
beta=coef(m)
vbeta0[is]=beta[1]
vbeta1[is]=beta[2]
cat("Mostra", is, "Beta0", vbeta0[is], "Beta1", vbeta1[is],
"Beta0_par", vbeta0_par[is], "Beta1_par", vbeta1_par[is],"\n")
}
```

```
mean(vbeta0); mean(vbeta1)
var(vbeta0); var(vbeta1)
ecm0=(mean(vbeta0)-1.5)^2+var(vbeta0); ecm0
ecm1=(mean(vbeta1)-3)^2+var(vbeta1); ecm1
mean(vbeta0_par)
var(vbeta0_par)
ecm0=(mean(vbeta0_par)-3)^2+var(vbeta0_par); ecm0
mean(vbeta1_par)
var(vbeta1_par)
ecm1=(mean(vbeta1_par)-3)^2+var(vbeta1_par); ecm1
```

Bibliografía

[1] Anderson-Bergman, C. (2017). icenReg: Regression Models for Interval Censored Data in R. Journal of Statistical Software, 81(12), 1-23. https://doi.org/10.18637/jss.v081.i12

[2] Barrodale, I., Roberts, F. D. K. (1973). An Improved Algorithm for Discrete l_1 Linear Approximation. SIAM Journal on Numerical Analysis, 10(5), 839-848. https://doi.org/10.1137/0710069

[3] Becker, N. G., Melbye, M. (1990). Use of a log-linear model to compute the empirical survival curve from interval-censored data, with application to data on tests for HIV-positivity. Australian Journal of Statistics, 33, 125-133. https://doi.org/10.1111/j.1467-842X.1991.tb00420.x.

[4] Betensky, R. A., Rabinowitz, D., Tsiatis, A. A. (2001). Computationally Simple Accelerated Failure Time Regression for Interval Censored Data. Biometrika, 88(1), 703-711. https://www.jstor.org/stable/2673440.

[5] Dempster, A. P., Laird, N. M., Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1), 01-38. https://www.jstor.org/stable/2984875.

[6] Fay, M. P. (2014). CRAN - Package intervals. CRAN. https://cran.r-project.org/web/packages/intervals/index.html.

[7] Gasto en consumo final de los hogares interior por CCAA y periodo. 2015. (2019, septiembre). INE. https://www.ine.es/jaxi/tabla.do?path=/t00/ICV/Graficos/dim1/l0/&file=117G1.px&type=pcaxis&L=0

[8] Kaplan, E. L., Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. Journal of the American Statistical Association, 53(282), 457-481. http://www.jstor.org/stable/2281868

[9] Koenker, R. (2005). Asymptotic Theory of Quantile Regression. Quantile Regression, 116-150. https://doi.org/10.1017/cbo9780511754098.005

[10] Koenker, R. (2008a). CRAN - Package quantreg. CRAN. http://cran.r-project.org/web/packages/quantreg/index.html

[11] Koenker, R. (2008b, junio). Censored Quantile Regression Redux. Journal of Statistical Software, 27(6), 1-25. http://www.jstatsoft.org/

[12] Koenker, R., Bassett Jr., G. (1982). Robust Tests for Heteroscedasticity Based on Regression Quantiles. Econometrica, 50(1), 43-61. http://www.jstor.org/stable/1912528

[13] Lawless, J. F., Babineau, D. (2006). Models for Interval Censoring and Simulation-Based Inference for Lifetime Distributions. Biometrika, 93(3), 671-686. http://www.jstor.com/stable/20441315

[14] Li, L., Pu, Z. (2003). Rank estimation of log-linear regression with interval-censored data. Lifetime Data Analysis, 9, 57-70. https://doi.org/10.1023/a:1021882122257

[15] Melbye, M., Biggar, R. J., Ebbesen, P., Sarngadharan, M. G., Weiss, S. H., Gallo, R. C., Blattner, W. A. (1984). Seroepidemiology of HTLV-III antibody in Danish homosexual men: prevalence, transmission, and disease outcome. BMJ, 289(6445), 573-575. https://doi.org/10.1136/bmj.289.6445.573

[16] Portnoy, S. (2013, diciembre). Censored Regression Quantiles. Journal of the American Statistical Association, 98(464), 1001-1012. http://www.jstor.org/stable/30045346

[17] Powell, J. L. (1983). Least absolute deviations estimation for the censored regression model. Econometrics, 25, 303-325. https://doi.org/10.1016/0304-4076(84)90004-6

[18] Rabinowjtz, D., Tsiatis, A., Aragon, J. (1995). Regression with interval-censored data. Biometrika, 82(3), 501-513. https://doi.org/10.1093/biomet/82.3.501

[19] Shiboski, S. C. (1998). Generalized additive models for current status data. Lifetime Data Analysis, 4, 29-50. https://doi.org/10.1023/a:1009652024999

[20] Stute, W. (1999). Nonlinear censored regression. Statistica Sinica, 9(4), 1089-1102. https://www.jstor.org/stable/24306638

[21] Sun, J., Sun, L. (2005). Semiparametric linear transformation models for current status data. The Canadian Journal of Statistics, 33, 85-96. https://doi.org/10.1002/cjs.5540330107

[22] Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables. Econometrica, 26(1), 24-36. http://www.jstor.org/stable/1907382

[23] Turnbull, B. W. (1976). The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. Journal of the Royal Statistical Society. Series B (Methodological), 38(3), 290-295. http://www.jstor.org/stable/2984980

[24] Younes, N., Lachin, J. (1997). Linked-based models for survival data with interval and continuous time censoring. Biometrics, 53, 1199-1211. https://doi.org/10.2307/2533490

[25] Zhang, Z. (2010). Interval censoring. Statistical Methods in Medical Research, 19, 53-70. https://doi.org/10.1177/0962280209105023

[26] Zhang, Z., Sun, L., Zhao, X., Sun, J. (2005). Regression analysis of interval censored failure time data with linear transformation models. The Canadian Journal of Statistics, 33, 61-70. https://www.jstor.org/stable/25046161

[27] Zhou, X., Feng, Y., Du, X. (2016). Quantile Regression for Interval Censored Data. Communications in Statistics - Theory and Methods, 45, 1-27. https://doi.org/10.1080/03610926.2015.1073317