TRABAJO DE FIN DE MÁSTER

Máster Interuniversitario en Estadística e Investigación Operativa

Título: Regresión cuantílica para la cuantificación del riesgo

Autor: Álvaro Cortés Ruiz

Directora: Catalina Bolancé Losilla

Departamento: Econometria, Estadística i Economia Aplicada

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

UNIVERSITAT DE BARCELONA

Facultat de Matemàtiques i Estadística
Universitat Politècnica de Catalunya

Tesis de máster

Regresión cuantílica para la cuantificación del riesgo

Álvaro Cortés Ruiz

Directora: Catalina Bolancé Losilla

Departament Econometria, Estadística i Economia aplicada

Dedicatorias

A mi madre, por hacerme la persona que soy, porque siempre has sido el faro que ha guiado mi camino.

A mi pareja Celia, por tu paciencia y amor incondicional, porque tú has hecho que todo tenga sentido.

A mi padre, mis tíos y mi primo, por siempre protegerme, porque habéis hecho que todo esto sea más sencillo.

Agradecimientos

A todas las personas que me han formado y transmitido todo su conocimiento a lo largo de estos años, en especial a Catalina Bolancé por su guía, su ayuda y su capacidad de transmitir la pasión y conocimiento sobre la estadística.

A todos los compañeros y amistades generadas estos años. Ha sido un placer vivir todos los momentos buenos y duros de estos años de formación.

Resumen

El modelo de regresión cuantílica de Koenker y Bassett (1978) ha tenido un gran impacto en los análisis del ámbito económico-financiero en general, y es de particular interés en la gestión y la cuantificación de riesgos.

Cuando nos adentramos en los datos que caracterizan la gestión y la cuantificación de riesgos vemos que se encuentran determinados por distribuciones de cola pesada y además los cuantiles de mayor interés son los más extremos, como por ejemplo el 99.5% o el 99.9%.

Mediante un estudio de simulación se demuestra que el modelo de Koenker y Bassett (1978) fracasa a la hora de estimar los coeficientes de un modelo lineal para los cuantiles más extremos; además, cuando se introducen nuevos métodos como la estimación con splines o estimación núcleo (kernel), y se comparan con el modelo de regresión cuantílica mediante el cálculo del error cuadrático medio de la predicción (ECMP) no hay una mejora de la predicción para los cuantiles más extremos. Por lo tanto, se constata la necesidad de proporcionar métodos alternativos que nos permitan mejorar la estimación de los cuantiles condicionales para niveles de confianza α próximos a 1 y con distribución de cola pesada.

Finalmente, se ha realizado un análisis sobre una cartera de pólizas de seguros de automóvil y de hogar mediante el modelo de regresión cuantílica de Koenker y Bassett (1978) que muestra la importancia para entender los factores causantes de los costes más extremos de la cartera. De este análisis, hemos podido extraer lo siguiente, por un lado, para los costes generados por las pólizas de automóvil, se muestra que el género tiene un efecto significativo, así como también lo tiene la edad de los conductores. Por otro lado, el modelo que analiza los costes de pólizas de hogar, destaca la gran influencia del lugar de residencia en los costes de los siniestros.

Palabras clave: cuantiles, regresion cuantílica, splines, estimación núcleo, kernel, cuantiles extremos, distribuciones de cola pesada, gestión y cuantificación de riesgos.

Abstract

The quantile regression model of Koenker and Bassett (1978) has had a great impact on the economic-financial analysis in general, and is of particular interest in the management and quantification of risks.

When we go into the data that characterize the management and quantification of risks we see that they are characterized by heavy tail distributions and also our quantiles of interest are the most extreme, such as 99.5% or 99.9%.

A simulation study shows that the Koenker and Bassett (1978) model fails to estimate the coefficients of a linear model for the most extreme quantiles; In addition, when new methods like splines or density estimation (kernel) are introduced and compared with the quantile regression model by calculating the predicted mean square error (PMSE) there is no improvement on the prediction of the most extreme quantiles. Therefore, it is necessary to provide alternative methods that allow us to improve the estimation of conditional quantiles for confidence levels α close to 1 with heavy tail distribution.

Finally, an analysis of a portfolio of automobile and home insurance policies has been carried out using the quantile regression model of Koenker and Bassett (1978), which shows the importance of understanding the factors causing the most extreme costs of an insurance portfolio. From this analysis, we have been able to extract the following conclusion, on one hand, for the costs generated by automobile policies, it is shown that gender has a significant effect, as does the age of drivers. On the other hand, the model that analyses the costs of home insurance policies, shows that the place of residence has a great influence on the costs of claims.

Keywords: quantiles, quantile regression, splines, density estimation, kernel, extreme quantiles, heavy tailed distributions, risk management and quantification

Índice de contenidos

Dedicatorias i
Agradecimientos ii
Resumen iii
Abstract iv
Índice de tablas e ilustraciones vi
1. Introducción 1
2. Marco teórico 5
2.1. Fundamentos del modelo de regresión cuantílica 5
2.2 Regresión cuantílica no paramétrica 8
2.3 Modelos aditivos de regresión cuantílica con splines 9
2.4 Regresión cuantílica no paramétrica mediante estimación núcleo (Kernel) 11
3. Estudio de simulación 12
3.1 Procedimiento para la generación de datos 13
3.2 Resultados análisis parámetros estimados por el modelo Koenker y Bassett (1978) 17
3.3 Resultados análisis para predicciones 24
4. Aplicación a datos reales 30
4.1. Descripción de los datos 30
4.2 Análisis de datos para las pólizas de automóvil 33
4.3 Análisis de datos para las pólizas de hogar 44
5. Conclusiones 54
Bibliografía 57
Apéndice A: Código R implementado en el estudio de simulación 59
Apéndice B: Código R implementado en datos de costes de pólizas de automóvil 66
Apéndice C: Código R implementado en datos de costes de pólizas de hogar 72

Índice de tablas e ilustraciones

Ilustración 1: Distribución de e bajo Pareto (1,8) 14
Ilustración 2: Distribución de e bajo Pareto (1,2) 15
Ilustración 3: Distribución de e bajo Log-normal (0,1)..... 15
Ilustración 4: Relación de x1 con suponiendo distribución Pareto (1,8)..... 17
Ilustración 5: Relación de x2 con suponiendo distribución Pareto (1,8)..... 18
Tabla 1: Error Cuadrático Medio de la estimación de parámetros del modelo bajo Pareto (1,8) 19
Ilustración 6: Relación de x1 con suponiendo distribución Pareto (1,2)..... 20
Ilustración 7: Relación de x2 con suponiendo distribución Pareto (1,8)..... 20
Tabla 2: Error Cuadrático Medio de la estimación de parámetros del modelo bajo Pareto (1,2) 21
Ilustración 8: Relación de x1 suponiendo distribución Log-normal(0,1)..... 22
Ilustración 9: Relación de x2 suponiendo distribución Log-normal(0,1)..... 22
Tabla 3: Error Cuadrático Medio de la estimación de parámetros del modelo bajo Log-normal(0,1)..... 23
Tabla 4: Comparación error cuadrático medio de la predicción para Pareto (1,8)..... 25
Tabla 5: Comparación error cuadrático medio de la predicción para Pareto (1,2)..... 26
Tabla 6: Comparación error cuadrático medio de la predicción para Log-normal (0,1)..... 27
Ilustración 10: Distribución costes asociados a pólizas de automóvil 31
Tabla 7: Cuantiles costes asociados a las pólizas de automóvil..... 31
Ilustración 11: Distribución costes asociados a pólizas de hogar..... 32
Tabla 8: Cuantiles costes asociados a las pólizas de automóvil..... 32
Tabla 9: Descriptivos variables cualtitativas para costes de pólizas de automóvil 33
Ilustración 12: Distribución de la edad del tomador del seguro después de limpieza..... 34
Ilustración 13: Distribución de la edad del primer conductor del seguro después de limpieza .. 34
Tabla 10: Descriptivos variables cualitativas para costes de pólizas de automóvil después de limpieza..... 35
Tabla 11: Matriz de correlaciones..... 35
Ilustración 14: Relación género y edad del primer conductor para costes de automóvil 36
Ilustración 15: Relación género y edad del cliente para costes de automóvil 36
Ilustración 16: Relación género para costes de automóvil 37
Ilustración 17: Costes pólizas de automóvil por disposición de pólizas 37
Ilustración 18: Costes pólizas de automóvil residentes grandes ciudades 38
Ilustración 19: Costes pólizas de automóvil residentes norte..... 38
Ilustración 20: Relación de coeficientes por cuantiles para costes de automóvil con la edad del primer conductor 39
Ilustración 21: Relación de coeficientes por cuantiles para costes de automóvil con la edad del primer conductor a partir del cuantil 0.9 40
Ilustración 22: Relación de coeficientes por cuantiles para costes de automóvil con la edad del tomador 41
Ilustración 23: Relación de coeficientes por cuantiles para costes de automóvil con la edad del tomar a partir del cuantil 0.9 42
Tabla 12: Resultados modelo de regresión cuantílica sobre costes de pólizas de automóvil..... 43
Tabla 13: Descriptivos variables cualitativas para costes de pólizas de hogar 45
Ilustración 24: Distribución de la edad del tomador de póliza de hogar después de limpieza 46
Tabla 14: Descriptivos variables continúas para costes de pólizas de hogar después de limpieza 46
Ilustración 25: Relación género y edad del cliente para costes de hogar 47
Ilustración 26: Relación género para costes de hogar 47
Ilustración 27: Costes pólizas de hogar por disposición de pólizas 48

Ilustración 28: Costes pólizas de hogar residentes en una gran ciudad..... 48
Ilustración 29: Costes pólizas de hogar residentes en el norte..... 49
Ilustración 30: Relación de coeficientes por cuantiles para costes de hogar..... 50
Ilustración 31: Relación de coeficientes por cuantiles para costes de hogar a partir del cuantil 0.9 51
Tabla 15: Resultados modelo de regresión cuantílica sobre costes de pólizas de hogar..... 52

1. Introducción

La técnicas de regresión lineal estándar analizan el efecto entre un conjunto de variables explicativas y una variable dependiente a partir del ajuste de la media condicional, $E(Y|X)$; por lo tanto, estas relaciones estimadas solo describen de forma parcial la relación de interés de la distribución condicional de Y.

En la regresión lineal estándar, se asume que la relación entre variables es la misma para todos los niveles, lo cual nos genera una visión parcial de los efectos de las variables explicativas sobre la variable dependiente. Esta visión limitada puede ser ampliada mediante el ajuste de la regresión lineal en otras medidas de posición como son los cuantiles, estos nos permiten aumentar la visión de las relaciones en diferentes puntos de la distribución. En esta visión más amplia de las relaciones entre variables se basará este trabajo. Específicamente, se explicarán teóricamente y se aplicarán técnicas paramétricas y no paramétricas en un contexto de datos con cola pesada y cuantiles extremos.

Como primer paso, se explicará el origen de la regresión cuantílica y sus propiedades. De forma análoga a la relación de la media condicional de la regresión lineal.

Koenker y Bassett (1978) introdujeron la idea de analizar la relación entre las variables explicativas y la variable dependiente para la mediana condicional,$Q_{0,5}(Y|X)$, entendida como el cuantil $\alpha = 0{,}5$, sea un nivel de confianza α. Este concepto se puede extender para todos los cuantiles α, donde α puede tomar valores en el intervalo $(0{,}1)$, que dividen Y en la proporción α por debajo y $1 - \alpha$ por arriba del cuantil. De esta forma, la regresión cuantílica conceptualmente pondera las distancias entre los datos predichos y reales y las minimiza. También podemos entender la relación estocástica de las variables de una forma más precisa mediante la descripción a diversos niveles de confianza. Además de dar una oportunidad de mayor entendimiento de la relación entre variable dependiente y variables explicativas, el modelo de regresión cuantílica añade más robustez frente a valores atípicos que el modelo de regresión estimado por el método de los mínimos cuadrados y, consecuentemente, tiene una mejor aplicación a un gran abanico de datos, donde éstos no se comportan como una distribución normal.

Como hemos comentado anteriormente, la regresión cuantílica tiene una amplia diversidad de aplicaciones en análisis reales que nos permiten entender mejor las relaciones entre diferentes tipologías de datos. Algunos ejemplos que han destacado en el tiempo son:

- En el campo de la economía de salud, se analizan a los bebés que al nacer tienen un peso inferior al promedio, ya que éste hecho es un detonante de enfermedades graves. Chernozhukov y Fernández-Val (2011) mediante un estudio empírico demuestran los factores que producen que un recién nacido tenga un peso anormalmente bajo. De esta forma es posible intentar atacar los detonantes del bajo peso y prevenir enfermedades sobre recién nacidos.

- En el campo de la eficiencia de la productividad, se estudia la frontera de la productividad $Q_{1-\alpha}(Y|X)$. De esta forma se describe los factores que explican los niveles de producción Y para las $(1-\alpha)x100$ empresas más productivas.

- El uso de la regresión cuantílica es el análisis de eventos sistemáticos adversos en los rendimientos de un banco. Se puede analizar el impacto de una gran caída del valor de mercado de una cartera en el rendimiento global del banco y de esta forma determinar los niveles de capital necesario para prevenir la bancarrota.

Como se puede observar en los ejemplos anteriores la regresión cuantílica se ha mostrado como una herramienta muy importante y ampliamente usada en el ámbito de la economía y la salud para el análisis de los regresores en la distribución condicional de las variables dependientes. Específicamente, como se ve en los ejemplos, la aplicación de la regresión cuantílica en las colas surge de forma natural en las aplicaciones financiera.

En los estudios de carácter económico-financiero encontramos típicamente datos donde las colas son muy pesadas y los cuantiles de interés son los más extremos, como por ejemplo en los análisis del valor en riesgo (VaR, value-at-risk) o la eficiencia de la productividad. Los datos y estudios del ámbito económico-financiero se centran mayoritariamente en variables que se pueden describir mediante una distribución de una ley-potencial donde algunos eventos solo ocurren con una frecuencia muy baja, esta relación fue descrita por primera vez por Vilfredo Pareto en 1895 para los datos de ingreso y riqueza y conllevó una gran serie de aplicaciones y la generación de nuevas teorías en

el tiempo, como es la de valores extremos. Un claro ejemplo de la importancia del análisis de esta tipología de datos y comportamientos, es el estudio de Jansen y de Vries (1991) donde demuestran que mediante el análisis de las colas de los retornos financieros de Estados Unidos la caída de los mercados de 1987 no fue un valor atípico sino un evento raro predecible con datos previos. La predicción de eventos raros pero con un alto impacto, es especialmente importante en la gestión de riesgos y, de forma concreta, en la gestión de riesgos financiero. Gracias al modelo de regresión cuantílica de Koenker y Bassett (1978), se han generado trabajos en el campo de la economía financiera de relieve, como por ejemplo el cálculo del valor en riesgo condicional (CVaR, conditional value-at-risk) (Chernozhukov y Umantsev, 2001; Engle y Manganelli, 2004).

Bajo estas premisas de la necesidad de trabajo y análisis del entorno económico-financiero con la visión de cuantiles condicionales se enmarca este trabajo. En esta situación, el estudio realizado intenta responder a las siguientes cuestiones; ¿Es el modelo clásico de Koenker y Bassett (1978) óptimo para el análisis?, ¿Han surgido nuevas técnicas que nos permitan una mejorar estimación y predicción?. Para resolver estas dudas compararemos el módelo clásico con dos métodos no paramétricos.

Para dar lugar a la resolución de las cuestiones anteriores, el trabajo se ordenará en cuatro secciones adicionales. En la segunda sección se describe el marco teórico de los métodos que comparamos; el primero presentado es el de regresión cuantílica de Koenker y Bassett (1978). Describimos también el método basado en splines y, finalmente, el método basado en estimación núcleo.

Para realizar el análisis planteado a lo largo de este trabajo, nos centraremos en la estimación de la regresión cuantílica $Q_\alpha(Y|X)$ cuando α está próximo a 1, por tanto, en la tercera sección, se describen los resultados de un estudio de simulación, que han sido organizados en dos apartados. En primer lugar, se muestran los resultados de la estimación de los coeficientes por parte del modelo de regresión cuantílica y, en segundo lugar, se presenta la comparación de las predicciones para los tres métodos mencionados anteriormente. Teniendo como objetivos de la simulación, analizar la calidad de la estimación de los coeficientes de la regresión de Koenker y Bassett (1978) y el error cuadrático medio de la predicción de todos los métodos. En general los estudios que se mostrarán en los siguientes apartados están focalizados en las aplicaciones enmarcadas

bajo este escenario donde tenemos una serie datos con una distribución con cola pesada y queremos estimar una serie de cuantiles extremos.

En la cuarta sección se muestran los resultados del análisis con datos reales de una cartera de pólizas de seguros, en el que se modelizan los costes asociados a pólizas de automóvil y los costes asociados a pólizas de hogar. El objetivo de esta sección es reafirmar la necesidad de incorporar nuevas técnicas basadas en cuantiles condicionales en el ámbito económico-actuarial y, específicamente, en la gestión de riesgos. Para ello se ha realizado un estudio de cuáles son los potenciales factores determinantes de los riesgos asociados a pólizas de seguro de hogar y pólizas de seguro de automóvil.

Finalmente, en el quinto apartado se resumen las principales conclusiones del trabajo.

2. Marco teórico

2.1. Fundamentos del modelo de regresión cuantílica

El modelo de regresión cuantílica básico fue introducido por Koenker y Bassett (1978) como una extensión del modelo de regresión en la media condicional a una aproximación que ensambla un conjunto de modelos para varias funciones de cuantiles condicionales. De forma alternativa al método de los mínimos cuadrados ordinarios, el modelo de regresión cuantílica se estima en base a la minimización asimétricamente ponderada de los errores absolutos y, particularmente, en el caso específico de la mediana como la minimización de la suma de los errores absolutos.

Si repasamos el modelo de regresión lineal, podemos representar la estimación de un vector de parámetros desconocidos, $\beta = (\beta_1, \beta_2, \dots, \beta_k)$, de una muestra de observaciones independientes de una serie de variables aleatorias $Y_1, Y_2, \dots, Y_i$ distribuidas de tal manera que

$$P(Y_i < y) = F(y - x_i\beta)$$

donde $x_i = (x_{i1}, \dots, x_{ik}), i = 1,2, \dots, n$, donde k es el número de variables explicativas y n el tamaño muestral, es un vector de valores de las variables explicativas y la distribución de F es desconocida. En el caso de que F sea conocida y coincida con la normal, se puede demostrar que el estimador de mínimos cuadrados es el estimador con varianza mínima para los estimadores insesgados. La principal debilidad del modelo es que trabaja bajo la hipótesis de normalidad de los errores causando una gran sensibilidad a valores atípicos. Una alternativa a esta debilidad, y además como modelo con capacidad para describir de forma más amplia las relaciones entre variable dependiente y variable explicativa surge el modelo de regresión cuantílica.

En el primer paso, podemos definir el cuantil muestral. Bajo el escenario donde $y_i, i = 1,2, \dots, n$ es una muestra aleatoria de n observaciones de la variable aleatoria Y con función de distribución F, entonces el cuantil muestral, donde $0 < \alpha < 1$, se puede definir como la solución del siguiente problema de minimización:

$$min_{b \in \mathbb{R}}[\sum_{i \in (i: y_i \geq b)} \alpha |y_i - b| + \sum_{i \in (i: y_i < b)} (1 - \alpha)|y_i - b|]$$

Podemos observar que α define el porcentaje relativo de observaciones por encima y $(1-\alpha)$ el porcentaje relativo de observaciones por debajo de y_i. El caso más conocido es el de $\alpha = \frac{1}{2}$ que se corresponde con la mediana, es decir dejamos por encima el 50% de las observaciones y por debajo el 50% restante.

Este concepto de cuantil empírico se puede extender al modelo lineal suponiendo que $x_i = (x_{i1}, \dots, x_{ik}), i = 1,2, \dots, n$, son los valores de las k variables explicativas para el caso i en la regresión $u_i = Y_i - x_i\beta$ con función de distribución F. La regresión cuantílica para una probabilidad α, donde $0 < \alpha < 1$, es la solución al siguiente problema de minimización:

$$min_{\beta \in \mathbb{R}^k}[\sum_{i \in (i:y_i \geq x_i\beta)} \alpha|y_i - x_i\beta_\alpha| + \sum_{i \in (i:y_i < x_i\beta)}(1-\alpha)|y_i - x_i\beta_\alpha|].$$

Bajo esta definición matemática lo que obtenemos es que, para el vector de parámetros estimados $\hat{\beta}_\alpha$ todas la observaciones por encima del hiperplano de $x_i\hat{\beta}_\alpha$ son ponderadoras por α, mientras que las que se encuentran por debajo son ponderadas por $(1-\alpha)$. Es importante remarcar que la idea de que se puede estimar la regresión cuantílica mediante la segmentación de Y en subconjuntos de acuerdo con su distribución no condicional y luego estimar la regresión de mínimos cuadrados es errónea. Esto conlleva a lo denominado como sesgo de la selección, es decir los resultados son erróneos a causa de un truncamiento de Y.

Además, el cuantil condicional se puede especificar de una forma lineal tal que:

$$Q_\alpha(y_i|x_i) = x_i\beta_\alpha.$$

Los efectos marginales son los coeficientes en el cuantil α:

$$\frac{\partial Q_\alpha(y_i|x_i)}{\partial x_i} = \beta_\alpha.$$

Los paramétros β_α estiman los cambios en el cuantil α de la variable dependiente y_i producido por los cambios de una unidad los valores de las variables en el vector x_i. Es importante destacar que se asume que después del cambio en x_i la variable dependiente y_i se mantiene en el mismo cuantil muestral.

Tal y como se demuestra en el trabajo de Koenker y Bassett (1978) la regresión cuantílica tiene las siguientes propiedades:

1. Si X, como la matriz de datos explicativos, tiene rango K entonces el conjunto de regresiones cuantilícas, $B^*(\alpha)$, tiene al menos un elemento de la forma:

$$B^*(\alpha) = X(h)^{-1}y(h)$$

2. En relación a los coeficientes estimados, $\hat{\beta}_\alpha$:
   a. $\hat{\beta}_\alpha(\alpha, \lambda\, y, X) = \lambda\hat{\beta}_\alpha(\alpha, y, X)$ para $\lambda \in [0, \infty)$
   b. $\hat{\beta}_\alpha(1-\alpha, \lambda\, y, X) = \lambda\hat{\beta}_\alpha(\alpha, y, X)$ para $\lambda \in -(\infty, 0]$
   c. $\hat{\beta}_\alpha(\alpha, y + X\gamma, X) = \lambda\hat{\beta}_\alpha(\alpha, y, X) + \gamma$ para $\gamma \in \mathbb{R}^k$
   d. $\hat{\beta}_\alpha(\alpha, y, XA) = A^{-1}\hat{\beta}_\alpha(\alpha, y, X)$ para A_{KxK}no singular

Por lo tanto, a y b establecen que los coeficientes son equivariantes en escala, es decir, si Y es reescalada por λ, entonces $\hat{\beta}_\alpha$ es reescalada en la misma proporción. La propiedad c es conocida como la propiedad de localización, donde, si $\hat{\beta}_\alpha$ es la solución de (y, X) entonces $\beta_\alpha + \gamma$ es la solución de $(y + X\gamma, X)$. Finalmente, d representa la equivarianza de reparametrización de la matriz de diseño. Todas estas propiedades se mantienen iguales que en el modelo de regresión lieanl pero se añade una quinta propiedad:

   e. $\widehat{Q_\alpha}(j(y)|X) = j(Q_\alpha(y|X))$ para $j(.)$ en $\mathbb{R}$

Es decir, los cuantiles condicionales son equivariantes frente a transformaciones crecientes de la variable dependiente.

3. Una de las propiedades más relevantes del modelo es la propiedad de robustez frente a valores atípicos. Esta propiedad añade el valor distintivo desde el punto de vista teórico más importante respecto al método de mínimos cuadrados. Si definimos los residuos como $u^* = y - X\beta^*$ y D es una matriz diagonal de tamaño con elementos no negativos, entonces si $\beta^*(\alpha) \in B^*(\alpha, y, X)$ entonces $\beta^*(\alpha) \in B^*(\alpha, X\beta^* + Du^*, X)$.

Mencionadas las principales características del modelo de regresión cuantílica podemos ver que tiene una serie de propiedades análogas al modelo de regresión lineal clásico. En este conjunto de propiedades, se añade una nueva más relevante en la cual se muestra una mayor robustez ante valores atípicos, por lo tanto, podemos concluir que la regresión

cuantílica no únicamente tiene un interés para el análisis de relaciones entre variable dependiente y explicativas, sino que también aporta nuevas propiedades desde el punto de vista teórico que nos pueden ser beneficiosas en el trabajo con datos distribuidos de forma no normal.

2.2 Regresión cuantílica no paramétrica

En las últimas décadas ha habido un elevado crecimiento del desarrollo tanto práctico como teórico de modelos no paramétricos. Estos se han visto representados en diversas técnicas y se han introducido en multitud de análisis, entre los cuales los cuantiles condicionales no ha sido una excepción.

En este marco, el abanico de soluciones propuestas por diversos autores ha ocupado la mayoría de técnicas conocidas y desarrolladas en las últimas décadas. Uno de los primeros modelos fue la aplicación de nearest neighbours. Contemporáneamente se introdujeron modelos basados en la estimación núcleo (kernel) como por ejemplo el expuesto en Chaudhuri (1991). Otra de las metodologías implementadas es la de White (1991), que introduce el uso de redes neuronales. El uso de splines bajo los marcos teóricos de Cole (1988) o Bloomfiled y Steiger (1983), formaría la última de las metodologías destacadas. Queremos destacar dos en su expansión y desarrollo en el campo de la regresión cuantílica y, consecuentemente, serán en estas dos metodologías en las que nos centraremos en esta sección. La primera corresponde con los modelos basados en splines introducidos por Cole (1988) y una posterior evolución propuesta por Koenker, Ng, y Portnoy (1994) con la mezcla de los modelos aditivos de Hastie y Tibshirani (1987), derivó en la metodología más usada actualmente. En segundo lugar destacan los modelos basados en kernels desarrolladas por Takeuchi et al. (2006).

En los próximos apartados, se explicaran de forma breve los marcos teóricos en los que se fundamentan los splines y la estimación núcleo, con el objetivo de poder entender cuál es la lógica e intuición que se usa y como se aplica a nuestro marco de análisis de cuantiles condicionales.

2.3 Modelos aditivos de regresión cuantílica con splines

En primer lugar, para mejorar la comprensión de la técnica, introduciremos el concepto de spline en el ámbito de las matemáticas.

De forma genérica, un spline es una función que está formada por varios polinomios, cada uno definido sobre un subintervalo. Estos se unen entre sí obedeciendo a ciertas condiciones de continuidad. Es decir, es una función continua que busca conectar los datos con un segmento polinomial de P grados. Estas funciones se aplican ampliamente en el suavizado de curvas. De forma concreta una función $l(x)$ es un spline en el intervalo $[a, b]$, si se cumple que, dada una partición de los datos en el intervalo

$$P = \{a = x_0 < x_1 \dots < x_n = b\},$$

de esta forma $l(x)$ es un polinomio en $[x_i < x_{i+1}]$, $i = 0,1, \dots, n$, donde los puntos representados por x_i, $i = 0,1, \dots, n-1$ se llaman nodos del spline.

A lo largo de esta sección y de las siguientes realizaremos un cambio de anotación para simplificar la exposición teórica, en vez de la anotación de Koenker y Bassett introducida anteriormente:

$$min_{\beta \in \mathbb{R}^k}[\sum_{i \in (i: y_i \geq x_i \beta)} \alpha |y_i - x_i \beta_\alpha| + \sum_{i \in (i: y_i < x_i \beta)} (1-\alpha)|y_i - x_i \beta_\alpha|].$$

Pasaremos a utilizar la siguiente representación del problema para simplificar:

$$min_{\xi \in \mathbb{R}^k} \sum_{i=1}^{n} \rho_\alpha (y_i - \xi),$$

donde $\rho_\alpha(u) = u\{\alpha - I(u < 0)\}$ es la función de comprobación de Koenker y Bassett (1978), es decir representa las ponderaciones α y $(1-\alpha)$.

Los modelos aditivos suponen una forma pragmática de modelizar los modelos no paramétricos mediante la restricción de los componentes no paramétricos a un espacio dimensional bajo. De esta forma podremos evitar el conocido efecto de "curse of dimensionality".

Igual que en el caso de la regresión cuantílica clásica, el modelo basado en splines también se puede traducir de forma análoga del modelo de splines para media

condicional. De forma que el modelo general para la media condicional se puede ver como:

$$min_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n}\big(y_i - g(x_i)\big)^2 + \lambda J(g),$$

donde g es la función spline, λ es el término de penalización y $J(.)$ es una función definida sobre el espacio de $g(x_i)$ .Con la misma idea, un estimador de cuantiles basados en splines para g puede venir dado por el problema de minimización siguiente:

$$min_{g\in\mathcal{G}} \frac{1}{n}\sum_{i=1}^{n}\rho_\alpha\{y_i - g(x_i)\} + \lambda\, J(g).$$

En este problema encontramos dos términos, el primero que mide la fidelidad a las observaciones, y el segundo término es una penalización sobre g par evitar el sobre-ajuste. Además, λ se define como un parámetro de ajuste que controla la magnitud del efecto de J.

Los primeros modelos no paramétricos de regresión cuantílica con el uso de splines lo propuso Cole (1988) a partir de la siguiente estimación de cuantiles mediante splines:

$$min_{g\in\mathcal{G}} \sum_{i=1}^{n}\rho_\alpha\{y_i - g(x_i)\} + \lambda \int \{g''(x)\}^2 dx.$$

Bajo este planteamiento $\lambda \in \mathbb{R}_+$ controla el grado de alisamiento del spline cúbico resultante.

Posteriormente a este trabajo, Koenker, Ng, y Portnoy (1994) introdujeron el reemplazo de la penalización $\{g''(x)\}^2$ por $|g''(x)|$, es decir cambiar el término de penalización a una forma basada en la variación total. El principal objetivo de esta variante es adaptar el trabajo para obtener un problema de programación línea, es decir de esta forma las soluciones con la forma de L_1 del término de penalización son splines lineales. Bajo este cambio el problema de minimización propuesto sería el siguiente:

$$min_{g\in\mathcal{G}} \sum_{i=1}^{n}\rho_\alpha\{y_i - g(x_i)\} + \lambda \int |g''(x)|\, dx.$$

Con $x_0 < x_1 \ldots < x_n < x_{n+1} = 1$, sobre el espacio de Sobolev W_1^2 de la función continua en $[0,1]$ con primeras derivadas continuas y segundas derivadas integrables. Este

modelo fue posteriormente extendido desde el modelo con una variable explicativa al modelo con dos variables explicativas por Koenker y Mizera (2003).

Esta lógica se puede ampliar a los modelos aditivos manteniendo como función de alisado la variación total de forma que el nuevo problema a minimizar tenga la forma:

$$min_{g\in\mathcal{G}} \frac{1}{n}\sum \rho_\alpha \left\{ y_i - \sum_{j=1}^{n} g_j(z_{ij}) \right\} + \sum_{j=1}^{n} \lambda \bigvee(\nabla g_j),$$

donde $\bigvee(\nabla g_j)$ representa la variación total del gradiente de la función g.

Como es habitual en los modelos no paramétricos, λ representa un parámetro de alisamiento el cual tiene un papel crucial para el balanceo entre la fidelidad del dato y la penalización. En caso de que λ sea muy grande, la penalización aplicada es muy alta y como consecuencia hay un sobre-alisamiento. En el caso contrario, cuando λ es muy pequeña, hay un sub-alisamiento debido a que se interpolan más los datos.

Teniendo en cuenta el gran impacto que tiene el parámetro λ en los resultados se definió un criterio de cálculo para el valor óptimo de este parámetro. Koener, Ng y Portnoy(1994) se basaron en el criterio de Schwarz (1978) para el cálculo del valor:

$$SIC(\lambda) = n\log(\hat{\sigma}(\lambda) + \frac{1}{2}p(\lambda)\log(n),$$

donde $\hat{\sigma}(\lambda) = n^{-1}\sum_{i=1}^{n} \rho_\alpha\,(y_i - \hat{g}(x,z))$, y $p(\lambda)$ es el número efectivo de dimensiones del modelo ajustado y se puede definir de la siguiente forma:

$$p(\lambda) = \sum_{i=1}^{n} \frac{\partial \hat{g}(x_i, z_i)}{\partial y_i}$$

2.4 Regresión cuantílica no paramétrica mediante estimación núcleo (Kernel)

Como método alternativo Takeuchi et al. (2006) plantearon la aplicación del uso de la estimación núcleo. Los métodos de regresión basados en kernels se fundamentan en la estimación de una función de regresión $g(x)$ mediante un método tipo núcleo. Esto se realiza mediante el uso de aquellas observaciones alrededor del punto de interés x para estimar la función $\tilde{g}(x)$ de tal manera que sea alisada en $\mathbb{R}^p$. Esto se consigue mediante

el uso de una función ponderada de x_i, que se denomina núcleo de la estimación $K(x_i, x_j)$, y depende de la distancia entre x_0 y la observación x_i, esa función se denomina núcleo de la estimación. Además estas funciones se ven controladas por un parámetro de ajuste λ que controla la influencia de las observaciones vecinas a considerar.

En un escenario parecido al explicado previamente en los modelos basados en splines, podemos definir el escenario inicial como el uso de la función de perdida definida en la regresión cuantílica

$$min_{g\in G}\frac{1}{n}\sum_{i=1}^{n}\rho_\alpha\{y_i - g(x_i)\} + \lambda\, J(g).$$

Para $J(g)$ la función seleccionada puede ser cualquier núcleo deseado. En el caso concreto que trabajaremos, nos basaremos en el Reproducing Kernel Hilbert Spaces (RKHS) introducido por Takeuchi et al. (2006).

Bajo el escenario definido en Takeuchi et al. (2006). Asumimos que $G = \{g = g' + b\colon g' \in \mathcal{H}, b \in \mathbb{R}\}$, donde $\mathcal{H}$ es RKHS para x_i con la función Kernel $K(.,.)$, y b es el intercepto de la función de regresión. Además, la norma de $\mathcal{H}$ se representa como $||.||_{\mathcal{H}}$. Bajo estos supuestos, la regresión cuantílica con una penalización compuesta por la norma al cuadrado, se obtiene a partir de resolver el siguiente problema de minimización:

$$min_{g\in G}\frac{1}{n}\sum_{n=1}^{n}\rho_\alpha\big(y_i - g(x_i)\big) + \lambda\big|\big|g'\big|\big|_{\mathcal{H}}^2.$$

3. Estudio de simulación

Con el objetivo de evaluar los diferentes métodos de estimación analizados en este trabajo, se ha realizado un estudio de simulación.

Para este estudio de simulación se han propuesto modelos alternativos para diferentes distribuciones de colas más o menos pesadas. Se han obtenido 500 réplicas de cada modelo con muestra de tamaño 100, 500 y 1000 observaciones. Se ha estimado en los cuantiles extremos condicionados bajo el comportamiento de los modelos propuestos. Para estudiar los modelos se realizan dos tipos de análisis.

En primer lugar, queremos evaluar la estimación del método de regresión cuantílica clásico a través de la comparación de los parámetros reales del modelo con los predichos. En este primer proceso se quiere ver como varía el error en el ajuste a medida que se usan distribuciones con colas más pesadas y cuantiles más extremos. La calidad de la estimación se analizará con el cálculo del sesgo y el error cuadrático medio (ECM) para diferentes tamaños muéstrales y cuantiles. En segundo lugar, queremos comparar la calidad de las predicciones de diferentes modelos estadísticos para diversos pares de observaciones. Para ello ajustaremos las 3 alternativas expuestas en el marco teórico y analizaremos la predicción a partir de las muestras generadas en un conjunto de observaciones, en concreto, estas son $x_i = (-0.5, -,0.5)$, $x_i = (0,0)$ y $x_i = (0.5,0.5)$.

Por lo tanto, como objetivo se quiere, por una parte, entender las deficiencias del modelo de regresión clásico para datos con colas pesadas y cuantiles extremos, que son situaciones frecuentes en datos de origen económico-financiero. Y por otra parte, cual es la mejoría de los nuevos modelos basados en metodologías no paramétricas, como splines y kernels, en términos de predicción en comparación con el modelo clásico de regresión cuantílica propuesto por Koenker y Bassett (1978).

3.1 Procedimiento para la generación de datos

Para la generación de los datos se ha utilizado un modelo basado en el estudio de simulación de Wang, Li y He (2012), en el cual también se realiza una comparación de modelos de regresiones cuantílicas. Los datos se han generado a partir del siguiente modelo:

$$y_i = x_{i1} + x_{i2} + (1 + rx_{i1})e_i, i = 1, \dots, n,$$

que está basado en dos variables explicativas que siempre definiremos con la misma distribución. Es importante destacar que para el modelo la variable x_{i2} toma un efecto constante sobre los cuantiles α, ya que el valor de su coeficiente es 1 independientemente del cuantil analizado.

Se ha definido las variables explicativas con distribución uniforme con valores entre -1 y 1:

$$x_{ij} \sim Uniforme(-1,1), j = 1,2,$$

e_i son variables aleatorias independiente e idénticamente distribuidas para la que se han definido tres distribuciones diferentes.

La primera distribución utilizada es de Pareto con cola ligera. Para ello se le ha definido un parámetro de escala, η, 1 y de forma, θ, 8. La distribución de Pareto tiene la siguiente función de densidad asociada:

$$f(e;\eta,\theta) = \frac{\theta\eta^{\theta}}{e^{\theta+1}}, \eta > 0, \theta > 0, e \geq \eta.$$

A continuación, en la Ilustración 1 se muestra un ejemplo de la distribución de e_i para una de las réplicas con n=1000.

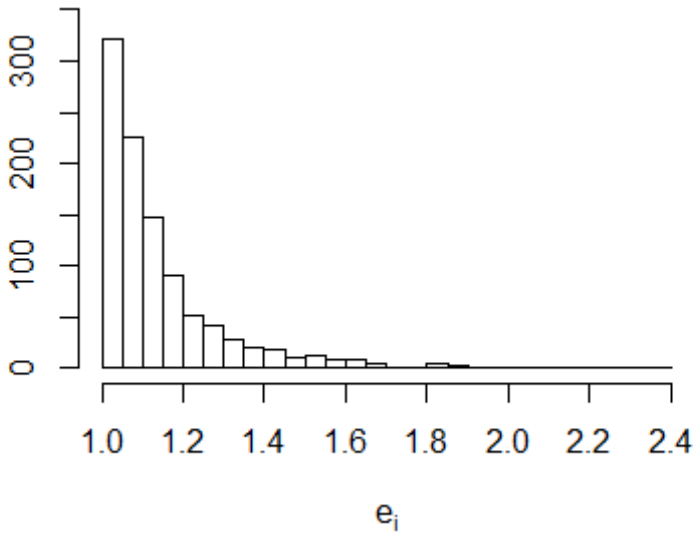

Ilustración 1: Distribución de e bajo Pareto (1,8)

Posteriormente se ha utilizado una distribución de Pareto de cola pesada con parámetros de escala, η, 1 y de forma, θ, 2. Definiendo la distribución de Pareto como se ha expresado anteriormente. A continuación, en la Ilustración 2 se muestra un ejemplo de la distribución de e_i para una de las réplicas con n=1000.

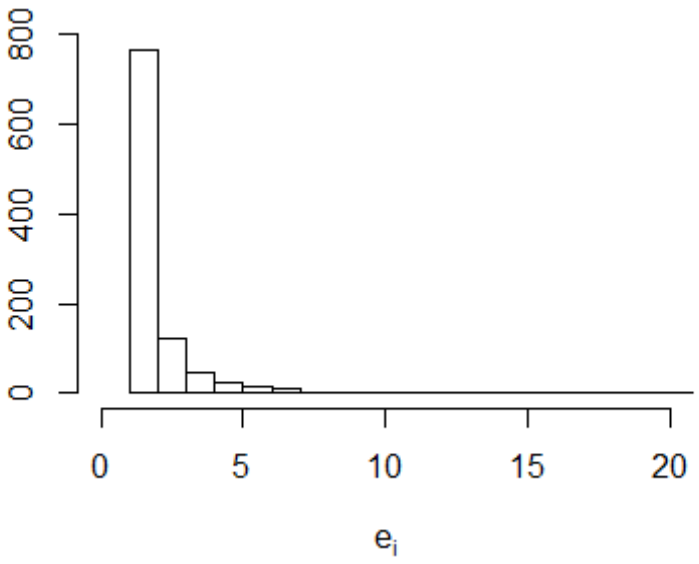

Ilustración 2: Distribución de e bajo Pareto (1,2)

La última distribución implementada para e_i es de tipo Log-normal con parámetros de media, μ, 0 y varianza, σ, 1. Se ha definido la función de distribución de la siguiente manera:

$$f(e;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(\ln(e)-\mu)^2/2\sigma^2}.$$

A continuación, en la Ilustración 3 se muestra un ejemplo de la distribución de e_i para una de las réplicas con n=1000.

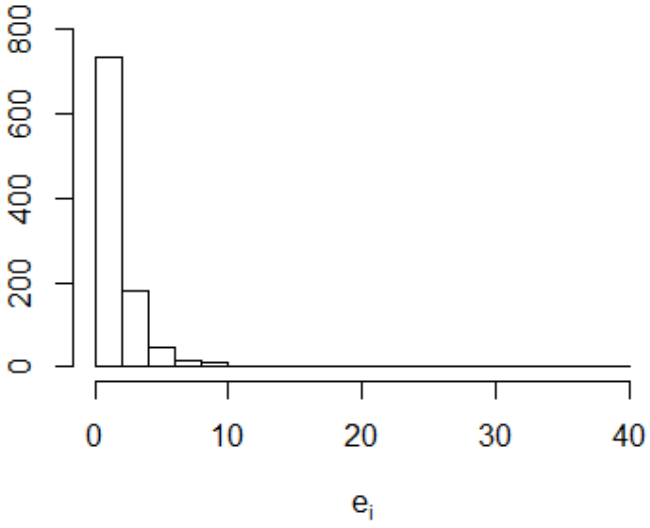

Ilustración 3: Distribución de e bajo Log-normal (0,1)

Con estas tres distribuciones, el objetivo es poder ver como varía el ajuste según el tipo de cola de la distribución.

Finalmente, r es una constante controlando el grado de heterocedasticidad.

Bajo este modelo y un cuantil, α_i, se puede definir el cuantil condicional de Y como:

$$Q_Y(\alpha|x_i) = \delta(\alpha) + x_i'\beta(\tau).$$

Los parámetros de la igualdad se definen como:

$x_{ij} = (x_{i1}, x_{i2})$, son las variables explicativas

$\delta(\alpha) = Q_e(\alpha)$, es el cuantil α de e_i y supone la constante del modelo, β_0,

$\beta(\alpha) = \big(\beta_1(\alpha), \beta_2(\alpha)\big) = (1 + rQ_e(\alpha), 1)$, es el vector de coeficiente de las variables explicativas para un cuantil α

$Q_e(\alpha)$ es el α cuantil de e_i.

Se han generado los datos con las distribuciones definidas para e_i y x_{ij} con $r = 0.9$, podemos decir que nos encontramos bajo una situación de heterocedasticidad. Los modelos implementados se han analizado bajo tres tamaños muéstrales distintos, $n = 100, 500, 1000$, y seis cuantiles distintos, $\alpha = 0.50, 0.70, 0.90, 0.95, 0.99, 0.995$, para un total de 500 réplicas generadas. Bajo estos escenarios se quiere ver el efecto al ampliar el tamaño muestral y al desplazar el análisis de cuantiles centrales a los extremos.

Para el primer paso donde se comparan los parámetros reales del modelo y los estimados, se han generado dos métricas: el sesgo y el error cuadrático medio. El sesgo se puede definir como:

$$Sesgo_\beta\big[\hat{\beta}\big] = E_{x|\beta}\big[\hat{\beta}\big] - \beta$$

y el error cuadrático medio (ECM) se define como:

$$ECM\big(\hat{\beta}\big) = Var\big(\hat{\beta}\big) + sesgo(\hat{\beta})^2.$$

Para evaluar la capacidad predictiva de los tres modelos, se obtiene el error cuadrático medio de la predicción. Esta métrica se ha definido como:

$$ECMP\big(g(x_i)\big) = \sum_{i=1}^{n}(E[\hat{g}\,(x_i) - g(x_i))^2 + \sum_{i=1}^{n} var(\hat{g}\,(x_i)).$$

3.2 Resultados análisis parámetros estimados por el modelo Koenker y Bassett (1978)

Como primera aproximación al análisis, se ha evaluado el modelo de regresión cuantílica clásica propuesto por Koenker y Bassett (1978). Para la distribución de Pareto(1,8) podemos observar en las ilustraciones 4 y 5 un ejemplo de datos para una de las réplicas con tamaño muestral de n=1000.

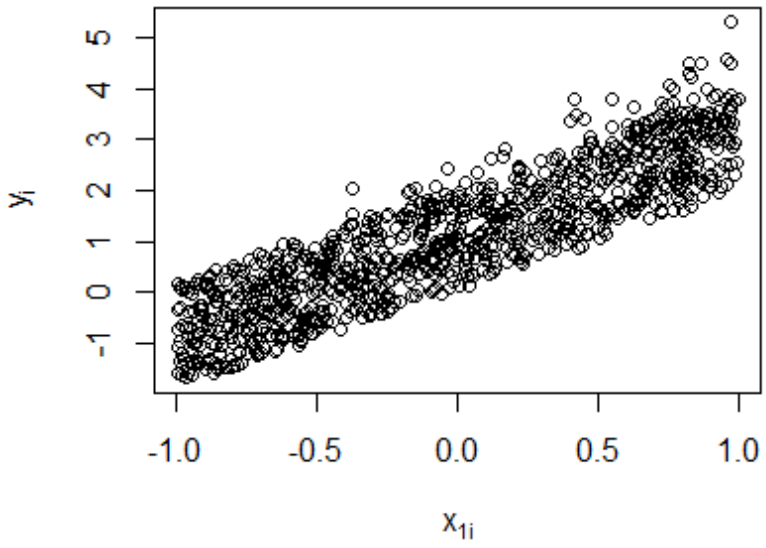

Ilustración 4: Relación de x1 con suponiendo distribución Pareto (1,8)

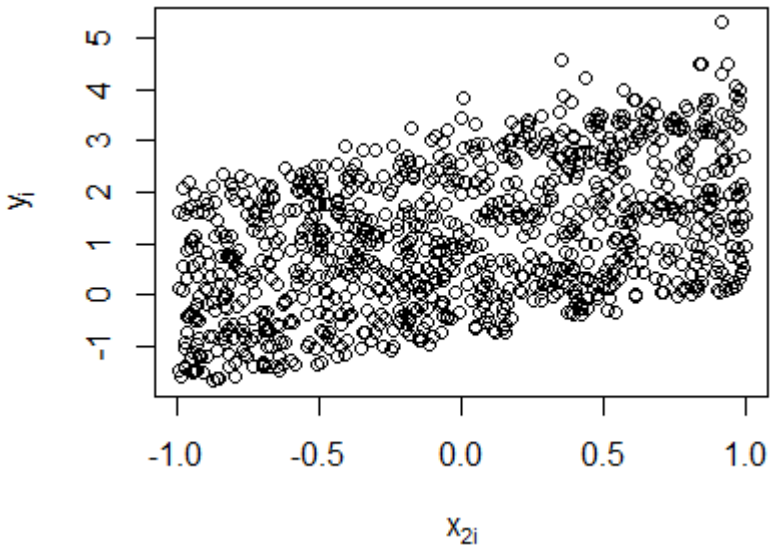

Ilustración 5: Relación de x2 con suponiendo distribución Pareto (1,8)

Como se observa en las ilustraciones 4 y 5, bajo la distribución de Pareto(1,8) no se obtienen datos extremos. Los resultados obtenidos para esta se muestran en la Tabla 1, los cuales muestran como a medida que el cuantil es más extremo el ECM aumenta.

	n=100				n=500				n=1000			
α=0.5	**Valor Real**	**Sesgo**	**Varianza**	**ECM**	**Valor Real**	**Sesgo**	**Varianza**	**ECM**	**Valor Real**	**Sesgo**	**Varianza**	**ECM**
β_0	1.089	0.001	0.000	0.000	1.095	0.000	0.000	0.000	1.090	0.000	0.000	0.000
β_1	1.980	-0.002	0.000	0.000	1.926	0.000	0.000	0.000	1.981	0.000	0.000	0.000
β_2	1.000	-0.001	0.000	0.000	1.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000
α=0.7												
β_0	1.167	0.001	0.000	0.000	1.172	0.000	0.000	0.000	1.164	0.000	0.000	0.000
β_1	2.050	-0.002	0.001	0.001	2.055	-0.001	0.000	0.000	2.047	0.000	0.000	0.000
β_2	1.000	-0.001	0.001	0.001	1.000	-0.001	0.000	0.000	1.000	0.000	0.000	0.000
α=0.9												
β_0	1.345	0.006	0.002	0.002	1.365	0.000	0.000	0.000	1.314	0.001	0.000	0.000
β_1	2.210	-0.001	0.004	0.004	2.229	-0.004	0.001	0.001	2.182	-0.002	0.001	0.001
β_2	1.000	0.001	0.002	0.002	1.000	0.000	0.000	0.000	1.000	0.001	0.000	0.000
α=0.95												
β_0	1.434	0.030	0.006	0.007	1.476	-0.003	0.001	0.001	1.417	0.001	0.001	0.001
β_1	2.290	0.015	0.011	0.011	2.329	-0.009	0.002	0.002	2.275	-0.004	0.001	0.006
β_2	1.000	-0.010	0.003	0.003	1.000	0.001	0.001	0.001	1.000	0.001	0.001	0.001
α=0.99												
β_0	1.570	0.044	0.078	0.080	1.735	-0.012	0.006	0.007	1.879	0.013	0.006	0.006
β_1	2.413	-0.051	0.114	0.117	2.562	-0.047	0.014	0.016	2.691	-0.017	0.018	0.018
β_2	1.000	-0.035	0.038	0.039	1.000	0.001	0.013	0.013	1.000	-0.006	0.006	0.006
α=0.995												
β_0	1.655	-0.031	0.078	0.079	1.862	-0.017	0.018	0.018	2.152	0.036	0.016	0.017
β_1	2.490	-0.012	0.114	0.114	2.675	-0.091	0.004	0.012	2.937	-0.017	0.029	0.029
β_2	1.000	-0.035	0.038	0.039	1.000	-0.011	0.003	0.003	1.000	-0.007	0.014	0.014

Tabla 1: Error Cuadrático Medio de la estimación de parámetros del modelo bajo Pareto (1,8)

Para la distribución de Pareto(1,2), en las ilustraciones 6 y 7 podemos encontrar un ejemplo de datos para una de las réplicas con tamaño muestral de n=1000.

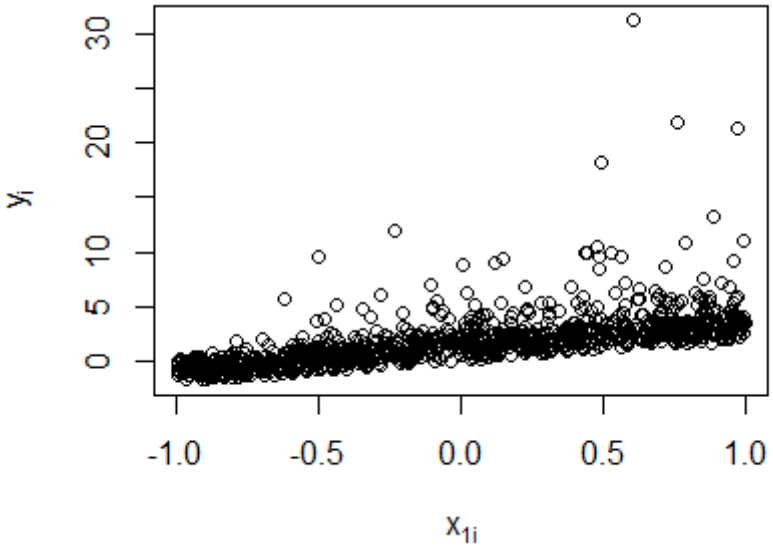

Ilustración 6: Relación de x1 con suponiendo distribución Pareto (1,2)

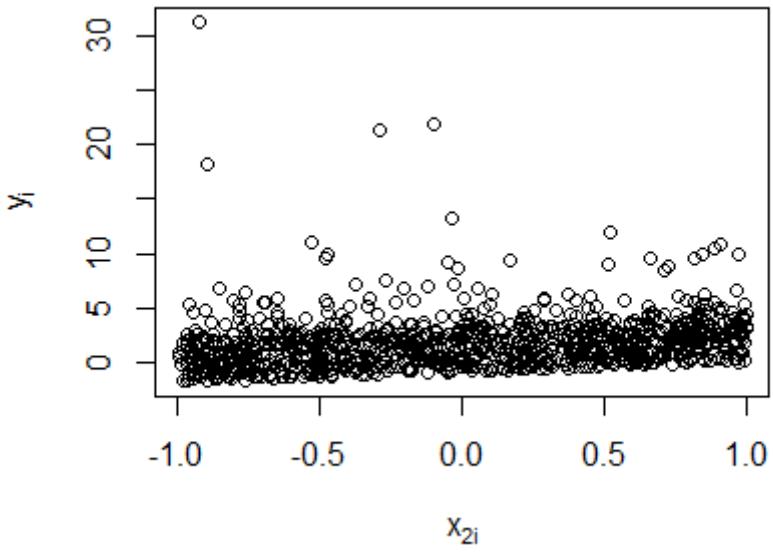

Ilustración 7: Relación de x2 con suponiendo distribución Pareto (1,8)

A diferencia de los datos generados suponiendo una distribución de Pareto(1,2), en las ilustraciones 6 y 7 se observa como han aparecido datos con valores extremos. Los resultados obtenidos para esta distribución se muestran en la Tabla 2, dónde se aprecia claramente como el ECM se incrementa a medida que lo hace el cuantil α. Además, se observa que el ECM aumenta respecto a los observados en la Tabla 1, ya que ahora hemos supuesto una distribución con una cola más pesada.

	n=100				n=500				n=1000			
α=0.5	**Valor Real**	**Sesgo**	**Varianza**	**ECM**	**Valor Real**	**Sesgo**	**Varianza**	**ECM**	**Valor Real**	**Sesgo**	**Varianza**	**ECM**
β_0	1.409	0.021	0.007	0.008	1.441	0.000	0.002	0.002	1.414	0.003	0.001	0.001
β_1	2.268	0.014	0.012	0.013	2.297	-0.006	0.003	0.003	2.273	0.002	0.002	0.002
β_2	1.000	-0.003	0.006	0.006	1.000	0.003	0.001	0.001	1.000	0.006	0.000	0.000
α=0.7												
β_0	1.857	0.056	0.028	0.031	1.889	-0.004	0.005	0.005	1.837	0.003	0.003	0.003
β_1	2.671	0.022	0.051	0.051	2.700	-0.015	0.010	0.010	2.653	0.001	0.005	0.005
β_2	1.000	-0.002	0.030	0.030	1.000	0.006	0.003	0.003	1.000	0.001	0.002	0.002
α=0.9												
β_0	3.278	0.204	0.295	0.337	3.478	0.030	0.083	0.084	2.985	0.013	0.021	0.021
β_1	3.950	-0.091	0.623	0.631	4.130	-0.025	0.141	0.142	3.686	-0.022	0.050	0.051
β_2	1.000	0.038	0.324	0.326	1.000	-0.015	0.046	0.047	1.000	-0.013	0.003	0.003
α=0.95												
β_0	4.233	0.399	1.883	2.041	4.758	0.062	0.427	0.431	4.037	0.053	0.149	0.152
β_1	4.809	-0.434	4.291	4.479	5.282	-0.084	0.645	0.652	4.633	-0.052	0.233	0.236
β_2	1.000	0.139	1.612	1.631	1.000	-0.004	0.220	0.220	1.000	-0.053	0.016	0.019
α=0.99												
β_0	6.082	2.644	90.486	97.476	9.075	0.380	4.817	4.961	12.486	0.069	3.993	3.998
β_1	6.474	-0.908	64.999	65.823	9.167	-0.844	8.999	9.711	12.238	0.320	4.769	4.871
β_2	1.000	0.601	67.271	67.632	1.000	-0.340	5.429	5.545	1.000	-0.144	1.917	1.938
α=0.995												
β_0	7.636	-0.207	90.486	90.528	12.134	2.271	32.200	37.359	21.469	1.693	1.541	4.407
β_1	7.872	-3.474	64.999	77.067	11.921	-0.956	29.336	30.249	20.322	0.436	20.600	20.790
β_2	1.000	0.601	67.271	67.632	1.000	-0.205	43.603	43.645	1.000	-0.298	10.542	10.630

Tabla 2: Error Cuadrático Medio de la estimación de parámetros del modelo bajo Pareto (1,2)

Para la distribución de Log-Normal(0,1), en las ilustraciones 8 y 9 podemos encontrar un ejemplo de datos para una de las réplicas con tamaño muestral de n=1000.

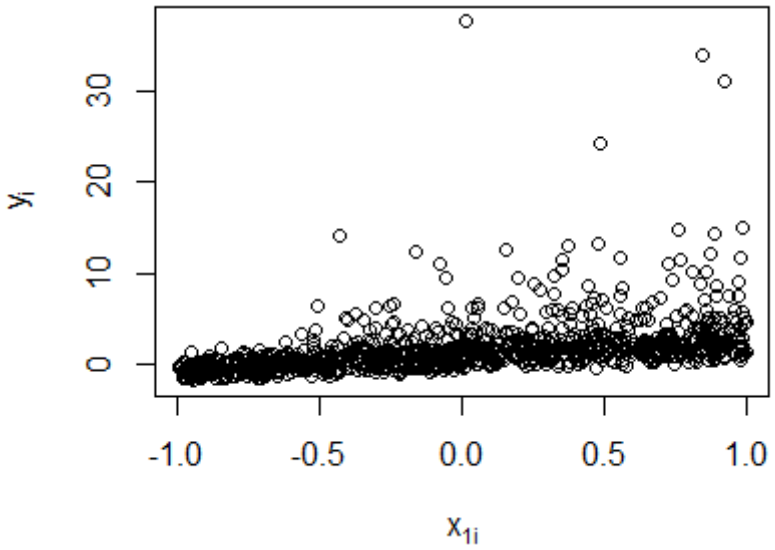

Ilustración 8: Relación de x1 suponiendo distribución Log-normal(0,1)

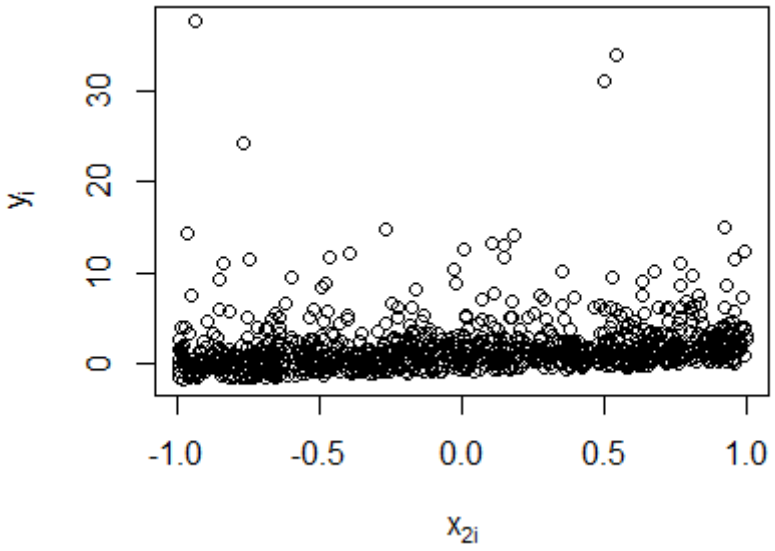

Ilustración 9: Relación de x2 suponiendo distribución Log-normal(0,1)

De igual forma que bajo el supuesto de una distribución de Pareto(1,2), observamos como en los datos han aparecido observaciones con valores extremos. Los resultados obtenidos para esta distribución que se muestran en la Tabla 3, estos reflejan de igual que en la Tabla 1 y Tabla 2 un incremento del ECM al incrementar el cuantil analizado.

	n=100				n=500				n=1000			
α=0.5	**Valor Real**	**Sesgo**	**Varianza**	**ECM**	**Valor Real**	**Sesgo**	**Varianza**	**ECM**	**Valor Real**	**Sesgo**	**Varianza**	**ECM**
β_0	1.219	0.040	0.002	0.004	0.996	0.008	0.006	0.006	1.090	0.001	0.001	0.001
β_1	2.097	0.158	0.046	0.071	1.896	0.011	0.010	0.010	1.981	-0.002	0.002	0.002
β_2	1.000	0.016	0.016	0.016	1.000	0.004	0.003	0.003	1.000	0.014	0.002	0.002
α=0.7												
β_0	2.065	0.047	0.063	0.065	1.611	0.011	0.014	0.014	1.838	-0.005	0.004	0.004
β_1	2.858	0.707	0.117	0.617	2.450	-0.001	0.027	0.027	2.654	-0.017	0.008	0.008
β_2	1.000	0.071	0.065	0.070	1.000	0.012	0.012	0.013	1.000	0.020	0.006	0.006
α=0.9												
β_0	4.161	0.050	0.412	0.414	3.123	0.004	0.121	0.121	3.818	0.011	0.080	0.080
β_1	4.745	0.003	0.705	0.705	3.811	-0.068	0.189	0.194	4.436	0.000	0.130	0.130
β_2	1.000	0.003	0.421	0.421	1.000	0.095	0.090	0.099	1.000	0.039	0.028	0.030
α=0.95												
β_0	5.907	0.287	1.451	1.533	3.986	0.087	0.334	0.342	5.272	0.085	0.256	0.263
β_1	6.316	0.079	2.431	2.437	4.587	-0.057	0.536	0.539	5.745	0.028	0.397	0.397
β_2	1.000	0.079	1.631	1.638	1.000	0.295	0.291	0.378	1.000	0.019	0.114	0.114
α=0.99												
β_0	7.839	1.976	55.220	59.126	7.603	0.181	3.061	3.093	9.877	0.310	1.503	1.599
β_1	8.055	0.474	67.650	67.874	7.843	-0.673	4.835	5.289	9.890	0.018	2.640	2.640
β_2	1.000	0.474	24.770	24.994	1.000	4.028	3.574	19.797	1.000	-0.102	1.481	1.491
α=0.995												
β_0	9.045	-0.366	55.220	55.354	9.589	0.490	6.440	6.680	13.320	0.922	5.130	5.979
β_1	9.141	0.474	67.650	67.874	9.630	-1.262	11.665	13.257	12.988	-0.037	10.330	10.331
β_2	1.000	0.474	24.770	24.994	1.000	10.063	8.471	109.737	1.000	-0.024	3.053	3.054

Tabla 3: Error Cuadrático Medio de la estimación de parámetros del modelo bajo Log-normal(0,1)

De forma general podemos concluir en base a los resultados obtenidos, bajo el escenario específico de esta simulación, que el modelo de Koenker y Bassett (1978) se ve afectado en términos de ajuste de los parámetros por el nivel de α seleccionado. Vemos claramente que a medida que se incrementa el cuantil de interés la desviación es cada vez mayor respecto al valor real del modelo. Para un tamaño muestral de $n = 100$ este problema se acentúa en nuestro marco a partir de $\alpha = 0.90$ y esto supone un problema para la aplicación del modelo en análisis del ámbito económico-financiero. Para tamaños muestrales superiores el problema se acentúa a partir de $\alpha = 0.99$ este efecto toma lugar de forma mucho más notoria para las dos distribuciones con colas más pesadas (Pareto(1,2) y Log-normal(0,1)).

En el siguiente apartado analizaremos la comparativa conjunta de las predicciones de los tres modelos.

3.3 Resultados análisis para predicciones

Bajo los métodos expuestos en el marco se ha realizado una comparación de los resultados de las predicciones para las 500 réplicas y las tres distribuciones definidas al inicio de esta sección y tres observaciones, $x_i = (-0.5, -0.5)$, $x_i = (0,0)$ y $x_i = (0.5,0.5)$.

Los resultados para Pareto(1,8), Pareto(1,2) y Log-normal(0,1) se muestran respectivamente en las Tablas 4, 5 y 6.

	n=100			n=500			n=1000		
x= (-0.5 . -0.5)	**Modelo clásico**	**Modelo con Kernels**	**Modelo con Splines**	**Modelo clásico**	**Modelo con Kernels**	**Modelo con Splines**	**Modelo clásico**	**Modelo con Kernels**	**Modelo con Splines**
α=0.5	0.247	**0.067**	0.259	0.250	0.279	**0.247**	**0.249**	0.253	0.252
α=0.7	**0.247**	0.705	0.251	0.249	**0.222**	0.243	**0.248**	0.288	0.251
α=0.9	**0.244**	4.034	0.264	0.248	0.367	**0.236**	0.248	**0.062**	0.245
α=0.95	**0.224**	6.502	0.229	0.243	2.165	**0.224**	0.246	0.475	**0.231**
α=0.99	**0.199**	9.738	0.675	**0.221**	9.045	0.421	**0.228**	6.935	0.438
α=0.995	**0.226**	10.125	0.358	**0.204**	11.462	0.147	0.216	9.736	**0.187**
x= (0.0)									
α=0.5	**0.000**	**0.000**	0.001	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
α=0.7	**0.000**	0.044	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	0.002	**0.000**
α=0.9	**0.001**	0.882	0.003	**0.000**	0.008	0.001	**0.000**	0.002	**0.000**
α=0.95	**0.004**	2.063	0.007	**0.000**	0.234	0.003	**0.000**	0.013	0.001
α=0.99	**0.005**	3.169	0.350	**0.004**	3.324	0.007	**0.002**	2.016	0.090
α=0.995	**0.016**	3.369	0.128	**0.009**	4.675	0.041	**0.004**	3.386	0.014
x= (0.5 . 0.5)									
α=0.5	0.249	**0.060**	0.221	**0.250**	0.339	0.258	0.249	0.262	**0.242**
α=0.7	0.251	**0.117**	0.242	**0.250**	0.354	0.268	0.249	0.293	**0.242**
α=0.9	0.259	**0.196**	0.212	**0.248**	0.334	0.284	**0.247**	0.335	0.255
α=0.95	0.294	0.399	**0.290**	**0.251**	0.367	0.316	**0.245**	0.359	0.288
α=0.99	0.382	0.298	**0.135**	0.254	0.759	**0.012**	0.259	0.514	**0.001**
α=0.995	0.225	0.319	**0.200**	**0.247**	1.113	0.491	**0.259**	0.726	0.364

Tabla 4: Comparación error cuadrático medio de la predicción para Pareto (1,8)

	n=100			n=500			n=1000		
x= (-0.5 . -0.5)	**Modelo clásico**	**Modelo con Kernels**	**Modelo con Splines**	**Modelo clásico**	**Modelo con Kernels**	**Modelo con Splines**	**Modelo clásico**	**Modelo con Kernels**	**Modelo con Splines**
α=0.5	0.236	**0.016**	0.294	**0.249**	0.388	0.265	0.251	**0.202**	0.256
α=0.7	**0.221**	2.383	0.246	0.250	0.261	**0.250**	**0.249**	0.265	0.286
α=0.9	0.264	4.982	**0.205**	0.218	0.357	**0.211**	**0.232**	0.249	0.291
α=0.95	**1.328**	13.679	1.893	**0.261**	2.945	0.344	**0.209**	0.211	0.242
α=0.99	28.849	39.733	**2.368**	**2.650**	5.796	31.143	**0.833**	23.953	22.418
α=0.995	**18.014**	40.839	46.109	49.988	76.611	**48.461**	**8.198**	41.998	13.803
x= (0.0)									
α=0.5	**0.003**	0.043	0.013	**0.000**	**0.000**	0.003	**0.000**	0.014	0.002
α=0.7	**0.012**	0.101	0.026	**0.001**	0.029	0.008	**0.001**	**0.001**	0.009
α=0.9	**0.168**	0.292	0.668	**0.026**	0.050	0.084	**0.006**	0.012	0.031
α=0.95	**0.845**	2.295	3.145	**0.059**	0.258	0.654	**0.040**	0.048	0.173
α=0.99	58.552	11.620	**4.228**	**1.517**	11.467	86.294	1.471	**0.008**	59.467
α=0.995	12.238	**12.174**	120.016	**27.319**	35.233	44.647	9.481	**0.450**	20.386
x= (0.5 . 0.5)									
α=0.5	0.289	**0.014**	0.202	0.251	0.276	**0.230**	0.258	**0.191**	0.262
α=0.7	0.361	**0.003**	0.357	0.248	**0.196**	0.268	0.257	**0.092**	0.192
α=0.9	0.912	**0.577**	2.529	**0.366**	0.476	0.546	0.287	**0.014**	0.325
α=0.95	3.831	**0.099**	9.019	**0.539**	0.944	1.949	0.439	**0.329**	0.957
α=0.99	172.625	**0.296**	7.152	**9.504**	91.482	170.139	**4.407**	24.587	113.798
α=0.995	75.578	**0.360**	305.401	**37.083**	157.267	92.349	25.617	**23.667**	42.466

Tabla 5: Comparación error cuadrático medio de la predicción para Pareto (1,2)

	n=100			n=500			n=1000		
	Modelo clásico	**Modelo con Kernels**	**Modelo con Splines**	**Modelo clásico**	**Modelo con Kernels**	**Modelo con Splines**	**Modelo clásico**	**Modelo con Kernels**	**Modelo con Splines**
x= (-0.5 . -0.5)									
α=0.5	0.245	**0.002**	0.068	0.249	0.415	**0.233**	**0.255**	0.354	0.294
α=0.7	0.258	**0.068**	0.232	0.245	**0.210**	0.256	**0.258**	0.367	0.289
α=0.9	0.310	6.974	**0.205**	**0.258**	0.962	0.422	0.265	**0.257**	0.379
α=0.95	**0.596**	9.527	0.824	**0.270**	8.621	0.355	**0.233**	2.265	0.281
α=0.99	26.146	**7.015**	12.043	**1.358**	84.636	15.472	**0.514**	49.914	17.507
α=0.995	13.449	**3.609**	25.524	**3.677**	91.335	5.346	**2.520**	69.621	4.001
x= (0.0)									
α=0.5	0.009	**0.008**	0.053	**0.001**	0.018	0.010	**0.000**	0.001	0.011
α=0.7	**0.025**	0.068	0.078	**0.003**	0.072	0.017	0.001	**0.000**	0.014
α=0.9	**0.119**	0.845	0.595	**0.024**	0.045	0.158	0.014	**0.003**	0.097
α=0.95	0.358	**0.093**	1.436	**0.100**	0.857	0.333	**0.063**	0.693	0.263
α=0.99	54.283	**3.462**	31.003	**1.334**	18.873	39.252	**0.420**	5.120	43.913
α=0.995	16.868	**11.378**	54.208	**2.934**	12.372	6.847	**3.304**	6.061	6.925
x= (0.5 . 0.5)									
α=0.5	0.354	**0.015**	0.667	0.270	**0.153**	0.352	0.260	**0.100**	0.190
α=0.7	0.435	**0.117**	0.652	0.274	**0.000**	0.326	0.252	**0.027**	0.245
α=0.9	0.730	**0.402**	2.433	0.343	**0.238**	0.353	0.338	**0.002**	0.417
α=0.95	**1.758**	6.690	4.285	0.729	**0.613**	1.178	0.603	**0.468**	1.289
α=0.99	124.957	**43.051**	59.375	4.540	**0.081**	73.942	**2.166**	5.830	81.875
α=0.995	**49.926**	79.750	126.376	8.471	**6.361**	15.418	**10.108**	9.132	16.745

Tabla 6: Comparación error cuadrático medio de la predicción para Log-normal (0,1)

Previamente al análisis de los resultados, es importante destacar que el modelo de Koenker y Bassett (1978) parte con ventaja ya que es un modelo lineal de igual forma que el modelo teórico.

Para los datos generados mediante una distribución de Pareto(1,2), en 28 de las 54 estimaciones realizadas el mejor modelo ha sido el de Koenker y Gassett (1978). La mayoría de las estimaciones donde este modelo se ha comportado mejor que los demás ha sido cuando el tamaño muestral es de 100 y 500. Bajo un tamaño muestral de 1000 el modelo basado en splines toma las mejores predicciones para algunos cuantiles. Si nos centramos en los más extremos, a partir del 0.9, el número de estimaciones en que el modelo basado en splines es mejor aumenta respecto a los cuantiles más bajos. En ningún caso para los cuantiles extremos el modelo basado en la estimación núcleo ha sido el mejor posible en predicción.

Para los datos generados mediante una distribución de Pareto(1,8), el modelo de regresión cuantílica vuelve a ser el más óptimo en estimaciones siendo mejor en 31 de las 54. Bajo esta distribución el segundo mejor modelo es el basado en estimación núcleo, el cual tiene una mejor eficacia especialmente cuando las distribuciones se realizan con un tamaño muestral igual a 1000 observaciones. Si nos centramos en los cuantiles extremos, vemos claramente un aumento de los casos exitosos por parte de las estimaciones realizadas con el método de estimación basado en la estimación núcleo.

Finalmente, para la distribución con datos generados mediante una distribución Log-normal(0,1), el modelo clásico es el mejor en 26 de las 54 estimaciones. Además, de igual manera que en los datos basados en la distribución de Pareto(1,8), para los cuantiles más extremos el método de estimación núcleo ganan una mayor proporción de casos en las que su estimación es mejor.

Estos resultados permiten afirmar claramente, que el modelo clásico de regresión cuantílica sufre de problemas para llevar a cabo una estimación correcta de los parámetros de los modelos para cuantiles muy extremos bajo distribuciones con colas pesadas. Esto supone un problema para la mayoría de cuantiles de interés en el ámbito de la cuantificación de riesgos, ya que los estos suelen situarse entre 0.99 y 0.999. Si introducimos las nuevas generaciones de modelos de cuantiles basados en métodos no paramétricos, no vemos una clara mejora en los resultados obtenidos en términos de

predicción de la variable dependiente. La degradación de la predicción persiste para los modelos no paramétricos planteados en éste trabajo.

Por lo tanto, un reto importante es proporcionar métodos alternativos que nos permitan mejorar la estimación de los cuantiles condicionales para niveles de confianza α próximos a 1 y con distribución de cola pesada.

4. Aplicación a datos reales

4.1. Descripción de los datos

A lo largo de este apartado se trabajará con unos datos de costes reales obtenidos de una empresa del sector asegurador ubicada en el mercado español. Estos datos ya han sido utilizados previamente en otros estudios, como por ejemplo en Bolancé y Vernic (2019). El número total de observaciones de la muestra inicial es de 162 clientes que poseen tanto pólizas de seguros de automóvil como de hogar vigentes en algún momento del período comprendido entre 2006-2015.

Con el objetivo final de analizar la elación entre los costes, o potencial riesgo de un cliente, y algunas variables explicativas se han analizado las dos variables siguientes: el coste de siniestros por automóvil anualizado y el coste de siniestros de hogar anualizados. Para obtener estas variables se ha computado el total de costes por cliente divido por el número total de días de exposición, vigencia de la póliza, y anualizado multiplicando por 365 días, es decir $\frac{Coste\ total}{Días\ de\ exposición}\ x\ 365\ días$. De esta forma, podemos obtener una medida para el coste e individuo independiente del tiempo de exposición total, ya que un cliente que ha tenido activa más tiempo una póliza ha podido potencialmente general más costes o siniestros o, al contrario, un cliente con un siniestro ha podido ser expulsado de la póliza y su período de exposición sea potencialmente más breve. Además, es necesario destacar que el coste total para las pólizas de automóvil se compone de la suma tanto de daños materiales como daños corporales, mientras que para pólizas de hogar solo se consideran los daños materiales dado que en este caso los daños corporales apenas existen.

Del total de la muestra, únicamente han registrado costes en relación a pólizas de automóvil un total de 12,203 observaciones y para costes relacionados con pólizas de hogar un total de 11,647 observaciones. Estas dos sub-muestras son las que se emplearán en las estimaciones que se muestran posteriormente.

La representación gráfica de la distribución de los costes asociados a pólizas de automóvil final es:

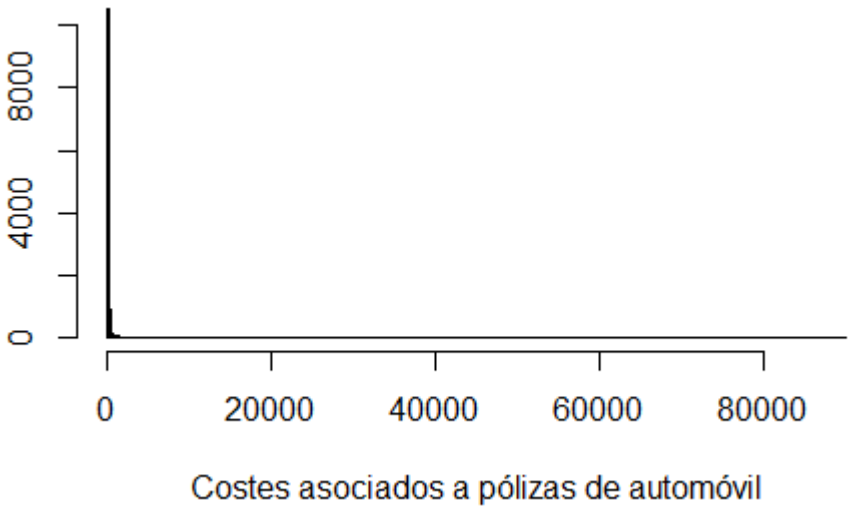

Ilustración 10: Distribución costes asociados a pólizas de automóvil

Como vemos los costes asociados a pólizas de seguro de automóvil tienen una cola muy pesada. A continuación, se muestra la relación de valores por cuantiles:

Cuantil	Coste €	Cuantil	Coste €
Mínimo	0.73	60%	70.33
5%	16.89	65%	79.73
10%	22.08	70%	90.44
15%	26.31	75%	110.46
20%	29.99	80%	137.15
25%	33.91	85%	175.32
30%	37.79	90%	258.71
35%	42.07	95%	473.01
40%	45.43	99%	1,735.03
45%	50.76	99.5%	2,824.82
50%	56.74	99.9%	7,019.79
55%	62.80	Máximo	8,9642.99

Tabla 7: Cuantiles costes asociados a las pólizas de automóvil

Para los costes asociados a pólizas de hogar observamos gráficamente la siguiente distribución:

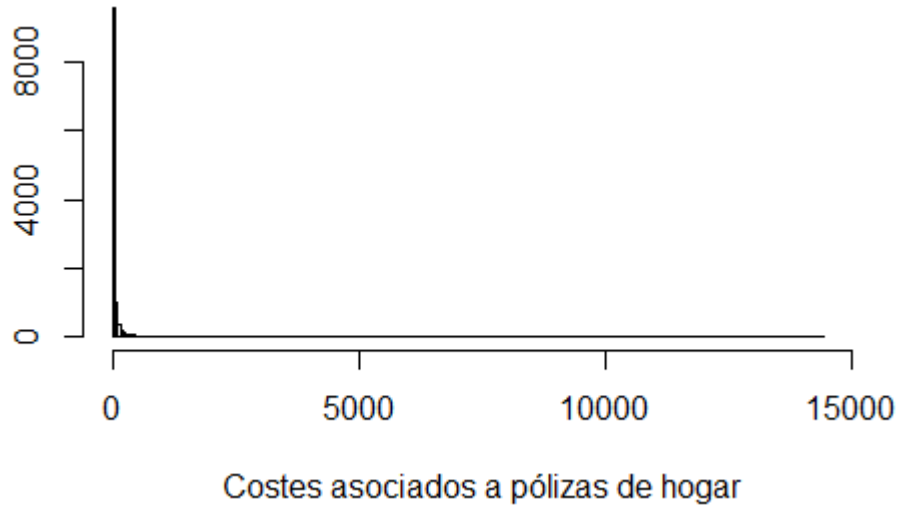

Ilustración 11: Distribución costes asociados a pólizas de hogar

Observamos que al igual que los costes de pólizas de automóvil los costes tienen una cola muy pesada, aunque algo menor que en el caso de seguros de automóvil. Esto se puede ver en la siguiente representación de los cuantiles:

Cuantil	Coste €	Cuantil	Coste €
Mínimo	0.034	60%	19.79
5%	2.95	65%	23.42
10%	4.19	70%	28.37
15%	5.38	75%	34.99
20%	6.38	80%	43.59
25%	7.39	85%	58.35
30%	8.7	90%	88.05
35%	10.33	95%	171.01
40%	12.08	99%	593.88
45%	13.39	99.5%	1,113.89
50%	14.79	99.9%	3,316.62
55%	16.67	Máximo	14,396.06

Tabla 8: Cuantiles costes asociados a las pólizas de automóvil

Las variables explicativas, o covariables, de las que se dispone son: el género, edad del tomador del seguro, edad del primer conductor (utilizada únicamente para los costes relacionados con pólizas de automóvil), si el tomador dispone de otras pólizas y las variables relacionadas con el lugar de residencia del cliente. La primera variable de residencia distingue entre residencia en ciudades grandes. Barcelona y Madrid, respecto al resto de ciudades; la segunda variable representa si el cliente reside en el norte o no; y

la tercera representa al resto del país. Estas variables han sido cogidas del último registro disponible del cliente a lo largo del período disponible entre 2006 y 2015.

4.2 Análisis de datos para las pólizas de automóvil

Como primer paso sobre el total de 12,203 observaciones con costes asociados a pólizas de automóvil se han eliminado 3 observaciones que no tenían informado el campo de edad del primer conductor del vehículo ni si disponía de otras pólizas o no. Posteriormente, se han obtenido los estadísticos descriptivos básicos para las covariables cuantitativas. A continuación, en la Tabla 9 se muestran estos descriptivos para nuestros datos:

	Edad cliente	Edad del primer conductor	Costes de pólizas de auto
Media	53.98	30.57	158.93
Desviación estándar	11.88	9.44	975.40
Mediana	53.47	30.60	57.46
Mínimo	-0.12	3.00	0.73
Máximo	339.67	64.00	89,642.99
Asimetría	1.49	0.05	67.01
Kurtosis	27.93	-0.30	5,875.73

Tabla 9: Descriptivos variables cualtitativas para costes de pólizas de automóvil

Se ha detecta que en los datos originales la edad del tomador del seguro comprende el rango de valores entre -0.12 y 339.67. Como es lógico, no puede haber clientes con una edad negativa o con una edad muy superior a los 120 años de edad, por lo tanto se ha determinado cortar los datos entre 16 y 100 años. Esto supone eliminar 21 observaciones con valores superiores a los 100 años de edad y 7 observaciones con edad inferior a los 16 años del tomar del seguro. La distribución final de la variable se puede observar en el histograma que se muestra en la Ilustración 12:

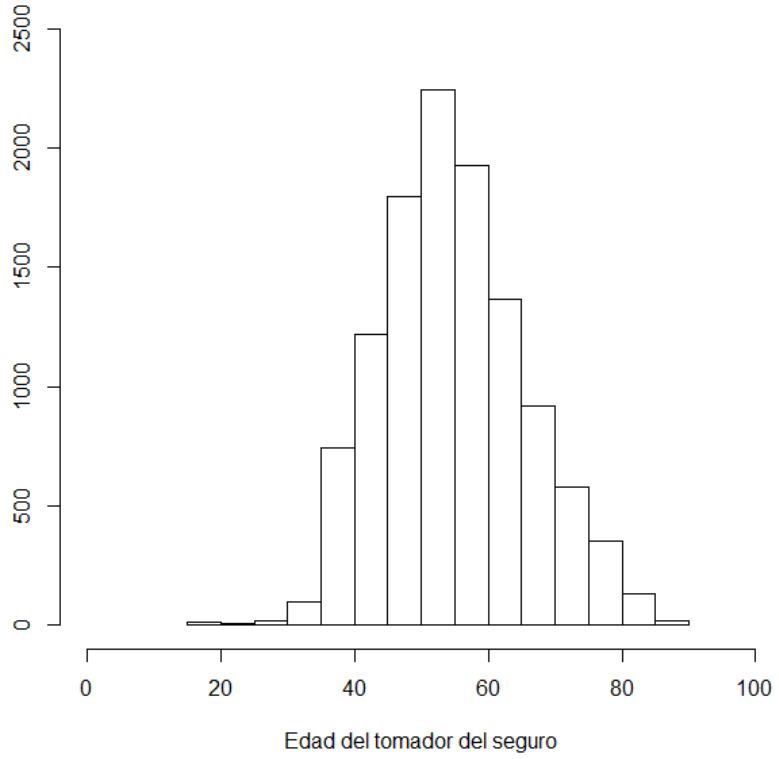

Ilustración 12: Distribución de la edad del tomador del seguro después de limpieza

Después de estas modificaciones los datos contienen observaciones donde la edad del primer conductor es inferior a 16 años. Durante el período vigente solo se podían conducir algunos vehículos a partir de los 16 años de edad y nunca antes, por lo tanto se han eliminado un total de 761 observaciones con una edad inferior a 16 años. Esto supone que el número de observaciones final es de 11,411. La distribución final de la variable se puede observar en la Ilustración 13:

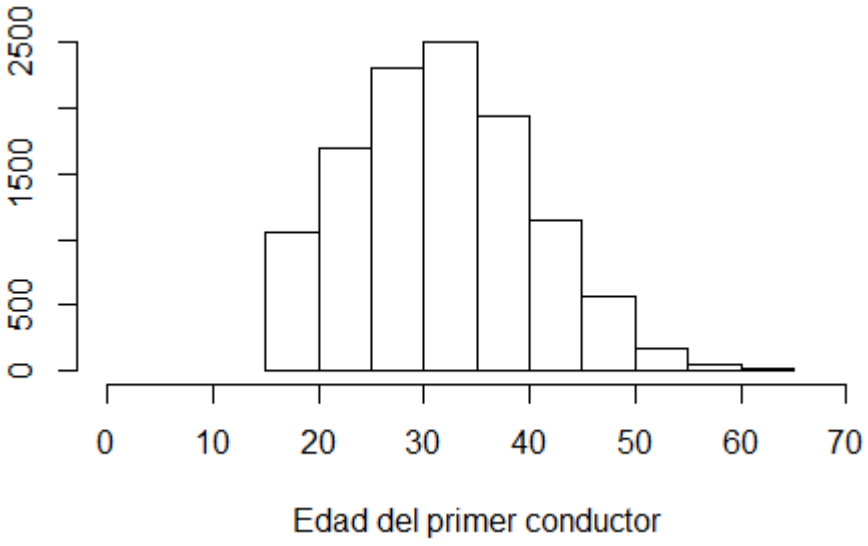

Ilustración 13: Distribución de la edad del primer conductor del seguro después de limpieza

A continuación, en la Tabla 10 se muestran, para las 3 variables cualitativas de nuestros datos, los descriptivos básicos después de las modificaciones:

	Edad cliente	Edad del primer conductor	Costes de pólizas de auto
Media	54.97	31.76	157.41
Desviación estándar	10.65	8.49	1.002.58
Mediana	54.20	31.50	56.73
Mínimo	16.50	16.00	0.73
Máximo	90.00	64.00	89.642.99
Asimetría	0.30	0.31	65.94
Kurtosis	-0.15	-0.35	5.628.02

Tabla 10: Descriptivos variables cualitativas para costes de pólizas de automóvil después de limpieza

Además, se puede analizar la matriz de correlaciones de estas tres variables, que se han mostrado en la Tabla 4:

	Edad del primer conductor	Edad del cliente	Costes de pólizas de auto
Edad del primer conductor	1.00	0.76	0.00
Edad del cliente	0.76	1.00	-0.01
Costes de pólizas de auto	0.00	-0.01	1.00

Tabla 11: Matriz de correlaciones

En relación con la edad del tomador del seguro, la primera conclusión que podemos extraer rápidamente es la elevada correlación entre la edad del primer conductor y la edad del tomador del seguro, dicho valor es de 0.76. A causa de esta elevada correlación se ha decidido realizar dos regresiones distintas, una con la edad del tomar del seguro y sin la edad del primer conductor como variable explicativa y otra regresión con la edad del primer conductor y sin la edad del tomador del seguro.

A continuación, en la Ilustración 14 observamos que en relación al género en el que el 83.7% de los clientes son hombres, destaca la interacción entre la variable de género y de edad del primer conductor, ya que observamos que para las mujeres hay una proporción mayor con una edad menor.

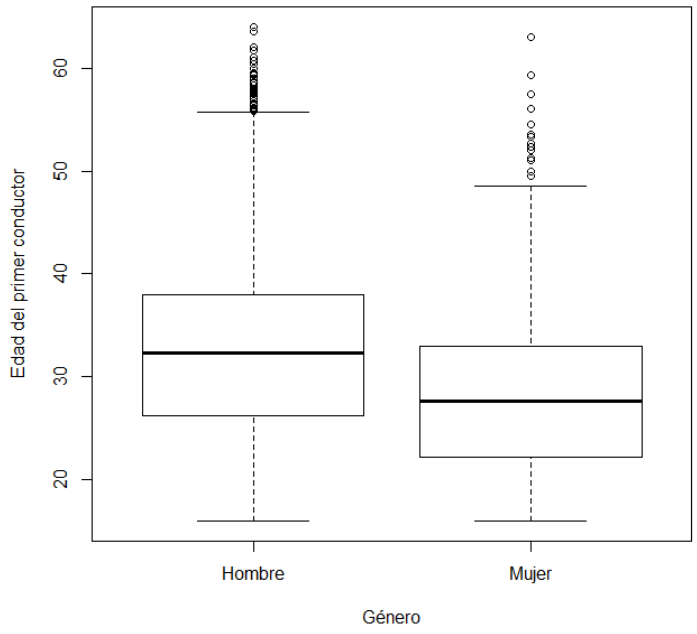

Ilustración 14: Relación género y edad del primer conductor para costes de automóvil

En la Ilustración 15 se observa que en cuanto a la relación con la edad del tomador del seguro la distribución es relativamente parecida para hombres y mujeres, aunque nuevamente se pueda apreciar una tendencia a una edad menor para las mujeres como se muestra a continuación:

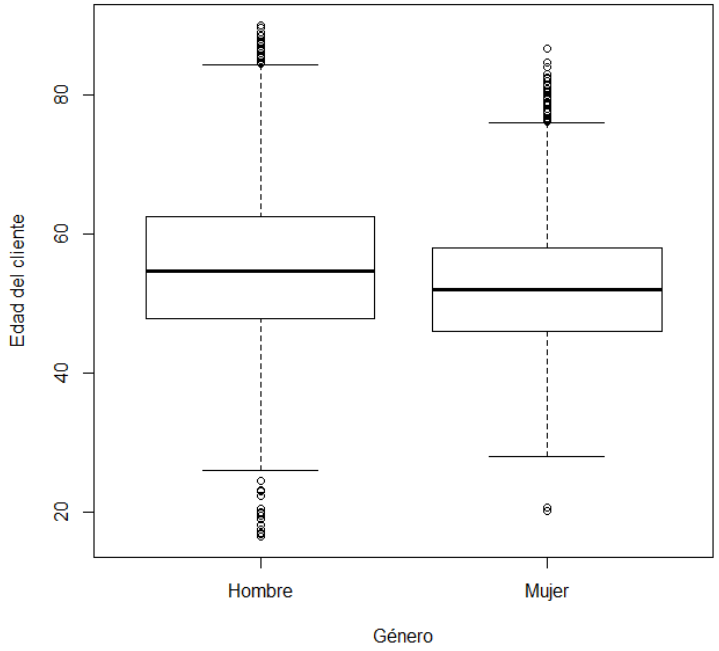

Ilustración 15: Relación género y edad del cliente para costes de automóvil

Finalmente, la relación con los costes en la Ilustración 16 parece que en promedio no existen diferencias entre los hombre y las mujeres.

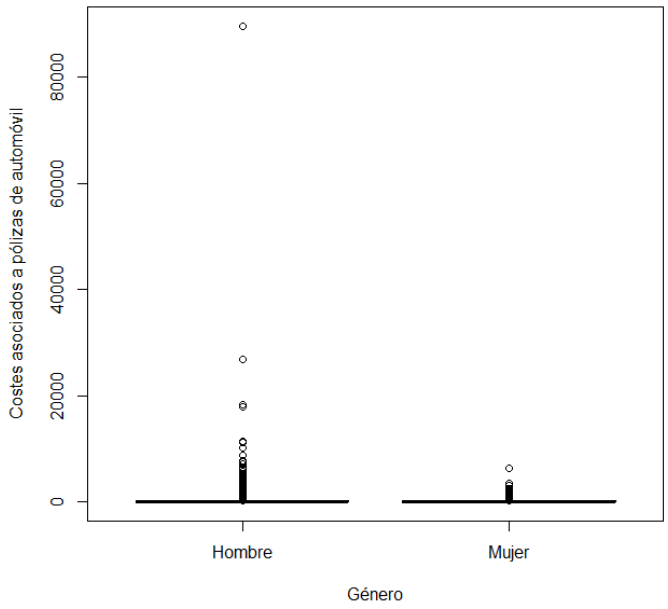

Ilustración 16: Relación género para costes de automóvil

En cuanto a si el cliente dispone de otras pólizas, podemos observar que la mayoría de los clientes no dispone ellas. En concreto el 79% de los clientes no dispone de otra póliza. En la Ilustración 17 podemos observar que aparentemente los clientes sin otras pólizas han podido producir unos costes mayores que los que sí disponen de otras pólizas. Principalmente. esta diferencia se observa en los valores más extremos.

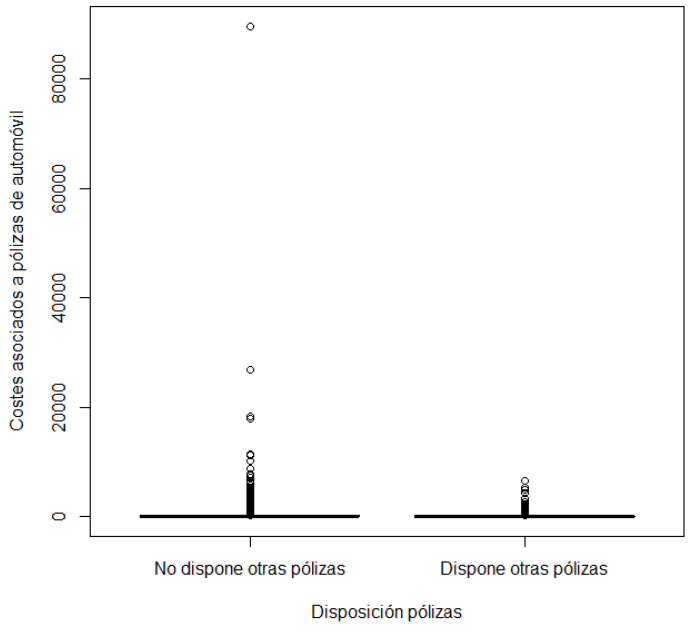

Ilustración 17: Costes pólizas de automóvil por disposición de pólizas

Posteriormente, si analizamos el lugar de residencia, primero podemos observar que más del 93% de la cartera de autos no reside en una gran ciudad. En la ilustración 18, vemos

que en promedio no es observan diferencias pero si se observan diferencias en los cuantiles extremos, donde los no residentes tienen un mayor coste.

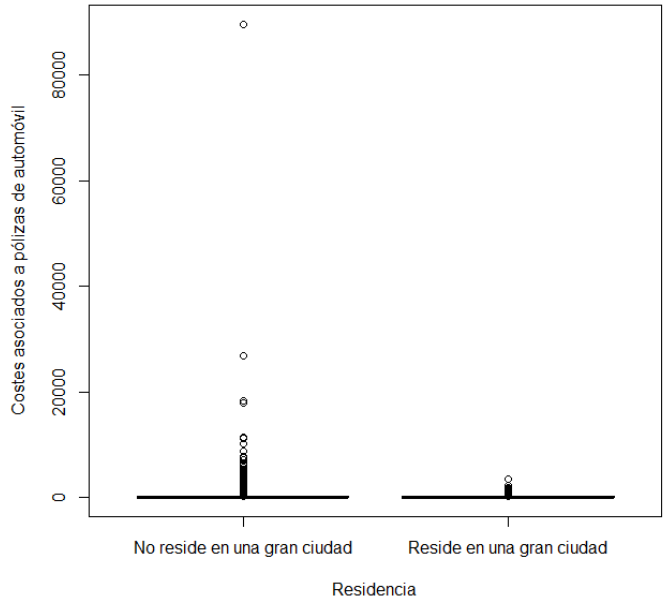

Ilustración 18: Costes pólizas de automóvil residentes grandes ciudades

La última variable a analizar corresponde a la residencia o no en el norte del territorio. En la ilustración 19, vemos que la distribución de costes es muy parecida para los residentes y no residente en el norte del territorio aunque se puede observar diferencias en los cuantiles extremos, donde los no residentes tienen un mayor coste.

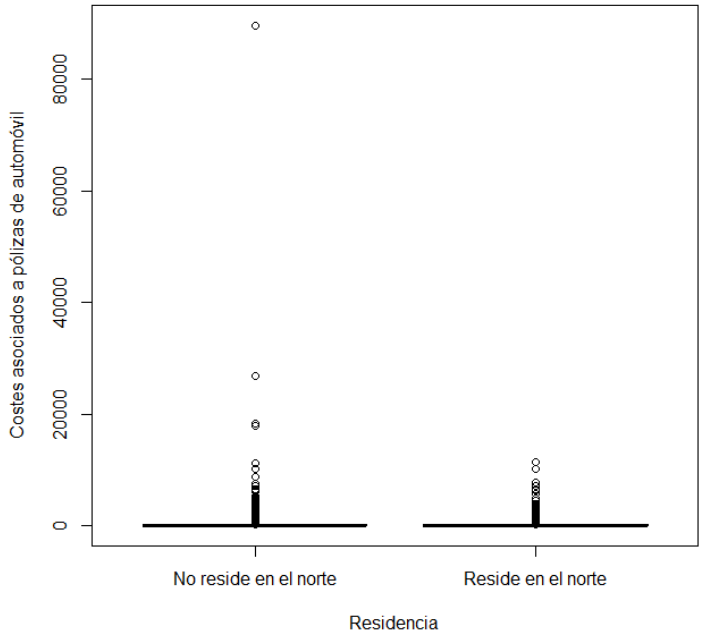

Ilustración 19: Costes pólizas de automóvil residentes norte

Si analizamos los datos mediantes el modelo de regresión cuantílica de Koenker y Bassett (1978), observamos las relaciones que se representan en la Ilustración 20 a lo largo de los cuantiles de los costes de pólizas de automóvil anualizados respecto a las variables descritas anteriormente para el modelo que excluye la edad del tomador del seguro.

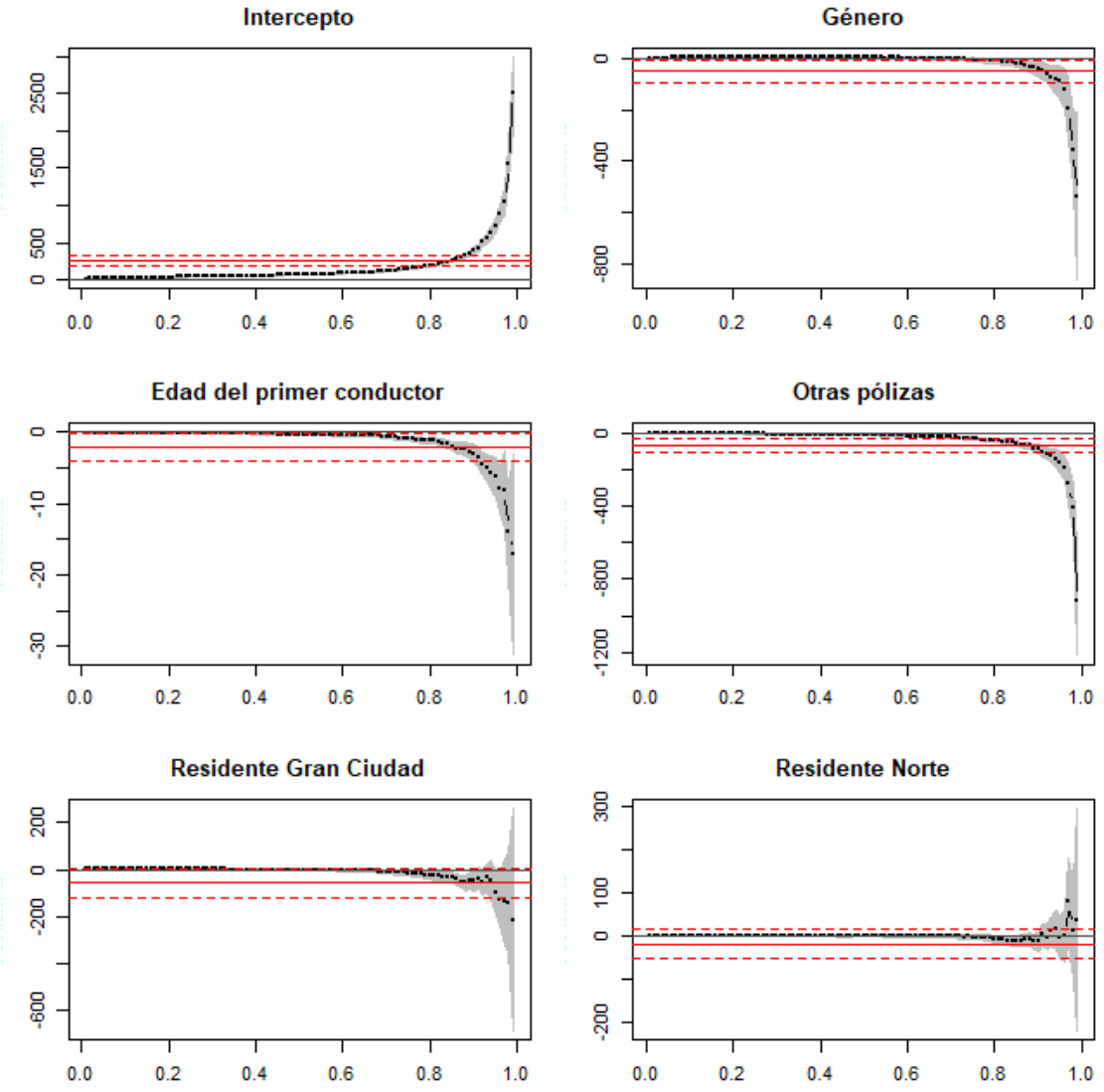

Ilustración 20: Relación de coeficientes por cuantiles para costes de automóvil con la edad del primer conductor

Observamos que para todas las variables a partir de los cuantiles más extremos hay un efecto diferente al efecto calculado por mínimos cuadrados que se encuentra indicado por la línea roja junto a las bandas de confianza por las líneas discontinuas rojas. Si nos centramos a partir del cuantil 0.90 observamos las relaciones representadas en la Ilustración 21.

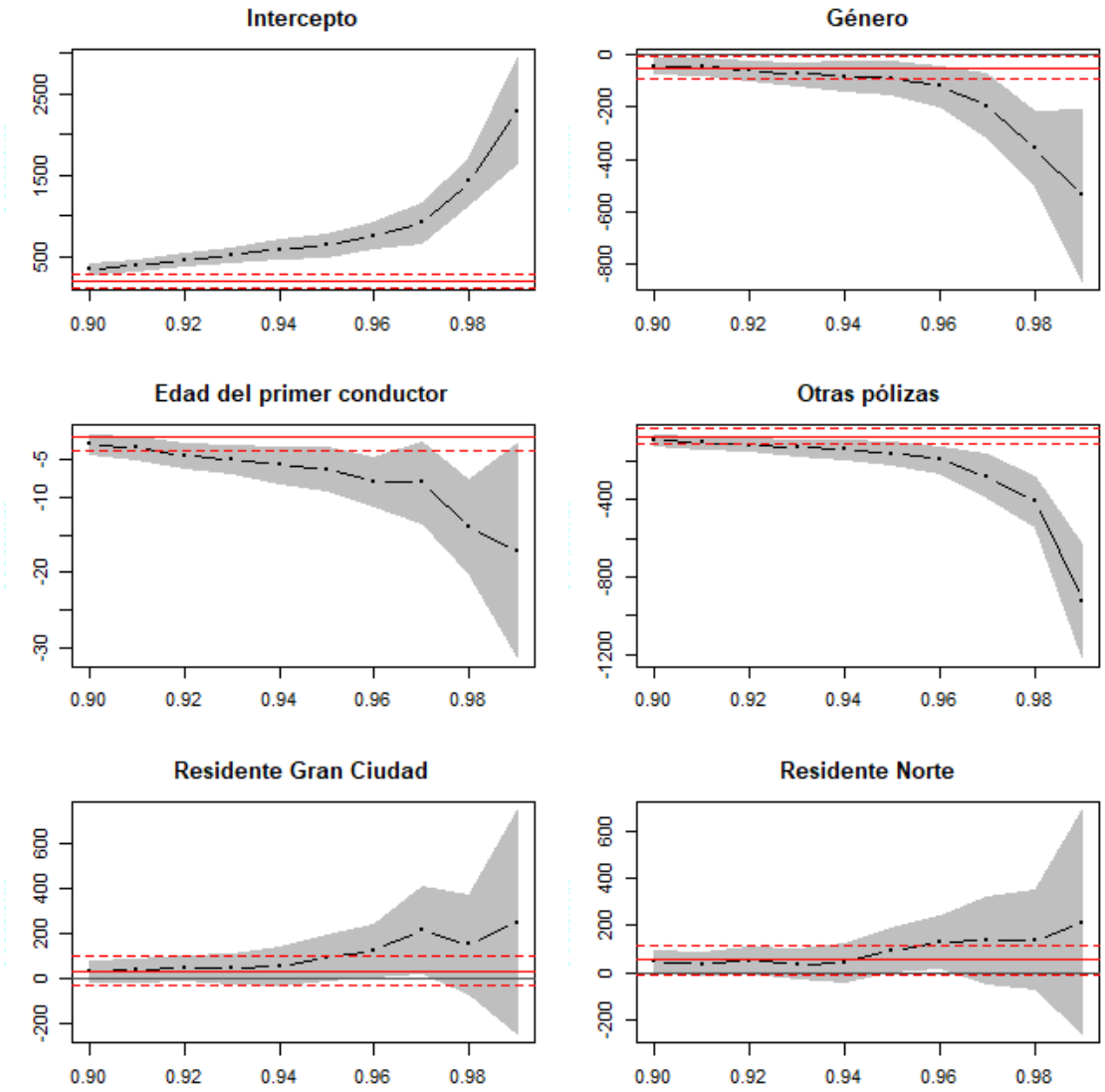

Ilustración 21: Relación de coeficientes por cuantiles para costes de automóvil con la edad del primer conductor a partir del cuantil 0.9

Visualmente se observa que el hecho de ser mujer tiene un efecto negativo superior al observado en la media sobre los costes, aproximadamente a partir del cuantil 96%. De la misma manera, a mayor edad del primer conductor el efecto negativo incrementa a partir del cuantil 92%. Tener otra póliza también tiene un efecto más negativo a partir del cuantil 92%. Para los efectos de zona se observa que los efectos no son muy diferentes a los que se producen a la media. Estos efectos no son estadísticamente significativos, ya que las bandas de confianza de la regresión cuantílica toman tanto valores positivos como negativos.

Si realizamos la misma regresión cuantílica pero reemplazamos la edad del primer conductor por la edad del tomador del seguro vemos las relaciones a lo largo de los cuantiles que se muestran en la ilustración 22.

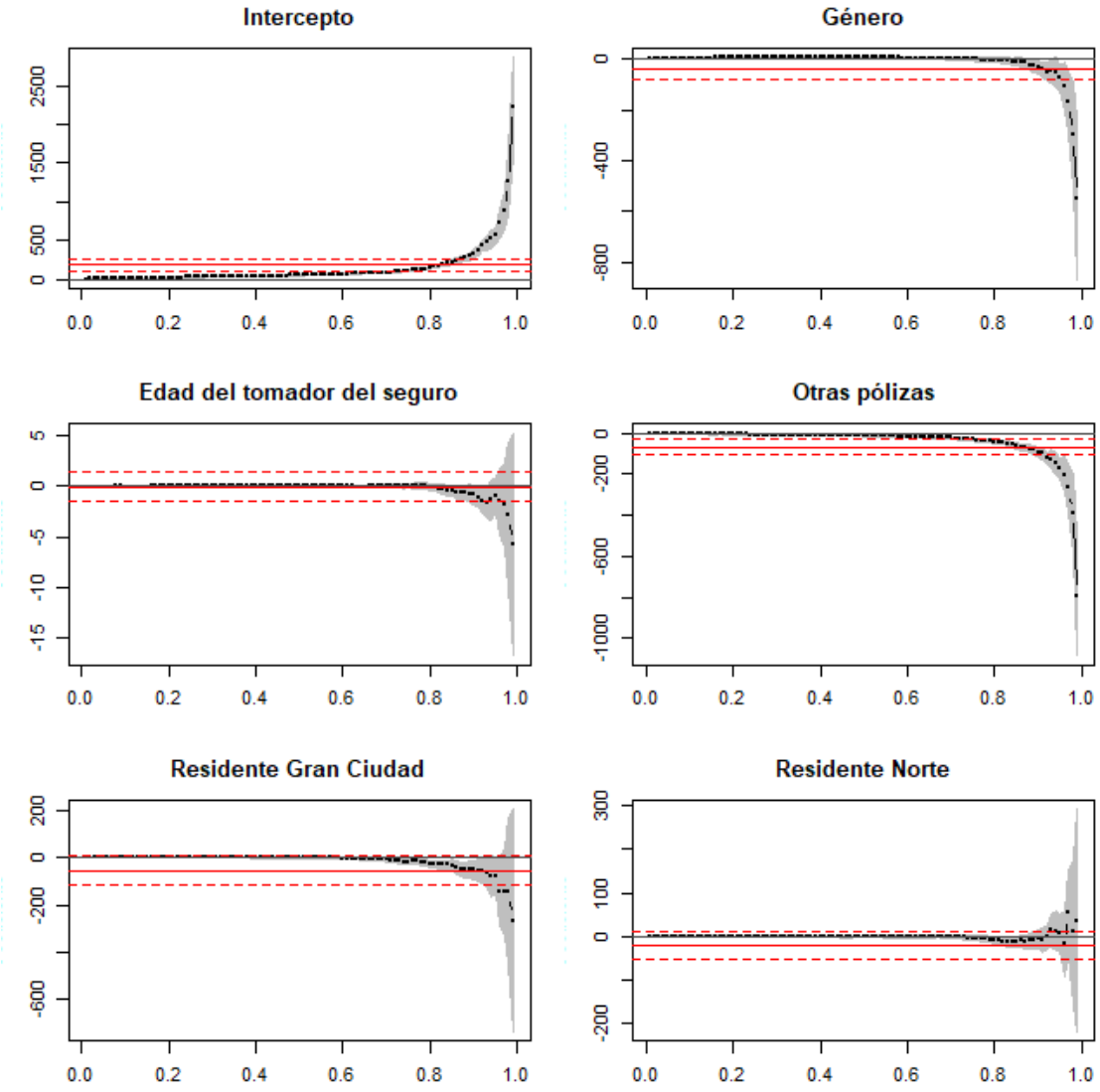

Ilustración 22: Relación de coeficientes por cuantiles para costes de automóvil con la edad del tomador

Observamos unas relaciones parecidas a las anteriores, donde la edad del tomar del seguro también tiene un efecto negativo sobre los costes. En la Ilustración 23 mostramos los resultados ampliados a partir del cuantil 90%.

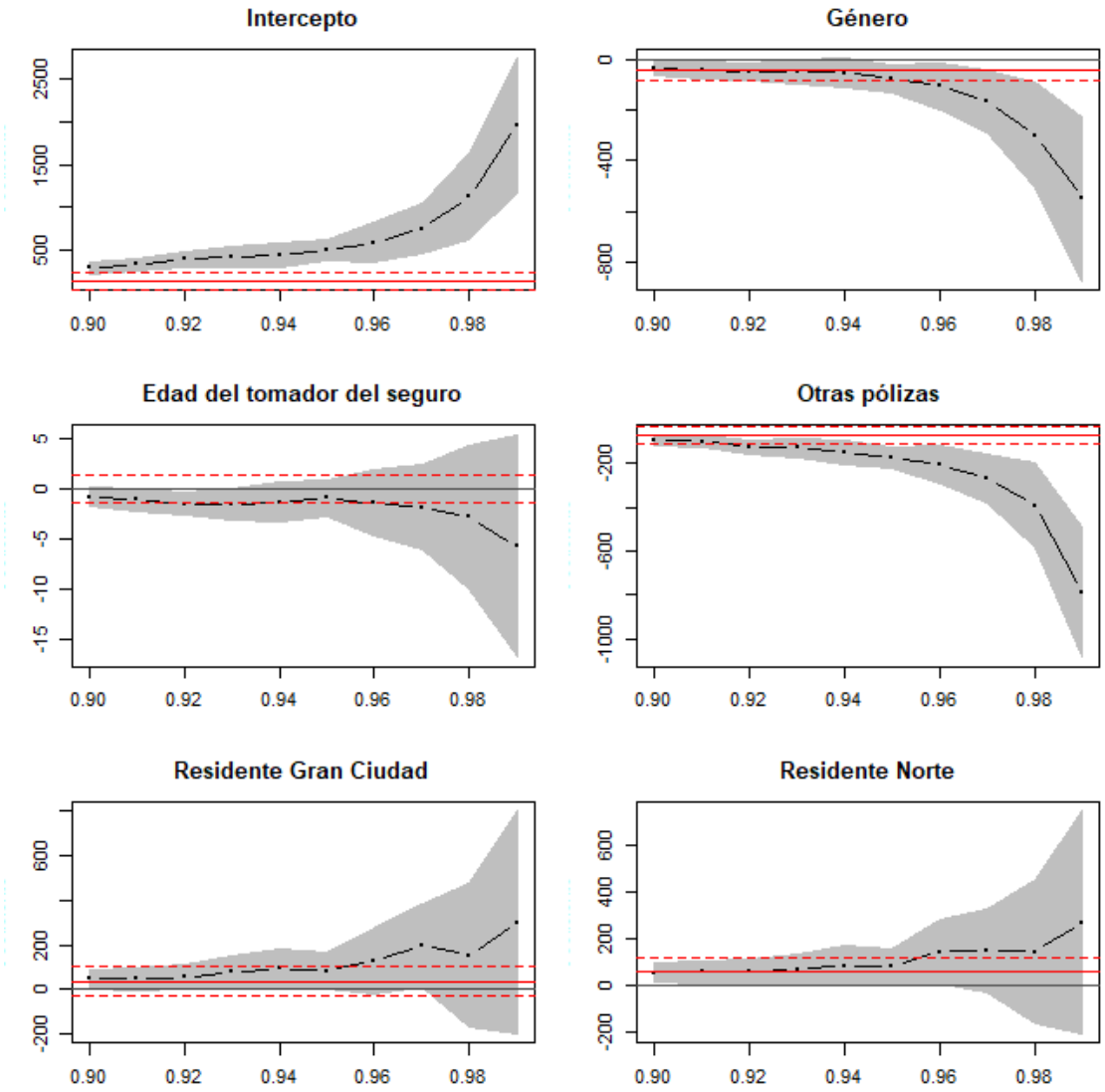

Ilustración 23: Relación de coeficientes por cuantiles para costes de automóvil con la edad del tomar a partir del cuantil 0.9

Vemos que la edad del tomador del seguro tiene un efecto negativo y diferenciado del efecto en la media a partir del cuantil 97%.

En la Tabla 12 se resumen los coeficientes por cuantil y modelo a partir del cuantil 95%. En negrita se destacan los efectos estadísticamente significativos.

	Modelo con edad conductor		Modelo con edad tomador	
	Coeficiente	**p-valor**	**Coeficiente**	**p-valor**
α=95%				
Intercepto	633.23	**0.00**	505.22	**0.00**
Género	-88.69	**0.02**	-75.75	**0.02**
Edad del tomador del seguro	-	-	-1.07	0.34
Edad del primer conductor	-6.27	**0.00**	-	-
Otras pólizas	-161.75	**0.00**	-168.11	**0.00**
Residente Gran Ciudad	92.83	0.12	82.36	0.10
Residente Norte	95.83	0.09	75.14	0.11
α=96%				
Intercepto	760.22	**0.00**	596.31	**0.00**
Género	-120.47	**0.01**	-104.52	0.08
Edad del tomador del seguro	-	-	-1.74	0.40
Edad del primer conductor	-8.01	**0.00**	-	-
Otras pólizas	-195.25	**0.00**	-202.95	**0.00**
Residente Gran Ciudad	125.70	0.07	133.22	0.15
Residente Norte	128.63	0.05	138.13	0.12
α=97%				
Intercepto	909.66	**0.00**	727.53	**0.00**
Género	-194.29	**0.01**	-186.50	**0.00**
Edad del tomador del seguro	-	-	-1.13	0.62
Edad del primer conductor	-8.07	**0.01**	-	-
Otras pólizas	-279.22	**0.00**	-275.55	**0.00**
Residente Gran Ciudad	215.18	0.06	197.15	0.08
Residente Norte	135.62	0.22	131.83	0.18
α=98%				
Intercepto	1426.00	**0.00**	1124.28	**0.00**
Género	-360.98	**0.00**	-300.75	**0.01**
Edad del tomador del seguro	-	-	-2.86	0.52
Edad del primer conductor	-13.96	**0.00**	-	-
Otras pólizas	-409.34	**0.00**	-389.83	**0.00**
Residente Gran Ciudad	152.02	0.25	152.69	0.44
Residente Norte	138.01	0.27	143.15	0.45
α=99%				
Intercepto	2287.51	**0.00**	1943.04	**0.00**
Género	-537.12	**0.00**	-551.92	**0.00**
Edad del tomador del seguro	-	-	-5.21	0.46
Edad del primer conductor	-17.10	**0.04**	-	-
Otras pólizas	-921.40	**0.00**	-789.49	**0.00**
Residente Gran Ciudad	249.63	0.40	302.88	0.32
Residente Norte	213.11	0.45	264.90	0.37

Tabla 12: Resultados modelo de regresión cuantílica sobre costes de pólizas de automóvil

Como se podía intuir en el análisis descriptivo, la variable de género tiene un efecto significativo en los costes, siendo estos menores para las mujeres en todos los cuantiles mostrados.

De igual forma que con el género, vemos claramente que a medida que aumenta la edad del conductor los costes disminuyen y se mantiene el efecto significativo para todos los cuantiles. El resultado de la edad del primer conductor es habitual y ampliamente conocido en los costes de pólizas de seguro, ya que hay una tendencia mayor a sufrir un accidente para los clientes más jóvenes.

Finalmente, disponer de otras pólizas reduce los costes producidos. El parámetro estimado asociado a esta variable se mantiene como significativa hasta el cuantil 98%. Disponer de otra póliza podría estar condicionado al propio comportamiento del cliente y de la compañía, es decir si el cliente no genera elevados costes, la compañía estará dispuesta a ofrecerle más productos, y como consecuencia tiene un impacto en los resultados.

Cuando analizamos el modelo con la variable edad del tomador del seguro podemos observar que su efecto no es significativo. Éste resultado es importante, ya que como hemos visto hay una alta correlación con la edad del primer conductor, la cual sí es significativa cuando se incluye en el modelo.

Como conclusión general para ambos modelos, destaca la poca significación del efecto de la zona de residencia.

4.3 Análisis de datos para las pólizas de hogar

Tras analizar el total de 11,646 observaciones con costes asociados a pólizas de hogar se ha eliminado una observación que no tenían informado el campo de edad del tomador del seguro ni si disponía de otras pólizas o no. Posteriormente, se ha comprobado los estadísticos descriptivos básicos para las covariables. A continuación, en la Tabla 13 se muestran los estadísticos descriptivos para las dos variables cualitativas relacionadas con los datos de seguro de hogar.

	Edad cliente	Costes de pólizas de hogar
Media	56.14	52.46
Desviación estándar	12.98	254.97
Mediana	55.33	14.80
Mínimo	-1.50	0.03
Máximo	115	14,396.06
Asimetría	0.34	27.59
Kurtosis	0.35	1,158.70

Tabla 13: Descriptivos variables cualitativas para costes de pólizas de hogar

Se ha detecta que en los datos originales la edad del tomador del seguro comprende el rango de valores entre -1.50 y 115. Como es lógico, no puede haber clientes con una edad negativa o con una edad muy superior a los 100 años de edad, por lo tanto se ha determinado cortar los datos entre 16 y 100 años. Esto supone eliminar 7 observaciones con edad inferior a los 16 años del tomar del seguro y 38 con una edad superior a 100 años. La distribución final de la variable se puede observar en la Ilustración 24.

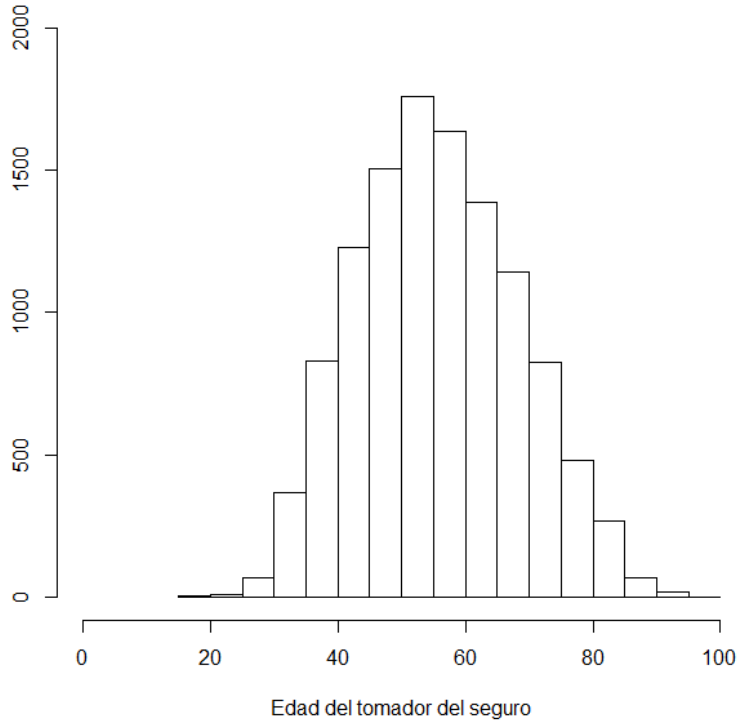

Ilustración 24: Distribución de la edad del tomador de póliza de hogar después de limpieza

A continuación, en la Tabla 14 se muestran para las 2 variables cualitativas de nuestros datos los descriptivos básicos después de las modificaciones:

	Edad cliente	Costes de pólizas de hogar
Media	56.00	52.54
Desviación estándar	12.56	255.43
Mediana	55.33	14.80
Mínimo	16.50	0.03
Máximo	96.50	14,396.06
Asimetría	0.18	27.52
Kurtosis	-0.48	1,154.87

Tabla 14: Descriptivos variables continúas para costes de pólizas de hogar después de limpieza

En relación al género, hemos observado que el 77.5% de los clientes son hombres. En cuanto a la relación con la edad del tomador del seguro la distribución es relativamente parecida para hombres y mujeres, en la Ilustración 25 nuevamente se puede apreciar cierta tendencia a una edad menor para las mujeres.

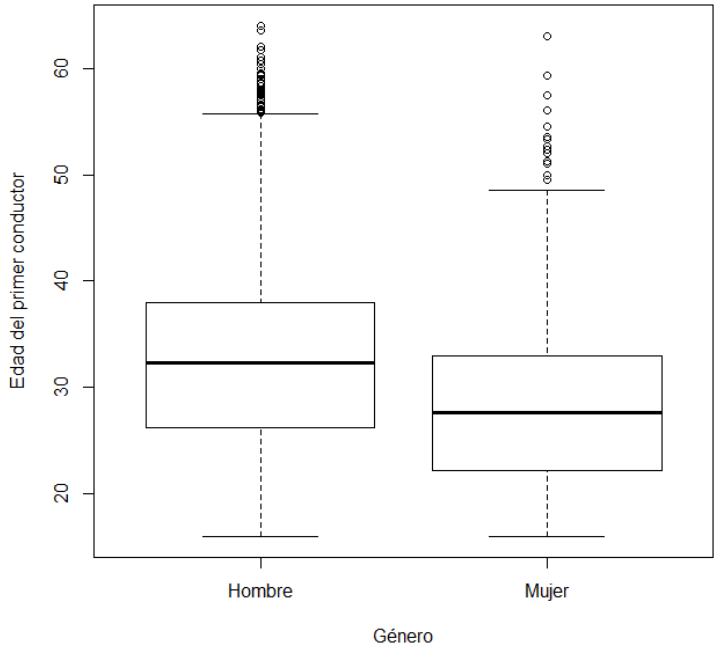

Ilustración 25: Relación género y edad del cliente para costes de hogar

Finalmente, en cuanto a la relación con los costes en la Ilustración 26 parece que los costes con mayores en los extremos para los hombres aunque se observa un valor muy extremo para las mujeres.

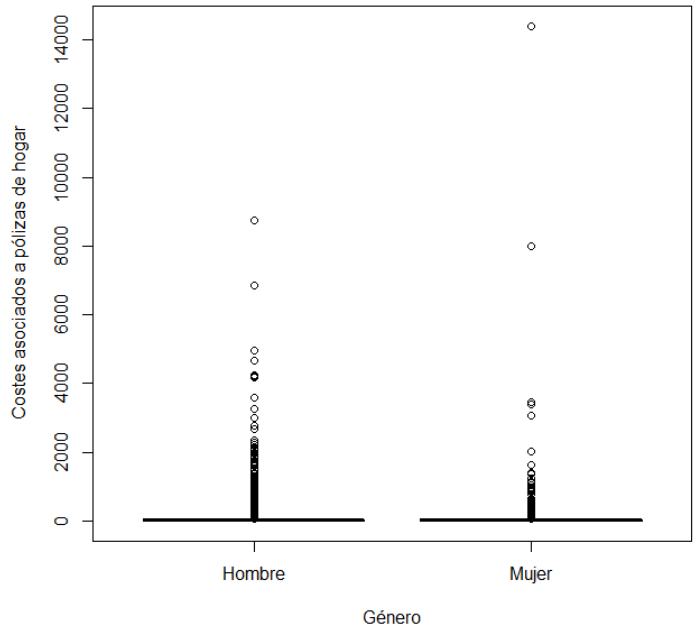

Ilustración 26: Relación género para costes de hogar

En cuanto a si el cliente dispone de otras pólizas, podemos observar que la mayoría de los clientes no dispone de otra póliza, en concreto el 81.5%. En la Ilustración 27, podemos

observar que aparentemente los clientes sin otras pólizas han podido producir unos costes más extremos que los que sí disponen de otras pólizas.

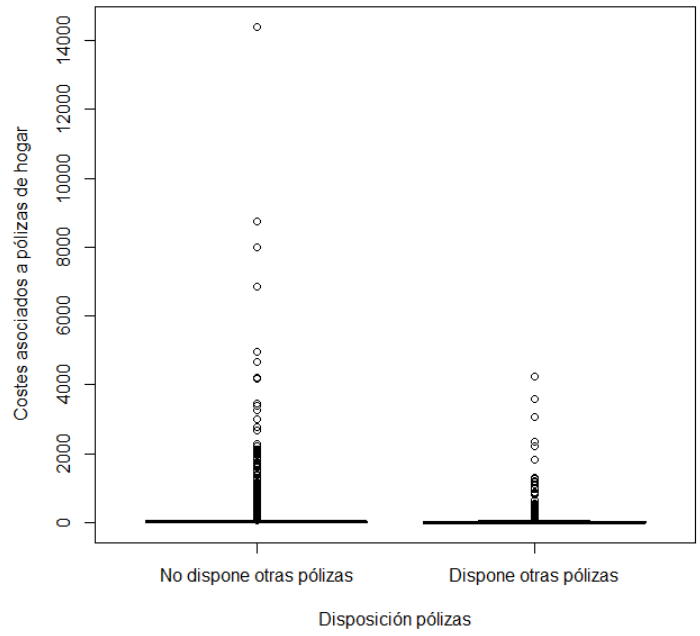

Ilustración 27: Costes pólizas de hogar por disposición de pólizas

Si analizamos el lugar de residencia, primero podemos observar que más del 79% de la cartera de hogar no reside en una gran ciudad. En la Ilustración 28 muestra como los costes son más extremos para los no residentes en una gran ciudad.

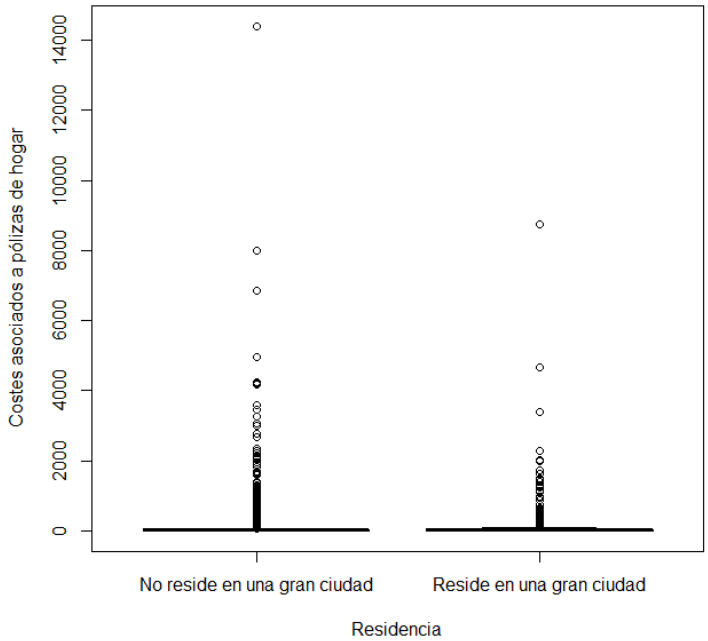

Ilustración 28: Costes pólizas de hogar residentes en una gran ciudad

La última variable a analizar corresponde a la residencia o no en el norte del territorio. En la Ilustración 29 vemos que la distribución de costes es muy parecida para los residentes y no residente en el norte del territorio.

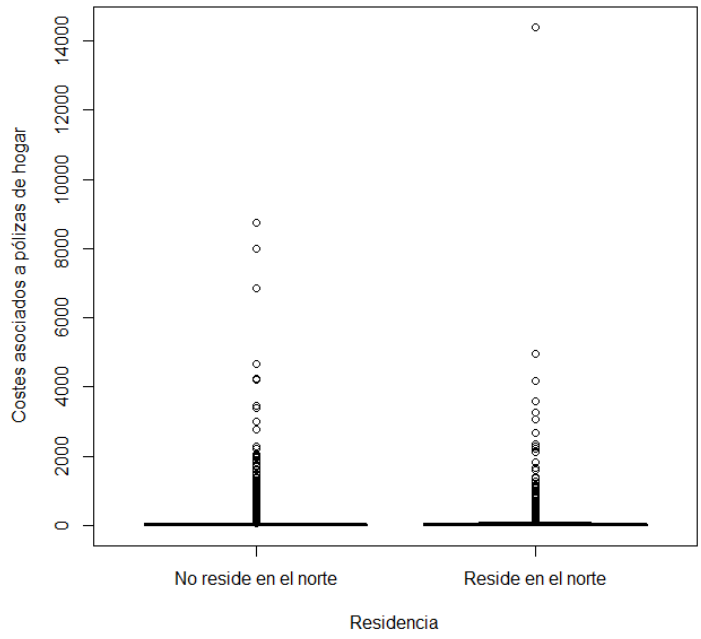

Ilustración 29: Costes pólizas de hogar residentes en el norte

Si analizamos los datos mediantes el modelo de regresión cuantílica de Koenker y Bassett (1978), observamos relaciones representadas en la Ilustración 30 a lo largo de los cuantiles de los costes de los siniestros de pólizas de hogar anualizados respecto a las variables descritas anteriormente.

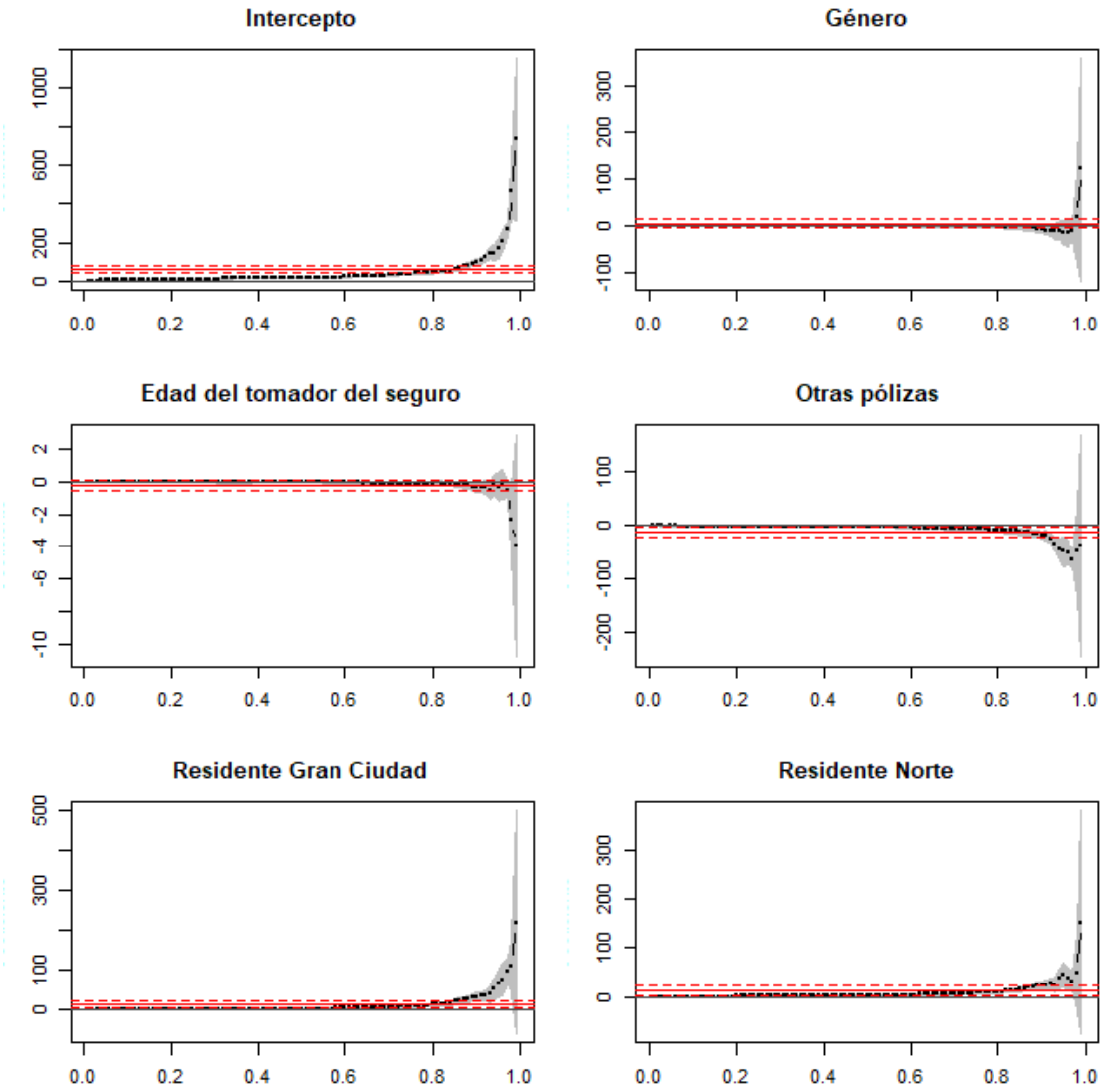

Ilustración 30: Relación de coeficientes por cuantiles para costes de hogar

Para todas las variables se observa que a partir de los cuantiles más extremos puede haber un efecto diferente al calculado por mínimos cuadrados, que se representa con la línea continua roja junto a sus las bandas de confianza, las líneas discontinuas rojas. Si realizamos la representación a partir del cuantil 0.90 observamos mejor los valores, tal y como se muestra en la Ilustración 31.

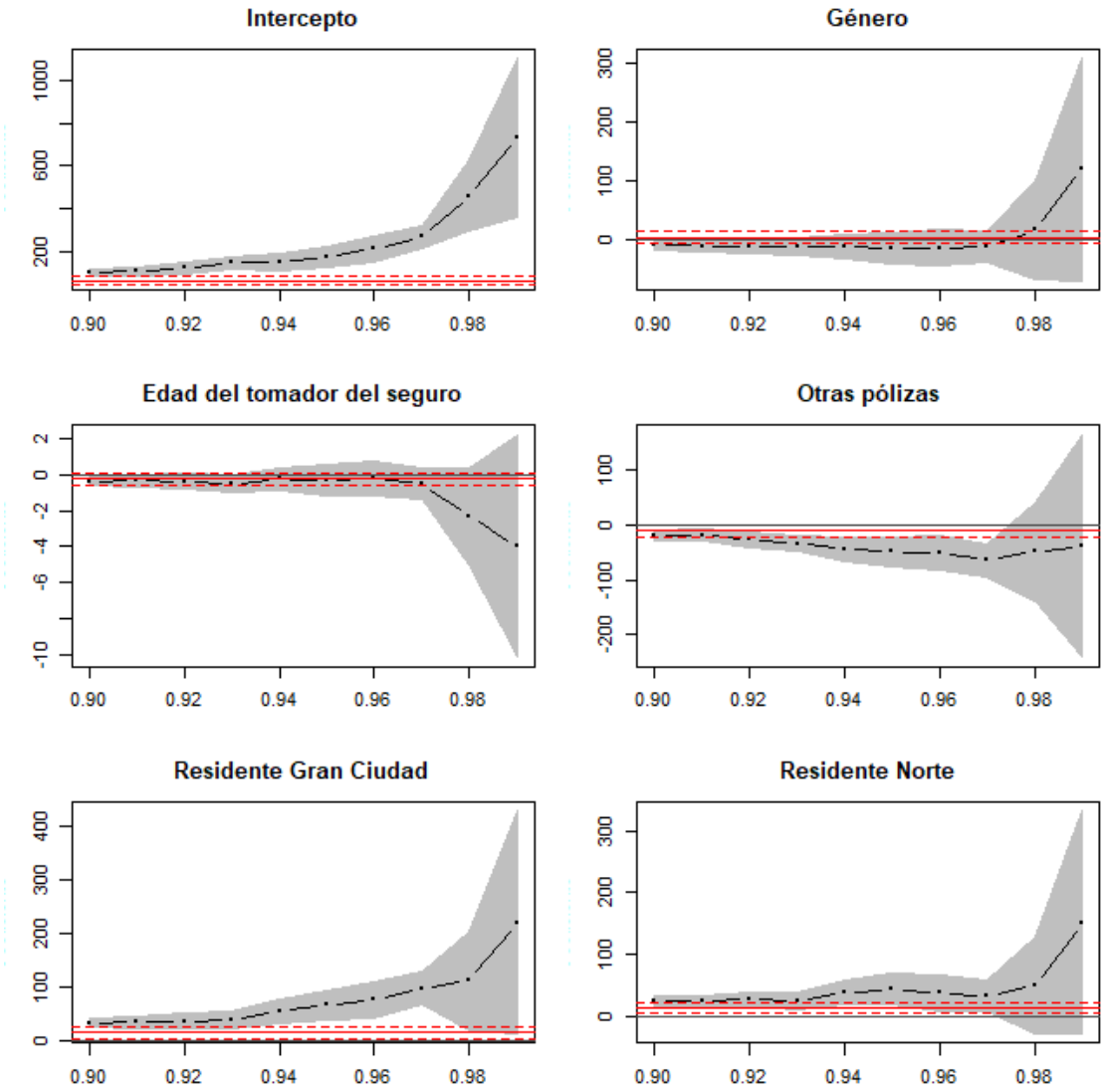

Ilustración 31: Relación de coeficientes por cuantiles para costes de hogar a partir del cuantil 0.9

En la Tabla 15 se resumen los coeficientes por cuantil y modelo a partir del cuantil 95% y hasta el 99%. En negrita se marcan los efectos que son estadísticamente significativos.

	Coeficiente	p-valor
α=95%		
Intercepto	165.65	**0.00**
Género	-12.23	0.45
Edad del tomador del seguro	-0.19	0.71
Otras pólizas	-46.11	**0.00**
Residente Gran Ciudad	67.54	**0.00**
Residente Norte	44.19	**0.00**
α=96%		
Intercepto	204.27	**0.00**
Género	-14.12	0.44
Edad del tomador del seguro	-0.15	0.80
Otras pólizas	-49.80	**0.01**
Residente Gran Ciudad	78.28	**0.00**
Residente Norte	38.14	**0.02**
α=97%		
Intercepto	257.32	**0.00**
Género	-7.99	0.64
Edad del tomador del seguro	-0.39	0.49
Otras pólizas	-65.20	**0.00**
Residente Gran Ciudad	98.66	**0.00**
Residente Norte	36.78	**0.02**
α=98%		
Intercepto	468.37	**0.00**
Género	17.30	0.72
Edad del tomador del seguro	-2.48	0.13
Otras pólizas	-48.98	0.36
Residente Gran Ciudad	113.96	**0.03**
Residente Norte	52.88	0.26
α=99%		
Intercepto	736.27	**0.00**
Género	123.99	0.29
Edad del tomador del seguro	-4.07	0.30
Otras pólizas	-41.79	0.74
Residente Gran Ciudad	220.19	0.09
Residente Norte	155.72	0.16

Tabla 15: Resultados modelo de regresión cuantílica sobre costes de pólizas de hogar

Para el modelo ajustado a los costes de hogar, observamos que a diferencia de lo que sucedía para los costes asociados a pólizas de automóvil, la distribución geográfica podría tener un efecto significativo. Vemos que residir en una gran ciudad o al norte del territorio incrementa los costes de los siniestros hasta el cuantil 98%. Este resultado podría estar relacionado con el hecho de que en una gran ciudad implica un valor del inmueble superior, y, por lo tanto, el coste a compensar será potencialmente superior.

Por otro lado, igual que en el modelo de costes de automóvil, la edad del tomador del seguro no resulta relevante para ningún cuantil. En cambio, en el caso del género el efecto sí resulta significativo para el modelo de costes de automóvil pero no para los del hogar.

Finalmente, se mantiene el hecho de que tener otras pólizas reduce el coste de una forma significativa hasta el cuantil 97%.

5. Conclusiones

Después del análisis realizado, podemos obtener diferentes conclusiones. En primer lugar, el modelo de regresión cuantílica tiene dos importantes ventajas respecto al modelo lineal en la media estimada por mínimos cuadrados ordinarios.

Desde un punto de vista teórico, el modelo de regresión cuantílica no se ve restringido a trabajar únicamente bajo datos con distribución normal. Como consecuencia, la regresión cuantílica puede ser aplicada a un gran número de análisis con datos no distribuidos de forma normal.

Desde un punto de vista práctico, el modelo de regresión cuantílica nos permite una mejor descripción de las relaciones entre las variables explicativas y variable dependiente, esto es debido a que podemos analizar cuáles son los factores influyentes en los valores fuera de la media.

El modelo de regresión cuantílica es por lo tanto una herramienta muy útil en la gestión y la cuantificación de riesgos. Pero es imprescindible confirmar la calidad de la estimación de los coeficientes bajo unos datos con distribución de cola pesada para una α extrema, característicos de la gestión y cuantificación de riesgos.

En el estudio realizado para medir la calidad de la estimación de los coeficientes, se demuestra que la cola de la distribución es más pesada, el modelo de regresión cuantílica sufre un deterioro de la estimación a medida en la que incrementa α. Esto supone un claro problema en el campo de la gestión y cuantificación de riesgos, ya que no se estiman correctamente los coeficientes para α extremas, que son justamente nuestro punto de interés.

Tras comprobar que el modelo de regresión cuantílica de Koenker y Bassett (1978), empeora sus resultados para distribuciones con cola pesada, estudiamos como alternativa dos modelos no paramétricos de regresión cuantílica, que son; el método basado en splines y el basado en estimación núcleo (kernel), además, en este caso no se asume una relación lineal entre la variable dependiente y las explicativas. Al ser comparados estos dos métodos con el del primer estudio mediante el cálculo de error cuadrático medio de la predicción, observamos que los nuevos modelos no paramétricos en general no proporcionan mejores predicciones que el clásico. Por tanto, en nuestro contexto específico preferimos utilizar el modelo de regresión cuantílica ya que obtenemos

mejores predicciones en un número mayor de casos en la simulación. Aunque para la distribución con cola más pesada y cuantiles muy extremos la estimación tipo núcleo proporciona mejores resultados con muestras de mayor tamaño.

Basando este modelo en los datos reales, podemos dividir el análisis en dos grupos, uno para los costes asociados a pólizas de automóviles y otro asociado a los costes de pólizas del hogar.

Los resultados para los costes de pólizas de automóviles nos indican que la variable de género tiene un efecto significativo para los cuantiles a partir del 95% y además distinto que al del modelo lineal. Esto significa que las mujeres generan unos costes menores al de los hombres, y además el efecto del género incrementa a mediada que lo hace el cuantil. De la misma manera sucede con aquellos clientes que disponen de otras pólizas contratadas, ya que tienen menos costes a diferencia de los que tienen una única póliza, esto podría ser debido a las propias políticas de las compañías aseguradoras, porque estas están dispuestas a ampliar el número de productos contratados para aquellos clientes con buen comportamiento en términos de costes. De forma contraria sucede con los clientes que generan elevados costes. Finalmente, es importante destacar las diferencias detectadas entre el efecto de la edad del primer conductor y la edad del tomador del seguro, puesto que la edad del primer conductor es significativa y la edad del tomador del seguro no lo es, esto es relevante ya que se ha detectado una correlación entre estas variables. Para la edad del primer conductor, se observa que a medida que va incrementando la edad se reducen los costes.

En cuanto el modelo utilizado para las pólizas de hogar, encontramos unas variables distintas a las del anterior. Surgen como variables significativas las variables de residencia. En primer lugar, los clientes residentes en grandes ciudades tienen unos mayores costes asociados hasta el cuantil 98%, de igual forma sucede para aquellos clientes residentes en el norte del territorio, pero hasta el cuantil 97%. Bajo mi punto de vista creo que los residentes en grandes ciudades generan más costes debido a que las viviendas en esas zonas tienen un mayor valor, dejando así sin efecto que la causa sea las características intrínsecas de las personas residentes. El único caso en la que la variable es significativa en ambos modelos es la posesión de otras pólizas de seguro.

Con todo esto podemos concluir que a pesar de que no es el mejor modelo, es muy útil para escribir efectos en valores extremos de los datos y poder entender mejor los factores generadores de los riesgos asociados a las carteras de pólizas de seguros.

Bibliografía

Bloomfield. P.. & Steiger. W. L. (1983). LAD Spline Fitting. In *Least Absolute Deviations* (pp. 131–151). https://doi.org/10.1007/978-1-4684-8574-5_5

Bolancé. C.. & Vernic. R. (2019). Multivariate count data generalized linear models: Three approaches based on the Sarmanov distribution. *Insurance: Mathematics and Economics*. *85*. 89–103. https://doi.org/10.1016/J.INSMATHECO.2019.01.001

Chaudhuri. P. (1991). Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation. *The Annals of Statistics*. *19*(2). 760–777. https://doi.org/10.1214/aos/1176348119

Chernozhukov. V.. & Fernandez-Val. I. (2011). Inference for Extremal Conditional Quantile Models. with an Application to Market and Birthweight Risks. *The Review of Economic Studies*. *78*(2). 559–589. https://doi.org/10.1093/restud/rdq020

Chernozhukov. Victor. & Umantsev. L. (2001). Conditional value-at-risk: Aspects of modeling and estimation. *Empirical Economics*. *26*(1). 271–292. https://doi.org/10.1007/s001810000062

Cole. T. J. (1988). Fitting Smoothed Centile Curves to Reference Data. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*. *151*(3). 385-418. https://doi.org/10.2307/2982992

de Haan. L. (1970). On regular variation and its application to the weak convergence of sample extremes. *Mathamatisch Centrum, Amsterdam.*

Engle. R. F.. & Manganelli. S. (2004). CAViaR. *Journal of Business & Economic Statistics*. *22*(4). 367–381. https://doi.org/10.1198/073500104000000370

Hastie. T.. & Tibshirani. R. (1987). Generalized Additive Models: Some Applications. *Journal of the American Statistical Association*. *82*(398). 371–386. https://doi.org/10.1080/01621459.1987.10478440

Jansen. D. W.. & de Vries. C. G. (1991). On the Frequency of Large Stock Returns: Putting Booms and Busts into Perspective. *The Review of Economics and*

Statistics. *73*(1). 18-24. https://doi.org/10.2307/2109682

Koenker. R.. & Bassett. G. (1978). Regression Quantiles. In *Econometrica* 46(1), 33-50. doi:10.2307/1913643

Koenker. R.. & Mizera. I. (2004). Penalized Triograms: Total Variation Regularization for Bivariate Smoothing. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*. 66(1). pp. 145–163. https://doi.org/10.2307/3647632

Koenker. R.. Ng. P.. & Portnoy. S. (1994). Quantile smoothing splines. *Biometrika*. *81*(4). 673–680. https://doi.org/10.1093/biomet/81.4.673

Schwarz. G. (1978). Estimating the Dimension of a Model. In *Source: The Annals of Statistics* (6).

Smith. R. L. (1992). *Introduction to Gnedenko (1943) On the Limiting Distribution of the Maximum Term in a Random Series*. https://doi.org/10.1007/978-1-4612-0919-5_13

Takeuchi. I.. Le. Q. V. Sears. T. D.. & Smola. A. J. (2006). Nonparametric Quantile Estimation. In *Journal of Machine Learning Research* (Vol. 7).

Wahba. G.. & Society for Industrial and Applied Mathematics. (1990). *Spline models for observational data*.

Woodroofe. M.. & Meyer. M. (2000). On the degrees of freedom in shape-restricted regression. *The Annals of Statistics*. *28*(4). 1083–1104. https://doi.org/10.1214/aos/1015956708

Apéndice A: Código R implementado en el estudio de simulación

```
#librerias
library(EnvStats)#para generar valeros de Pareto
library(quantreg) #para estimar por regresion cuantilica Koeneker
library(kernlab) #para modelos no paramétricos
library(splines) #para splines

#numero de replicas
nrep<-500
#tamaño muestral
n<-1000
#parametro de heteroskedacity. cuando es 0 es homoskedastico
r<-0.9
#estimamos por regresion cuantilica los cuantiles:
tau<-c(0.5.0.7.0.90.0.95.0.99.0.995)

#generamos x1 y x2 como nomales y el parámetro e
x1<-matrix(nrow=n.ncol=nrep)
x2<-matrix(nrow=n.ncol=nrep)
e<-matrix(nrow=n.ncol=nrep)
set.seed(1)
for(i in 1:nrep){
  x1[.i]<-runif(n.-1.1)
  x2[.i]<-runif(n.-1.1)
  e[.i]<-rpareto(n.location=1.shape=2)#rlnorm(n.0.1)



}
hist(e[.1]. ylab="". xlab=expression('e'[i]). main="".breaks=20
```

```
    . ylim=c(0.800).xlim=c(0.20))
#modelo generador de y
y<-matrix(nrow=n.ncol=nrep)
for(i in 1:nrep){y[.i]<-x1[.i]+(1+r*x1[.i])*e[.i]+x2[.i]}

#histograma descriptivos
plot(x1[.1].y[.1].xlab=expression('x'[1][i]).ylab=expression('y'[i]))
plot(x2[.1].y[.1].xlab=expression('x'[2][i]).ylab=expression('y'[i]))
#definimos los parametros reales del modelo
beta_1_real<-vector()
beta_1_list<-matrix(nrow=nrep.ncol=length(tau))
beta_2_real<-rep(1.length(tau)) #por construcción del modelo
beta_0_real<-vector()
beta_0_list<-matrix(nrow=nrep.ncol=length(tau))
for(j in 1:nrep){
  for(i in 1:length(tau)){
    beta_0_real[i]<-quantile(e[.j].tau[i])
    beta_1_real[i]<-(1+r*quantile(e[.j].tau[i]))}
  beta_0_list[j.]<-beta_0_real
  beta_1_list[j.]<-beta_1_real}

#definimos los valores para la predicción
x_test<-matrix(nrow=3.ncol=2)
x_test[1.]<--0.5;x_test[2.]<-0;x_test[3.]<-0.5
x_test<-as.data.frame(x_test)
colnames(x_test)<-c('x1_fit'.'x2_fit')
y_test<-matrix(nrow=3.ncol=length(tau))
y_test_list<-list()
for(k in 1:nrep){
  for(j in 1:nrow(x_test)){
    for(i in 1:length(beta_0_real)){
```

```
    y_test[j.i]<-beta_0_list[k.][i]+beta_1_list[k.][i]*x_test[j.1]
    +beta_2_real[i]*x_test[j.2]}}
  y_test_list[[k]]<-y_test}

#estimamos el modelo
beta_0<-matrix(0.nrow=nrep.ncol=length(tau));colnames(beta_0) <- tau
beta_1<-matrix(0.nrow=nrep.ncol=length(tau));colnames(beta_1) <- tau
beta_2<-matrix(0.nrow=nrep.ncol=length(tau));colnames(beta_2) <- tau

#ajustamos regresión cuantílica clásica
predict_rq<-list()
for(i in 1:nrep){
  y_fit<-y[.i]
  x1_fit<-x1[.i]
  x2_fit<-x2[.i]
  fit<-rq(y_fit~x1_fit+x2_fit. tau=tau)
  predict_rq[[i]]<-predict(fit.newdata=x_test)
  fit_coef<-fit$coefficients
  for (j in 1:length(tau)){
    beta_0[i.j]<-fit_coef[1.j]
    beta_1[i.j]<-fit_coef[2.j]
    beta_2[i.j]<-fit_coef[3.j]
  }
}

#calculamos sesgo. y varianza del los parámetros del modelo
sesgo_beta_0<-colSums((beta_0-beta_0_list))/nrep
varianza_beta_0<-vector()
for(i in 1:length(tau)){
  varianza_beta_0[i]<-var(beta_0[.i])}
mse_beta_0<-sesgo_beta_0^2+varianza_beta_0
```

```
sesgo_beta_1<-colSums((beta_1-beta_1_list))/nrep
varianza_beta_1<-vector()
for(i in 1:length(tau)){
  varianza_beta_1[i]<-var(beta_1[.i])}
mse_beta_1<-sesgo_beta_1^2+varianza_beta_1

sesgo_beta_2<-colSums((beta_2-beta_2_real))/nrep
varianza_beta_2<-vector()
for(i in 1:length(tau)){
  varianza_beta_2[i]<-var(beta_2[.i])}
mse_beta_2<-sesgo_beta_1^2+varianza_beta_2

#cálculo del MSE de la prediccion
mse_0<-matrix(nrow=nrep.ncol=length(tau))
mse_1<-matrix(nrow=nrep.ncol=length(tau))
mse_2<-matrix(nrow=nrep.ncol=length(tau))

for(i in 1:nrep){
  mse_0[i.]<-(predict_rq[[i]][1.]-y_test_list[[i]][1.])^2
  mse_1[i.]<-(predict_rq[[i]][2.]-y_test_list[[i]][2.])^2
  mse_2[i.]<-(predict_rq[[i]][3.]-y_test_list[[i]][3.])^2}

mse_0<-colMeans(mse_0)
mse_1<-colMeans(mse_1)
mse_2<-colMeans(mse_2)

#modelo aditivo:
predict_qss_0.5<-list()
predict_qss_0.7<-list()
predict_qss_0.9<-list()
```

```
predict_qss_0.95<-list()
predict_qss_0.99<-list()
predict_qss_0.995<-list()
predict_qss<-matrix(nrow=nrow(x_test).ncol=length(tau))
predict_qss_list<-list()

for(i in 1:nrep){
  data<-matrix(nrow=n.ncol=3)
  data[.1]<-y[.i]
  data[.2]<-x1[.i]
  data[.3]<-x2[.i]
  colnames(data)<-c('y'.'x1_fit'.'x2_fit')
  data<-as.data.frame(data)
  fit_qss_0.5<-rqss(y ~ qss(cbind(x1_fit.x2_fit)). data=data.tau=0.50)
  predict_qss_0.5[[i]]<-predict(fit_qss_0.5.x_test)
  predict_qss[.1]<-predict_qss_0.5[[i]]
  fit_qss_0.7<-rqss(y ~ qss(cbind(x1_fit.x2_fit)). data=data.tau=0.7)
  predict_qss_0.7[[i]]<-predict(fit_qss_0.7.x_test)
  predict_qss[.2]<-predict_qss_0.7[[i]]
  fit_qss_0.9<-rqss(y ~ qss(cbind(x1_fit.x2_fit)). data=data.tau=0.90)
  predict_qss_0.9[[i]]<-predict(fit_qss_0.9.x_test)
  predict_qss[.3]<-predict_qss_0.9[[i]]
  fit_qss_0.95<-rqss(y ~ qss(cbind(x1_fit.x2_fit)). data=data.tau=0.95)
  predict_qss_0.95[[i]]<-predict(fit_qss_0.95.x_test)
  predict_qss[.4]<-predict_qss_0.95[[i]]
  fit_qss_0.99<-rqss(y ~ qss(cbind(x1_fit.x2_fit)). data=data.tau=0.99)
  predict_qss_0.99[[i]]<-predict(fit_qss_0.9.x_test)
  predict_qss[.5]<-predict_qss_0.99[[i]]
  fit_qss_0.995<-rqss(y ~ qss(cbind(x1_fit.x2_fit)). data=data.tau=0.995)
  predict_qss_0.995[[i]]<-predict(fit_qss_0.995.x_test)
  predict_qss[.6]<-predict_qss_0.995[[i]]
```

```
  predict_qss_list[[i]]<-predict_qss}

#cálculo del MSE
mse_0_qss<-matrix(nrow=length(predict_qss_list).ncol=length(tau))
mse_1_qss<-matrix(nrow=length(predict_qss_list).ncol=length(tau))
mse_2_qss<-matrix(nrow=length(predict_qss_list).ncol=length(tau))

for(i in 1:length(predict_qss_list)){
  mse_0_qss[i.]<-(predict_qss_list[[i]][1.]-y_test[1.])^2
  mse_1_qss[i.]<-(predict_qss_list[[i]][2.]-y_test[2.])^2
  mse_2_qss[i.]<-(predict_qss_list[[i]][3.]-y_test[3.])^2}

mse_0_qss<-colMeans(mse_0_qss)
mse_1_qss<-colMeans(mse_1_qss)
mse_2_qss<-colMeans(mse_2_qss)

#generamos regresion no paramétrica: kernel lineal
predict_kqr_0.5<-list()
predict_kqr_0.7<-list()
predict_kqr_0.9<-list()
predict_kqr_0.95<-list()
predict_kqr_0.99<-list()
predict_kqr_0.995<-list()
predict_kqr<-matrix(nrow=nrow(x_test).ncol=length(tau))
predict_kqr_list<-list()
for(i in 1:nrep){
  x_fit<-matrix(nrow=n.ncol=2)
  x_fit[.1]<-x1[.i]
  x_fit[.2]<-x2[.i]
  y_fit<-y[.i]
  fit_kqr_0.5<-kqr(x_fit. y_fit. tau =0.5. kernel = "rbfdot")
```

```
  predict_kqr_0.5[[i]]<-predict(fit_kqr_0.5.newdata=x_test)
  predict_kqr[.1]<-predict_kqr_0.5[[i]]
  fit_kqr_0.7<-kqr(x_fit. y_fit. tau =0.7. kernel = "rbfdot")
  predict_kqr_0.7[[i]]<-predict(fit_kqr_0.7.newdata=x_test)
  predict_kqr[.2]<-predict_kqr_0.7[[i]]
  fit_kqr_0.9<-kqr(x_fit. y_fit. tau =0.9. kernel = "rbfdot")
  predict_kqr_0.9[[i]]<-predict(fit_kqr_0.9.newdata=x_test)
  predict_kqr[.3]<-predict_kqr_0.9[[i]]
  fit_kqr_0.95<-kqr(x_fit. y_fit. tau =0.95. kernel = "rbfdot")
  predict_kqr_0.95[[i]]<-predict(fit_kqr_0.95.newdata=x_test)
  predict_kqr[.4]<-predict_kqr_0.95[[i]]
  fit_kqr_0.99<-kqr(x_fit. y_fit. tau =0.99. kernel = "rbfdot")
  predict_kqr_0.99[[i]]<-predict(fit_kqr_0.99.newdata=x_test)
  predict_kqr[.5]<-predict_kqr_0.99[[i]]
  fit_kqr_0.995<-kqr(x_fit. y_fit. tau =0.995. kernel = "rbfdot")
  predict_kqr_0.995[[i]]<-predict(fit_kqr_0.995.newdata=x_test)
  predict_kqr[.6]<-predict_kqr_0.995[[i]]
  predict_kqr_list[[i]]<-predict_kqr}

#cálculo del MSE
mse_0_kqr<-matrix(nrow=length(predict_kqr_list).ncol=length(tau))
mse_1_kqr<-matrix(nrow=length(predict_kqr_list).ncol=length(tau))
mse_2_kqr<-matrix(nrow=length(predict_kqr_list).ncol=length(tau))

for(i in 1:length(predict_kqr_list)){
  mse_0_kqr[i.]<-(predict_kqr_list[[i]][1.]-y_test[1.])^2
  mse_1_kqr[i.]<-(predict_kqr_list[[i]][2.]-y_test[2.])^2
  mse_2_kqr[i.]<-(predict_kqr_list[[i]][3.]-y_test[3.])^2}
mse_0_kqr<-colMeans(mse_0_kqr)
mse_1_kqr<-colMeans(mse_1_kqr)
mse_2_kqr<-colMeans(mse_2_kqr)
```

Apéndice B: Código R implementado en datos de costes de pólizas de automóvil

```
rm(list=ls())


#librerias
library(readxl) #para generar valeros de Pareto
library(quantreg) #para estimar por regresion cuantilica Koeneker
library(dplyr) #para limpiar datos
library(ggplot2)
library(GGally)
library(plyr)
library(tidyverse)
#importamos los datos de hogar
costes_auto <-
read_excel("C:/Users/alvar/Desktop/CURSOS/MESIO/TFM/Correcto/datos/Claims_3d_def.xls
x".
                    sheet = "Coste auto limpio")
head(costes_auto)


#cambiamos la variable para hacerla binaria. otras pólizas=1 . no hay otras pólizas=0
costes_auto$client_other_npol<- ifelse( costes_auto$client_other_npol>0.
                                1. costes_auto$client_other_npol )


#arreglamos los tipos de variables
costes_auto[.c(3.4.9)] <- lapply(costes_auto[.c(3.4.9)] . as.numeric)
costes_auto[.c(1.2.5.6.7.8)] <- lapply(costes_auto[.c(1.2.5.6.7.8)] . as.character)



#densidad de los costes
windows()
hist(costes_auto$CosteAuto. breaks=500.xlim=c(0.90000)
    .xlab="Costes asociados a pólizas de automóvil"
    .ylab="
    .main=")
```

```
quantile(costes_auto$CosteAuto. probs=seq(0.1.0.05))


#Miramos si hay missings

(missings<-sapply(costes_auto. function(x) sum(is.na(x))))

which(is.na(costes_auto$firstdriver_agelicense_ag ))

costes_auto<-costes_auto[-(which(is.na(costes_auto$firstdriver_agelicense_ag))).] #eliminamos los missings

(missings<-sapply(costes_auto. function(x) sum(is.na(x)))) #comprobamos que no hay más missings


#Descriptivos

(descriptive <- psych:: describe(costes_auto[.c(3.4.9)]. skew = T))#solo las continuas


#vemos que la edad hay negativas o inferiores a 18 años. filtramos para que sean mayores a 18

hist(costes_auto$client_age. xlab="Edad del tomador del seguro"

    . ylab="Número observaciones"

    .main="Distribución de la edad del tomador del seguro original")


costes_auto<-costes_auto %>%

 filter(client_age>=16) %>%

 select(client_id.          client_sex.       client_age.       firstdriver_agelicense_ag.

       client_other_npol. zone1.  zone2.  zone3.  CosteAuto)


#También vemos que el máximo de edad son 339 años. lo limitaremos a 114 que es el sexto valor más alto

costes_auto<-costes_auto %>%

 filter(client_age<=114) %>%

 select(client_id.          client_sex.       client_age.       firstdriver_agelicense_ag.

       client_other_npol. zone1.  zone2.  zone3.  CosteAuto)


hist(costes_auto$client_age. xlab="Edad del tomador del seguro"

    . ylab=""

    .main=""

    .xlim=c(0.120)
```

```
    .cex=25)

#revisamos de nuevo los valores
(descriptive <- psych:: describe(costes_auto[.c(3.4.9)]. skew = T))

#la edad del primer conductor también registra valores extraños como 3 años.
#como la edad mínima para conducir son 16 (motocicletas) ponemos esa edad
sort(costes_auto$firstdriver_agelicense_ag. decreasing=F)
hist(costes_auto$firstdriver_agelicense_ag)
costes_auto<-costes_auto %>%
  filter(firstdriver_agelicense_ag>=16) %>%
  select(client_id.          client_sex.        client_age.        firstdriver_agelicense_ag.
         client_other_npol. zone1.   zone2.   zone3.   CosteAuto)

windows()
hist(costes_auto$firstdriver_agelicense_ag. xlab="Edad del primer conductor"
    . ylab=""
    .main=""
    .xlim=c(0.70))

#revisamos de nuevo los valores
(descriptive <- psych:: describe(costes_auto[.c(3.4.9)]. skew = T))

#descriptivos categoricas
summary(costes_auto[.c(1.2.5.6.7.8)])

#gráficos descriptivos de las variables

windows()
ggpairs(costes_auto[.c(2:5.9)])

#matriz de correlaciones entre continuas

round(cor(costes_auto[.c(3.4.9)]).2)
```

```
#genero
costes_auto$client_sex<-as.factor(costes_auto$client_sex)
costes_auto$client_sex<-revalue(costes_auto$client_sex.
                        c("Man"="Hombre". "Woman"="Mujer"))
prop.table(table(costes_auto$client_sex))
windows()
boxplot( costes_auto$firstdriver_agelicense_ag~costes_auto$client_sex.
       xlab='Género'. ylab='Edad del primer conductor'. main='')
boxplot( costes_auto$client_age~costes_auto$client_sex.
       xlab='Género'. ylab='Edad del cliente'. main='')
boxplot( costes_auto$CosteAuto~costes_auto$client_sex.
       xlab='Género'. ylab='Costes asociados a pólizas de automóvil')

#otras pólizas
costes_auto$client_other_npol<-as.factor(costes_auto$client_other_npol)
costes_auto$client_other_npol<-revalue(costes_auto$client_other_npol.
      c("0"="No dispone otras pólizas". "1"="Dispone otras pólizas"))
prop.table(table(costes_auto$client_other_npol))
windows()
boxplot( costes_auto$CosteAuto~costes_auto$client_other_npol
       .xlab='Disposición pólizas'
       .ylab='Costes asociados a pólizas de automóvil')

#zona2
costes_auto$zone2<-as.factor(costes_auto$zone2)
costes_auto$zone2<-revalue(costes_auto$zone2.
     c("0"="No reside en el norte". "1"="Reside en el norte"))
prop.table(table(costes_auto$zone2))
windows()
boxplot( costes_auto$CosteAuto~costes_auto$zone2.
       xlab='Residencia'. ylab='Costes asociados a pólizas de automóvil')

#zona1
```

```
costes_auto$zone1<-as.factor(costes_auto$zone1)
costes_auto$zone1<-revalue(costes_auto$zone1.
                    c("0"="No reside en una gran ciudad"
                     . "1"="Reside en una gran ciudad"))
prop.table(table(costes_auto$zone1))
windows()
boxplot( costes_auto$CosteAuto~costes_auto$zone1.
       xlab='Residencia'
       . ylab='Costes asociados a pólizas de automóvil')


#Ajustamos modelo con la edad del primer conductor
fit<-rq(CosteAuto~client_sex+firstdriver_agelicense_ag+client_other_npol
      +client_other_npol+zone1+zone2
      .data=costes_auto. tau=seq(0.01.0.99. by=0.01))
summary<-summary(fit. se='iid')
windows()
nombres<-c("Intercepto"."Género"
        . "Edad del primer conductor"
        ."Otras pólizas"
        ."Residente Gran Ciudad"
        ."Residente Norte")
plot(summary. main=nombres.xlab="Cuantil". ylab="Coeficiente")

#Ajustamos modelo a partir del 0.90
fit_0.90<-rq(CosteAuto~client_sex+firstdriver_agelicense_ag
      +client_other_npol+zone2+zone3
      .data=costes_auto. tau=seq(0.90.0.99. by=0.01))
summary_0.90<-summary(fit_0.90. se='iid')
windows()
plot(summary_0.90. main=nombres.xlab="Cuantil". ylab="Coeficiente")

#Ajustamos modelo. con edad del tomador seguro
fit2<-rq(CosteAuto~client_sex+client_age
```

```
      +client_other_npol+zone1+zone2
      .data=costes_auto. tau=seq(0.01.0.99. by=0.01))
summary2<-summary(fit2. se='iid')
windows()
nombres_2<-c("Intercepto"."Género"
        . "Edad del tomador del seguro"
        ."Otras pólizas"
        ."Residente Gran Ciudad"
        ."Residente Norte")
plot(summary2. main=nombres_2.xlab="Cuantil". ylab="Coeficiente")

#Ajustamos modelo a partir del 0.90
fit2_0.90<-rq(CosteAuto~client_sex+client_age
          +client_other_npol+zone2+zone3
          .data=costes_auto. tau=seq(0.90.0.99. by=0.01))
summary2_0.90<-summary(fit2_0.90. se='iid')
windows()
plot(summary2_0.90. main=nombres_2.xlab="Cuantil". ylab="Coeficiente")
```

Apéndice C: Código R implementado en datos de costes de pólizas de hogar

```
#librerias
library(readxl) #para generar valeros de Pareto
library(quantreg) #para estimar por regresion cuantilica Koeneker
library(dplyr) #para limpiar datos
library(ggplot2)
library(GGally)
library(plyr)
library(tidyverse)

#importamos los datos de hogar
costes_hogar <-
read_excel("C:/Users/alvar/Desktop/CURSOS/MESIO/TFM/Correcto/datos/Claims_3d_def.xls
x".
                    sheet = "Coste Hogar limpio")
head(costes_hogar)

#cambiamos la variable para hacerla binaria. otras pólizas=1 . no hay otras pólizas=0
costes_hogar$client_other_npol<- ifelse( costes_hogar$client_other_npol>0.
                                1. costes_hogar$client_other_npol )

#arreglamos los tipos de variables
costes_hogar[.c(3.4.9)] <- lapply(costes_hogar[.c(3.4.9)] . as.numeric)
costes_hogar[.c(1.2.5.6.7.8)] <- lapply(costes_hogar[.c(1.2.5.6.7.8)] . as.factor)
#eliminamos la edad del primer conductor
costes_hogar<-costes_hogar[.-4]

#densidad de los costes
windows()
hist(costes_hogar$CosteHogar. breaks=500.ylim=c(0.10000)
    .xlab="Costes asociados a pólizas de hogar"
    .ylab="
    .main=")
```

```
lines(density(costes_hogar$CosteHogar))
quantile(costes_hogar$CosteHogar. probs=seq(0.1.0.05))

#Miramos si hay missings
(missings<-sapply(costes_hogar. function(x) sum(is.na(x))))
which(is.na(costes_hogar$client_age))
costes_hogar<-costes_hogar[-(which(is.na(costes_hogar$client_age))).] #eliminamos los
missings

#Descriptivos
(descriptive <- psych:: describe(costes_hogar[.c(3.8)]. skew = T))#solo las continuas

#vemos que la edad hay negativas o inferiores a 18 años. filtramos para que sean mayores a 18
costes_hogar<-costes_hogar %>%
  filter(client_age>=16) %>%
  select(client_id.
         client_sex.
         client_age.
         client_other_npol. zone1.   zone2.   zone3.   CosteHogar)

#edad del tomador
windows()
hist(costes_hogar$client_age. xlab="Edad del tomador del seguro"
     . ylab=""
     .main=""
     .xlim=c(0.120)
     .ylim=c(0.2000)
     .cex=25)

#revisamos de nuevo los valores
(descriptive <- psych:: describe(costes_hogar[.c(3.8)]. skew = T))

#descriptivo de categoricas
summary(costes_hogar[.c(1.2.4.5.6.7)])
```

```
#gráficos descriptivos de las variables

windows()
ggpairs(costes_hogar[.c(2:4.8)])

#genero
costes_hogar$client_sex<-as.factor(costes_hogar$client_sex)
costes_hogar$client_sex<-revalue(costes_hogar$client_sex.
                        c("Man"="Hombre". "Woman"="Mujer"))
prop.table(table(costes_hogar$client_sex))
windows()
boxplot( costes_hogar$client_age~costes_hogar$client_sex.
      xlab='Género'. ylab='Edad del cliente'. main='')
boxplot( costes_hogar$CosteHogar~costes_hogar$client_sex.
      xlab='Género'. ylab='Costes asociados a pólizas de hogar')
#otras pólizas
costes_hogar$client_other_npol<-as.factor(costes_hogar$client_other_npol)
costes_hogar$client_other_npol<-revalue(costes_hogar$client_other_npol.
                              c("0"="No dispone otras pólizas". "1"="Dispone otras pólizas"))
prop.table(table(costes_hogar$client_other_npol))
windows()
boxplot( costes_hogar$CosteHogar~costes_hogar$client_other_npol
      .xlab='Disposición pólizas'
      .ylab='Costes asociados a pólizas de hogar')
#zona1
costes_hogar$zone1<-as.factor(costes_hogar$zone1)
costes_hogar$zone1<-revalue(costes_hogar$zone1.
                      c("0"="No reside en una gran ciudad"
                       . "1"="Reside en una gran ciudad"))
prop.table(table(costes_hogar$zone1))
windows()
boxplot( costes_hogar$CosteHogar~costes_hogar$zone1.
      xlab='Residencia'
```

```
        . ylab='Costes asociados a pólizas de hogar')

#zona2
costes_hogar$zone2<-as.factor(costes_hogar$zone2)
costes_hogar$zone2<-revalue(costes_hogar$zone2.
                    c("0"="No reside en el norte". "1"="Reside en el norte"))
prop.table(table(costes_hogar$zone2))
windows()
boxplot( costes_hogar$CosteHogar~costes_hogar$zone2.
        xlab='Residencia'. ylab='Costes asociados a pólizas de hogar')

#Ajustamos modelo
fit<-rq(CosteHogar~client_sex+client_age
        +client_other_npol+zone1+zone2
        .data=costes_hogar. tau=seq(0.01.0.99. by=0.01))
summary<-summary(fit. sd='iid')
summary
nombres_2<-c("Intercepto"."Género"
           . "Edad del tomador del seguro"
           ."Otras pólizas"
           ."Residente Gran Ciudad"
           ."Residente Norte")
windows()
plot(summary. main=nombres_2.xlab="Cuantil". ylab="Coeficiente")


#Ajustamos modelo a partir del 0.90
fit_0.90<-rq(CosteHogar~client_sex+client_age
           +client_other_npol+zone1+zone2
           .data=costes_hogar. tau=seq(0.90.0.99. by=0.01))
summary_0.90<-summary(fit_0.90. se='iid')
summary_0.90
windows()
plot(summary_0.90. main=nombres_2.xlab="Cuantil". ylab="Coeficiente")
```