Máster en Ciencias Actuariales y Financieras

TRABAJO FINAL DE MÁSTER

Título: Análisis de los factores de riesgo en el seguro de automóvil mediante la regresión cuantílica y *expected shortfall*

Autoría: Wanting Li

Tutoría: Miguel Santolino

Curso académico: 2022-2023

UNIVERSITAT DE BARCELONA
Facultat d'Economia i Empresa

Màster de Ciències Actuarials i Financeres

Facultad de Economía y Empresa

Universidad de Barcelona

Trabajo Final de Máster

Máster en Ciencias Actuariales y Financieras

Análisis de los factores de riesgo en el seguro de automóvil mediante la regresión cuantílica y *expected shortfall*

Autoría:

Wanting Li

Tutoría:

Miguel Santolino

El contenido de este documento es de exclusiva responsabilidad del autor, quien declara que no ha incurrido en plagio y que la totalidad de referencias a otros autores han sido expresadas en el texto.

AGRADECIMIENTO

Quiero expresar mi más sincero agradecimiento a mi familia por su constante apoyo y su amor incondicional que me ha impulsado a seguir adelante en cada etapa de mi formación. También quiero agradecer a mi tutor, Miguel Santolino, por su inagotable paciencia, orientación y valiosa ayuda durante la realización de este trabajo. Asimismo, deseo expresar mi gratitud a todos mis profesores del Máster por sus enseñanzas y amabilidad.

A todos ustedes, muchas gracias.

RESUMEN

Este trabajo se enfoca en el estudio de las técnicas de regresión cuantílica y regresión conjunta de regresión cuantílica con *expected shortfall*. El objetivo principal es superar las limitaciones de la regresión lineal estándar en un contexto de datos con colas pesadas. Con este fin, se aplicarán ambas regresiones en una base de datos de seguros de automóviles, con el propósito de mejorar la comprensión de las relaciones entre variables en diferentes puntos de la distribución de la variable dependiente, que en este caso es el coste total de siniestros, y realizar predicciones más precisas.

Palabras claves: Seguros de automóvil, Regresión cuantílica, Valor en riesgo, Pérdida esperada, Regresión conjunta.

ABSTRACT

This work focuses on the study of quantile regression techniques and the joint regression of quantile regression with expected shortfall. The main objective is to overcome the limitations of standard linear regression in a heavy-tailed data context. To achieve this, both regressions will be applied to a car insurance database, with the purpose of improving the understanding of the relationships between variables at different points of the dependent variable distribution, which in this case is the total cost of claims, and making more accurate predictions.

Keywords: Automobile insurance, Quantile regression, Value at Risk, Expected shortfall, Joint regression.

ÍNDICE

1. INTRODUCCIÓN

El presente trabajo se centra en el estudio de las técnicas de regresión cuantílica y regresión conjunta de regresión cuantílica con *expected shortfall*. Además de sus aplicaciones en un contexto de datos con cola pesada y cuantiles extremos, donde sus usos resultan de especial interés.

La regresión lineal estándar es una técnica comúnmente utilizada para analizar el efecto de un conjunto de variables explicativas sobre una variable dependiente, ajustando la media condicional, es decir, la relación media entre las variables. Sin embargo, esta técnica tiene limitaciones, ya que asume que la relación entre las variables explicativas y la variable dependiente es constante a lo largo de toda la distribución de la variable dependiente, lo que no siempre es cierto en la realidad.

Para superar estas limitaciones, una alternativa es la regresión cuantílica, una técnica que permite modelar las relaciones entre variables en diferentes puntos de la distribución de la variable dependiente. Otra alternativa es la regresión conjunta de regresión cuantílica con *expected shortfall* (una medida de riesgo financiero que tiene en cuenta los cuantiles extremos de la distribución). Esta medida está adquiriendo cada vez más importancia, como lo indica el Comité de Supervisión Bancaria de Basilea[1]: "El uso de ES (*expected shortfall*) ayudará a garantizar una captura más prudente del 'riesgo de cola' y la adecuación de capital durante períodos de estrés significativo en los mercados financieros[2]".

Por lo tanto, la estructura del trabajo se divide en dos partes principales. En la primera parte, se lleva a cabo un estudio exhaustivo de las teorías relacionadas con las técnicas de regresión cuantílica y regresión conjunta de regresión cuantílica con *expected shortfall*, lo cual ayuda a establecer una base sólida para comprender las técnicas que se aplicarán posteriormente. En la segunda parte del trabajo, se realiza la aplicación práctica de estas técnicas en una base de datos de seguros de automóviles. Se preprocesan los datos y se estudian las variables, para luego ajustar los modelos de regresión cuantílica y regresión conjunta. Se analizan diferentes puntos de la distribución de la variable dependiente con el objetivo de obtener una comprensión más completa de las relaciones existentes. Al finalizar el trabajo, se presenta una conclusión que resume los *insights* obtenidos durante el desarrollo del trabajo.

[1] Véase Basel Committeeon Banking Supervision publication: Minimum capital requirements for market risk

[2] Situación de incertidumbre económica

2. METODOLOGÍA

2.1. CUANTIL

El concepto de cuantil fue introducido por Kendall (1940)[3], quien acuñó este término para hacer referencia a la división equitativa de una distribución de probabilidad. Por ejemplo, el cuantil 50%, conocido como la mediana, divide la distribución en dos partes iguales. Dado que los cuantiles tienen un papel fundamental en el presente trabajo, es apropiado recordar la definición precisa de los mismos.

La inversa generalizada y la función de cuantiles [9]:

1. La inversa generalizada de una función creciente $T: \mathbb{R} \to \mathbb{R}$, denotada como $T^{\leftarrow}(y)$, se define como el ínfimo (*infimum*) de los valores x en el dominio de *T*, tal que $T(x) \geq y$. En otras palabras, es el valor más pequeño de x en $\mathbb{R}$, para el cual T(x) es mayor o igual que y.

   Matemáticamente, la definición de la inversa generalizada de T se expresa como:

   $$T^{\leftarrow}(y) := inf\{x \in \mathbb{R}:\ T(x) \geq y\}$$

   Es importante destacar que si no existe ningún valor x en $\mathbb{R}$ que cumpla con la condición $T(x) \geq y$, entonces se utiliza la convención de que el ínfimo de un conjunto vacío es $+\infty$ (infinito positivo).

2. La función de cuantiles o función cuantil de una función de distribución acumulada (FDA) F, denotada como $F^{\leftarrow}$, se define como la inversa generalizada de F. Específicamente, para un valor de α en el intervalo (0, 1), el α-cuantil de F se define como el ínfimo de los valores x en el dominio de F, tal que $F(x) \geq \alpha$.

   Matemáticamente, la definición de la función de cuantiles se expresa como:

   $$F^{\leftarrow}(\alpha) := inf\{x \in \mathbb{R}:\ F(x) \geq \alpha\}$$

   Donde $F^{\leftarrow}(\alpha)$ representa la función cuantil de F evaluada en α, y $Q_{\alpha}(F)$ se utiliza mucho para denotar el α-cuantil de F, que es equivalente a $F^{\leftarrow}(\alpha)$.

2.2. VaR (*Value at Risk*) y ES (*Expected shortfall*)

El VaR (*Value at Risk*) es una medida de riesgo ampliamente utilizada en la gestión de riesgos financieros [9]. Desde una perspectiva probabilística, el VaR puede entenderse como un simple cuantil de la distribución de pérdidas. Por lo tanto, se define como el cuantil generalizado $F^{\leftarrow}(\alpha)$:

[3] Véase https://es.wikipedia.org/wiki/Cuantil#

$$\mathrm{VaR}_{\alpha}(Y) := F^{-1}(\alpha) := inf\{x \in \mathbb{R}: F(x) \geq \alpha\}$$

Siendo Y una variable aleatoria que representa la pérdida asociada a una determinada posición financiera y F su función de distribución.

El VaR es una medida estadística que establece una cantidad de pérdida máxima esperada en un intervalo de tiempo determinado y con cierto nivel de confianza. Sin embargo, en ocasiones, las pérdidas extremas pueden ser de gran preocupación para las empresas, por lo tanto, es importante complementar el VaR con otras medidas de riesgo,

El ES (*Expected shortfall*) o también conocido como VaR Condicional, por otro lado, proporciona información adicional sobre las pérdidas esperadas que pueden exceder el nivel máximo establecido por el VaR. Se puede definir de la siguiente forma:

$$\mathrm{ES}_{\alpha}(Y) = \frac{1}{1 - \alpha}\int_{\alpha}^{1} \mathrm{VaR}_{u}(Y)\mathrm{d}u$$

para $\alpha \in (0,1)$.

Si la función de distribución de Y es continua en su α-cuantil, el *expected shortfall* se puede simplificar como la esperanza de Y, condicionada a que Y sea mayor o igual que su α-cuantil.

$$\mathrm{ES}_{\alpha}(Y) = \frac{E\big(Y; Y \geq q_{\alpha}(Y)\big)}{1 - \alpha} = E(Y|Y \geq VaR_{\alpha}(Y))$$

Por lo tanto, dado un nivel de confianza $\alpha \in (0,1)$, entonces, $\mathrm{ES}_{\alpha}(Y) \geq \mathrm{VaR}_{\alpha}(Y)$.

Para realizar una comparación adecuada entre VaR y *expected shortfall*, es importante mencionar algunas propiedades deseables en una medida de riesgo [8][10]:

Se establece:

ρ es una función de valor real que representa la medida de riesgo.

L es una variable aleatoria que representa las pérdidas de carácter económico.

$\mathcal{M}$ es un espacio lineal de variables aleatorias.

1. Existencia de un recargo de seguridad positivo:

   Para todo $L \in \mathcal{M}$, se cumple que $\rho(L) \geq E[L]$.

   Esto quiere decir que es fundamental incluir un recargo de seguridad positivo en una medida de riesgo, ya que de esta manera se tiene en cuenta la incertidumbre y se protege contra eventos imprevistos.

2. Propiedad de "*no rip-off*":

   Para todo $L \in \mathcal{M}$, se cumple que $\rho(L) \leq \min\{l | F_L(l) = 1\}$.

   Esto quiere decir que la medida de riesgo tampoco debe asignar un valor excesivamente alto a la posible pérdida, ya que esto podría conducir a una sobreestimación del riesgo y, en consecuencia, a una toma de decisiones conservadora e inflexible. Por lo tanto, es importante buscar un equilibrio al evaluar el riesgo.

3. Monotonía:

   Para $L_1, L_2 \in \mathcal{M}$, si $L_1 \leq L_2$, entonces $\rho(L_1) \leq \rho(L_2)$.

   Esto implica que si una variable aleatoria L_1 es menos riesgosa que otra variable aleatoria L_2, entonces la medida de riesgo $\rho(L_1)$ debe ser menor o igual que la medida de riesgo $\rho(L_2)$. Desde un enfoque económico, este axioma es evidente: las posiciones que generan mayores pérdidas requieren más capital de riesgo.

4. Invariancia en translación:

   Para todo $L \in \mathcal{M}$ y para cualquier $I \in \mathbb{R}$, se cumple que $\rho(L + I) = \rho(L) + I$.

   Esto implica que si se desplaza una variable aleatoria L en una cantidad I, el riesgo asociado $\rho(L)$ también se desplaza en la misma cantidad I. Desde un enfoque económico, este axioma establece que al agregar o restar una cantidad determinada I a una posición que resulta en una pérdida L, se alteran los requisitos de capital exactamente en esa cantidad.

5. Subaditividad:

   Para todo $L_1, L_2 \in \mathcal{M}$, se cumple que $\rho(L_1 + L_2) \leq \rho(L_1) + \rho(L_2)$.

   Esto significa que el riesgo total asociado a la suma de dos variables aleatorias L_1 y L_2 debe ser menor o igual que la suma de los riesgos individuales $\rho(L_1)$ y $\rho(L_2)$. La razón detrás de este axioma se resume en la afirmación de Artzner et al. (1999) [2] de que "una fusión no crea riesgo adicional". Es decir, se argumenta que el riesgo puede reducirse mediante la diversificación, un principio ampliamente aceptado en finanzas y economía.

6. Homogeneidad positiva:

   Para todo $L \in \mathcal{M}$ y para todo $\lambda > 0$, se cumple que $\rho(\lambda L) = \lambda\rho(L)$.

   Esto significa que si se escala una variable aleatoria L por un factor positivo λ, el riesgo asociado $\rho(L)$ también se escala proporcionalmente. Este axioma refleja la idea de que el riesgo es proporcional al tamaño de la posición y se justifica fácilmente si se asume que el axioma anterior se cumple:

La subaditividad implica que, para cualquier número natural n, se tiene:

$$\rho(nL) = \rho(L + \cdots + L) \leq n\rho(L)$$

Dado que no existe diversificación entre las pérdidas en esta cartera (todas *L* tienen una correlación perfecta de 1), es razonable exigir que se cumpla la igualdad.

Los últimos cuatro axiomas establecen las propiedades que una medida de riesgo coherente debe cumplir.

Otra propiedad interesante es la propiedad de aditividad comonotónica, la cual establece que cuando las variables aleatorias L_1 y L_2 son comonótonas, es decir, se mueven en la misma dirección, la medida de riesgo de su suma es igual a la suma de las medidas de riesgo individuales. Matemáticamente se define como:

Para todo $L_1, L_2 \in \mathcal{M}$, se cumple que $\rho(L_1 + L_2) = \rho(L_1) + \rho(L_2)$.

Si una medida de riesgo coherente satisface además la propiedad de aditividad comonotónica, se la denomina medida de riesgo espectral (Acerbi, 2002) [1].

Aunque el VaR es monótono, invariante a la traslación y homogéneo positivo como cuantil de la distribución de pérdidas, su incapacidad para cumplir con la subaditividad lo limita como una medida coherente de riesgo y, por lo tanto, no se considera una medida de riesgo espectral (Artzner et al.1999) [2]. Por otro lado, el *expected shortfall* se destaca como una mejor alternativa, ya que es una medida de riesgo espectral que cumple con la coherencia y la aditividad comonótona.

2.3. REGRESIÓN CUANTÍLICA

La regresión lineal múltiple se enfoca únicamente en la estimación de la media de la variable de respuesta, lo cual no siempre proporciona una visión completa de los datos. Por otro lado, también es importante tener en cuenta que, en algunos conjuntos de datos, la media puede no ser un buen representante de los datos. Aquí hay algunos ejemplos de conjuntos de datos para los cuales la media puede no ser la estadística apropiada para estimar:

1. Datos con valores atípicos: Si un conjunto de datos tiene valores extremos o atípicos (puntos de datos observados que difieren notablemente del resto de los datos), estos pueden afectar significativamente el cálculo de la media. En tales casos, la media puede no reflejar adecuadamente el centro de los datos, y la mediana puede ser una medida más apropiada de tendencia central.

2. Datos sesgados: cuando la distribución de los datos es sesgada y la mayoría de los datos se agrupan en un extremo de la distribución, la media puede no ser una buena medida de tendencia central. Esto es especialmente común en situaciones en las que

la variable de interés tiene una distribución de cola pesada (*long tail*), como es el caso de los costes de siniestros de seguros. En tales casos, la moda o la mediana pueden ser más apropiadas que la media para representar tendencia central de la distribución de los datos.

3. Datos multi-modales: significa que hay más de un valor que ocurre con alta frecuencia. En un conjunto de datos bimodal con dos picos aproximadamente idénticos, la regresión lineal múltiple puede estimar un valor en el punto medio entre los dos picos, lo que obviamente no representa adecuadamente los datos. En estos casos, sería muy necesario explorar otros modelos de regresión más adecuados para obtener una representación más precisa de la relación entre las variables.

La regresión cuantílica es una alternativa propuesta por Koenker y Basset (1978) [6] que se enfoca en la estimación de diferentes percentiles de la distribución de la variable de respuesta, en contraposición a la media, como se hace en la regresión lineal por mínimos cuadrados ordinarios (MCO). Sin embargo, una de las principales diferencias entre ellas es que la regresión cuantílica no hace suposiciones sobre la distribución de la variable de respuesta, a diferencia de la MCO que asume que los errores de la regresión, es decir, las diferencias entre los valores observados de la variable dependiente y los valores predichos por el modelo siguen una distribución normal con media cero y varianza constante para cualquier valor dado de las variables (Otero y Reyes, 2012) [11]. Además, en caso de la presencia de datos atípicos o sesgados, la regresión cuantílica también puede ser más robusta que la MCO al no enfocar únicamente en la media.

La forma de especificar el modelo de regresión cuantílica es la siguiente:

$$y_i = X_i\beta_\alpha + u_i$$

donde y_i es el valor de la variable dependiente de la observación i, $X_i = (X_{i1}, \dots, X_{ik})$ son las variables explicativas, β_α son los parámetros de la estimación correspondiente al cuantil α, y u_i es la perturbación aleatoria correspondiente al cuantil α.

Partiendo de la suposición de que:

$$Q_\alpha(u_i \mid X_i) = 0$$

El modelo de regresión de cuantiles se puede reformular de la siguiente manera:

$$Q_\alpha(y_i \mid X_i) = X_i\beta_\alpha$$

siendo α los cuantiles, tales que $\alpha \in (0, 1)$.

Otra diferencia importante es que la MCO estima los coeficientes de la regresión que minimizan la suma de los errores al cuadrado, mientras que la regresión cuantílica busca minimizar la suma ponderada de las diferencias absolutas entre los valores observados y los valores predichos en diferentes percentiles de la distribución de la variable de respuesta.

Para obtener los parámetros β_α, simplemente se debe resolver el siguiente problema de optimización lineal de minimización.

$$\min_{\beta_\alpha}[\sum_{i:\{y_i \geqslant X\beta_\alpha\}} \alpha|y_i - X_i\beta_\alpha| + \sum_{i:\{y_i < X\beta_\alpha\}} (1-\alpha)|y_i - X_i\beta_\alpha|]$$

Los parámetros β_α indican cómo varía el α-ésimo cuantil de la variable respuesta y_i cuando se modifica en una unidad los valores de las variables regresoras X_i, manteniendo constantes las demás variables.

En el contexto de la regresión cuantílica, existen diversos métodos algorítmicos utilizados para calcular el ajuste del modelo (Koenker, 2005) [7]:

- "br" (Barrodale y Roberts): el método clásico, un método eficiente para problemas de tamaño moderado y puede computar la regresión cuantílica completa. Permite calcular intervalos de confianza para los parámetros estimados.

- "fn" (Frisch-Newton): un método adecuado para problemas más grandes que utiliza el algoritmo del punto interior.

- "pfn" (Frisch-Newton con preprocesamiento): una variante del método "fn" que utiliza una estrategia de preprocesamiento descrita en el estudio de Portnoy y Koenker (1997), un método adecuado para problemas con un gran número de observaciones y una dimensión paramétrica modesta.

- "sfn" (Frisch-Newton con álgebra dispersa): una variante del método "fn" que aprovecha el álgebra dispersa para calcular los iterados con una matriz de diseño dispersa, un método eficiente para problemas grandes con una dimensión paramétrica grande.

- "fnc" (Frisch-Newton con restricciones lineales): una variante del método "fn" que permite especificar restricciones lineales en los coeficientes estimados.

- "conquer" (Conquer): un método para problemas muy grandes, especialmente aquellos con una dimensión paramétrica grande. Utiliza el paquete "conquer" de R y el método de bootstrap para el cálculo de intervalos de confianza.

- "lasso": una variante del método "fn" que implementa la penalización lasso y la penalización de desviación absoluta recortada suavizada de Fan y Li, respectivamente.

En un estudio realizado por Koenker (2005) [7], se demostró que el algoritmo simplex propuesto por Barrodale y Roberts resulta ser altamente competitivo cuando el tamaño de la muestra es inferior a 25,000. Sin embargo, para tamaños de muestra más grandes, el algoritmo Frisch-Newton se destaca por su mayor eficiencia. Asimismo, se observó que la realización de un preprocesamiento de los datos otorga ventajas significativas en casos

en los que el tamaño de la muestra supera los 100,000. No obstante, estas ventajas se atenúan en cierta medida al aumentar la dimensión paramétrica del modelo.

En resumen, estos métodos ofrecen diferentes enfoques para resolver el problema de regresión cuantílica y la elección entre ellos depende de las características específicas del problema, como el tamaño de la muestra, la dimensión paramétrica del modelo y las restricciones que se apliquen.

2.4. REGRESIÓN CONJUNTA (VaR & ES)

Una propiedad matemática importante que deben cumplir las medidas de riesgo para poder modelar de manera efectiva la distribución de pérdidas financieras y estimar los parámetros de los modelos de riesgo que es la elicitabilidad. Un funcional estadístico, como la media o la mediana, se denomina elicitable si se pueden obtener estimaciones consistentes de los parámetros de la distribución de tal manera que dichos parámetros estimados sean los únicos minimizadores de una función *scoring* específica (Fissler et al., 2016) [4].

Sea Y una variable aleatoria con una distribución desconocida F. La siguiente definición formaliza la noción de una *scoring* consistente para $\mathrm{T}_\alpha(F)$ (un funcional de F) (Ziegel et al., 2017) [14].

Una función *scoring* $\mathrm{S} : \mathrm{A}_0 \times \mathbb{R} \to \mathbb{R}$ es una función tal que $\int \mathrm{S}(x, y)\mathrm{d}F(y)$ existe para todo $F \in \mathcal{F}_1, x \in \mathrm{A}_0$. En este contexto, la clase $\mathcal{F}_1$ se refiere a las distribuciones que tienen media finita y cuantiles únicos. Se dice que la función *scoring* S es consistente para T_α si se cumple

$$\mathbb{E}(\mathrm{S}(\mathrm{T}_\alpha(F), Y)) \leq \mathbb{E}(\mathrm{S}(x, Y))$$

para todo $x \in \mathrm{A}_0$ y todas las variables aleatorias Y con distribución en $\mathcal{F}_1$. La función *scoring* S es estrictamente consistente si se cumple la igualdad $x = \mathrm{T}_\alpha(F)$.

En el caso del VaR, se considera elicitable, es decir, se puede definir como el minimizador de la esperanza de una función *scoring* adecuada.

En cuanto al *expected shortfall*, a diferencia del VaR, no cumple con la propiedad de 1-elicitable según estudios realizados por Weber (2006) [12] y Gneiting (2011) [5]. Se ha mostrado que no existe una función *scoring* (o una función de pérdida) estrictamente consistente $S : \mathbb{R}^2 \to \mathbb{R}$ tal que, para cualquier variable aleatoria Y con media finita, se cumpla:

$$ES_\alpha(Y) = \arg min_{e \in \mathbb{R}} \mathbb{E}[S(e, Y)]$$

mientras que el VaR lo cumple para la mayoría de las clases de distribuciones F y una posible función *scoring* estrictamente consistentes para el VaR tiene la siguiente forma:

$$S_V(v,x) = (\mathbb{1}\{y \leq v\} - \alpha)\big(G(v) - G(y)\big)$$

donde G es una función estrictamente creciente.

Esto ha sido criticado como una limitación importante para los procesos de *backtesting* del ES, ya que no se puede evaluar su capacidad predictiva de manera consistente y objetiva. Pero, de hecho, excepto por la esperanza, todas las medidas de riesgo espectral no cumplen con la propiedad de 1-elicitable, como se observa en el estudio de Ziegel (2015) [13].

Sin embargo, se han propuesto métodos alternativos para aproximar el *expected shortfall* y hacer posible su *backtesting*, como la elicibilidad conjunta propuesta por Fissler et al. (2016) [4], que demuestran que el *expected shortfall* es conjuntamente elicitable con el VaR.

$$\mathrm{T}_\alpha(F) = (VaR_\alpha(F), ES_\alpha(F))'$$

La elicitabilidad conjunta se refiere a la capacidad de una función de pérdida para permitir la estimación conjunta de varios parámetros de la distribución de pérdidas. En el trabajo de Dimitriadis y Bayer (2019) [3] se ha introducido un modelo de regresión conjunta para el cuantil y *expected shortfall* de una variable de respuesta dada un conjunto de covariables. Se ha propuesto un estimador M[4] utilizando una clase de funciones de *scoring* que cumple con las propiedades de la elicitabilidad conjunta para los parámetros de regresión conjuntos de ambos modelos. Además, se ha demostrado que estos estimadores son consistentes y asintóticamente normales.

$$(\mathrm{VaR}_\alpha\,(Y), \mathrm{ES}_\alpha(Y)) = \underset{(v,e)\in\mathbb{R}^2}{arg\,min}\,\mathbb{E}[S_{V,E}(v,e,Y)]$$

donde $S_{V,E}$ son las posibles funciones *scoring*.

Basándose en este resultado de elicitabilidad conjunta, se puede definir el modelo de regresión conjunta de la siguiente forma:

Sea $Y \in \mathbb{R}$ una variable aleatoria que describe el rendimiento obtenido en un periodo de tiempo dado en una cartera de inversión, donde un rendimiento negativo, $Y < 0$, corresponde a una pérdida. Y dado los covariables $X \in \mathbb{R}^p$ y $\beta_0^q, \theta_0^e \in \mathbb{R}^p$ que son los parámetros desconocidos para el cuantil y el ES, respectivamente.

$$Y = X_q^T\beta_0^q + u^q \qquad and \qquad Y = X_e^T\theta_0^e + u^e$$

donde asumen que:

$$Q_\alpha(u^q|X) = 0 \qquad and \qquad ES_\alpha(u^e|X) = 0$$

[4] Véase la definición de Estimador M: https://es.frwiki.wiki/wiki/M-estimateur

De esta manera, se garantiza que el modelo conjunto esté correctamente especificado, en el sentido de que:

$$Q_\alpha(Y|X) = X_q^T \beta_0^q \qquad and \qquad ES_\alpha(Y|X) = X_e^T \theta_0^e$$

Con esta construcción del modelo, se permite que los modelos de cuantil y ES dependan de diferentes vectores de covariables X_q y X_e, respectivamente. Sin embargo, es importante asegurarse de que el cuantil no dependa de las covariables del ES. Esto implica que al incluir covariables únicamente en el modelo de ES o únicamente en el modelo de cuantil, se asume que estas covariables no tienen un efecto directo en la estimación del cuantil condicional. Esta suposición es necesaria para garantizar la correcta especificación del modelo y la interpretación adecuada de los resultados.

Asimismo, el procedimiento de estimación M para el vector de parámetros de regresión conjunta $\beta_0^q \; y \; \theta_0^e$, basado en la utilización de la función *scoring* conjunta estrictamente consistente para la estimación del cuantil y el ES propuesta por Fissler et al. (2016) [4], tiene la siguiente forma:

Se establece:

G_1 : es una función creciente, integrable y dos veces diferenciable.

a : es una función integrable pero generalmente se establece en cero porque solo depende de Y.

$\mathcal{G}_2$: es una función diferenciable tres veces de manera continua tal que tanto $\mathcal{G}_2$ como su derivada G_2 ($\mathcal{G}_2^{(1)} = G_2$) y $G_2^{(1)}$ sean estrictamente positivas.

Y $G_1, G_2, \mathcal{G}_2, a : \mathbb{R}^2 \to \mathbb{R}$.

La función *scoring* conjunta tiene la siguiente forma:

$$\begin{aligned} S(Y, X, \beta, \theta) = {} & (\mathbb{1}_{\{Y \le X_q' \beta^q\}} - \alpha) G_1(X_q' \beta^q) - \mathbb{1}_{\{Y \le X_q' \beta^q\}} G_1(Y) \\ & + G_2(X_e' \theta^e)(X_e' \theta^e - X_q' \beta^q + \frac{(X_q' \beta^q - Y)\mathbb{1}_{\{Y \le X_q' \beta^q\}}}{\alpha}) \\ & - \mathcal{G}_2(X_e' \theta^e) + a(Y) \end{aligned}$$

Una posible interpretación sería: el primer sumando en (2) es una función *scoring* estrictamente consistente para VaR, dada en (1), y por lo tanto solo depende del cuantil. Por otro lado, el segundo sumando se puede considerar como una combinación de componentes que dependen tanto del cuantil como del *expected shortfall*. Esta estructura ilustra el hecho de que el ES en sí mismo no es 1-elicitable, pero puede ser considerado 2-elicitable junto con el VaR.

Y los estimadores M ($\widetilde{\beta_0^q} \; y \; \widetilde{\theta_0^e}$) correspondientes para las covariables, que exhiben consistencia y normalidad asintótica, se definen de la siguiente forma:

$$\begin{pmatrix}\tilde{\beta}\\ \tilde{\theta}\end{pmatrix} \in \underset{\beta,\theta\in\Theta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} S(Y_i, X_i, \beta, \theta)$$

donde $\Theta \subseteq \mathbb{R}^p$es el espacio de parámetros, asumiendo que es compacto, convexo y tiene interior no vacío.

3. APLICACIONES DE LAS REGRESIONES

3.1. DESCRIPCIÓN DE LA BASE DE DATOS

Los datos utilizados en el presente estudio provienen de la librería CASdatasets[5] del programa R y se encuentran en la sección "freMTPL" (French Motor Third-Part Liability datasets), la cual contiene conjuntos de datos de responsabilidad civil de terceros para automóviles en Francia. Específicamente, se utilizan dos conjuntos de datos: freMTPL2freq y freMTPL2sev.

Se ha realizado una combinación de los conjuntos de datos freMTPL2freq y freMTPL2sev utilizando la función *merge()* de R, utilizando la columna de ID de póliza como clave de unión. Esta operación ha permitido fusionar la información sobre las características de riesgo y el número de reclamaciones por póliza del conjunto de datos freMTPL2freq con la información sobre la gravedad de las reclamaciones y el ID de la póliza correspondiente del conjunto de datos freMTPL2sev.

Estos conjuntos de datos recopilan características de riesgo de 677,991 pólizas de responsabilidad civil de terceros para vehículos motorizados, observadas principalmente durante un año. Sin embargo, dado que la variable respuesta es el costo de siniestros, se han eliminado aquellas pólizas que no tuvieron siniestros, lo que ha resultado en una base de datos que contiene únicamente 24,943 pólizas con siniestros.

El conjunto de datos incluye un total de 13 variables que describen diferentes aspectos relacionados con las pólizas y los siniestros. Estas variables son:

1. **IDpol**: Identificador único de cada póliza de responsabilidad civil de terceros.
2. **ClaimNb**: Número de siniestros durante el periodo de exposición
3. **Exposure**: Periodo de exposición de la póliza, en años
4. **VehPower**: Potencia del vehículo
5. **VehAge**: Edad del vehículo, en años

[5] Véase la documentación oficial en el repositorio de los datos: http://cas.uqam.ca/pub/

6. **DrivAge**: Edad del conductor principal del vehículo asegurado, en años, considerando que en Francia se puede conducir un automóvil a partir de los 18 años

7. **BonusMalus**: Bonus/Malus, entre 50 y 350: <100 significa bonus, >=100 significa malus en Francia

8. **VehBrand**: Marca del coche

9. **VehGas**: Tipo de combustible del coche, diesel o combustible regular

10. **Area**: Comunidad donde reside el conductor varía en densidad, desde "A" para áreas rurales hasta "F" para centros urbanos

11. **Density**: Densidad de habitantes en la ciudad de residencia del conductor (número de habitantes por kilómetro cuadrado)

12. **Region**: Región de la póliza en Francia (basada en la clasificación de 1970-2015)

13. **ClaimAmount**: Costo total de los siniestros ocurridos en la póliza

Donde "ClaimAmount" es la variable respuesta que pretendemos predecir en nuestros modelos.

Se presentan a modo de ilustración las primeras 6 pólizas de la base de datos.

	IDpol	ClaimNb	Exposure	VehPower	VehAge	DrivAge	BonusMalus	VehBrand	VehGas	Area	Density	Region	ClaimAmount
1	139	1	0.75	7	1	61	50	B12	Regular	F	27000	Ile-de-France	303.00
2	190	1	0.14	12	5	50	60	B12	Diesel	B	56	Basse-Normandie	1981.84
3	414	1	0.14	4	0	36	85	B12	Regular	E	4792	Ile-de-France	1456.55
4	424	2	0.62	10	0	51	100	B12	Regular	F	27000	Ile-de-France	10834.00
5	463	1	0.31	5	0	45	50	B12	Regular	A	12	Midi-Pyrenees	3986.67
6	606	1	0.84	10	6	54	50	B12	Diesel	D	583	Provence-Alpes-Cotes-D'Azur	1840.14

Ilustración 1. Representación de las primeras 6 filas de la base de datos. Fuente: Elaboración propia

Esta base de datos se utilizará para llevar a cabo análisis estadísticos y desarrollar modelos predictivos, como regresión de cuantiles, regresión conjunta de cuantiles y *expected shortfall*, con el objetivo de obtener *insights* valiosos sobre los factores de riesgo en el contexto de los seguros de automóviles.

En la ilustración 2 se muestran los principales indicadores estadísticos de las variables numéricas, incluyendo la cantidad de observaciones, la media, la desviación estándar, el valor mínimo y máximo. Por otro lado, en la ilustración 3 se presenta un resumen de las variables categóricas.

Statistic	N	Mean	St. Dev.	Min	Max
Exposure	24,943	0.693	0.314	0.003	2.000
VehPower	24,943	6.469	2.013	4	15
VehAge	24,943	7.386	5.174	0	99
DrivAge	24,943	45.140	14.659	18	99
BonusMalus	24,943	64.931	19.865	50	228
Density	24,943	1,984.353	4,119.796	2	27,000
ClaimAmount	24,943	2,401.845	30,265.360	1.000	4,075,401.000

Ilustración 2. Resumen de las variables numéricas. Fuente: Elaboración propia

	skim_type	skim_variable	n_missing	complete_rate	factor.n_unique
1	factor	VehBrand	0	1	11
2	factor	VehGas	0	1	2
3	factor	Area	0	1	6
4	factor	Region	0	1	21

Ilustración 3. Resumen de las variables categóricas. Fuente: Elaboración propia

Al analizar los resultados, podemos concluir que la base de datos no contiene valores faltantes (*NA*) en ninguna de las variables, ya que el número de observaciones (N) es consistente para todas ellas. Además, el rango de valores de cada variable es razonable. Estos indicadores nos proporcionan una visión general de las características y distribución de los datos analizados.

3.1.1. Exploración Estadística de Variables Numéricas

La primera variable que analizamos es "ClaimNb", que representa el número de reclamos presentados por los asegurados. Según la ilustración 4, la mayoría de los asegurados (23,569) ha tenido un reclamo, mientras que un número menor ha presentado múltiples reclamos, siendo el valor máximo de reclamos igual a 16.

ClaimNb	Freq
1	23569
2	1299
3	62
4	5
5	2
6	1
8	1
9	1
11	2
16	1

Ilustración 4. Tabla de frecuencias de la variable "ClaimNb". Fuente: Elaboración propia

Esta distribución es un resultado esperado en el contexto de un seguro de automóviles, ya que es común que la mayoría de los asegurados tengan un historial de reclamos limitado.

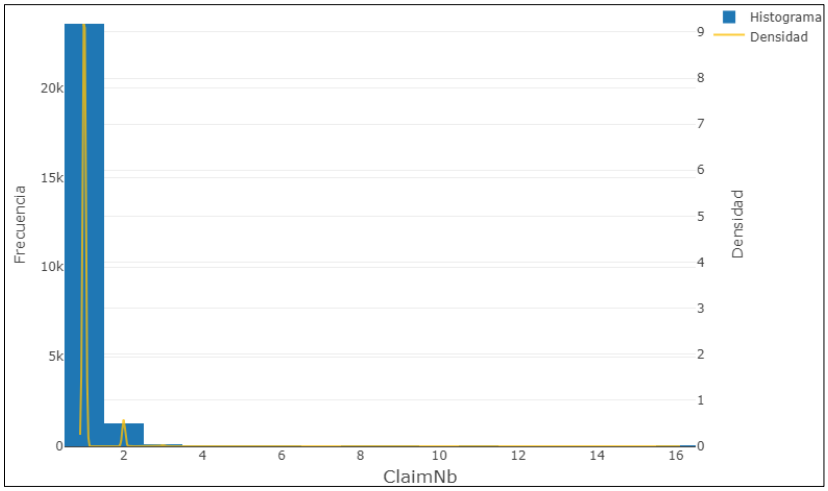

Ilustración 5. Histograma de la variable "ClaimNb". Fuente: Elaboración propia

Al observar el histograma de esta variable, podemos notar que el pico más alto se encuentra en el extremo izquierdo del histograma y que se visualizan algunos valores atípicos en el extremo derecho del histograma. Esto confirma la conclusión obtenida de la tabla de frecuencias. Además, la curva de densidad muestra una distribución sesgada hacia la derecha, lo que indica que la media es mayor que la mediana.

La exposición en el seguro de automóviles se refiere a la duración en la que el conductor está protegido por la póliza y puede presentar reclamos en caso de siniestros. Como se observa en la ilustración 6, la mayoría de los asegurados tienen una exposición cercana a 1 año, que es la duración habitual de un contrato de seguro.

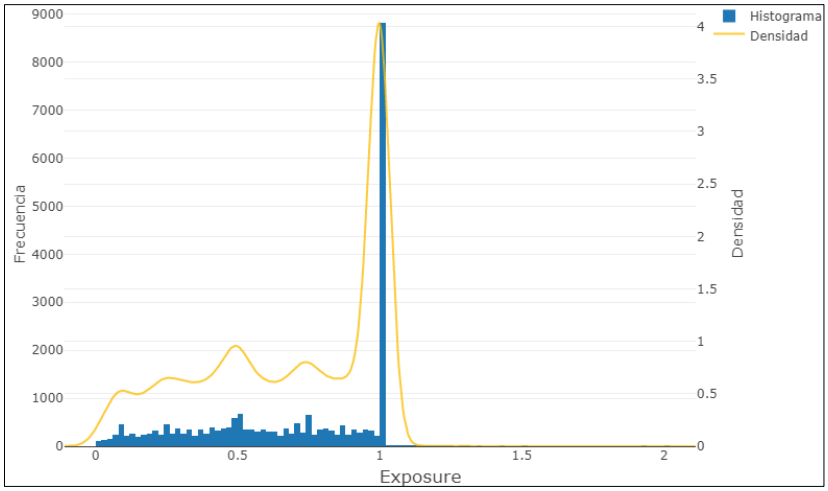

Ilustración 6. Histograma de la variable "Exposure". Fuente: Elaboración propia

Según las estadísticas descriptivas de la ilustración 7, el rango de valores de la variable "Exposure" va desde 0.003 hasta 2 años, con un promedio de 0.693 años, lo que equivale aproximadamente a 253 días. Además, la desviación estándar es relativamente baja en comparación con el promedio, lo que indica que los valores de exposición están bastante concentrados alrededor de la media.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.Dev
0.003	0.450	0.760	0.693	1	2	0.314

Ilustración 7. Descripción estadística de la variable "Exposure". Fuente: Elaboración propia

En realidad, la duración de la exposición en sí misma no es un indicador directo de los costos de los reclamos. es decir, los costos de los reclamos no están directamente influenciados por la duración del período de exposición. No obstante, sería interesante crear una variable llamada "Frequency" que represente la relación entre el número de reclamos y la duración de la exposición. Esta variable nos permitiría calcular el número de reclamos anuales para todas las pólizas, asumiendo un comportamiento homogéneo a lo largo del año.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St.Dev	Asimetría	Curtosis
0.500	1	1.350	3.224	2.440	365	10.070	16.404	395.735

Ilustración 8. Descripción estadística de la variable "Frequency". Fuente: Elaboración propia

La ilustración 8 proporciona información sobre la variable "Frequency", la cual varía desde un mínimo de 0.500 hasta un máximo de 365. Además, se observa una asimetría positiva en los datos y una curtosis extremadamente alta. Esto revela que la distribución de la variable es muy sesgada hacia la derecha y tiene una concentración muy alta de datos en el rango inferior, con muy pocos valores más altos.

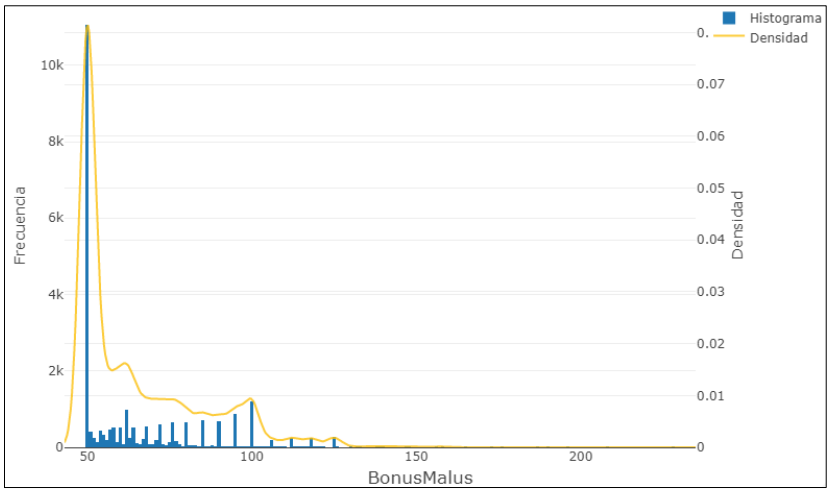

Ilustración 9. Histograma de la variable "BonusMalus". Fuente: Elaboración propia

El "Bonus/Malus" es un sistema utilizado en Francia para calcular las primas de seguro de automóviles. Se basa en un sistema de bonificación y penalización según el historial de conducción del asegurado. Un valor de "BonusMalus" inferior a 100 se considera "Bonus" y significa que el conductor tiene un historial de conducción seguro y sin

siniestros. Por otro lado, un valor igual o superior a 100 se considera "Malus" y significa que el conductor ha tenido siniestros. Según el histograma, se puede observar que los datos están concentrados en el lado izquierdo, lo cual indica que la gran mayoría de los conductores tienen un buen historial de conducción.

	Casos	Porcentaje
Bonus	22,607	90.63%
Malus	2,336	9.37%

Ilustración 10. Tabla de frecuencia de la variable "BonusMalus". Fuente: Elaboración propia

La ilustración 10 muestra que hay 22,607 casos (90.63%) clasificados como "Bonus" y 2,336 casos (9.37%) clasificados como "Malus". Esto significa que la mayoría de los conductores se benefician de descuentos en las primas de seguro. Por otro lado, los conductores clasificados como "Malus" enfrentan aumentos en las primas debido a reclamaciones de siniestros anteriores.

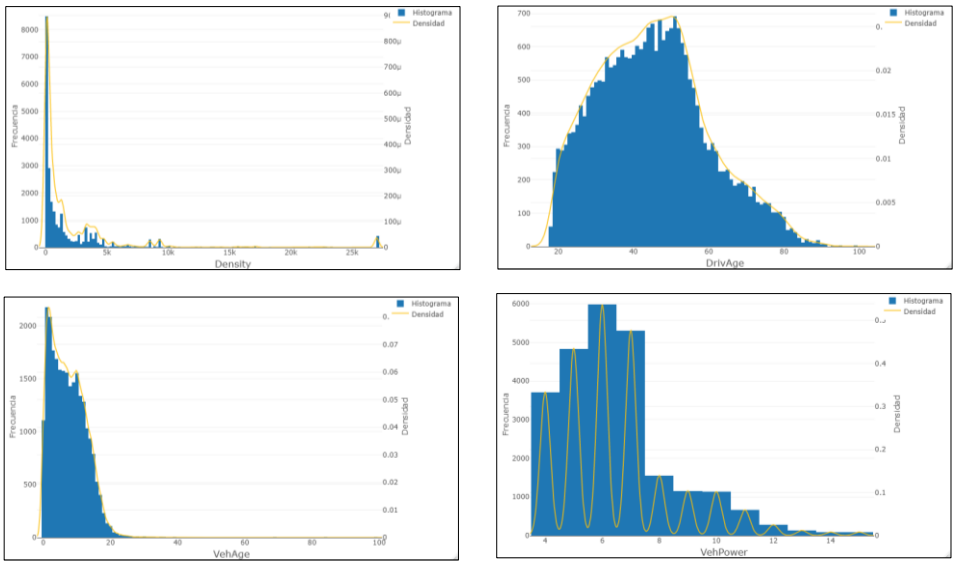

Ilustración 11. Histograma de las variables "Density", "DrivAge", "VehAge" y "VehPower". Fuente: Elaboración propia

Al observar la ilustración 11, podemos notar que la distribución de la densidad de habitantes es asimétrica y sesgada hacia la derecha, lo cual indica que la mayoría de los conductores residen en áreas con una densidad de habitantes relativamente baja.

En relación a la edad del conductor asegurado, se observa que la variable cubre un rango amplio, desde los 18 hasta los 99 años, lo cual concuerda con la distribución de edades de la población de conductores en general. Por otro lado, el histograma muestra una distribución aproximadamente simétrica y de forma de campana, lo que sugiere que la edad de los conductores en el conjunto de datos está distribuida de manera relativamente equilibrada. La concentración se encuentra cerca de los valores medios de la muestra, alrededor de los 45 años.

Para la edad del vehículo y la potencia, se puede observar que ambas presentan asimetría hacia la derecha. Esto indica que la mayoría de los vehículos de las pólizas analizadas son relativamente nuevos. Además, en cuanto a la potencia, aunque existen algunos vehículos con una potencia muy alta, la mayoría de los coches tienen una potencia media o baja.

Variable dependiente: "ClaimAmount"

La variable dependiente en esta base de datos es "ClaimAmount", que representa el costo total de siniestros y se distribuye de la siguiente manera:

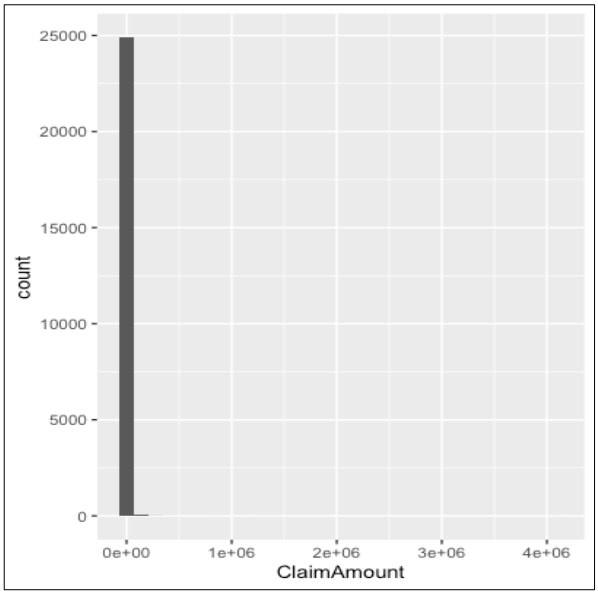

Ilustración 12. Histograma de la variable "ClaimAmount". Fuente: Elaboración propia

La ilustración 12 muestra la distribución de los valores de "ClaimAmount" en la base de datos. Sin embargo, debido a la presencia de un valor extremo muy alto, la escala del eje X se ha extendido para incluir este valor, lo que hace que la mayoría de los valores se agrupen en el lado izquierdo del histograma y que la parte derecha del histograma parezca vacía. Este patrón indica que la mayoría de los valores de "ClaimAmount" están concentrados en un rango estrecho, pero hay un valor atípico o *outlier* que se encuentra significativamente por encima de los demás valores. Sin embargo, esto es precisamente lo que se desea analizar, ya que el interés del trabajo radica en analizar la cola de la derecha de la distribución y observar cómo se comportan las regresiones con valores extremos.

Los siguientes seis perfiles corresponden a las pólizas con los costos de siniestros más altos.

	IDpol	ClaimNb	Exposure	VehPower	VehAge	DrivAge	BonusMalus	VehBrand	VehGas	Area	Density	Region	ClaimAmount	Frequency
7488	1120377	1	0.22	9	13	19	100	B2	Regular	B	93	Centre	4075400.6	4.55
2843	110846	2	0.43	6	13	20	100	B1	Regular	C	203	Centre	1404185.5	4.65
12730	2141337	1	0.32	4	14	18	100	B2	Regular	D	863	Rhone−Alpes	1301172.6	3.12
17084	3122016	1	0.91	7	7	40	63	B11	Diesel	E	9307	Rhone−Alpes	774411.5	1.10
9765	2008127	3	0.36	4	2	57	50	B4	Regular	D	1217	Rhone−Alpes	399213.7	8.33
15354	3025890	2	0.09	7	1	36	50	B12	Diesel	A	38	Champagne−Ardenne	382955.1	22.22

Ilustración 13. Las seis pólizas con mayor coste de siniestros. Fuente: Elaboración propia

n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
24,943	2,401.845	30,265.360	1,172	1,182.931	458.642	1	4,075,401.000	4,075,400.000	106.130	13,489.760	191.634

Ilustración 14. Estadísticas descriptivas de la variable "ClaimAmount". Fuente: Elaboración propia

La ilustración 14 muestra varias estadísticas descriptivas de la variable respuesta "ClaimAmount". En particular, se puede observar que los datos presentan una dispersión considerable, ya que la desviación estándar es alta y el rango es amplio, lo que indica una variabilidad significativa en los valores. La mediana, que es el valor que se encuentra en el centro de los datos, es relativamente pequeña en comparación con la media.

La media truncada (trimmed) y la desviación absoluta mediana (mad) son medidas robustas que proporcionan una estimación de la tendencia central y la dispersión, respectivamente. Estas medidas son menos sensibles a los valores atípicos. Al observar los valores, vemos que son considerablemente menores que la media y la desviación estándar, lo cual indica que la distribución es altamente sesgada.

Además, el coeficiente de asimetría es positivo, lo que sugiere una asimetría hacia la derecha en la distribución de los datos. Por último, el coeficiente de curtosis es elevado, lo que indica una distribución con picos pronunciados y colas pesadas.

A continuación, se muestra el *boxplot* y el *QQ-plot* de la variable 'ClaimAmount', ya que estos gráficos son muy útiles cuando se desea visualizar la distribución y detectar posibles valores atípicos en los datos.

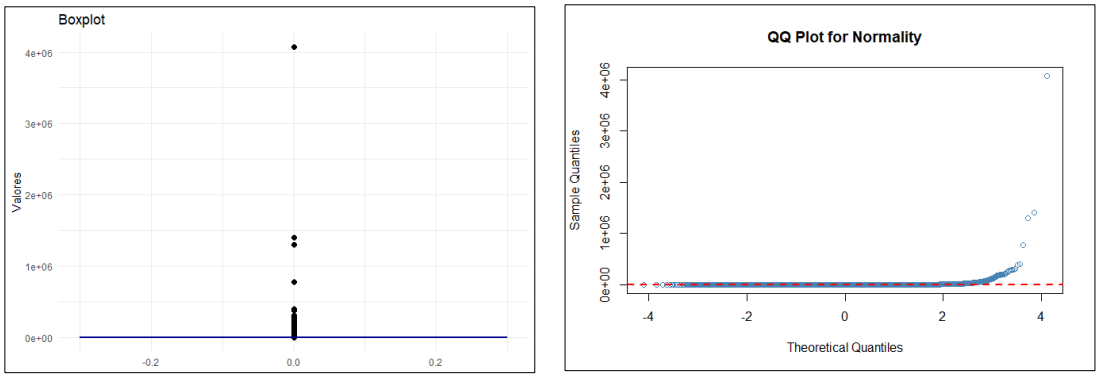

Ilustración 15. Boxplot y QQ-plot de la variable "ClaimAmount". Fuente: Elaboración propia

El *boxplot* revela una distribución con una presencia notable de valores atípicos muy altos. Además, el *QQ-plot* indica claramente que la distribución no se ajusta bien a una distribución normal, con una desviación notable en la parte final.

	statistic	p.value	method
1	(A = 8553.28068709451)	3.7e-24	Anderson-Darling normality test
2	(D = 0.468386464682361)	0	Lilliefors (Kolmogorov-Smirnov) normality test

Ilustración 16. Test de normalidad para la variable "ClaimAmount". Fuente: Elaboración propia

Las pruebas de normalidad Anderson-Darling y Lilliefors (Kolmogorov-Smirnov) realizadas en la variable "ClaimAmount" del conjunto de datos "freMTPL2" muestran resultados similares. Ambas pruebas indican que los datos no siguen una distribución normal, ya que los p-valores son extremadamente pequeños. Por lo tanto, se puede concluir que la variable respuesta "ClaimAmount" no proviene de una población con distribución normal.

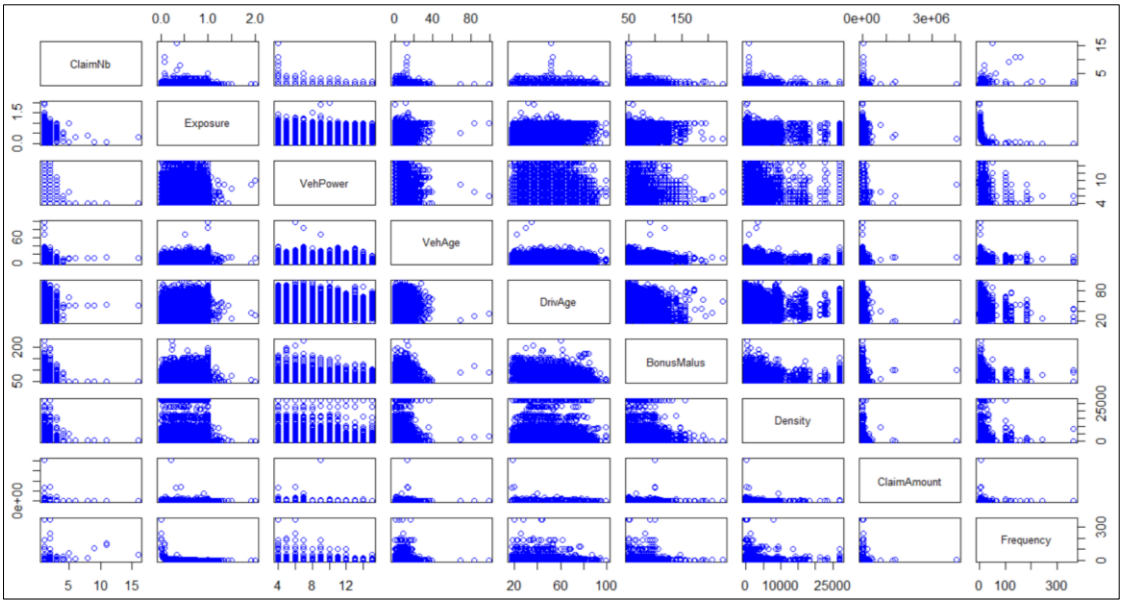

Ilustración 17. Representación gráfica de dispersión entre las variables numéricas. Fuente: Elaboración propia

En la ilustración 17, se emplean gráficos de dispersión para representar visualmente los valores de dos variables numéricas distintas. La posición de cada punto en el eje horizontal y vertical indica un par de valores correspondientes a las dos variables. Esto nos permite observar las relaciones entre nuestras variables numéricas.

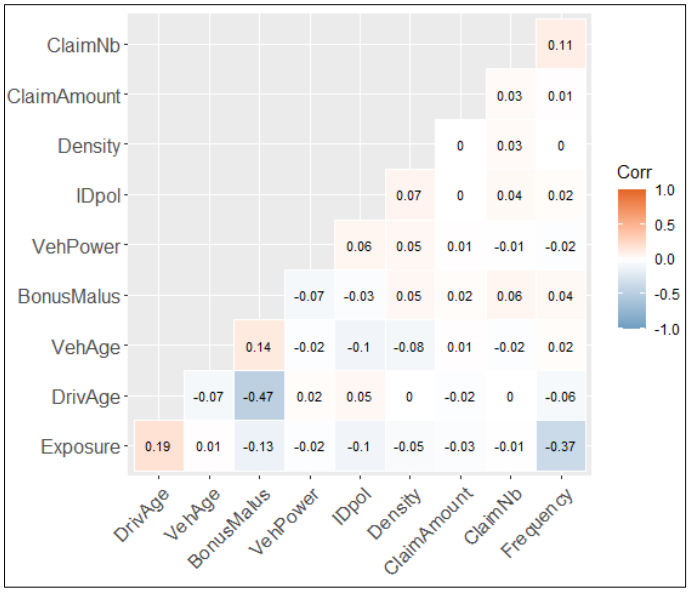

Ilustración 18. Representación gráfica de las correlaciones entre las variables numéricas. Fuente: Elaboración propia

En el gráfico de correlaciones, se destaca una correlación positiva (0.19) entre la exposición y la edad del conductor, mientras que el bonus/malus muestra una correlación negativa (-0.47) con la edad del conductor. Sin embargo, la mayoría de las correlaciones entre las variables son muy débiles, lo que sugiere una baja dependencia entre ellas.

3.1.2. Exploración Estadística de Variables Categóricas

Ahora procederemos a analizar las cuatro variables categóricas: “VehBrand”, “VehGas”, “Area” y “Region”.

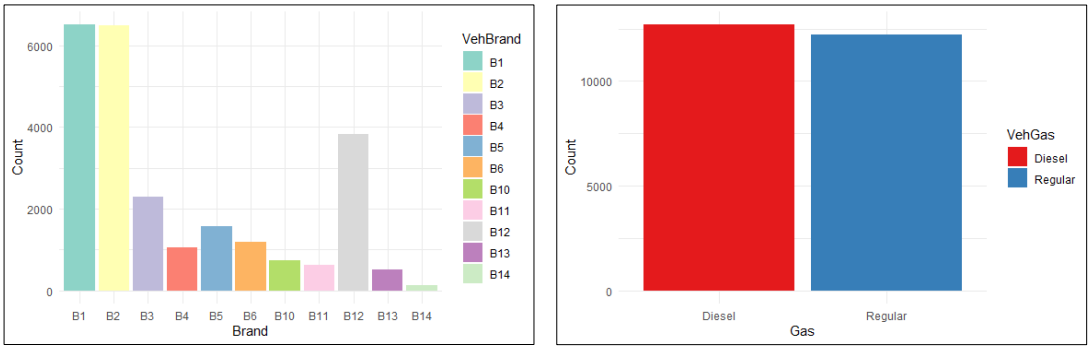

Ilustración 19. Frecuencia de las categorías de las variables “VehBrand” y “VehGas”. Fuente: Elaboración propia

Con la ilustración 19, se puede apreciar que las marcas más comunes en las pólizas son B1, B2 y B12. Además, se observa una distribución equilibrada de pólizas entre coches de Diesel y gasolina regular.

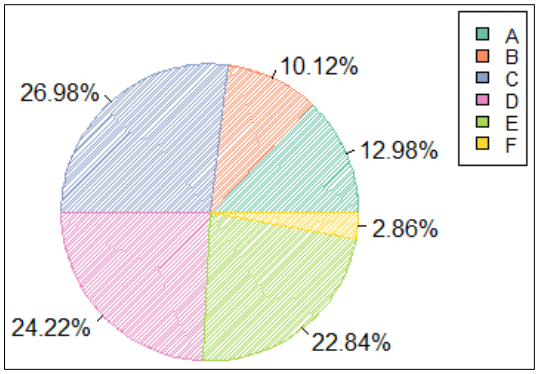

Ilustración 20. Pie Chart de la variable "Area". Fuente: Elaboración propia

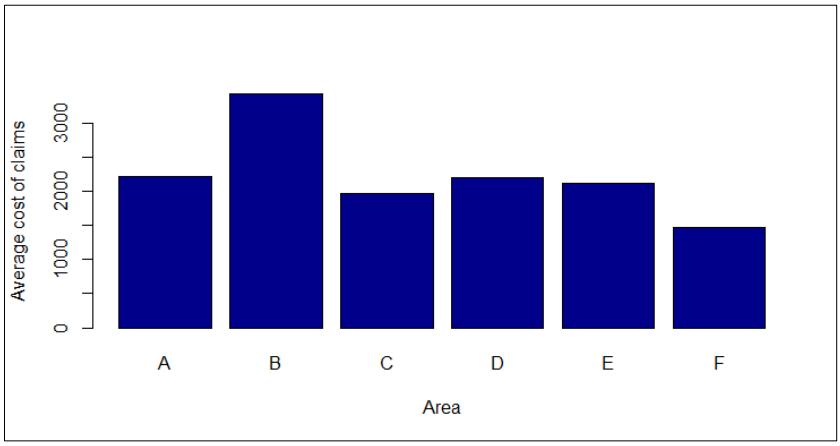

Ilustración 21. Representación gráfica del coste medio de siniestros para cada área de la variable "Area". Fuente: Elaboración propia

La ilustración 20 representan la cantidad acumulada de pólizas en diferentes áreas. El área C tiene el mayor número de pólizas (26.98%), mientras que el área F tiene el menor número (2.86%). En el *barplot* de la ilustración 21, se observa que el área B tiene el mayor coste medio por siniestro, mientras que el área con menor coste medio por siniestro es el área F.

Region	Freq	Porcentaje
Alsace	83	0.33%
Aquitaine	985	3.95%
Auvergne	135	0.54%
Basse-Normandie	429	1.72%
Bourgogne	328	1.31%
Bretagne	1793	7.19%
Centre	6262	25.11%
Champagne-Ardenne	70	0.28%
Corse	109	0.44%
Franche-Comte	36	0.14%
Haute-Normandie	209	0.84%
Ile-de-France	2392	9.59%
Languedoc-Roussillon	934	3.74%
Limousin	184	0.74%
Midi-Pyrenees	350	1.4%
Nord-Pas-de-Calais	1324	5.31%
Pays-de-la-Loire	1498	6.01%
Picardie	302	1.21%
Poitou-Charentes	765	3.07%
Provence-Alpes-Cotes-D'Azur	2753	11.04%
Rhone-Alpes	4002	16.04%

Ilustración 22. Tabla de frecuencia de la variable "Region". Fuente: Elaboración propia

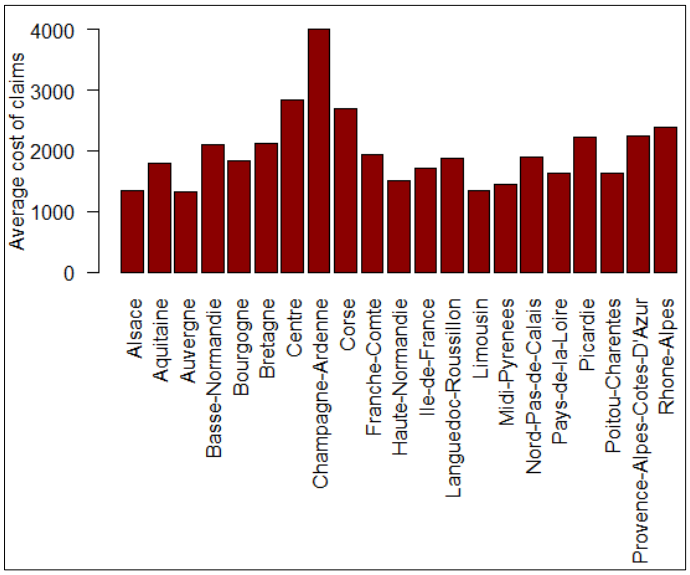

Ilustración 23. Representación gráfica del coste medio de siniestros para cada región de la variable "Region". Fuente: Elaboración propia

En el análisis de las regiones, se observa que Centre, Ile-de-France, Provence-Alpes-Cotes-D'Azur y Rhone-Alpes son las regiones con mayor número de pólizas con siniestros. Por otro lado, Alsace, Champagne-Ardenne y Franche-Comte son las regiones con menor número de pólizas con siniestros. En cuanto al coste medio por siniestro, se destaca que Champagne-Ardenne tiene el mayor coste medio, seguida de Centre y Corse, mientras que Auvergne presenta el coste medio más bajo.

3.2. RESULTADOS DE LA REGRESIÓN CUANTÍLICA

A continuación, ajustaremos un modelo de regresión cuantílica utilizando "ClaimAmount" como variable de respuesta y las variables "Frequency", "VehPower", "VehAge", "DrivAge", "BonusMalus", "VehBrand", "VehGas", "Area", "Density" y "Region" como variables predictoras para los percentiles 50, 75, 90 y 95.

tau: 0.50	Value	Pr(> | t|)
(Intercept)	1,081.686	0
Frequency	1.860	0.002
VehPower	0.777	0.017
VehAge	0.434	0.003
DrivAge	0.771	0
BonusMalus	0.715	0
VehBrandB2	7.108	0.064
VehBrandB3	7.553	0.206
VehBrandB4	0.531	0.945
VehBrandB5	-2.784	0.566
VehBrandB6	-5.355	0.292
VehBrandB10	6.173	0.524
VehBrandB11	-5.444	0.659
VehBrandB12	43.009	0
VehBrandB13	16.444	0.040
VehBrandB14	-4.262	0.879
VehGasRegular	-9.948	0
AreaB	0.662	0.827
AreaC	0.045	0.986
AreaD	2.076	0.398
AreaE	-2.565	0.448
AreaF	-28.559	0.017
Density	0.0005	0.351
RegionAquitaine	-5.016	0.730
RegionAuvergne	-9.169	0.520
RegionBasse-Normandie	-32.493	0.034
RegionBourgogne	-8.153	0.578
RegionBretagne	-20.056	0.165
RegionCentre	-26.525	0.063
RegionChampagne-Ardenne	-2.008	0.893
RegionCorse	-2.766	0.870
RegionFranche-Comte	4.367	0.925
RegionHaute-Normandie	-8.842	0.564
RegionIle-de-France	-5.558	0.695
RegionLanguedoc-Roussillon	-7.228	0.619
RegionLimousin	-25.561	0.152
RegionMidi-Pyrenees	-15.677	0.330
RegionNord-Pas-de-Calais	-6.617	0.643
RegionPays-de-la-Loire	-29.390	0.040
RegionPicardie	-7.282	0.650
RegionPoitou-Charentes	-28.255	0.052
RegionProvence-Alpes-Cotes-D'Azur	-7.353	0.610
RegionRhone-Alpes	-18.322	0.202

tau: 0.75	Value	Pr(> | t|)
(Intercept)	1,178.600	0.0001
Frequency	46.901	0
VehPower	5.170	0.085
VehAge	-6.259	0.00000
DrivAge	0.451	0.330
BonusMalus	1.956	0.002
VehBrandB2	18.805	0.201
VehBrandB3	71.999	0.106
VehBrandB4	-9.424	0.800
VehBrandB5	-16.450	0.277
VehBrandB6	-29.579	0.156
VehBrandB10	49.812	0.500
VehBrandB11	154.321	0.177
VehBrandB12	-50.420	0.003
VehBrandB13	69.696	0.471
VehBrandB14	-20.227	0.638
VehGasRegular	-10.746	0.338
AreaB	7.475	0.631
AreaC	8.457	0.535
AreaD	-6.902	0.652
AreaE	16.523	0.547
AreaF	-169.280	0.039
Density	0.006	0.081
RegionAquitaine	30.143	0.921
RegionAuvergne	-119.021	0.689
RegionBasse-Normandie	-118.032	0.692
RegionBourgogne	-75.342	0.807
RegionBretagne	-109.376	0.712
RegionCentre	-116.265	0.695
RegionChampagne-Ardenne	-77.937	0.805
RegionCorse	548.748	0.449
RegionFranche-Comte	183.521	0.542
RegionHaute-Normandie	-124.414	0.675
RegionIle-de-France	-52.522	0.860
RegionLanguedoc-Roussillon	220.090	0.471
RegionLimousin	-105.569	0.735
RegionMidi-Pyrenees	-86.113	0.772
RegionNord-Pas-de-Calais	-50.011	0.867
RegionPays-de-la-Loire	-122.563	0.679
RegionPicardie	-106.416	0.725
RegionPoitou-Charentes	-127.247	0.668
RegionProvence-Alpes-Cotes-D'Azur	203.658	0.497
RegionRhone-Alpes	-41.399	0.889

tau: 0.90	Value	Pr(> | t|)
(Intercept)	1,890.984	0
Frequency	157.011	0
VehPower	28.098	0.122
VehAge	-53.999	0
DrivAge	-1.640	0.544
BonusMalus	6.386	0.006
VehBrandB2	158.552	0.082
VehBrandB3	195.837	0.184
VehBrandB4	-262.401	0.234
VehBrandB5	-107.010	0.460
VehBrandB6	-501.357	0.00001
VehBrandB10	581.370	0.012
VehBrandB11	410.679	0.135
VehBrandB12	-367.864	0.005
VehBrandB13	375.631	0.236
VehBrandB14	-158.989	0.597
VehGasRegular	-35.799	0.615
AreaB	-91.909	0.492
AreaC	37.198	0.739
AreaD	131.668	0.282
AreaE	129.810	0.440
AreaF	-629.357	0.222
Density	0.011	0.564
RegionAquitaine	863.236	0.0001
RegionAuvergne	-258.026	0.269
RegionBasse-Normandie	620.125	0.005
RegionBourgogne	528.429	0.104
RegionBretagne	465.484	0.004
RegionCentre	289.153	0.015
RegionChampagne-Ardenne	-34.665	0.937
RegionCorse	3,443.173	0.093
RegionFranche-Comte	572.565	0.635
RegionHaute-Normandie	646.366	0.295
RegionIle-de-France	340.134	0.025
RegionLanguedoc-Roussillon	792.234	0.008
RegionLimousin	-110.812	0.528
RegionMidi-Pyrenees	350.462	0.327
RegionNord-Pas-de-Calais	604.522	0.003
RegionPays-de-la-Loire	40.657	0.803
RegionPicardie	267.369	0.120
RegionPoitou-Charentes	252.833	0.251
RegionProvence-Alpes-Cotes-D'Azur	1,508.135	0
RegionRhone-Alpes	455.243	0.001

tau: 0.95	Value	Pr(> | t|)
(Intercept)	1,179.181	0.264
Frequency	290.803	0
VehPower	102.446	0.021
VehAge	-83.126	0.00000
DrivAge	-1.828	0.744
BonusMalus	13.170	0.016
VehBrandB2	118.752	0.492
VehBrandB3	43.785	0.881
VehBrandB4	-447.369	0.036
VehBrandB5	-170.127	0.592
VehBrandB6	-429.263	0.181
VehBrandB10	1,239.487	0.038
VehBrandB11	726.066	0.204
VehBrandB12	356.673	0.309
VehBrandB13	763.231	0.211
VehBrandB14	-531.527	0.723
VehGasRegular	-86.388	0.540
AreaB	-90.412	0.768
AreaC	-67.616	0.717
AreaD	151.150	0.536
AreaE	536.971	0.110
AreaF	-831.734	0.304
Density	0.011	0.745
RegionAquitaine	2,040.183	0.027
RegionAuvergne	-15.821	0.986
RegionBasse-Normandie	1,362.685	0.164
RegionBourgogne	2,040.747	0.030
RegionBretagne	2,018.756	0.029
RegionCentre	1,519.254	0.084
RegionChampagne-Ardenne	442.683	0.937
RegionCorse	6,271.379	0.00000
RegionFranche-Comte	5,704.351	0.659
RegionHaute-Normandie	1,678.540	0.058
RegionIle-de-France	1,487.030	0.127
RegionLanguedoc-Roussillon	3,138.253	0.012
RegionLimousin	518.823	0.557
RegionMidi-Pyrenees	1,365.778	0.355
RegionNord-Pas-de-Calais	2,328.455	0.020
RegionPays-de-la-Loire	1,294.101	0.151
RegionPicardie	2,083.105	0.034
RegionPoitou-Charentes	1,594.925	0.097
RegionProvence-Alpes-Cotes-D'Azur	4,456.781	0.00000
RegionRhone-Alpes	1,934.243	0.029

Ilustración 24. Summary de los modelos cuantílicos con tau 0.50, 0.75, 0.90 y 0.95. Fuente: Elaboración propia

Como sabemos, los valores p indican la significancia estadística de los coeficientes estimados. Un valor p bajo (generalmente menor que 0.05) indica que existe evidencia

estadística suficiente para rechazar la hipótesis nula de que el coeficiente es igual a cero. Esto indica que la variable independiente correspondiente tiene un efecto significativo en la variable dependiente en ese cuantil específico. Por otro lado, un valor p alto (mayor que 0.05) implica que la variable independiente no tiene un efecto significativo en la variable dependiente en ese cuantil específico. Por lo tanto, con respecto a la ilustración 24, podemos realizar la siguiente interpretación:

- **Intercept** es estadísticamente significativo para los cuantiles 0.5, 0.75 y 0.90 pero no lo es para el cuantil 0.95.

- **Frequency** es significativa y positivamente relacionada con el coste de siniestros en todos los cuantiles analizados. Para el cuantil 0.50, el coeficiente de la variable "Frequency" tiene un valor de 1.860. Esto indica que, un aumento de una unidad en la variable "Frequency" se relaciona, en promedio, con un incremento de 1.860 en el coste de siniestros.

- **VehPower** es estadísticamente significativo para los cuantiles 0.5 y 0.95 pero no lo es para los cuantiles 0.75 y 0.90. Para el cuantil 0.50, el coeficiente de la variable "VehPower" tiene un valor de 0.777. Esto indica que, un aumento de una unidad en la variable "VehPower" se relaciona, en promedio, con un incremento de 0.777 en el coste de siniestros.

- **VehAge** es significativa en todos los cuantiles, con relación positiva en el cuantil 0.5 y negativa en los demás. Para el cuantil 0.50, el coeficiente de la variable "VehAge" tiene un valor de 0.434. Esto indica que, un aumento de una unidad en la variable " VehAge" se relaciona, en promedio, con un incremento de 0.434 en el coste de siniestros.

- **DrivAge** es estadísticamente significativo solo para el cuantil 0.5. En este punto específico del rango de cuantiles, a medida que la edad del conductor aumenta, se espera que el coste de siniestros también aumente.

- **BonusMalus** es significativa y positivamente relacionada con el coste de siniestros en todos los cuantiles analizados. Para el cuantil 0.50, el coeficiente de la variable "BonusMalus" tiene un valor de 0.715. Esto indica que, un aumento de una unidad en la variable "BonusMalus" se relaciona, en promedio, con un incremento de 0.715 en el coste de siniestros.

- **VehBrand** es significativa en todos los cuantiles**.** Específicamente, se observa que en el cuantil 0.5, las marcas "B12" y "B13" muestran una relación significativa con el costo de siniestros. En el cuantil 0.75, la marca "B12" presenta un impacto significativo en el costo de siniestros. En el cuantil 0.90, se observa que las marcas "B6", "B10" y "B12" tienen coeficientes estimados significativos en relación al coste de siniestros. Por último, en el cuantil 0.95, "B4" y "B10" son las marcas con coeficientes significativos.

- **VehGas** es estadísticamente significativo solo para el cuantil 0.5. En este punto específico del rango de cuantiles, cuando se utiliza combustible regular, se espera una disminución en el coste de siniestros.

- **Area** es estadísticamente significativo para los cuantiles 0.5 y 0.75 pero no lo es para los cuantiles 0.90 y 0.95. En los cuantiles 0.5 y 0.75, se observa que cuando el conductor vive en el área F, el coste de siniestros tiende a ser menor.

- **Density** no muestra significancia estadística en ninguno de los cuantiles analizados.

- **Region** es estadísticamente significativo para los cuantiles 0.5, 0.90 y 0.95 pero no lo es para el cuantil 0.75. Específicamente, se observa que en el percentil 50 de los casos, los titulares de pólizas en las regiones "Basse-Normandie" y "Pays-de-la-Loire" tienden a tener reclamaciones de menor cuantía en comparación con la región de referencia "Alsace". En el cuantil 0.90, los titulares de pólizas en las regiones "Aquitaine", "Basse-Normandie", "Bretagne", "Centre", "Ile-de-France", "Languedoc-Roussillon", "Nord-Pas-de-Calais", "Provence-Alpes-Cotes-D'Azur" y "Rhone-Alpes" tienden a tener reclamaciones de mayor cuantía en comparación con la región de referencia "Alsace". En el cuantil 0.95, los titulares de pólizas en las regiones "Aquitaine", "Bourgogne", "Bretagne", "Corse", "Languedoc-Roussillon", "Nord-Pas-de-Calais", "Picardie", "Provence-Alpes-Cotes-D'Azur" y "Rhone-Alpes" tienden a tener reclamaciones de mayor cuantía en comparación con la región de referencia "Alsace".

Para comprobar si los coeficientes de regresión cuantílica para diferentes cuantiles son realmente diferentes, se puede utilizar una prueba de hipótesis basada en la función ANOVA (Análisis de Varianza). Esta técnica permite comparar los modelos de regresión cuantílica y determinar si existen diferencias significativas en los coeficientes estimados entre los cuantiles. En este caso, se realizará una comparación entre el cuantil 0.50 y el cuantil 0.95.

```
Quantile Regression Analysis of Deviance Table

Model: ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge + BonusMalus + VehBrand + VehGas + Area + Density + Region
Joint Test of Equality of Slopes: tau in {  0.5 0.95  }

  Df Resid Df F value    Pr(>F)
1 42    49844  21.669 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ilustración 25. Test Anova para los modelos con tau 0.50 y 0.95. Fuente: Elaboración propia

Basándonos en el resultado de la prueba de ANOVA, si podemos rechazar la hipótesis nula, esto indica que hay una diferencia significativa en los coeficientes entre los cuantiles que se están comparando. Esto significa que los efectos de las variables independientes en la variable dependiente varían significativamente entre estos dos cuantiles.

A continuación, se realiza una estimación de la regresión lineal múltiple y una comparación de los coeficientes de regresión de cada cuantil en el rango de 0.05 a 0.95 con el coeficiente de regresión lineal múltiple obtenido mediante mínimos cuadrados ordinarios, junto con sus intervalos de confianza del 95%.

```
Call:
lm(formula = ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge +
    BonusMalus + VehBrand + VehGas + Area + Density + Region,
    data = freMTPL2)

Residuals:
    Min      1Q  Median      3Q     Max
 -17477   -1984   -1076    -176 4068710

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -7.377e+02  3.659e+03  -0.202   0.8402
Frequency                       3.974e+01  1.910e+01   2.080   0.0375 *
VehPower                        1.154e+02  1.022e+02   1.130   0.2585
VehAge                          1.684e+01  4.073e+01   0.413   0.6793
DrivAge                        -2.032e+01  1.504e+01  -1.351   0.1767
BonusMalus                      2.392e+01  1.115e+01   2.146   0.0319 *
VehBrandB2                      7.515e+02  5.324e+02   1.411   0.1582
VehBrandB3                     -4.372e+00  7.407e+02  -0.006   0.9953
VehBrandB4                      2.695e+02  1.010e+03   0.267   0.7897
VehBrandB5                     -3.525e+02  8.509e+02  -0.414   0.6787
VehBrandB6                     -5.551e+02  9.582e+02  -0.579   0.5624
VehBrandB10                     1.804e+02  1.220e+03   0.148   0.8824
VehBrandB11                     1.075e+03  1.294e+03   0.830   0.4063
VehBrandB12                     3.968e+02  6.885e+02   0.576   0.5643
VehBrandB13                     1.380e+02  1.402e+03   0.098   0.9216
VehBrandB14                    -3.371e+02  2.716e+03  -0.124   0.9012
VehGasRegular                   5.175e+02  3.964e+02   1.306   0.1917
AreaB                           1.246e+03  8.144e+02   1.530   0.1261
AreaC                          -8.716e+01  6.726e+02  -0.130   0.8969
AreaD                           2.951e+01  7.139e+02   0.041   0.9670
AreaE                          -2.109e+02  9.359e+02  -0.225   0.8217
AreaF                          -1.214e+03  3.271e+03  -0.371   0.7105
Density                         4.019e-02  1.347e-01   0.298   0.7655
RegionAquitaine                 5.036e+02  3.463e+03   0.145   0.8844
RegionAuvergne                 -2.315e+02  4.232e+03  -0.055   0.9564
RegionBasse-Normandie           6.961e+02  3.637e+03   0.191   0.8482
RegionBourgogne                 3.297e+02  3.730e+03   0.088   0.9296
RegionBretagne                  7.164e+02  3.407e+03   0.210   0.8334
RegionCentre                    1.464e+03  3.358e+03   0.436   0.6627
RegionChampagne-Ardenne         5.470e+03  4.925e+03   1.111   0.2667
RegionCorse                     2.609e+03  4.426e+03   0.590   0.5555
RegionFranche-Comte             4.236e+02  6.047e+03   0.070   0.9442
RegionHaute-Normandie           8.887e+01  3.931e+03   0.023   0.9820
RegionIle-de-France             4.104e+02  3.409e+03   0.120   0.9042
RegionLanguedoc-Roussillon      7.043e+02  3.475e+03   0.203   0.8394
RegionLimousin                  1.778e+02  4.019e+03   0.044   0.9647
RegionMidi-Pyrenees             7.134e+01  3.702e+03   0.019   0.9846
RegionNord-Pas-de-Calais        5.816e+02  3.430e+03   0.170   0.8654
RegionPays-de-la-Loire          1.491e+02  3.418e+03   0.044   0.9652
RegionPicardie                  4.580e+02  3.755e+03   0.122   0.9029
RegionPoitou-Charentes          3.918e+01  3.509e+03   0.011   0.9911
RegionProvence-Alpes-Cotes-D'Azur 1.145e+03  3.374e+03   0.340   0.7342
RegionRhone-Alpes               1.049e+03  3.361e+03   0.312   0.7549
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ilustración 26. Summary del modelo MCO. Fuente: Elaboración propia

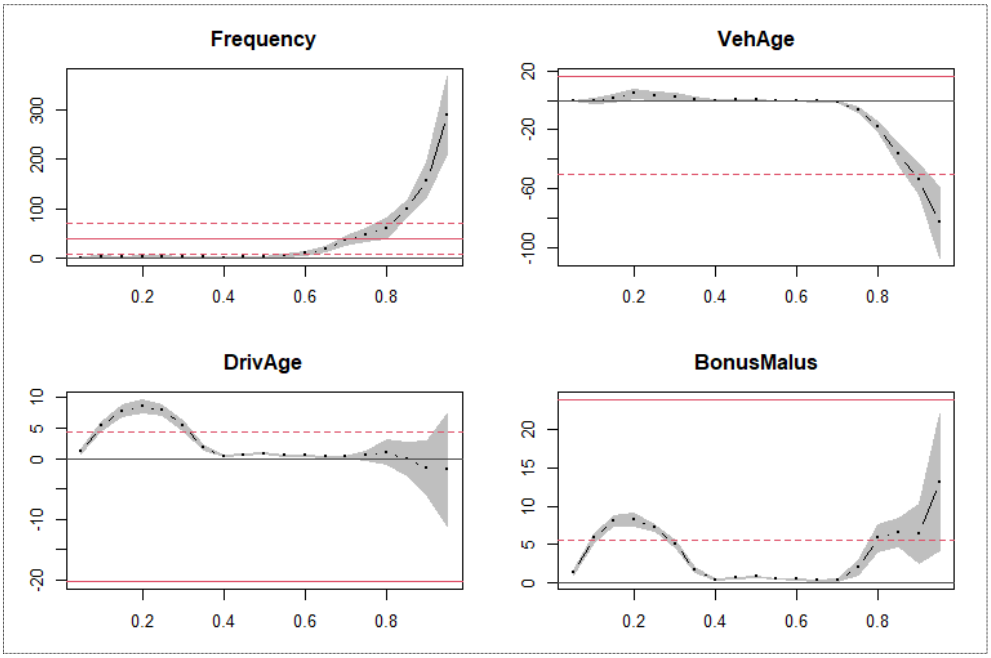

Ilustración 27. Comparación de los coeficientes de los cuantiles con los de MCO. Fuente: Elaboración propia

La línea roja continua representa el coeficiente obtenido mediante MCO, mientras que las líneas rojas discontinuas representan los intervalos de confianza del 95% asociados. Los puntos negros corresponden a los coeficientes estimados para cada cuantil mediante la regresión cuantílica. Observamos que las cuatro variables que muestran diferencias estadísticamente significativas entre los coeficientes estimados de la regresión cuantílica y los coeficientes estimados mediante MCO son "Frequency", "VehAge", "DrivAge" y "BonusMalus".

En las variables "Frequency" y "VehAge", se observan diferencias en la relación con la variable respuesta "ClaimAmount" entre la regresión cuantílica y la regresión de mínimos cuadrados ordinarios en los cuantiles más altos. Esto indica claramente que la asociación entre estas variables y la variable respuesta varían en diferentes cuantiles de la distribución.

Para la variable "DrivAge", el coeficiente de la regresión cuantílica es más alto que el coeficiente de la regresión de mínimos cuadrados ordinarios en los cuantiles entre 0.1 y 0.3, esto significa que la relación entre la variable "DrivAge" y la variable respuesta puede ser más fuerte en esos cuantiles inferiores. Dicho de otro modo, a medida que la edad del conductor aumenta en esos cuantiles, se espera que el coste de siniestros también aumente.

Por último, en la variable "BonusMalus" se puede observar que el impacto de la bonificación o penalización del asegurado puede ser menor en los cuantiles intermedios en comparación con los demás cuantiles analizados.

En realidad, la discrepancia significativa entre el coeficiente de la regresión cuantílica con tau=0.50 y el coeficiente de la regresión lineal múltiple ya indica la presencia de una alta asimetría en los datos. Como se ha mencionado en la parte teórica, cuando los datos son muy sesgados, la regresión lineal ya no es representativa y no es capaz de capturar

adecuadamente la relación entre las variables. Por lo tanto, es necesario utilizar otros métodos, como la regresión cuantílica, para obtener estimaciones más precisas.

3.3. RESULTADOS DE LA REGRESIÓN CONJUNTA VAR & ES

Para facilitar la estimación computacional de la regresión conjunta de cuantiles con *expected shortfall*, se eliminan las variables categóricas de la fórmula y se utilizan solo las variables numéricas. Esto se debe a que las variables categóricas requieren una codificación adicional, lo que puede complicar el cálculo de los coeficientes. Además, para evitar limitaciones asociadas con las operaciones computacionales, se realiza una normalización dividiendo la variable respuesta por 1000. Asimismo, se ha transformado la variable respuesta el coste de siniestros en valores negativos, ya que este modelo considera que $Y < 0$ representa una pérdida.

Como se sabe:

$$\mathrm{VaR}_{\alpha}(Y) = -\mathrm{VaR}_{1-\alpha}(-Y)$$

Si se desea estimar el percentil 95 del costo de siniestros, es necesario utilizar un valor de *alpha* de 0.05.

A continuación, se estima el modelo de regresión seleccionando el mismo conjunto de variables explicativas para la ecuación de regresión cuantílica y la ecuación del *expected shortfall*, por lo tanto, $X = (X_q, X_e) = (Z, Z)$.

Variable explicativa	**Descripción**
Frequency	Número de siniestros anualizados
VehPower	Potencia del vehículo
VehAge	Edad del vehículo
DrivAge	Edad del conductor principal del vehículo asegurado
BonusMalus	Bonus/Malus, entre 50 y 350: <100 significa bonus, >=100 significa malus en Francia
Density	Densidad de habitantes por kilómetro cuadrado en la ciudad de residencia del conductor

Ilustración 28. Explicación de las variables explicativas. Fuente: Elaboración propia

Las funciones de especificación G_1 utilizadas en la *función scoring* son las siguientes[6]:

$\boldsymbol{G_1}\ (\mathbf{1}):\ G_1(z) = z$

$\boldsymbol{G_1}\ (\mathbf{2}):\ G_1(z) = 0$

Las funciones de especificación $\mathcal{G}_2$ utilizadas en la *función scoring* son las siguientes[6]:

$\boldsymbol{\mathcal{G}_2}\ (\mathbf{1}):\ \mathcal{G}_2(z) = -1/z \quad con\ (z < 0)$

$\boldsymbol{\mathcal{G}_2}\ (\mathbf{2}):\ \mathcal{G}_2(z) = 0.5/\sqrt{z} \quad con\ (z < 0)$

$\boldsymbol{\mathcal{G}_2}\ (\mathbf{3}):\ \mathcal{G}_2(z) = 1/z\text{^}2 \quad con\ (z < 0)$

$\boldsymbol{\mathcal{G}_2}\ (\mathbf{4}):\ \mathcal{G}_2(z) = 1/(1 + \exp(-z))$

$\boldsymbol{\mathcal{G}_2}\ (\mathbf{5}):\ \mathcal{G}_2(z) = \exp(z)$

Se ha estimado la regresión conjunta utilizando cada combinación de las funciones de especificación mencionadas anteriormente para el cuantil 0.05. Como resultado, se han obtenido cinco modelos con parámetros significativos utilizando las siguientes funciones:

1. $\boldsymbol{G_1(z) = z\ y\ \mathcal{G}_2(z) = -1/z}$

Quantile Coefficients

	Estimate	**Std. Error**	**t value**	**Pr(> | t |)**	
bq_0	-2.2314e+00	8.2267e-01	-2.7124	0.006684	**
bq_1	-4.5646e-01	3.7793e-02	-12.0779	< 2.2e-16	***
bq_2	-1.4493e-01	5.3415e-02	-2.7132	0.006667	**
bq_3	9.6488e-02	2.4693e-02	3.9075	9.35e-05	***
bq_4	-3.6047e-03	7.9456e-03	-0.4537	0.650069	
bq_5	-1.6581e-02	8.5248e-03	-1.9450	0.051786	.
bq_6	9.5987e-06	1.9230e-05	0.4991	0.617679	

Ilustración 29. Resultados de la estimación de la regresión conjunta (g1=1 y g2=1). Fuente: Elaboración propia

[6] Véase API for esreg: https://rdrr.io/cran/esreg/api/

Expected Shortfall Coefficients

	Estimate	Std. Error	t value	Pr(> | t |)	
be_0	19.93917162	10.04150138	1.9857	0.047080	*
be_1	-6.77677565	0.89662815	-7.5581	4.232e-14	***
be_2	-1.06838617	0.63465363	-1.6834	0.092307	.
be_3	-0.19233100	0.25438997	-0.7560	0.449628	
be_4	-0.06811146	0.09631580	-0.7072	0.479469	
be_5	-0.25894611	0.09474418	-2.7331	0.006278	**
be_6	0.00011863	0.00036663	0.3236	0.746256	

Ilustración 30. Resultados de la estimación de la regresión conjunta (g1=1 y g2=1). Fuente: Elaboración propia

En la salida del modelo, se presentan los coeficientes estimados para el cuantil y para el *expected shortfall*, junto con sus errores estándar, valores t y valores p asociados. Los asteriscos indican la significancia estadística de cada coeficiente[7]. Como podemos observar, a un nivel de significancia de 0.05, las variables significativas son "Frequency", "VehPower" y "VehAge" para la estimación de cuantil, y "Frequency" y "BonusMalus" para la estimación de *expected shortfall*. Por lo tanto, a la tabla 29 y la tabla 30, podemos realizar la siguiente interpretación:

- **Frequency**: el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 456.46 unidades por cada incremento de una unidad en la variable "Frequency" y el $ES_{0.95}$ del coste de siniestros aumenta aproximadamente en 6776.77 unidades por cada incremento de una unidad en la variable "Frequency".

- **VehPower:** el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 144.93 unidades por cada incremento de una unidad en la variable "VehPower".

- **VehAge:** el cuantil 0.95 del coste de siniestros disminuye aproximadamente en 96.49 unidades por cada incremento de una unidad en la variable "VehAge".

- **BonusMalus:** el $ES_{0.95}$ del coste de siniestros aumenta aproximadamente en 258.95 unidades por cada incremento de una unidad en la variable "BonuMalus".

[7] Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

2. $\boldsymbol{G_1(z) = z\ y\ G_2(z) = 0.5/\sqrt{z}}$

Quantile Coefficients

	Estimate	**Std. Error**	**t value**	**Pr(> | t |)**	
bq_0	-1.9773e+00	7.9300e-01	-2.4934	0.012658	*
bq_1	-5.2762e-01	4.4014e-02	-11.9875	< 2.2e-16	***
bq_2	-1.6039e-01	5.2583e-02	-3.0502	0.002289	**
bq_3	9.6304e-02	2.1020e-02	4.5816	4.638e-06	***
bq_4	-5.7796e-03	7.7407e-03	-0.7466	0.455287	
bq_5	-1.7101e-02	7.8605e-03	-2.1756	0.029596	*
bq_6	2.1535e-05	2.0121e-05	1.0703	0.284502	

Ilustración 31. Resultados de la estimación de la regresión conjunta (g1=1 y g2=2). Fuente: Elaboración propia

Expected Shortfall Coefficients

	Estimate	**Std. Error**	**t value**	**Pr(> | t |)**	
be_0	8.26666436	10.07425976	0.8206	0.41190	
be_1	-4.06126368	0.67346896	-6.0304	1.659e-09	***
be_2	-0.92876989	0.67911664	-1.3676	0.17145	
be_3	-0.36122219	0.24358535	-1.4829	0.13810	
be_4	0.04404488	0.11874234	0.3709	0.71069	
be_5	-0.23423238	0.09418721	-2.4869	0.01289	*
be_6	0.00028520	0.00039076	0.7298	0.46549	

Ilustración 32. Resultados de la estimación de la regresión conjunta (g1=1 y g2=2). Fuente: Elaboración propia

En este caso, podemos observar que, a un nivel de significancia de 0.05, las variables significativas son "Frequency", "VehPower", "VehAge" y "BonusMalus" para los coeficientes de cuantil, y "Frequency" y "BonusMalus" para los coeficientes de *expected shortfall*. Por lo tanto, a la tabla 31 y la tabla 32, podemos realizar la siguiente interpretación:

- **Frequency**: el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 527.62 unidades por cada incremento de una unidad en la variable "Frequency" y el $ES_{0.95}$ del coste de siniestros aumenta aproximadamente en 4061.26 unidades por cada incremento de una unidad en la variable "Frequency".

- **VehPower:** el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 160.39 unidades por cada incremento de una unidad en la variable "VehPower".

- **VehAge:** el cuantil 0.95 del coste de siniestros disminuye aproximadamente en 96.30 unidades por cada incremento de una unidad en la variable "VehAge".
- **BonusMalus:** el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 17.10 unidades por cada incremento de una unidad en la variable "BonuMalus" y el $ES_{0.95}$ del coste de siniestros aumenta aproximadamente en 234.2324 unidades por cada incremento de una unidad en la variable "BonuMalus".

3. $\boldsymbol{G_1(z) = z\ y\ G_2(z) = 1/z\text{^}2}$

Quantile Coefficients

	Estimate	Std. Error	t value	Pr(> | t |)	
bq_0	-2.3151e+00	9.0240e-01	-2.5655	0.010307	*
bq_1	-3.3667e-01	4.1153e-02	-8.1808	2.953e-16	***
bq_2	-1.6025e-01	5.5837e-02	-2.8700	0.004109	**
bq_3	1.0182e-01	3.6881e-02	2.7607	0.005771	**
bq_4	-5.4957e-03	8.8981e-03	-0.6176	0.536830	
bq_5	-1.8197e-02	8.7818e-03	-2.0721	0.038269	*
bq_6	3.0170e-06	1.8597e-05	0.1622	0.871126	

Ilustración 33. Resultados de la estimación de la regresión conjunta (g1=1 y g2=3). Fuente: Elaboración propia

Expected Shortfall Coefficients

	Estimate	Std. Error	t value	Pr(> | t |)	
be_0	2.4246e+01	2.2042e+01	1.1000	0.2713598	
be_1	-1.0235e+01	2.7062e+00	-3.7821	0.0001559	***
be_2	-7.0759e-01	1.1080e+00	-0.6386	0.5230817	
be_3	6.9960e-02	5.0902e-01	1.1374	0.8906844	
be_4	-1.3617e-01	2.1058e-01	-0.6466	0.5178693	
be_5	-2.5386e-01	2.2060e-01	-1.1508	0.2498407	
be_6	-5.2829e-04	6.0121e-04	-0.8787	0.3795683	

Ilustración 34. Resultados de la estimación de la regresión conjunta (g1=1 y g2=3). Fuente: Elaboración propia

Se observa que, a un nivel de significancia de 0.05, las variables significativas son “Frequency”, “VehPower”, “VehAge” y “BonusMalus” para los coeficientes de cuantil, y “Frequency” para los coeficientes de *expected shortfall*. Por lo tanto, con respecto a la tabla 33 y la tabla 34, podemos realizar la siguiente interpretación:

- **Frequency**: el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 336.67 unidades por cada incremento de una unidad en la variable "Frequency" y el $ES_{0.95}$ del coste de siniestros aumenta aproximadamente en 10235.10 unidades por cada incremento de una unidad en la variable "Frequency".
- **VehPower:** el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 160.25 unidades por cada incremento de una unidad en la variable "VehPower".
- **VehAge:** el cuantil 0.95 del coste de siniestros disminuye aproximadamente en 101.82 unidades por cada incremento de una unidad en la variable "VehAge".
- **BonusMalus:** el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 18.20 unidades por cada incremento de una unidad en la variable "BonuMalus".

4. $\boldsymbol{G_1(z) = 0 \ y \ G_2(z) = -1/z}$

Quantile Coefficients

	Estimate	Std. Error	t value	Pr(> | t |)	
bq_0	-1.1572e+00	1.3746e+00	-0.8419	0.39988	
bq_1	-1.2344e+00	9.4700e-02	-13.0351	< 2e-16	***
bq_2	-1.5420e-01	6.6109e-02	-2.3325	0.01968	*
bq_3	8.4417e-02	4.4196e-02	1.9101	0.05614	.
bq_4	-6.2248e-03	1.0893e-02	-0.5715	0.56769	
bq_5	-9.8995e-03	1.9846e-02	-0.4988	0.61791	
bq_6	3.1651e-05	2.2605e-05	1.4002	0.16146	

Ilustración 35. Resultados de la estimación de la regresión conjunta (g1=2 y g2=1). Fuente: Elaboración propia

Expected Shortfall Coefficients

	Estimate	Std. Error	t value	Pr(> | t |)	
be_0	19.18428934	11.00336640	1.7435	0.08126	.
be_1	-6.80855149	1.06814277	-6.3742	1.871e-10	***
be_2	-1.05682094	0.66546046	-1.5881	0.11228	
be_3	-0.19695095	0.28766555	-0.6847	0.49357	
be_4	-0.06035063	0.10132901	-0.5956	0.55145	
be_5	-0.25179658	0.11503833	-2.1888	0.02862	*
be_6	0.00014997	0.00037421	0.4008	0.68860	

Ilustración 36. Resultados de la estimación de la regresión conjunta (g1=2 y g2=1). Fuente: Elaboración propia

En este caso, se observa que, a un nivel de significancia de 0.05, las variables significativas son “Frequency” y “VehPower” para los coeficientes de cuantil, y “Frequency” y “BonusMalus” para los coeficientes de *expected shortfall*. Por lo tanto, con respecto a la tabla 35 y la tabla 36, podemos realizar la siguiente interpretación:

- **Frequency**: el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 1234.42 unidades por cada incremento de una unidad en la variable "Frequency" y el $ES_{0.95}$ del coste de siniestros aumenta aproximadamente en 6808.55 unidades por cada incremento de una unidad en la variable "Frequency".
- **VehPower:** el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 154.20 unidades por cada incremento de una unidad en la variable "VehPower”.
- **BonusMalus:** el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 251.80 unidades por cada incremento de una unidad en la variable "BonuMalus".

5. $\boldsymbol{G_1(z) = 0 \ y \ G_2(z) = 0.5/\sqrt{z}}$

Quantile Coefficients

	Estimate	**Std. Error**	**t value**	**Pr(> | t |)**	
bq_0	-1.1754e+00	8.0122e-01	-1.4671	0.142374	
bq_1	-6.8344e-01	5.5261e-02	-12.3676	< 2.2e-16	***
bq_2	-1.3827e-01	5.2978e-02	-2.6100	0.009060	**
bq_3	9.3835e-02	2.2527e-02	4.1655	3.118e-05	***
bq_4	-1.2505e-02	7.8073e-03	-1.6017	0.109242	
bq_5	-2.2045e-02	8.1902e-03	-2.6917	0.007114	**
bq_6	2.8195e-05	1.9913e-05	1.4159	0.156823	

Ilustración 37. Resultados de la estimación de la regresión conjunta (g1=2 y g2=2). Fuente: Elaboración propia

Expected Shortfall Coefficients

	Estimate	Std. Error	t value	Pr(> | t |)	
be_0	8.79925488	10.22506001	0.8606	0.38949	
be_1	-4.08349006	0.70305981	-5.8082	6.393e-09	***
be_2	-0.90195810	0.68189303	-1.3227	0.18594	
be_3	-0.33563340	0.25263385	-1.3285	0.18401	
be_4	0.03677521	0.11732806	0.3134	0.75395	
be_5	-0.24123967	0.09592678	-2.5148	0.01192	*
be_6	0.00029156	0.00038589	0.7555	0.44993	

Ilustración 38. Resultados de la estimación de la regresión conjunta (g1=2 y g2=2). Fuente: Elaboración propia

Como podemos ver, a un nivel de significancia de 0.05, las variables significativas son "Frequency", "VehPower", "VehAge" y "BonusMalus" para los coeficientes de cuantil, y "Frequency" y "BonusMalus" para los coeficientes de *expected shortfall*. Por lo tanto, con respecto a la tabla 37 y la tabla 38, podemos realizar la siguiente interpretación:

- **Frequency**: el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 683.44 unidades por cada incremento de una unidad en la variable "Frequency" y el $ES_{0.95}$ del coste de siniestros aumenta aproximadamente en 4083.49 unidades por cada incremento de una unidad en la variable "Frequency".

- **VehPower:** el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 138.27 unidades por cada incremento de una unidad en la variable "VehPower".

- **VehAge:** el cuantil 0.95 del coste de siniestros disminuye aproximadamente en 93.83 unidades por cada incremento de una unidad en la variable "VehAge".

- **BonusMalus:** el cuantil 0.95 del coste de siniestros aumenta aproximadamente en 22.05 unidades por cada incremento de una unidad en la variable "BonuMalus" y el $ES_{0.95}$ del coste de siniestros aumenta aproximadamente en 241.24 unidades por cada incremento de una unidad en la variable "BonuMalus".

Comparación de los modelos

	MSE ($VaR_{0.95}$)	MSE ($ES_{0.95}$)
Modelo 1	940.700	6,307.000
Modelo 2	948.200	3,161.000
Modelo 3	932.000	12,642.000
Modelo 4	1,075.000	6,346.000
Modelo 5	967.100	3,177.000

Ilustración 39. MSE de los coeficientes de VaR y ES de cada modelo. Fuente: Elaboración propia

En la tabla 39, podemos observar que los parámetros de ES exhiben un MSE (error cuadrático medio) más grande en comparación con los de cuantil. Este hecho es razonable, ya que el ES se sitúa en la cola extrema derecha de la distribución de pérdidas y, por lo tanto, su estimación está sujeta a un mayor error.

	MSE	SSE
Modelo 1	3623.612	180767494
Modelo 2	2054.782	102504848
Modelo 3	6786.977	338575137
Modelo 4	3710.770	185115490
Modelo 5	2071.963	103361943

Ilustración 40. Comparaciones de MSE y SSE de los modelos. Fuente: Elaboración propia

En la ilustración 40 se muestran los MSE y SSE (suma residual de cuadrados) de cada modelo, y se observa que los modelos 2 y 5 tienen menores residuos en comparación con los otros modelos.

$$\textbf{Model 2}: \ G_1(z) = z \ y \ G_2(z) = 0.5/\sqrt{z}$$

$$\textbf{Model 5}: \ G_1(z) = 0 \ y \ G_2(z) = 0.5/\sqrt{z}$$

Quantile Coefficients

	Modelo 2		**Modelo 5**	
	Estimate	*Signf.*	*Estimate*	*Signf.*
bq_0	-1.9773e+00	*	-1.1754e+00	
bq_1	-5.2762e-01	***	-6.8344e-01	***
bq_2	-1.6039e-01	**	-1.3827e-01	**
bq_3	9.6304e-02	***	9.3835e-02	***
bq_4	-5.7796e-03		-1.2505e-02	
bq_5	-1.7101e-02	*	-2.2045e-02	**
bq_6	2.1535e-05		2.8195e-05	

Ilustración 41. Comparación de los coeficientes entre el modelo 2 y el modelo 5. Fuente: Elaboración propia

Expected Shortfall Coefficients

	Modelo 2		**Modelo 5**	
	Estimate	*Signf.*	*Estimate*	*Signf.*
be_0	8.26666436		8.79925488	
be_1	-4.06126368	***	-4.08349006	***
be_2	-0.92876989		-0.90195810	
be_3	-0.36122219		-0.33563340	
be_4	0.04404488		0.03677521	
be_5	-0.23423238	*	-0.24123967	*
be_6	0.00028520		0.00029156	

Ilustración 42. Comparación de los coeficientes entre el modelo 2 y el modelo 5. Fuente: Elaboración propia

Al comparar los coeficientes estimados de los dos modelos, observamos que son muy similares. Además, a un nivel de significancia de 0.05, ambos modelos presentan las mismas variables explicativas significativas, las variables significativas son “Frequency”, “VehPower”, “VehAge” y “BonusMalus” para los coeficientes de cuantil, y “Frequency” y “BonusMalus” para los coeficientes de *expected shortfall*.

Ahora se utiliza el paquete ggplot2 para crear un gráfico de dispersión de los valores predichos versus los valores reales para los dos modelos con mejor rendimiento.

Modelos 2:

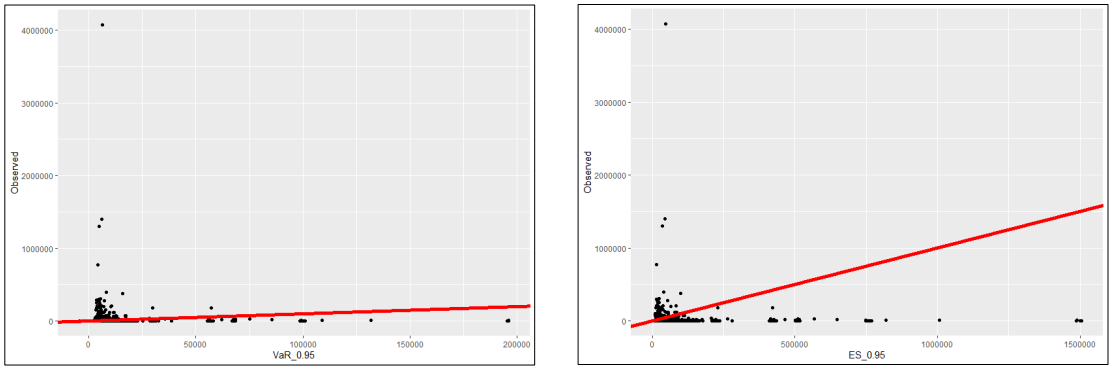

Ilustración 43. Observed vs Predicted del modelo 2. Fuente: Elaboración propia

Modelos 5:

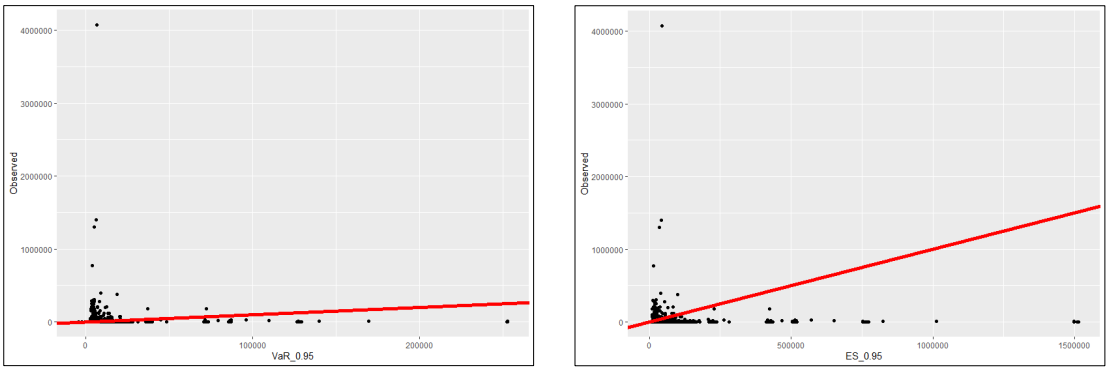

Ilustración 44. Observed vs Predicted del modelo 5. Fuente: Elaboración propia

Observamos que no existe mucha diferencia en las predicciones de los dos modelos, ya que la función de especificación $\mathcal{G}_2$ de ambos modelos es la misma. En cuanto al eje de las predicciones de *expected shortfall* vemos que es mucho más largo que el de los cuantiles, es decir, el rango de valores predichos de *expected shortfall* es más amplio. Esto indica que el modelo de *expected shortfall* ha logrado captar los costes potenciales que se encuentran más allá del nivel de riesgo especificado por el VaR (95% en nuestro caso) y se cumple la criterio de que *expected shortfall* es mayor que el VaR, tal como se ha definido en la parte teórica.

4. CONCLUSIÓN

A lo largo de este estudio sobre la regresión cuantílica y la regresión conjunta de VaR y *expected shortfall*, se ha explorado la teoría de estos modelos, se ha realizado un análisis descriptivo de los diversos factores de riesgo en el contexto de los seguros de automóviles, como la edad del conductor y la potencia del vehículo. Además, se han aplicado los modelos de regresión cuantílica y regresión conjunta de VaR y *expected shortfall* a esta base de datos.

El objetivo inicial era estimar el coste de siniestros utilizando ambos enfoques, y se encontró que la regresión cuantílica fue capaz de proporcionar estimaciones adecuadas. Esto se debe a que la base de datos analizada presentaba una alta asimetría en los datos. Se observaron numerosas variables con coeficientes significativos en la regresión cuantílica y se encontraron coeficientes distintos entre los diferentes cuantiles de la distribución, lo cual indica la importancia de considerar la regresión cuantílica para capturar las variaciones en la variable respuesta.

Sin embargo, al analizar la regresión conjunta regresión cuantílica con *expected shortfall*, se encontró que la mayoría de los coeficientes estimados no fueron significativos. Esto indica que la mayoría de las variables utilizadas en el modelo conjunto no tuvieron un impacto estadísticamente significativo en la predicción de *expected shortfall*. Aunque este resultado podría, en un principio, parecer desalentador, es importante destacar que el modelo de *expected shortfall* logró mejorar significativamente las predicciones en comparación con el modelo de cuantil al predecir las pérdidas extremas.

La mejora en las predicciones del modelo de *expected shortfall* se debe a su capacidad para capturar eventos extremos más allá de lo que puede captar el modelo de cuantil. Al tener en cuenta estos eventos extremos, el modelo de *expected shortfall* proporciona una medida más conservadora y abarca un mayor nivel de riesgo. Esto puede ser especialmente relevante en situaciones donde la protección contra eventos inesperados es fundamental para las compañías de seguros.

Además, al combinar ambos enfoques, la regresión conjunta de VaR y *expected shortfall* nos permite identificar las pólizas que representan un mayor riesgo, lo cual facilita el ajuste adecuado de la prima pura correspondiente. De esta manera, las compañías de seguros obtienen una evaluación más completa sobre el riesgo de su cartera de pólizas, lo que les permite tomar decisiones más informadas y precisas en términos de suscripción, gestión de riesgos, etc.

5. ANEXOS

```
################################
## Instalacion de librerias ##
################################
install.packages("CASdatasets")
install.packages("plotly")
install.packages("lattice")
install.packages("nortest")
install.packages("grid")
install.packages("dplyr")
install.packages("usethis")
install.packages("quantreg")
install.packages("esreg")
install.packages("ggplot2")
install.packages("xtable")
install.packages("skimr")
install.packages("devtools")
install.packages("stargazer")
install.packages("psych")
install.packages("nortest")
install.packages("RColorBrewer")
install.packages("magrittr")
install.packages("glue")
install.packages("tidyverse")
install.packages("ggcorrplot")
install.packages("zoo")
install.packages("xts")
```

```
install.packages("sp")
install.packages("gridExtra")

######################
## Abrir librerias ##
######################
library(CASdatasets)
library(plotly)
library(lattice)
library(nortest)
library(grid)
library(dplyr)
library(usethis)
library(quantreg)
library(esreg)
library(ggplot2)
library(xtable)
library(skimr)
library(devtools)
library(stargazer)
library(psych)
library(nortest)
library(RColorBrewer)
library(magrittr)
library(glue)
library(tidyverse)
library(ggcorrplot)
```

```
library(zoo)
library(xts)
library(sp)
library(gridExtra)

##############################
## Importar bases de datos ##
##############################
freMTPL2freq <- read.csv("freMTPL2freq.csv", header=TRUE, stringsAsFactors=TRUE)
freMTPL2sev <- read.csv("freMTPL2sev.csv", header=TRUE, stringsAsFactors=TRUE)
freMTPL2 <- merge(x = freMTPL2freq, y = freMTPL2sev)

#SOLO LAS PÓLIZAS CON SINIESTROS
summary(freMTPL2$ClaimNb)
summary(freMTPL2$ClaimAmount)

i <- 1
while (i <= nrow(freMTPL2)) {
 if(freMTPL2$ClaimNb[i]>1){
  freMTPL2$ClaimAmount[i] <- sum(freMTPL2$ClaimAmount[i:(i+freMTPL2$ClaimNb[i]-1)])
  freMTPL2 <- freMTPL2[-c((i+1):(i+freMTPL2$ClaimNb[i]-1)), ]
 }
 i <- i + 1
}

rownames(freMTPL2) <- NULL
```

```
pdf("data.pdf", height=18, width=14.5)
grid.table(head(freMTPL2), gp=gpar(fontsize=8))
dev.off()

a <- skim(freMTPL2)
stargazer(a[1:4,c(1:4,6)],
        summary = FALSE,
        type = "text",
        out = "data_stargazer.txt")

freMTPL2 <- freMTPL2[,-1]

############################
## ANÃLISIS DESCRIPTIVO ##
############################
############################
## Variables numericas ##
############################
#Seleccionar las variables numéricas
variables_numericas <- freMTPL2 %>%
  select_if(is.numeric)

summary(variables_numericas)

#Histograma
for (variable in colnames(variables_numericas)) {
  x <- variables_numericas[[variable]]
```

```
  fit <- density(x)

  fig <- plot_ly(x = x, type = "histogram", name = "Histograma") %>%

    add_lines(x = fit$x, y = fit$y,type = "scatter",fill = "tozeroy", yaxis = "y2", name = "Densidad",
fillcolor = 'rgba(255, 255, 255, 0)', line = list(color = "rgba(252, 192, 11, 0.8)")) %>%

    layout(title = list(text = variable, y = 0.01, x = 0.5, xanchor = 'center', yanchor =  'bottom'),
yaxis = list(title = "Frecuencia"), yaxis2 = list(overlaying = "y", side = "right", title = "Densidad"))


  print(fig)

}


#ClaimNb

summary(freMTPL2$ClaimNb)

tabla <- table(freMTPL2$ClaimNb)

a <- as.data.frame(tabla)

colnames(a) <- c("ClaimNb", "Freq")

stargazer(a, summary = FALSE, type = "text", out = "data_stargazer.txt")


a <- c(summary(freMTPL2$Exposure),"St.Dev"=sd(freMTPL2$Exposure))

stargazer(a,

        summary = FALSE,

        type = "text",

        out = "data_stargazer.txt")


#Exposure

a <- data.frame(t(c(summary(freMTPL2$Exposure), sd(freMTPL2$Exposure))))

colnames(a) <- c("Min.", "1st Qu.", "Median", "Mean", "3rd Qu.", "Max.", "St.Dev")

a <- round(a, 3)
```

```
pdf("data.pdf", height=18, width=14.5)
grid.table(a)
dev.off()

#Frequency
freMTPL2$Frequency <- round(freMTPL2$ClaimNb/freMTPL2$Exposure,2)
x <- freMTPL2$Frequency
a <- c(summary(x),"St.Dev"=sd(x), "Asimetría"=skewness(x), "Curtosis"=kurtosis(x))
stargazer(a, summary = FALSE, type = "text", out = "data_stargazer.txt")

skim(freMTPL2$Frequency)
pdf("data.pdf", height=18, width=14.5)
grid.table(a)
dev.off()


#Crear una columna adicional para indicar si es bonus o malus
freMTPL2 <- freMTPL2 %>%
  mutate(Tipo = ifelse(BonusMalus < 100, "Bonus", "Malus"))
tabla <- table(freMTPL2$Tipo)
porcentaje_bonus <- prop.table(tabla)["Bonus"] * 100
porcentaje_malus <- prop.table(tabla)["Malus"] * 100
tabla_con_porcentajes <- data.frame(Casos = c(tabla["Bonus"], tabla["Malus"]),
                                    Porcentaje = c(porcentaje_bonus, porcentaje_malus))
tabla_con_porcentajes <- round(tabla_con_porcentajes, 2)
tabla_con_porcentajes$Porcentaje <- paste0(tabla_con_porcentajes$Porcentaje, "%")
```

```
tabla_con_porcentajes

stargazer(tabla_con_porcentajes, summary = FALSE, type = "text", out = "data_stargazer.txt")


#Density, DrivAge, VehAge y VehPower

summary(variables_numericas)

sapply(variables_numericas, sd)

sapply(variables_numericas, skewness)

sapply(variables_numericas, kurtosis)

sapply(variables_numericas, jarque.bera.test)



#ClaimAmount

##histograma

hist(freMTPL2$ClaimAmount)


##top6

top_6_max <- freMTPL2[order(freMTPL2$ClaimAmount, decreasing = TRUE), ][1:6, ]

pdf("data.pdf", height=18, width=14.5)

grid.table(top_6_max)

dev.off()

sort(as.numeric(freMTPL2$ClaimAmount),decreasing = TRUE)

summary(freMTPL2$ClaimAmount)


##estadísticas descriptivas

a <- describe(freMTPL2$ClaimAmount)

stargazer(a[-1], summary = FALSE, type = "text", out = "data_stargazer.txt")
```

```
##test de normalidad

stargazer(rbind(broom::tidy(ad.test(freMTPL2$ClaimAmount)),broom::tidy(lillie.test(freMTPL2
$ClaimAmount))), summary = FALSE, type = "text", out = "data_stargazer.txt")


##boxplot y qqplot

par(mfrow = c(1, 2))

ggplot() +

  geom_boxplot(aes(y = freMTPL2$ClaimAmount), width = 0.6, color = "darkblue", outlier.color
= "black", outlier.shape = 16, outlier.size = 2) +

  labs(title = "Boxplot", y = "Valores") +

  scale_y_continuous(limits = c(0, max(freMTPL2$ClaimAmount))) +

  theme_minimal()

qqnorm(freMTPL2$ClaimAmount,main = 'QQ Plot for Normality',col = 'steelblue')

qqline(freMTPL2$ClaimAmount, col = 'red', lwd = 2, lty = 2)


#XYPlot

par(mar=c(1,1,1,1))

pairs(cbind(variables_numericas,Frequency=freMTPL2$Frequency), col="blue",
main="Scatterplots")


#CorrPlot

variables_numericas <- cbind(variables_numericas, Frequency = freMTPL2$Frequency)

ggcorrplot(round(cor(variables_numericas),2), hc.order = TRUE, type = "lower",

        lab = TRUE, lab_size = 3,

        outline.col = "white",

        ggtheme = ggplot2::theme_gray,

        colors = c("#6D9EC1", "white", "#E46726"))
```

```
############################
## Variables categoricas ##
############################
#VehBrand
freMTPL2$VehBrand <- factor(freMTPL2$VehBrand, levels = paste0("B", c(1:6,10:14)))
freq_table <- table(freMTPL2$VehBrand)
df_freq <- as.data.frame(freq_table)
colnames(df_freq) <- c("VehBrand", "Freq")
colors <- brewer.pal(length(levels(freMTPL2$VehBrand)), "Set3")
ggplot(df_freq, aes(x = VehBrand, y = Freq, fill = VehBrand)) +
  geom_bar(stat = "identity") +
  xlab("Brand") +
  ylab("Count") +
  scale_fill_manual(values = colors) +
  theme_minimal()

#VehGas
freq_table <- table(freMTPL2$VehGas)
df_freq <- as.data.frame(freq_table)
colnames(df_freq) <- c("VehGas", "Freq")
colors <- brewer.pal(length(levels(freMTPL2$VehGas)), "Set1")
ggplot(df_freq, aes(x = VehGas, y = Freq, fill = VehGas)) +
  geom_bar(stat = "identity") +
  xlab("Gas") +
  ylab("Count") +
  scale_fill_manual(values = colors) +
```

```
 theme_minimal()


#Area

color <- brewer.pal(length(levels(freMTPL2$Area)), "Set2")

pie(table(freMTPL2$Area),

   labels = paste(round(proportions(table(freMTPL2$Area)) * 100,2),"%",sep=""),

   col = color, density = 50, angle = 45)


legend("topright", levels(freMTPL2$Area), cex = 0.8, fill = color)


##Gráfico de barras (el promedio de claimamount por area)

promedio_por_area <- aggregate(ClaimAmount/ClaimNb ~ Area, data = freMTPL2, FUN =
mean)

promedios <- promedio_por_area$`ClaimAmount/ClaimNb`

barplot(promedios, names.arg = promedio_por_area$Area, xlab = "Area", ylab = "Average cost
of claims", col = "darkblue")


#Region

tabla <- table(freMTPL2$Region)

a <- as.data.frame(tabla)

a <- cbind(a, paste(round(proportions(table(freMTPL2$Region)) * 100,2),"%",sep=""))

colnames(a) <- c("Region", "Freq", "Porcentaje")

stargazer(a, summary = FALSE, type = "text", out = "data_stargazer.txt")


#Gráfico de barras (el promedio de claimamount por region)

promedio_por_region <- aggregate(ClaimAmount/ClaimNb ~ Region, data = freMTPL2, FUN =
mean)

promedios <- promedio_por_region$`ClaimAmount/ClaimNb`
```

```
par(mar=c(15,4,1,4))
barplot(promedios, names.arg = promedio_por_region$Region,
      ylab = "Average cost of claims",
      col = "darkred", las=2)



####################################
## Aplicación de las regresiones ##
####################################
#Regresiones cuantilicas
## Definir niveles de confianza deseados
niveles_confianza <- c(0.5, 0.75, 0.9, 0.95)

## Ajustar regresiones cuantilicas para diferentes niveles de confianza
models <- list()
for (i in 1:length(niveles_confianza)) {
  tau <- niveles_confianza[i]
  model <- rq(ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge + BonusMalus +
VehBrand + VehGas + Area + Density + Region, data = freMTPL2, tau = tau)
  models[[i]] <- model
}

##Imprimir resumenes de los modelos
for (i in 1:length(models)) {
  cat("Resumen del modelo para nivel de confianza tau =", niveles_confianza[i], ":\n")
  print(summary(models[[i]]))
  cat("\n")
}
```

```
##Salida de los modelos
a <- summary(models[[1]])
a <- a$coefficient[,c(1,4)]
stargazer(a, summary = FALSE, type = "text", out = "data_stargazer.txt")

a <- summary(models[[2]])
a <- a$coefficient[,c(1,4)]
stargazer(a, summary = FALSE, type = "text", out = "data_stargazer.txt")

a <- summary(models[[3]])
a <- a$coefficient[,c(1,4)]
stargazer(a, summary = FALSE, type = "text", out = "data_stargazer.txt")

a <- summary(models[[4]])
a <- a$coefficient[,c(1,4)]
stargazer(a, summary = FALSE, type = "text", out = "data_stargazer.txt")

##Comparar dos modelos
anova(models[[1]], models[[4]])

#regresion lineal
modelo.rl<- lm(ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge
        + BonusMalus + VehBrand + VehGas + Area + Density
        + Region, data=freMTPL2)
summary(modelo.rl)
```

```
#regresion cuantilica con tau desde 0.05 hasta 0.95

model.rq <- rq(ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge

        + BonusMalus + VehBrand + VehGas + Area + Density

        + Region, data=freMTPL2, tau=seq(0.05, 0.95, by=0.05))


plot(summary(model.rq), parm = c("Frequency","VehPower","VehAge","DrivAge"))

plot(summary(model.rq), parm =
c("BonusMalus","VehBrandB2","VehBrandB3","VehBrandB4"))

plot(summary(model.rq), parm =
c("VehBrandB5","VehBrandB6","VehBrandB10","VehBrandB11"))

plot(summary(model.rq), parm =
c("VehBrandB12","VehBrandB13","VehBrandB14","VehGasRegular"))

plot(summary(model.rq), parm = c("AreaB","AreaC","AreaD","AreaE"))

plot(summary(model.rq), parm = c("AreaF","Density","RegionAquitaine","RegionAuvergne"))

plot(summary(model.rq), parm = c("RegionBretagne","RegionBasse-
Normandie","RegionBourgogne","RegionCentre"))

plot(summary(model.rq), parm = c("RegionChampagne-
Ardenne","RegionCorse","RegionFranche-Comte","RegionHaute-Normandie"))

plot(summary(model.rq), parm = c("RegionIle-de-France","RegionLanguedoc-
Roussillon","RegionLimousin","RegionMidi-Pyrenees"))

plot(summary(model.rq), parm = c("RegionNord-Pas-de-Calais","RegionPays-de-la-
Loire","RegionPicardie","RegionPoitou-Charentes"))

plot(summary(model.rq), parm = c("RegionProvence-Alpes-Cotes-D'Azur","RegionRhone-
Alpes"))


##Cuatro variable con diferencia significativa

plot(summary(model.rq), parm = c("Frequency","VehAge","DrivAge","BonusMalus"))
```

```
#Regresiones conjunta VaR & ES

freMTPL2$ClaimAmount <- freMTPL2$ClaimAmount/1000

freMTPL2$ClaimAmount <- -freMTPL2$ClaimAmount

##Modelo 1

modelrqes1 <- esreg(ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge +
BonusMalus + Density,

        freMTPL2,

        alpha = 0.05,

        g1 = 1,

        g2 = 1)

##Obtener un resumen del modelo 1

summary(modelrqes1)


##Modelo 2

modelrqes2 <- esreg(ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge +
BonusMalus + Density,

         freMTPL2,

         alpha = 0.05,

         g1 = 1,

         g2 = 2)

##Obtener un resumen del modelo 2

summary(modelrqes2)


##Modelo 3

modelrqes3 <- esreg(ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge +
BonusMalus + Density,

         freMTPL2,

         alpha = 0.05,
```

```
        g1 = 1,
        g2 = 3)
##Obtener un resumen del modelo 3
summary(modelrqes3)

##Modelo 4
modelrqes4 <- esreg(ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge +
BonusMalus + Density,
        freMTPL2,
        alpha = 0.05,
        g1 = 1,
        g2 = 4)
##Obtener un resumen del modelo 4
summary(modelrqes4)

##Modelo 5
modelrqes5 <- esreg(ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge +
BonusMalus + Density,
        freMTPL2,
        alpha = 0.05,
        g1 = 1,
        g2 = 5)
##Obtener un resumen del modelo 5
summary(modelrqes5)

##Modelo 6
modelrqes6 <- esreg(ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge +
BonusMalus + Density,
```

```
            freMTPL2,
            alpha = 0.05,
            g1 = 2,
            g2 = 1)
##Obtener un resumen del modelo 6
summary(modelrqes6)


##Modelo 7
modelrqes7 <- esreg(ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge +
BonusMalus + Density,
            freMTPL2,
            alpha = 0.05,
            g1 = 2,
            g2 = 2)
##Obtener un resumen del modelo 7
summary(modelrqes7)


##Modelo 8
modelrqes8 <- esreg(ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge +
BonusMalus + Density,
            freMTPL2,
            alpha = 0.05,
            g1 = 2,
            g2 = 3)


##Obtener un resumen del modelo 8
summary(modelrqes8)
```

```
##Modelo 9

modelrqes9 <- esreg(ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge +
BonusMalus + Density,

                    freMTPL2,

                    alpha = 0.05,

                    g1 = 2,

                    g2 = 4)


##Obtener un resumen del modelo 9

summary(modelrqes9)


##Modelo 10

modelrqes10 <- esreg(ClaimAmount ~ Frequency + VehPower + VehAge + DrivAge +
BonusMalus + Density,

                    freMTPL2,

                    alpha = 0.05,

                    g1 = 2,

                    g2 = 5)

##Obtener un resumen del modelo 10

summary(modelrqes10)


#Coefficientes normalizados

options("scipen"=100, "digits"=4)

summary(modelrqes1)$coefficients_q*-1000

summary(modelrqes1)$coefficients_e*-1000

summary(modelrqes2)$coefficients_q*-1000

summary(modelrqes2)$coefficients_e*-1000

summary(modelrqes3)$coefficients_q*-1000
```

```
summary(modelrqes3)$coefficients_e*-1000

summary(modelrqes6)$coefficients_q*-1000

summary(modelrqes6)$coefficients_e*-1000

summary(modelrqes7)$coefficients_q*-1000

summary(modelrqes7)$coefficients_e*-1000


#Tabla de residuos

##MSE VaR & ES

t_error1 <- data.frame(VaR_0.95=c(mean(residuals(modelrqes1)[,1]^2),

                      mean(residuals(modelrqes2)[,1]^2),

                      mean(residuals(modelrqes3)[,1]^2),

                      mean(residuals(modelrqes6)[,1]^2),

                      mean(residuals(modelrqes7)[,1]^2)),

                  ES_0.95=c(mean(residuals(modelrqes1)[,2]^2),

                      mean(residuals(modelrqes2)[,2]^2),

                      mean(residuals(modelrqes3)[,2]^2),

                      mean(residuals(modelrqes6)[,2]^2),

                      mean(residuals(modelrqes7)[,2]^2)))

stargazer(t_error1, summary = FALSE, type = "text")


##MSE Y SSE de los modelos

t_error2 <- data.frame(MSE=c(mean(residuals(modelrqes1)^2),

                     mean(residuals(modelrqes2)^2),

                     mean(residuals(modelrqes3)^2),

                     mean(residuals(modelrqes6)^2),

                     mean(residuals(modelrqes7)^2)),

                 SSE=c(sum(residuals(modelrqes1)^2),
```

```
                    sum(residuals(modelrqes2)^2),
                    sum(residuals(modelrqes3)^2),
                    sum(residuals(modelrqes6)^2),
                    sum(residuals(modelrqes7)^2)))
stargazer(t_error2, summary = FALSE, type = "text", out = "data_stargazer.txt")

#Predicted versus Observed
data_mod <- data.frame(Predicted = predict(modelrqes2)*(-1000),
                       Observed = freMTPL2$ClaimAmount*(-1000))

ggplot(data_mod,
       aes(x = Predicted.1,
           y = Observed)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linewidth = 2) +
  labs(x = "VaR_0.95")

ggplot(data_mod,
       aes(x = Predicted.2,
           y = Observed)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
```

```
            linewidth = 2) +
  labs(x = "ES_0.95")


data_mod <- data.frame(Predicted = predict(modelrqes7)*(-1000),
                       Observed = freMTPL2$ClaimAmount*(-1000))
ggplot(data_mod,
       aes(x = Predicted.1,
           y = Observed)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linewidth = 2) +
  labs(x = "VaR_0.95")


ggplot(data_mod,
       aes(x = Predicted.2,
           y = Observed)) +
  geom_point() +
  geom_abline(intercept = 0,
              slope = 1,
              color = "red",
              linewidth = 2) +
  labs(x = "ES_0.95")
```

6. REFERENCIAS

[1] Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26, 1505-1518. https://doi.org/10.1016/S0378-4266(02)00281-9

[2] ARTZNER, P., DELBAEN, F., EBER J. y HEATH, D. (1999). Coherent Measures of Risk. *Mathematical Finance*, 203-228.

[3] Dimitriadis, T. y Bayer, S. (2019). A Joint Quantile and Expected shortfall Regression Framework. Electronic Journal of Statistics. https://doi.org/10.1214/19-ejs1560

[4] Fissler, T., Ziegel, J. F. y Gneiting, T. (2016). Expected shortfall is jointly elicitable with value at risk – implications for backtesting. *Risk Magazine*.

[5] Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106.746-762.

[6] Koenker, R. y Bassett, G. W. (1978). Regression Quantiles. *Econometrica*, 46, 33-50. https://doi.org/10.2307/1913643

[7] Koenker, R. (2005). Quantile Regression in R: A Vignette. *En Cambridge University Press eBooks*, 295-316. https://doi.org/10.1017/cbo9780511754098.011

[8] Marmol Jimenez Mª. T. *Solvència a curt termini de carteres de riscos,*97-98. Universidad de Barcelona.

[9] McNeil, A. J., Frey, R. y Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques, and Tools.* Princeton University Press. Estados Unidos.

[10] Ortiz-Gracia, L. *Quantitative Finance,* 19. Universidad de Barcelona.

[11] Otero, J. V., y Reyes, B. S. (2012). *Regresión cuantílica: estimación y contrastes*.

[12] Weber, S. (2006). Distribution invariant risk measures, information, and dynamic consistency. *Mathematical Finance*, 16, 419-441.

[13] Ziegel, J. F. (2015). Higher order elicitability and Osband's principle. *Annals of Statistics*, 44. https://doi.org/10.1214/16-aos1439

[14] Ziegel, J. F., Krueger, F., Jordan, A. y Fasciati, F. (2017). *Murphy Diagrams: Forecast Evaluation of Expected shortfall.*