*Anales del Instituto de Actuarios Españoles, 4ª época, 25, 2019 /101-117*

# QUANTILE REGRESSION AS A STARTING POINT IN PREDICTIVE RISK MODELS

# REGRESIÓN CUANTÍLICA COMO PUNTO DE PARTIDA EN LOS MODELOS PREDICTIVOS PARA EL RIESGO

Albert Pitarque, Ana María Pérez-Marín[1], Montserrat Guillen

Dept. Econometría, Riskcenter-IREA, Universidad de Barcelona, Av. Diagonal, 690, 08034 Barcelona

Fecha de recepción: 1 de agosto de 2019
Fecha de aceptación: 23 de octubre de 2019

**Abstract**

Given a risk level or tolerance, quantile regression is a predictive model that fits the corresponding percentile of the continuous response variable. Given a fixed percentage value, we identify the effect of each predictor variable in the cumulative distribution up to that level of the dependent variable. In this article, we show how this methodology can be used in motor insurance data analysis and we propose an extension of quantile regression inspired by the need to predict the expectation of the conditional tail. To this end, specific R routines have been developed and a resampling procedure has been implemented to approximate standard errors. The main conclusion is that this type of models allows us to analyze which factors affect accident risk and can be used to mitigate or to evaluate risk in the insurance field.

**Keywords:** predictive modelling, value-at-risk, tail value at-risk, optimization, resampling

Se agradece la financiación recibida del proyecto del Ministerio de Economía y Competitividad y FEDER, ECO2016-76203-C2-2-P del programa ICREA Academia y de las ayudas Fundación BBVA para equipos de investigación en big data.
[1] Autor para correspondencia: amperez@ub.edu

DOI 10.26360/2019_5

**Resumen**

Dado un nivel o tolerancia de riesgo, la regresión cuantílica es un modelo predictivo que ajusta el correspondiente percentil de la variable respuesta continua. Fijado un determinado valor porcentual, se identifica el efecto de cada variable predictora en la distribución acumulada hasta ese nivel de la variable dependiente. En este artículo mostramos cómo puede utilizarse esta metodología en el análisis de datos en el seguro de automóvil y proponemos una extensión de la regresión cuantílica inspirada en la necesidad de predecir la esperanza de la cola condicional. Para ello se han desarrollado rutinas específicas en R y se ha implementado un procedimiento de remuestreo para la aproximación de los errores estándar. La principal conclusión es que este tipo de modelos permite analizar qué factores inciden en el riesgo de accidente y pueden ser utilizados para mitigarlo o para valorarlo en el ámbito asegurador.

**Palabras clave:** modelización predictiva, valor en riesgo, valor en riesgo de la cola, optimización, remuestreo.

## 1. Introducción, conceptos previos y motivación

Este trabajo trata de modelos orientados a la predicción del valor en riesgo, o lo que es lo mismo, el percentil o el cuantil y de otras medidas de riesgo de una variable dependiente en función de un conjunto de variables explicativas. A diferencia de los modelos de regresión habituales, donde el interés reside en predecir el valor promedio, en el caso de la regresión cuantílica, se desea conocer el riesgo estimado condicionado a los valores concretos de los factores explicativos y para un nivel de tolerancia dado, por ejemplo, el 95% o el 99%.

Históricamente, uno de los primeros autores que habló del concepto de regresión fue Roger J. Boscovich en el siglo XVIII. Este físico y matemático restringía la media de los residuos de la regresión a ser cero y para estimar el efecto de una variable sobre otra, planteaba minimizar la suma de los residuos en valor absoluto. Este tipo de regresión fue conocido como LAD (*Least Absolute Deviations*) y más tarde fue generalizada por Pierre-Simon Laplace para el caso de más variables. En definitiva, este antecedente, que es anterior a Carl F. Gauss con el método de los mínimos cuadrados ordinarios, trataba de ajustar la mediana de la variable dependiente, que no es más que el cuantil al 50%. A finales del siglo XIX, Francis Y. Edgeworth proporcionó un algoritmo para ajustar regresiones de forma que la suma de

las desviaciones absolutas, o residuos en valor absoluto, fuera mínima. Sin embargo, dicho método y algunas generalizaciones posteriores no se consolidaron porque requerían disponer de mejores métodos de optimización y capacidad computacional. Hasta bien entrado el siglo XX, los estudios que utilizaban la regresión cuantílica se centraban en estimar una regresión para la mediana. Entre los primeros en estudiar la regresión para diferentes cuantiles se encuentra un artículo de Koencker y Bassett (1978), cuyas propuestas han ido evolucionando hasta hoy.

El valor del $\tau$-cuantil de una variable aleatoria continua, *Y*, es aquel valor $c_\tau$ para el que se cumple que la probabilidad de que la variable no tome valores superiores al mismo es igual a $\tau$, es decir $P(Y \leq c_\tau) = \tau$ donde $c_\tau$ es el $\tau$-cuantil o el percentil $\tau$, que recibe el nombre de valor en riesgo al nivel $\tau$ en el ámbito asegurador y financiero y se denota $VaR_\tau(Y)$.

De igual modo, se define el valor del riesgo de la cola (*Tail Value at Risk*) al nivel $\tau$ para la variable *Y*, $TVaR_\tau(Y)$, como la esperanza de la cola condicionada, es decir, el valor esperado de la variable *Y* por encima de $c_\tau$. Esta medida de riesgo también se denomina Expected Shortfall ($ES_\tau$) o Conditional Value at Risk ($CVaR_\tau$) y se corresponde con la esperanza de los valores que superan el $VaR_\tau$ y puede expresarse como:

$$TVaR_\tau(\mathrm{Y}) = E(Y|Y > c_\tau) = \frac{1}{1-\tau}\int_{c_\tau}^{\infty} yf(y)dy, \qquad (1)$$

donde $f(y)$ es la función de densidad de probabilidad de la variable aleatoria *Y*. Pueden consultarse más detalles en Hardy (2006), quien realiza una introducción a las medidas de riesgo con aplicaciones actuariales.

El problema que planteamos en este trabajo es el de estimar un modelo de regresión para ambas medidas de riesgo. Hasta el momento, la regresión cuantílica, que cubriría la primera de las dos medidas, está resuelto, pero no así el segundo caso.

De los trabajos en los que se ha aplicado la regresión cuantílica muy pocos estudian datos del ámbito asegurador. Kudryavtsev (2009) aplica la regresión cuantílica a la tarificación de un seguro de robo, analizando la severidad de la pérdida. Pitt (2006) se centra en el ámbito de los seguros de protección de ingresos. Nuestra motivación es ir más allá en la implementación y desarrollo de la metodología y mostrar su creciente interés en las ciencias actuariales.

En la sección 2 de este artículo, se introduce la regresión cuantílica y se explica cómo se determina la bondad de ajuste de estos modelos. A

continuación, en la sección 3, se describe qué metodología se ha propuesto para realizar la aproximación al TVaR. En la sección 4 se presentan los datos utilizados en este trabajo y los resultados que se han obtenido. Por último, en la sección 5, se comentan las principales conclusiones.

## 2. Introducción a la regresión cuantílica

Para comprender la regresión cuantílica, hay que recordar que la regresión lineal es una aproximación que fija una relación entre la variable respuesta (o variable dependiente) y una o más variables explicativas (o variables independientes) con la expresión siguiente:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \qquad (2)$$

donde $Y_i$ corresponde a la variable dependiente para el *i*-ésimo caso de la muestra ($i=1,\ldots,n$) y $X_{ji}$, las correspondientes observaciones de las $k$ variables explicativas, siendo $j=1\ldots,k$. Se tiene en cuenta un término de perturbación $\varepsilon_i$ que recoge las desviaciones respecto de la media. En este caso, dado que se asume que el término de perturbación está centrado en cero, se obtiene que:

$$E(Y_i \,|X_{1i}, \ldots, X_{\mathrm{k}i}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}. \qquad (3)$$

Los coeficientes se estiman por mínimos cuadrados ordinarios de forma que:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, S(\beta) \qquad (4)$$

donde $S(\beta) = ||Y - X\beta||^2$ representa una distancia entre el vector de observaciones de la variable dependiente y el de predictores lineales calculados como la combinación lineal (3) para cada componente. Con la norma euclídea, $S(\beta)$ corresponde a la suma de los residuos al cuadrado.

En la regresión cuantílica, se desea encontrar una relación entre el cuantil de la variable dependiente, condicionado a los valores de las variables explicativas de manera que:

$$VaR_\tau(Y_i \,|X_{1i}, \ldots, X_{ki}) = \beta_0^\tau + \beta_1^\tau X_{1i} + \beta_2^\tau X_{2i} + \ldots + \beta_k^\tau X_{ki}. \qquad (5)$$

De ese modo, se puede demostrar (ver Koenker y Bssett, 1978) que.

$$\widehat{\beta^{\tau}} = \underset{b}{\operatorname{argmin}} \left[ \sum_{Y_i \geq X_i' b} \tau |Y_i - X_i' \mathrm{b}| + \sum_{Y_i < X_i' b} (1-\tau) |Y_i - X_i' \mathrm{b}| \right]. \quad (6)$$

A diferencia de lo que se sabe para los estimadores de los coeficientes de la regresión por mínimos cuadrados ordinarios que siguen una distribución t-Student, para los coeficientes de la regresión cuantílica no se conoce la distribución exacta de los estimadores. Aun así, bajo ciertas condiciones, se ha estudiado que $\sqrt{n}\left(\widehat{\beta^{\tau}} - \beta^{\tau}\right)$ tiende a una distribución Normal. Se suele decir que la función objetivo en (6) se corresponde con la suma *n* componentes, donde cada una de ellas puede expresarse como:

$$\rho_{\tau}(Y_i - X_i' \mathrm{b}) = \tau(Y_i - X_i' \mathrm{b}) I_{\{Y_i \geq X_i' b\}} + (\tau - 1)(Y_i - X_i' \mathrm{b}) I_{\{Y_i < X_i' b\}} =$$

$$= (Y_i - X_i' \mathrm{b})(\tau - I_{\{Y_i < X_i' b\}}), \quad (7)$$

siendo $I_{\{.\}}$ una función indicadora que vale 1 si la condición del subíndice se cumple y 0 en caso contrario.

Koenker y Machado (1999) propusieron una expresión para medir la bondad de ajuste en la regresión cuantílica basada en la comparación de las funciones objetivo del modelo de regresión cuantílica estimado y del modelo restringido que sólo incluye el término independiente. En concreto, sea

$$\hat{V}(\tau) = \sum_{y=1}^{n} \rho_{\tau}(Y_i - X_i' \widehat{\beta^{\tau}}) \quad (10)$$

el valor de la función objetivo del modelo completo que incluye todos los parámetros y

$$\tilde{V}(\tau) = \sum_{y=1}^{n} \rho_{\tau}(Y_i - \beta_0^{\tau}) \quad (11)$$

el valor de la función objetivo del modelo que sólo incluye el término independiente. Entonces, la medida de bondad de ajuste propuesta por Koenker y Machado (1999) es

$$R^1(\tau) = 1 - \hat{V}(\tau)/\tilde{V}(\tau) \quad (12)$$

que es el análogo al $R^2$ del modelo de regresión lineal múltiple.

## 3. Metodología propuesta para aproximar la regresión al $TVaR$

Si bien la medida del riesgo utilizada en la regresión cuantílica es el valor en riesgo al nivel $\tau$, uno de sus principales problemas de es que al tomar sólo el valor del cuantil, es una medida del riesgo que no tiene en cuenta las pérdidas superiores al mismo. En cambio, como hemos visto en la sección 1, el $TVaR_\tau$ (esperanza de los valores que superan el $VaR_\tau$) sí los tiene en cuenta.

Nuestra aportación parte de la base de los artículos de Koenker sobre la regresión cuantílica y, habiendo comprobado mediante los recientes resultados de Fissler y Ziegel (2016) y Acerbi y Székely (2014) que se puede aplicar la regresión cuantílica sobre otras medidas de riesgo, se establece una nueva función de pérdida similar a la que se optimiza para el cuantil que permitirá ajustar una regresión cuantílica paramétrica cercana al $TVaR_\tau$. Dicha regresión es una extensión de la regresión cuantílica para un cuantil $\tau$ y por eso la manera de calcular cuáles son los efectos de las variables es bastante similar a la del estimador de la regresión cuantílica.

Para calcular los valores de los coeficientes de la regresión cuantílica, es necesario minimizar la función objetivo (6), que pondera el valor absoluto de las desviaciones de acuerdo con el nivel $\tau$ prefijado. Partiendo de la definición de *TVaR* en (1), se puede interpretar el *TVaR* como una esperanza matemática y para calcular la esperanza de una variable aleatoria continua con función de densidad *f(x)*, se utiliza la expresión clásica: $E(X) = \int_{-\infty}^{\infty} xf(x)dx$.

Cuando se desea calcular la esperanza de toda la variable aleatoria, los límites de la integral abarcan todo el dominio, es decir, van desde $-\infty$ hasta $+\infty$ para variable aleatorias no acotadas.

En nuestro caso, el interés se encuentra en la esperanza de los valores que se encuentran por encima del valor $c_\tau$. Una expresión alternativa a (1) es:

$$TVaR_\tau(\mathrm{Y}) = \frac{1}{1-\tau}\int_\tau^1 VaR_{\mathrm{u}}(Y)du.$$

Como no es posible obtener directamente una función objetivo que permita tener el análogo a (6) para el $TVaR_\tau$ lo que hacemos es considerar la función objetivo $\rho_\tau$ en (7) que se utiliza para estimar los coeficientes para la regresión cuantílica, como una densidad y calcular la esperanza en la cola, es decir integrando desde $\tau$ hasta 1. Resolviendo esta integral se obtiene la siguiente expresión (ver Pitarque, 2019):

$$\int_{\tau}^{1} \tau\rho_{\tau}(Y - X\beta)d\tau = (Y - X\beta)\int_{\tau}^{1} \tau(\tau - I_{(Y-X\beta)<0})d\tau$$
$$= (Y - X\beta)\left(\frac{1}{3} - \frac{I_{(Y-X\beta)<0}}{2} - \frac{\tau^3}{3} + \frac{\tau^2 I_{(Y-X\beta)<0}}{2}\right). \quad (8)$$

Finalmente, los coeficientes del modelo inspirado en el $TVaR$ se obtienen al minimizar la siguiente función objetivo:

$$\beta_{\tau}\widehat{(TVaR)} =$$

$$\underset{\beta}{\operatorname{argmin}} \sum_{y=1}^{n} \left[(Y_i - X_i'\beta)\left(\frac{1}{3} - \frac{I_{(Y_i - X_i'\beta)<0}}{2} - \frac{\tau^3}{3} + \frac{\tau^2 I_{(Y_i - X_i'\beta)<0}}{2}\right)\right] \quad (9)$$

La función objetivo y su minimización se ha programado en *R* y la obtención de sus errores estándar se ha efectuado mediante un método de remuestreo. Además, se ha adaptado un cálculo de la bondad de ajuste análogo al de la medida propuesta por Koenker y Machado (1999) en la regresión cuantílica.

## 4. Datos y resultados

La base de datos utilizada en este trabajo sirve para modelizar la distancia que recorren en exceso de velocidad los conductores de una muestra de asegurados de una entidad española. La tarificación en función del pago por uso se está extendiendo en todo el mundo y, a pesar de que no hay todavía mucha tradición, la posibilidad de contratar este tipo de pólizas, donde el precio está en función del patrón de conducción, es previsible que se generalice en los próximos años. Para este tipo de seguros, la información sobre los patrones de conducción se recoge a través de un dispositivo que graba los datos básicos telemáticos tales como la velocidad o el kilometraje. La muestra utilizada en este trabajo es de 7,691 conductores jóvenes (de 18 a 32 años) con este tipo de póliza del automóvil en vigor durante el año 2010. La razón de que no haya conductores de más edad en esta base de datos es que la entidad aseguradora sólo comercializó este seguro entre gente joven. La información correspondiente a cada conductor incluye un total de seis variables, que se presentan en la Tabla 1. Anteriormente, diversos autores han realizado trabajos de otra índole con estos mismos datos. En concreto, Boucher, Côte, y Guillen (2017) estudiaron la transformación de los factores de riesgo; Ayuso, Guillen, y Pérez-Marín (2016a, 2016b) analizaron los efectos de la distancia recorrida hasta que se produce el primer accidente viendo además la diferencia en la intensidad y patrones de conducción entre hombres y mujeres. Finalmente, Guillen, Nielsen, Ayuso, y Pérez-Marín

(2019) trataron cómo diseñar tarifas que tengan en cuenta la presencia elevada de ceros en el número de siniestros declarados por los asegurados y como usar la información de los datos telemáticos.

Tabla 1
*Variables de la Base de Datos*

| Variable | Descripción |
|---|---|
| Toler_km | Total de kilómetros recorridos por encima de los límites de velocidad durante el año 2010. |
| lnKm | Logaritmo del total de kilómetros recorridos durante 2010. |
| Porc_urba | Porcentaje de kilómetros recorridos por vías urbanas |
| Porc_noct | Porcentaje de kilómetros recorridos en horarios nocturno. |
| Edad | Edad del conductor a 1 de enero de 2010. |
| Sexo | 1 = hombre, 0 = mujer. |

Tabla 2
*Estadísticos Descriptivos*

| Variable | Media | Desv. Est. | Mediana | Mín. | Máx. | Asimetría | Curtosis |
|---|---|---|---|---|---|---|---|
| Toler_km | 1,400.12 | 2,008.67 | 689.42 | 0.00 | 23,500.19 | 3.70 | 23.75 |
| lnKm | 9.26 | 0.76 | 9.37 | -0.37 | 10.96 | -1.95 | 13.38 |
| Porc_urba | 26.36 | 14.29 | 23.47 | 0.00 | 100 | 1.56 | 6.33 |
| Porc_noct | 7.01 | 6.10 | 5.30 | 0.00 | 46.34 | 1.03 | 4.16 |
| Edad | 24.77 | 2.82 | 24.61 | 18.11 | 31.56 | 0.10 | 2.23 |

La Tabla 2 muestra los estadísticos descriptivos de las variables de la base de datos. Un 50.88% de los individuos son hombres, es decir, hay aproximadamente el mismo número de hombres que de mujeres. El rango de edad de los conductores se sitúa entre los 18 y los 32 años, con la media de edad situada en los 24 años. El coeficiente de asimetría es muy bajo por lo tanto la variable es bastante simétrica. En relación con la variable dependiente *Toler_km*, que mide el total de kilómetros recorridos con exceso de velocidad, como se observa a la Tabla 2, los conductores suelen circular por encima del límite de velocidad bastantes kilómetros al años, pero existen casos extremos siendo la media es muy superior a la mediana. Sin embargo, hay muchos casos en que el conductor no ha superado en ningún momento el límite de velocidad. Este hecho también puede ser comprobado mirando el coeficiente de asimetría que es positivo y muy elevado. La curtosis es muy alta debido a que la mayor parte de los conductores no supera el límite de

velocidad y si lo hacen, lo hacen durante pocos kilómetros. También se observa que hay valores muy elevados puesto que el máximo es de 23,500.19 kilómetros por encima del límite de velocidad. Es probable que estos valores tan elevados estén relacionados con la zona de conducción por ejemplo con límites de velocidad muy bajos y poca congestión, o la edad de los conductores, puesto que se cree que cuanto más joven es el conductor más conduce por encima de los límites de velocidad. Otra posible razón puede ser la profesión del conductor, dado que determinadas profesiones requieren muchos desplazamientos. Aun así, estas suposiciones deben ser contrastadas.

Los kilómetros totales conducidos se introducen en el modelo en escala logarítmica. En media, se recorren unos 10,000 kilómetros durante el periodo en el que se registran sus datos de conducción (un año), y la mediana es bastante similar. Observamos el valor máximo correspondiente a una distancia anual es de 57,000 kilómetros. El coeficiente de asimetría es negativo y bastante elevado e indica que hay cierta asimetría por la izquierda, indicativo de que hay más valores bajos que elevados. La curtosis de esta variable también es muy elevada, lo que indica que la distribución es más apuntada que la normal y una gran parte de las observaciones toman valores similares.

Respecto a las zonas de conducción más habituales, se observa que los conductores usan sobretodo vías interurbanas, dado que en promedio el porcentaje de conducción por vía urbana se sitúa sólo en el 26% del total del kilometraje registrado. No obstante, el valor máximo es igual a 100, lo que indica que hay conductores que sólo conducen por áreas urbanas, probablemente sólo utilizan su vehículo para ir de casa al trabajo en zonas metropolitanas. Esta variable también presenta cierta asimetría por la derecha y un coeficiente de curtosis elevado.

Respecto a la franja horaria de conducción, los conductores de la muestra no suelen conducir por la noche puesto que el promedio de kilómetros recorridos en horario nocturno es del 7%. El valor más elevado en este caso es 46.34% de desplazamientos nocturnos, que puede ser por varias razones; o bien el conductor trabaja por la noche o es una persona, posiblemente joven, que utiliza el coche por ocio. Esta variable también presenta cierta asimetría por la derecha, pero no muy elevada. La curtosis también es positiva, pero menos elevada que la observada en el resto de variables.

### *4.1 Regresión cuantílica*

En esta sección presentamos los resultados del ajuste de la regresión cuantílica al análisis del número de kilómetros recorridos con exceso de velocidad. En la Tabla 3 se presentan los valores de los coeficientes estimados para los diferentes cuantiles, y para el caso $\tau = 0{,}9$ se incluyen además los p-valores correspondientes. Para su obtención se ha utilizado la función *rq* del paquete *quantreg* de R (Koenker, Portnoy, Ng, Zeileis, Grosjean, y Ripley, 2018).

Tabla 3
*Coeficientes Estimados del Modelo de Regresión Cuantílica y sus Errores Estándar entre Paréntesis*

| **Variable** | $\boldsymbol{\beta_{lm}}$ | $\boldsymbol{\beta_{rq}^{0.25}}$ | $\boldsymbol{\beta_{rq}^{0.5}}$ | $\boldsymbol{\beta_{rq}^{0.75}}$ | $\boldsymbol{\beta_{rq}^{0.90}}$ | **p-valor** |
|---|---|---|---|---|---|---|
| Constante | -8,120.85 | -2,824.98 (148.44) | -4,588.69 (273.60) | -6,281.67 (470.82) | -6,451.74 (1,032.88) | < 0.0001 |
| lnKm | 1,062.54 | 359.81 (14.90) | 603.55 (27.89) | 894.77 (44.62) | 1,086.12 (90.46) | < 0.0001 |
| Porc_Vurba | -21.31 | -2.95 (0.45) | -9.11 (0.86) | -21.36 (1.91) | -38.64 (3.32) | < 0.0001 |
| Porc_Noctur | 5.35 | 3.18 (1.18) | 3.49 (2.31) | 4.07 (5.38) | 19.69 (12.64) | 0.12 |
| Edad | 1.37 | -2.77 (2.43) | -0.52 (4.70) | 2.42 (9.48) | 2.19 (20.47) | 0.91 |
| Sexo | 329.98 | 97.51 (14.07) | 204.34 (27.82) | 364.79 (58.18) | 582.63 (141.87) | < 0.001 |
| Bondad de ajuste | - | 0.1937 | 0.1380 | 0.5114 | 0.6924 | |

$B_{rq}^{\tau}$ representa los coeficientes del modelo para el percentil $\tau$ obtenidos con la función $rq$, así como el valor de la bondad de ajuste. $B_{lm}$ representa los coeficientes del modelo de regresión lineal múltiple.

En la Figura 1 se representan los gráficos de la evolución de los efectos de las variables. Observando los gráficos de la Figura 1 se observa que, en relación con el término independiente, no se puede considerar que su efecto sea igual al de la constante del modelo lineal. Este efecto siempre es negativo y decrece de forma más o menos constante hasta el cuantil 0.75. A partir del cuantil 0.75 el valor del coeficiente sigue decreciendo pero lo hace de una forma menos brusca.

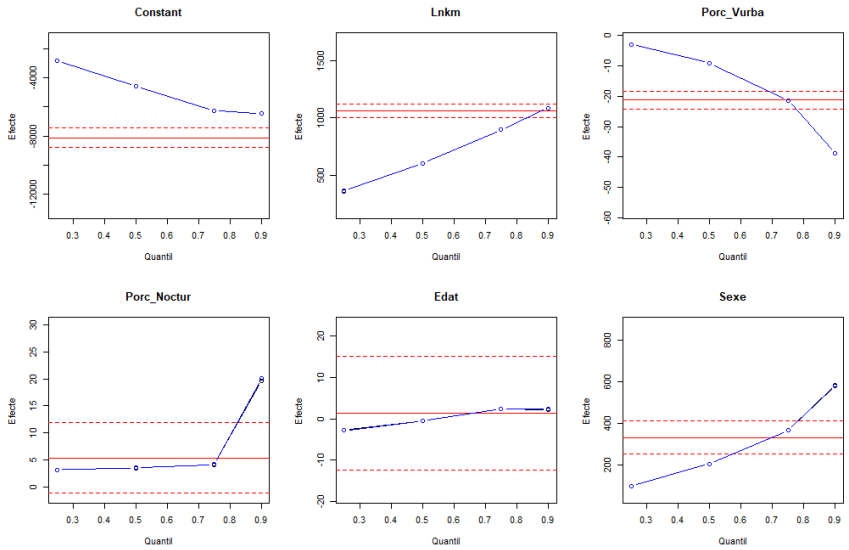


*Figura 1*. Gráficos de la evolución de los efectos de las variables a medida que aumenta el nivel del cuantil en la Base de Datos de Seguros utilizando la Regresión Cuantílica.

Para la variable que indica el número total de kilómetros recorridos el efecto es creciente durante todos los cuantiles y bastante constante. Desde el cuantil 0.25 hasta el 0.75 se puede considerar que este es diferente del efecto del modelo lineal. Para el cuantil 0.9 el efecto sí que se considera igual al del modelo lineal y resulta significativo.

Para la variable que indica el porcentaje de kilómetros recorridos por vía urbana, el efecto va decreciendo cada vez más a medida que el cuantil aumenta. Para el cuantil 0.25, este efecto es prácticamente nulo y a medida que aumenta el cuantil toma valores cada vez más negativos. Para los cuantiles 0.25 y 0.5, el efecto de esta variable hace disminuir menos el cuantil de la variable respuesta que el efecto estimado para el modelo lineal. Para el cuantil 0.75, el efecto es igual que el del modelo lineal y para el

cuantil 0.9 el efecto es significativo y vuelven a existir diferencias respecto al modelo lineal.

Para la variable que indica el porcentaje de kilómetros conducidos por la noche el efecto es prácticamente el mismo que el del modelo lineal a pesar de que va aumentando lentamente a medida que aumenta el cuantil. A partir del cuantil 0.75 este crecimiento se hace más pronunciado llegando a diferenciarse del efecto del modelo lineal. No obstante, para el cuantil 0.9 la conducción nocturna no tiene efecto significativo.

Respecto a las variables que tienen relación con las características personales del conductor, vemos que en cuanto a la edad, el efecto prácticamente no varía para diferentes niveles. Para los dos primeros cuantiles que se han estudiado esta variable tiene efecto negativo y para los cuantiles 0.75 y 0.9 tiene efecto positivo Aun así en ninguno de los casos se puede considerar que este efecto sea diferente al del modelo lineal. Para el cuantil 0.9 el efecto de la edad no es significativo. En relación a la variable sexo, el comportamiento es el inverso del de la variable que recoge la conducción por vía urbana, a medida que el cuantil va aumentando, el efecto aumenta en mayor medida. Para los cuantiles 0.25 y 0.5 el efecto es menor que el del modelo lineal y se puede considerar diferente de éste. Para el cuantil 0.75 el efecto se puede considerar igual y para el cuantil 0.9 se puede volver a considerar diferente y mayor, siendo además significativo.

Finalmente, la Tabla 3 también proporciona los valores de la la bondad de ajuste siguiendo la fórmula (12) propuesta por Kroencker y Machado (1999). Lo más destacable es que el ajuste tiende a mejorar a medida que aumenta el nivel de $\tau$, con una ligera disminución en la mediana.

### *4.2 Regresión para el TVaR*

En esta sección se muestran los resultados del ajuste del *TVaR* a la misma base de datos. La Tabla 4 recoge los resultados del ajuste del modelo lineal así como del *TVaR* para diferentes niveles de $\tau$, en concreto, $\tau$ = (0.25; 0.5; 0.75; 0.9). Los efectos para $\tau$ = 0,9 se analizan de forma más detallada, incluyéndose en este caso también los p-valores correspondientes.

Observamos que la constante toma valores negativos para todos los modelos ajustados. Por lo que respecta al logaritmo de los kilómetros recorridos, vemos que su coeficiente es positivo. Para $\tau$ = 0.9 el coeficiente toma valor 1,091.75 y es significativo. Eso significa que cuando el logaritmo del

kilometraje total aumenta en una unidad (por lo tanto el total del kilometraje se multiplica por 2.718) el *TVaR* al nivel 90% de la distancia recorrida por encima de los límites de velocidad aumenta en 1,091.75 km.

Tabla 4
*Coeficientes Estimados del Modelo de Regresión Cuantílica para el Modelo Inspirado en el Tvar y sus Errores Estándar entre Paréntesis*

| **Variable** | $\boldsymbol{\beta_{lm}}$ | $\boldsymbol{\beta_{TVaR}^{0.25}}$ | $\boldsymbol{\beta_{TVaR}^{0.5}}$ | $\boldsymbol{\beta_{TVaR}^{0.75}}$ | $\boldsymbol{\beta_{TVaR}^{0.90}}$ | **p-valor$_{0.90}$** |
|---|---|---|---|---|---|---|
| Constante | -8.120,85 | -5.936,59<br>(165,93) | -6.753,75<br>(237,93) | -7.050,46<br>(394,48) | -6.194,13<br>(1117,77) | < 0.0001 |
| lnKm | 1.062,54 | 831,87<br>(16,49) | 953,45<br>(20,80) | 1.102,14<br>(37,44) | 1.091,75<br>(77,07) | < 0.0001 |
| Porc_Vurba | -21,31 | -18,41<br>(0,56) | -22,59<br>(0,95) | -34,79<br>(1,11) | -54,68<br>(4,00) | < 0.0001 |
| Porc_Noctur | 5,35 | 4,78<br>(1,57) | 4,90<br>(2,84) | 16,75<br>(4,20) | 26,90<br>(9,31) | < 0.01 |
| Edad | 1,37 | 0,70<br>(2,82) | 7,13<br>(5,00) | 7,99<br>(7,47) | 52,29<br>(21,56) | 0.02 |
| Sexo | 329,98 | 285,04<br>(19,05) | 369,19<br>(30,22) | 526,06<br>(47,71) | 907,82<br>(116,33) | < 0.0001 |
| Bondad de ajuste | - | 0,4698 | 0,5390 | 0,6682 | 0,8011 | |

$\beta_{TVar}^{\tau}$ representa los coeficientes del modelo para el *Tvar* al nivel τ, así como los valores de la Bondad de Ajuste. $\beta_{lm}$ representa los coeficientes del modelo de Regresión Lineal Múltiple.

Respecto a la circulación por vía urbana, vemos como el coeficiente asociado toma valores negativos en todos los modelos. Eso quiere decir que cuanto más se circula por vía urbana, el porcentaje de kilómetros circulados por encima de los límites de velocidad disminuye, lo cual tiene sentido dado que es más difícil sobreasar los límites circulando por la ciudad. Para $\tau = 0.9$ el coeficiente resulta significativo y podemos decir que si el porcentaje de circulación por vía urbana aumenta en un punto porcentual, el *TVaR* al 90% disminuye en 54.68 kilómetros.

Por lo que respecta a la conducción nocturna, el coeficiente es positivo en todos los modelos, por lo que contribuye a aumentar el riesgo de circular por encima del límite de velocidad. En concreto, para $\tau = 0.9$ el coeficiente resulta significativo y en concreto, cuando el porcentaje de circulación nocturna aumenta en un punto porcentual el *TVaR* al 90% aumentara en 26.90 kilómetros.

Por últimos comentamos los resultados para las variables edad y sexo. La edad tiene parámetro positivo, por lo que a más edad aumenta el *TVaR*. El parámetro es significativo para $\tau$ = 0.9, y en ese caso, un incremento de un año en la edad provoca un incremento del *TVaR* al 90% de 52.29 kilómetros. Ser hombre también aumenta el valor del *TVaR* para cualquier $\tau$. Para $\tau$ = 0.9 el coeficiente es significativo y nos indica que el *TVaR* al 90% de los hombres es 907,82 kilómetros mayor que el *TVaR* al 90% de las mujeres.

En la Figura 2 se muestra la evolución de los efectos de las variables explicativas en función del nivel $\tau$ del *TVaR*. Vemos como a medida que aumenta $\tau$ la constante del modelo toma valores cada vez más negativos acercándose al efecto de la constante en el modelo lineal, pero sin llegar a encontrarse dentro de banda de confianza. Al pasar de $\tau = 0.75$ a $\tau = 0.9$ el valor de la constante aumenta.

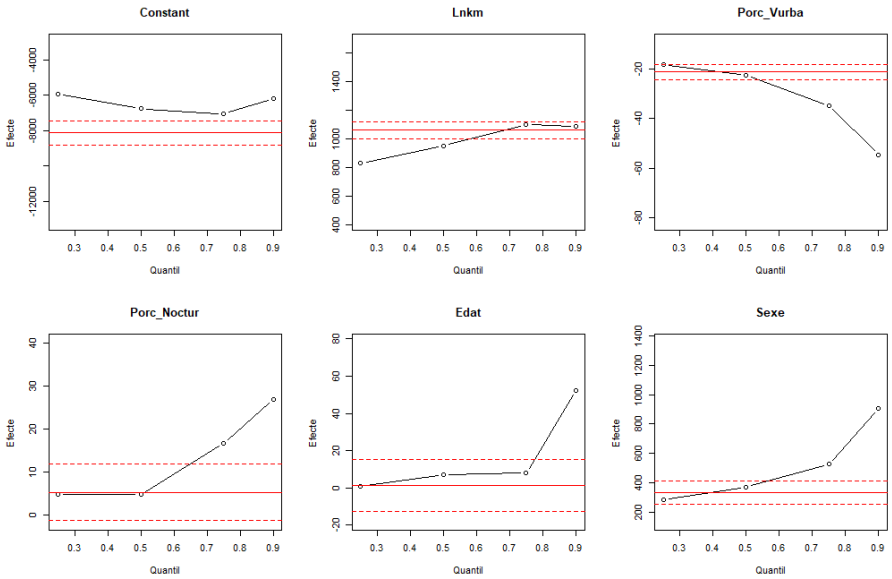


*Figura 2*. Gráficos de la Evolución de los Efectos de las Variables a Medida que Aumenta el Cuantil Utilizando los Modelos que Predicen el *TVaR*.

Respecto a la variable logaritmo del kilometraje total, vemos como su efecto va aumentando a medida que aumenta $\tau$, llegando a coincidir con el efecto del modelo lineal para valores de $\tau$ entre el 0.75 y 0.9. Tal vez este efecto

queda contrarrestado con el de la constante, de modo que el efecto conjunto podría equivaler al del modelo clásico.

Por lo que respecta a la circulación por vía urbana, su efecto coincide con el del modelo lineal para valores de $\tau$ como máximo iguales a 0.5. A partir de ese momento, al incrementar el percentil el efecto toma valores cada vez más negativos. En relación con la circulación nocturna, se observa cómo también se da la misma coincidencia con el modelo lineal para valores de $\tau$ inferiores a 0.5, luego el efecto aumenta considerablemente.

Por lo que respecta a la edad, existe una gran coincidencia de los efectos con los correspondientes al modelo lineal, dado que se encuentran entre las bandas de confianza para valores de $\tau$ inferiores a 0.75. A partir de ese valor, el efecto aumenta pronunciadamente. Finalmente, la variable sexo tiene un efecto coincidente con el del modelo lineal para valores de $\tau$ inferiores a 0.5, y luego aumenta también de manera muy pronunciada.

Finalmente, la tabla 4 también muestra los valores de la de la bondad de ajuste. En este caso observamos que para valores pequeños de $\tau$ la bondad de ajuste baja pero que, a medida que aumenta el nivel de $\tau$, también lo hace el valor de la bondad de ajuste hasta llegar a ser de 0.8011, lo cual indica un buen ajuste.

## 5. Conclusiones

En este trabajo se ha analizado una manera de predecir el valor en riesgo y el valor en riesgo de la cola. Además se ha mostrado que la utilización de la regresión cuantílica no proporciona los mismos resultados que si se hubiera utilizado un modelo de regresión lineal clásico para predecir el valor esperado. En concreto, en este trabajo hemos mostrado como la regresión cuantílica puede utilizarse para estimar el valor en riesgo y la aproximación al valor en riesgo en la cola para cada asegurado en función de sus características.

La aplicación resulta interesante porque esta metodología permite medir el riesgo que presenta un determinado asegurado no tanto respecto al conjunto total de la muestra, sino respecto a los conductores que son similares a él. En particular en nuestro ejemplo práctico, conducir una determinada cantidad de kilómetros por encima de los límites de velocidad resulta ser más o menos arriesgado en función del resto de características del individuo y de su patrón

de conducción, siendo especialmente importante el total de kilómetros que se recorran a lo largo del año.

En definitiva, la aplicación práctica de este trabajo en el caso de los modelos del TVaR permite establecer cuál es el promedio de kilómetros recorridos con exceso de velocidad por ejemplo para el nivel del 90%, a partir de ciertas características medidas en las variables independientes. De esta forma, si un conductor se acerca a ese valor, o incluso si lo sobrepasa, se podría identificar que su comportamiento es anómalo y, llegado el caso de tener una tarificación basada en el uso, se puede llegar a establecer una penalización en el precio.

El seguro del automóvil basado en el uso, aparte de tener en cuenta el total de kilómetros recorridos, deberá adaptar la tarifa pagada al patrón de conducción y, en el caso del exceso de velocidad que se modeliza en este artículo, parece natural que premie a los conductores cuyo perfil corresponda a niveles de riesgo bajos.

## Referencias

Acerbi, C., y Szekely, B. (2014). Back-testing expected shortfall. *Risk*, 27(11), 76-81.

Ayuso, M., Guillen, M., y Pérez-Marín, A. M. (2016a). Using GPS data to analyse the distance travelled to the first accident at fault in pay-as-you-drive insurance. *Transportation Research Part C: Emerging Technologies*, 68, 160-167.

Ayuso, M., Guillen, M., y Pérez-Marín, A. M. (2016b).Telematics and gender discrimination: some usage-based evidence on whether men's risk of accidents differs from women's. *Risks*, 4(2), 1-10.

Boucher, J-P., Côté, S., y Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks*, 5(4), 54. https://doi.org/10.3390/risks5040054

Fissler, T., y Ziegel, J. F. (2016). Higher order elicitability and Osband's principle. *The Annals of Statistics*, 44(4), 1680-1707.

Guillen, M., Nielsen, J. P., Ayuso, M., y Pérez-Marín, A. M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk Analysis*, 39(3), 662-672.

Hardy, M. R. (2006). *An introduction to risk measures for actuarial applications*. SOA Syllabus Study Note.

Koenker, R., y Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1), 33-50.

Koenker, R., y Machado, A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, 94(448), 1296-1310.

Koenker, R., Portnoy, S., Ng, P. T., Zeileis, A., Grosjean, P., y Ripley, B. D. (2018). *Package 'quantreg'*. Cran R-project.

Kudryavtsev, A. A. (2009). Using quantile regression for rate-making. *Insurance: Mathematics and Economics*, 45(2), 296-304.

Pitarque, A. (2019). *Regresió quantílica en la gestió de riscos*. Trabajo de final de máster Universitat Politècnica de Catalunya y Universitat de Barcelona.

Pitt, D. G. W. (2006). Regression quantile analysis of claim termination rates for income protection insurance. *Annals of Actuarial Science*, 1(2), 345-357.