UNIVERSITAT DE BARCELONA

Essays on Estimation, Prediction and Evaluation of Insurance Risk

Albert Pitarque Méndez

RSITAT DE
ELONA

2022

PhD in Business | Albert Pitarque Méndez

UNIVERSITAT DE
BARCELONA

PhD in Business

Essays on Estimation, Prediction and Evaluation of Insurance Risk

Albert Pitarque Méndez

LIBERTAS PERFVNDET
OMNIA LVCE

UNIVE
BARC

PhD in Business

Thesis title:

Essays on Estimation, Prediction and Evaluation of Insurance Risk

PhD student:

Albert Pitarque Méndez

Advisor:

Montserrat Guillén Estany

Date:
September 2022

UNIVERSITAT DE BARCELONA

Acknowledgment

I want to express my gratitude to my supervisor, Dr. Montserrat Guillén for sharing with me all her experience and guiding me with a lot of patience through this journey. I also want to give thanks to all members of the Riskcenter research group who always found time to help me when I needed. Finally, I want to give a special mention to my family who supported me unconditionally.

Contents

List of Figures

List of Tables

Introduction

Risk is defined as the possibility of an adverse or harmful event occurring. In a business environment, risk can be understood as the exposure to having an unexpectedly large economic loss. Despite the importance of risk analysis, it is only in recent decades that the methodology used to carry out this type of analysis has begun to evolve to take advantage the huge source of statistical data now available.

A company's profit and losses have a distribution similar to the Normal distribution, where losses are around zero. This thesis focuses on the risk of a negative effect and, as such, considers losses as the positive values of the distribution.

There are numerous methodologies that study risk, many of which have only been developed in recent years. This thesis uses the methodology of quantile regression. This methodology is based on the linear regression that looks for a relationship between one or several explanatory variables and the mean of a single response variable. In the case of quantile regression the relationship of the variables is developed in relation to the quantile of the distribution or, in other words, the effect of the variables on the extreme values of the distribution of the variable of interest.

Quantile regression was developed in 1979 by Koenker, R. and Basset, G. (1979), and throughout the decade that followed, the majority of articles have a theoretical approach. Several papers from the 1990s apply quantile regression in practical cases. Although these are fairly simple models, these study the effects of exogenous variables on various quantiles. Poterba, J.M. and Rueben, K. (1994) compare wage differences between employees and their private sector counterparts between 1979 and 1992. They fit different regressions depending on sex and use the explanatory variables of level of education, experience and year. Buchinsky, M. (1998) studies the salary differences between women over the years 1968 to 1990, taking into account the size of the family, income, race, education, experience, and number and age of children. Eide, E. and Showalter, M.H. (1998) fit quantile regressions to study the relationship between the quality of schools and the results of their students. They take characteristics of the schools as diverse explanatory variables, such as the length of the school year, the ratio between teachers and students and the number of enrollments.

With the onset of the 21st century, studies begin to appear that apply quantile regression together with other methodologies thus obtaining more powerful results,

with most continuing to focus on studying salary difference. Martins, P.S. and Pereira, P.T. (2004) analyse the wage gap based on education in sixteen countries. They apply a very simple model consisting only of the variables education and experience. Melly, B. (2005) also focuses on analyzing the wage gap in the United States between 1973 and 1989, using a more elaborate model including variables on education, experience, race, region, industry, and interactions between variables. Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006) approach the quantile regression as a minimization of a function of loss of the mean of the weighted squared errors and study the salary difference in the 1980s, 1990s, and 2000s in the United States.

During these years, studies also appear with slightly different themes. Koenker, R and Geling, O. (2001) make an analysis of survival using quantile regression. They use as a basis a study that monitors the mortality of fruit flies and conclude that quantile regression refines several of the conclusions reached in that study. Fattouh, B., Scaramozzino, P., and Harris, L. (2005) fit quantile regressions to compare the debt-to-equity ratios of South Korean firms between 1992 and 2001, and discuss how South Korean firms choose their capital structures by comparing firms in different quantiles in terms of debt-equity ratio. They use variables such as firm size, liquidation values, financial assets, and year. Coad, A., and Rao, R. (2006)use quantile regression to study the market values of firms as a function of the innovation they apply. They use a database from the National Bureau of Economic Research (NBER) and a Compustat database.

The variety of applications where quantile regression is used is expanding. For example, Kaza, N. (2010) studies the effect of different variables, such as the size of a house, its location or its age to consider its energy consumption. He applies quantile regression for all deciles of the distribution of the variable of interest. Behr, A. (2010) presents quantile regression as a way of approximating Farrell's technical efficiency scores, which is a measure of productivity efficiency. In this case, he focuses on looking at productivity for data for German banks grouped into commercial banks, saving banks and cooperative banks. Hung, W. T., Shang, J. K., and Wang, F. C. (2010) analyse, through quantile regression, the determinants of the price of a hotel room in Taiwan. They also apply ordinary least squares regression to compare the results.

More recent papers are much more varied and go beyond the field of economics, applying quantile regression together with more complex methodologies. Liao, W.C. and Wang X. (2012) use quantile regression to study house prices in China. To perform the study they also use models of spatial econometrics. Tareghian, R. and Rasmussen, P.F. (2013) use climate models to study rainfall in various seasons of the Canadian climate, using different climatic variables such as wind, pressure

or humidity. To select which variables are useful for each weather station, they use a Bayesian method adapted to quantile regression. Daniel-Spiegel, E., et al. (2013) attempt to improve the analysis of fetal growth charts using quantile regression. They use a database of women who are between 12 and 42 weeks pregnant and fit univariate quantile regressions for head diameter, femur length, and circumference of the baby's abdomen.

Briollais, L. and Durrieu, G. (2014) study the use of a quantile regression extension focusing on copulae to analyse the genomic loci that contribute to the levels of mRNA or protein expression. Marrocu, E., Paci, R., and Zara, A. (2015) fit linear regression models and quantile regression models to study the main determinants of the expenditures of tourists who visited Sardinia between April and October 2012. These are based on data obtained through a survey that asks questions such as the size of the group, the length of stay or mode of transportation. They find significant effects in the variables, especially for high quantiles. Lin, B. and Xu, B. (2018) study which factors affect emissions of CO2 using the data of the provinces of China. They fit regressions for 6 different quantiles dividing between groups with high, medium and low emission. They take into account variables such as financial capacity, urbanization and economic growth among others. Niemierko, R., Toppel, J. and Tränkler, T. (2019) use data from German households to study the cost of heating and propose the use of quantile regression through a D-vine copula to study the factors that affect different levels of energy efficiency and study how the effects change between the homes that save the most energy and those that waste the most.

In the field of insurance, the application of quantile regression is still very limited. Besides the scientific papers mentioned in this thesis, and which were pioneers at the time, we found very few that apply quantile regression in the field of insurance. Kudryavtsev, A. A. (2009) proposed quantile regression in insurance pricing and showed a comparison between traditional methods. Heras, A., Moreno, I. and Vilar-Zanón, J. L. (2018) also discussed this methodology, proposing the analysis of the total cost of claims and showing that the QPP (Quantile Premium Principle) improves the EVPP (Expected Value Premium principle) for the premium security surcharge. The same methodology was extended by Baione, F. and Biancalana, D. (2021). In recent years, the application of quantile regression has been generalized and we can find applications from insurance against earthquakes (Pai, J., Li, Y., Yang, A. and Li, C., 2022), to the analysis of insurance fraud (Li, H., Song, Q. and Su, J., 2021) to health insurance (Nortey, E. N., Pometsey, R., Asiedu, L., Iddi, S. and Mettle, F. O., 2021), as well as insurance intended to cover cybersecurity (Eling, M., Jung, K. and Shim, J., 2022) and agricultural insurance (Yu, J., Sumner, D. A. and Lee, H., 2021).

During my final doctorate year I worked in the automobile department of an

insurance company, dealing with data closely related to the subject of the doctorate. Although for reasons of confidentiality I cannot show any type of data or result, I was able to gain experience in the sector and compare the theoretical work presented in this thesis with what is done on a practical level in insurance companies.

The first and most important consideration is that each company has its own policies when it comes to pricing an insurance of any kind. There are companies that prefer to focus on capturing clients with safe profiles, who have little risk of accidents and build a client portfolio in which there are few claims to pay, even if returns are lower. In contrast, there are other companies that try to capture the maximum number of clients by taking a higher risk, but covering it with a larger client base. No one strategy is better than the other and as in other fields within private companies, the methodologies applied or the way they are applied will depend on the risk aversion that each company has.

Although we know that insurance companies work with large amounts of data, we are not aware of the millions of observationsused when applying pricing methodologies and grouping the insured. This has its positive and its negative side. On the one hand, very precise models can be adjusted which closely reflect the personal reality of each insurance contract. On the other hand, many processes take hours to run and powerful software is required to work with such a massive amount of data. It is, therefore, very important that the entire methodology that is developed for actuarial analysis is very agile and optimized to the maximum so that the computation times are as short as possible.

The method for reflecting how an insured in the automobile industry drives in the policy premium is an issue that insurance companies in Spain are reluctant to pursue because people are very protective of giving their personal data. So, although this type of data is being used, the sample size is quite small and this makes it difficult to draw firm conclusions. There are other countries that do work with this type of data in more detail and affirm that the variables obtained from this method provide very important information when pricing car insurance.

Although these data is very useful, it would be a mistake to base the whole pricing procedure on a model that only contemplates how the driver drives and does not consider other variables that are included in current models. As mentioned earlier databases of this type contain a huge amount of data, not only the number of observations but also the number of variables available in the internal databases. The driving skills of the driver, therefore, should be treated as information complementary to current models.

A device can be installed in a vehicle which, apart from providing data, also has an automatic emergency warning system in the event of an accident. This device is linked to the driver's mobile phone. In the event of a collision, a call to the driver

is made to check the magnitude of the accident. In the case that the driver does not answer the call, a call is made to the hospital to report the accident. The installation of this device is very complicated mechanically and most people are reluctant to install it in their vehicles. Drivers, through an app installed on their own mobile phone, receive three different scores that evaluate different factors of their driving and a general score and are able to consult them at any time.

There are many companies that are dedicated to marketing this product. The problem is that there are numerous simultaneous data sources and insurance companies do not have contracts with all of them. In addition, there are many vehicle brands that, in their latest models, already have such a device installed. There exists, therefore, a huge number of sources of information with various configurations.

From my own experience, I have noticed several differences between the models that are applied in the thesis and those applied in insurance companies. As already mentioned the devices used by companies contain many variables, although the one used in this thesis contains a very reduced number. However, there are also other differences worth highlighting in this study. The models applied in this thesis lead me to conclude that sex is an important variable in determining the risk of accident. Men are more prone to accidents than women and, combined with the age of the driver, this variable plays a significant role. It is suspected that gender actually determines very different driving habits and these may be responsible for a greater or lesser accident rate. In Spain, as a result of the transposition of the corresponding European directive, this variable cannot be applied by law since it produces discrimination between the sexes.

Another of the explanatory variables that have been used in the models proposed throughout the thesis is the percentage of urban driving of the driver. The importance attached to this variable is much higher in the company than in the model that we apply in the thesis since here we only use it as an explanatory variable and depending on the quantile in which we are applying the regression this variable does not turn out to have a significant effect. In the information collected by the device, this variable not only comes in the form of a percentage, but the client is also given a score.

Finally, the driver also receives a score on acceleration and braking. In my opinion, this variable is just as relevant as speeding above the speed limit when studying, not only the risk of having a car accident, but also the seriousness of the accident. In this thesis I only work with the variable of speed since information of acceleration and braking was not available. However, this is a variable that should be taken into account for future research.

In all the chapters of the thesis a series of models that share notation are used. Y_i corresponds to the value of the response variable of the model for the observation

$i = 1, \ldots, n$ in a set of n data. X_{ij} corresponds to the matrix containing the value of the explanatory variable $j = 1, \ldots, k$ of observation i. The objective of the models is to find how values of the variables of matrix X affect values of the response variable Y. This effect is represented by β_j. The model also contains an error term for the effect that is not represented by the explanatory variables and is defined as ϵ_i.

The objective of the thesis is to look more closely into the methodology of quantile regression and its generalizations in the analysis of insurance data. For this, the study proposes a series of solutions to the problems mentioned above (assessment of the risk of loss and analysis of the influence of certain risk factors and, ultimately, pricing) explained in three sections. The first section focuses on the development of explanatory models for the afore mentioned risk measures. This section is made up of Chapters 2 and 3. The second section focuses on developing a score for drivers depending on their car accident risk. This is covered by Chapter 4. The third section, Chapter 5, focuses on the use of distributions to predict the risk of bad driving habits.

The following describes each chapter in more detail. Chapter 2 is titled: Quantile Regression as a Starting Point in Predictive Risk Models. I propose a methodology to fit a regression inspired by the conditional tail expectation (CTE) based on quantile regression. In this chapter, I explain in detail what quantile regression is and how it adjusts for the effects of variables. I adjust the VaR and the CTE for different quantiles and study the evolution of the estimated parameters. I show that the developed methodology is a good approximation to estimate risk measures more complex than VaR.

Chapter 3 is titled: An algorithm to fit conditional tail expectation regression models for vehicle excess speed in driving data. In this chapter I present a model that enables regression models to be adjusted for the tail expectation that are a natural generalization of quantile regression models. Here I consider a linear relationship between covariates and show that quantile regression identifies risky drivers by modelling quantiles of distance driven yearly above the speed limit. I observe that the linear relationship between variables causes some adjusting problems; theseare studied in the over the chapters that follow the thesis.

Chapter 4 is titled: Joint Generalized Quantile and Conditional Tail Expectation Regression for Insurance Risk Analysis. Here I develop a method that adjusts the VaR and the CTE in a two-part procedure. When fitting models that define a linear relationship between the parameters, an error in the prediction of the risk measures is found. To correct this, I introduce a link function that allows the non-linear relationship between variables to be studied. The resulting model type which I call, risk regression, provides a better fit of the CTE than the model methodology above and has the potential to be developed for other risk measures.

Chapter 5 is titled: "Interpolation of Quantile Regression to Estimate Driver's Risk of Traffic Accident Based on Excess Speed". A risk score of having a car accident can be assigned adjusting a quantile regression for each quantile and comparing in which value τ it approximates the most to the observed value of the variable of interest. Fitting a large number of regressions is not feasible from the point of view of computational time and becomes a problem when dealing with models with lots of data and variables, which is normal nowadays. In this chapter, I develop a methodology where the values of the quantile regressions that I did not adjust can be interpolated and I compute the score by fitting a small number of regressions, obtaining acceptable predictions.

Chapter 6 is titled: "A Sarmanov Distribution with Beta Marginals: An application to motor insurance pricing." In this chapter, I propose the use of a bivariate model using a Sarmanov distribution to predict risk. The Sarmanov distribution uses marginal distribution regressions for which I select Beta regressions. I adjust three models with different variables and study how they improve the results. I establish three customer profiles and study the relationship of the variables depending on the customer profile.

Chapters included in this doctoral thesis have been published and can be found in:

1. Pitarque, A, Pérez-Marín, A.M., Guillen, M. (2019) "Regresión cuantílica como punto de partida en los modelos predictivos para el riesgo." *Anales del Instituto de Actuarios Españoles*, 25, 77-117.

2. Guillen, M., Bermúdez, LL., Pitarque, A. (2021) "Joint generalized quantile and conditional tail expectation regression for insurance risk analysis."*Insurance: Mathematics and Economics*, 99, 1-8.

3. Pitarque, A., Guillen, M. (2022) "Interpolation of Quantile Regression to Estimate Driver's Risk of Traffic Accident Based on Excess Speed." *Risks*, 10, 1, 19.

4. Bolancé, C., Guillen, M. Pitarque, A. (2020) "A Sarmanov Distribution with Beta Marginals: An Application to Motor Insurance Pricing" *Mathematics*, 8, 11

I also contributed on the following publications related with the thesis topic.

1. Uribe, J.M. and Guillen, M. (2020) "Quantile Regression for Cross-Sectional and Time Series Data: Applications in Energy Markets Using R", Springer Nature.

2. Pitarque, A., Guillen, M. "An algorithm to fit conditional tail expectation regression models for vehicle excess speed in driving data" (2020) *CARMA 2020: 3rd International Conference on Advanced Research Methods and Analytics*.

Finally, I participated in the congress "2020 Actuarial Research Conference" with the presentation: Pitarque, A., Guillen, M. (2020) "Joint Generalized Quantile and Conditional Tail Expectation Regression for Insurance Risk Analysis" 55th Actuarial Research Conference (ARC 2020), Nebraska, August 10-12.

Quantile Regression as a Starting Point in Predictive Risk Models

1 Introduction

This paper looks at models oriented towards predicting value at risk or, in other words, the percentile or quantile and other risk measures of a response variable depending on multiple explanatory variables. Unlike other regression models where the objective is to predict the mean value, in the case of quantile regression, we want to know the estimated risk conditioned for given values of explanatory factors within a tolerance level, for example, 95% or 99%.

Historically, one of the first authors to discuss the regression concept was Roger J. Boscovich in the 18th century. This physicist and mathematician limited the mean of the residuals of the regression to 0 and proposed minimizing the sum of absolute values in order to estimate the effects of one variable on another. This type of regression was named LAD regression (Least Absolute Deviations) and later was generalized by Pierre-Simon Laplace for multiple variables. This antecedent, which is prior to Carl F. Gauss and his Ordinary Least Square regression, aimed to adjust the median of a response variable which is equal to the quantile at 50%. At the end of the 19th century, Francis Y. Edgeworth proposed an algorithm to adjust regressions so that the sum of the absolute value of the residuals was minimum. However, this method and other posterior generalizations were not consolidated because more computational power was required to obtain better optimization algorithms. Until the middle of the 20th century, studies that used quantile regression focused on estimating a regression for the median. Among the first authors to study regressions for different quantiles were Koenker and Bassett (1978), whose propositions have continued to evolve.

The τ-quantile value of a continuous random variable, Y, is that value c_τ at which the probability that the response variable is equal to or no greater than τ, namely, $P(Y \leq c_\tau) = \tau$. Where c_τ is the τ^{th} quantile or percentile τ, in insurance and finance, this is called value at risk at level τ and denoted as $VaR_\tau(Y)$.

Similarly, it is defined as the tail value at risk at level τ for variable Y, $TVaR_\tau$,

is defined as the expected value of the conditional tail. In other words, the expected value of variable Y that is greater than c_τ. This risk measure is also called Expected Shortfall (ES_τ) or Conditional Value at Risk ($CVaR_\tau$). This corresponds to the mean of the values that are greater than VaR_τ and can be expressed by:

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

where $f(y)$ is the density function of the probability of the random variable Y. More details can be found in Hardy (2006) who introduces risk measures with actuarial applications.

The issue that we address in this paper is how to estimate a regression model for both risk measures. Until now, quantile regression that covers the first of the two risk measures has been resolved but as yet not the second.

In other papers where quantile regression has been applied, only a few have been oriented towards motor insurance. Kudryavtsev (2009) uses quantile regression to price a robbery insurance analysing loss severity. Pitt (2006) focuses on income protection insurance. Our intention is to llok more closely at the use and development of methodology and its value in the field of actuarial science.

In Section 2 of this paper, we introduce quantile regression and explain how goodness of fit is determined in this type of model. Next, in Section 3, we describe the methodology that we propose to use approximate TVaR. In Section 4, we present the data used in this paper and the results obtained. Lastly, in Section 5, we present the main conclusions.

1.1 Introduction to quantile regression

To understand quantile regression, we must first appreciate that a linear regression model is an approximation that fixes a linear relation between the response variable (or dependent variable) and one or more explanatory variables (or independent variables) with the following expression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i, \tag{2}$$

where Y_i corresponds with the dependent variable for the i-th case of the sample($i = 1, \cdots, n$) and X_{ji}, the corresponding observations of the k explanatory variables, being $j = 1, \cdots, k$. A disturbance term ϵ_i is considered and it captures all deviations from the mean. In this case, given that the disturbance term is centred at zero, this can be expressed by the following equation:

$$E(Y_i|X_{1i}, \cdots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}. \tag{3}$$

Model coefficients are estimated using ordinary least square method (OLS) so:

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

where $S(\beta) = ||Y - X\beta||^2$ represents a distance between the vector of the dependent variable and the linear predictors calculated as the linear combination (3) for each component.With the Euclidean norm, $S(\beta)$ correspond to the sum of the squared residuals. In quantile regression, we want to find a relation between the quantile of the dependent variable, given values of the explanatory variables, so that:

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

In this way, it can be shown (see Koenker and Bassett, 1978) that:

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

While the coefficient estimators of the OLS regression that follow a Student's t-distribution are known, for quantile regression coefficients the exact distribution of its estimators is unknown. However, under certain conditions, studies have shown that $\sqrt{n}\left(\hat{\beta^\tau} - \beta^\tau\right)$ tends to a Normal distribution. It is understood that the objective function in (6) corresponds to the sum of n components, where each of them can be expressed as:

$$\begin{aligned} \rho_\tau(Y_i - X_i'\beta) = \tau(Y_i - X_i'\beta) I_{\{Y_i \geq X_i'\beta\}} + \\ (1 - \tau(Y_i - X_i'\beta) I_{\{Y_i < X_i'\beta\}} = (Y_i - X_i'\beta)(\tau - I_{\{Y_i < X_i'\beta\}}) \end{aligned} \tag{7}$$

being $I_{\{\cdot\}}$ an indicator function that equals 1 if the condition of the subindex is accomplished and 0 otherwise.

Koenker and Machado (1999) proposed an expression to measure the goodness of fit in quantile regression based on a comparison between the objective functions of the estimated quantile regression model and a restricted model that only includes the independent term. More concretely, being

$$\hat{V}(\tau) = \sum_{y=1}^{n} \rho_\tau(Y_i - X_i'\hat{\beta^\tau}) \tag{8}$$

the value of the objective function of the complete model that includes all param-

eters and

$$\tilde{V}(\tau) = \sum_{y=1}^{n} \rho_\tau(Y_i - X_i'\beta^\tau) \tag{9}$$

the value of the objective function of the model that only includes the independent term. The goodness of fit measure, therefore, proposed by Koenker and Machado(1999)is

$$R^1(\tau) = 1 - \hat{V}(\tau)/\tilde{V}(\tau) \tag{10}$$

which is analogous to the R^2 of the linear regression model.

2 Proposed methodology to approximate regression to TVaR

Although the risk measure used in quantile regression is the value at risk at level τ, one of the main problems is that since only one quantile value is taken, it is a risk measure that does not consider losses greater than itself. However, in Section 1 showed that $TVaR_\tau$ (expected value of the values that are bigger than VaR_τ) can do this.

Our contribution takes as a starting point the papers presented by Koenker regarding quantile regression and, having referred to the recent results offered by Fissler and Ziegel (2016) and Acerbi and Székely (2014) that quantile regression can be applied to other risk measures, we establish a new loss function similar to the one that is optimized for the quantile that will allow a parametric quantile regression to be adjusted close to $TVaR_\tau$. This regression is an extension of the quantile regression for quantile τ, which is why the way to calculate the effects of variables is similar to the estimator of quantile regression.

In order to calculate the values of the coefficients of quantile regression, it is necessary to minimize the objective function (6) that ponders the absolute value of deviations according to the pre fixed τ level. Based on the definition of TVaR in (1), this can be interpreted as an expected value, and to calculate the expected value of a continuous random variable with density function f(x), the classic expression $E(X) = \int_{-\infty}^{\infty} xf(x)dx$ can be used.

To calculate the expected value of the entire random variable, the integral limits include all the domain i.e., the limits go from $-\infty$ to $+\infty$ for non-bounded random variables. However, our interest lies in the expected value of the values greater than the value $c\tau$. An alternate expression to (1) is:

$$TVaR_\tau(Y) = \frac{1}{1-\tau}\int_\tau^1 VaR_u(Y)du. \tag{11}$$

Since it is not possible to directly obtain an objective function that allows the analogous value to (6) for $TVaR_\tau$ to be obtained, we consider the objective function $\rho\tau$ in (7), used to estimate the coefficients for quantile regression, as a density and calculate the expected value in the tail of the distribution i.e., integrating from τ to 1. Solving this integral we obtain the following expression (see Pitarque, 2019):

$$\begin{aligned}\int_\tau^1 \tau\rho_\tau(Y - X\beta)d\tau &= (Y - X\beta)\int_\tau^1 \tau(\tau - I_{(Y-X\beta)<0})d\tau \\ &= (Y - X\beta)\left(\frac{1}{3} - \frac{I_{(Y-X\beta)<0}}{2} - \frac{\tau^3}{3} + \frac{\tau^2 I_{(Y-X\beta)<0}}{2}\right).\end{aligned} \tag{12}$$

Finally, the coefficients of the model inspired in $TVaR$ are obtained when minimizing the following objective function:

$$\beta_\tau(\hat{TVaR}) = \underset{\beta}{\operatorname{argmin}} \sum_{y=1}^n \left[(Y_i - X_i'\beta)\left(\tfrac{1}{3} - \tfrac{I_{(Y_i - X_i'\beta)<0}}{2} - \tfrac{\tau^3}{3} + \tfrac{\tau^2 I_{(Y_i - X_i'\beta)<0}}{2}\right)\right] \tag{13}$$

The objective function and its minimization has been programmed in R and its standard errors obtained using a resample method. We also adapt a calculation of the goodness of fit measure analogous to the measure proposed by Koenker and Machado (1999) in quantile regression.

3 Data and results

Data in this paper are used to model the distance driven above the speed limit by a sample of drivers insured by a Spanish entity. The use of a "pay as you drive" insurance programme is being offered in more and more countries and, despite it being a novel idea, the possibility of taking out this type of policy, where the price depends on driving patterns, is predicted to be fairly common in the coming years. This insurance programme, requires information on the driving patterns obtained through a device fitted to the vehicle which records telematics basic data, such as speed or number of kilometres. The sample used in this study contains information on 7.691 young drivers (aged between 18 and 32 years old) that took out this type of policy during 2010. The reason that older people are not included in this database is that the insurance company only offered this insurance policy to younger people.

Information corresponding to each driver includes a total of six variables as shown in Table 1. Various authors have already used this database to carry out studies. For example, Boucher et al. (2017) studied the transformation of risk factors. Ayuso et al. (2016a) and (2016b) analysed the effects of distance driven up until the first crash car occurs also considering the difference in intensity and driving patterns depending on the genre. Finally, Guillen et al. (2018) studied how to design premiums that consider a high presence of zeros in the number of declared crashes by the insured drivers and how to use telematics data.

Table 1: Variables of the database

Variable	**Description**
Toler_km	Total of kilometres driven above the posted speed limit during 2010.
lnKm	Logarithm of the total of kilometres driven during 2010.
Porc_urba	Percentage of kilometres driven on urban roads.
Porc_noct	Percentage of kilometres driven at night.
Age	Age of driver at 1 January, 2010.
Gender	1 = Male, 0 = Female.

Table 2: Descriptive analysis

Variable	**Mean**	**Std.**	**Median**	**Minimum**	**Maximum**	**Asymmetry**	**Kurtosis**
Toler_km	1400.12	2008.67	689.42	0.00	23500.19	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
Age	24.77	2.82	24.61	18.11	31.56	0.10	2.23

Table 2 shows the descriptive analysis of the variables of the database. A 50,88% of the drivers are male which means that the database includes approximately the same number of women and men. The age range is between 18 and 32 years old, with the mean of age being 24 years old. The asymmetry coefficient is very low which indicates that the variable is symmetrical. In Table 2 the dependent variable Toler_km, which measures the total number of kilometres driven above the speed limit, shows that drivers usually drive a great number of kilometres above the speed limit each year but due to some extreme cases in the databae the mean is much higher than the median. Nevertheless, a lot of drivers have never gone above the posted speed limit. This fact is affirmed by the asymmetry coefficient, which is high and positive. Kurtosis is high because the majorith of drivers do not go above the posted speed limit or, if they do, it is not for a long period of time. However, the

total kilometres covered while speeding is high, at 23.500,19 kilometres . It is likely that these very high values are related to the driving zone, for example with very low speed limits and little congestion, or the age of the driver, since it is believed that the younger the driver, the higher the tendency to speed above the posted speed limit. Another possible reason may be the profession of the driver, since certain professions require more journeys by car. Nevertheless, these assumptions need to be tested.
The total kilometres driven are entered into the model on a logarithmic scale. On average, about 10.000 kilometres are traveled during the period in which the driving data is recorded (one year), and the median is quite similar. We observe that the maximum value corresponding to an annual distance is 57.000 kilometres. The coefficient of asymmetry is negative and quite high and suggests that there is some asymmetry to the left, indicating that there are more low values than high values. The kurtosis of this variable is also very high, which indicates that the distribution is sharper than the normal distribution and a large proportion of the observations take similar values. The most common driving area, is seen to be the use of interurban roads, with on average 26 % of the total mileage recorded. However, the maximum value is equal to 100, which indicates that there are drivers who only drive in urban areas, probably only using their vehicle to go from home to work in metropolitan areas. This variable also presents some asymmetry to the right and a high kurtosis coefficient. With regard to time of driving, drivers in the sample do not usually drive at night since the average percentage of kilometres travelled at night is 7 %. The maximum value obtained for night driving is 46.34 %, which could be for various reasons; either the driver works at night or is a person, possibly young, who uses the car for leisure. This variable also shows some asymmetry on the right, but not very high. The kurtosis is also positive, but less high than that observed in the other variables.

3.1 Quantile regression

In this section we present the results of the adjustment of quantile regression to the analysis of the number of kilometres driven above the speed limit. In Table 3 we present the values of the estimated coefficients for every quantile and for $\tau = 0{,}9$ we also include the corresponding p-values. To obtain these we use the function *rq* of the *quantreg* R package (Koenker *et al.*, 2018).

Table 3: Estimated coefficients for quantile regression and its standard errors between parenthesis. β_{rq}^{τ} represents the coefficients for τ-th percentile obtained using rq funcion. β_{lm} represents the linear model estimated coefficients.

Variable	β_{lm}	$\beta_{rq}^{0.25}$	$\beta_{rq}^{0.50}$	$\beta_{rq}^{0.75}$	$\beta_{rq}^{0.90}$	**p-value**
Constant	-8120.85	-2824.98 (148.44)	-4588.69 (273.60)	-6281.67 (470.82)	-6451.74 (1032.88)	<0.0001
lnKm	1062.54	359.81 (14.90)	603.55 (27.89)	894.77 (44.62)	1086.12 (90.46)	<0.0001
Porc_urba	-21.31	-2.95 (0.45)	-9.11 (0.86)	-21.63 (1.91)	-38.64 (3.32)	<0.0001
Porc_noct	5.35	3.18 (1.18)	3.49 (2.31)	4.07 (5.38)	19.69 (12.64)	0.12
Age	1.37	-2.77 (2.43)	-0.52 (4.70)	2.42 (9.48)	2.19 (20.47)	0.91
Gender	329.98	97.51 (14.07)	204.34 (27.82)	364.79 (58.18)	582.63 (141.87)	<0.0001
Goodness of fit	-	0.1937	0.1380	0.5114	0.6924	

In Figure 1 we present the evolution of the effects of the explanatory variables.

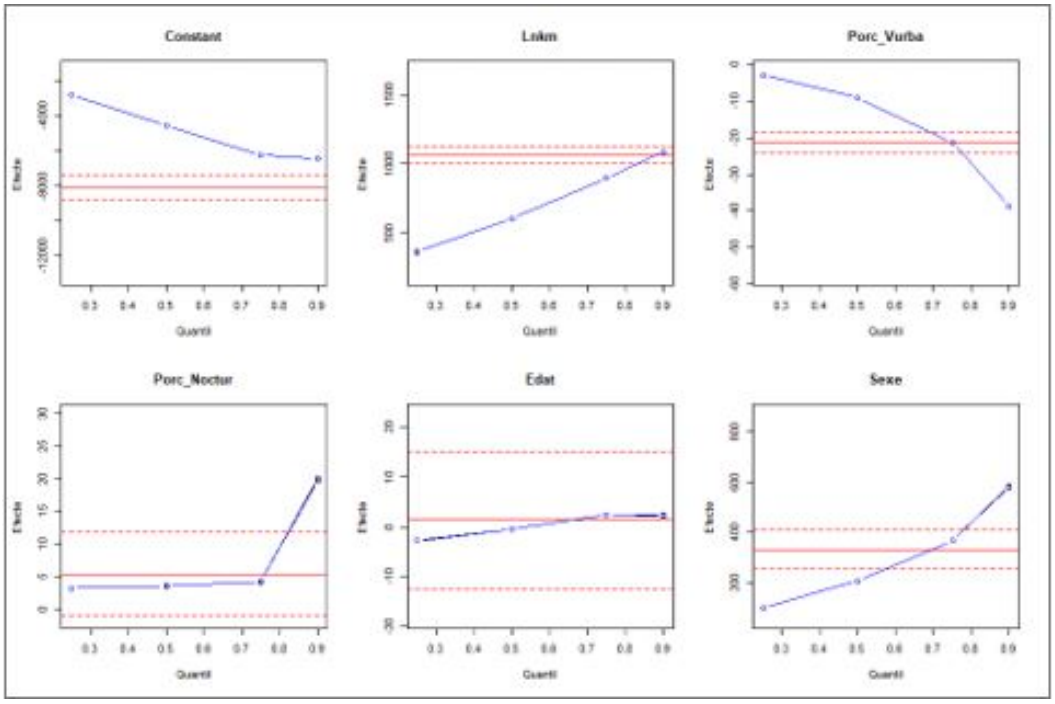

Figure 1: Graphs of the evolution of variable effects as τ quantile increases using quantile regression.

The graphs in Figure 1 show that, in relation to the constant term, we cannot consider it to be the same as the estimated effect using a linear model. This effect is negative for all values of τ and decreases constantly until quantile 0,75. From quantile 0,75 the effect keeps decreasing but less abruptly. For the variable that indicates the total number of kilometres driven the effect is increasing for all quantiles and more or less constant. From quantile 0,25 to 0,75 this effect can be considered different to the effect estimated using linear regression and for quantile 0,9 the effect is considered equal to that estimated by linear regression and significantly different from 0. For the variable that indicates the percentage of kilometres travelled by urban road, the effect decreases as the quantile increases. For the 0,25 quantile, this effect is practically zero and as the quantile increases it takes on increasingly negative values. For quantiles 0,25 and 0,5, the effect of this variable reduces the quantile of the response variable less than the effect estimated for the linear model. For the 0,75 quantile, the effect is the same as that of the linear model and for the 0,9 quantile the effect is significant and there are again differences with respect to the linear model. For the variable that indicates the percentage of kilometres driven at night, the effect is practically the same as the one estimated using the linear model, although it increases slowly as the quantile increases. From the 0.75 quantile, this growth becomes more pronounced, differentiating itself from the effect of the linear model. However, for the 0.9 quantile, night driving has no significant effect.

Regarding the variables that are related to the personal characteristics of the driver, we see that in terms of age, the effect practically does not vary for different levels. For the first two quantiles, this variable has a negative effect and for the 0,75 and 0,9 quantiles it has a positive effect. Even so, in none of the cases can this effect be considered to be different from that of the linear model. For the 0,9 quantile the effect of age is not significant. In relation to the gender variable, the behaviour is the inverse of the variable that represents urban driving; as the quantile increases, the effect increases to a greater extent. For the 0,25 and 0,5 quantiles the effect is less than that of the linear model and can be considered different from it. For the 0,75 quantile the effect can be considered the same and for the 0.9 quantile it can be considered different and greater again, being also significant. Finally, Table 3 also provides the values of the goodness of fit following the formula (12) propose by Koenker and Machado(1999). Most notably, the fit tends to improve as the τ level increases, with a slight decrease in the median.

3.2 TVaR regression

In this section we show the results for the adjustment of TVaR in the same database. Table 4 presents the results of the adjustment of the linear model as well as the model based on TVaR for different τ levels, specifically, $\tau = (0.25; 0.5; 0.75; 0.9)$. The effects for τ = 0.9 are analysed in a more detailed way, including in this case the corresponding p-values.

We can see that the constant takes negative values for all the adjusted models. With respect to the logarithm of total kilometres driven, we see that its coefficient is positive. For τ = 0,9 the coefficient value is 1091.75 and is significant. This means that when the logarithm of the total number of kilometres increases a unit (thus the total of kilometres driven is multiplied by 2.718) TVaR at level 90% of the distance driven above the posted speed limit increases by 1091.75 km.

With regard to the driving on urban roads, we see that the associated coefficient takes negative values in all models. This means that the more the vehicle is driven on urban roads, the less is the TVaR of number of kilometres of speeding which makes sense given that it is more difficult to go over the posted speed limit in built-up areas. For τ = 0,9 the coefficient is significant; if the percentage of urban driving increases by one point, the TVaR at 90% decreases by 54,68 kilometres.

With night driving, the coefficient is positive for all models and leads to an increase of the risk of speeding. Specifically, for τ = 0,9, the coefficient is significant; with one percentage increase the TVaR of speeding at 90% increases by 26.90 kilometres.

Age is seen to be a positive parameter in that the older the driver, the greater the

Table 4: Estimated coefficients for quantile regression model based on TVaR and its standard errors between parenthesis. β^{τ}_{TVaR} represents the coefficients for TVaR at level τ. β_{lm} represents the linear model estimated coefficients.

Variable	β_{lm}	$\beta^{0.25}_{TVaR}$	$\beta^{0.50}_{TVaR}$	$\beta^{0.75}_{TVaR}$	$\beta^{0.90}_{TVaR}$	**p-value**
Constant	-8120.85	-5936.59 (165.93)	-6753.75 (237.93)	-7050.46 (394.48)	-6194.13 (1117.77)	<0.0001
lnKm	1062.54	831.87 (16.49)	953.45 (20.80)	1102.14 (37.44)	1091.75 (77.07)	<0.0001
Porc_urba	-21.31	-18.41 (0.56)	-22.59 (0.95)	-34.79 (1.11)	-54.68 (4.00)	<0.0001
Porc_noct	5.35	4.78 (1.57)	4.90 (2.84)	16.75 (4.20)	26.90 (9.31)	<0.01
Age	1.37	0.70 (2.82)	7.13 (5.00)	7.99 (7.47)	52.29 (21.56)	0.02
Gender	329.98	285.04 (19.05)	369.19 (30.22)	526.06 (47.71)	907.82 (116.33)	<0.0001
Goodness of fit	-	0.4698	0.5390	0.6682	0.8011	

TVaR. The parameter is significant for τ = 0,9; one year increment in age represents an increment in the TVaR at 90% of 52.29 kilometres. Being a male driver represents a significant increase in TVaR; for τ = 0,9 the coefficient is significant and indicates that male TVaR at 90% is 907,82 kilometres higher than for female.

Figure 2 shows the evolution of the effects of the explanatory variables depending on level τ of the TVaR. We see that as τ rises, the Constant effect becomes more negative, approaching the Constant value of the linear model but never entering into its confidence interval. Between τ = 0,75 and τ = 0,9, constant value increases.

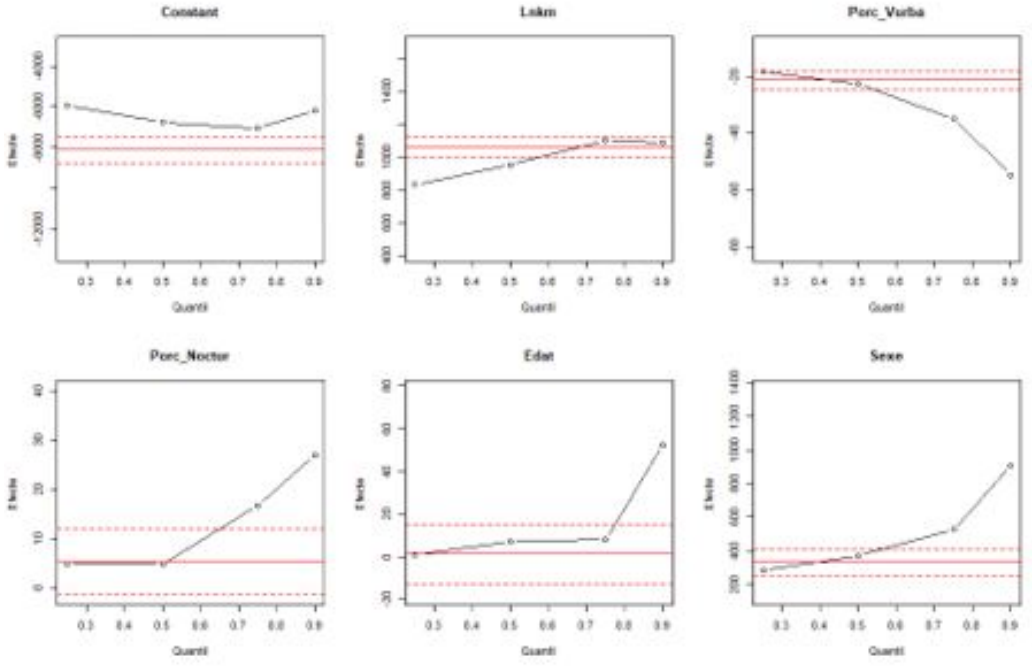

Figure 2: Graphs of the evolution of variable effects as τ quantile increases using quantile regression to predict TVaR.

With respect to the logarithm of number of kilometres driven during the year, we can see that its effect increases as τ does until the effect matches the effect of the linear model for values of τ between 0,75 and 0,9. It could be that this effect is countered by the Constant effect, so that the conjoint effect is equal to the classical model.

With respect to distance driven in urban areas, this effect is the same as in the linear model for τ values less or equal to 0,5. From this point, when the percentile increases, the effect takes more negative values. In relation yo night driving, it can be seen that the effect is also the same as in the linear model until τ is equal to 0,5 and then the effect increases.

A huge coincidence is seen in age, with the effects corresponding to the linear model, given that it lies between the confidence interval for τ values under 0,75. From this value, the effect increases drastically. Finally, the gender variable has the same effect in the linear model and in models with τ less than 0,5 and from this point, the effect also increases drastically.

Finally, Table 4 also shows goodness of fit values. In this case we observe that for small values of τ, goodness of fit decreases but as τ increases, goodness of fit also does until it reaches 0,8011 which indicates a good adjustment.

4 Conclusions

This paper sets out to analyse a way of predicting the value at risk and the value at risk in the tail. We sohw that the use of quantile regression does not provide the same results as when using an ordinary least square regression to predict the expected value. Specifically, in this study we show how quantile regression can be used to estimate the value at risk and the approximation to the value at risk of the tail for each insured driver depending on his/her driving patterns. The application of this methodology allows the risk that a specific driver represents to be assessed, not only with regard to the set of the sample but to other drivers with similar patterns. For example, our practical example shows that, driving above the posted speed limit for a number of kilometres can be more, or less, risky, depending on the other patterns of the same driver, in particular the total number of kilometres driven during the year. This practical application, in the case of TVaR, enables the mean of the number of kilometres driven above the posted speed limit to be calculated, for example, for the level of 90%, from certain characteristics of independent variables. So, if a driver is coming close to this value or if he/she is likely to exceed it, this anomalous behaviour can be identified and, with a "pay as you drive" insurance programme, a premium could be added to the price as a way of penalizing the driving pattern. Car insurance based on use, apart from considering the total number of driven kilometres, should adapt the price to the driving patterns. In the case of speeding, used as an example in this paper, it seems only fair to reward those drivers with low risk levels.

An algorithm to fit conditional tail expectation regression models in driving data for vehicle excess speed

1 Introduction

The analysis of data collected from vehicles in motion is an emerging area in transportation research. The reason for its growing interest is the opportunity it offers to improve road safety and to develop fairer ways of calculating motor insurance prices. The aim of this paper is to propose new models for risk analysis. We present an algorithm that allows regression models to be adjusted for the tail expectation, which are a natural generalization of quantile regression models. Unlike the classical linear model, which finds the effects of covariates on the mean of a response variable, quantile regression identifies the effects on the quantile of the response. Tail expectation regressions can model conditional average responses above a given conditional quantile. In our case study, we show that quantile regression identifies risky drivers by modelling quantiles of distance driven yearly above the posted speed limits. The quantile order is fixed at high levels, such as 95%. We denote as c_τ the quantile at level τ (τ between 0 and 1) of a variable response Y. By definition, the probability that Y is greater or equal to c_τ is equal to τ. Quantiles are used in areas such as finance, insurance and risk analysis, where they are usually referred to as τ – Value at Risk (VaR_τ). Another risk measure is the Expected Shortfall (ES_τ) also known as Conditional Tail Expectation (CTE_τ) or Tail Value at Risk ($TVaR_\tau$). This is defined as:

$$TVaR_\tau(Y) = E(Y|Y > c_\tau). \tag{1}$$

Quantile regression and tail expectation regression specify VaR_τ and $TVaR_\tau$, respectively, as a linear combination of regressors.

2 Methodology

The starting point for this study is quantile regression. Quantile regression is an extension of the linear regression that is especially interesting when the response variable has asymmetry, for instance, when there is a substantial difference between the conditional mean and the conditional median. As is widely known, the median is robust to the presence of outliers, while the mean is not. Koenker and Bassett (1978) proposed an optimization framework to fit quantile regressions. Here, a new procedure to estimate the tail expectation model is presented and implemented in open source software R.

A classical linear regression model is represented as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \cdots \beta_k X_{ki} + \epsilon_i, \tag{2}$$

where Y_i is the response variable for the i^{th} individual (i = 1,$\cdots$, n), X_{ji} represents the value of the i^{th} observation of explanatory variable j (j = 1, $\cdots$, k) and β_j is the j^{th} parameter. The i^{th} linear predictor is defined as $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \cdots \beta_k X_{ki}$. The error term, ϵ_i, is the part of the response variable that cannot be explained by the covariates. Parameter β_0 is known as the intercept and it is usually included in the model, so it can be assumed that the error term has expectation equal to zero. Model (1) is usually estimated by ordinary least squares (OLS), i.e. by minimizing the sum of squared residuals:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} f_i(\beta), \tag{3}$$

where $f_i(\beta) = (Y_i - X_i\beta)^2$ represents the difference between the observed response and the linear predictor.

Quantile regression assumes that the quantile at level τ of the response equals a linear combination of the regressors:

$$VaR_\tau(Y_i|X_{j1}, \cdots, X_{ji}) = \beta_0^\tau + \beta_1^\tau X_{1i} + \beta_2^\tau X_{2i} + \cdots + \beta_k^\tau X_{ki}. \tag{4}$$

Coefficient estimates are obtained as follows (see Koenker and Bassett, 1979; Koenker and Machado 1999):

$$\hat{\beta^\tau} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \left[\rho_i^\tau (Y_i - X_{ij}\beta_j\right]. \tag{5}$$

where $\rho_{i\tau}$ represents a loss function of the quantile, which is equal to τ when $Y_i - X_i\beta$ is greater or equal to 0, and $\tau - 1$ otherwise. The standard deviation of the estimated coefficients can be calculated following the bootstrap method (Chernick,

2011; Hestenberg, 2011).

The specification of tail expectation regression is defined as:

$$TVaR_{\tau}\left(Y_i|X_{j1},\cdots,X_{ji}\right)=\beta_0^{\tau}+\beta_1^{\tau}X_{1i}+\beta_2^{\tau}X_{2i}+\cdots+\beta_k^{\tau}X_{ki}. \quad (6)$$

Acerbi and Szekely (2014) recently proposed a loss function to estimate the conditional tail expectation using the quantile. Despite developing this method theoretically, these authors did not consider a linear predictor. In the field of risk analysis, databases are large. This is the reason why we focus on studying the optimization underlying the estimation procedure. Computational time remains a challenge.

3 Data

Information about different characteristics of 9.614 car drivers was collected during 2010 by an insurance company, using a telematics device. Driving data measure patterns of vehicles in motion, such as distance driven, vehicle speed, time of day, and zone (urban versus nonurban). For privacy reasons, GPS localization data are not recorded. A definition of the variables is presented in Table 1. Drivers are aged between 18 and 35 years because the insurance company offered a "pay as you drive" motor policy only to young drivers. Boucher et al. (2017) studied the transformation of the risk factors with the same dataset; Ayuso et al.(2016a, 2016b) compared the driving patterns between male and female drivers; Guillen et al. (2019) proposed new methods to calculate the price of motor insurance. Pitarque et al. (2019) used quantile regression to analyse risk of having an accident.

Table 1: Definition of the variables in the insurance dataset (9.614 observations in 2010)

Variable	Description
Toler_km	Total number of kilometres driven exceeding the posted limit
lnKm	Logarithm of the total kilometres driven
P_urban	Percentage of kilometres driven in urban areas
P_night	Percentage of kilometres driven at night
Age	Age of driver on 1st of January, 2010
Male	Gender of driver (1 = male, 0 = female)

A descriptive analysis of the data is presented in Table 2. Skewness equal to 3.64 is one of the most relevant features of total distance driven above the posted speed limits during one year. This means that while most drivers have low levels of excess

speeding, a few show larger values. However, all factors total driving distance, urban driving and night driving should be considered before drawing conclusions.

Table 2: Descriptive statistics in the insurance dataset (9.614 observations in 2010)

	Mean	**Median**	**Minimum**	**Maximum**	**Standard deviation**	**Skewness**
Toler_km	1398.21	689.23	0.00	23500.19	1995.37	3.64
lnKm	9.27	9.37	-0.37	10.96	0.75	-1.87
P_urban	26.29	23.39	0.00	100.00	14.18	1.03
P_night	7.02	5.31	0.00	78.56	6.13	1.68
Age	24.78	24.63	18.11	35.00	2.82	0.11

4 Results

A simple quantile regression with only one explanatory variable is adjusted to model the percentage of kilometres driven above the speed limit with $\tau = 0.9$ as a function of the percentage of kilometres driven in urban areas. The tail expectation regression is also fitted. Parameter estimates are not displayed for brevity. The results are shown graphically in Figure 1. Quantile regression at the 0.9 level indicates that when there is an increase of 1% in the percentage of kilometres driven in urban areas, the Value at Risk of the percentage of kilometres driven above the speed limit decreases by 0.35% and the average beyond the quantile level decreases 52 basis points.

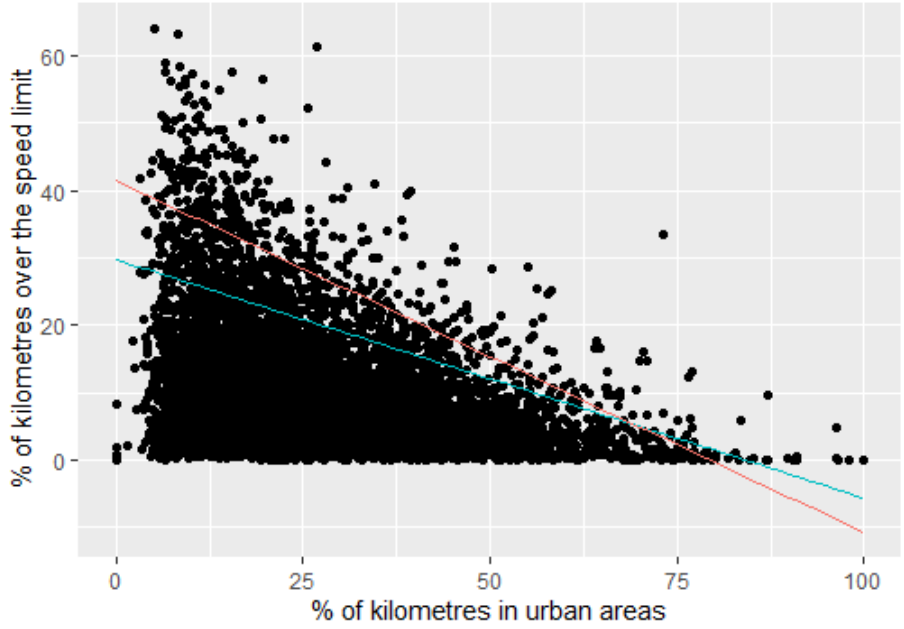

Figure 1: Graph of the relation between the percentage of kilometres driven above the speed limit and the percentage of kilometres driven in urban areas in the insurance dataset. The blue line represents a 90% quantile regression line and the red line represents a 90% tail expectation regression.

In the multivariate case, the total number of kilometres driven above the speed limit as the response variable is analysed for quantile levels $\tau = 0.5$ (median) and $\tau = 0.9$ (upper decile). A linear regression model is also estimated to compare the coefficient estimates.

Coefficient and standard deviation estimates are calculated using the *quantreg* package of R (Koencker et al., 2019). Standard errors are computed from 3000 replications with samples of the same length as the original sample with replacement, so that a comparison between models can be analyzed. Table 3 presents results for the linear regression, the quantile regression and the tail expectation regression together with the goodness-of-fit statistic. As in the univariate case extrapolation of the linear specifications can produce abnormalities such as negative predictions or values of the conditional tail expectation lower than its corresponding quantile level. A summary is reported in Table 4.

Table 3: Results of models of linear regression (OLS), quantile regression (VaR) and tail expectation regression (TVaR) for quantile levels τ = 0.5 and τ = 0.9 in the insurance dataset. In parenthesis, the standard errors of the estimated coefficients.

Variable	**OLS**	$\boldsymbol{VaR}_{0.5}$	$\boldsymbol{TVaR}_{0.5}$	$\boldsymbol{VaR}_{0.9}$	$\boldsymbol{TVaR}_{0.9}$
	-8082.51	-4496.53	-11708.92	-6418.11	-14068.39
Intercept					
	(309.95)	(186.02)	(843.57)	(742.98)	(3505.13)
	1064.51	597.60	1588.38	1074.66	2229.62
lnKm					
	(26.51)	(19.32)	(86.59)	(64.46)	(364.14)
	-21.87	-9.19	-39.72	-39.59	-86.08
P_urban					
	(1.39)	(0.62)	(2.16)	(2.34)	(7.14)
	7.54	5.41	11.99	21.76	26.56
P_night					
	(2.93)	(1.82)	(6.10)	(9.80)	(19.21)
	-1.13	-2.56	0.96	5.16	7.71
Age					
	(6.26)	(3.26)	(11.09)	(15.24)	(37.13)
	328.01	206.76	528.84	574.08	913.63
Male					
	(35.89)	(19.01)	(66.51)	(103.97)	(223.48)
R2	0.25	0.14	0.17	0.20	0.49

Table 4: Percentage of cases where the predicted TVaR is lower than the predicted VaR and percentage of cases where the predicted TVaR is negative in the insurance database. Two quantile levels are considered, τ = 0.5 and τ = 0.9.

$\%\boldsymbol{TVaR}_{0.5} < \boldsymbol{VaR}_{0.5}$	8.20%
$\%\boldsymbol{TVaR}_{0.5} < \mathbf{0}$	7.41%
$\%\boldsymbol{TVaR}_{0.9} < \boldsymbol{VaR}_{0.9}$	6.48%
$\%\boldsymbol{TVaR}_{0.9} < \mathbf{0}$	3.60%

The implementation of a routine to estimate the coefficients for the tail expectation regression can be compared with the VaR regression computation. An evaluation of computational time is presented in Table 5. The difference between TVaR regression and VaR regression is about double time both for the parameter estimates and the standard error. In both cases, the parameter estimates are obtained in less than 0.2 seconds for our working sample of almost 10 thousand cases and six coefficients. The most relevant result is the time needed to compute the standard errors, which is quite low given the number of replicates. The quantile level did not affect computational time required.

Table 5: Computational time comparison in our case study.

Output generated	**Computational time**
Estimation of the VaR coefficients	0.088 seconds
Estimation of the standard deviation of the VaR coefficients	2,618 minutes
Estimation of the ES coefficients	0.175 seconds
Estimation of the standard deviation of the ES coefficients	5,410 minutes

5 Conclusions

In this paper, an innovative method is implemented that generalizes quantile regression in order to study risky drivers. The study ss done using a database of approximately 10.000 observations, which contain a highly skewed response variable. This is a typical feature of risk analysis problem settings. In the case of the bivariate regression, the results show that the percentage of kilometres driven in urban areas influences the risk of exceeding speed limits. Specifically, each additional percent point driven in an urban area reduces the highest decile of the percentage of distance

driven above the speed limits by 0,35%. This decrease is emphasized in the case of the tail expectation where an increase of 1% in the percentage of kilometres driven in urban areas reduces the expected percentage of kilometres driven above the speed limit by 52 basis points, for those drivers that are in the top decile.

In the multivariate case similar conclusions are drawn from quantile regression and tail expectation regression for quantile levels 0.5 and 0.9. Some problems arose when applying the models for an "in-sample" prediction exercise. In a few cases, the tail expectation was lower than the value provided by the quantile, or even negative. This could be a result of the simplicity of the linear specification and further research should be carried out to develop possible solutions to this issue. Despite these problems, the computational time of the estimation procedure to obtain the coefficient estimates is low, so the routine for the tail expectation regression created here is not excessively slow. The computational time for the standard errors is also relatively low, taking into account that the bootstrap method iterates the estimation in a large number of sample replicates.

For future studies, other methods to calculate the standard errors of the coefficient estimates should be investigated so that computational effort does not increase too much with sample size. Certainly with the bootstrap method, there are currently several possible alternatives that seem suitable for addressing this problem. Another area for further analysis is larger datasets and tuning the parameters of the bootstrap method to estimate coefficients and standard errors in a reasonable computational time window.

Joint Generalized Quantile and Conditional Tail Expectation Regression for Insurance Risk Analysis

1 Introduction

Our predictive modelling focuses on Value at Risk (VaR) and the Conditional Tail Expectation (CTE). While classical linear regression finds the effects of covariates on the mean of a response variable via a linear predictor, quantile regression focuses on the VaR of the response, and CTE regression links covariates to the conditional average responses beyond a quantile. We consider the usual case in insurance, where risk concentrates on positive losses. In finance, where risk focuses on negative returns, the usual risk measure is Expected Shortfall (ES) rather than CTE because ES looks at the lack of resources needed to cope with unexpected negative outcomes. In finance, CTE regression is known as ES regression. We call these models *risk regressions*, in general.

Risk regressions have not been popular in insurance because of the technical difficulty of fitting the models. However, they may be extremely useful for identifying factors that influence the worst case outcomes. There are many examples of loss random variables that are asymmetric and right skewed, where risk is located at higher quantiles. A prime example is that of accident severity.

The first complication in risk regression lies in establishing a suitable score function, similar to the sum of squared residuals in the least squares method for linear regression. However, such a score function is not always possible to find. The second complication is that the existing models may produce negative predictions, or even a predicted CTE imcompatible with the predicted VaR[1]. Our contribution solves these two issues and proposes generalized quantile and CTE regressions.

A risk measure is called *elicitable* (Gneiting, 2011) if a scoring function exists

[1]Note that for positive responses, CTE should be larger than the VaR

such that the expected score under a distribution takes its unique minimum at the risk value of the distribution. Wang and Ziegel (2015) and Kou and Peng (2016) have shown that distortion risk measures are rarely elicitable.

Koenker and Bassett Jr (1978) provided a score function for quantile regression (VaR regression) and initiated a methodology that has become increasingly popular over the years (see, Koenker, 2017). Gneiting (2011) showed that the ES is not a 1-elicitable risk measure, which means that there is no score function that can be minimized to obtain ES (or CTE) alone. Therefore, it is not possible to estimate ES regression in the same way as quantile regression. However, Fissler and Ziegel (2016) have established the joint elicitability of VaR and ES risk measures under some regularity conditions for random variables that take negative values and capture left-side risks, and they have shown that the corresponding joint score function is not unique. One particular case is the score function proposed by Acerbi and Szekely (2014). Dimitriadis and Bayer (2019) have analysed possible choices for the family of score functions put forward by Fissler and Ziegel (2016), finding that they have the property of positive homogeneity such that linear rescaling of the input variable does not alter the ranking of losses.

All of the above studies have dealt with left-side risks. However, in insurance and actuarial applications, economic losses are defined as positive. As a result, insurance risk analysis concentrates on large positive values, which are naturally located on the right side of the distribution. We therefore work with positive, right-side risks. This implies a change of sign that is often confusing when drawing on sources that use the other convention. We distinguish CTE regression, for positive right tails, from ES regression, for negative left tails, and we convey all of our results in terms of the risk analysis of random variables defined on the positive real semi-axis. Many recent results for ES regression and forecasting can be easily, but cautiously, rewritten for positive values by changing the sign of the response variable and establishing low quantile levels, such as 5% instead of 95%.

We present predictive models for positive risks. Here, we propose an algorithm to fit generalized CTE regression as a correction of generalized quantile regression, where we include a link function similar to a Generalized Linear Model (GLM) setting. As a result, we can predict the expectation of the dependent variable in the tail of the distribution for given values of the explanatory variables. We specify the model so that we can compare the results of generalized quantile regression and CTE regression without any inconsistencies between the predicted values of the two regressions. We guarantee non-negative predictions and naturally restrict CTE to exceed VaR by definition. Standard errors for the parameter estimates can be found via bootstrap. In the linear case, standard errors can be approximated with a large sample as in Dimitriadis and Bayer (2019).

2 Notation and Basics

Before delving into our regression modelling framework, let us formally define VaR and CTE. Value-at-risk at level τ, $\tau \in (0,1)$, also known as the $(\tau \times 100)$-th percentile or τ-quantile, is defined as follows:

$$VaR_\tau(Y) = inf\left\{y \in \mathbb{R}^+ : F_Y(y) > \tau\right\},$$

where F_Y(y) corresponds to the cumulative distribution function of a continuous non-negative random variable Y. VaR does not consider observations beyond the quantile, but it is one of the most popular measures to analyse risk since it is simple and easy to understand.

To account for observations in the tail, CTE averages the extreme values of the distribution function. This risk measure in continuous variables is also known as Tail Value at Risk (TVaR) or Conditional Value at Risk (CVaR) and is the mean of the values that exceed the VaR. CTE is defined as follows:

$$CTE_\tau(Y) = \mathbb{E}\left[Y|Y > VaR_\tau\right].$$

Definition 3.1. A risk measure $\varphi(Y)$ of a random variable Y is elicitable when it minimizes the expected value of a scoring function, $S(\varphi, Y)$. So, an estimator of an elicitable $\varphi(Y)$ comes from $\hat{\varphi} = \arg\min_{\phi} \mathbb{E}[S(\varphi, Y)]$.

In practice, for a sample $Y_1, ..., Y_n$ of size n, an estimator $\hat{\varphi}$ can be found minimizing $\sum_{i=1}^{n} S(\varphi, Y_i)$. While $VaR_\tau(Y)$ is elicitable, $CTE_\tau(Y)$ is not, essentially because $VaR_\tau(Y)$ is needed to define $CTE_\tau(Y)$. However, they are jointly elicitable under regularity conditions (see, Fissler and Ziegel, 2016).

3 Predictive Models for VaR and CTE

The starting point of our study is quantile regression (Koenker and Bassett Jr, 1978). Even though quantile regression is a relatively new methodology, an increasing number of applications exist in a wide variety of fields (see, Uribe and Guillen, 2020, for an overview of recent methods and R implementation) .

Quantile regression is an extension of linear regression that is especially interesting when the response variable has asymmetry, for instance, when there is a substantial difference between the conditional mean and the conditional median. As is widely known, the median is robust to the presence of outliers, while the mean is not. Risk analysis actually focuses on quantile regression for large τ-quantiles.

To fix notation, let us consider a classical linear regression model for n observations and k covariates, which is specified as:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i, \tag{1}$$

where Y_i is the response variable for the i-th individual ($i = 1, \ldots, n$), X_{ji} represents the value of the i-th observation of explanatory variable j ($j = 1, \ldots, k$) and β_j is the j-th parameter. The i-th linear predictor is defined as $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$ or as $X_i'\beta$ in its matrix form. The error term, ε_i, is the part of the response variable, Y_i, that cannot be explained by the covariates. Parameter β_0 is the intercept and it is usually included in the model so that it can be assumed that the error term has expectation equal to zero, $\mathbb{E}(\varepsilon_i) = 0$. Thus, linear regression under the previous assumption sets $\mathbb{E}(Y_i|X_i) = X_i'\beta$.

Model (1) is estimated using ordinary least squares (OLS) by minimizing the sum of squared residuals,

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} S_i(\beta)$$

where $S_i(\beta) = (Y_i - X_i'\beta)^2$, so this is a score function that represents the difference between the observed and predicted responses of the i-th observation. The expectation is approximated by the sample mean, and the term $(1/n)$ is omitted because it does not affect the minimization.

In GLM, McCullagh and Nelder (1989) introduce a link function $g(\cdot)$, with $g^{-1}(\cdot) = F(\cdot)$, and the conditional expectation is specified as $\mathbb{E}(Y_i|X_i) = F(X_i'\beta)$. The choice of $F(\cdot)$ is not completely free as it must satisfy regularity conditions for likelihood maximization. In addition, $F(\cdot)$ is usually dictated by the nature of Y (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). For example, for a discrete count dependent variable, a Poisson distribution is assumed, with the canonical link function $g(z) = ln(z)$, so that $F(z) = \exp(z)$ and the conditional expectation is always positive.

4 Quantile regression

Unlike linear regression, which estimates the effect of each explanatory variable on the mean of the response variable, quantile regression establishes the effect of explanatory variables on the quantile of the response variable. We can specify the τ-quantile regression model at level $\tau \in (0, 1)$ as follows:

$$Y_i = \beta_0^\tau + \beta_1^\tau X_{1i} + \beta_2^\tau X_{2i} + \cdots + \beta_k^\tau X_{ki} + \varepsilon_i^\tau,$$

where we assume that $VaR_\tau(\varepsilon_i^\tau) = 0$ and β^τ is the vector of unknown parameters. There is no assumption made about the distribution of Y_i, and this is the reason why quantile regression is sometimes called *semiparametric*.

Alternatively, we can write quantile regression as a link between the τ-quantile of Y_i and a linear combination of the regressors, i.e. the linear predictor:

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

In short, $VaR_\tau(Y_i|X_i) = X_i'\beta^\tau$. Koenker and Bassett Jr (1978) proposed an optimization framework to fit quantile regressions. Basically, the parameter estimates can be obtained as the solution of the following optimization problem (see, Koenker and Bassett Jr, 1982; Koenker and Machado, 1999):

$$\hat{\beta}^\tau = \arg\min_{\beta^\tau} \sum_{i=1}^{n} \rho_q^\tau(Y_i - X_i'\beta^\tau), \tag{3}$$

where ρ_q^τ represents the score function of the τ-quantile, which is equal to $\tau(Y_i - X_i'\beta^\tau)$ when $(Y_i - X_i'\beta^\tau)$ is greater than or equal to 0, and $(\tau-1)(Y_i - X_i'\beta^\tau)$ otherwise. The standard error of the estimated coefficients can be calculated following the bootstrap method (see, Chernick, 2011; Hesterberg, 2011).

With no loss of generality, we may introduce link function $F^v(\cdot)$ in (2). So, as opposed to quantile *linear* regression, or simply quantile regression, we can define the *generalized quantile regression* as:

$$VaR_\tau(Y_i|X_i) = F^v(X_i'\beta^\tau),$$

where $F^v(\cdot)$ is monotone and twice continuously differentiable to meet GLM assumptions. For example, we can choose $F^v(z) = \exp(z)$ to guarantee that the predictions are positive. This is exactly the generalized quantile regression that is later implemented in our case study, using

$$VaR_\tau(Y_i|X_i) = \exp(X_i'\beta^\tau). \tag{4}$$

In a simultaneous and independent work, Dimitriadis and Schnaitmann (2019) also introduce link functions.

5 CTE regression

If it is possible to establish a relationship between the explanatory variables and VaR, it should also be possible to do so with other risk measures. The specification

of a conditional tail expectation *linear* regression is:

$$CTE_\tau(Y_i|X_{1i},\ldots,X_{ki}) = \gamma_0^\tau + \gamma_1^\tau X_{1i} + \gamma_2^\tau X_{2i} + \cdots + \gamma_k^\tau X_{ki}, \tag{5}$$

where γ^τ corresponds to the parameters for the effects of the explanatory variables on the expectation above the conditional quantile $VaR_\tau(Y_i|X_i)$. Equivalently, we can use an error term whose CTE_τ equals zero. To ease notation, we write $CTE_\tau(Y_i|X_i) = X_i'\gamma^\tau$ and we assume that (2) and (5) have the same regressors, but we could define a set X_i^q for (2) and another possibly overlapping set X_i^e (5) as in Dimitriadis and Bayer (2019).

With no loss of generality, we may also introduce a link function $F^e(\cdot)$ in (5). We only need the same monotonicity and regularity conditions as before. So, the *generaralized* CTE regression is denoted as $CTE_\tau(Y_i|X_i) = F^e(X_i'\gamma^\tau)$. The *generalized* CTE regression that is implemented subsequently in our case study is:

$$CTE_\tau(Y_i|X_i) = \exp(X_i'\gamma^\tau). \tag{6}$$

5.1 Extensions of generalized VaR and CTE regressions

We have now introduced the link function in the VaR and CTE models. This is the reason why we include the word “generalized" in the name of our models. The choice of the link function has to do with the domain of the response variable, which is non-negative. Unlike in GLM, we do not have to specify a link between the canonical parameter of the exponential distribution of the dependent variable and the linear predictor. In risk regression, our choice of a link function is guided by the need to provide predictions that stay in the domain of our variable of interest.

One of the earliest attempts to introduce a connection between VaR and ES and the predictors appears in a recent study by Taylor (2019). His proposal is to set:

$$CTE_\tau(Y_i|X_i) = VaR_\tau(Y_i|X_i)(1 + X_i'\gamma^\theta)$$

to force CTE to exceed VaR. With the above form, however, this is not necessarily true.

Our proposal is to choose exponential links in order to guarantee that predictions are positive. In addition, we also consider the following specification:

$$CTE_\tau(Y_i|X_i) = F^v(X_i'\hat{\beta}^\tau) + F^e(X_i'\eta^\tau), \tag{7}$$

where $\hat{\beta}^\tau$ is the corresponding term in the generalized quantile regression and η^τ is the vector of unknown parameters that guide the additive term in the CTE regres-

sion. This specification guarantees that $CTE_\tau(Y_i|X_i) \leq VaR_\tau(Y_i|X_i)$ for all given $\tau \in (0,1)$ and that the predictions of VaR and CTE, conditional on X_i, are always positive. We call this specification the *generalized additive term* CTE regression.

6 Estimation Procedure

As mentioned before, the parameter estimates in quantile regression are obtained via the optimization problem as in (3). We do not reproduce the details on how to obtain the score function because this has already been developed extensively in Koenker and Bassett Jr (1982).

Nadarajah et al. (2014) reviewed the estimation methods for CTE in the univariate case, but they are not suitable for the inclusion of regressors. Unfortunately, there is no stand-alone score function parallel to equation (3) to find the parameter estimates of a conditional tail expectation regression. However, Acerbi and Szekely (2014) proposed a way to obtain VaR and CTE together using a score function that relates both risk measures.

Fissler and Ziegel (2016) showed that there are an infinite number of score functions to achieve the joint elicitability of VaR and CTE, but they did not introduce regressors. In order to estimate the effects of the explanatory variables on CTE, we take the score function proposed by Acerbi and Szekely (2014) as a starting point. Dimitriadis and Bayer (2019), based on the work by Fissler and Ziegel (2016), conducted a simulation study in which they showed that some particular choices in the score function might have better small-sample properties than others. Dimitriadis and Bayer (2019) created an R package, *esreg*, which can be used to fit linear quantile and ES regressions in (2) and (5). To fit risk regression on large positive values, the implementation needs to be adapted: the sign of the dependent variable, the level, $(1-\tau)$, and the sign of the resulting parameters have to be reversed.

In addition to the problem of the joint score function to be minimized, a few additional problems arise in practice when fitting VaR and CTE regressions. The first such problem concerns numerical instability, i.e. the fact that local minima may be found. This is the reason why Dimitriadis and Bayer (2019) recommend iterative local metaheuristics inspired by Lourenço et al. (2003). The issue is that this optimization is stochastic, because there is a small random noise alteration of the solution to refine the search for a minimum. So, to obtain the same results, one should always remember to use the same seed in the random number generation. The second problem is that the predicted values may lie outside the plausible range for the identity link, namely $F^v(z) = z$ and $F^e(z) = z$. A prediction is implausible when, given the exogenous information or the covariates, the predicted CTE

does not exceed the predicted VaR, or when a predicted conditional risk measure is negative when we are only considering a non-negative dependent variable.

6.1 Score minimization for VaR and CTE regression

To propose a suitable estimation method, we consider the joint score function established by Fissler and Ziegel (2016) in the univariate setting. In our notation we have positive outcomes and a large τ, for example $\tau = 0.90$, whereas in their original study, the returns were negative and the focus was on low τ levels.

Based on the results obtained by Fissler and Ziegel (2016), the general score function for VaR and CTE regressions (2) and (5) for a non-negative Y_i and linear predictors $X_i'\beta^\tau$ and $X_i'\gamma^\tau$ is:

$$\begin{aligned}\rho(Y_i, X_i, \beta^\tau, \gamma^\tau) = {} & I(Y_i \geq X_i'\beta^\tau)\left[G_1(-X_i'\beta^\tau) - G_1(-Y_i)\right] \\ & + G_2(-X_i'\gamma^\tau)\left[X_i'\beta^\tau - X_i'\gamma^\tau - \frac{(Y_i - X_i'\beta^\tau)}{1-\tau} I(Y_i \geq X_i'\beta^\tau)\right] \\ & - \mathcal{G}_2(-X_i'\gamma^\tau) + a(-Y_i). \end{aligned} \tag{8}$$

where G_2 is the first derivative of $\mathcal{G}_2$. Functions $G_1(\cdot)$ and $\mathcal{G}_2$ must satisfy some regularity conditions. Also, $a(\cdot)$ can be eliminated in the optimization procedure, but it should be carefully selected to guarantee that $\rho(Y_i, X_i, \hat{\beta}^\tau, \hat{\gamma}^\tau) > 0$ for the goodness-of-fit calculation (Koenker and Machado, 1999, see). A common choice is $a(z) = (1-\tau)G_1(z) + \mathcal{G}_2(z)$, (see, Dimitriadis and Bayer, 2019).

To obtain joint estimates from (8), the following optimization problem needs to be solved for a sample (Y_i, X_i), $i = 1, ..., n$:

$$(\widehat{\beta^\tau}, \widehat{\gamma^\tau}) = \underset{\gamma^\tau \beta^\tau}{\arg\min} \sum_{i=1}^{n} \rho(Y_i, X_i, \beta^\tau, \gamma^\tau). \tag{9}$$

The proposal put forward by Acerbi and Szekely (2014) is equivalent to setting $G_1(z) = -Wz^2/2)$ and $\mathcal{G}_2(z) = (1-\tau)z^2/2$, where W is a constant so that WVaR>CTE. This guarantees the required regularity conditions, namely that $G_2(\eta)v/(1-\tau) + G_1(v)$ is a strictly increasing function of v, a VaR_τ, and η is its corresponding CTE_τ. But the choice of W is unclear. Dimitriadis and Bayer (2019) suggest either $G_1(z) = 0$ or $G_1(z) = z$, like Fissler and Ziegel (2016), and they also propose five options[2] for $\mathcal{G}_2(\cdot)$.

An indirect estimator for CTE regression is also presented in Dimitriadis and Bayer (2019), where it is called the *oracle* estimator of CTE regression. We obtain

[2]Some of the choices of G_1 and $\mathcal{G}_2$ were unstable in our implementation and no standard error estimates could be obtained.

estimates of β^τ via the quantile regression score in (3), $\hat{\beta}_O^\tau$, and then minimize the sum of squares of conditional residuals $(Y_i - X_i\hat{\gamma}_O^\tau)$ only for those observations that satisfy $Y_i > X_i\hat{\beta}_O^0$. However, this procedure is not recommended for small samples or extreme quantiles, due to the small number of observations beyond the quantile. We denote the oracle estimator as $(\hat{\beta}_O^\tau, \hat{\gamma}_O^\tau)$.

6.2 Two-step procedure for linear CTE regression

We propose the use of a two-step process to solve (9). First, we estimate β^τ via the quantile regression score in (3) as $\hat{\beta}_O^\tau$ and then we find $\widehat{\gamma^\tau}$:

$$\widehat{\gamma^\tau} = \arg\min_{\gamma^\tau} \sum_{i=1}^{n} \rho_{AS}(Y_i, X_i, \hat{\beta}_O^\tau, \gamma^\tau), \tag{10}$$

where the score function is taken from Acerbi and Szekely (2014), and we follow our positive sign convention:

$$\begin{aligned}\rho_{AS}(Y_i, X_i, \hat{\beta}_O^\tau, \gamma^\tau) = (1-\tau)\left(\frac{(X_i'\gamma^\tau)^2}{2} + W\frac{(X_i'\hat{\beta}_O^\tau)^2}{2} - X_i\gamma^\tau X_i'\hat{\beta}_O^\tau\right)\\ + I(Y_i \geq X_i'\hat{\beta}_O^\tau)\left(-X_i'\gamma^\tau(Y_i - X_i'\hat{\beta}_O^\tau) + W\frac{Y_i^2 - (X_i'\hat{\beta}_O^\tau)^2}{2}\right)\\ + (1-\tau)(W-1)Y_i^2/2,\end{aligned} \tag{11}$$

where $I(Y_i \geq X_i\hat{\beta}_O^\tau)$ equals 1 if $Y_i \geq X_i\hat{\beta}_O^\tau$ and equals 0 otherwise. W is a fixed constant that is selected as before, but this has no impact on the minimization.

In our second step, $\widehat{\gamma^\tau}$ is fixed and the minimization of (10) is on β^τ to refine the quantile regression estimate part. However, this should be done carefully to avoid numerical instability, for instance, by using partial gradient descent.

Standard errors for the linear CTE regression can be found via bootstrap or with the asymptotic approximation provided by Dimitriadis and Bayer (2019). In addition, Taylor (2019) has recently proposed a semiparametric approach to estimate VaR and ES regression, but no asymptotic statistical theory is available.

Following Theorem 2.6 and the notation in Dimitriadis and Bayer (2019), we can approximate the variance and covariance matrix of the estimator for the linear CTE model as follows:

$$\Lambda_{22}^{-1} C_{22} \Lambda_{22}^{-1},$$

$\Lambda_{22} = \mathbb{E}(X_i X_i' G_2^{(1)}(-X_i'\gamma^\tau))$, $G_2^{(1)}$ is the derivative of G_2,

$$C_{22} = X_i G_2^{(1)}(-X_i'\gamma^\tau)^2 \mathbb{V}\left(X_i'\gamma^\tau - X_i'\beta^\tau - (1-\tau)^{-1}(Y_i - X_i'\beta^\tau)I(Y_i - X_i'\beta^\tau)\right) X_i'.$$

Then C_{22}, the asymptotic variance and covariance term of the covariance matrix for the estimator of the CTE model using (11) in the linear case, is approximated as:

$$\sum_{i=1}^{n} X_i \mathbb{V}\left(X'\gamma^\tau - X'\beta^\tau - (1-\tau)^{-1}(Y - X'\beta^\tau)I(Y - X'\beta^\tau)\right) X_i'.$$

The scalar term, $\mathbb{V}\left(X'\gamma^\tau - X'\beta^\tau - (1-\tau)^{-1}(Y - X'\beta^\tau)I(Y - X'\beta^\tau)\right)$, can be approximated as:

$$\left(X_i'\gamma^\tau - X_i'\beta^\tau\right)^2 - (1-\tau)^{-2}\mathbb{V}(Y_i - X_i'\beta^\tau)I(Y_i - X_i'\beta^\tau).$$

6.3 Two-step procedure for generalized CTE regression

In all the previous settings, an identity link with the linear predictor has been assumed. When we replace $X_i'\beta^\tau$ and $X_i'\gamma^\tau$ by the generalized terms using monotone transformations $F^v(\cdot)$ and $F^e(\cdot)$ in (8) and (11), then we obtain the new score functions to be minimized. The asymptotic statistical theory for the linear case is no longer valid for generalized specifications. In the generalized case, we propose a bootstrap method (Efron and Tibshirani, 1994). To obtain the bootstrap estimates, B samples are generated, that is, for each $b = 1, \ldots, B$, a resample of the original data (Y_i, X_i) is considered for all $i = 1, \ldots, n$. Then the bootstrapped parameter estimate is the average of the estimates obtained in the replication process and the bootstrapped covariance matrix is given by the sample covariance over all bootstrapped parameter estimates.

7 Case Study: Predicting the Risk of Driving over the Speed Limit

An increasing number of companies are starting to work with telematics data in order to fit a better price for motor insurance by analysing driving patterns. For this study, we used a database containing information on $9,618$ car drivers aged between 18 and 35 years in 2010. The data contain information on the distance driven over one year, the type of roads, the time of day and the distance driven above the posted speed limit. The definitions of the variables appear in Table 1. The data have been used in previous studies together with claims information. Boucher et al. (2017) have analysed the simultaneous effect of the distance traveled and exposure time on the risk of accident by using Generalized Additive Models (GAM), while Ayuso et al. (2016b) have compared the driving patterns between male and female drivers and Guillen et al. (2019a) have proposed new methods to calculate the price of

Table 1: Definition of variables in the telematics data set for 2010.

Variable	Description*
Speed_km**	Total number of kilometres driven over the speed limit
lnKm	Logarithm of the total number of kilometres driven
P_urban	Percentage of kilometres driven in urban areas
P_night	Percentage of kilometres driven at night
Age	Age of driver
Male	Gender of driver (1 = male, 0 = female)

*Distances driven are measured over one year

**P_speed is the proportion (percentage) of total kilometres driven above the speed limit. P_speed= 100 $\times$ Speed_km /exp(lnKm)

Table 2: Descriptive analysis of the continuous variables in the telematics data set for 2010 ($n = 9,618$).

	Mean	Median	Min.	Max.	Std. Dev.	Skewness
Speed_km	1398.21	689.23	0.00	23500.19	1995.37	3.64
lnKm	9.27	9.37	-0.37	10.96	0.75	-1.87
P_urban	26.29	23.39	0.00	100.00	14.18	1.03
P_night	7.02	5.31	0.00	78.56	6.13	1.68
Age	24.78	24.63	18.11	35.00	2.82	0.11

motor insurance. Pitarque et al. (2019) have used quantile regression to analyse the risk of having an accident and Pérez-Marín et al. (2019b) have analysed speeding.

Our variable of interest is the total number of kilometres driven over the speed limit, Speed_km, which is highly positively skewed. Following previous analyses, we consider the distance driven on a logarithmic scale. The descriptive statistics appear in Table 2. There are 4,873 male and 4,741 female drivers in the sample.

7.1 Results for a bivariate analysis

We present a simple model with one covariate. Our objective is to show the pitfals of existing methods and the advantage of our proposal in an illustrative example. We model the percentage of kilometres driven above the speed limit as a function of the percentage driving in urban areas. Thus, our initial predictive model for risk establishes a linear relationship between the percentage of total distance driven above the posted speed limit, P_speed, computed as (Speed_km $\times$ 100)/ $\exp(\text{lnKm})$, and the percentage driven in urban areas, P_urban. The parameter estimates for linear quantile and CTE regression together with their standard errors have been found

using our estimation approach.[3] A simple linear regression (the details are omitted) finds a negative relationship between P_speed and P_urban, since the slope equals -0.178 (p-value < 0.001), which means that the higher the proportion of driving in urban areas, the lower the proportion of driving above the speed limit. This was expected, because urban areas tend to be more congested than non-urban areas and the possibility of exceeding the speed limit is therefore reduced by traffic. However, we also expect the slope and the intercept to change when looking at the median regression and quantiles with $\tau > 0.5$. Table 3 shows the results for the identity link corresponding to model (2) for VaR (linear quantile regression) and to model (5) for CTE (linear CTE regression) and for the exponential link corresponding to model (4) for VaR (generalized quantile regression) and model (6) for CTE (generalized CTE regression).

[3] A table showing the results obtained with the *esreg* package for linear models and the oracle estimator for $\tau = 0.50, 0.75, 0.90, 0.95$ is available from the authors

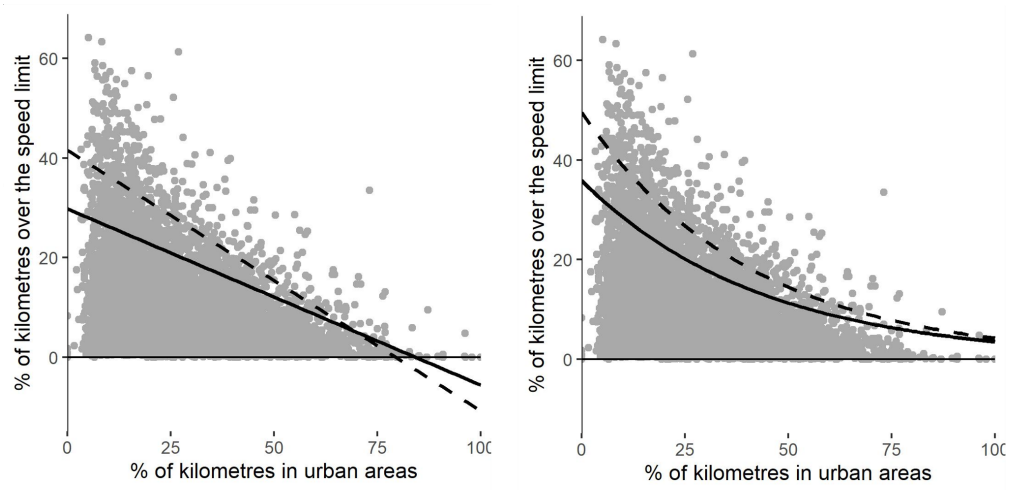

Figure 1: Linear (left) and generalized (right) quantile regression for VaR (solid) and CTE (dashed) of the percentage of distance driven above the speed limit as a function of the percentage of urban driving, at $\tau = 0.9$.

As Figure 1 (left) shows, the 90th-quantile regression finds a linear relationship between the exogenous variable and the response. In the left plot, the problem lies in the extremely large values of the exogenous variable where the predicted values of both risk measures in the linear quantile regression and the linear CTE regression are sometimes negative (0.15% of cases for VaR and 0.24% of cases for CTE) and also where CTE is predicted to be lower than VaR, in 1.02% of cases. The right plot presents the results of the generalized quantile regression (4) and the generalized CTE regression (6).

As can be seen on the right in Figure 1, the main difference is that it is impossible to have negative predictions. The estimation results using our two-step minimization (10) appear in Table 3.

Table 3: Model results for the percentage of distance driven above the speed limit, at quantile levels $\tau = 0.50, 0.75, 0.90$ and 0.95, as a function of the percentage of urban driving. Identity link (upper) and exponential link (lower). Standard errors in parenthesis.

$$VaR_\tau(P\_speed_i|P\_urban_i) = \beta_0^\tau + \beta_1^\tau P\_urban_i$$
$$CTE_\tau(P\_speed_i|P\_urban_i) = \gamma_0^\tau + \gamma_1^\tau P\_urban_i$$

	τ 0.5	0.75	0.90	0.95
$\hat{\beta}_0$	9.322***	18.329***	29.793***	37.334 ***
	(0.116)	(0.146)	(0.306)	(0.382)
$\hat{\beta}_1$	-0.107***	-0.221***	-0.353***	-0.434***
	(0.002)	(0.004)	(0.007)	(0.009)
$\hat{\gamma}_0$	22.822***	31.538***	41.549***	47.472***
	(0.226)	(0.295)	(0.417)	(0.501)
$\hat{\gamma}_1$	-0.300***	-0.413***	-0.522***	-0.565***
	(0.006)	(0.008)	(0.012)	(0.015)
Goodness-of-fit (R^2)	0.003	0.022	0.088	0.189
Score$\times 10^3$	44,711.26	23,147.33	9,167.40	4,229.30
Score$_0\times 10^3$	44,867.09	23,669.75	10,055.20	5,217.83

$$VaR_\tau(P\_speed_i|P\_urban_i) = \exp(\beta_0^\tau + \beta_1^\tau P\_urban_i)$$
$$CTE_\tau(P\_speed_i|P\_urban_i) = \exp(\gamma_0^\tau + \gamma_1^\tau P\_urban_i)$$

	τ 0.5	0.75	0.90	0.95
$\hat{\beta}_0$	2.412***	3.109***	3.580***	3.798***
	(0.018)	(0.015)	(0.014)	(0.015)
$\hat{\beta}_1$	-0.022***	-0.024***	-0.023***	-0.023***
	(0.001)	(0.001)	(0.001)	(0.001)
$\hat{\gamma}_0$	3.310***	3.639***	3.903***	4.028***
	(0.015)	(0.015)	(0.015)	(0.016)
$\hat{\gamma}_1$	-0.025***	-0.026***	-0.025***	-0.023***
	(0.001)	(0.001)	(0.001)	(0.001)
Goodness-of-fit (R^2)	0.012	0.023	0.087	0.155
Score$\times 10^3$	3,601.57	1,657.90	505.68	176.99
Score$_0\times 10^3$	3,645.02	1,696.77	554.16	209.46

Score$_0$ is the value of the score function in a model with only intercepts.
p-value $<1\%$ ***, $<5\%$ **and $<10\%$ *.

7.2 Results for a multivariate analysis

Our aim is now to model the total kilometres driven above the posted speed limit by considering all other covariates to identify risky drivers who exceed the posted speed limit. We use a generalized quantile and generalized CTE regression in order to avoid negative predictions. We therefore, use an exponential link as in (4) and (6). In addition, we estimate the model where the CTE regression part is an additive term, using the specification presented in (7). We prefer the latter for interpreting the effects of covariates on the tail average, as opposed to the quantile effects.

Table 4 below presents the parameter estimates for the generalized quantile regression in (4) and the generalized CTE regression in (6). We omit the results for the linear case because they produce predictions that are out of scope (1.93% of predicted cases for VaR and 3.60% of cases for CTE are negative and CTE is predicted to be lower than VaR, in 6.48% of cases). Table 5 presents the generalized with additive term CTE regression in (7), so that we can interpret the quantile effects and the additional effects for the tail conditional expectation.

Table 4: Model results for distance driven above the speed limit as a function of total distance driven, percentage night driving, percentage urban driving, age and gender at quantile levels $\tau = 0.50, 0.75, 0.90$ and 0.95. Standard errors in parenthesis.

$$VaR_\tau(Y_i|X_i) = \exp(X_i'\beta^\tau)$$
$$CTE_\tau(Y_i|X_i) = \exp(X_i'\gamma^\tau)$$

	τ			
	0.5	0.75	0.9	0.95
$\beta_{\text{Intercept}}$	-5.247***	-3.541***	-2.494***	-1.884***
	(0.157)	(0.168)	(0.163)	(0.125)
β_{lnKm}	1.320***	1.207***	1.141***	1.094***
	(0.014)	(0.014)	(0.014)	(0.010)
$\beta_{\text{P\_urban}}$	-0.015***	-0.019***	-0.020***	-0.020***
	(0.001)	(0.001)	(0.001)	(0.001)
$\beta_{\text{P\_night}}$	0.004**	0.003**	0.001	0.002
	(0.001)	(0.001)	(0.001)	(0.002)
β_{Age}	-0.011***	-0.007**	-0.001	0.002
	(0.003)	(0.003)	(0.002)	(0.002)
β_{Male}	0.290***	0.246***	0.174***	0.123***
	(0.014)	(0.014)	(0.015)	(0.015)
$\gamma_{\text{Intercept}}$	-4.385***	-3.529***	-2.802***	-2.279***
	(0.351)	(0.372)	(0.380)	(0.403)
γ_{lnKm}	1.364***	1.303***	1.237***	1.180***
	(0.038)	(0.040)	(0.041)	(0.044)
$\gamma_{\text{P\_urban}}$	-0.021***	-0.022***	-0.021***	-0.020***
	(0.001)	(0.001)	(0.002)	(0.002)
$\gamma_{\text{P\_night}}$	-0.002	-0.002	-0.002	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)
γ_{Age}	-0.023***	-0.019***	-0.011***	-0.005
	(0.004)	(0.004)	(0.004)	(0.005)
γ_{Male}	0.140***	0.086**	0.043	0.049
	(0.026)	(0.028)	(0.028)	(0.027)
Goodness-of-fit (R^2)	0.029	0.110	0.472	0.996

p-value $<1\%$ ***, $<5\%$ **and $<10\%$ *.

Table 5: Model results for distance driven above the speed limit, as a function of total distance driven, percentage night driving, percentage urban driving, age and gender at quantile level $\tau = 0.90$. Standard errors in parenthesis.

$$VaR_{0.9}(Y_i|X_i) = \exp(X_i'\beta^{0.9})$$
$$CTE_{0.9}(Y_i|X_i) = \exp(X_i'\beta^{0.9}) + \exp(X_i'\eta^{0.9})$$

Variable	$\hat{\beta}^{0.9}$	$\hat{\eta}^{0.9}$
Intercept	-2.494***	-8.500***
	(0.242)	(0.238)
lnKm	1.141***	0.784***
	(0.021)	(0.025)
P_urban	-0.020***	-0.018***
	(0.001)	(0.001)
P_night	0.001	-0.018***
	(0.001)	(0.001)
Age	-0.001	-0.045***
	(0.004)	(0.002)
Male	0.174***	-0.373***
	(0.026)	(0.015)
Goodness-of-fit (R^2)	0.467	

p-value <1% ***, <5% **and <10% *.

An important factor that must be considered when jointly modelling two different risk measures like VaR and CTE is that there is a possibility that an explanatory variable has an impact on one but not the other. In other words, when considering the mean of the worst cases, CTE does not necessarily depend on the same factors as VaR. In Table 4, we see that gender has a positive coefficient in the quantile part, meaning that male drivers have a higher predicted quantile than female drivers at all levels, but we see no significant association between the tail expectation, i.e., the expected driving distance in excess of the speed limit and gender at the 90th and 95th percentiles, all other variables being equal. So, in the top decile, the quantile parameter is higher for males than for females (quantile parameter positive and significant), the tail average distance driven above the speed limit does not differ for the two groups of drivers (CTE parameter not significant).

Table 5 presents the generalized VaR regression and the generalized CTE regression as an additive term as in (7) for the 90th quantile. We want to interpret the results for the top decile of risky drivers which is why we fix $\tau = 0.90$. Here, $CTE_{0.9}(Y_i|X_i) = \exp(X_i'\beta^{0.9}) + \exp(X_i'\eta^{0.9})$. We argue that with this specification we can see the additional influence of each regressor on the tail average. For

example, when looking at the results in Table 5, we conclude that an increment of 1% of the total distance (lnKm) causes an increase of 1.141% in the $\text{VaR}_{0.9}$ of kilometres driven over the speed limit and an additional increase of 1.784% in the mean kilometres for those drivers exceeding the $\text{VaR}_{0.9}$, all other variables being equal. In addition, we see that the effect of age is negative on the CTE regression part, meaning that the average distance driven above the speed limit by drivers in the top decile, $\tau = 0.9$, diminishes with age, whereas age does not preclude them from being in the top risk decile. i.e., the age parameter is not a significant parameter in the quantile regression part. Here again, we see the impact of gender with opposite signs on the quantile part and the CTE additive term part which indicates, as before, that in the top decile the difference in the average distance above the speed limits between male and female drivers at the top decile vanishes. Both the percentage of night driving and the percentage of urban driving have negative effects on the tail average, so the higher the percentage of night driving and urban driving, the lower the tail average distance driven above the speed limit in the top decile, given that we have set $\tau = 0.90$.

In Table 6, VaR and CTE are predicted at level $\tau = 0.9$ using model (7) for the first six observations in our dataset using the multivariate models with exponential links. Note that each driver has a different 90th-percentile and CTE prediction because they depend on the driver's characteristics. Note also that the fifth observation stands out. This particular driver has an observed total speeding distance equal to 2,009.42, which is well above the predicted 90th percentile for drivers with the same characteristics and his observation is almost equal to the tail conditional expectation at level 0.9. This can be used as an indicator of risky driving, as it is widely known that speeding is positively correlated with accident occurrence. The situation is quite different for all other drivers and especially for the third, fourth and sixth drivers, who drive at a much less risky speed than the predicted 90th quantile.

Table 6: Observed distance driven above the speed limit over one year, predicted $\text{VaR}_{0.9}$ and $\text{CTE}_{0.9}$ for the first six observations in the telematics data set.

Observation	Speed_km	Predicted $\text{VaR}_{0.9}$	Predicted $\text{CTE}_{0.9}$
1	4,212.34	9,875.67	12,897.10
2	3,647.30	4,902.82	6,405.09
3	808.59	5,913.95	7,101.61
4	966.69	7,743.66	9,632.31
5	2,009.42	1,681.38	2,077.91
6	187.67	1,024.24	1,093.68

In Figure 2, all the observations versus the CTE predictions are compared at different τ levels. The black dots indicate the observations that exceed the mean of the worst cases, $(1-\tau)$. This serves to identify risky drivers. These drivers have more distance driven above the speed limit than the average of the tail, at the 50th (top left), 75th (top right), 90th (bottom left) and 95th (bottom right) percentile levels. The grey dots indicate the remaining observations.

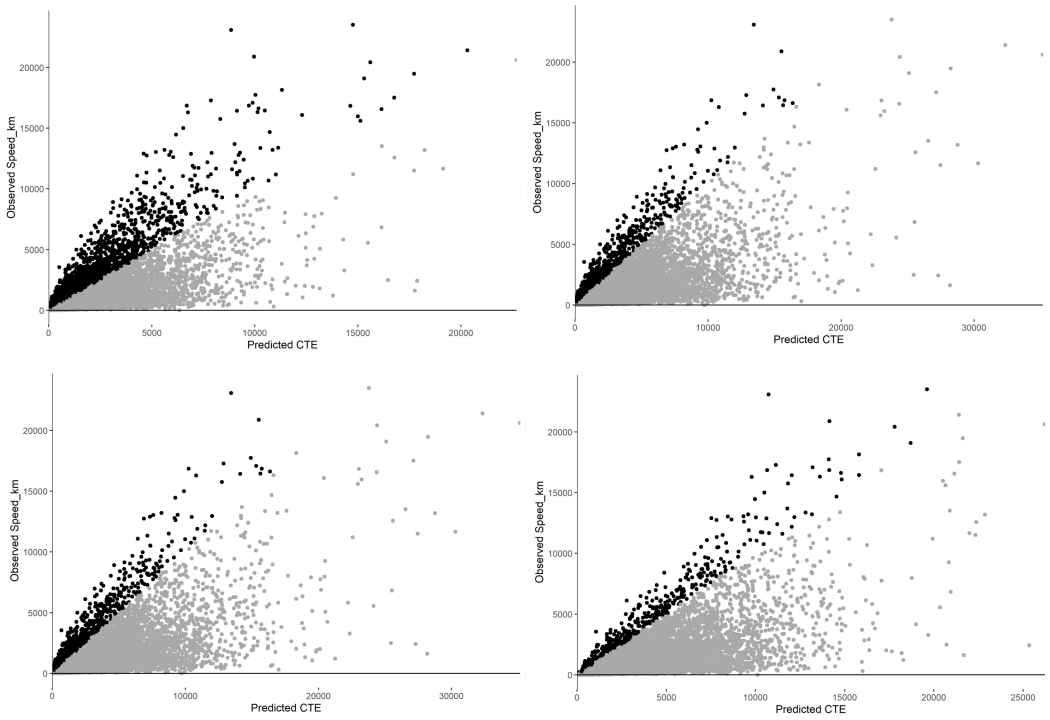

Figure 2: Observed total distance driven above the speed limit (y-axis) versus predicted CTE (x-axis) at $\tau = 0.5$ (top left), 0.75 (top right), 0.90 (bottom left), 0.95 (bottom right). Black dots indicate drivers whose observed distance exceeds the corresponding CTE prediction. Other drivers are displayed in grey.

8 Conclusions

This paper has proposed solutions to the prediction of VaR and CTE for positive losses. In our view, CTE considers values at the extremes and is therefore more informative than VaR. When we adjusted the linear regression versions, we observed that there were predictions that did not fall within a plausible range of the response variable and that the predictions for CTE were greater than for VaR, which is not possible.

We have shown that CTE predictive modelling is helpful in locating risky drivers in a telematics dataset. These methods are easy to implement and can guide risk analysis when there is exogenous information to be considered on the right side of the distribution of a positive response variable.

Our case study shows that risk regression can be used to identify of bad drivers and may guide portfolio selection in motor insurance companies once a level of risk appetite has been selected.

The paper also opens up new lines of research. If it is possible to estimate the effect of covariates for a non-elicitable risk measure such as CTE, it should be possible to follow a similar process to predict other risk measures, or to implement other machine learning methodologies to identify the effects of covariates on a risk measure.

Interpolation of Quantile Regression to Estimate Driver's Risk of Traffic Accident Based on Excess Speed

1 Introduction

Our paper focuses on adjusting the risk level of drivers in car insurance data. Quantile regression adjusts the effects of multiple covariates as function of the Value-at-Risk (VaR) of the response variable. In other words, quantile regression estimates a relation between covariates and the risk of having a high value in the response variable. Value-at-risk at level τ, $\tau \in (0,1)$, also known as τ-quantile is defined as:

$$VaR_\tau(Y) = inf\{y \in \mathbb{R}^+ : F_Y(y) > \tau\}, \tag{1}$$

where $F_Y(y)$ corresponds to the cumulative distribution function of a continuous random variable Y. This measure is the most common in risk analysis in spite of its simplicity.

In recent decades, motor insurance companies have been interested in personalizing prices for their customers. There are different methods to adjust the price. Pay as you drive (PAYD) method determines the price of the insurance depending on the distance driven during one year. This method does not consider the driving patterns of the customer. Pay how you drive (PHYD) method considers the driving patterns of the customer to determine the price of the insurance. Data are collected over the course of a year through a device set up in the motor of the car. The newest method is Manage how you drive (MHYD) method. This is similar to PHYD but in this case, data is obtained almost in real time. Because of the cost, companies are being slower to offer this. Although pricing in insurance has different methodologies, risk regressions have not been popular because of the complexity of fitting models.

In this paper we model the number of kilometres driven above the posted speed limit as function of different characteristics related to the driving patterns of the customers. We also include gender and age in the model. We consider that our variable of interest is strongly related to the risk of having a traffic accident and

should affect the pricing of the insurance. Using quantile regression we compare the predicted values of the response variable for each quantile level with the observed value. The quantile level τ that minimizes this difference will indicate the risk profile of the observation. A τ value close to 1 will indicate that this observation has a high risk of having a high number of kilometres driven above the speed limit. The possibility of identifying a high-risk customer before selling a new policy will allow insurance companies to take measures to protect themselves from extra risk.

The main problem with this method is that a regression must be adjusted for every quantile to obtain the best τ predictions. When this methodology is applied to databases with a large number of variables and to models with a lot of variables, it requires a huge computational time to obtain results. To overcome this, we adjust a reduced number of models and propose an extrapolation method to calculate the effects of those regressions that are not adjusted. We also want to determine the minimum of regressions required to obtain accurate predictions of the risks of the drivers. Obtaining fast predictions of the risk is crucial for providing a good service and the methodology proposed in this paper offers a huge improvement in terms of computational time.

We also study the evolution of the extrapolated values depending on the number of regressions adjusted. We see that for extreme values, the number of regressions adjusted has high impact on the accuracy of the predictions. As the number of regressions adjusted increases, the fit is better.

The paper proceeds as follows. First, we present a literature review of papers that study risk of traffic accident from different points of view. In the next section, we present quantile regression starting from the basic linear regression model and the extrapolation method used to estimate the covariates effects. After methodology, we present data used in this paper and in the following subsections we show the results obtained adjusting a model with multiple covariates and determining the minimum regressions required to obtain adequate results. The last section offers the conclusions.

2 Literature review

Most of the research on risk focuses on studies related to health or the economy. However, there are a number of studies that look at risk in the field of insurance. This brief literature review presents papers from the past five years that study the risk of having a traffic accident.

Smith Smith (2016) studies the relation between having an accident and the driving patterns of drivers. He also considers the risk taken while driving and driving

when the conductor is tired. He carries out a qualitative study using a survey completed in the UK. Another similar study is by Singh Singh (2017) who, from information of multiple traffic accidents in India, assesses which scenarios are likely to be a higher risk of road accident. He considers the weather and location of the crash and discusses solutions to lowering the number of accidents.

Guillen et al. (2020) study the use of reference charts to estimate the percentiles of the distance driven at high speed. Reference charts are usually used to study the weight and height of children and they propose an approach based on telematics data. The authors adjust quantile regression models at different quantiles using variables that show the driving patterns of the sample. They find that total distance driven, gender and percentage of urban driving have a significnt role in explaining the total distance driven above the speed limit over a year. They also found that the relation between the total distance driven above the posted speed limit and the distance driven in total have an exponential relation which causes a difference in the reference chart, allowing for a better fit of the percentiles of the observations in the sample.

Sun et al. (2020) adjust ordinary least square and binary logistic regressions to calculate a driving risk score on different drivers using internet of vehicles (IoV) data. Usage-based insurance is a new methodology based on IoV to customize insurance prices. However, this method requires clear identification of risky drivers. They find that revolutions per minute, speed, braking and accelerations are important variables for the prediction of risky variables while GPS related variables do not provide much information. In both linear and logistic regressions, the number of mistakes made by the prediction system is very low.

Pérez-Marín et al. (2019) also study the risk of speeding adjusting quantile regression at different quantile levels. They use information related to the driving patterns of drivers, in line with in Guillen et al. (2020), but they focus more on the differences of the effects of explanatory variables for different regressions. They conclude that total distance driven, night driving, urban driving, gender and age are important factors in the risk of speeding and propose quantile regression as a methodology when calculating premiums for car insurance. Guillen et al. (2019) study the high number of drivers with zero claims adjusting a zero-inflated Poisson model. Their goal is to propose methodology to improve the design of insurance. They adjust different models for all reported claims and those claims where the driver is at fault obtaining a better adjustment on the latter. In the case of the study of all claims, gender, driving experience, vehicle age, power of the vehicle, distance driven at high speed and urban driving are significant. In the second model, a similar result is obtained but neither gender nor power are significant. The authors highlight the importance of distance driven when analysing risk of accidents and

discuss the addition of different prices depending on distance in pay-as-you drive insurances.

A significant number of papers attempt to relate traffic accidents with an specific disease. Gohardehi et al. Gohardehi et al. (2018), for example, review papers that study if toxoplasmosis has any influence on the risk of having a traffic accident. They use conclusions of studies carried out in different countries to evaluate if the disease is an impact factor. Huppert et al. Huppert et al. (2019) study the risk of road accident in drivers diagnosed in the last five years with some diseases that can cause vertigo. These drivers were not diagnosed at the time that they took out insurance of the vehicle. Matsuoka et al. Matsuoka and Saji (2019) study if there is any correlation between traffic accidents and epileptic drivers that have sleep-related problems. They consider driver characteristics and the type of epilepsy that they suffer.

Closer to the aim of this paper, are other studies that try to examine which factors affect the risk of crash adjusting mathematical models. In the case of Mao et al. Mao et al. (2019) a multinomial logistic regression is adjusted to identify which factors affect to the risk of having a traffic accident in China. They consider four different types of crashes depending on the collision characteristics, and separate the factors into six categories. Rovsěk et al. Rovšek et al. (2017) identify the risk factors adjusting a Classification and Regression Tree (CART) using data collected from Slovenia which provide information about the conditions when the accident happen. Lu et al. Lu et al. (2016) study the agents that affect the severity of traffic accidents adjusting an ordered logit model. The data contain information about different traffic accidents that occur in different Shanghai tunnels and include characteristics of the driver, time, weather conditions and site features.

3 Proposed Methodology

Lets consider the linear model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_k i,$$

where Y_i represents the response variable for the *i*-th individual (*i = 1...n*), X_{ji} represents the value of the *i*-th observations of the explanatory variable j (*k = 1...k*), β_j is the parameter for the *j*-th explanatory variable and ϵ_i represents the residuals of the model that follows E[ϵ|Xβ] = 0. This model allows the effects for multiple covariates to be adjusted with respect to the mean of the response variable. In order to calculate the estimations of β parameters, an optimization problem is solved. This problem is known as Ordinary Least Square (OLS) and minimizes the sum of

the squared residuals.

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} S_i(\beta), \tag{2}$$

where $S_i(\beta) = (Y_i - X_i'\beta)^2$ and represents the diference between the predicted and observed values of the response variable.

3.1 Quantile Regression

In risk analysis, studying regression for extreme positive values of the response variable can be of great interest. These values correspond to high values of the quantile $\tau \in (0,1)$. Koenker and Bassett (1978) Koenker and Bassett Jr (1978) propose an extension of the linear regression called quantile regression. Quantile regression adjusts the effects of the explanatory variables for the τ-th quantile of the variable of interest. This methodology works extremely well when the response variable is not symmetrical because the median of a random variable is robust to the outliers of the distribution while the mean is affected. We can specify the regression model at level τ as follows:

$$Y_i = \beta_0^\tau + \beta_1^\tau X_{1i} + \beta_2^\tau X_{2i} + \cdots + \beta_k^\tau X_{ki} + \epsilon_i^\tau, \tag{3}$$

where $VaR_\tau(\epsilon_i^\tau) = 0$ and β^τ is the vector of unknown parameters. Quantile regression can also be specified as a relation between the τ-th quantile of Y_i and the linear combination of the covariates with the equation:

$$VaR_\tau(Y_i|X_{1i},\cdots,X_{ki}) = \beta_0^\tau + \beta_1^\tau X_{1i} + \beta_2^\tau X_{2i} + \cdots + \beta_k^\tau X_{ki} + \epsilon_i^\tau \tag{4}$$

Koenker and Bassett (1982) Koenker and Bassett Jr (1982) and Koenker and Machado (1999) Koenker and Machado (1999) specify the adapted optimization problem to adjust the quantile regression model.

$$\beta^\tau = \arg\min_{\beta^\tau} \sum_{i=1}^{n} \rho_q^\tau(Y_i - X_i'\beta^\tau), \tag{5}$$

where ρ_q^τ represents a score function of the quantile that equals $\tau(Y_i - X_i'\beta^\tau)$ when $(Y_i - X_i'\beta^\tau) \geq 0$ and $(\tau - 1)(Y_i - X_i'\beta^\tau)$ otherwise.

3.2 Extrapolating parameters

Let us suppose that, instead of adjusting $VaR_\tau(Y_i|X_i'\beta^\tau)$, we want to identify at which quantile τ corresponds to every observation depending on its observed vari-

ables. To adjust the quantile value we solve the following optimization problem

$$\hat{\tau}_i = \arg\min_{\tau_i}(Y_i - X_i\beta^{\tau})^2, \tag{6}$$

which represents the difference between the observed value of the variable Y and the predicted value of the response variable for quantile τ. The ideal scenario to solve this problem would be to dispose of an adjusted regression adjusted for each quantile τ because we would have more predicted values to compare with the observed one. This is non-viable when the database has a massive number of observations and when we are trying to adjust a model with a big number of variables because of the required computational time. In this paper we want to determine how many quantile regressions are required to get accurate predictions of τ without modelling for all quantiles and to save time.

We adjust 99 regressions (1 for each entire number between 1 and 99) and select $m < 99$ regressions that are in a similar distance from each other. I.e. if m=6, $\tau = (0.01, 0.2, 0.4, 0.6, 0.8, 0.99)$. To extrapolate the values of β^{τ} for the not selected τ, we use the formula:

$$\hat{\beta}^{\tau_0+c} = \frac{\beta^{\tau_0} + c(\beta^{\tau_1} - \beta^{\tau_0})}{\tau_1 - \tau_0}. \tag{7}$$

In this formula values of τ are multiplied by 100 to have integer quantiles. We will select two consecutive values of the vector of quantiles to estimate values of β^{τ} between them. τ_0 represents the lower selected quantile, τ_1 represents the upper selected quantile and c is the difference between the lower quantile and the value τ that we are extrapolating. c must accomplish $c \in (0, (\tau_1 - \tau_0))$.

To compare the obtained predictions extrapolating regressions and predictions adjusting all possible regressions we calculate the Mean Square Error (MSE) using the formula:

$$MSE = \frac{\sum_{i=1}^{n}(\tau_i^{99} - \tau_i^{m})^2}{n}, \tag{8}$$

where n represents the number of observations, τ_i^{99} represents the predicted tau for the i-th observation adjusting 99 regressions and τ_i^{m} represents the predicted tau for the i-th observation adjusting m regressions.

3.3 Other methodology to extrapolate parameters

Using the methodology proposed to extrapolate the coefficients, we choose m regressions with τ values that are equidistant. Thus, we are adjusting the same amount of regressions at parts of the distribution of the response variable which have a lot

of information and at those parts that do not. The following figure shows that for the lower quantiles the distribution function rises steeply, then the increase is constant and for higher quantiles the increase is exponential again.

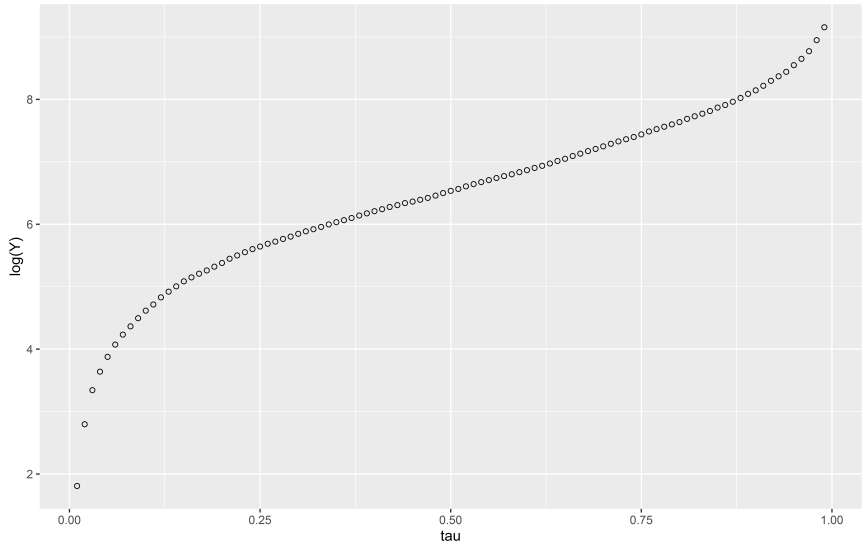

Figure 1: Distribution of log(Y) dependent on quantile τ.

Bearing in mind that when we work with quantile regression we are establishing a linear relation between the effects of the variables, it is important to determine at which zones of the distribution more regressions should be adjusted. In the case of our data, we need to adjust the major number of regressions for lower quantiles because this is where the biggest increase of the distribution appears. Also, we need to adjust the minimum number of regressions for the central part of the distribution because the increase is linear. Using this methodology, it is really important to select appropiate τ values because they have a huge impact on the results. Changing the amount of regressions adjusted at each part of the distribution or even changing the distance between τ values selected lead to poor results for the adjustment of the risk. In this paper we do not propose any specific methodology to identify which regressions provide the best results, rather we make an arbitrary selection that improves the results of choosing equidistant quantiles.

4 Data and results

Telematics data have become more relevant in recent years. In the field of motor insurance, they are used to adjust personalized prices for customers, depending on

their driving patterns. In this paper we use a database containing information of 9.614 drivers aged between 18 and 35 years. The information contained on each observation includes distance driven, time of day, type of road and distance driven above the posted speed limit. It also contains information on the drivers such as age and gender. All information contained in the database was collected during the year 2010. The description of used variables can be found in Table 1.

Table 1: Definition of variables in the telematics data set for 2010.

Variable	Description
Speed_km*	Total number of kilometres driven over the speed limit
lnKm	Logarithm of the total number of kilometres driven
P_urban	Percentage of kilometres driven in urban areas
P_night	Percentage of kilometres driven at night
Age	Age of driver
Male	Gender of driver (1 = male, 0 = female)

*P_speed is the proportion (percentage) of total kilometres driven above the speed limit. P_speed= 100 $\times$ Speed_km /exp(lnKm)

Table 2 presents a descriptive analysis of the variables used in this study. The sample contains 4.873 male and 4.741 female. We use total distance driven in a logarithmic scale. Our variable of interest is the number of kilometres driven above the speed limit and we can see that this is positively skewed.

Table 2: Descriptive analysis of the continuous variables in the telematics data set for 2010 ($n = 9,618$).

	Mean	Median	Min.	Max.	Std. Dev.	Skewness
Speed_km	1398.21	689.23	0.00	23500.19	1995.37	3.64
lnKm	9.27	9.37	-0.37	10.96	0.75	-1.87
P_urban	26.29	23.39	0.00	100.00	14.18	1.03
P_night	7.02	5.31	0.00	78.56	6.13	1.68
Age	24.78	24.63	18.11	35.00	2.82	0.11

Other studies also used this database. Ayuso et al.(2016)(Ayuso et al., 2016b) compared driving patterns between male and female drivers and Guillen et al.(2019)Guillen et al. (2019a) proposed new methodology to determine the insurance price. Boucher et al.(2017)Boucher et al. (2017) analysed the effects of distance driven and the exposure time to the risk of having a traffic accident using Generalized Additive Models (GAM). Pitarque et al.(2019)Pitarque et al. (2019)

used quantile regression to analyse the risk of traffic accident and Pérez-Marín (2019)Pérez-Marín et al. (2019b) analysed speeding.

4.1 Multivariate case study

Our objective is to model the total number of kilometres driven above the speed limit to identify which observations correspond to risky drivers, given their driving patterns. We work with the logarithm of the response variable. In this case, we compare τ predictions for each driver using the parameters estimated adjusting a regression for every quantile and those estimated adjusting a lower number of models. First, Figure 2 show τ predictions change as the number of adjusted models increase.

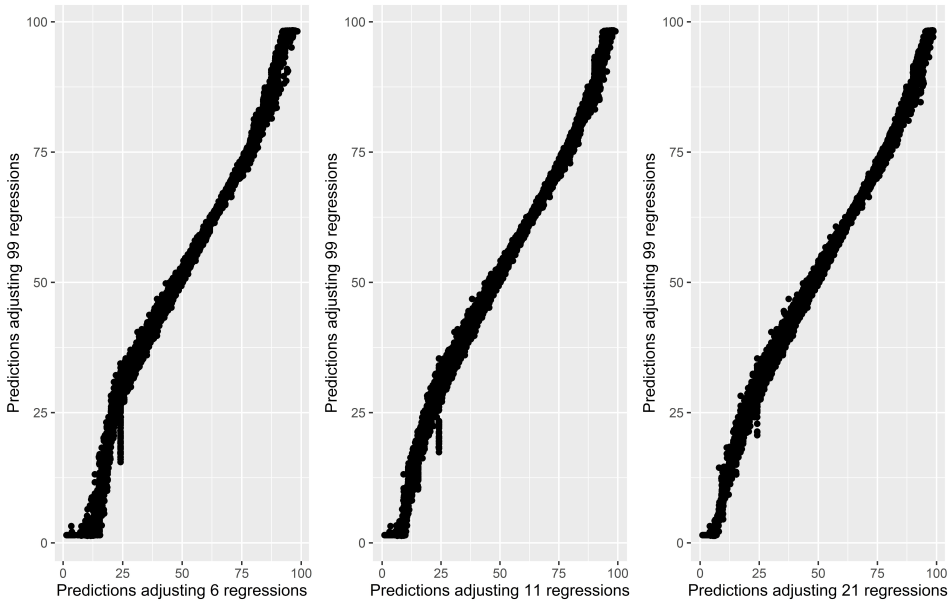

Figure 2: Multivariate predictions.

The horizontal axis represents the value of τ predictions obtained adjusting m regressions and vertical axis represents predictions obtained adjusting 99 regressions. We see that adjusting 6 models, there is a problem when predicting lower quantile values. In these cases, the predicted value is larger than it should be. There are also some differences in the predictions of the larger quantiles but to lesser magnitude. Adjusting 11 regressions, predictions for $\tau \in (0.4, 0.85)$ are more accurate than before and the problem of predicting extreme values of τ relaxes, especially

for large values. Adjusting 21 regressions, we see that in general all predictions are more accurate but there are still some discrepancies for lower quantile predictions. To determine how many regressions are necessary to obtain a good fit we present Figure 3 that shows the behaviour of the MSE as we adjust more regressions. Some of its corresponding values appear in Table 3.

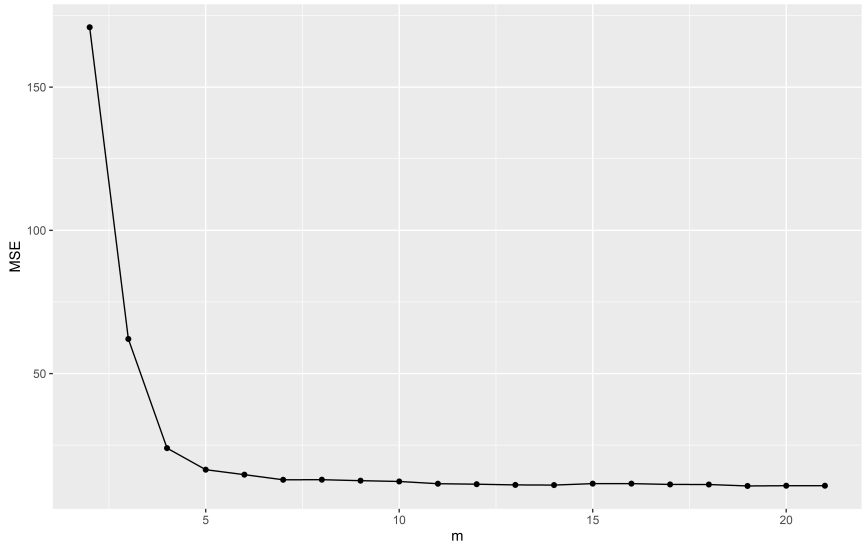

Figure 3: MSE values for predictions adjusting m regressions.

Table 3: MSE values for predictions adjusting m regressions.

m	2	6	9	11	13	15	18	21
MSE	170.890	14.713	12.609	11.545	11.116	11.573	11.264	10.836

In both Figure 3 and Table 3 we see that MSE values decrease sharply as the number of adjusted regressions increases and then stabilizes around m = 10. Despite there being, for m = 11, problems predicting lower quantiles, in general, there is no significant improvement in the results of the predictions. For upper quantiles, these predictions are accurate, which is positive in terms of risk analysis. As seen in Figure 5, the evolution of β^τ parameters is not necessarily linear. Depending on which m quantiles are selected to model, increases or decreases in MSE are produced. To study the possible causes of the prediction problems, in Figure 4 we present the evolution of β^τ parameters and the extrapolation, adjusting m = 13 regressions that had the lowest MSE value for $m < 15$.

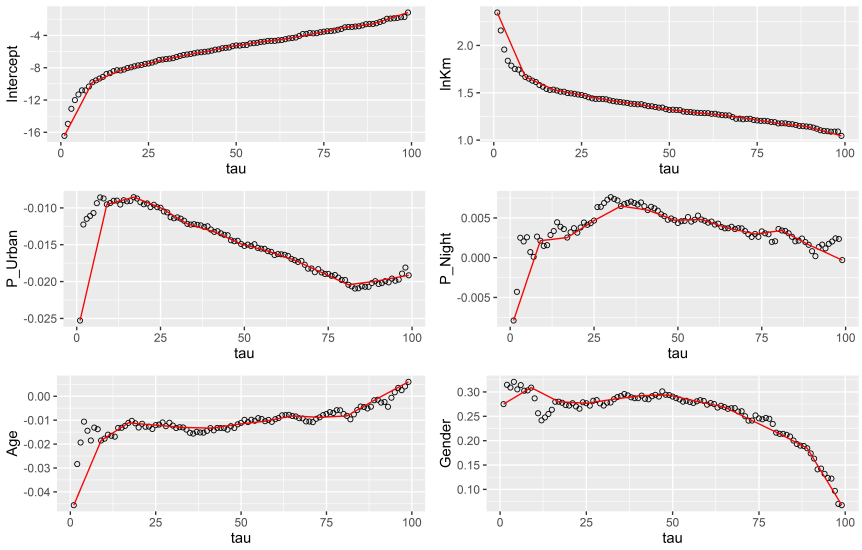

Figure 4: Comparison between β^{τ} estimation.

Figure 4 shows that, in general for all covariates, extrapolations do not fit the parameters for lower quantiles well. In all cases except total distance driven, the extrapolated value is lower than the adjusted value for m = 99. Since the extrapolated value is lower than it should be, a larger value of τ is required to obtain the minimum value for $Y_i - X_i'\beta^{\tau}$. For upper quantiles, the unique variable that has problems of fit is the percentage of night driving. Adjusting m=13 regressions, it takes thirteen seconds to obtain all predictions while adjusting m=99, the computational time is two minutes. This would suppose a major improvement in terms of computational time for larger of databases containing more observations.

Considering the distribution on 1, we decide to adjust m=13 regressions selected as follows: 5 regressions for $\tau \in [0.01, 0.20)$, 4 regressions for $\tau \in [0.20, 0.80)$ and 4 regressions for $\tau \in [0.80, 0.99)$. In this example, unbalancing distance between τ values gives us more accurate extrapolations for β^{τ} parameters for the extreme values of τ. Observing the new comparison between β^{τ} estimation, we see that the problem for the estimation of lower quantile coefficients that appeared in Figure 5 is resolved and an acceptable adjustment for the rest of the quantile coefficients is maintained.

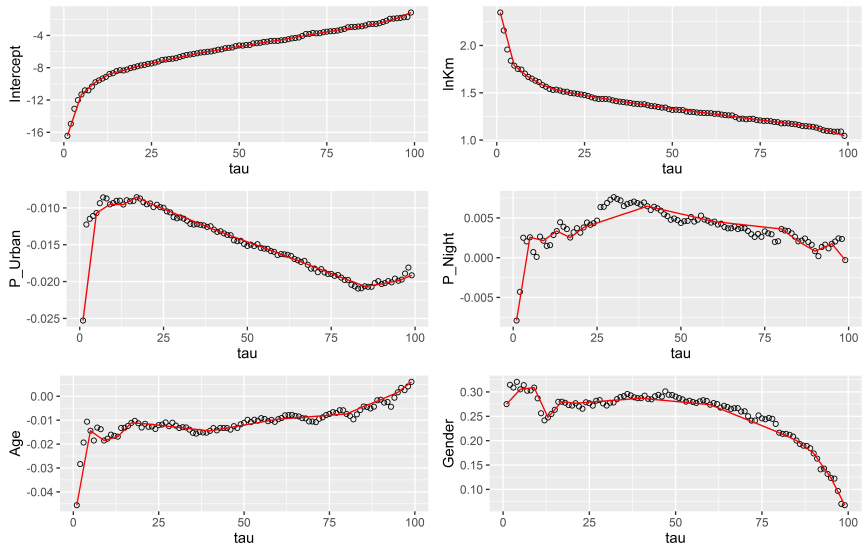

Figure 5: Comparision between β^{τ} estimation with unbalanced τ values.

The multivariate predictions show a good fit for all quantiles especially for the extreme values of τ, this being the aim of unbalancing distances between the selected τ values. MSE is also affected positively when we extrapolate β^{τ} with this methodology. Compared with 11,116 (value obtained from Table 3) MSE now equals 10,057, which indicates an improvement on the quality of our predictions.

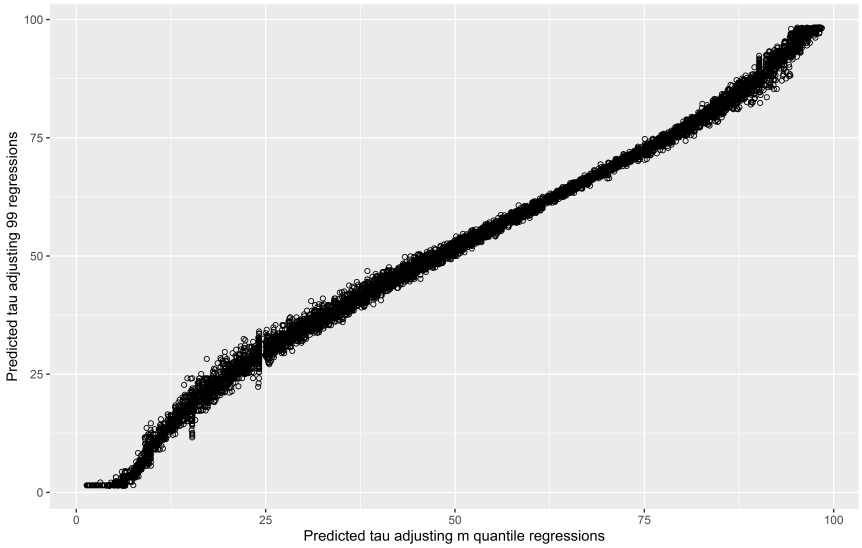

Figure 6: Comparision between β^{τ} estimation with unbalanced τ values.

5 Conclusion

In this paper, we propose methodology to extrapolate β^{τ} parameters of quantile regression without adjusting a regression for all quantile levels τ. When offering insurance policies, it is really important to detect which drivers hold a major risk of having a traffic accident or bad driving patterns. Adjusting a quantile regression for each quantile to check which quantile level is more similar to the observation has a high computational time cost. This cost is accentuated when the adjusted model has more variables or the database has a large number of observations. We see that, for 9.614 observations, the computational time is drastically reduced.

We show that our extrapolation adjusting a reduced number of regressions, gives good results when predicting quantile τ. Adjusting 10 regressions is enough to obtain good predictions, and that increasing the number of regressions would provide few benefits. Nevertheless, the evolution of β^{τ} parameters should be studied in each case to select an appropriate number of models to adjust.

We see that as the number of regressions adjusted increases, β^{τ} extrapolated parameters are closer to the adjusted parameters. However, for lower quantiles there is always a small error that provokes some deviation in our predictions. Applying the different methodology proposed in this paper we show that if we take τ values observing the distribution of the response variable we can obtain better predictions on the quantile of each observation. Although in this study the improvement is

huge, this is very sensitive to which regressions we are choosing to adjust and how many for each part of the distribution.

This paper opens up new fields of research. In our case we establish a linear relation between the extrapolated β^{τ}. Finding new extrapolation methods would allow correction of the prediction mistakes for extreme quantiles. Further study is required into how this extrapolation method improves computational time in data that contains more variables and observations.

A Sarmanov Distribution with Beta Marginals: An Application to Motor Insurance Pricing

1 Introduction

We analyse a bivariate model based on the Sarmanov distribution with marginal Beta distributions. These marginals are specified based on a generalized linear model (Beta-GLM) or Beta regression as defined by Ferrari and Cribari-Neto (2004). The objective is to fit data defined in the $(0, 1)$ interval.

Many authors have analysed bivariate Beta distributions [see, for example, (Gupta and Wong, 1985; Olkin and Liu, 2003; Olkin and Trikalinos, 2015; Arnold and Ng, 2011)]. However, these distributions pose several difficult challenges: their generalization to higher dimensions and their specification as a generalized linear model are not straightforward. The Sarmanov distribution provides a way to address these challenges.

Originally, the Sarmanov distribution in its bivariate form was introduced by Sarmanov (1966), its multivariate version was suggested by Joe and Xu (1996) and was generalized by Bairamov et al. (2011). Its use to model the bivariate behaviour of random variables with a marginal $Beta(\alpha, \beta)$ distribution was proposed by Gupta and Wong (1985). These authors defined the five parameter bivariate Beta distribution from what is known as Morgenstern's distribution (Morgenstern, 1956) with marginal Beta, which is a particular case of the Sarmanov distribution.

The bivariate Sarmanov distribution is characterized by its flexibility in the marginal distributions and, furthermore, given that its functional form establishes that the marginals are clearly separated from the dependency model, the specification in terms of a bivariate generalized linear model turns out to be natural. Generalizing from two dimensions to higher dimensions is simple—(see (Bolancé and Vernic, 2019) for an example of a trivariate Sarmanov distribution specified as a generalized linear model with Negative Binomial marginals).

In this work, we show an application of the bivariate Sarmanov distribution with

Beta marginals generalised linear model to predict two of the most relevant telematics variables in motor insurance (Guillen et al., 2019a). Telematics variables are obtained from GPS/inertial devices installed in vehicles and they provide an abundant source of information to motor insurers. In our case study, a bivariate model is specified, for the proportion of kilometres driven above the posted speed limit and the proportion of kilometres driven at night. These two variables seem to be related, but researchers have not yet been able to find a good way to understand their association. The explanatory variables are the characteristics of the insured policyholder and the vehicle. The database used in our application has already been analysed in various works published in statistical, transport and risk analysis journals (see (Ayuso et al., 2019; Pérez-Marín et al., 2019b; Pesantez-Narvaez et al., 2019; Pérez-Marín et al., 2019a; Pérez-Marín and Guillen, 2019; Guillen et al., 2019a; Sun et al., 2020)). In all previous studies, the two telematics variables that we analyse here were used as predictors of the accident rate, and they were assumed to be uncorrelated.

In subsection 2, the new bivariate Sarmanov model is specified and the particular case with marginal Beta-GLM with a domain in the $(0,1)$ interval is analysed; the estimation method is also discussed. The results of our case study are shown in subsection 3. Finally, subsection 4 contains the conclusions.

2 The Models

Let (Y_1, Y_2) be a bivariate random vector that, for convenience, is defined in $(0,1)^2$. Its distribution depends on a set of k quantitative or binary covariates, whose values are represented by the vector $x_j = \left(x_{1j}, ..., x_{kj}\right)'$, $j = 1, 2$, where $x_{1j} = 1$ is a constant term. The relationship between Y_j and the covariates is given by the linear predictor $x_j'\beta^j$, where $\beta^j = \left(\beta_1^j, ..., \beta_k^j\right)'$, $j = 1, 2$, are vectors of parameters to be estimated. To simplify the notation, the covariates are assumed to be the same for $j = 1$ and $j = 2$, and so the vector of explanatory variables is denoted as x. The bivariate probability density function (pdf) associated with the Sarmanov distribution is:

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2|x'\beta^1, x'\beta^2) = & \quad f_1(y_1|x'\beta^1) f_2(y_2|x'\beta^2) \\ \times & \quad \left[1 + \omega\phi_1(y_1|x'\beta^1)\phi_2(y_2|x'\beta^2)\right], \; y_1, y_2 \in (0,1) \end{aligned} \quad (1)$$

where ω is the dependence parameter and ϕ_j, $j = 1,2$, are bounded kernel functions. For the function defined in (2) to be a pdf, the following conditions must hold:

$$\int_0^1 \phi_j(y_j|x'\beta^j) f_j(y_j|x'\beta^j)\, dy_j = 0, j = 1,2 \tag{2}$$

and

$$1 + \omega\phi_1(y_1|x\beta^1)\phi_2(y_2|x\beta^2) \geq 0, \forall (y_1,y_2) \in (0,1)^2. \tag{3}$$

For given values of $x'\beta^j$, $j = 1,2$, we define:

$$m_j\left(x'\beta^j\right) = \inf_{0<y_j<1} \phi_j\left(y_j|x'\beta^j\right) \quad \text{and} \quad M_j\left(x'\beta^j\right) = \sup_{0<y_j<1} \phi_j\left(y_j|x'\beta^j\right), j = 1,2.$$

Taking into account the condition defined in (3), bounds can be defined for the dependency parameter ω. However, as this parameter does not depend on the linear predictor, new extreme values are defined as: $m_j^\star = \max_{\forall x'\beta^j} m_j(x'\beta^j)$ and $M_j^\star = \min_{\forall x'\beta^j} M_j(x'\beta^j)$, so that the bounds of the dependency parameter are:

$$\max\left\{-\frac{1}{m_1^\star m_2^\star}, -\frac{1}{M_1^\star M_2^\star}\right\} \leq \omega \leq \min\left\{-\frac{1}{m_1^\star M_2^\star}, -\frac{1}{M_1^\star m_2^\star}\right\}. \tag{4}$$

The previous condition holds for every vector of covariates x, which implies that the dependency parameter must be located within the narrowest bounds. In practice, we will assume that the vectors observed in the sample dataset lead to the entire domain of values of linear predictors $x'\beta^j$, $j = 1,2$. In the insurance context, where we will discuss our illustration, we assume that all possible risk profiles that can be insured by the company are already present in the portfolio.

For each vector of covariate observations x, we can also obtain the covariance between the dependent variables as:

$$cov(Y_1,Y_2) = \omega v_1(x) v_2(x), \tag{5}$$

where $v_j(x) = \int_0^1 y_j\phi_j(y_j|x\beta^j) f_j(y_j|x'\beta^j)\, dy_j$, $j = 1,2$. The correlation is obtained by dividing by the product of standard deviations.

There exist many possible specifications for the kernel functions ϕ_j, $j = 1,2$ (see (Bahraoui et al., 2015) [for a description of kernel functions proposed in the literature). When fitting the bivariate Beta distribution without covariates, Gupta and Wong (1985) propose a kernel function such as $\phi_j = 2F_j - 1$, where F_j is the cumulative distribution function (cdf). This specification has the advantage that the bounds for the dependency parameter are given by $-1 \leq \omega \leq 1$ for any vector

x. However, the previous model does not allow obtaining closed expressions for some magnitudes of interest, such as the conditioned moments. In this work, we propose to use kernels $\phi_j = y_j^r - E(Y_j^r)$, where r is a value to be determined by the analyst. Next, some results obtained for the particular case of the Sarmanov distribution with marginal $Beta(\alpha, \beta)$ distribution with $r = 1$ are analyzed. These cases intuitively correspond to a situation of linear dependency, controlled by the dependence parameter ω.

2.1 The Bivariate Beta GLM Model

The pdf of a random variable Y with $Beta(\alpha, \beta)$ distribution, with parameters $\alpha,\ \beta > 0$, is:

$$f_Y(y;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} = \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

and its cdf is:

$$F_Y(y;\alpha,\beta) = \frac{B(y,\alpha,\beta)}{B(\alpha,\beta)},$$

where $\Gamma(\cdot)$ and $B(\cdot,\cdot)$ are the Gamma and Beta functions, respectively, and $B(y,\cdot,\cdot)$ is the incomplete Beta function.

The Beta regression was proposed by Ferrari and Cribari-Neto (2004), with the reparametrization $\mu = \frac{\alpha}{\alpha+\beta}$ and $\psi = \alpha + \beta$, so that:

$$f(y;\mu,\psi) = \frac{1}{B\left(\mu\psi,(1-\mu)\psi\right)} y^{\mu\psi-1}(1-y)^{(1-\mu)\psi-1},$$

where $E(Y) = \mu$, with $0 < \mu < 1$, and $V(Y) = \frac{\mu(1-\mu)}{(1+\psi)}$, with $\psi > 0$, where ψ^{-1} is the scale parameter. We note that, given the values of μ and ψ, it holds that $V(Y) < 0.25$. The specification as GLM is defined as (note that we use $\mu(x)$ to emphasize that μ depends on the linear predictor):

$$g\left[\mu(x)\right] = x'\beta,$$

where $g[\cdot]$ is a link function that can be defined in different ways, in this work, we use the logit link, $g\left[\mu(x)\right] = \log\left[\frac{\mu(x)}{1-\mu(x)}\right]$.

To simplify the notation from now on, we eliminate the linear predictors in the conditioned part. The pdf associated with the bivariate random vector (Y_1, Y_2) with a Sarmanov distribution and Beta GLM marginals that will be called the Sarmanov-

Beta-GLM is ():

$$\begin{aligned} f_{Y_1,Y_2}(y_1,y_2) &= \frac{1}{B\left(\mu_1(x)\psi_1,(1-\mu_1(x))\psi_1\right)} y_1^{\mu_1(x)\psi_1-1}(1-y_1)^{(1-\mu_1(x))\psi_1-1} \\ &\times \frac{1}{B\left(\mu_2(x)\psi_2,(1-\mu_2(x))\psi_2\right)} y_2^{\mu_2(x)\psi_2-1}(1-y_2)^{(1-\mu_2(x))\psi_2-1} \\ &\times \left[1+\omega(y_1-\mu_1(x))(y_2-\mu_2(x))\right],\ y_1,y_2\in(0,1). \end{aligned} \tag{6}$$

For the previous expression to be a pdf, the dependency parameter must be located within the bounds defined in (4), which, for the kernel functions that we propose, are:

$$\begin{aligned} &\max\left\{-\frac{1}{\max\limits_{\forall x'\beta^1}(-\mu_1(x))\max\limits_{\forall x'\beta^2}(-\mu_2(x))},-\frac{1}{\min\limits_{\forall x'\beta^1}(1-\mu_1(x))\min\limits_{\forall x'\beta^2}(1-\mu_2(x))}\right\} \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \leq\omega\leq \\ &\min\left\{-\frac{1}{\min\limits_{\forall x'\beta^1}(1-\mu_1(x))\max\limits_{\forall x'\beta^2}(-\mu_2(x))},-\frac{1}{\max\limits_{\forall x'\beta^1}(-\mu_1(x))\min\limits_{\forall x'\beta^2}(1-\mu_2(x))}\right\}. \end{aligned} \tag{7}$$

The bivariate cdf associated with a Sarmanov-Beta-GLM is obtained directly from the double integral of the bivariate pdf defined in (6):

$$\begin{aligned} F_{Y_1,Y_2}(y_1,y_2) &= \frac{B\left(y_1,\psi_1\mu_1(x),(1-\mu_1(x))\psi_1\right)}{B\left(\psi_1\mu_1(x),(1-\mu_1(x))\psi_1\right)} \times \frac{B\left(y_2,\psi_2\mu_2(x),(1-\mu_2(x))\psi_2\right)}{B\left(\psi_2\mu_2(x),(1-\mu_2(x))\psi_2\right)} \\ &\times\left[1+\omega\left(\frac{B\left(y_1,\psi_1\mu_1(x)+1,(1-\mu_1(x))\psi_1\right)}{B\left(\psi_1\mu_1(x),(1-\mu_1(x))\psi_1\right)}-\mu_1(x)\frac{B\left(y_1,\psi_1\mu_1(x),(1-\mu_1(x))\psi_1\right)}{B\left(\psi_1\mu_1(x),(1-\mu_1(x))\psi_1\right)}\right)\right. \\ &\left.\times\left(\frac{B\left(y_2,\psi_2\mu_2(x)+1,(1-\mu_2(x))\psi_2\right)}{B\left(\psi_2\mu_2(x),(1-\mu_2(x))\psi_2\right)}-\mu_2(x)\frac{B\left(y_2,\psi_2\mu_2(x),(1-\mu_2(x))\psi_2\right)}{B\left(\psi_2\mu_2(x),(1-\mu_2(x))\psi_2\right)}\right)\right], \end{aligned} \tag{8}$$

where $y_1,y_2\in(0,1)$.

Proposition 1. The conditioned pdf is:

$$\begin{aligned} f_{Y_1|Y_2}(y_1|Y_2=y_2) &= \frac{1}{B\left(\mu_1(x)\psi_1,(1-\mu_1(x))\psi_1\right)} y_1^{\mu_1(x)\psi_1-1}(1-y_1)^{(1-\mu_1(x))\psi_1-1} \\ &\times \left[1+\omega(y_1-\mu_1(x))(y_2-\mu_2(x))\right],\quad y_1,y_2\in(0,1) \end{aligned} \tag{9}$$

and similarly for $f_{Y_2|Y_1}(y_2|Y_1=y_1)$. Integrating the previous expression, the con-

ditional cdf is obtained as

$$F_{Y_1|Y_2}(y_1|Y_2=y_2) = F_1(y_1)\times \qquad [1+\omega\,(y_2-\mu_2(x))\,(1-\mu_1(x))] \\ - \quad \omega\,(y_2-\mu_2(x))\,\frac{y_1(1-y_1)}{\psi_1\mu_1(x)}f_1(y_1),\ y_1,y_2\in(0,1) \quad . \tag{10}$$

Proof. The conditioned pdf is obtained directly as

$$f_{Y_1|Y_2}(y_1|Y_2=y_2) = \frac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_2}(y_2)}.$$

Integrating the result of $f_{Y_1|Y_2}(y_1|Y_2=y_2)$ in (9), we obtain:

$$F_{Y_1|Y_2}(y_1|Y_2=y_2) = \int_0^{y_1} f_1(t)dt+\omega\,(y_2-\mu_2(x))\int_0^{y_1} f_1(t)\,(t-\mu_1(x))\,dt \\ = \quad F_1(y_1)+\omega\,(y_2-\mu_2(x))\left[\frac{B\,(y_1,\psi_1\mu_1(x)+1,(1-\mu_1(x))\psi_1)}{B\,(\psi_1\mu_1(x),(1-\mu_1(x))\psi_1)}-\mu_1(x)F_1(y_1)\right]. \tag{11}$$

In addition, since

$$\begin{aligned} & & \frac{B\,(y_1,\psi_1\mu_1(x)+1,(1-\mu_1(x))\psi_1)}{B\,(\psi_1\mu_1(x),(1-\mu_1(x))\psi_1)} \\ = & \frac{B\,(y_1,\psi_1\mu_1(x),(1-\mu_1(x))\psi_1)}{B\,(\psi_1\mu_1(x),(1-\mu_1(x))\psi_1)} - \frac{y_1^{\mu_1(x)\psi_1}(1-y_1)^{(1-\mu_1(x))\psi_1}}{\psi_1\mu_1(x)B\,(\psi_1\mu_1(x),(1-\mu_1(x))\psi_1)} \\ = & F_1(y_1)-\frac{y_1(1-y_1)}{\psi_1\mu_1(x)}f_1(y_1), \end{aligned}$$

then, by substituting the previous expression in (2.1), then (1) follows directly. □

The conditioned quantile is obtained from the inverse of expression (1), for which a numerical method (such as Newton's method) can be used.

Proposition 2. The conditional expectation is:

$$E\,(Y_1|Y_2=y_2) = \mu_1(x)+\omega(y_2-\mu_2(x))V\,(Y_1|x)\,, \tag{12}$$

where $V\,(Y_1|x)=\frac{\mu_1(x)(1-\mu_1(x))}{(\psi_1+1)}$ is the variance, which also depends on the vector of covariates. Similarly, $E\,(Y_2|Y_1=y_1)$ can be found.

Proof. The conditional expectation is obtained directly by solving the integral:

$$\begin{aligned}
E\left(Y_1|Y_2=y_2\right) = & \int_0^1 y_1 f_{Y_1|Y_2}(y_1|Y_2=y_2)\,dy_1 \\
= & \int_0^1 y_1 f_{Y_1}(y_1)\,dy_1 \times (1+\omega(y_1-\mu_1(x))(y_2-\mu_2(x))) \\
= & \int_0^1 y_1 f_{Y_1}(y_1)\,dy_1 \\
& +\omega(y_2-\mu_2(x))\left(\int_0^1 y_1^2 f_{Y_1}(y_1)\,dy_1 - \mu_1(x)\int_0^1 y_1 f_{Y_1}(y_1)\,dy_1\right) \\
= & \mu_1(x)+\omega\left(y_2-\mu_2(x)\right)\left(E(Y_1^2|x)-\mu_1(x)^2\right) \\
= & \mu_1(x)+\omega\left(y_2-\mu_2(x)\right)V\left(Y_1|x\right).
\end{aligned}$$

Likewise, the corresponding result is obtained for $E\left(Y_2|Y_1=y_1\right)$. □

Proposition 3. From (5), the conditional covariance which depends on the vector of covariates x is:

$$cov(Y_1,Y_2)=\omega V(Y_1)V(Y_2)=\omega\frac{\mu_1(x)(1-\mu_1(x))}{(\psi_1+1)}\frac{\mu_2(x)(1-\mu_2(x))}{(\psi_2+1)} \tag{13}$$

and the correlation is:

$$corr(Y_1,Y_2)=\omega\sqrt{\frac{\mu_1(x)(1-\mu_1(x))}{(\psi_1+1)}}\sqrt{\frac{\mu_2(x)(1-\mu_2(x))}{(\psi_2+1)}}. \tag{14}$$

Proof. Note that the covariance and the correlation are calculated directly if, in expression (5), we see that:

$$\begin{aligned}
v_j(x) &= \int_0^1 y_j\phi_j(y_j|x\beta^j)f_j(y_j|x'\beta^j)\,dy_j \\
&= \int_0^1 y_j(y_j-\mu_j(x))f_j(y_j|x'\beta^j)\,dy_j = E(Y_j^2|x)-\mu_j(x)^2, j=1,2
\end{aligned}$$

□

The dependence parameter of the model proposed in Gupta and Wong (1985), which uses kernel functions $\phi_j = 2F_j - 1$, $j = 1,2$, is located in the interval $-1 \leq \omega \leq 1$ and is the same for all x. Our proposal bounds the dependence parameter to the narrowest interval among those obtained from all x. However, the advantage of our proposal is that our model allows for obtaining closed expressions for some magnitudes of interest such as bivariate moments (covariance) and conditional moments. In the numerical analysis subsection, we also compare the

correlations estimated from our model and that of Gupta and Wong (1985).

2.2 Estimation

In practice, we start from a bivariate sample of n observations. Let us denote the sample information as (Y_{i1}, Y_{i2}), $i = 1, ..., n$, where for each i we know the values of the covariates $X_i = (X_{i1}, ..., X_{ik})'$. Our objective is to estimate the parameter vectors β^j, the scale parameters, ψ_j and $j = 1, 2$, and the dependency parameter ω, from the maximization of the logarithm of the likelihood function associated with the Sarmanov distribution:

$$\begin{aligned} l\left(\beta^1, \beta^2, \psi_1, \psi_2, \omega\right) &= \sum_{i=1}^{n} \log f_1(Y_{i1}|X_i'\beta^1) + \sum_{i=1}^{n} \log f_2(Y_{i2}|X_i'\beta^2) \\ &+ \sum_{i=1}^{n} \log\left(1 + \omega\phi_1(Y_{i1}|X_i'\beta^1)\phi_2(Y_{i2}|X_i'\beta^2)\right) \\ &= l_1(\beta^1, \psi_1) + l_2(\beta^2, \psi_2) + l_{12}(\omega, \beta^1, \beta^2, \psi_1, \psi_2), \quad (15) \end{aligned}$$

The maximization of (15) cannot be carried out directly without considering that the parametric space is restricted and, in addition, as it was shown in expression (4), the bounds of the dependence parameter are closely related to the parameters of the marginals. Thus, in the maximization process, infeasible solutions will often be reached unless a careful numerical procedure is specifically designed. One way to address these difficulties is to rely on the IFM (Inference from Margin) method that has been widely used in the estimation of copulas see (Joe and Xu, 1996) [for a review]. For the estimation of the Sarmanov distribution, the IFM was already used by Bolancé and Vernic (2019) for the case of GLM marginals with Negative Binomial distributions.

The IFM method is implemented as follows:

Inicialization. The parameters for the marginals are estimated as:

$$\left(\hat{\beta}^{1(0)}, \hat{\psi}_1^{(0)}\right) = \max_{\beta^1, \psi_1} l_1(\beta^1, \psi_1) \quad (16)$$

$$\left(\hat{\beta}^{2(0)}, \hat{\psi}_2^{(0)}\right) = \max_{\beta^2, \psi_2} l_2(\beta^2, \psi_2). \quad (17)$$

For the initial estimation, function `betareg()` of `betareg` R package is used. With these parameters of the marginals, we start the iterative process in the two steps described below.

Step 1. Given the estimated marginal parameters in iteration $m-1$ and taking into account the limits of the dependence parameter ω defined in (4), with function

`optim()` and the `L-BFGS-B` method using `R` , we estimate ω from the maximization of the likelihood function given fixed values of the marginal parameters, which is:

$$\hat{\omega}^{(m)} = \max_{\omega} l_{\omega|12}\left(\omega|\hat{\beta}^{1(m-1)}, \hat{\beta}^{2(m-1)}, \hat{\psi}_1^{(m-1)}, \hat{\psi}_2^{(m-1)}\right), \tag{18}$$

where $l_{\omega|12}$ is the likelihood as a function of ω given the estimated parameters for the marginals in iteration $m-1$.

Step 2. Given the estimated dependency parameter $\hat{\omega}^{(m)}$ in step 1, the marginal parameters are re-estimated in iteration m as:

$$\left(\hat{\beta}^{1(m)}, \hat{\psi}_1^{(m)}, \hat{\beta}^{2(m)}, \hat{\psi}_2^{(m)}\right) = \max_{\beta^1,\psi_1,\beta^2,\psi_2} l_{12|\omega}(\beta^1, \psi_1, \beta^2, \psi_2|\hat{\omega}^{(m)}), \tag{19}$$

where $l_{12|\omega}$ is the likelihood as a function of the marginal parameters given the dependence parameter estimated in step 1. The above maximization is also performed with function `optim()` and the `L-BFGS-B` method of `R`.

Steps 1 and 2 described above are repeated until reaching the convergence criterion based on the differences between parameter estimates obtained in two consecutive iterations.

Remark 1. In the initialization process, if the dependent variables contain zeros or ones, the following correction $\tilde{Y}_j = (Y_j * (n-1) + 0.5)/n,\ j = 1, 2$ was proposed by Smithson and Verkuilen (2006).

In practice, the algorithm described above is based on the optimization of conditional likelihood functions and not on the likelihood function defined in (15). However, in the last stage, the parameters estimated with the IFM method can be used as initial parameters in the process of maximizing the full likelihood function defined in (15). For this purpose, function `optim()` and method `L-BFGS-B` of `R` are used again.

Remark 2. To estimate the Sarmanov model proposed by Gupta and Wong (1985), it is not necessary to use the two-step process, since the bounds of the dependence parameter do not depend on the parameters of the marginal distributions.

3 Numerical Analysis

We analyse a database corresponding to a car insurance portfolio, in which part of the variables have been measured via a telematic system. The objective of our analysis is to model the joint behaviour of the percentage of kilometres driven above the posted speed limits (Y_1) and percentage of kilometres driven at night (Y_2). It is

well known that both variables are related to the risk of having an accident. In Table 1, we show the main descriptive statistics of the dependent variables and the covariates used in the modelling process. For the estimation of the Sarmanov-Beta-GLM, the dependent variables have been transformed as indicated in Remark 1 in subsection 2.2. Furthermore, to avoid very low coefficient values due to the scale of some covariates, variables age (X_1), age of driving license (X_2) and age of the vehicle (X_5) have been divided by 10; the vehicle power variable (X_6) is divided by 100 and the total annual distance driven in kilometres (X_7) is divided by 1000. In addition, note that, in this study, we have included a variable denoting the driver's gender (X_3) and an indicator of private garage (X_4) as covariates.

The last row of Table 1 shows the Pearson correlation between the two dependent variables. This correlation is compared with the corresponding parameter estimate obtained from the Sarmanov model with marginal Beta proposed here and with the one proposed by Gupta and Wong (1985), from now on the GW model. With this objective, Table 2 shows the dependence parameters estimated with both models, and the AIC and BIC statistics without including the covariates and including them. Using expression (14) and without covariates, from the dependence parameters $\hat{\omega} = 14.883$, it can be deduced that the estimated correlation is 0.0601, which is within the confidence interval of the Pearson correlation as shown in the last row of Table 1. On the contrary, if we use the five parameter Beta distribution, the (residual) correlation that is obtained from the numerical calculation of expression (5) is practically zero. This means that the association is captured by the bivariate model. Comparing both models, with and without covariates, using the AIC and BIC statistics, the results of Table 2 show that the fit is better for the model proposed here than it is for the GW model.

Table 1: Definition of variables and descriptive statistics: mean, standard deviation (STD), minimum (Min) and Maximum (Max). The last row shows the linear correlation between dependent variables and a confidence interval at the 95% level.

Variable	Description	Mean	STD	Min	Max
Y_1	Percentage of kilometres driven above the speed limit	0.063	0.068	0.000	0.704
Y_2	Percentage of kilometres driven at night	0.069	0.064	0.000	1.000
X_1	Age of the driver	27.565	3.094	19.849	36.904
X_2	Age if driver License	7.174	3.053	1.810	15.910
X_3	Gender (=1 Men, =0 Women)	0.489	0.500	0.000	1.000
X_4	Night parking (=1 yes, 0=no)	0.774	0.418	0.000	1.000
X_5	Age of the vehicle	8.749	4.174	1.938	20.468
X_6	Power of the vehicle in Horse Power (HP)	97.226	27.772	12.000	500.000
X_7	Total Km	7159.510	4191.753	1.590	50,035.560
ρ	Pearson correlation between dependent variables (CI)		0.070 (0.057,0.082)		

Table 3 shows the results of our Sarmanov-Beta-GLM using different vectors of covariates. Model I includes all the explanatory variables, among which we have the

Table 2: Estimated dependence from Sarmanov-Beta models and goodness of fit criteria

		$\hat{\omega}$ (p-Value)		AIC	BIC
Proposed Model	No covariates	14.883	(<0.001)	−171,282.2	−171,241.5
	With all covariates	2.388	(0.055)	−177,508.8	−177,354.4
GW Model	No Covariates	0.002	(0.346)	−171,165.4	−171,124.8
	With all covariates	0.002	(0.356)	−177,497.2	−177,342.8

age (X_1), the age of the driving license (X_2) and the total distance driven annually (X_7), these three variables are associated with driving experience. To analyze the robustness of the results, in Model II, age (X_1) is eliminated, and, in addition, in Model III, the age of a driver's license (X_2) is also eliminated. The results of Model I show that the effect of age is negative on both Y_1 and Y_2 that the effect of the driver's license age is positive on Y_1 and negative on Y_2 and the effect of total distance, X_7, is positive on both dependent variables. By eliminating age (X_1) in Model II, the signs of the parameters associated with X_2 and X_7 are maintained, although the value is smaller in the case of X_2 and remains practically the same for X_7. After eliminating variables X_1 and X_2, we see that the effect of the total annual distance driven remains practically the same. If we observe the effects of the rest of covariates, these are practically the same in models I, II, and III. A man driver (X_3) with a powerful vehicle (X_6) would have larger Y_1 and Y_2 than the rest, all other characteristics being the same. However, using parking at night (X_4) has a positive effect on the percentage of speeding distance (Y_1) and a negative effect on the percentage of night-time driving (Y_2); the opposite happens with the age of the vehicle (X_5). The effect of X_5 indicates that, when the vehicle is older, drivers tends to diminish the percent of speed driving, while night-time driving is larger.

To visualize the dependence between Y_1 and Y_2 in different quantiles, the following three examples of insured drivers are graphically analysed:

- **Profile 1** corresponds to a 27-year-old man, who drives about 7000 kilometres per year, with a 7-year-old driving license, with parking, with a vehicle of about 8 years and 100 HP.
- **Profile 2** corresponds to a 20-year-old man, who drives about 4000 kilometres per year, with a 2-year-old driving license, with parking, with a vehicle of about 2 years and 75 HP.
- **Profile 3** corresponds to a 36-year-old man, who drives about 10,000 kilometres per year, with a 15-year-old driving license, without parking, with a vehicle of about 15 years and 200 HP.

Table 3: Parameter estimates (p-values) for the Sarmanov-Beta models and goodness of fit statistics.

	Model I		**Model II**		**Model III**	
	Y1	**Y2**	**Y1**	**Y2**	**Y1**	**Y2**
Cons.	−3.055 (<0.001)	−2.556 (<0.001)	−3.819 (<0.001)	−2.975 (<0.001)	−3.796 (<0.001)	−3.061 (<0.001)
X_1	−0.339 (<0.001)	−0.185 (<0.001)		-		-
X_2	0.294 (<0.001)	−0.052 (0.018)	0.048 (0.002)	−0.187 (<0.001)		-
X_3	0.097 (<0.001)	0.274 (<0.001)	0.107 (<0.001)	0.281 (<0.001)	0.109 (<0.001)	0.274 (<0.001)
X_4	0.108 (<0.001)	−0.031 (0.007)	0.107 (<0.001)	−0.031 (0.007)	0.107 (<0.001)	−0.031 (0.007)
X_5	−0.043 (0.001)	0.055 (<0.001)	−0.043 (0.001)	0.055 (<0.001)	−0.043 (0.001)	0.055 (<0.001)
X_6	0.653 (<0.001)	0.077 (<0.001)	0.654 (<0.001)	0.079 (<0.001)	0.664 (<0.001)	0.038 (0.027)
X_7	0.045 (<0.001)	0.035 (<0.001)	0.046 (<0.001)	0.035 (<0.001)	0.046 (<0.001)	0.035 (<0.001)
ϕ_1	18.480 (<0.001)		18.300 (<0.001)		18.294 (<0.001)	
ϕ_2	14.823 (<0.001)		14.782 (<0.001)		14.703 (<0.001)	
ω	2.388 (0.055)		2.325 (0.059)		2.214 (0.060)	
AIC	−177,508.8		−177,238.5		−177,113.5	
BIC	−177,354.4		−177,100.3		−176,991.6	

hline

Profile 1 represents the average insured individual of the portfolio; Profile 2 is a younger man driver, less experienced than Profile 1 and with a newer and less powerful vehicle; finally, Profile 3 is an older man driver, more experienced than Profile 1 and an older and more powerful vehicle. Figure 1 represents different quantiles of the variable kilometres driven above the speed limit (Y_1) in the y-axis given the values of the percentage of kilometres driven at night (Y_2) for Profile 1 in the x-axis. Quantiles have been obtained from the expression (1). Note that, if the dependence parameter was zero, all the curves would remain constant. The adjusted dependence structure results in the represented conditional quantiles having a negative nonlinear relationship and, furthermore, the curves for the different quantile levels are non-parallel. Figure 1 indicates that, for Profile 1, the higher the percentage of kilometres driven at night (Y_2), the greater the caution in driving and, therefore, the lower the percentage of distance driven above the speed limits (Y_1). The same quantiles at 75% (plot on the left) and 95% (plot on the right) confidence levels are represented in Figure 2. These plots show that the curves are non-parallel and that Profile 3 is the most risky, followed by Profiles 1 and 2.

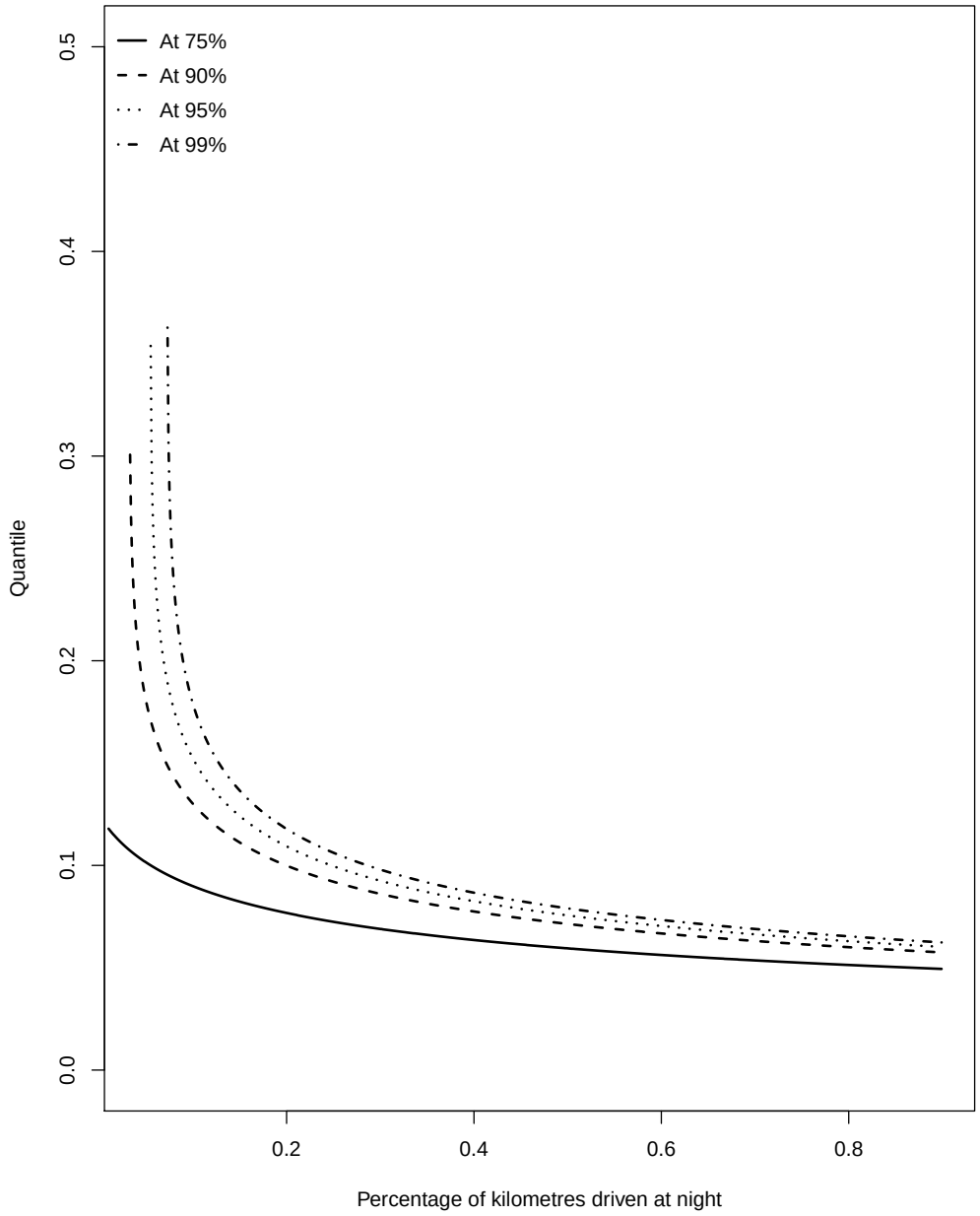

Figure 1: Quantiles of percentage of kilometres driven over the speed limit (Y_1) in the y-axis for Profile 1 given the values of percentage of kilometres driven at night (Y_2) in the x-axis.

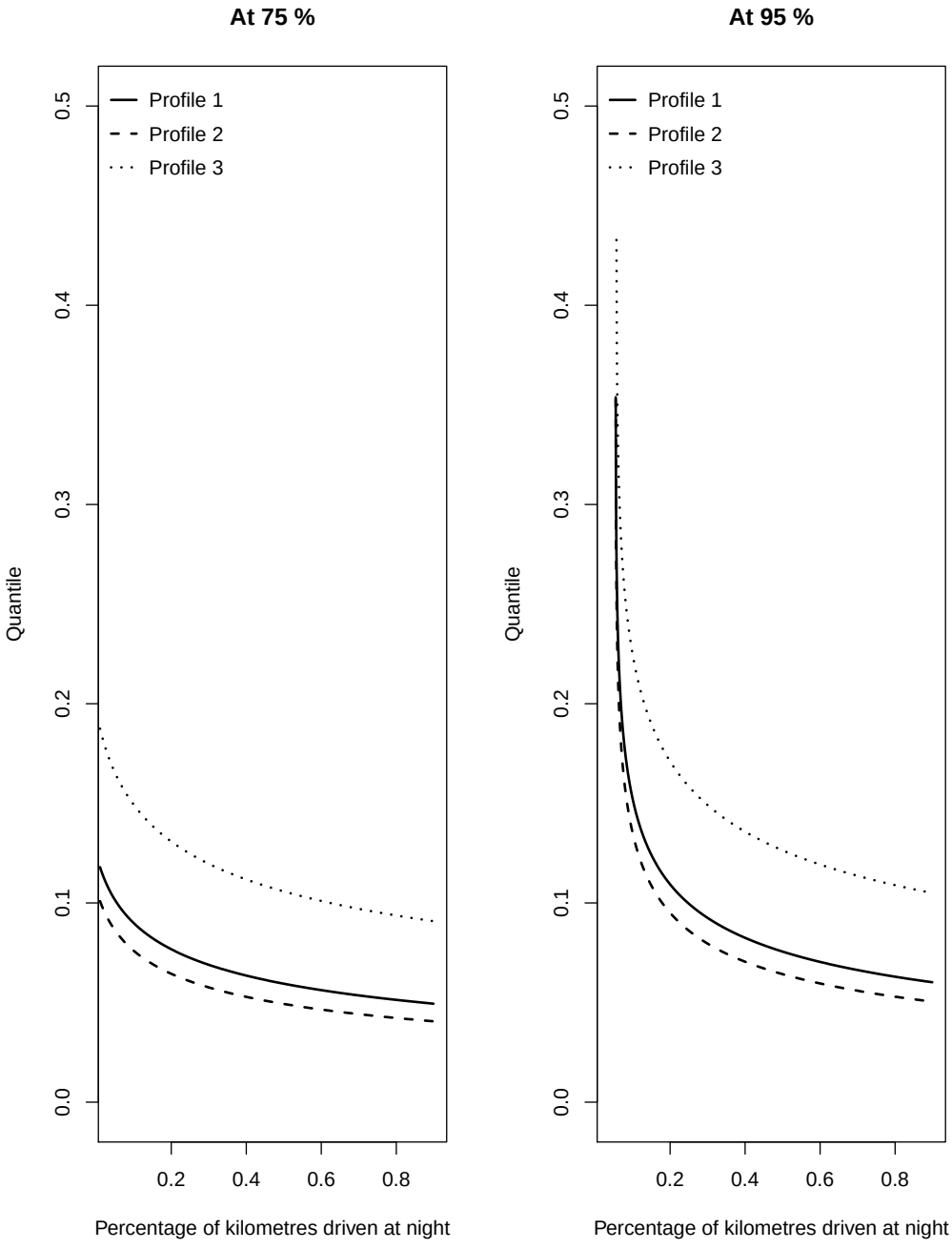

Figure 2: Quantiles of percentage of kilometres driven over the speed limit (Y_1) for each driver profile given the values of percentage of kilometres driven at night (Y_2), (**left**) 75% level and (**right**) 95% level.

4 Conclusions

We have developed a bivariate model based on the Sarmanov distribution with marginal Beta GLM which has allowed us to model two important variables in modern motor insurance telematics databases. Our model is an alternative to a proposal previously made by Gupta and Wong (1985) based on what is known as Morgenstern's distribution, which is a particular case of the Sarmanov distribution. Our proposal allows for obtaining closed expressions for some magnitudes of interest, such as the bivariate cdf and conditioned moments, covariance and correlation, which are fundamental in risk analysis. We have shown that our Sarmanov-Beta-GLM model presents better fits than previous proposals also based on the Sarmanov distribution.

The results of our case study have shown that, for a specific example, although the dependence parameter is positive, which directly implies that, in the mean, the relationship between the conditioned mean and the values of the variable that conditions is positive, the conditional quantiles show that the relationship between the conditioned quantile, and the value of the conditioning variable may be negative for high quantile levels, a result that is consistent with the expected behaviour of drivers.

Conclusions

The objective of the thesis is to look more closely at the methodology of quantile regression and its generalizations in the analysis of insurance work. For this, we use the telematics data, a type of data that is being used more and more in the world of insurance because it allows information to be obtained, practically in real time. The study proposes a series of solutions for analysing of the influence of various accident risk factors when driving based on quantile regression, which allows a quantile of the distribution of the variable of interest to be studied in a precise way. To expand on the study of risk, other methodologies are also applied, such as bivariate Sarmanov distribution, which allows two risk factors to be studied in the same distribution.

The first chapter presents the foundations for the study of the thesis. I noted how the results obtained to adjust the VaR (Value at Risk) and an approximation for the TVaR (Tail Value at Risk) were not the same if we used a linear regression or quantile regression when fitting the model. When studying the risk based on the number of kilometers driven over the speed limit, I determined the most important variable to be the total number of kilometres driven during the year. In the case of models fitted to TVaR, I noted that I could set the average kilometres driven over the speed limit for a specific level based on drivers' driving patterns. If this level is close to the average or even just exceeds it, this driver can be identified as risky. If a pricing model such as pay how you drive were used, the insurance company could add a premium to the price of the insurance because the driver falls into the risk category.

In the second chapter, I fit the TVaR instead of the approximation from the previous chapter using quantile regression. Here I consider a linear relation between the explanatory variables, which produce some mistakes in the prediction of the risk measures. For example, the value of the predictions appear in a range that is not feasible with some predictions for the TVaR lower than for the VaR, which is not possible by definition. To correct these mistakes, in the third chapter I propose the use of an additive model for both risk measures. The model for the TVaR turns out to be a very powerful tool for identifying drivers with a high risk of traffic accident through the number of kilometres driven above the speed limit. It is also a really easy model to implement. This type of model allows the different types of insured

customers to be grouped and gives a better profile of drivers in the portfolio. Many of the scientific papers that use this methodology do not share the same notation. Some of them consider that high risk values are the highest of the distribution while others consider that risky values are the lower ones. In our case I propose a common notation, which is more natural for the analysis of accident cost or loss models, where the worst profiles are located on the right side of the distribution.

In the fourth chapter I propose a method to interpolate β coefficients of the quantile regression. This methodology is useful to speed up the process of scoring the risk of the drivers based on quantile regression by forfeiting some accuracy in the predictions. This method is based on adjusting a reduced number of regressions and, considering that the coefficients estimated follow a linear increase, performing an interpolation between the estimations. In this chapter, I find it really important to determine correctly at which quantiles the quantile regression is adjusted depending on the distribution of the response variable. I also patent that it must be known how many regressions are adjusted. The more regressions, the better the predictions will be but, the computational time will rise. In this chapter I achieve a balance where the computational time is very low and the predictions obtained are sufficiently accurate. Thus, with the proposed process, I manage to establish an agile and reasonably accurate method that can be implemented in real life cases with few calculus difficulties.

In the final chapter, I look at different methodologies apart from quantile regression. Here I use a Bivariate Sarmanov distribution that allows two risk factors to be studied at the same time, depending on how are distributed. In this chapter, I use as risk factors the number of kilometres driven above the speed limit and the percentage of kilometres driven at night time. I assume that both risk factors follow a marginal beta distribution. There are other studies that use this type of distribution, but the model adjusted in this chapter shows better results that the ones obtained in previous approximations. The adjusted model allows expressions to be obtained for some interesting measures, such as the conditional moments, covariance, correlation and the cumulative distribution function of both response variables. Despite the positive dependency of the parameters used in the study, the conditional quantiles show that the relation between the conditional quantile and the value of the conditional variable is negative for upper quantiles which, although this is an expected result, is novel in being able to obtain bivariate risk measures not conditional on other covariates.

In this thesis, the application of quantile regression to study which driver characteristics affect the risk factor, such as speeding has been studied in depth. I start with a model that approximated TVaR, adjust an additive model to study two risk measures and develop a methodology that allows drivers to be rated depending on

their risk of having a traffic accident, using their driving patterns. I also widen the study and apply other methodologies to study other risk factors not related to speeding. I develop agile methods that can be adapted to the huge volume of data that insurance companies use nowadays.

Chapters one, two and three were the first ones to be written. Then chapter five was written and finally chapter four is the most recently written one, opening up a number of future areas of study. The writing of the final chapter corresponds with my incorporation into an insurance company, which gave me a more in depth view of how these data are treated and enabled me to evaluate possible limitations more closely.

This thesis, apart from providing methodologies to study risk measures beyond classical VaR, opens up new research lines that will allow risk studies to be carried out in a more precise and comprehensive way.

In the second and third chapters of the thesis, I adjust a quantile regression to the TVaR that is not a direct quantile. In this way, I prove that it is possible to adjust more complex risk measures, even considering that it can be expressed conditional on the covariates. A new possible study, starting from the same point as this study, could be to adjust other risk measures apart from the TVaR. In addition modelling methodologies other than quantile regression could be adapted to adjust complex risk measures, for example, changing the loss function used on the adjustment.

In the fourth chapter of the thesis, I develop a methodology that approximates variable coefficients for the quantiles where an adjustment is not made through an interpolation that considers that the effects change constantly as the quantile increases. As shown by the evolution of the effects graph, the increase or decrease of the coefficients is not lineal but neither is this always the case. I note that, mainly for extreme quantiles this variation can be more volatile, and this is the case in the study analysed in this chapter. Finding another way to approximate the values of the coefficients that are not approximated using quantile regression would allow to solve the detected approximation problems and even reduce the required number of regressions to interpolate all the coefficients.

Related to this number of necessary coefficients, other interesting topics appear. In the paper, I study how the distribution function of the response variable was varying and, depending on the growth rate, I adjust a higher number of regressions where the distribution vary more abruptly and less where the increase is linear. I establish three intervals of quantiles where the increases are different and for each interval I adjust a different number of regressions equally spaced. Another possible study could be to find a methodology that allows the number of quantiles that provides an optimum approximation to be determined and for which quantiles the quantile regression should be adjusted. A possible way to do this could be study-

ing the derivative function of the response variable distribution and selecting the adjustment points, depending on the value of the derivative.

Moving away from the methodological part, I find a broad range of possibilities for the application of the processes to other factors. For motor insurance, I focused on the risk factor of the number of kilometres driven above the speed limit but there are other factors that affect the risk of having a traffic accident, such as sudden braking or accelerating. It would be interesting to see the results when applying the methodology proposed in the thesis to other risk factors, even those of a different nature, such as sudden acceleration. In the latter, this is a recount variable that counts discrete values in contrast to the variables used in this study that are of a continuous nature.

Moving on to a completely different field outside of the world of insurance, it would be interesting to apply quantile regression in other disciplines, in particular the field of healthcare. An example of this could be in the study of experimental medicine when treating certain diseases by detecting those factors that affect the mortality risk or the risk of developing health difficulties or severe secondary effects, particularly when the response variable is of a continuous nature, such as, certain clinical analyses. Not only could quantile regression be applied in this type of study but also, more advanced methodologies that enable the identification of the mean effect of a normal dose and also its influence on the response quantile. For example, let us assume that dispensing a standard dose of a drug increases the expected value of a parameter, such as glucose. The idea is rather than studying how the medicine affects patients on the mean, to study patients that already have a high level of glucose, for example at quantile 95%. Moreover, in this type of field, it is common to study the evolution of patients, dividing them not only by gender but also creating different groups and having a control group. Developing a new methodology that allows the results of quantile regressions between groups to be compared could provide very interesting results that consider scenarios with high risk. This could become a very promising area of study in years to come.

Bibliography

Aarts, L. and Van Schagen, I. (2006). Driving speed and the risk of road crashes: A review. *Accident Analysis & Prevention*, 38(2):215–224.

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

Angrist, J., Chernozhukov, V., and Fernández-Val, I. (2006). Quantile regression under misspecification, with an application to the us wage structure. *Econometrica*, 74(2):539–563.

Arnold, B. and Ng, H. (2011). Flexible bivariate beta distributions. *Journal of Multivariate Analysis*, 102(8):1194–1202.

Ayuso, M., Guillen, M., and Nielsen, J. (2019). Improving automobile insurance ratemaking using telematics: incorporating mileage and driver behaviour data. *Transportation*, 46(3):735–752.

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

Bahraoui, Z., Bolancé, C., Pelican, E., and Vernic, R. (2015). On the bivariate sarmanov distribution and copula. an application on insurance data using truncated marginal distributi. *SORT*, 39(2):209–230.

Baione, F. and Biancalana, D. (2021). An application of parametric quantile regression to extend the two-stage quantile regression for ratemaking. *Scandinavian Actuarial Journal*, 2021(2):156–170.

Bairamov, I., Altinsoy, B., and Kerns, J. (2011). On generalized sarmanov bivariate distributions. *TWMS Journal of Applied and Engineering Mathematics*, 1(1):86–97.

Behr, A. (2010). Quantile regression for robust bank efficiency score estimation. *European Journal of Operational Research*, 200(2):568–581.

Bolancé, C. and Vernic, R. (2019). Multivariate count data generalized linear models: Three approaches based on the sarmanov distribution. *Insurance: Mathematics and Economics*, 85:89–103.

Boucher, J., Côté, S., and Guillen, M. (2017). Exposure as duration and distance in telematics motor insurance using generalized additive models. *Risks*, 5:54.

Briollais, L. and Durrieu, G. (2014). Application of quantile regression to recent genetic and-omic studies. *Human genetics*, 133(8):951–966.

Chen, L. and Zhou, Y. (2020). Quantile regression in big data: A divide and conquer based strategy. *Computational Statistics & Data Analysis*, 144:106892.

Chernick, M. (2011). *Bootstrap methods: A guide for practitioners and researchers*, volume 619. John Wiley & Sons.

Chernozhukov, V., Fernández-Val, I., and Melly, B. (2020). Fast algorithms for the quantile regression process. *Empirical economics*, pages 1–27.

Coad, A. and Rao, R. (2006). Innovation and market value: a quantile regression analysis. *Economics Bulletin*, 15(13).

Daniel-Spiegel, E., Weiner, E., Yarom, I., Doveh, E., Friedman, P., Cohen, A., and Shalev, E. (2013). Establishment of fetal biometric charts using quantile regression analysis. *Journal of Ultrasound in Medicine*, 32(1):23–33.

Dimitriadis, T. and Bayer, S. (2019). A joint quantile and expected shortfall regression framework. *Electronic Journal of Statistics*, 13(1):1823–1871.

Dimitriadis, T. and Schnaitmann, J. (2019). Forecast encompassing tests for the expected shortfall. *Available at SSRN 3436728*.

Efron, B. and Tibshirani, R. (1994). *An introduction to the bootstrap*. CRC press.

El Mazouri, F., Abounaima, M., and Zenkouar, K. (2019). Data mining combined to the multicriteria decision analysis for the improvement of road safety: case of france. *Journal of Big Data*, 6:5.

Eling, M., Jung, K., and Shim, J. (2022). Unraveling heterogeneity in cyber risks using quantile regressions. *Insurance: Mathematics and Economics*, 104:222–242.

Elliott, M., Armitage, C., and Baughan, C. (2003). Drivers' compliance with speed limits: an application of the theory of planned behavior. *Journal of Applied Psychology*, 88(5):964.

Fattouh, B., Scaramozzino, P., and Harris, L. (2005). Capital structure in south korea: a quantile regression approach. *Journal of Development Economics*, 76(1):231–250.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815.

Fissler, T. and Ziegel, J. (2016). Higher order elicitability and osband's principle. *The Annals of Statistics*, 44(4):1680–1707.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762.

Gohardehi, S., Sharif, M., Sarvi, S., Moosazadeh, M., Alizadeh-Navaei, R., Hosseini, S., Amouei, A., Pagheh, A., Sadeghi, M., and Daryani, A. (2018). The potential risk of toxoplasmosis for traffic accidents: A systematic review and meta-analysis. *Experimental parasitology*, 191:19–24.

Guillen, M., Nielsen, J., Ayuso, M., and Pérez-Marín, A. (2019a). The use of telematics devices to improve automobile insurance rates. *Risk analysis*, 39:662–672.

Guillen, M., Nielsen, J., Ayuso, M., and Pérez-Marín, A. (2019b). The use of telematics devices to improve automobile insurance rates. *Risk analysis*, 39(3):662–672.

Guillen, M., Nielsen, J. P., and Pérez-Marín, A. M. (2021). Near-miss telematics in motor insurance. *Journal of Risk and Insurance*.

Guillén, M., Pérez-Marín, A., and Alcañiz, M. (2020). Percentile charts for speeding based on telematics information. *Accident Analysis & Prevention*, 150:105865.

Gupta, A. and Wong, C. (1985). On three and five parameter bivariate beta distributions. *Metrika*, 32(1):85–91.

Hardy, M. (2006). An introduction to risk measures for actuarial applications. *SOA Syllabus Study Note*, 19.

Heras, A., Moreno, I., and Vilar-Zanón, J. (2018). An application of two-stage quantile regression to insurance ratemaking. *Scandinavian Actuarial Journal*, 2018(9):753–769.

Hesterberg, T. (2011). Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(6):497–526.

Huang, Y. and Meng, S. (2019). Automobile insurance classification ratemaking based on telematics driving data. *Decision Support Systems*, 127:113156.

Hung, W., Shang, J., and Wang, F. (2010). Pricing determinants in the hotel industry: Quantile regression analysis. *International Journal of Hospitality Management*, 29(3):378–384.

Huppert, D., Straube, A., Albers, L., von Kries, R., and Obermeier, V. (2019). Risk of traffic accidents after onset of vestibular disease assessed with a surrogate marker. *Journal of neurology*, 266:3–8.

Joe, H. and Xu, J. (1996). The estimation method of inference functions for margins for multivariate models.

Kaza, N. (2010). Understanding the spectrum of residential energy consumption: A quantile regression approach. *Energy policy*, 38(11):6574–6585.

Koenker, R. (2017). Quantile regression: 40 years on. *Annual Review of Economics*, 9:155–176.

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, 46:33–50.

Koenker, R. and Bassett Jr, G. (1982). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica: Journal of the Econometric Society*, 50:43–61.

Koenker, R. and Geling, O. (2001). Reappraising medfly longevity: a quantile regression survival analysis. *Journal of the American Statistical Association*, 96(454):458–468.

Koenker, R. and Machado, J. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94:1296–1310.

Koenker, R., Portnoy, S., Ng, P., Zeileis, A., Grosjean, P., and Ripley, B. (2018). Package 'quantreg'. *Cran R-project. org*.

Kou, S. and Peng, X. (2016). On the measurement of economic tail risk. *Operations Research*, 64(5):1056–1072.

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

Li, H., Song, Q., and Su, J. (2021). Robust estimates of insurance misrepresentation through kernel quantile regression mixtures. *Journal of Risk and Insurance*, 88(3):625–663.

Liao, W. and Wang, X. (2012). Hedonic house prices and spatial quantile regression. *Journal of Housing Economics*, 21(1):16–27.

Lourenço, H., Martin, O., and Stützle, T. (2003). Iterated local search. In *Handbook of metaheuristics*, pages 320–353. Springer.

Lu, J., Xing, Y., Wang, C., and Cai, X. (2016). Risk factors affecting the severity of traffic accidents at shanghai river-crossing tunnel. *Traffic injury prevention*, 17:176–180.

Mao, X., Yuan, C., Gan, J., and Zhang, S. (2019). Risk factors affecting traffic accidents at urban weaving sections: Evidence from china. *International journal of environmental research and public health*, 16:1542.

Marrocu, E., Paci, R., and Zara, A. (2015). Micro-economic determinants of tourist expenditure: A quantile regression approach. *Tourism Management*, 50:13–30.

Martins, P. and Pereira, P. (2004). Does education reduce wage inequality? quantile regression evidence from 16 countries. *Labour economics*, 11(3):355–371.

Matsuoka, E. and Saji, M. Kanemoto, K. (2019). Daytime sleepiness in epilepsy patients with special attention to traffic accidents. *Seizure*, 69:279–282.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman Hall/CRC.

Melly, B. (2005). Decomposition of differences in distribution using quantile regression. *Labour economics*, 12(4):577–590.

Morgenstern, D. (1956). Einfache beispiele zweidimen-sionaler verteilungen. *Mitteilingsblatt Math. Stat.*, 8:234–235.

Nadarajah, S., Zhang, B., and Chan, S. (2014). Estimation methods for expected shortfall. *Quantitative Finance*, 14(2):271–291.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Niemierko, R., Töppel, J., and Tränkler, T. (2019). A d-vine copula quantile regression approach for the prediction of residential heating energy consumption based on historical data. *Applied energy*, 233:691–708.

Nortey, E., Pometsey, R., Asiedu, L., Iddi, S., and Mettle, F. (2021). Anomaly detection in health insurance claims using bayesian quantile regression. *International Journal of Mathematics and Mathematical Sciences*, 2021.

Olkin, I. and Liu, R. (2003). A bivariate beta distribution. *Statistics & Probability Letters*, 62(4):407–412.

Olkin, I. and Trikalinos, T. (2015). Constructions for a bivariate beta distribution. *Statistics & Probability Letters*, 96:54–60.

Pai, J., Li, Y., Yang, A., and Li, C. (2022). Earthquake parametric insurance with bayesian spatial quantile regression. *Insurance: Mathematics and Economics*, 106:1–12.

Pérez-Marín, A., Ayuso, M., and Guillen, M. (2019a). Do young insured drivers slow down after suffering an accident? *Transportation research part F: traffic psychology and behaviour*, 62:690–699.

Pérez-Marín, A. and Guillen, M. (2019). Semi-autonomous vehicles: Usage-based data evidences of what could be expected from eliminating speed limit violations. *Accident Analysis & Prevention*, 123:99–106.

Pérez-Marín, A., Guillen, M., Alcañiz, M., and Bermúdez, L. (2019b). Quantile regression with telematics information to assess the risk of driving above the posted speed limit. *Risks*, 7:80.

Pesantez-Narvaez, J., Guillen, M., and Alcañiz, M. (2019). Predicting motor insurance claims using telematics data—xgboost versus logistic regression. *Risks*, 7(2):70.

Pitarque, A., Pérez Marín, A., and Guillen, M. (2019). Regresión cuantílica como punto de partida en los modelos predictivos para el riesgo. *Anales del Instituto de Actuarios Españoles*, 4:77–117.

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

Qi, B., Liu, P., Ji, T., Zhao, W., and Banerjee, S. (2018). Drivaid: Augmenting driving analytics with multi-modal information. In *2018 IEEE Vehicular Networking Conference (VNC)*, pages 1–8. IEEE.

Rovšek, V., Batista, M., and Bogunović, B. (2017). Identifying the key risk factors of traffic accident injury severity on slovenian roads using a non-parametric classification tree. *Transport*, 32:272–281.

Sarmanov, O. (1966). Generalized normal correlation and two-dimensional frechet classes. *Doclady Soviet Math*, 168:596–599.

Shepelev, V., Aliukov, S., Glushkov, A., and Shabiev, S. (2020). Identification of distinguishing characteristics of intersections based on statistical analysis and data from video cameras. *Journal of Big Data*, 7:1–23.

Singh, S. K. (2017). Road traffic accidents in india: issues and challenges. *Transportation research procedia*, 25:4708–4719.

Smith, A. (2016). A uk survey of driving behaviour, fatigue, risk taking and road traffic accidents. *BMJ open*, 6.

Smithson, M. and Verkuilen, J. (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological methods*, 11(1):54.

Sun, S., Bi, J., Guillen, M., and Pérez-Marín, A. (2020). Assessing driving risk using internet of vehicles data: an analysis based on generalized linear models. *Sensors*, 20:2712.

Sun, S., Bi, J., Guillen, M., and Pérez-Marín, A. (2021). Driving risk assessment using near-miss events based on panel poisson regression and panel negative binomial regression. *Entropy*, 23(7):829.

Tareghian, R. and Rasmussen, P. (2013). Statistical downscaling of precipitation using quantile regression. *Journal of hydrology*, 487:122–135.

Taylor, J. (2019). Forecasting value at risk and expected shortfall using a semi-parametric approach based on the asymmetric laplace distribution. *Journal of Business & Economic Statistics*, 37(1):121–133.

Ting Lee, M. (1996). Properties and applications of the sarmanov family of bivariate distributions. *Communications in Statistics-Theory and Methods*, 25(6):1207–1222.

Uribe, J. and Guillen, M. (2020). *Quantile Regression for Cross-Sectional and Time Series Data Applications in Energy Markets Using R*. Springer.

Valenzuela, C., Valencia, A., White, S., Jordan, J., Cano, S., Keating, J., Nagorski, J., and Potter, L. (2014). An analysis of monthly household energy consumption among single-family residences in texas, 2010. *Energy Policy*, 69:263–272.

Wang, R. and Ziegel, J. (2015). Elicitable distortion risk measures: A concise proof. *Statistics & Probability Letters*, 100:172–175.

Weidner, W., Transchel, F. W., and Weidner, R. (2016). Classification of scale-sensitive telematic observables for riskindividual pricing. *European Actuarial Journal*, 6:3–24.

Yu, J., Sumner, D., and Lee, H. (2021). Premium rates and selection in specialty crop insurance markets: Evidence from the catastrophic coverage participation. *Food Policy*, 101:102079.