*3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)*
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11512

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

**Albert Pitarque, Montserrat Guillen**
Riskcenter UB, Spain.

***Abstract***

*An algorithm to fit regression models aimed at predicted the average responses beyond a conditional quantile level is presented. This procedure is implemented in a case study of insured drivers covering almost 10,000. The aim is to predict the expected yearly distance driven above the posted speed limits as a function of driving patterns such as total distance, urban and night percent driven. Gender and age are also controlled. Results are analyzed for the median and the top decile. The conclusions provide evidence of factors influencing speed limit violations for risky drivers and they are interesting to price motor insurance and implement road safety policies. The efficiency of the algorithm to fit tail expectation regression is compared to quantile regression. Computational time doubles for tail expectation regression compared to quantile regression. Standard errors are estimated via bootstrap methods. Further considerations regarding in-sample predictive performance are discussed. In particular, further restrictions should be imposed in the model specification to avoid prediction outside the plausible range.*

***Keywords:*** *Telematics; quantile regression; insurance; tail value-at-risk; traffic safety.*

## 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 possibility to obtain safety improvements on the road and to develop new ways to calculate motor insurance prices. The aim of this paper is to propose new models for risk analysis. We present an algorithm that allows adjusting regression models for the tail expectation that 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 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 $c_\tau$ the quantile at the level $\tau$ ($\tau$ between 0 and 1) of a variable response Y. By definition, the probability that Y is greater or equal to $c_\tau$ is equal to $\tau$. Quantiles are used in areas such as finance, insurance and risk analysis, where they are usually referred to as $\tau$ – Value at Risk ($VaR_\tau$). Another risk measure is the Expected Shortfall ($ES_\tau$) also known as Conditional Tail Expectation ($CTE_\tau$) or Tail Value at Risk ($TVaR_\tau$). It is defined as:

$$TVaR_\tau(Y) = E(Y|Y > c_\tau). \quad (1)$$

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

## 2. Methodology

The starting point for this work 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 it 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 it is implemented in open source software R.

A classical linear regression model is represented as follows:

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

where $Y_i$ is the response variable for the $i^{th}$ individual (i = 1, ... , n), $X_{ji}$ represents the value of the $i^{th}$ observation of explanatory variable j (j = 1, ... , k) and $\beta_j$ is the $j^{th}$ parameter. The $i^{th}$ linear predictor is defined as $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \ldots \beta_k X_{ki}$. The error term, $\varepsilon_i$, is the part of the response variable that cannot be explained by the covariates. Parameter $\beta_0$ is

known as 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. Model (1) is usually estimated by ordinary least squares (OLS), i.e. by minimizing the sum of squared residuals:

$$\hat{\beta} = arg \min_{\beta} \sum_{i=1}^{n} f_i(\beta)\,, \tag{3}$$

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 $\tau$ of the response equals a linear combination of the regressors:

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

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

$$\widehat{\beta^\tau} = arg \min_{\beta} \sum_{i=1}^{n} [\rho_i^\tau (Y_i - X_{ij}\beta_j)]. \tag{5}$$

where $\rho_i^\tau$ represents a loss function of the quantile, which is equal to $\tau$ when $Y_i - X_i\beta$ is greater or equal than 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:

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

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 studying the optimization underlying the estimation procedure is of outmost interest. 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 the 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 of kilometres driven |
| P_urban | Percentage of kilometres driven in urban areas |
| P_night | Percentage of kilometres driven during the night |
| Age | Age of the driver at 1st of January, 2010 |
| Male | Gender of the 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 of them present large values. However, it is necessary to consider total driving distance, urban driving and night driving to extract 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.

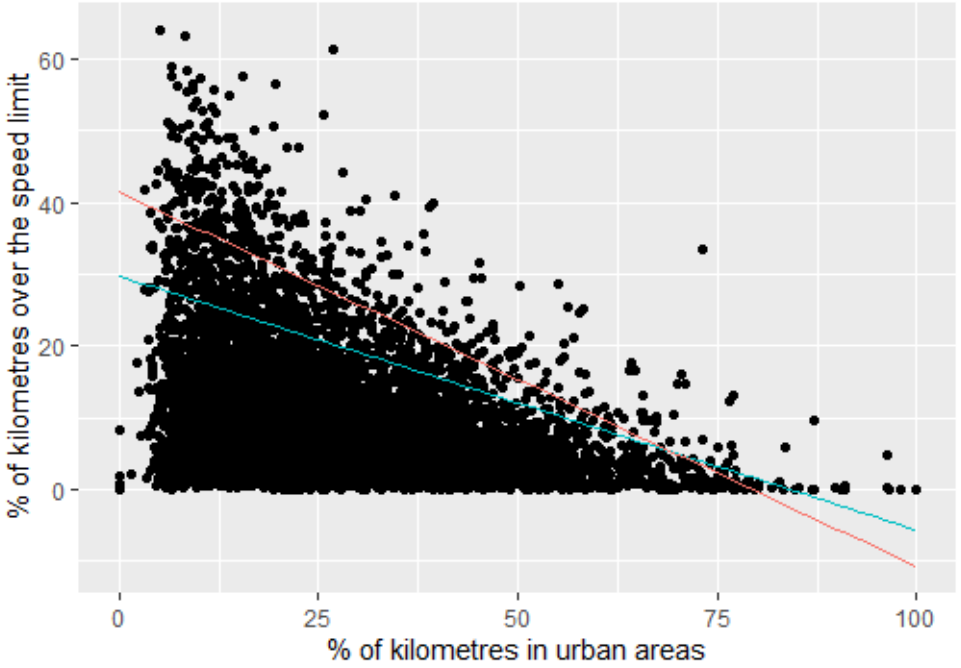


*Figure 1. Graph of the relation between the percentage of kilometres driven above the speed limit and the percentage of kilometres driven in urban areas in the insurance dataset. Blue line represents a 90% quantile regression line and red line represents the a 90% tail expectation regression.*

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 were computed from 3.000 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. Models results of linear regression (OLS), quantile regression (VaR) and tail expectation regression (TVaR) for quantile levels $\tau = 0.5$ and $\tau = 0.9$ in the insurance dataset. In parenthesis, the standard errors of the estimated coefficients.**

| Variable | OLS | $VaR_{0.5}$ | $TVaR_{0.5}$ | $VaR_{0.9}$ | $TVaR_{0.9}$ |
|---|---|---|---|---|---|
| Intercept | -8082.51 | -4496.53 | -11708.92 | -6418.11 | -14068.39 |
| | (309.95) | (186.02) | (843.57) | (742.98) | (3505.13) |
| lnKm | 1064.51 | 597.60 | 1588.38 | 1074.66 | 2229.62 |
| | (26.51) | (19.32) | (86.59) | (64.46) | (364.14) |
| P_urban | -21.87 | -9.19 | -39.72 | -39.59 | -86.08 |
| | (1.39) | (0.62) | (2.16) | (2.34) | (7.14) |
| P_night | 7.54 | 5.41 | 11.99 | 21.76 | 26.56 |
| | (2.93) | (1.82) | (6.10) | (9.80) | (19.21) |
| Age | -1.13 | -2.56 | 0.96 | 5.16 | 7.71 |
| | (6.26) | (3.26) | (11.09) | (15.24) | (37.13) |
| Male | 328.01 | 206.76 | 528.84 | 574.08 | 913.63 |
| | (35.89) | (19.01) | (66.51) | (103.97) | (223.48) |
| $R^2$ | 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 $\tau = 0.5$ and $\tau = 0.9$.**

| | |
|---|---|
| **% $TVaR_{0.5} < VaR_{0.5}$** | 8.20% |
| **% $TVaR_{0.5} < 0$** | 7.41% |
| **% $TVaR_{0.9} < VaR_{0.9}$** | 6.48% |
| **% $TVaR_{0.9} < 0$** | 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

An innovative method that generalizes quantile regression in order to study risky drivers was implemented. The study was done using a database containing 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 exceed speed limits. In particular, 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 52 basis points the expected percentage of kilometres driven above the speed limit, 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 those problems, the computational time of the estimation procedure to obtain the coefficient estimates is low, so the routine for the tail expectation regression that was 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. Specially with the bootstrap method, there are currently several possible alternatives

that seem suitable to our 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.

## References

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

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

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

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

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

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

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

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

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

Koenker, R., Portnoy, S., Ng, P. T., Zeileis, A., Grosjean, P., & Ripley, B. D. (2019). Package 'quantreg'. https://cran.r-project.org/web/packages/quantreg/

Pitarque, A., Pérez-Marín, A. M., & Guillen, M. (2019). Quantile regression as a starting point in predictive risk models. *Anales del Instituto de Actuarios Españoles, 4ª época*, 25, 2019 /101-117