TECHNICAL PAPER

Applying Quantile Regression to Ratemaking: A Measured Approach

Last update: July 2023

Contents

Relevant Products and Releases

- SAS® Dynamic Actuarial Modeling

Introduction

Our study focused on the insurance ratemaking process, specifically on modeling the frequency of claims, with an emphasis on incorporating telematics information in automobile insurance data. Telematics information can reveal crucial associations between drivers' behavior on the road and the number of claims incurred. The study used quantile regression to estimate the frequency of claims (Pérez-Marín et al. 2019; Kudryavtsev 2009), which determines the conditional median (or other quantiles) of the target variable, as opposed to its mean, as is commonly done in the widely used method of generalized linear modeling (Nelder and Wedderburn 1972). The comparison of these two methods forms the basis of multiple sections of this paper, which documents our study.

The study fully utilized the extensive ratemaking functionality of SAS® Dynamic Actuarial Modeling software. The first two sections of the paper discuss this software and introduce the data that we used for our analysis. The third section delves into the fundamental theory of quantile regression models and generalized linear models (Koenker and Bassett 1978; McCullagh and Nelder 1989). The paper then highlights the advantages of using quantile regression for the ratemaking process, such as in addressing heteroscedasticity or robustness to outliers. Next, a practical example demonstrates the modeling of claims frequency by using automobile insurance data. First, it compares the predictions that are produced by generalized linear modeling and the 0.5 quantile of quantile regression by using various measures of model quality. Then, it describes the practical application of higher quantiles to obtain predictions that incorporate safety loadings or identify high-risk drivers.

SAS® Dynamic Actuarial Modeling

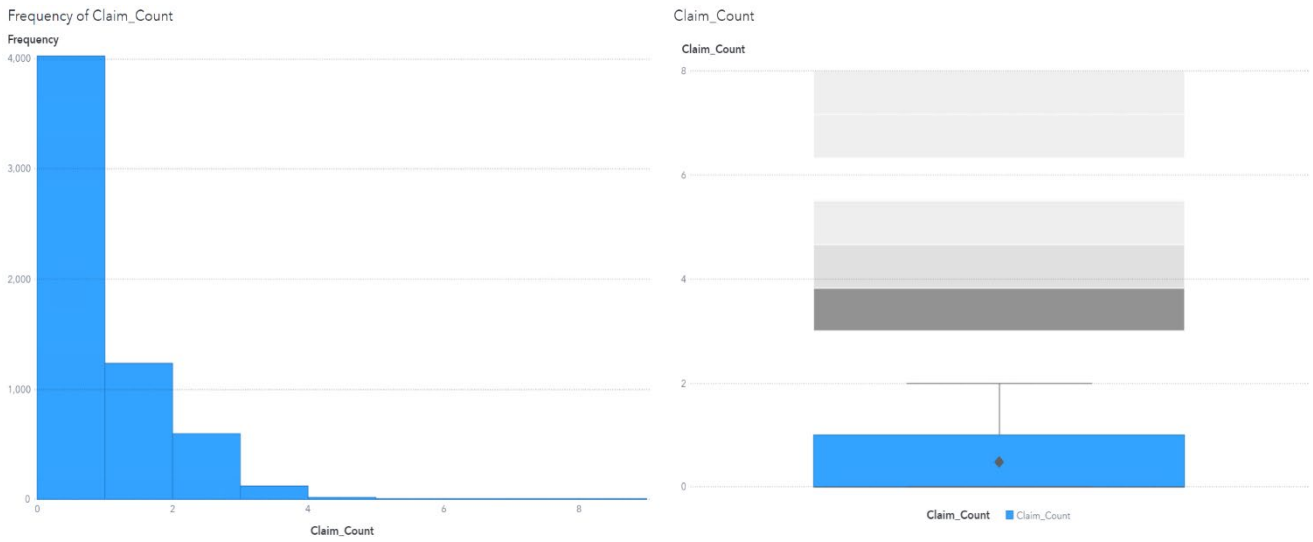
SAS Dynamic Actuarial Modeling software provides functionality to support end-to-end pricing in the insurance industry. It enables insurance companies to enhance their modeling agility and accuracy by leveraging industry-leading modeling and analytical capabilities. SAS Dynamic Actuarial Modeling delivers real-time quotations based on customizable model parameters and decision factors.

The first component of this software tool involves managing data and transforming the data set into the desired analytical base table (ABT) format, as well as assessing the quality of the data. This component is not presented in this paper; we commence directly with the prepared ABT. The second component is data exploration and visualization, which we use in the "Data Introduction" section to illustrate the characteristics and typical properties of insurance data. The third component involves constructing the desired model by performing variable transformations, grouping, and selecting the type of model and its parameters. This component is used in the "Practical Example" section of the paper. The last component of SAS Dynamic Actuarial Modeling software entails back-testing and implementing a pricing algorithm; this component was not included in our study.

Data Introduction

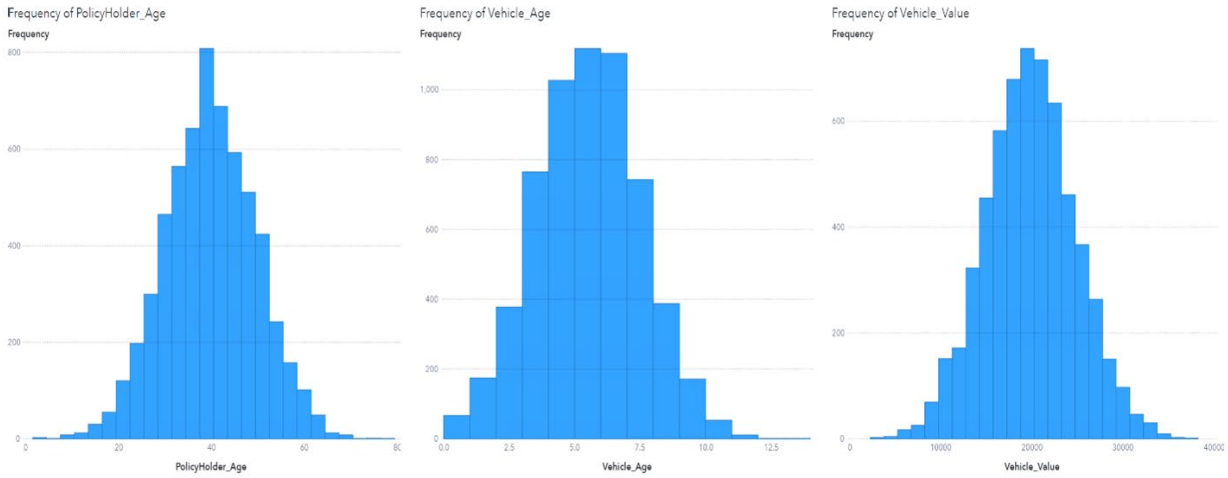
The data set that was used in the analysis contains 6,006 observations and describes a count of claims for automobile insurance. The count of claims is represented by the variable Claim_Count, which is the target variable that we want to explain. As shown in Figure 1, the plot of the target variable has a typical shape for a claim count, with a high proportion of zeros.

Figure 1. *Frequency Histogram (left) and Box Plot (right) of the Target Variable, Claim_Count*



The explanatory variables are divided into three groups. The first group, shown in Figure 2, contains general information about drivers and vehicles, including numerical variables such as PolicyHolder_Age, Vehicle_Age, and Vehicle_Value and the categorical variable Vehicle_Type.

Figure 2. *Frequency Distributions of the Input Variables PolicyHolder_Age, Vehicle_Age, and Vehicle_Value*



The second group of explanatory variables, shown in Figures 3 and 4, consists of telematics information that describes the behavior of drivers (Location, Harsh_Accel, Harsh_Brakes, Harsh_Lateral, Night_Driving). “Harsh” driving can indicate inexperience, as well as distracted or aggressive driving. As shown in Figure 3, the distribution of these numerical variables is often right-skewed. The scatter plots in Figure 4 reveal a positive relationship between the telematics variables and the target variable. The third group of explanatory variables consists of a single variable, Exposure_Amt, which is treated as an offset; it is not shown here.

Figure 3. *Frequency Distributions of the Input Variables Harsh_Accel, Harsh_Brakes, Harsh_Lateral, and Night_Driving*

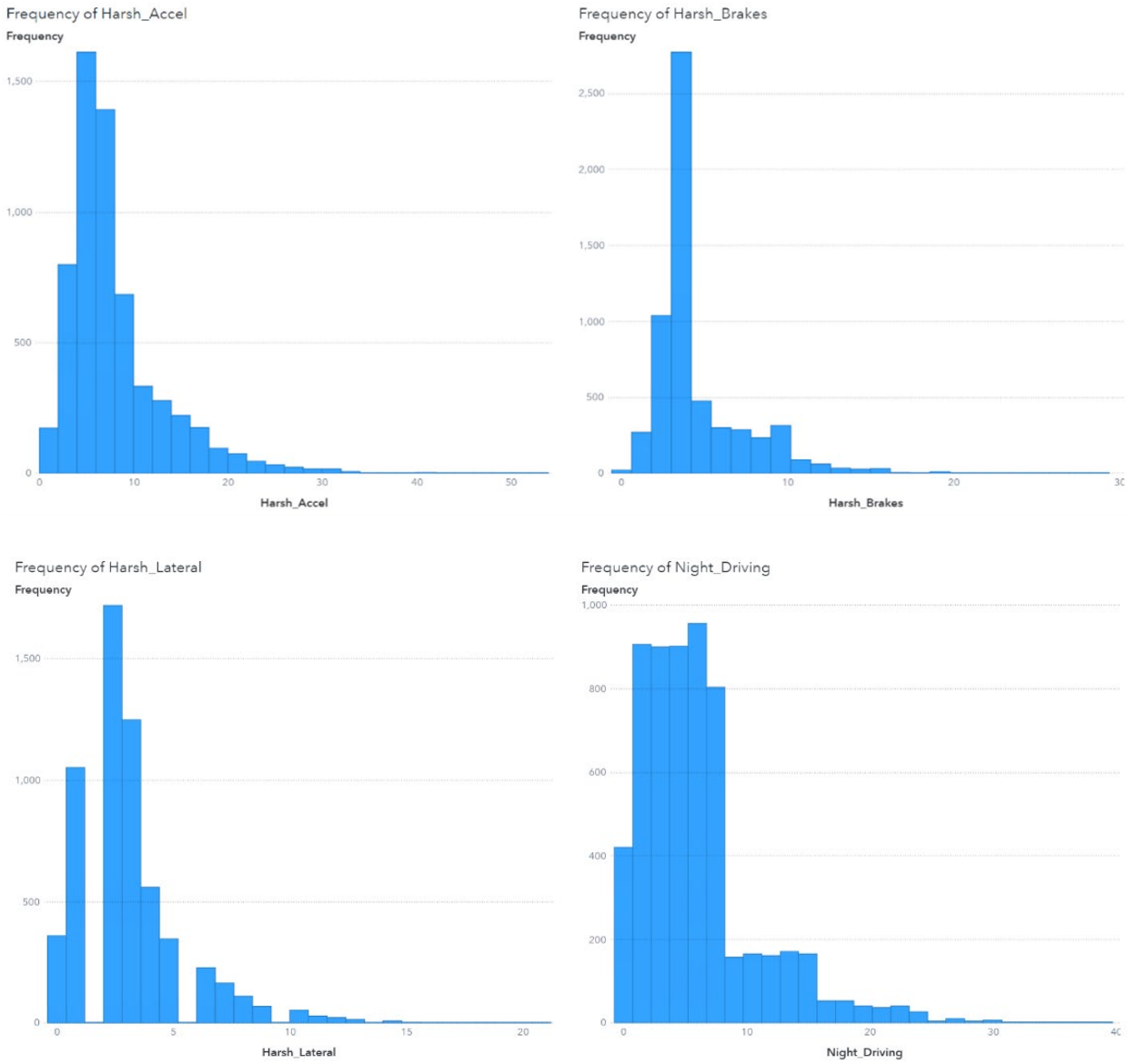


Figure 4. *Relationship between the Target Variable (Claim_Count) and the Telematics Input Variables (Harsh_Accel, Harsh_Brakes, Harsh_Lateral, and Night_Driving)*

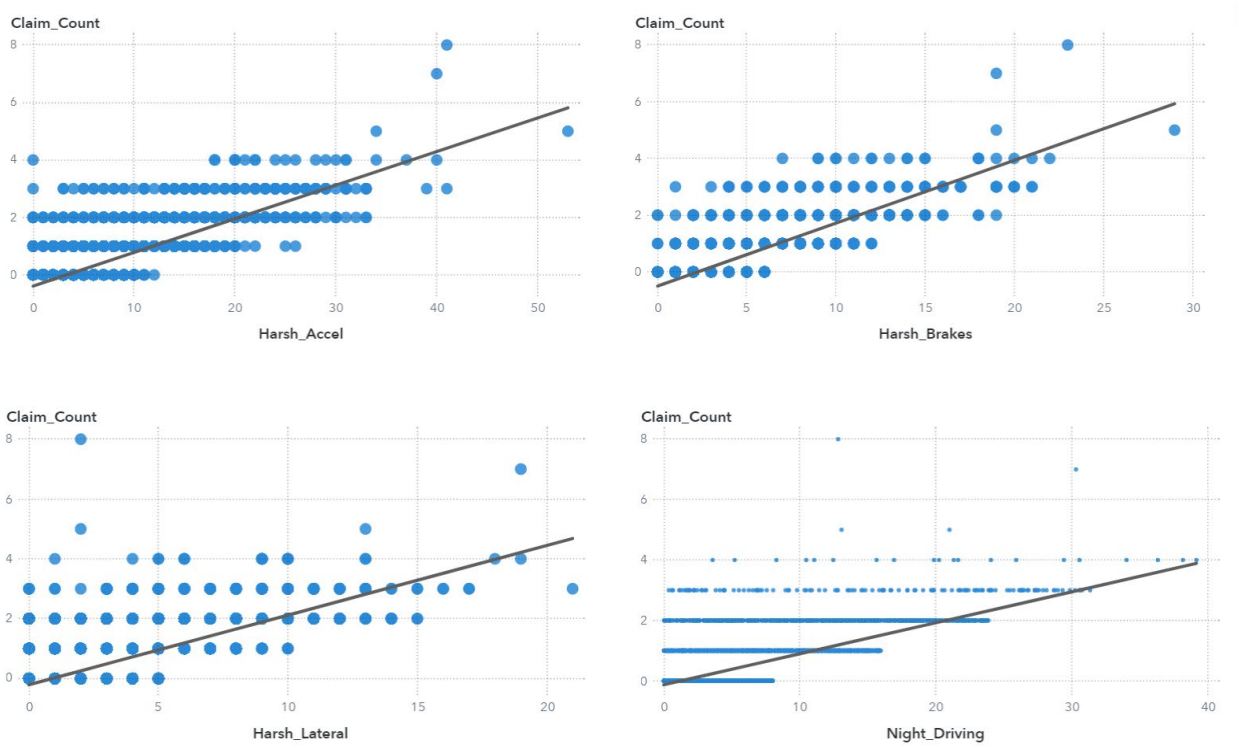


Table 1 shows the complete list of variables in the data set that was used in our study.

Table 1. *Input and Target Variables*

Variable	Description
Claim_Count	Number of claims per policyholder exposure
PolicyHolder_Age	Age of policyholder
Vehicle_Age	Age of insured car
Vehicle_Value	Value of vehicle
Vehicle_Type	Type of vehicle (SUV, sedan, sports car)
Location	Location of driving (urban, suburban, rural)

Variable	Description
Harsh_Accel	Harsh acceleration (occurs when a driver uses more power than necessary to proceed from a stop)
Harsh_Brakes	Harsh braking (occurs when a driver applies excessive force to stop a vehicle)
Harsh_Lateral	Lateral acceleration (occurs when there is a sudden or abrupt lateral movement of a vehicle)
Night_Driving	Number of kilometers driven at night
Exposure	Exposure amount (portion of the year during which the policyholder was covered, indicating whether the coverage was for the entire year or for a shorter period)

Basic Definitions

Generalized Linear Models

A generalized linear model is a generalization of ordinary linear regression. Generalized linear modeling enables the linear model to be related to the dependent variable through a link function. In fact, linear regression can be seen as a special case of generalized linear modeling in which the identity function is the link function, because the relationship between the predictor and the response is linear and requires no transformation.

Let $(y_i, \boldsymbol{x}_i)$ be a member of the set of observations $(i = 1, \dots, n)$, where y_i is a dependent variable in the regression equation and $\boldsymbol{x}_i = (x_{i1}, \dots, x_{im})$ is a row vector of independent variables.

In generalized linear modeling, the target variable y_i is modeled as a random variable that follows a probability distribution that belongs to the exponential family of distributions (such as normal, Poisson, and gamma). The model is given by the formula $E(y_i|\, \boldsymbol{x}_i) = \mu_i = g^{-1}(\boldsymbol{x_i\beta})$, where $E(y_i|\, \boldsymbol{x}_i)$ is the expected value of y_i conditional on $\boldsymbol{x}_i$ and $\boldsymbol{x_i\beta}$ is the linear predictor, a linear combination of the unknown parameter $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{m-1})$. The function $g(.)$, which is called the link function, provides the relationship between the linear predictors and the mean of the distribution function. The best estimates of regression coefficients are obtained by maximizing the likelihood function (Nelder and Wedderburn 1972). The usual choice for the link function in the ratemaking process is the natural logarithm (log-link function). The dependent variable is usually highly skewed to the right, and it follows, for example, the Poisson or negative binomial distribution for the number of claims or the gamma distribution for the severity of claims.

Quantile Regression

Let Y be a real-valued random variable with the cumulative distribution function $F_Y(y) = P(Y \leq y)$. The τth quantile of Y is given by $Q_\tau(Y) = F_Y^{-1}(\tau) = \inf\{y: F_Y(y) \geq \tau\}$, where $\tau \in (0, 1)$.

Let $(y_i, \boldsymbol{x}_i)$ be a member of the set of observations $(i = 1, \dots, n)$, where y_i is a dependent variable in the regression equation and $\boldsymbol{x}_i = (x_{i1}, \dots, x_{im})$ is a row vector of independent variables. Then the quantile regression model is

given by the formula $Q_\tau(y_i|\boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}^{(\tau)}$, indicating the conditional quantile of the random variable y_i for probability τ, provided that the vector of the regressors $\boldsymbol{x}_i$ and $\boldsymbol{\beta}^{(\tau)} = (\beta_0^{(\tau)}, \dots, \beta_{m-1}^{(\tau)})$ is the corresponding column vector of regression coefficients. The goal of quantile regression is to find the best estimation $\widehat{\boldsymbol{\beta}}^{(\tau)}$ of the vector $\boldsymbol{\beta}^{(\tau)}$. For a large enough n, the estimation could be obtained by solving the minimalization problem (Koenker and Bassett 1978):

$$\min_{\boldsymbol{\beta}^{(\tau)}} \frac{1}{n} \left\{ \sum_{i:y_i \geq x_i\beta^{(\tau)}} \tau \left| y_i - \boldsymbol{x_i}\boldsymbol{\beta}^{(\tau)} \right| + \sum_{i:y_i < x_i\beta^{(\tau)}} (1-\tau) \left| y_i - \boldsymbol{x_i}\boldsymbol{\beta}^{(\tau)} \right| \right\} \qquad (1)$$

For $\tau = 0.5$, the minimalization problem is called least absolute deviation regression or median regression (Karst 1958):

$$\min_{\boldsymbol{\beta}^{(\tau)}} \frac{1}{n} \left\{ \sum_i \frac{1}{2} \left| y_i - \boldsymbol{x_i}\boldsymbol{\beta}^{(\tau)} \right| \right\}$$

Advantages of Quantile Regression

Generalized linear modeling is a well-known approach that actuaries have used for many years, despite its limitations regarding the character of real insurance portfolios. However, most of these limitations can be eliminated by using quantile regression, which has a number of advantages over generalized linear modeling.

Handling Heterogeneity

In generalized linear models, heteroscedasticity is typically addressed by specifying an appropriate variance function or variance structure. This allows the model to account for the varying levels of dispersion in the response variable across different levels of the predictors. Generalized linear models are commonly used when the focus is on modeling the conditional mean of the response variable.

Quantile regression (QR), on the other hand, directly models the relationship between predictors and different quantiles of the response variable, not just the conditional mean. By estimating the conditional quantiles, QR provides information about the entire distribution of the response variable, including the variability at different points of the distribution. This makes QR particularly useful when you are dealing with heteroscedasticity and when you are interested in understanding how predictors influence different parts of the response distribution.

In the context of heteroscedasticity, the main difference between generalized linear modeling and quantile regression is that generalized linear modeling focuses on modeling the mean response and accommodating heteroscedasticity through variance functions, whereas quantile regression directly models the conditional quantiles and captures heteroscedasticity implicitly by estimating quantiles at different points of the response distribution.

Robustness to Outliers

Outliers, such as large or catastrophic losses, can have a significant impact on analysis that uses generalized linear modeling, because this method assumes the absence of outliers in the data (McCullagh and Nelder 1989). The

presence of outliers can lead to biased parameter estimates and inaccurate predictions, because the model becomes overly influenced by the extreme values (Rousseeuw and Leroy 1988).

Quantile regression, on the other hand, is more robust to the presence of outliers, because it focuses on estimating conditional quantiles rather than the mean (Koenker and Xiao 2003). QR is less sensitive to extreme values because it minimizes the sum of asymmetrically weighted absolute residuals, thus reducing the impact of outliers on the model (Huber 1981). This robustness to outliers enables QR to provide a more accurate representation of the underlying risk factors and the relationships between insured objects and their corresponding losses, even in the presence of extreme events (Koenker 2005).

Minimizing the Sum of Asymmetrically Weighted Absolute Residuals

Let's narrow our focus to the topic of minimizing the sum of asymmetrically weighted absolute residuals (Buchinsky 1998). The objective function of QR is formulated using a check function, $\rho\tau(\varepsilon)$, which assigns asymmetric weights to positive and negative residuals according to the quantile of interest, τ (see Formula 1). When you estimate the τth quantile, the check function assigns a weight of τ to the positive residuals and a weight of $(\tau - 1)$ to the negative residuals. This means that the weights change according to the quantile that is being estimated and according to the sign of the residual.

For example, when you estimate the median ($\tau = 0.5$), the check function assigns equal weights of 0.5 to both positive and negative residuals. This leads to a balanced minimization of the absolute residuals above and below the median. In contrast, when you estimate the upper quantiles, such as $\tau = 0.9$, the check function assigns a larger weight of 0.9 to positive residuals and a smaller weight of 0.1 to negative residuals. This reflects the focus on minimizing the residuals above the 0.9 quantile, which is the main objective of estimating an upper quantile. These asymmetric weights ensure that the QR model is robust to outliers, because the objective function is less sensitive to extreme values than the mean-based objective function that is used in least squares regression.

Additionally, QR's ability to model multiple quantiles of the conditional distribution enables analysts to examine the entire distribution of losses, providing valuable insights into the tail behavior and the potential for large losses in insurance portfolios. This robustness to outliers is particularly important for the insurance industry, because it helps analysts better understand and manage the risk exposure that is associated with catastrophic events, ultimately leading to more informed decision-making and risk management practices (Fahrmeir and Tutz 2001).

Independence Assumptions: The Flexibility of Quantile Regression

In statistical analysis, an assumption of independence between observations is often made to simplify modeling and analysis. However, real-world data are rarely completely independent, because there can be hidden dependencies or shared characteristics among observations. Traditional methods such as generalized linear modeling rely on the assumption of strict independence. However, quantile regression offers a unique advantage by providing a flexible framework that allows for the relaxation of the independence assumption (Koenker and Hallock 2001; Machado and Silva 2014).

Unlike other regression methods, quantile regression does not explicitly model the full dependence structure among observations. Instead, it focuses on estimating the conditional quantiles of the response variable on the basis of the predictor variables.

By estimating the conditional quantiles directly, quantile regression can capture the relationship between the predictors and different parts of the response distribution, even when there is a weaker form of dependence or

there are violations of the strict independence assumption. It does not rely heavily on the precise nature of the dependence structure among the observations.

This flexibility arises from the fact that quantile regression estimates the quantiles conditionally, meaning that it considers the relationship between the predictors and specific points of the response distribution. This enables it to capture the variability and patterns in different parts of the distribution, regardless of the exact dependence structure.

For example, in a scenario where there is some residual correlation or clustering among the observations, quantile regression can still provide valid estimates of the conditional quantiles. It can adapt to variations in the dependence structure and provide insights into how the predictors influence different parts of the response distribution.

Practical Example

Comparison of Generalized Linear Model and 0.5 Quantile Regression Model

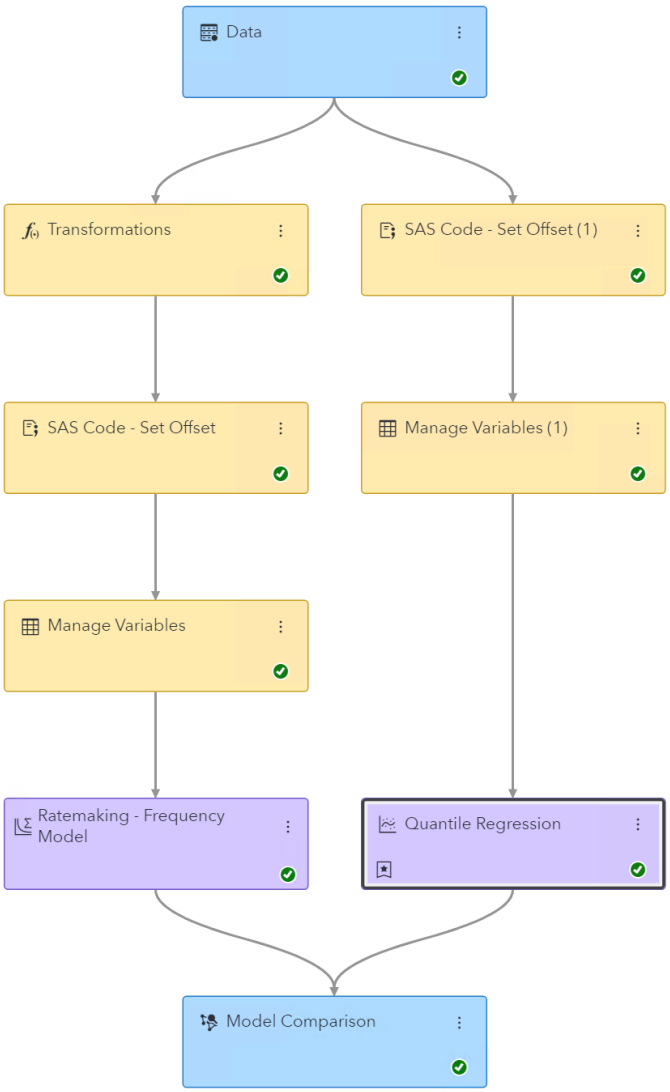
In this practical example, we use our automobile insurance data set and create a model to estimate the claim count for automobiles. Figure 5 lists the variables in the data set, along with their type, role, and level.

***Figure 5**. Input Variables and Target Variable for the Quantile Regression Model*

Variable Name ↑	Label	Type	Role	Assess for Bias	Level
Claim_Count		Numeric	Target		Interval
Exposure		Numeric	Input		Interval
Harsh_Accel		Numeric	Input		Interval
Harsh_Brakes		Numeric	Input		Interval
Harsh_Lateral		Numeric	Input		Interval
Location		Character	Input		Nominal
Night_Driving		Numeric	Input		Interval
PolicyHolder_Age		Numeric	Input		Interval
Vehicle_Age		Numeric	Input		Interval
Vehicle_Type		Character	Input		Nominal
Vehicle_Value		Numeric	Input		Interval

First, we compare the results of a generalized linear model that has a Poisson target distribution to a quantile regression model for quantile 0.5. Our pipeline for the models is shown in Figure 6.

***Figure 6**. Pipeline for Two Models: Generalized Linear Model and Quantile Regression Model*



The pipeline is divided into two subpipelines: one for ratemaking models and the other for quantile regression. The first nodes are for data mining and preprocessing. The Transformation node is used to transform variables according to our needs. This transformation is applied to the variable Exposure because it must be transformed in the same way as the target variable. The logarithmic Exposure variable is to be offset in the next node, the SAS Code node. In the quantile regression subpipeline, the role of offset is assigned to the original Exposure variable. The Manage Variables node provides final adjustments to the roles of variables (input, rejected, offset, and so on). Finally, we have two supervised learning nodes to compare. The first is called the Ratemaking – Frequency Modeling node. This model fits a parametric distribution model for frequency of loss data and is based on the generalized linear model. We chose the Poisson distribution as the distribution of the input variable. The second model is created using quantile regression with a 0.5 quantile.

Comparison of the two models, which is shown in Figure 7, reveals that quantile regression achieves the best fit. The selected criterion for determining the champion model is the lowest average squared error, but you can see that quantile regression wins for each possible model quality criterion. The second-best model is the Poisson generalized linear model, but its average squared error is more than eight times greater than that of the quantile regression model.

***Figure 7**. Model Quality Criterion Statistics for the Generalized Linear and Quantile Regression Models*

Model Comparison

Champion	Name	Algorithm Name	Average Squared Error
★	Quantile Regression	Quantile Regression	0.1047
	Ratemaking - Frequency Model	Count GLM	0.8965

Data Role	Number of Observations	Root Average Squared Error	Root Mean Absolute Error	Root Mean Squared Logarithmic E...
TEST	601	0.3235	0.4702	0.1930
TEST	601	0.9468	0.6464	0.3174

Both models have selected telematics input data as the data that have the greatest impact on the model. Other variables are not considered to be significant for our prediction. This choice was based on the stepwise selection method, which aims to improve the model's performance as measured by the Schwarz Bayesian criterion (SBC). The output shown in Figure 8 displays the results of quantile regression.

***Figure 8**. Significant Input Variables Chosen by the Stepwise Selection Method*

The SAS System

The QTRSELECT Procedure

Quantile Level = 0.5
Selection Details

Selection Summary			
Step	**Effect Entered**	**Number Effects In**	**SBC**
0	**Intercept**	1	-9927.649
1	**Harsh_Brakes**	2	-13220.967
2	**Harsh_Accel**	3	-14535.940
3	**Harsh_Lateral**	4	-15157.912
4	**Night_Driving**	5	-15587.209*
*** Optimal Value Of Criterion**			

Stepwise selection stopped because adding or removing an effect does not improve the SBC criterion.

The model at step 4 is selected where SBC is -15587.2.

Selected Effects: Intercept Harsh_Accel Harsh_Brakes Harsh_Lateral Night_Driving

Another important component of the quantile regression results is the table of parameter estimates, shown in Figure 9, which we are particularly interested in.

***Figure 9**. Parameter Estimates for the Quantile Regression Model's Input Variables*

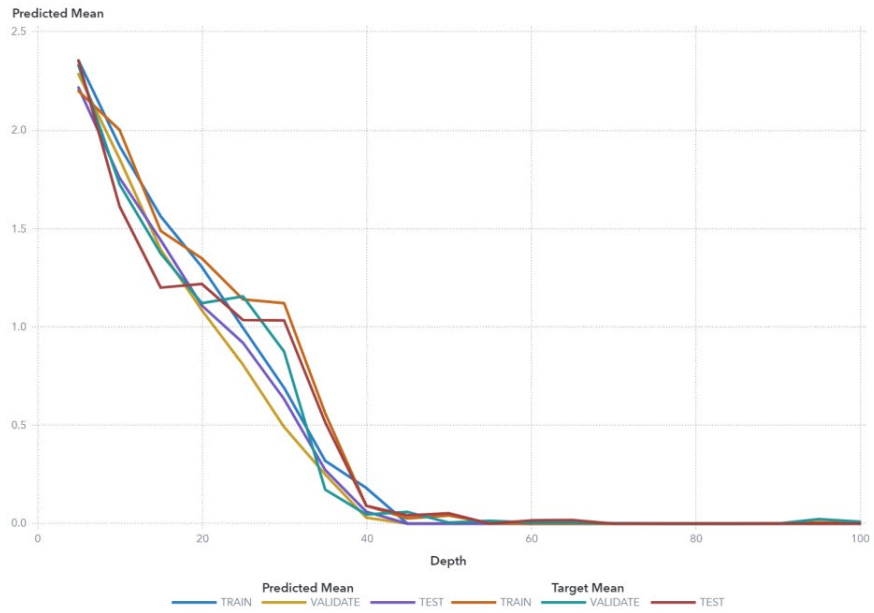
Parameter Estimates

Effect	Parameter	t Value	Sign	Estimate	Absolute Esti...	Standard Error	Pr > |t|	Degrees of Fr...
Intercept	Intercept	65.5526	-	-0.7368	0.7368	0.0112	0	1
Harsh_Brakes	Harsh_Brakes	31.5274	+	0.1023	0.1023	0.0032	0.0000	1
Harsh_Accel	Harsh_Accel	26.1498	+	0.0450	0.0450	0.0017	0.0000	1
Harsh_Lateral	Harsh_Lateral	21.1418	+	0.0728	0.0728	0.0034	0.0000	1
Night_Driving	Night_Driving	16.5127	+	0.0276	0.0276	0.0017	0.0000	1

We can observe from Figure 9 that all selected parameters are statistically significant according to their *p*-values. Harsh_Brakes is the most influential variable (excluding the Intercept), whereas Night_Driving has the least impact on the prediction.

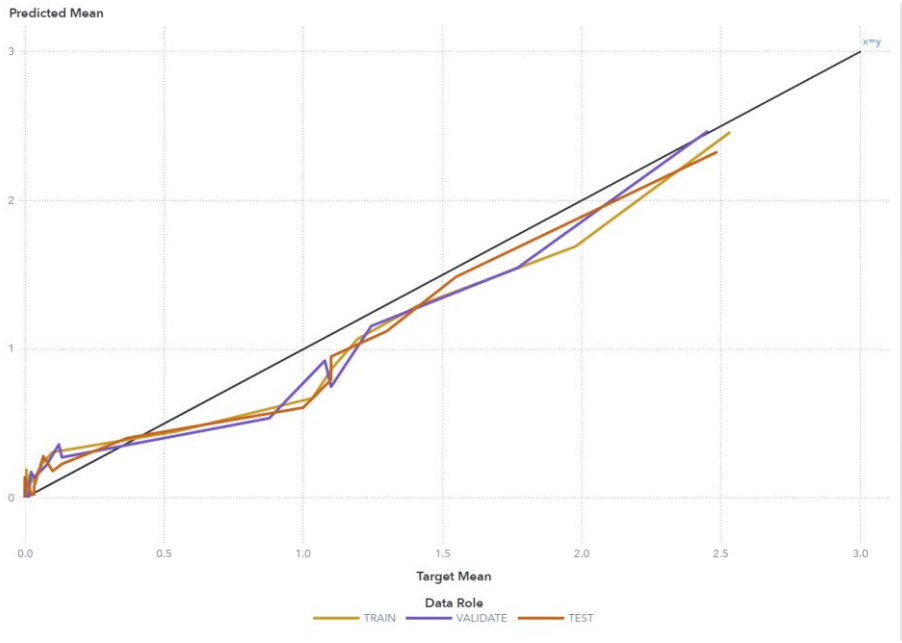
In addition, we analyzed a plot of actual and predicted variables by depth, shown in Figure 10. In this plot, each partition of the data is sorted in descending order by the predicted target variable, P_Claim_Count, for the actual target variable, Claim_Count. The data are then divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the means of the predicted target and actual target are calculated and plotted for each quantile (depth in increments of 5). The largest difference between the actual and predicted target means is 0.392; it occurs for the test partition at a depth of 30. The plot demonstrates that the predicted values of the target variable are relatively close to the actual values. For comparison, the largest difference between the actual and predicted target means for the Poisson generalized linear model is 1.305.

***Figure 10**. Plot of Actual and Predicted Variables by Depth for the Quantile Regression Model*



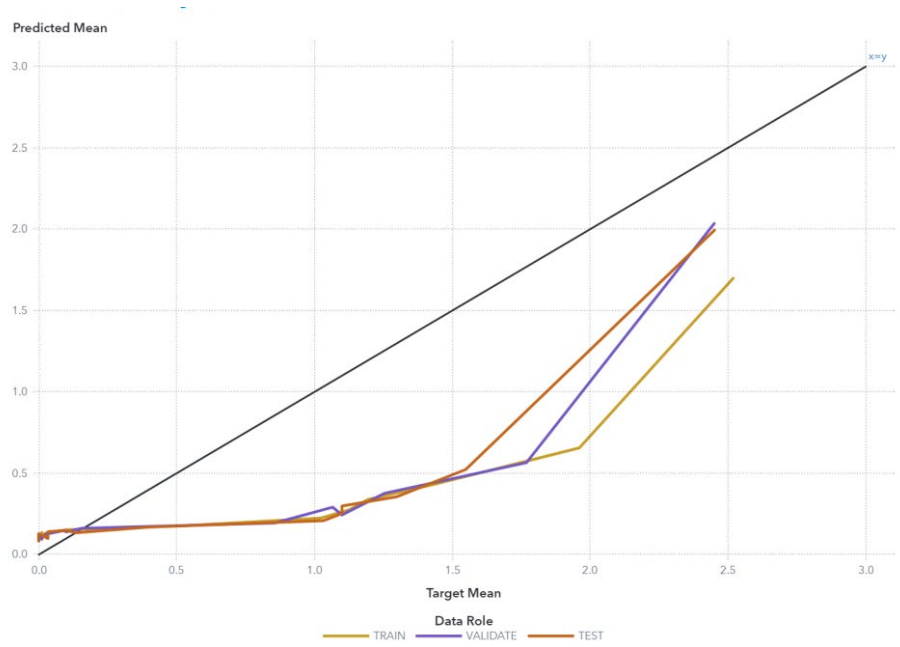
We also examined another plot, of the predicted mean versus the actual mean, shown in Figure 11, which offers a different representation of the data than Figure 10. In Figure 11, the predicted mean is plotted against the target mean. Partitions are sorted and divided in the same manner as in Figure 10. The black diagonal line in the plot indicates the points where the predicted mean and the actual mean are equal. Consequently, the points for a perfect model would correspond to that line. Points that are plotted above the line signify overprediction of the target; points that are plotted below the line indicate underprediction of the target. Although our predicted values closely align with the actual values, there is some overprediction of lower means and a slight underprediction farther along the black diagonal line.

Figure 11. *Plot of Predicted Mean versus Actual Mean for the Quantile Regression Model*



For comparison, Figure 12 shows the same type of plot for the generalized linear model.

***Figure 12**. Plot of Predicted Mean versus Actual Mean for the Generalized Linear Model*



We can also compare the mean of the predicted Claim_Count value for these two methods to the actual mean, which is 0.48. The mean of the target variable that is estimated using generalized linear modeling is 0.27. Both the mean plot (Figure 10) and the plots of the predicted mean versus the actual mean (Figures 11 and 12) exhibit a substantial underestimation of the target variable; the mean is underestimated by 44%. This underestimation could result in a final premium amount that is insufficient to cover all actual losses, potentially leading to a devastating impact on an insurance company. To construct an appropriate generalized linear model, we would need to transform skewed input variables, address outliers, and account for the heteroscedasticity of residuals. In contrast, the mean of the predicted variable that is calculated using quantile regression is 0.44, which is considerably closer to the actual mean, although it remains slightly lower.

Independence from Distribution

In the previous section, the only transformed input variable to have the same transformation as the target variable was Exposure. Because telematics variables are highly skewed to the right, we applied a transformation and analyzed how it affected the results. The next step is to apply this transformation to telematics variables and observe the change in the results for both models. Figure 13 shows the pipeline for these models.

***Figure 13**. Pipeline for Generalized Linear and Quantile Regression Models with Transformation of Input Variables*

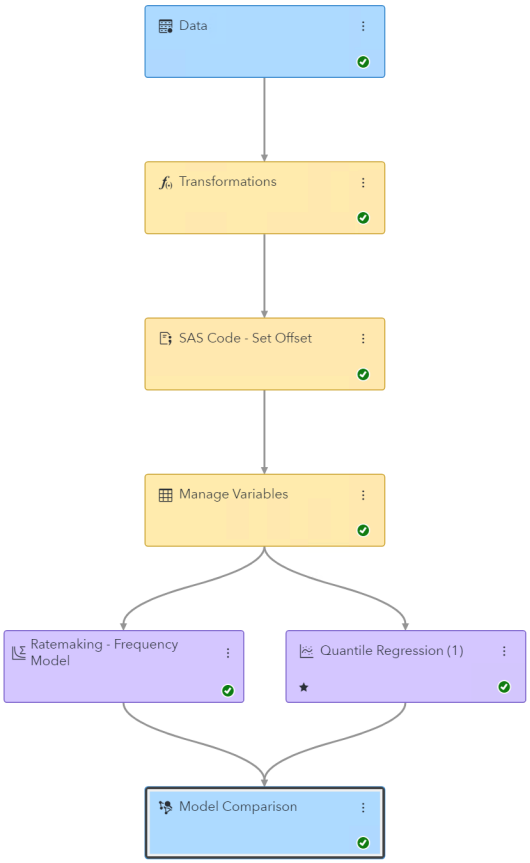


Figure 14 shows significant decrease in average squared error for the generalized linear model. This statistic is almost three times lower than the one for the same model with untransformed variables. On the other hand, average squared error slightly increased for the quantile regression model, so there is no need to transform variables for this model to achieve better results.

***Figure 14**. Average Squared Error for Generalized Linear and Quantile Regression Models after Transforming Input Variables*

Model Comparison

Champion	Name	Algorithm Name	Average Squared Error
★	Quantile Regression (1)	Quantile Regression	0.1911
	Ratemaking - Frequency Model	Count GLM	0.3013

Robustness to Outliers

An important point of comparison between quantile regression and generalized linear modeling is how robust each method is to outliers. As mentioned earlier, quantile regression is considered to be more robust to the presence of outliers, but we wanted to test this comparison on our automobile insurance data. To do so, we added a new observation to our data with an extreme value of the Claim_Count variable; this value is 200. We then recalculated our quantile regression and generalized linear models to see how the predicted value of Claim_Count changed, looking at the mean of the predicted Claim_Count value for both models. Before adding the outlier, the mean of the generalized linear model's predictions was 0.26837. After adding the outlier, this value increased by almost 21%, to 0.32398. This significant increase in mean prediction is due to the sensitivity of this model to outliers, which can pull the estimated coefficients to extreme values and reduce the quality of the model. In the insurance industry, this can lead to higher premium rates for certain policyholders, which are influenced by the extreme value of a single individual. It can also cause problems when you apply the model to different data sets. In our comparison, the mean of the predicted Claim_Count value for the quantile regression model changed only slightly, from 0.43529 to 0.4369—an increase of less than 1%. This demonstrates the robustness of quantile regression to outliers, because QR is less affected by extreme values and thus can provide more stable predictions. This makes it a useful modeling technique for applications such as insurance, where accurate predictions are essential to risk assessment and pricing.

Comparison of Predictions for Different Quantiles

Now let's examine the predictions that quantile regression generates for the 0.5, 0.75, 0.95, and 0.99 quantiles. Figures 15−19 illustrate how the plot of predicted versus actual values changes with higher quantiles. For the 0.5 quantile, as described earlier, our model's predicted target values are for the most part slightly underpredicted. However, at higher quantiles, the predicted lines progressively shift above the line that represents the actual values. Although we can still observe some underprediction for a small portion of the predicted values for the 0.75 quantile, most of the values align with our expectations that the predicted values for the 0.75 quantile should be higher than the actual values of the target variable. The lines that represent predicted values are entirely above the line of actual values for both the 0.95 and 0.99 quantiles.

***Figure 15**. Predicted versus Actual Values for the 0.5 Quantile*

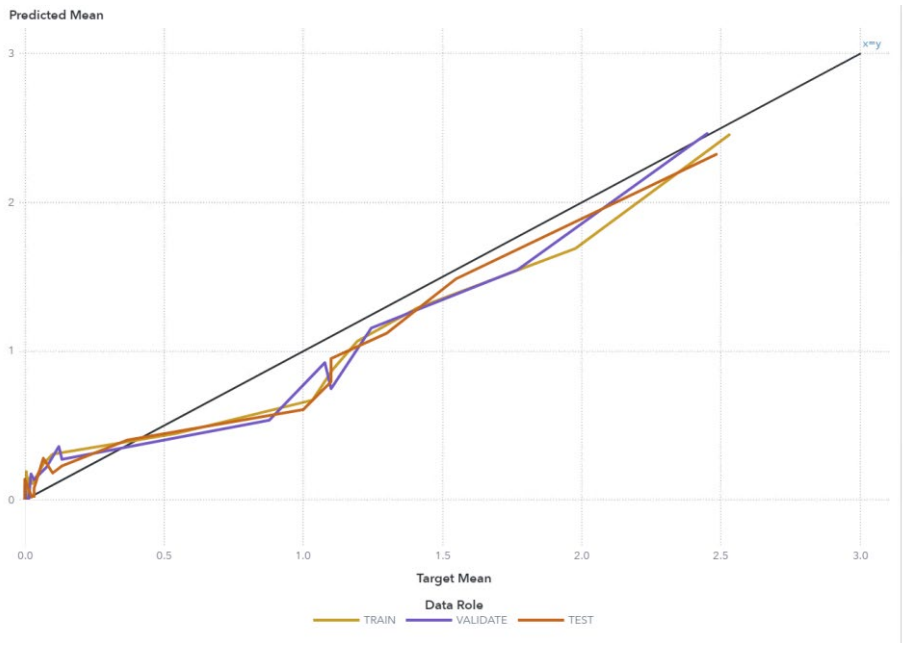


***Figure 16**. Predicted versus Actual Values for the 0.75 Quantile*

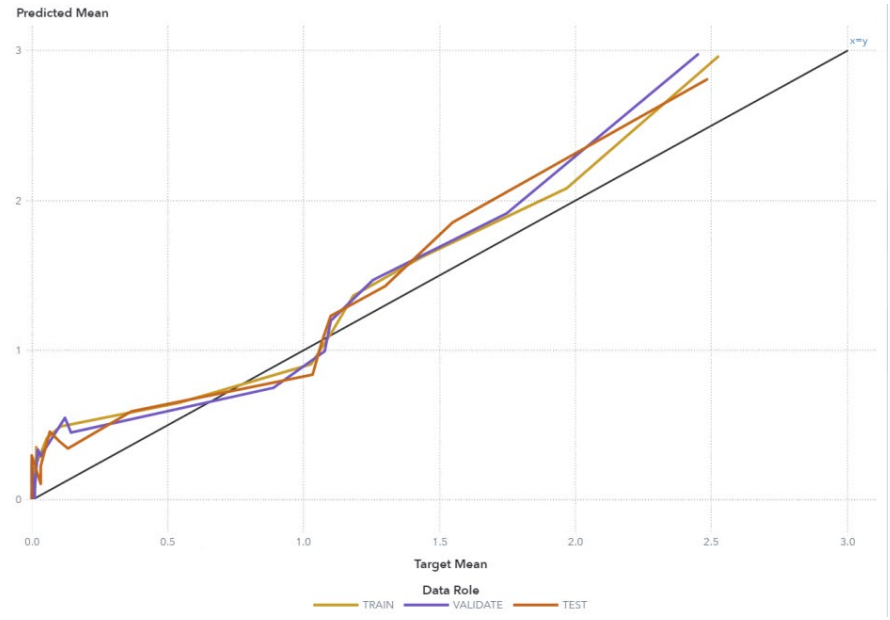


Figure 17. *Predicted versus Actual Values for the 0.95 Quantile*

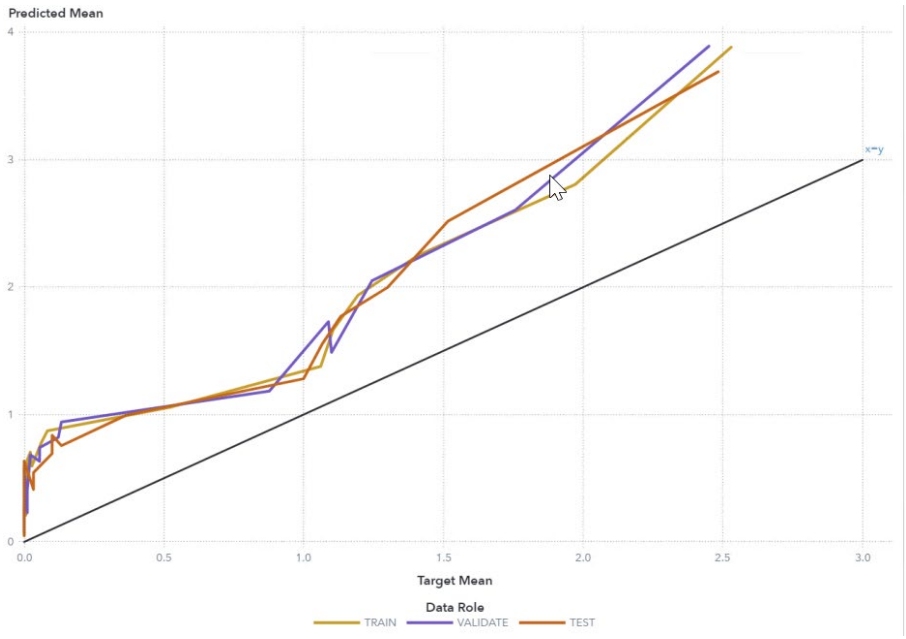
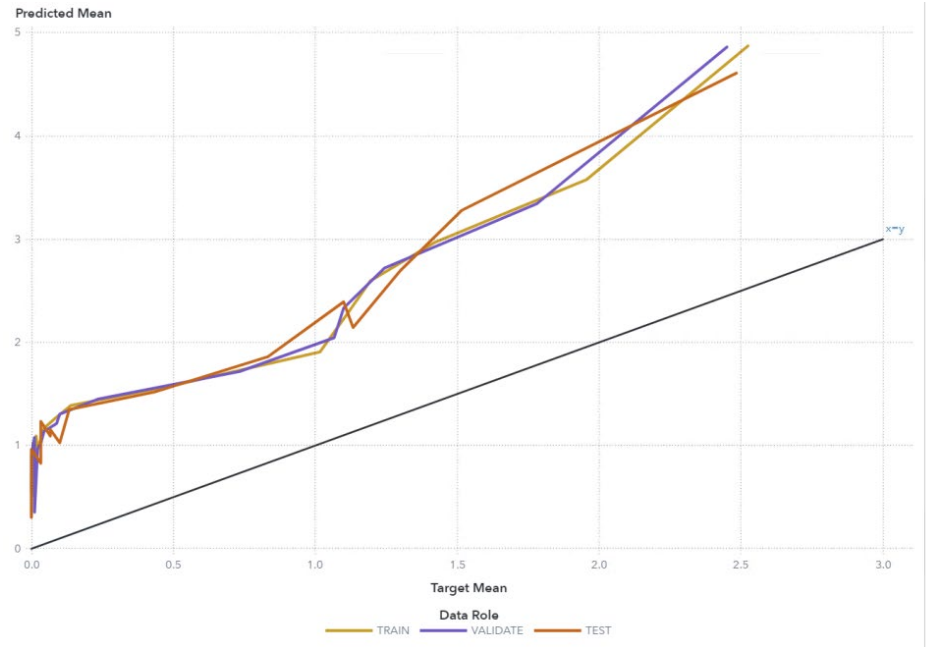


Figure 18. *Predicted versus Actual Values for the 0.99 Quantile*



We now examine the most significant predictors of the four quantiles that we investigated. As previously observed, telematics variables are the most relevant for the 0.5 quantile (Figure 19). Specifically, the variable Harsh_Brakes has the greatest impact on the target variable, Claim_Count. As Figure 20 shows, the variables that are selected and their order of importance remain the same for the 0.75 quantile, and very similar effects are estimated. However, the estimates for all variables are slightly higher than those for the 0.5 quantile.

Figure 19*. Parameter Estimates for the 0.5 Quantile*

Effect	Parameter	t Value	Sign	Estimate	Absolute Estimate ↓	Standard Error	Pr > |t|
Intercept	Intercept	65.5526	-	-0.7368	0.7368	0.0112	0
Harsh_Brakes	Harsh_Brakes	31.5274	+	0.1023	0.1023	0.0032	0.0000
Harsh_Lateral	Harsh_Lateral	21.1418	+	0.0728	0.0728	0.0034	0.0000
Harsh_Accel	Harsh_Accel	26.1498	+	0.0450	0.0450	0.0017	0.0000
Night_Driving	Night_Driving	16.5127	+	0.0276	0.0276	0.0017	0.0000

Figure 20*. Parameter Estimates for the 0.75 Quantile*

Effect	Parameter	t Value	Sign	Estimate	Absolute Estimate ↓	Standard Error	Pr > |t|
Intercept	Intercept	43.1802	-	-0.7135	0.7135	0.0165	0
Harsh_Brakes	Harsh_Brakes	25.5715	+	0.1212	0.1212	0.0047	0.0000
Harsh_Lateral	Harsh_Lateral	14.2494	+	0.0749	0.0749	0.0053	0.0000
Harsh_Accel	Harsh_Accel	19.4300	+	0.0507	0.0507	0.0026	0.0000
Night_Driving	Night_Driving	15.3438	+	0.0351	0.0351	0.0023	0.0000

For the 0.95 quantile, the order of variables that have the most significant effect remains unchanged, as shown in Figure 21. However, two new variables have emerged as important attributes: Vehicle_Age and PolicyHolder_Age. Despite having the lowest estimated values, especially PolicyHolder_Age, and the highest *p*-values, these variables still have valid estimates because the *p*-values remain acceptable at a 5% significance level. In comparison to the 0.75 quantile, the estimates for all variables have slightly increased.

Figure 21*. Parameter Estimates for the 0.95 Quantile*

Effect	Parameter	t Value	Sign	Estimate	Absolute Estim... ↓	Standard Error	Pr > |t|
Intercept	Intercept	10.1085	-	-0.6108	0.6108	0.0604	0.0000
Harsh_Brakes	Harsh_Brakes	15.4647	+	0.1405	0.1405	0.0091	0.0000
Harsh_Lateral	Harsh_Lateral	9.6897	+	0.1026	0.1026	0.0106	0.0000
Harsh_Accel	Harsh_Accel	12.9727	+	0.0646	0.0646	0.0050	0.0000
Night_Driving	Night_Driving	7.9548	+	0.0392	0.0392	0.0049	0.0000
Vehicle_Age	Vehicle_Age	2.5860	-	-0.0180	0.0180	0.0070	0.0097
PolicyHolder_Age	PolicyHolder_Age	2.1893	+	0.0025	0.0025	0.0012	0.0286

As we increase the quantile estimate, the number of relevant parameter estimates tends to increase as well. For the 0.99 quantile, shown in Figure 22, the most important attribute is Vehicle_Type, followed by the telematics variables. At the bottom of the table are Vehicle_Age, PolicyHolder_Age, and Vehicle_Value. However, most non-telematics variables have a relatively high *p*-value, which means that we cannot consider them to be reliable estimates. If we remove all estimates that have a high *p*-value from our table, as shown in Figure 23, telematics variables once again become the most important. One key difference between the 0.99 quantile and other quantiles is that the 0.99 quantile does not consider the Intercept as having an effect on the model.

Figure 22. *Parameter Estimates for the 0.99 Quantile*

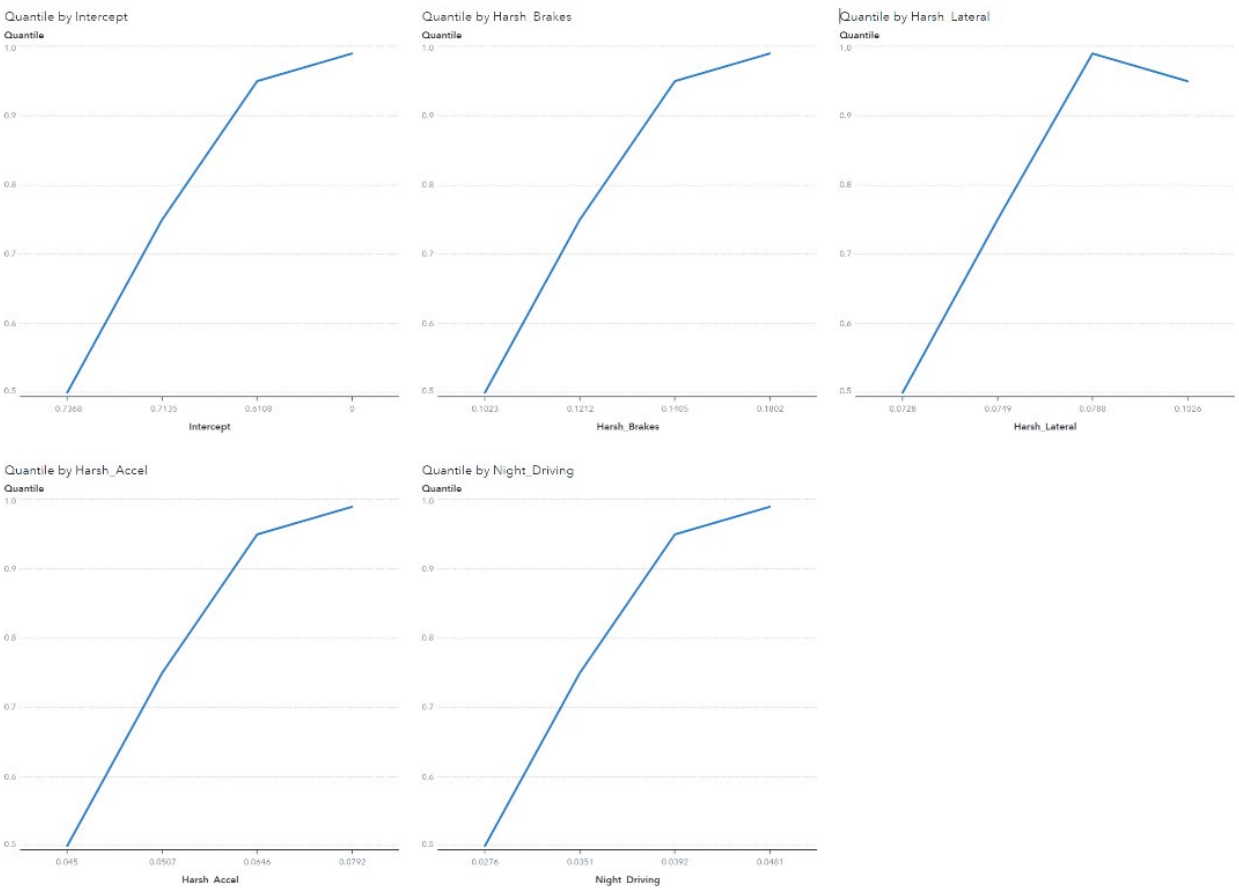
Effect	Parameter	t Value	Sign	Estimate	Absolute Estim... ↓	Standard Error	Pr > |t|
Vehicle_Type	Vehicle_Type sports car	1.1637	-	-0.3068	0.3068	0.2637	0.2446
Vehicle_Type	Vehicle_Type sedan	0.8846	-	-0.2379	0.2379	0.2689	0.3764
Vehicle_Type	Vehicle_Type SUV	0.7125	-	-0.1899	0.1899	0.2665	0.4762
Harsh_Brakes	Harsh_Brakes	7.1157	+	0.1847	0.1847	0.0260	0.0000
Harsh_Lateral	Harsh_Lateral	3.3304	+	0.0807	0.0807	0.0242	0.0009
Harsh_Accel	Harsh_Accel	5.7637	+	0.0795	0.0795	0.0138	0.0000
Night_Driving	Night_Driving	4.0246	+	0.0541	0.0541	0.0134	0.0001
Vehicle_Age	Vehicle_Age	2.6080	-	-0.0484	0.0484	0.0186	0.0091
PolicyHolder_Age	PolicyHolder_Age	0.7815	+	0.0031	0.0031	0.0040	0.4346
Vehicle_Value	Vehicle_Value	0.7806	-	0.0000	0.0000	0.0000	0.4351

Figure 23. *Parameter Estimates for the 0.99 Quantile without the High p-Values*

Effect	Parameter	t Value	Sign	Estimate ↓	Absolute Estimate	Standard Error	Pr > |t|
Harsh_Brakes	Harsh_Brakes	7.7099	+	0.1802	0.1802	0.0234	0.0000
Harsh_Accel	Harsh_Accel	6.2723	+	0.0792	0.0792	0.0126	0.0000
Harsh_Lateral	Harsh_Lateral	3.3690	+	0.0788	0.0788	0.0234	0.0008
Night_Driving	Night_Driving	4.0427	+	0.0481	0.0481	0.0119	0.0001
Vehicle_Age	Vehicle_Age	5.9906	-	-0.0736	0.0736	0.0123	0.0000

Figure 24 illustrates the evolution of estimates of telematics variables across quantiles. As shown, all variables follow a similar increasing trend.

***Figure 24**. Evolution of Estimates of Telematics Variables across the Quantiles*



Use of Higher Quantiles of Quantile Regression

How can quantiles higher than 0.5 be beneficial in analyzing an insurance portfolio, given that the primary objective of the modeling process is to estimate the actual values of the target variable as accurately as possible? One valuable application of higher quantiles is in the treatment of safety loadings.

In insurance, calculating a net premium by using generalized linear modeling entails the analysis of two factors: expected pure premium (pure cost of risk) and safety loadings (Klugman, Panjer, and Willmot 2012). Safety loadings are charges that are added to the pure premium to guarantee that insurers have adequate funds to cover any losses that might occur during the policy period. They can be estimated as a proportion of one of the moments of the loss distribution. Quantile regression enables us to compute the pure premium in a single step by simply modeling one of the quantiles (Koenker 2005). We can use results from the previous section and set the 0.75 percentile as our benchmark for safety loadings. This implies that instead of selecting the 0.5 quantile model as the final model, we can choose the 0.75 quantile model, which is slightly overestimated. This approach automatically incorporates safety loadings in our final estimates, eliminating the need for separate calculations. It is possible to directly control the

value of safety loadings by choosing the probability as a parameter of the model. The mean value of the predicted 0.75 quantile of the target variable is 0.64, which overestimates Claim_Count by 31%.

Another advantage of quantile regression is its utility in identifying high-risk drivers. We can use the findings from the previous section and examine the estimated 0.99 quantile of the target variable, denoted as P_Claim_Count. The study of the risk factors at different quantile levels offers the possibility of a more granular risk segmentation. In the context of risk segmentation for ratemaking, quantile regression offers a significant advantage in terms of granular risk segmentation. By estimating conditional quantiles specific to different risk segments, QR lets insurers differentiate pricing and underwriting strategies at a more refined level. This enables a more precise assessment of risk for customers who have varying risk profiles, leading to more tailored pricing models. QR's ability to identify and capture the heterogeneity in risk within a given population enhances insurers' ability to segment customers accurately and assign appropriate premiums according to their specific risk characteristics.

Analyzing extreme quantiles can greatly enhance tail-risk estimation, which is crucial for insurers to comprehend the losses that they might incur in the face of catastrophic events. Tail risk specifically refers to the probability of extreme events occurring at the tail end of a probability distribution. These are rare events, but their impact on an insurer's financial stability or profitability can be significant. Tail-risk estimation is a valuable tool for insurers, because it helps them determine the necessary level of reinsurance to cover potential losses that arise from such extreme events. By using this estimation, insurers can also evaluate their risk exposure and devise strategies to mitigate or manage tail risk.

Suggestions for Further Examination

Quantile regression has already demonstrated its value in the ratemaking modeling process, but its full potential has yet to be fully realized. One promising area for further exploration is the combination of this method with other modeling techniques, such as generalized additive models (GAMs), which can capture nonlinear and complex relationships between variables. Another candidate for integration is copulas, which can model the dependence structure between variables separately from their marginal distributions—a valuable tool when the marginal distributions are complex or unknown.

Conclusion

Our study focused on the advantages of quantile regression over generalized linear modeling in the ratemaking process. We used the SAS actuarial tool SAS Dynamic Actuarial Modeling software to perform all the necessary analysis in the study. We began by introducing our telematics data on automobile drivers, followed by a brief overview of quantile regression and generalized linear modeling theory, and then we highlighted the key benefits of quantile regression.

In a practical example, we compared quantile regression to generalized linear modeling and demonstrated that the model that was created by quantile regression was more accurate than the one created by generalized linear modeling when applied to our data. Quantile regression is distribution-free, meaning that there is no need to transform any of the variables or to determine the correct distribution for the target variable. We also showed the robustness of quantile regression to outliers. This robustness can lead to more accurate predictions and premium calculations, improving underwriting performance and profitability. We demonstrated how insurance companies can take advantage of higher quantiles. In the first place, higher quantiles can be used to improve risk segmentation. Higher quantiles provide a more comprehensive understanding of the relationship between policyholder characteristics and claim outcomes across different quantiles of the loss distribution. This leads to more accurate

risk segmentation, enabling insurers to price policies more accurately and competitively. By using higher quantiles, insurers can also identify policyholders who pose a higher risk than others, enabling them to impose penalties, adjust premiums, or limit coverage for higher-risk drivers, potentially reducing claim costs and increasing profitability. Another valuable application of extreme quantiles is the improvement of tail-risk estimations. Insurers can manage their capital more effectively by ensuring adequate capitalization to cover extreme events while optimizing capital allocation and profitability. Quantile regression is also useful for calculating safety loadings, which is another segment where insurers can benefit from this method. It enables them to estimate safety loadings more accurately by directly modeling the higher quantiles of the loss distribution. Accurate safety loadings guarantee that insurers have enough funds to cover potential losses during the policy period, reducing the likelihood of financial distress and improving long-term profitability. By using quantile regression, insurers can save significant effort in calculating safety loadings and optimize their capital allocation. All these benefits make quantile regression a suitable method for the ratemaking process, because it can significantly improve the efficiency, profitability, and accuracy of an insurance company's modeling process. These improvements can ultimately lead to better decision-making and more effective pricing strategies, which can help insurance companies remain competitive and achieve their financial goals.

References

Buchinsky, M. 1998. "The Dynamics of Changes in the Female Wage Distribution in the USA: A Quantile Regression Approach." *Journal of Applied Econometrics* 13:1−30. Available at https://doi.org/10.1002/(SICI)1099-1255(199801/02)13:1<1::AID-JAE474>3.0.CO;2-A.

Fahrmeir, L., and Tutz, G. 2001. "Models for Multicategorical Responses: Multivariate Extensions of Generalized Linear Models." In *Multivariate Statistical Modelling Based on Generalized Linear Models*, 69−137. New York: Springer. Available at https://link.springer.com/chapter/10.1007/978-1-4757-3454-6_3.

Huber, P. J. 1981. *Robust Statistics*. Hoboken, NJ: Wiley. Available at http://dx.doi.org/10.1002/0471725250.

Karst, O. J. 1958. "Linear Curve Fitting Using Least Deviations." *Journal of the American Statistical Association* 53:118−132. Available at https://www.tandfonline.com/doi/abs/10.1080/01621459.1958.10501430.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. 2012. *Loss Models: From Data to Decisions*. 4th ed. Hoboken, NJ: Wiley.

Koenker, R. 2005. *Quantile Regression*. New York: Cambridge University Press. Available at https://doi.org/10.1017/CBO9780511754098.

Koenker, R., and Bassett, G., Jr. 1978. "Regression Quantiles." *Econometrica* 46:33−50. Available at https://www.jstor.org/stable/1913643.

Koenker, R., and Hallock, K. F. 2001. "Quantile Regression." *Journal of Economic Perspectives* 15:143−156. Available at https://www.aeaweb.org/articles?id=10.1257/jep.15.4.143.

Koenker, R., and Xiao, Z. 2003. "Inference on the Quantile Regression Process." *Econometrica* 70:1583−1612. Available at https://doi.org/10.1111/1468-0262.00342.

Kudryavtsev, A. A. 2009. “Using Quantile Regression for Rate-Making.” *Insurance: Mathematics and Economics* 45:296−304. Available at https://doi.org/10.1016/j.insmatheco.2009.07.010.

Machado, J. A. F., and Silva, J. M. C. S. 2005. “Quantiles for Counts.” *Journal of the American Statistical Association* 100:1226−1237. Available at https://doi.org/10.1198/016214505000000330.

McCullagh, P., and Nelder, J. A. 1989. *Generalized Linear Models*. 2nd ed. London: Chapman and Hall.

Nelder, J. A., and Wedderburn, R. W. M. 1972. “Generalized Linear Models.” *Journal of the Royal Statistical Society, Series A* 135:370−384. Available at https://doi.org/10.2307/2344614.

Pérez-Marín, A. M., Guillen, M., Alcañiz, M., and Bermúdez, L. 2019. “Quantile Regression with Telematics Information to Assess the Risk of Driving above the Posted Speed Limit.” *Risks* 7:80. Available at https://doi.org/10.3390/risks7030080.

Rousseeuw, P. J., and Leroy, A. M. (1988). “A Robust Scale Estimator Based on the Shortest Half.” *Statistica Neerlandica* 42:103−116. Available at https://doi.org/10.1111/j.1467-9574.1988.tb01224.x.

Release Information Content Version: 1.0. July 2023.

To contact your local SAS office, please visit: sas.com/offices

® indicates USA registration. Other brand and product names are trademarks of their respective companies. Copyright © SAS Institute Inc. All rights reserved.