# Quantile Regression

By

Luyang Fu, Ph. D., FCAS, State Auto Insurance Company

Cheng-sheng Peter Wu, FCAS, ASA, MAAA, Deloitte Consulting

# Agenda

- Overview of Predictive Modeling for P&C Applications
- Quantile Regression: Background and Theory
- Two Commercial Line Case Studies – Claim Severity Modeling and Loss Ratio Modeling
- Q&A

# Overview of Predictive Modeling for P&C Applications

# Overview of PM for P&C Applications

- Modeling Techniques:
  - Ordinary Least Square and Linear Regression:
    - Normal distribution assumption
    - Linear relationship between target and covariates
    - Predict the mean of the target against the covariates
  - GLM:
    - Expansion of distribution assumptions with exponential family distributions,: frequency - Poisson, severity - Gamma, pure premium and loss ratio - Tweedie
    - Linear relationship between the mean of the target and the covariates through the link function
  - Minimum Bias/General Iteration Algorithm:
    - Iterative algorithm
    - Essentially derives the same results as GLM
    - Categorical predictive variables only
    - Linear relationship between target and covariates
  - Neural Networks:
    - Non-linear regression: a series of logit functions to approximate the nonlinear relationship between target and covariates
    - Originated in the data mining field; a curve fitting technique; non parametric assumptions
    - Different cost or error functions can be selected for minimization, for example, a least square error term or an absolute error function
    - One popular algorithm is the backward propagation of errors

# Overview of PM for P&C Applications

- Modeling Techniques:
  - MARS:
    - Non-linear regression using series of “hockey stick” function to approximate the nonlinear relationship between the target variable and the covariates
    - Can test interaction between covariates
  - CART:
    - A recursive partitioning technique - regression trees for continuous target and classification trees for categorical target variables
    - Originated in the data mining field
    - Different error functions can be used, such as the least square error function, Gini, entropy, etc.
    - Strong for categorical variables, but weak for continuous variables with a linear relationship with the target
- Underlying Algorithms for Different Techniques:
    - Statistical modeling techniques, such as GLM and OLS, attempt to maximize the likelihood functions for the underlying distribution
    - Non-statistical modeling techniques, such as Nnet, CART, and MARS, attempt to minimize a pre-determined error function, and the error functions can include a least square error function, an absolute error function, a Gini function, etc. Different optimization techniques are deployed to find the solutions

# Overview of PM for P&C Applications

- Types of Model for P&C Applications:
  - Binary target models for retention, cross sale, or marketing:
    - Logistic regression
    - Target variable is a binary variable with a logit distribution assumption
  - Frequency and severity models
    - Popular for class plan optimization
    - Poisson for frequency and Gamma for severity
    - Becoming the standard approach for personal line pricing
    - Used when data and exposure information are in a good condition
  - Loss cost and pure premium models:
    - Long history with the minimum bias technique
    - One step modeling instead of 2 steps frequency and severity modeling
    - Tweedie distribution assumption can be applied, but may not fit the data well
    - Software support of the Tweedie assumption is not popular yet

# Overview of PM for P&C Applications

- Types of Model for P&C Applications:
  - Loss ratio modeling:
    - More popular for commercial line applications
    - Tweedie assumption can be applied as well, but may not fit the data well
    - Applied when data challenges are significant for the frequency or pure premium models, for example, exposure not accurate or not homogenous
    - May not focusing on the "predictive value" of the target, but the "ranked value" of the target, for example, the worst X% of policies for cancellation, or the best Y% for the most preferred company placement
    - Application is not for setting class plan factors or rates, but for underwriting, such as tiering, company placement, credit-debit assignment, new business rejection/acceptance, or renewal business cancellation

# Quantile Regression – Background and Theory

# Quantile Regression – Background

- Originated in the Econometric field by Roger Koenker and Gilbert Bassett from University of Illinois:
  - Roger Koenker and Gilber Bassett, “Regression Quantiles”, *Econometrica*, (1978)
- Traditional modeling, such as OLS and GLM, is to model the conditional mean of the target variable against the covariates, while Quantile Regression is to model conditional percentiles of the target variable against the covariates.
  - For example, driver gender may not impact the mean of claim severity, but may have a significant impact on the 95% percentile of the severity
- The technique has been used in other industries and researches, such as ecology, healthcare, and financial economics, where data is volatile and extremes are important.

# Quantile Regression – Advantages

- No distribution assumptions:
  - Severity: Gamma, Lognormal, or Pareto?
  - Pure premium or loss ratio: Tweedie?
- Robust:
  - Unlike OLS or GLM, it is robust in handling extreme value points and outliers for the target
  - Insensitive (equi-variant) to any monotonic transformations of the target variable
  - The regression coefficients do not vary by the capping on the target variable for most of the percentiles
- Comprehensive:
  - A more "complete" picture of the relationship between the target and the covariates

# Quantile Regression – Theory

- OLS with a Least Square Error Function -
  - The parameter estimates give the relationship between the mean value of the target against the covariates
  - For OLS, the parameters estimates by minimizing the least square function is equivalent, asymptotically, to the parameters estimates by maximizing the normal likelihood function

$$y = f(x) = \alpha_0 + \sum_{j=1}^{m} \alpha_j x_j$$

$$E(Y|X) = \alpha x'$$

$$\hat{\alpha} = argmin_{\alpha \in R^p} \sum_{i=1}^{n} (y_i - \alpha x_i')^2$$

  - In general, the GLM types of modeling will predict the mean of the target variable given the covariates

}

# Quantile Regression – Theory

- Quantile regression -
  - Predict the ξth percentile, instead of the mean, of the target variable against the covariates.
  - The ξth percentile of a random variable, Y is defined as:

$$Q(\xi) = \inf\{y : F(y) \geq \xi\}$$

  - Conditional quantile function of Y given covariates of X:

$$Q(\xi | X = x) = \alpha(\xi) x_i'$$

  - Let's start to predict the median, the 50th percentile, then, instead of minimizing the least square error term, we will minimize the absolute error function (also known as $L_1$ regression):

$$\hat{\alpha} = argmin_{\alpha \in R^p} \left( \sum_{i=1}^{n} |y_i - \alpha x_i'| \right)$$

  - To further conduct the ξth quantile regression, we will minimize the following error function:

$$\hat{\alpha} = argmin_{\alpha \in R^p} \left[ \sum_{i \in \{i : y_i \geq \alpha x_i^{i}\}} \xi \, |y_i - \alpha x_i^{i}| + \sum_{i \in \{i : y_i < \alpha x_i^{i}\}} (1 - \xi) |y_i - \alpha x_i^{i}| \right]$$

# Quantile Regression – Theory

- Algorithms to Solve Quantile Regression:
  - The error function for minimization can be transformed into the standard “Linear Programming” type of dual problems for minimization and maximization.
  - Then, linear programming algorithms can be applied to solve the parameters for Quantile Regression:
    - Simplex method: classical, less efficient, stable
    - Interior point method: fast, may not converge
    - Smoothing algorithm: fast, may not converge

# Quantile Regression – Theory

- Confidence Interval Calculation for Quantile Regression:
  - Since it is a non-parametric approach, no distribution function can be used to calculate the confidence interval
  - Three alternative algorithms to estimate the confidence interval:
    - Sparsity function: direct, fast, but not robust if data is not i.i.d.
    - Inversion of Rank Tests: computation intensive due to using simplex algorithm.
    - Markov Chain Marginal Bootstrapping: unstable for small data sets.

# Quantile Regression – Background and Theory

- Summary of Quantile Regression:
  - By definition, the target variable needs to be a continuous variable, not a categorical or mixed variable
  - Minimizing a quantile error function
  - More complete understanding of the impact of covariates on the dependent variable across the whole distribution, not just the mean of the dependent variable
  - Uses Linear Programming algorithms to estimate the regression parameters
  - Confidence intervals cannot be estimated with known distribution functions. Instead, several different algorithms can be used to estimate the confidence intervals
  - Widely applied in other research areas
  - Several statistical software can supports quantile regression, such as SAS (Proc Quantreg), R, Stata, and Matlab

# Cast Study #1 – BI Claim Severity Model

# Cast Study #2 – Loss Ratio Modeling

# Conclusions:

- For the two case studies, the technique proves to be useful:
  - More complete understanding of the impact of covariates on the target, especially toward the extreme ends of the distribution.
  - Yield more stable, robust, and stronger  results for the loss ratio modeling compared to OLS and GLM
- Can be another tool added to our modeling tool set
- Potential insurance applications in severity modeling, loss ratio modeling, reinsurance pricing, value-at-risk analysis, and capital allocation

# Q&A