University of Razi

Statistics Branch

Insurance premium prediction via Gradient Tree Boosted Tweedy Compound Poisson Model

By:
Mohanna Mosabeygi

Supervisor:
Dr. Maryam Sharafi
Dr. Reza Hashemi

January 2020

Contents

List of Figures

List of Tables

Abstract

Our research is applied in terms of purpose. Because the proposed model lays solutions to improve the premium determination and generally improve the performance of insurance companies. We offer model forecasting methods to determine the premium rate, That detects data exploration and modeling. Among these methods, the accelerated gradient is a method in composite Poisson model. Since the main variables and interaction effects used in the models are, therefore, a tree accelerated gradient algorithm with the name *TDboost* offer visited. also for data with a large zero accumulation. The methods will be provided to make the premium forecast possible. First, we will discuss the definitions and concepts required in insurance science. o we introduce and examine the accelerated gradient tree model. In the third chapter, we implement a model for the survey of the database composite Poisson with insurance studies data. In the fourth chapter, we will analyze and compare non-parametric models using data sets, And finally, we will conclude our suggestions and conclusions.

Keywords: Generalized Linear Model, Gradient Boosted Models, Tweedy Compound Poisson Model, Zero-Inflated Compound Poisson Regression Model.

1 Foreword:

This thesis is organized into four distinct chapters. The initial chapter lays the groundwork, introducing key concepts that are essential for understanding the rest of the work. Chapter two delves into various approaches for predicting non-linear models, including a thorough examination of generalized linear models, aggregate models, and methods involving tree analysis. The third chapter shifts the spotlight to the Tweedy Composite Poisson Model, specifically examining its combination with the tree-acceleration method for effective insurance premium computation, highlighting the use of the TD-boost algorithm. The influential work of Yang and colleagues in 2017 is also acknowledged here, particularly their introduction of non-parametric models that adeptly tackle model selection challenges. Finally, the fourth chapter applies the methods discussed earlier within the Composite Poisson framework, utilizing R software and associated packages for an in-depth analysis.

2 Chapter 1: Introduction to Basic Definitions a Concepts

3 Introduction to Insurance Fundamentals

The concept of 'insurance' in Persian connotes safeguarding against unexpected incidents and damages, resonating with the terms 'assurance' and 'provision' in English, French, and Arabic. Historically, the Persian nomenclature for insurance evolved from the notion of 'No insurance.' According to Iran Insurance Law (Article 1, 1316), insurance contracts are agreements providing financial protection or a predetermined sum subsequent to a specific event, in return for a premium payment.

3.1 Exploring the Evolution of Insurance Science

This chapter provides an extensive review of significant transformations in the field of insurance science. It acquaints the reader with diverse insurance types, their respective applications, and the complexities of insurance premiums, encompassing their classification and calculation. The chapter aims to impart a thorough comprehension of the insurance concept, its contractual foundations, and the pivotal role of premiums within the larger framework of insurance science and its practical implementations.

4 Fundamental Concepts in Insurance

4.1 The Insured and the Insurer

This segment explicates the primary roles within the insurance domain. 'The Insured' refers to the individual or entity who secures insurance coverage by paying premiums. As policyholders, they may obtain this coverage either for themselves or on others' behalf. Conversely, 'The Insurer' denotes the insurance firm stipulated in the policy, responsible for compensating the insured as per the policy stipulations.

4.2 Comprehending Insurance Policies

An insurance policy is a legal agreement delineating the obligations and rights of both the insurer and the insured under defined circumstances. It incorporates both 'general conditions,' applicable to all insurance types, and 'specific

conditions,' customized to individual contracts. Key elements of an insurance policy include the agreement date, party identities, insurance subject, risk coverage, policy tenure, premium amount, and enumerated exclusions.

4.3 Classification of Insurance Types

Insurance can be categorized into various segments, each addressing distinct requirements:

- Personal Insurance: Encompasses life, health, and related aspects.
- Property Insurance: Provides protection against property-related losses.
- Liability Insurance: Pertains to legal liabilities.
- Special Policies: Encompasses unique insurance types such as bank loans, fund insurance, horse insurance, and others.

5 Indemnity and Premiums in Insurance

5.1 The Concept of Indemnity

In the realm of insurance, 'Indemnity' pertains to the compensation provided for losses incurred due to insured events. It is a fundamental principle that underpins the insurance mechanism, ensuring financial restitution for policyholders.

5.2 The Role of Insurance Premiums

Insurance premiums represent a crucial source of revenue for insurance companies. These premiums, paid by the insured, are calculated based on various factors such as the policy's duration, associated risks, and coverage extensions. Premiums are bifurcated into 'gross premiums' and 'net premiums.' Gross premiums encompass net premiums along with additional charges like administrative fees and commissions, while net premiums are exclusive of these additional costs. Essentially, the premium reflects the cost of risk coverage.

6 Insurance Premiums: Key Aspects and Determinants

6.1 Centrality of Insurance Premiums

The insurance premium is pivotal in the financial sustenance of insurers, derived from the sale of insurance policies. Accurate premium determination is thus a critical and intricate task in the insurance sector.

6.2 Employer-Paid Premium

In certain insurance models, premium payment responsibilities are vested in employers. Non-compliance in timely premium payment attracts financial penalties, such as a monthly surcharge of 2% and an additional fee for failing to submit requisite beneficiary lists.

6.3 Defining the Premium Level

The 'premium level' refers to the agreed-upon insurance premium amount stipulated in the policy contract. This figure is determined based on specific conditions and requirements outlined within the policy.

6.4 Actuarial Fairness in Premium Calculation

Actuarial fairness plays a critical role in setting insurance premiums. Premium assessments are conducted by considering various factors, including the insured's age, gender, medical history, and familial health background. This approach ensures that premiums are proportionately adjusted in accordance with the associated risks

7 Determinants of Additional Premium Costs in Insurance

Insurance premiums are subject to escalation under various circumstances, including the extension of coverage duration, increase in the insured sum, regulatory changes in premium rates, and expansion of coverage to additional

risks. This section elucidates the factors contributing to the augmentation of basic insurance premiums into comprehensive premiums.

7.1 Factors Contributing to Premium Escalation

- Extension of coverage duration or increase in the insured sum.
- Regulatory adjustments in premium rates, as stipulated in insurance legislation, such as corrections for inadvertent errors (Article 13) or increased risks during the coverage period (Article 16).
- Inclusion of coverage for additional perils beyond the principal risk, for instance, natural disaster risks in fire insurance policies.[11]

7.2 Components of Comprehensive Premiums

To arrive at a comprehensive premium, several additional charges are incorporated, covering operational costs and service provision in insurance. These components include:

Commissions: A portion of the premium paid to brokers by insurers as compensation for policy sales. The rate varies depending on the insurance type, such as fire or auto insurance.

Acquisition Costs: Expenses related to acquiring new policyholders.

Premium Taxes: Taxes levied on the premiums.

Administrative Expenses: Allocated for operational and general corporate expenses, including executive salaries and public service charges.

Margin for Contingencies and Profit: Provision for unexpected events and profit margins.

7.3 Role of Statistics in Premium Calculation

The process of determining insurance premiums heavily relies on estimating the likelihood of future events, thus highlighting the significance of statistical science and probability theory. This statistical analysis is integral to the calculation of premiums, ensuring they are reflective of the actual risk and potential occurrences.

8 Enhancement of Predictive Accuracy in Insurance

Increased trial numbers significantly enhance predictive accuracy, reducing the discrepancy between predicted outcomes and actual events. Key factors to consider include:

- The estimated likelihood of a specific event's occurrence in the future.
- The comparison of an individual's or institution's risk probability against the average for a given risk category.

For instance, in auto insurance, drivers at fault in accidents often face higher premiums. However, elements such as managerial performance might outweigh the vehicle type in premium assessment. Consequently, low-risk drivers may qualify for reduced premiums. Similarly, in fire insurance, a sandwich shop owner may pay higher premiums than an office owner due to the greater risk associated with their business operations.

9 Evolving Models and Predictive Techniques in Insurance

9.1 Progressive Models in Premium Calculation

Recent advancements in insurance modeling include dynamic approaches such as calculating car insurance premiums based on annual mileage. This usage-based pricing strategy reflects a move towards more tailored premium calculation.

9.2 Predictive Modeling and Policyholder Behavior

Modern predictive models, like logistic regression, are used to forecast policyholder behaviors, including the propensity to switch insurers. These insights are vital for refining customer retention strategies and adjusting premiums. However, over-reliance on specific methods, like software logs or Pareto distributions, can lead to inaccurate predictions and adverse outcomes.

9.3 Upcoming Focus on Analytical and Forecasting Methods

The following chapter will explore analysis and forecasting methods in the insurance industry, emphasizing the statistical tools and models essential for these processes.

9.4 Gamma Distribution Function A Core Analytical Tool

The Gamma distribution function is fundamental in analyzing lifetime data and probabilistic scenarios. Along with the Poisson distribution, it forms a vital part of the statistical distribution family used in Bayesian estimation and insurance risk assessment. The Gamma distribution, when used as a prior distribution with the Poisson distribution as its conjugate prior, demonstrates its applicability beyond simple factorial functions. Here is the mathematical representation of the Gamma distribution function:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$$

$$f(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)} = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$$

It is defined which assumes that the expectation and variance of the gamma distribution are:

$$\begin{aligned} E(X) =& \frac{1}{\Gamma(\alpha)} \int_0^\infty (\lambda x)(\lambda x)^{\alpha-1} e^{-\lambda x} dx \\ =& \frac{1}{\lambda\Gamma(\alpha)} \int_0^\infty \lambda e^{-\lambda x} (\lambda x)^\alpha dx \\ =& \frac{\Gamma(\alpha+1)}{\lambda\Gamma(\alpha)} = \frac{\alpha}{\lambda} \end{aligned}$$

$$
\begin{aligned}
E(X^2) =& \frac{1}{\Gamma(\alpha)} \int_0^\infty \lambda x^2 e^{-\lambda x} (\lambda x)^{\alpha-1} dx \\
=& \frac{1}{\lambda^2 \Gamma(\alpha)} \int_0^\infty \lambda e^{-\lambda x} \lambda x^{\alpha+1} dx \\
=& \frac{\Gamma(\alpha+2)}{\lambda^2 \Gamma(\alpha)} \\
=& \frac{\alpha(\alpha+1)}{\lambda^2}
\end{aligned}
$$

10 Statistical Distributions and Their Properties

10.1 Variance of Gamma Random Variables

The variance of a gamma-distributed random variable X can be expressed as:

$$
\begin{aligned}
\text{var}(X) &= E(X^2) - [E(X)]^2 \\
&= \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} \\
&= \frac{\alpha}{\lambda^2}
\end{aligned}
$$

where α is the shape parameter and λ is the rate parameter.

10.2 The Dirac Delta Function

The Dirac delta function is a generalized function centered at $x = 0$. It is defined as infinitely large at $x = 0$ and zero elsewhere. Mathematically, it is represented as:

$$
f(x) = \begin{cases} \infty & \text{if } x = 0, \\ 0 & \text{if } x \neq 0. \end{cases}
$$

11 Poisson Distribution

In statistical theory, the Poisson distribution is used to model the number of events in a fixed interval of time or space. The variable X, representing

the number of successes in this interval, is considered a Poisson-distributed random variable. Variants of the Poisson distribution include classical Poisson, compound Poisson, heterogeneous Poisson, point Poisson, and truncated Poisson. The distribution is named after the French mathematician Siméon Denis Poisson.

The probability mass function of a Poisson distribution is given by:

$$f_x(x) = p(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

The characteristics of the Poisson distribution are as follows:

$$\begin{aligned} E(X) &= \lambda, \\ \operatorname{var}(X) &= \lambda, \\ \text{coefficient of variation (cv)} &= \frac{1}{\sqrt{\lambda}}. \end{aligned}$$

11.1 Poisson distribution

The Poisson distribution is a probabilistic model widely used in various fields such as statistics, physics, and engineering. It provides a framework for estimating the likelihood of a given number of events occurring within a fixed interval, assuming these events happen at a constant rate and independently of the time since the last event. This paper delves into the mathematical underpinnings of the Poisson distribution, exploring its cumulative function and the torque generating function. Furthermore, the concept of the Poisson process, an extension of the Bernoulli process, is elucidated. This process is distinguished by its continuous nature and is defined by specific criteria which classify a counting process as a Poisson process.

11.2 Introduction to the Poisson Distribution: A Mathematical Overview

This section introduces the Poisson distribution, outlining its significance in probability theory and its applicability in modeling the frequency of events. The mathematical formulation of the distribution is presented, emphasizing its reliance on the parameter λ, which represents the average number of events per interval.

11.3 Cumulative Function of the Poisson Distribution

The cumulative function, $F(x)$, of the Poisson distribution is explored in detail. This function, defined as $F(x) = P(X \leq x)$. , calculates the probability of observing up to x events. The function is expressed in terms of the exponential function and the gamma function, providing a comprehensive understanding of its behavior and implications.

$$F(x) = P(X \leq x) = \sum_{t=0}^{x} \frac{e^{-\lambda}\lambda^{t}}{t!} = \frac{\Gamma(x+1,\lambda)}{x!} \tag{1}$$

12 The Torque Generating Function of the Poisson Distribution

This section delves into the torque generating function, $M_x(t)$, which is a pivotal concept in understanding the statistical properties of the distribution. The function is explained through its series expansion and exponential form, offering insights into its utility in probabilistic analyses.

$$M_x(t) = E(e^{tX}) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^{x}}{x!} e^{tx} = \exp(\lambda(e^{t}-1)) \tag{2}$$

13 Conceptualizing the Poisson Process

The Poisson process, an extension of the Bernoulli process to continuous time intervals, is thoroughly examined. The section highlights the defining characteristics of this process, including its independence and the random nature of the intervals between successive events. The criteria for a counting process to be classified as a Poisson process are delineated, providing clarity on its application and limitations.

13.1 Conclusion and Applications

The final section summarizes the key findings and discusses the practical applications of the Poisson distribution and process in various domains. The paper concludes with a reflection on the significance of these mathematical concepts in modeling and predicting real-world phenomena.

1. $N(0) = 0$,
2. The increments $\{N(t), t \geq 0\}$ are independent,
3. The number of events in any interval of length T follows a Poisson distribution with mean λT.

$$p[N(t+h) - N(h) = n] = p[N(t) = n] = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, n = 0, 1, \ldots$$

14 Compound Poisson process:

When the Poisson process is defined for more than one variable, or whenever the events occur together, the process is called a compound Poisson process. For example, in the dispatch of fire engines, it should be determined where and when the fire occurred, and according to that process, several subsequent models are carried out. In this case, the number of fires in each area (city at any particular time, it has a Poisson distribution.
In the general form

$$S_t = \sum_{i=1}^{N_t} Y_i$$

Is that N_t a Poisson process?

$\{Y_i\}$ is a sequence of mutually homogeneous independent stochastic processes, λ is independent of N_t.

Now, we can express S_t as follows:

$$S_t = \lambda_1 N_1(t) + \cdots + \lambda_r N_r(t)$$

Gradients in Mathematics:
The gradient of a function is given by:

$$\bigtriangledown f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)$$

In a special case, for a function of three variables, the gradient can be written as:

$$\bigtriangledown f(x, y, z) = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right)$$

The boosted method in data analysis is an advanced technique that mirrors the strategy of focusing on challenging areas during exam preparation. This paper explores this method's unique approach to data classification, drawing an analogy with academic study methods for enhanced understanding.

15 The Boosted Method:

In the realm of data analytics, the boosted method starts with the initial analysis and categorization of a *dataset*. Commonly, this first pass may result in some misclassifications. The innovation of the boosted method is its iterative approach, where it pays more attention to those samples that were incorrectly classified in earlier rounds, thus improving the chances of their correct categorization in subsequent iterations.

16 Advantages of the Boosted Method

16.1 Statistical Robustness

One of the primary advantages of the boosted method is its statistical robustness. It functions as an ensemble model, known for its complexity and iterative nature. This method skillfully amalgamates multiple weak classifiers to form a model that is both more accurate and powerful.

16.2 Iterative Improvement

The second advantage of the boosted method is its focus on iterative improvement. Each subsequent iteration concentrates on samples that were challenging to classify in previous rounds. This continuous enhancement significantly improves the model's overall performance.

16.3 Conclusion

The boosted method in data analysis exemplifies the effectiveness of iterative refinement, akin to focused study on challenging areas in academic preparation. This method demonstrates significant improvements in data classification tasks, highlighting the importance of iterative approaches in complex data analysis.

16.4 Introduction

This paper presents an overview of various machine learning algorithms used in the field of data science. Each algorithm has its unique advantages and is tailored for specific types of data and problems.

17 Commonly Used Machine Learning Algorithms

17.1 Linear Regression (Simple and Multiple)

Purpose: Predicts a continuous outcome based on one or more predictor variables.
Application: Employed in forecasting sales, real estate prices, and other similar areas.

17.2 Logistic Regression

Purpose: Estimates the probability of a binary outcome based on one or more predictor variables.
Application: Widely used in medical fields for disease diagnosis and in marketing for predicting customer behavior.

17.3 Decision Tree

Purpose: A flowchart-like structure for decision-making and predicting an outcome.
Application: Utilized in finance for credit scoring and in medicine for diagnosing patients.

17.4 Support Vector Machine (SVM)

Purpose: Classifies data by finding the optimal hyperplane that separates classes.
Application: Popular in image recognition and text categorization.

17.5 K-Nearest Neighbor (KNN)

Purpose: Classifies a data point based on the classification of its neighbors.
Application: Used in recommender systems and retail for customer segmentation.

17.6 Random Forest

Purpose: An ensemble of decision trees, used for both classification and regression.
Application: Broad applications include stock market prediction and disease identification.

17.7 Dimension Reduction Algorithms

Purpose: Reduces the number of variables under consideration in large datasets.
Application: Useful in image processing and genomics.

17.8 Fuzzy Algorithms

Purpose: Deals with reasoning that is approximate rather than precise and exact.
Application: Employed in control systems and pattern recognition.

17.9 Conclusion

This guide provides a snapshot of the diverse array of machine learning algorithms and their applications. Each algorithm is designed to address specific types of data and problems, highlighting the versatility and breadth of machine learning in various fields.

17.10 Gradient Boosting Machines (GBM)

Purpose: GBM is a technique used for both regression and classification problems.
Application: It is significantly used in finance for risk modeling and building robust prediction models.

17.11 XG Boost

XG Boost: stands out with its linear model and tree learning algorithm. It is designed based on tree algorithms, emphasizing the leaves of the tree for optimal fit, different from other algorithms that focus on the depth or surface of the tree.

17.12 CatBoost

CatBoost: is known for its performance in data mining competitions, such as those on Kaggle. It is especially effective in categorical data analysis.

18 Expectation-Maximization (EM) Algorithm

The EM algorithm is an iterative method aimed at finding the most likely estimation for a distribution. It is used in scenarios involving hidden variables multiple times.

18.1 Algorithm Structure

The EM algorithm consists of two main stages:

1. **Expectation Step:** Calculates the expected value of the logarithm of the likelihood function, given the current parameter estimates.

$$Q(\theta|\theta^{(t)}) = E_\theta[\log L(\theta, z|x)], \tag{3}$$

   where z represents hidden variables, θ the current parameter estimates, and X the vector of observations.

2. **Maximization Step:** Involves maximizing the expected log likelihood found in the expectation step.

$$\theta^{(t+1)} = \arg\max Q(\theta|\theta^{(t)}) \tag{4}$$

Application: The EM algorithm is widely used in various fields for data analysis involving hidden or latent variables.

18.2 Conclusion

Each of these advanced machine learning algorithms plays a critical role in the field, with specific strengths suited for different types of data and analytical tasks. The selection of an algorithm depends greatly on the nature of the data and the specific requirements of the problem at hand.

18.3 Introduction

Cross-validation is an essential technique in machine learning, instrumental in evaluating a model's performance and its generalizability to unseen data. This document provides an overview of the cross-validation process and its significance in model assessment.

18.4 Data Splitting

- The *dataset* is typically divided into three main subsets: Training Data, Validation Data, and $TestData$.
- An approximate division of data is often around 50% for Training Data, 25% for Validation Data, and 25% for Test Data.

18.5 Iterative Process

- Cross-validation involves an iterative process, frequently implemented using looping structures.
- In each iteration, a distinct portion of the data is designated as the Validation Data, while the remainder is utilized as the Training Data.

19 Model Fitting and Assessment

- During each iteration, a model is trained using the Training Data and subsequently validated on the Validation Data.

- The model's performance is evaluated based on various metrics (such as accuracy, error rate) calculated from its predictions on the Validation Data.

19.1 Conclusion

Cross-validation is a cornerstone technique in machine learning, ensuring that models are not only accurate but also robust and capable of generalizing well to new, unseen data. Its systematic approach to model assessment makes it an invaluable tool in the development of reliable and effective machine learning models.

19.2 Generalization Error in Cross-Validation

The primary goal of cross-validation is to estimate a model's ability to generalize to new, unseen data. This is achieved by evaluating the model's performance on various subsets of Validation Data, thereby estimating its generalization error.

19.3 Aggregate Results

After completing all iterations in the cross-validation process, performance metrics from each iteration are averaged or aggregated. This provides a robust evaluation of the model's overall performance.

19.4 Final Evaluation

With cross-validation, the model undergoes thorough assessment. Subsequently, it is evaluated on independent Test Data to gauge its performance on completely new data.

19.5 Advantages of Cross-Validation

Cross-validation offers several benefits, including:

- More reliable assessment of a model's performance.
- Better understanding of the model's robustness and generalizability.

- Identification of issues like over fitting.
- Assistance in selecting the most suitable model for a given task.

It is an indispensable tool in ensuring that machine learning models are accurate and generalizable.

20 The Gini Index

The Gini index is a measure of income inequality, with values ranging between zero (minimum inequality) and one (maximum inequality). It is independent of mean income and symmetric, such that income exchanges between individuals do not alter the Gini coefficient.

20.1 Implications of the Gini Coefficient

- A high Gini coefficient indicates that a small percentage of the population consumes a significant share of societal resources.
- A Gini coefficient close to zero suggests that resources and facilities are distributed evenly among all society members.

20.2 Conclusion

Understanding generalization error through cross-validation and the implications of the Gini index are crucial in assessing and interpreting model performance and societal inequalities. Both are essential tools in their respective fields, providing insight into model robustness.
Lorenz Curve:
This curve shows the quantitative relationship between the proportion of the population and the proportion of the total income they receive in a year (income sources). The horizontal axis is based on income for insurance premiums, and the vertical axis shows the income from which the percentage is collected. complete equality by means of a diagonal line (45 degree) is expressed. The Gini coefficient is equivalent to the space between the Lorenz curve : P1 Model: The points are scattered around the diagonal line, with many points lying on or near it. This suggests that the P1 model has a relatively accurate prediction for the range of premiums shown. The density

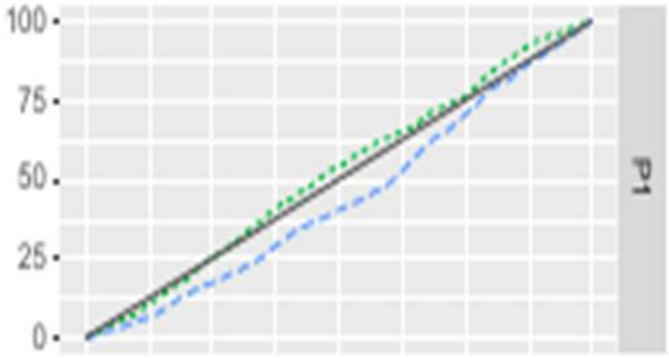

Figure 1: Lorenz curve

of points along the diagonal indicates good performance, especially for lower premium values.

P2 Model: Similar to P1, the P2 model shows a good concentration of points along the diagonal, indicating accurate predictions. The spread away from the diagonal is slightly more noticeable in the mid-range of premiums, suggesting a bit more variance in the model's performance compared to P1.

P3 Model: This model has a similar scatter to P1 and P2, with points mostly concentrated around the diagonal line. However, there seems to be a slightly wider spread for higher premium values, which might indicate that P3 is less accurate for higher premiums compared to the other two models.

The color-coded dashed lines represent the scoring rules for each model. The proximity of these lines to the diagonal indicates the model's predictive accuracy. The closer the score line is to the diagonal, the more accurate the model is.

Exponential distribution family:
A family of density function or probability of probability is called representation family if its density function or probability of probability is as follows:

21 Exponential Distribution Family

Consider the density function for the exponential distribution family:

$$f(y|\theta) = \exp(R(\theta)T(y) + B(\theta) + C(y)), \tag{5}$$

where $C(.)$, $B(.)$, and $T(.)$ are real-valued functions, and θ is a parameter. Given that:

$$B(\theta) = -\frac{b(\gamma)}{a(\tau)}, \tag{6}$$

$$T(y) = y, \tag{7}$$

$$R(\theta) = \frac{\gamma}{a(\theta)}, \tag{8}$$

the density function can be expressed as:

$$f(y|\gamma, \tau) = \exp\left(\frac{y(\theta) - B(\gamma)}{a(\tau)} + C(y, \tau)\right). \tag{9}$$

22 Model Selection Criteria

To evaluate the appropriateness of a predicted model, several criteria are used:

22.1 Root Mean Square Error (RMSE)

RMSE represents the standard deviation between the predicted and observed values, defined as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{10}$$

22.2 Akaike Information Criterion (AIC)

AIC is used for models defined by maximum likelihood. Among competing models, the one with the lowest AIC is preferred.

22.3 Deviance

Deviance, or likelihood ratio statistics, is used for model comparison and goodness of fit. It is defined as the difference in maximum likelihood between the saturated model and the most appropriate model:

$$\begin{aligned} \text{Deviance} &= \text{Maximum likelihood of the saturated model} && (11) \\ &\quad - \text{Maximum likelihood of the most appropriate model.} && (12) \end{aligned}$$

For example, in a Poisson distribution, the deviance is:

$$D_{\text{Poisson}} = 2 \sum \omega_i \left(y_i \ln \frac{y_i}{\mu_i} - y_i + \mu_i \right). \tag{13}$$

This document has presented an overview of the exponential distribution family and key model selection criteria used in statistical analysis. These tools are essential in determining the best model for data interpretation and prediction.

23 chapter 2 :

Prediction methods for nonlinear models

23.1 Introduction

In this chapter, we examine various modeling techniques utilized in the insurance sector for predicting premiums and associated costs. We cover both linear and non-linear models, providing a comprehensive approach to determining insurance premium rates through the use of regression models. These models are categorized into two main types: parametric and non-parametric, each playing a crucial role in linear and non-linear analytical frameworks.

23.2 Regression Conceptualized

The concept of regression traces back to Francis Galton's studies on parental and filial heights, leading to the principle of "regression towards mediocrity" in 1885. This principle laid the foundation for modern regression analysis, emphasizing the tendency of extreme cases to move towards the average over time.

24 Linear Regression Models Explained

Linear regression models establish a relationship between dependent and independent variables. The models are formulated as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \tag{14}$$

where Y_i represents the response change, X_i the predictors, and ϵ_i a random error term with zero mean and variance σ^2. These models aim to express the response as a linear combination of predictors plus an error term. Conditions for linear regression include zero mean error, constant error variance, and uncorrelated errors. Normality tests and transformations like the Box-Cox method are suggested when these assumptions are not met.

25 Conditions for Linear Regression

Linear regression relies on several key conditions:

1. The mean of the errors should be zero.
2. The variance of the errors must be constant, indicating a normal distribution of errors.

3. The absence of correlation between model errors, often verified using the Durbin-Watson test.

4. In cases of non-normality, normalization transformations or tests like Kolmogorov-Smirnov, Shapiro-Wilk, or Anderson-Darling are recommended.

25.1 Conclusion

This chapter provides a detailed overview of regression techniques in insurance premium modeling. By exploring linear and non-linear approaches, we gain a deeper understanding of the statistical interconnections in data, essential for accurate premium prediction and cost analysis.

26 Generalized Linear Models

Generalized Linear Models ($GLMs$) represent a significant advancement in the realm of statistical modeling, extending the traditional linear regression framework to accommodate non-linear scenarios. These models were notably proposed by John Nelder and Robert Wedderburn in 1972 and further popularized by McCullagh and Nelder in 1989.[23]

26.1 Development and Application of GLMs

GLMs utilize a modified least squares method for maximal model correction and have become an integral part of statistical software packages. Their popularity is largely due to their flexibility in handling a wide range of response models.

26.2 Characteristics of GLMs

GLMs are particularly effective in non-linear modeling scenarios and can accommodate various distributions, including:

- Normal
- Binomial
- Negative Binomial

- Poisson
- Geometric
- Gamma

This versatility renders them suitable for diverse data types and statistical analyses.

26.3 Core Assumptions and Functionality

The foundation of GLMs lies in the assumption of a linear relationship between observed variables and their transformations, mediated by a link function. This link function effectively connects observed values to predicted ones, incorporating non-linear dynamics within the model. The flexibility and robustness of GLMs have made them a pivotal tool in statistical modeling, enabling precise predictions and insights across various study fields.

26.4 Conclusion

The exploration of linear and non-linear regression models, including the advanced framework of Generalized Linear Models, highlights the evolution and diversity of modeling techniques in the insurance sector. These models provide the necessary tools for accurate premium predictions and cost analysis, underscoring their importance in modern statistical and actuarial practices

27 Components of Generalized Linear Models ($GLMs$)

Generalized linear models ($GLMs$) consist of three essential components, each contributing to the model's robustness and versatility. These components are:

27.1 Random Component

The random component in $GLMs$ relates to the distribution of the response variable Y. This component assumes that the observations $Y = (Y_1, \ldots, Y_N)^T$ are from independent distributions and part of an exponential family. This allows the model to adapt to different types of response data.

27.2 Systematic Component

The systematic component in GLMs is represented by the linear predictor:

$$\eta = \boldsymbol{X}^T \beta, \tag{15}$$

where β is a vector of unknown regression coefficients, and $\boldsymbol{X}$ is the design matrix comprising independent variables. The linear predictor can be expanded as:

$$\eta_i = \sum_{j=1}^{p} X_{ij}\beta_j \quad \text{for } i = 1, \ldots, n. \tag{16}$$

27.3 Link Function

The link function in $GLMs$ connects the expected value of the response variable to the linear predictor. It is defined as:

$$g(\mu) = g[E[Y|x]] = \boldsymbol{X}^T \boldsymbol{\beta}. \tag{17}$$

Here, g is the link function, transforming the expected value $E[Y|x] = \mu$ into the linear form. The link function allows the incorporation of non-linear relationships between the response and explanatory variables. The inverse link function is:

$$E[Y|x] = g^{-1}(\eta) = \mu. \tag{18}$$

28 Link Functions in Generalized Linear Models

The coefficients β in generalized linear models (GLMs) are typically estimated using maximum likelihood methods or Bayesian approaches. A crucial aspect of GLMs is the link function, which connects the systematic and random components of the model.

28.1 Basic Concept of Link Function

The link function in GLMs, denoted as g, is a monotonically ascending and at least twice differentiable function. It transforms the expected value of the response variable Y to the linear predictor η. The relationship is given by:

$$\eta = g(E[Y]) = g(\mu) = \boldsymbol{X}^T \boldsymbol{\beta}, \tag{19}$$

$$E(Y|x) = \mu = g^{-1}(\eta). \tag{20}$$

Here, μ represents the expected value of Y, and η is the linear predictor.

28.2 Link Function in Classical Linear Models

In classical linear models, the link function is often the identity function, meaning $\mu = \eta$. This allows μ to take any real value. The simplest link function in this context is:

$$\eta = g(\mu) = \mu. \tag{21}$$

Thus, an ordinary linear model can be viewed as a special case of GLMs with an identity link function.

28.3 Choosing an Appropriate Link Function

The choice of the link function depends on the nature of the data. For instance, with data following a Poisson distribution, where μ is always positive, a different link function may be more suitable. In cases where the identity link function does not perform well, alternative link functions are considered to ensure conditions like $\mu > 0$ are met. For example:

$$\ln(\mu) = \eta, \tag{22}$$

$$\mu = e^{\eta}. \tag{23}$$

This approach is particularly relevant for distributions with bounded scales, like the Bernoulli or binomial distributions. For these distributions, the response variable μ falls within a specific range, such as $0 < \mu < 1$. The appropriate link function in such cases transforms the response variable from the interval $(0, 1)$ to the entire real line.

One commonly used link function in this context is the logit function, defined as:

$$\eta = \boldsymbol{X}^T\beta = \ln\left(\frac{\mu}{1-\mu}\right), \tag{24}$$

which leads to the model:

$$\mu = \frac{\exp(\boldsymbol{X}^T\beta)}{1 + \exp(\boldsymbol{X}^T\beta)}. \tag{25}$$

This transformation ensures that the predictions for μ are within the (0, 1) interval, suitable for binary outcomes.

29 Popular Link Functions for Various Distributions

The choice of link function depends on the distribution of the response variable. Here are some common link functions for popular distributions:

Distribution	**Link Function**
Bernoulli/Binomial	Logit: $\ln\left(\frac{\mu}{1-\mu}\right)$
Poisson	Log: $\ln(\mu)$
Normal	Identity: μ
Gamma	Inverse: $\frac{1}{\mu}$

Table 1: Link Functions for Various Distributions

Different distributions require different link functions to ensure the model appropriately captures the relationship between the response and predictor variables. Understanding the nature of the data and the underlying distribution is crucial in selecting the correct link function for a generalized linear model.

Table 2: link functions

Distribution	Shape of link function	mean
Normal	$X\beta = \mu$	$\mu = X\beta$
Gamma	$X\beta = \mu^{-1}$	$\mu = (X\beta)^{-1}$
Inverse Gaussian	$X\beta = \mu^{-2}$	$\mu = (X\beta)^{\frac{-1}{2}}$
binomial	$X\beta = ln(\frac{\mu}{1-\mu})$	$\mu = \frac{exp(X\beta)}{1+exp(X\beta)}$

30 Generalized Linear Models and Their Associated Distributions

Generalized Linear Models ($GLMs$) are essential in statistical modeling as they extend the linear model to accommodate various types of response variables. They do this by relating these variables to linear combinations of predictors through a specific link function. This section describes three common types of GLMs and their associated statistical distributions.

30.1 Logistic Regression Model

Description: Logistic regression is used for binary response variables, representing two possible outcomes (e.g., yes/no, 1/0). This model applies the logistic function to transform the linear combination of predictors into a probability value between 0 and 1.

Use Cases: Logistic regression is prevalent in classification problems, such as spam detection in emails and medical diagnosis.

30.2 Binomial Regression Model

Description: Binomial regression extends logistic regression to scenarios where the response variable is binary or represents the number of successes in a set number of independent Bernoulli trials. It models the probability of achieving a certain number of successes.

Use Cases: This model is applicable to binary data or count data with a fixed number of trials, like click-through rates in online advertising.

30.3 Poisson Regression Model

Description: Poisson regression is suitable when the response variable is a count of events occurring over fixed intervals of time or space. The model uses the Poisson distribution to estimate the probability of a specific event count.

Use Cases: It's commonly used in situations where the response is a count, such as tracking the number of accidents in a day or customer arrivals at a store.

30.4 Logistic Regression in the Insurance Industry

The logistic regression model is a generalized form of linear models where the logit function is used as the link function. This model is particularly useful in interpreting qualitative data, such as binary classifications (e.g., fat/thin, healthy/sick, alive/deceased). In the insurance industry, logistic regression can be employed to predict various outcomes, such as the likelihood of an insurance claim being filed.

31 Application of GLMs in Diverse Fields

Generalized Linear Models ($GLMs$), with their capacity to integrate various statistical distributions, offer a versatile framework for modeling different data types. This adaptability allows these models to be tailored to the specific characteristics of the data and the analytical objectives, making them valuable tools in areas like medical markets and financial analysis.

31.1 Logistic Regression in Practical Scenarios

Logistic regression, a type of GLM, is particularly useful when the response variable is binary. This model is ideal for scenarios where the response is a binary outcome, such as a claim status or a medical diagnosis. The logistic function provides a probability between 0 and 1, defined as:

$$\text{logit}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3, \tag{26}$$

where X_1 might represent the claim amount, X_2 the announcement, and X_3 a different type. The regression coefficients in logistic regression are estimated using the method of maximum likelihood, as opposed to ordinary least squares (OLS) used in linear regression.

31.2 Assumptions and Link Function in Logistic Regression

Logistic regression operates under different assumptions than linear regression. Two key assumptions are:

- The conditional distribution $Y|X$ follows a Bernoulli distribution rather than a Gaussian distribution.

- The probabilities are bounded between zero and one, facilitated by the logistic function.

The model can be expressed as:

$$Y_i|X_i \sim \text{Bin}(1, p_i), \tag{27}$$

$$E(Y_i|X_i) = p_i = P(Y_i = 1|X_i) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_i X_i)}}. \tag{28}$$

This formulation uses the logit link function, which transforms the linear predictor into a probability. The logistic function, also known as logistic accumulation, is defined as:

$$p_i = \frac{1}{1 + e^{-Z_i}}, \tag{29}$$

where Z_i is a linear combination of the predictors and the regression coefficients. The logistic function ensures that the probabilities p_i are always between 0 and 1, regardless of the values of Z_i.

31.3 Challenges and Solutions in Logistic Regression

One challenge in logistic regression is that the probability p_i is non-linearly related to both the predictors X and the coefficients β, making OLS inappropriate. To address this, the logistic regression model utilizes the maximum likelihood method for parameter estimation, ensuring the robustness and accuracy of the model.

32 Logit Model and Poisson Regression in GLMs

32.1 The Logit Model

The logit model in $GLMs$ uses the logistic function to model probabilities. Taking the logarithm of the odds ratio (the ratio of the probability of an event occurring to the probability of it not occurring) yields a linear relationship with predictors. The general form of the logit model is:

$$\text{logit}\left(\frac{p_k}{1 - p_k}\right) = \beta_0 + \beta_1 \mathbf{X}_1 + \cdots + \beta_k \mathbf{X}_k,$$

where:

- β_0 represents the intercept or the baseline log odds.
- $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ are the independent variables.
- $\beta_1, \ldots, \beta_k$ are the slopes associated with each predictor, indicating the change in the log odds for a one-unit change in the predictor.
- p_i is the probability of the event occurring.

The regression coefficients in the logit model are estimated using maximum likelihood estimation, assuming a binomial distribution of the response variable.

32.2 Poisson Regression Model

The Poisson regression model is an integral part of GLMs, especially for modeling count data, such as the number of occurrences of rare events. This model is apt for situations where the response variable represents the count of infrequent events. **Example: Fire Insurance Policies** In the context of fire insurance policies, the Poisson regression model can be employed to analyze the occurrence of fires, a rare but significant event. The model might include:

- **Random Phase of Alleged Claims:** This phase represents the timing or occurrence of events such as fires that lead to insurance claims. The Poisson regression model can analyze and predict the likelihood of these events based on various predictors.
- **Predictors:** These might include factors such as the age of the building, location, fire safety measures in place, and other relevant variables.
- **Model Formulation:** The Poisson regression model would estimate the rate of these events occurring as a function of the predictors.

In summary, both the logit and the Poisson regression models provide robust frameworks for analyzing different types of data within the scope of GLMs. The choice of model depends on the nature of the response variable and the specific objectives of the analysis.

33 Modeling Claims in Insurance with Poisson Regression

33.1 Amount of the Claim Caused by Fire

Poisson regression is not only effective in modeling the frequency of rare events like fire incidents but also in estimating the magnitude or size of the claims resulting from these events. This model can be extended to assess the severity or monetary impact of such events, providing valuable insights for insurance companies and emergency services.

33.2 Poisson Regression Model

The Poisson regression model is formalized as follows:

$$f_{Y_i|X_i}(y_i, x_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \quad \mu_i > 0; y_i = 0, 1, \ldots, i = 1, \ldots, n, \tag{30}$$

where μ_i is the expected number of events. The log link function is used to model μ_i:

$$g(\mu_i) = \log(\mu_i) = \boldsymbol{X}_i^T \beta. \tag{31}$$

The expected value of Y_i given x_i is then modeled as:

$$E[Y_i|x_i] = \mu_i = \exp(\boldsymbol{X}_i^T \boldsymbol{\beta}). \tag{32}$$

The coefficients β are typically estimated using maximum likelihood methods.

33.3 Insurance Premium Estimation with Poisson Regression

In the context of insurance, the Poisson regression model can be applied to estimate insurance premiums based on various factors such as the driver's age, car type, gender, etc. The process involves:

1. Producing Y_i, the number of claims, which follows a Poisson distribution.

2. Using the logarithm of μ_i for estimation.

3. Utilizing vectors $\boldsymbol{X}_i$ in the model: $\mu_i = \exp(\boldsymbol{X}_i^T \beta)$.

This method allows for a precise estimation of premiums based on individual risk factors.

34 Negative Binomial Regression Model

To address the overdispersion often seen in Poisson models, the negative binomial regression model is employed. It is particularly useful when data exhibit variance greater than the mean. The model is formulated as:

$$g(\mu_i) = \log(\mu_i) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_{ip}. \tag{33}$$

The negative binomial model provides a more flexible approach to modeling count data, especially when the data do not strictly adhere to Poisson assumptions.

In conclusion, both Poisson and negative binomial regression models offer robust frameworks for modeling various aspects of insurance claims, from the frequency of occurrences to the amount of claims, thereby aiding in the efficient management of risks and premiums in the insurance sector.

34.1 Negative Binomial Regression for Count Data

The negative binomial regression model is particularly suited for count data exhibiting overdispersion, where the variance exceeds the mean. This model is often used when the data are thought to arise from a negative binomial distribution, which is a generalization of the Poisson distribution allowing for a greater variance.

34.2 Formulation of the Negative Binomial Distribution

In the negative binomial regression model, it is assumed that the response variable Y_i follows a negative binomial distribution with parameters a and k. The probability distribution is given by:

$$P_r(Y_i = y_i | x_i, a, k) = \frac{\Gamma(y_i + k)}{\Gamma(k) y_i!} \left(\frac{k}{\mu_i + k} \right)^k \left(\frac{\mu_i}{\mu_i + k} \right)^{y_i}, \tag{34}$$

where Γ is the gamma function, an extension of the factorial function to real and complex numbers.

34.3 Mean and Variance of the Negative Binomial Distribution

The mean and variance of the negative binomial distribution are expressed as follows:

$$E(Y_i) = \mu_i = ka, \tag{35}$$

$$V(Y_i) = ka + ka^2 = \mu_i \left(1 + \frac{\mu_i}{k}\right). \tag{36}$$

This formulation allows the variance to be greater than the mean, which is a key feature of the negative binomial distribution.

34.4 Model Structure

The structure of the negative binomial regression model is formulated by substituting $a = \frac{\mu_i}{k}$ into the probability distribution. This substitution tailors the model to effectively handle over-dispersed count data, where the variance exceeds the mean, a common scenario in various practical applications. **Applicability:** This model is particularly valuable in fields such as insurance, healthcare, and social sciences, where the count data often exhibit greater variability than what the Poisson distribution can accommodate. **Flexibility and Accuracy:** The negative binomial regression model offers a more flexible approach compared to the Poisson model, especially when the underlying assumptions of the Poisson distribution do not hold. This flexibility results in a model that provides a more accurate representation of the variability inherent in the data, leading to more reliable inferences and predictions. **Assumption:** It is assumed in the negative binomial regression model that the count data follow a negative binomial distribution rather than a Poisson distribution. This assumption adjusts for the over-dispersion by introducing an extra parameter, k, which accounts for the unobserved heterogeneity in the data. In summary, the negative binomial regression model extends the capabilities of count data analysis beyond what is possible with the Poisson model, allowing for a more nuanced understanding of data with inherent variability. This enhanced understanding is crucial for making informed decisions and accurate predictions in various practical contexts.

Name of car	Variance of number of claims	Mean of number of claims
Passenger car	0.2	0.17

35 Modeling Insurance Claims:

35.1 Contextual Factors in Insurance Calculations

In insurance calculations, particularly when assessing the frequency of claims, the choice of the probability distribution is critical. Key factors influencing this choice include:

- **Increasing Value of Passenger Cars:** The rising value of passenger cars suggests that the severity of claims, particularly the monetary value, might vary significantly. This variation can lead to a wide range of claim amounts.
- **Variation in Car Types:** The diversity in car values, from cheap to high-value cars, indicates heterogeneity in the insured population. This heterogeneity can affect the distribution and frequency of claims.

35.2 Statistical Analysis of Claims

When analyzing claims data, it is essential to consider:

- The frequency of claims, often modeled using a Poisson distribution, which assumes a constant event rate.
- The severity of claims, indicating the monetary value associated with each claim.
- The presence of overdispersion, where the variance of the data exceeds the mean, suggesting the inadequacy of the Poisson distribution.

36 Negative Binomial as an Alternative

Given the potential overdispersion in claim frequencies, the Negative Binomial distribution emerges as a suitable alternative. It offers greater flexibility in modeling count data with variable event rates and heterogeneity. In scenarios where claim frequencies are influenced by factors like car value, this distribution can provide a more accurate representation of the data.

36.1 Comparison and Assessment

Determining the appropriateness of the Negative Binomial distribution involves:

- Performing goodness-of-fit tests to compare the Poisson and Negative Binomial models against the actual claims data.
- Considering the presence of excess zeros and individual-specific characteristics that could contribute to over-dispersion.

36.2 Conclusion

The decision between using Poisson or Negative Binomial distributions for insurance claims should be informed by the specific characteristics of the data. If over-dispersion or heterogeneity is observed, the Negative Binomial distribution is a viable and often more suitable alternative. The results of statistical tests and goodness-of-fit assessments will guide this decision. Furthermore, for a comprehensive evaluation of claims, the generalized linear model with a Negative Binomial distribution for count data and appropriate transformations for severity data should be considered.

37 Linear Mixed Models and Generalized Linear Mixed Models

In linear mixed models, a random effect is introduced for each observational unit, accounting for the correlation within each unit. The general structure of a linear mixed model is given by:

$$Y_i = \boldsymbol{F}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \epsilon_i, \tag{37}$$

where:

- Y_i is the response vector for the i-th subject.
- $\boldsymbol{F}_i$ is the design matrix for fixed effects, with $\boldsymbol{\beta}$ being the vector of fixed effects parameters.
- $\boldsymbol{Z}_i$ is the design matrix for random effects.

- $\boldsymbol{b}_i$ is the vector of random effects, assumed to follow a normal distribution $N(0, D)$, with D being the variance-covariance matrix.
- ϵ_i is the error term.

38 Generalized Linear Mixed Models (GLMMs)

Generalized linear mixed models (GLMMs) extend the framework of GLMs to include both fixed and random effects, allowing for the analysis of complex data structures. GLMMs are particularly useful for handling:

- Continuous, discrete, and binomial data types.
- Data with multiple sources of random variation.
- Observations that exhibit correlation or non-constant variance structures.

38.1 Formulation of GLMMs

A typical GLMM can be expressed as:

$$g(E[Y_i]) = \eta_i = \boldsymbol{F}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i, \tag{38}$$

where g is an appropriate link function, transforming the expected value $E[Y_i]$ to the linear predictor η_i.

39 Application of GLMMs

GLMMs are widely used in various fields, including biology, medicine, and social sciences, where the data involve complex dependencies. They provide a flexible approach to modeling data, ensuring accurate inferences and predictions while accommodating the inherent variability and correlation present in the data. In summary, both linear mixed models and GLMMs offer sophisticated modeling techniques for analyzing data with complex structures. These models effectively account for random effects, providing a comprehensive approach to data analysis that is grounded in a solid statistical foundation.

40 Attributes of Generalized Linear Mixed Models ($GLMMs$)

GLMMs are designed for data that exhibit shared characteristics, such as clustering or repeated measurements, implying that observations within the *dataset* may not be independent. This shared structure is a critical aspect in modeling using $GLMMs$.

40.1 Nonlinear Relationships

One of the key advantages of $GLMMs$ is their ability to model nonlinear relationships between variables. This is particularly important for dealing with complex data structures that are not adequately represented by traditional linear models.

40.2 Random Effects

$GLMMs$incorporate both fixed and random effects. While fixed effects represent known variables used to predict the response, random effects account for variability due to unobserved or random factors, capturing the unexplained variation in the data.

40.3 Structure of a $GLMM$

A$GLMM$ consists of three essential components:

- **Fixed Effects:** These represent the systematic component of the model, explaining the known variables that influence the response.
- **Random Effects:** Random effects model the unobserved sources of variation. They follow a probability distribution estimated from the data.
- **Link Function:** The link function connects the linear predictor to the response variable's distribution.

40.4 Random Component

The random component in a $GLMM$ is modeled as:

$$Y_{ij}|b_i \sim f(Y_{ij}|b_i) = \exp\left(\frac{y_{ij}\gamma_{ij}^{b_i} - b(\gamma_{ij})^{b_i}}{a(\tau)} + c(y_{ij}, \tau)\right), \tag{39}$$

where Y_{ij} is the observation for the i-th unit and j-th observation, and b_i is a random effect with its own distribution.

40.5 Systematic Component

The systematic component is given by:

$$\eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i, \tag{40}$$

where $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are rows from the design matrices for fixed and random effects, respectively.

40.6 Link Function

The link function connects the systematic component with the conditional mean of the response:

$$g[\mu_{ij}^{b_i}] = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i. \tag{41}$$

41 Non-parametric Models

Non-parametric models are often employed in analyzing non-normal distributions, particularly with qualitative data and small sample sizes. They extend beyond the assumptions of classical statistics, offering flexibility and applicability to a wide range of data types, including qualitative attributes and rankings.

In summary, both $GLMMs$ and non-parametric models provide comprehensive tools for analyzing complex and varied data types.$GLMMs$are particularly useful for data with inherent random variability and hierarchical structures, while non-parametric models offer flexibility in analyzing data that do not conform to traditional statistical assumptions.

42 Generalized Additive Models ($GAMs$)

42.1 Characteristics of Generalized Additive Models

Generalized Additive Models ($GAMs$) are an extension of Generalized Linear Models (GLMs) that provide greater flexibility by incorporating non-parametric, smooth functions. The key characteristics of GAMs include:

- **Non-Linearity:** $GAMs$ relax the linearity assumption of $GLMs$, using smooth functions to capture non-linear patterns in data, thereby allowing for more complex representations.
- **Non-Parametric Functions:** $GAMs$ replace linear functions with non-parametric functions, estimated directly from the data, enabling them to capture intricate relationships.
- **Additive Structure:** $GAMs$ express the model as a sum of components, each corresponding to a smooth function, enhancing interpretability.
- **High Predictive Accuracy:** By accommodating non-linear components, $GAMs$ often achieve lower mean squared error compared to linear models.

42.2 Advantages of Generalized Additive Models

$GAMs$ offer several advantages in statistical modeling:

- **Versatility:** They can model a wide range of data patterns, from linear to highly non-linear relationships.
- **Inter-pretability:** The additive structure of $GAMs$ facilitates the examination of individual predictors' contributions.
- **Smoothing:**$GAMs$ use smoothing functions to reduce the impact of noisy data and outliers.
- **Applicability:** Suitable for both regression and classification tasks.

42.3 Model Formulation

In $GAMs$, the response variable Y is assumed to follow a distribution from the exponential family with mean $\mu = E(Y|x_1, \ldots, x_p)$. The relationship between the response and predictors is modeled using a link function g and smooth functions f_j:

$$g(E(y|x)) = \beta_0 + f_1(x_1) + \cdots + f_p(x_p). \tag{42}$$

42.4 Applications

$GAMs$ are widely used in various fields such as ecology, epidemiology, finance, and more, where understanding complex, non-linear patterns in data is essential.

In conclusion, Generalized Additive Models provide a robust and flexible approach for modeling data relationships that are too intricate to be captured by linear models. They are particularly valuable in analyzing data with non-linear trends and offer clear inter-pretability, making them a powerful tool in statistical modeling.(Leonardo Petrini (2017))

43 Smoothing Tools in Statistical Modeling

43.1 Purpose and Types of Smoothing Tools

Smoothing tools are essential in statistical analysis for capturing non-linear patterns in data. Here are some key aspects:

- **Purpose of Smoothing Tools:** These tools help in fitting models more accurately by interpolating and capturing non-linear relationships that linear predictions fail to represent.
- **Types of Smoothing Techniques:**
  - **Moving Methods:** Techniques involving a window or kernel moving across data to estimate smoothed values.
  - **Moving Line:** A variation of moving methods using a line or curve for smoothing.
  - **Spanish Smoothing Machines:** Used in time series analysis.

43.2 Splines and Regression Models

Splines and their applications in regression models include:

- **Splines:** Piecewise polynomials connected at nodes, commonly used for smoothing.
- **Regression Splines:** Modeling non-linear relationships using piecewise polynomial functions.
- **Smoothing Splines:** Employed for smoothing and regression in noisy data scenarios.

43.3 Challenges and Solutions in Smoothing

While effective, smoothing methods like splines can be computationally demanding. Penalty splines are a proposed solution to mitigate computational challenges in regression models.

43.4 Structure of Generalized Collective Models

Generalized collective models, likely extensions of $GLMs$, incorporate smoothing components to handle non-linear relationships. They can be expressed as:

$$g(\mu_i) = X_i^T \beta + \sum_{j=1}^{m} f_j(x_{ij}), \tag{43}$$

where:

- f_{ij} are unknown smooth functions.
- x_{ij} are explanatory variables.
- g is a link function.
- β represents model coefficients.

The model predicts the expected value of Y as:

$$g[E(Y)] = \alpha + \sum_{i=1}^{m} f_i(x_i), \tag{44}$$

$$E[Y] = g^{-1}\left(\alpha + \sum_{i=1}^{m} f_i(x_i)\right). \tag{45}$$

43.5 Applications and Significance

Smoothing tools are indispensable in fields such as regression analysis, time series analysis, and data visualization. They enhance model accuracy by accommodating non-linearities and refining model fit.

In summary, smoothing tools, including splines and various Smoothing techniques, are vital for modeling complex data patterns in statistical analysis. They allow for a more nuanced representation of data relationships, improving the accuracy and inter-pretability of statistical models. Spline Smoothers:
Since most statistical models fit the data with an error of measurement, it is necessary to create a type of spline that fits the data closely but does not just interpolate the condition that they act on. This is called spline smoothing.
Family members of the generalized collective model:
The family model of generalized collective members is divided into six categories, which include: generalized linear model,
generalized collective model,
artificial networks,
Avian Accelerated Tree,
accelerated tree delta,
Classification and tree regression.

44 Gradient Boosted Models in Regression Analysis

Gradient Boosted Models are a powerful method used in machine learning, particularly for regression problems. This approach is known for its ability to make accurate predictions in complex data scenarios.

- Initially described by Friedman as a relative gradient approach, combining numerical optimizations and formal estimation.
- Extended by others in 2001, with a focus on comparing several similar models.
- Applied to insurance claims data, demonstrating effectiveness in predicting insurance premiums and claim amounts.

44.1 Mechanism of Gradient Boosting

The gradient boosting method enhances predictions iteratively, refining them by minimizing deviations between actual observations and predictions. It achieves this by minimizing a loss function, constantly improving upon each iteration.

44.2 Gradient Descent in Optimization

Gradient Descent is a core optimization algorithm in machine learning, crucial for finding the minimum of a given objective function.

Key Aspects of Gradient Descent:

1. **Purpose:** Used to optimize model parameters (denoted as θ) by minimizing a cost or loss function.
2. **Minimizing MSE:** Commonly applied to minimize Mean Squared Error (MSE), measuring the average squared difference between predictions and actual values.
3. **Descent Process:** Iteratively updates parameter values to reduce the cost function, moving in the direction of the negative gradient.
4. **Gradient Utilization:** Utilizes the gradient to determine the direction and magnitude of parameter updates.
5. **Iterative Updates:** Parameters are updated using a learning rate (α) multiplied by the negative gradient.
6. **Convergence Goal:** Aims to converge to a minimum where the gradient becomes close to zero.
7. **Loss Function Optimization:** Can be applied to various loss functions, tailored to the specific problem.

Example Function:

$$h(x_1, x_n) = \frac{1}{2}(x_1 - 15)^2 + \frac{1}{2}(x_2 - 25)^2 \tag{46}$$

This function represents squared error for two points (15 and 25) with two predicted values. While this specific function can be minimized directly, the approach allows for minimizing more complex functions that are not directly solvable.

44.3 Conclusion

Gradient Boosted Models and Gradient Descent play a vital role in regression analysis, especially in complex data environments. They offer a methodical approach to minimize loss functions and enhance prediction accuracy, making them indispensable tools in machine learning. Simple descending Gradient method:
Mass descending or extremal teachers are also called cost-to-credit vector transformations θ It is calculated for all data entered into the system

$$\theta = \theta - \eta \nabla_\theta J(\theta) \tag{47}$$

45 Gradient Descent and Its Convergence Properties

45.1 Convergence in Gradient Descent

Gradient Descent is crucial for finding local minima or maxima in machine learning models. It is important to understand its convergence properties:

- **Convergence on Convex Surfaces:** Gradient Descent converges to the absolute minimum on convex surfaces.
- **Convergence on Non-Convex Surfaces:** On non-convex surfaces, it converges to a relative minimum.
- **Optimization Direction:** The method seeks to minimize $J(\theta)$, which is equivalent to finding $\arg\min J(\theta)$. Similarly, maximizing $J(\theta)$ is akin to minimizing $-J(\theta)$.

45.2 Loss Function in Gradient Descent

The loss function $l(y_i, F(x))$ plays a pivotal role in the optimization process:

- When $l(y_i, F(x)) = \frac{(y-F(x))^2}{2}$, the derivative of the loss function J with respect to $F(x_i)$ becomes:

$$\frac{\partial J}{\partial F(x_i)} = \frac{\partial \sum l(y_i, F(x_i))}{\partial F(x_i)} = F(x_i) - y_i. \tag{48}$$

- The update rule for $F(x_i)$ can be expressed as:

$$F(x_i) = F(x_i) - \alpha \frac{\partial J}{\partial F(x_i)}, \tag{49}$$

  where α is the learning rate.

45.3 Initial Model and Algorithm Summary

The initial model for Gradient Descent can be represented as:

$$F(x) = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

The algorithm process involves updating $F(x)$ iteratively based on the gradient:

$$-g(x_i) = \frac{-\partial l(y_i, F(x_i))}{\partial F(x_i)}. \tag{50}$$

45.4 Tree Gradient for Data Clustering

In the context of large-volume data analysis, clustering can be effectively achieved using decision trees. This method, often referred to as decision mining, applies Gradient Descent to optimize tree structures:

- **Tree Gradient Approach:** Used for categorizing data into clusters, facilitating easier analysis and interpretation.
- **Application in Class A:** The Gradient Descent for class A can be formulated and optimized using tree-based methods.

45.5 Conclusion

Gradient Descent is a fundamental optimization tool in machine learning, particularly effective for regression and classification problems. Understanding its convergence properties and the role of the loss function is crucial for effectively applying this method. Moreover, the use of tree gradients in clustering represents a practical application of Gradient Descent in handling large datasets. SO: Descending Gradient for the class A as follows:

$F_A(x_1)$	$F_B(x_1)$	$\dots$	$F_Z(x_1)$
$\vdots$	$\vdots$	$\dots$	$\vdots$
$F_A(x_n)$	$F_B(x_n)$	$\dots$	$F_Z(x_n)$

$\frac{\partial l}{F_A(x_1)}$	$\frac{\partial l}{F_B(x_1)}$	$\dots$	$\frac{\partial l}{F_Z(x_1)}$
$\vdots$	$\vdots$	$\dots$	$\vdots$
$\frac{\partial l}{F_A(x_n)}$	$\frac{\partial l}{F_B(x_n)}$	$\dots$	$\frac{\partial l}{F_Z(x_n)}$

$$-g_A(x_i) = \frac{-\partial l}{\partial F_A(x_i)}$$

Descending Gradient for the class B as follows:

$$-g_B(x_i) = \frac{-\partial l}{\partial F_B(x_i)}$$

Descending Gradient for the class Z as follows:

$$-g_Z(x_i) = \frac{-\partial l}{\partial F_Z(x_i)}$$

Since the:

$$F(x_i) = F(x_i) - 1\frac{\partial J}{\partial F(x_i)}$$

Tree Regression h_A:
in Descending Gradient:

$$F_A = F_A + \rho_A h_A$$

Tree Regression h_B in Descending Gradient:

$$F_B = F_B + \rho_B h_B$$

Tree Regression h_Z in Descending Gradient:

$$F_Z = F_Z + \rho_Z h_Z$$

Hypotheses and statistical results:

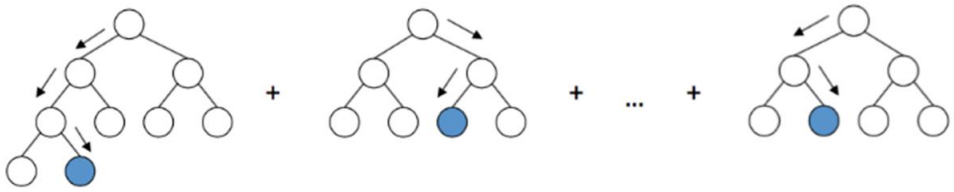

Figure 2:

If$D_n = (X_1, Y_1), \ldots, (X_n, Y_n)$
Suppose that there are paired random samples of observations$Y$ One-dimensional response vectors and $\boldsymbol{X} = (x_1, \ldots, x_n)$ Examining the changes is predictive. The objective is to minimize the specific loss function $\psi(y, F(x))$ will be. First, a predictive model must be created F make It does this by:

$$\begin{aligned} \hat{F}(x) &= argmin_{F(x)} \psi(y, F(x)) \\ &= argmin_{F(x)} \sum_{i=1}^{n} l(y_i, F(x_i)) \\ &= argmin_{F(x)} E_{y,x}[l(y, F(x))] \\ &= argmin_F E_x[E_y[l(y, f(x))]|x] \end{aligned} \tag{51}$$

[8] Note that the calculation F can be obtained by minimizing the empirical risk function

$$min_F \frac{1}{n} \sum_{i=1}^{n} \psi(y_i, F(x_i)) \tag{52}$$

Combined functions in F did is to do this, iterative methods are used. Search with experiential and line features It is as follows for each step, which is explained below:$\hat{F}$

$$\hat{F}_m(x) = \hat{F}_{m-1}(x) + \nu \beta^{[m]} h(x, \epsilon^{[m]}) \tag{53}$$

$h(x, \epsilon^{[m]})$ It means tree regression.
So that:

m: number of repetitions,
ν: size of each step (shrinkage coefficient)
Calculating the rate of decline as the work answer:

$$z_i = -\frac{\partial}{\partial F(x_i)} \psi(y_i, F(x_i))|_{F(x_i)=\hat{F}(x_i)}$$

Optimization in numerical space:
Due to the M The repetition stage takes weeks to form the writings:

$$\hat{\theta} = \sum_{i=1}^{n} \hat{\theta}_i$$

It is an area that helps to identify methods.
The simplest method of determination is the accelerated quadratic method. assuming that N Up data from points (x_i, y_i) There is $J(\theta)$ is the loss function for each observation.

$$J(\theta) = \sum_{i=1}^{N} \psi(y_i, f(x_i, \hat{\theta})) \tag{54}$$

In this case, t is considered an iterative step. The estimate $\hat{\theta}$ at step t is as follows:
1. Start with an initial estimate $\hat{\theta}_0$.
2. Update the estimate at step t using the formula:

$$\hat{\theta}_t = \sum_{i=1}^{t-1} \hat{\theta}_i$$

3. The gradient of the objective function $J(\theta)$ is given by:

$$\nabla J(\theta) = [\nabla J(\theta_i)] = \left[\frac{\partial j(\theta)}{\partial j(\theta_i)}\right] \tag{55}$$

where $\theta = \hat{\theta}^{[t]}$.
Optimization in space:
A new function $h(x_i; \theta)$ is proposed, suggesting the following gradient descent update:

$$g_t(x) = E_y \left[\frac{\partial \psi(y, F(x))}{\partial F(x)} \mid x\right]$$

where $F(x) = \left[\hat{F}(x)\right]^{[t-1]}$.
Secretaries actually tell us where to go, with the condition that $F(x) = [\hat{F}(x)]^{[t-1]}$. Note that you have conditions that you offer can be associated with $-g(x)$ chose. Evaluation now $F(X)$ is as follows:
1.Consider$\hat{F}(x)$
2.Calculation of descending averages $-g(x)$
3. $h(x_i;\theta)$ usually belongs to the class of basis functions x with sides θ that's mean:

$$h_1(x) = -g(x)$$

Simultaneously for each i:

$$h_i(x) = -g(x_i)$$

(Although this work seems difficult, it will be more suitable than any other way) 4.$\rho_t = argmin_\rho \sum_{i=1}^{N} \psi[y_i, \hat{F}_{t-1}(x_i) + \rho h(x_i, \theta_t)]$
In fact, the optimal size of each step is:

$$\rho_1 = argminL(y, \hat{F}_{t-1}(x_1) + \rho h_1(x))$$

5.End
Algorithm (ls-Boost): (squared error):
Suppose the loss function $l(y, F) = \frac{(y-F)^2}{2}$ and $\hat{y}_i$be equivalent to reduction. Then:

$$\begin{aligned} \hat{y}_i &= -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \\ &= y_i - F^{[m-1]}(x_i) \\ F_0(x) &= \hat{y} \end{aligned}$$

In the second step:

$$\hat{y}_i = y_i - F_{m-1}(x_i)$$

And in the last step:

$$F_m(x) = F_{m-1}(x) + \rho h(x, a_m)$$

45.6 Draft:

Linear model:

$$E(Y|X) = f(x) = \sum_{j=1}^{p} \beta_j X_j \tag{56}$$

$$E(Y|X) = f(x) = \sum_{j=1}^{p} f_j(X_j)$$

Accelerated school model:

$$E(Y|X) = f(x) = \sum_{j=1}^{p} \beta_t h(x, a_t)$$

Basically, it starts with a simple regression tree and calculates the residuals. Then a residual model is placed and the collected model is calculated. Calculate the new remainder and continue.
Accelerated Gradient Tree at a glance:
A simple example of accelerated tree freezing: (Ben Germain 2017)
From a set consisting of (x_i, y_i) With $n = 9$ Has used.
Objective: To predict the age of each person based on the level of interest in video games, enjoyment of gardening, and the person's ability to wear hats.
A: If the loss function includes the square error: (the goal is to minimize the square error)

Table 3: To predict the age of each person based on the level of interest in video games, enjoyment of gardening, and the person's ability to wear hats.

Age	Enjoy gardening	Video Games	Wearing a cap
13	False	True	True
14	False	True	False
15	False	True	False
25	True	True	True
35	False	True	True
49	True	False	False
68	True	True	True
71	True	False	False
73	True	False	True

As we can see, the age group is the most interested in gardening.

Younger people are interested in video games, and the age of those who like to wear hats cannot be guessed, or they are so-called noise because it is a random thing. Now it is possible to make a detailed classification and consider another form of data as follows: Now start with a weak learner

Table 4: more suitable Tree

Individual interest	FALSE	TRUE
Enjoy gardening	$[13, 14, 15, 35]$	$[25, 49, 68, 71, 73]$
Video Games	$[49, 71, 73]$	$[13, 14, 15, 25, 35, 68]$
Wearing a cap	$[14, 15, 49, 71]$	$[13, 25, 35, 68, 73]$

that cannot be a regression tree on the data and build a model based on it. This tree model that has been produced has branched out on the interest in gardening. Tree1: A branch of a tree on a gardening hobby If you normally continue with trees like decision making methods, this time you should continue to wear hats on specific features and play the tree video game, this method would be more appropriate: More suitable tree map: from tree 1: Now, with the help of the errors of the first tree, they have built the second regression tree:

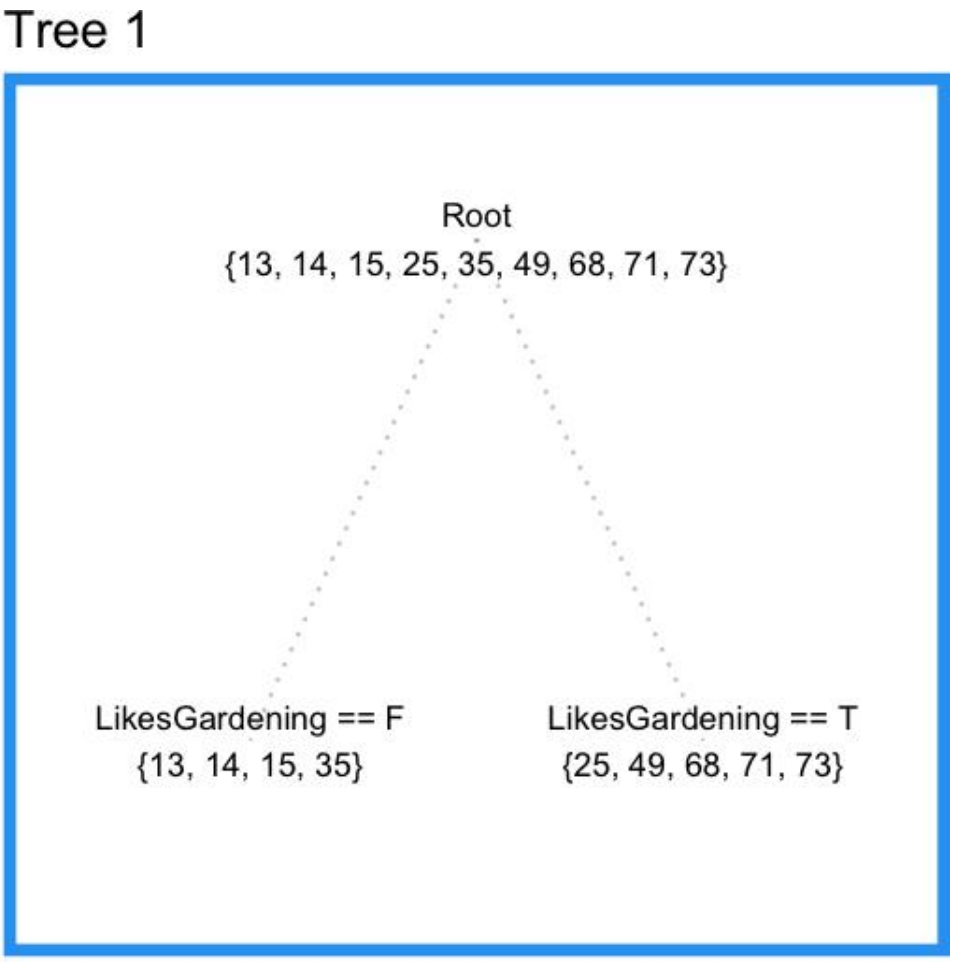

Figure 3: A branch of a tree on a gardening hobby

Table 5: The answers obtained from the tree 1

Age	Prediction of the first tree	Rest of tree1
13	19.25	−6.25
14	19.25	−5.25
15	19.25	−4.25
25	57.2	−32.2
35	19.25	15.75
49	57.2	−8.2
68	57.2	10.8
71	57.2	13.8
73	57.2	15.8

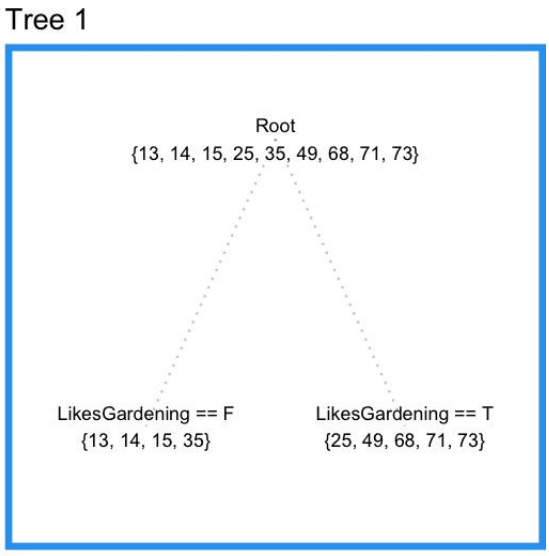

Figure 4: Second regression tree with branching on interest in video games

This is while the second tree predicts the amount of errors of the first tree and calculates the remaining amount of errors for the second tree.

45.7 Tree-Based Model Evaluation Using Mean Squared Error (MSE)

In a tree-based model, data values are assigned to the left and right branches of the tree. In this example, the values are distributed as follows:

Table 6: Residual amount of error with squared loss

Age	First Tree Prediction	First Tree Residual	Second Tree Prediction	Final Residual
13	19.25	-6.25	15.68	3.683
14	19.25	-5.25	15.68	1.683
15	19.25	-4.25	15.68	0.6833
25	57.2	-32.2	53.63	28.63
35	19.25	15.75	15.68	-19.32
49	57.2	-8.2	64.33	15.33
68	57.2	10.8	53.63	-14.37
71	57.2	13.8	64.33	-6.667
73	57.2	15.8	64.33	-8.667

- **Left Branch Value:** -3.567 represents the value assigned to all data on the left branch.
- **Right Branch Value:** 7.377 represents the value assigned to all data on the right branch.

45.8 MSE Calculation for the First Tree

The Mean Squared Error (MSE) is a crucial metric for evaluating the accuracy of tree-based models. For the first tree in this example, the MSE is calculated as:

$$\text{MSE (First Tree)} = 1993.55. \tag{57}$$

This value reflects the average squared difference between the predicted values by the tree and the actual data values.

45.9 Combined MSE Calculation

When combining the MSE values from multiple trees, the calculation is adjusted to include the contributions of each tree. In this case, the combined MSE is:

$$\text{Combined MSE} = 1764.57. \tag{58}$$

This value represents the total mean squared error when the predictions from the first tree are included.

45.10 Final Error Estimation

To estimate the final error, the predicted value from the first tree is added to the total value of predictions. The resulting value is then used to calculate the final error, which includes the contribution of the first tree and any subsequent trees used in the model. This process helps in refining the model's accuracy by iteratively minimizing the MSE.

45.11 Conclusion

Evaluating tree-based models using MSE is a critical step in understanding their predictive performance. By analyzing individual and combined MSE values, one can gauge the effectiveness of the model and make necessary adjustments to improve its accuracy. The iterative process of adding tree predictions and recalculating the error forms the basis of refining tree-based models in machine learning.

Table 7: Sum of predictions with squared error

Tree SSE	combined SSE
1994	1765

Accelerated Gradient Algorithm at a Glance: (Square Error)
1.Build the right model for the data.

$$F_1(x) = y$$

2.Make a model for the remainder.

$$h_1(x) = y - F_1(x)$$

3.produced a new model.

$$F_2(x) = F_1(x) + h_1(x)$$

Improve this idea by adding more models that correct the errors of the previous model I can generalize.

$$F_m(x) = F_{m-1}(x) + h_{m-1}(x)$$

Our task is to find at every stage

$$h_m(x) = y - F_m(x)$$

To get the squared error: We start with:

$$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y_i, \gamma) = argmin_\gamma \sum_{i=1}^{n} (\gamma - y_i)^2 = \frac{1}{n} \sum_{i=1}^{n} y_i \qquad (59)$$

Now:

$$F_{m+1}(x) = F_m(x) + h_m(x) = y$$

46 Iterative Refinement in Gradient Boosting and Loss Function Optimization

46.1 Optimizing Model Iterations

One of the crucial questions in the application of gradient boosting methods is determining the optimal number of iterations for model refinement. The process involves iteratively modifying the model to improve accuracy.

Determining Optimal Iterations:

- The best approach to ascertain the number of iterations required is through cross-validation.
- Cross-validation involves testing the model with different numbers of iterations and evaluating performance on separate validation data sets.
- This method helps in identifying the point of diminishing returns where additional iterations no longer contribute to significant improvements in model accuracy.

46.2 Handling Absolute Error Loss

When the loss function involves absolute error, the optimization strategy differs from that used for residual-based loss functions.

Approach for Absolute Error Loss:

- Instead of focusing on residuals (differences between predicted and actual values), attention is given to the locators of the loss function.
- The model optimization process targets minimizing the absolute discrepancies between predicted and actual values.
- This approach often requires specialized optimization techniques that consider the non-differentiable nature of absolute error.

46.3 Conclusion

The iterative process in gradient boosting requires careful consideration of the number of iterations and the nature of the loss function. Cross-validation serves as an effective tool for optimizing the number of iterations, ensuring the model achieves the best possible performance without overfitting. Additionally, when dealing with absolute error loss, the focus shifts to minimizing absolute discrepancies, which may require different optimization strategies. Understanding these nuances is crucial in developing robust and accurate gradient-boosted models.

$$F_0(x) = argmin_\gamma \sum_{i=1}^{n} L(y, \gamma)$$

Table 8: If Square Error

Age	$F_0(x)$	Residual 0
13	35	-22
14	35	-21
15	35	-20
25	35	-10
35	35	0
49	35	14
68	35	33
71	35	36
73	35	38

Here is the middle y equal to 35 dollars loss function $L(y_i, F(x_i))$ is that y_i: Observations and $F(x_i)$. It is predicted that the loss function comes from the absolute value of the difference between these two values.
Since we need to find a prediction with the least loss of this sum

Table 9: Final Accelerated Gradient model with squared error

AGE	$F_0(x)$	rest	h_0	γ_0	$F_1(x)$	Rest 1	h_1	γ_1	F_2
13	40.33	-27.33	21.08	1	19.25	-6.25	-3.567	1	15.68
14	40.33	-26.33	21.08	1	19.25	-5.25	-3.567	1	15.68
15	40.33	-25.33	21.08	1	19.25	-4.25	-3.567	1	15.68
25	40.33	-15.33	16.87	1	57.2	-32.2	-3.567	1	53.63
35	40.33	-5.33	-21.5	1	19.25	15.75	-3.567	1	15.68
49	40.33	8.667	16.87	1	57.2	-8.2	7.133	1	64.33
68	40.33	27.67	16.87	1	57.2	10.8	-3.567	1	53.63
71	40.33	30.67	16.87	1	57.2	13.8	7.133	1	64.33
73	40.33	32.67	16.87	1	57.2	15.8	7.133	1	64.33

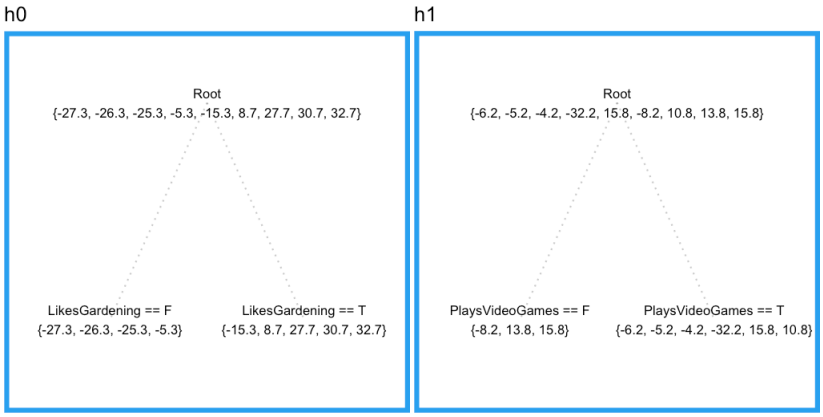

Figure 5: Tree branching with the help of residual error of the second tree

Table 10: The final accelerated Gradient model with absolute error is given in the following table:

Age	$F_0(x)$	Rest	h_0	γ_0	$F_1(x)$	Rest1	h_1	γ_1	F_2
13	35	-1	-1	20.5	14.5	-1	-0.3333	0.75	14.25
14	35	-1	-1	20.5	14.5	-1	-0.3333	0.75	14.25
15	35	-1	-1	20.5	14.5	1	-0.3333	0.75	14.25
25	35	-1	0.6	55	68	-1	-0.3333	0.75	67.75
35	35	-1	-1	20.5	14.5	1	-0.3333	0.75	14.25
49	35	1	0.6	55	68	-1	0.3333	9	71
68	35	1	0.6	55	68	-1	-0.3333	0.75	67.75
71	35	1	0.6	55	68	1	0.3333	9	71
73	35	1	0.6	55	68	1	0.3333	9	71

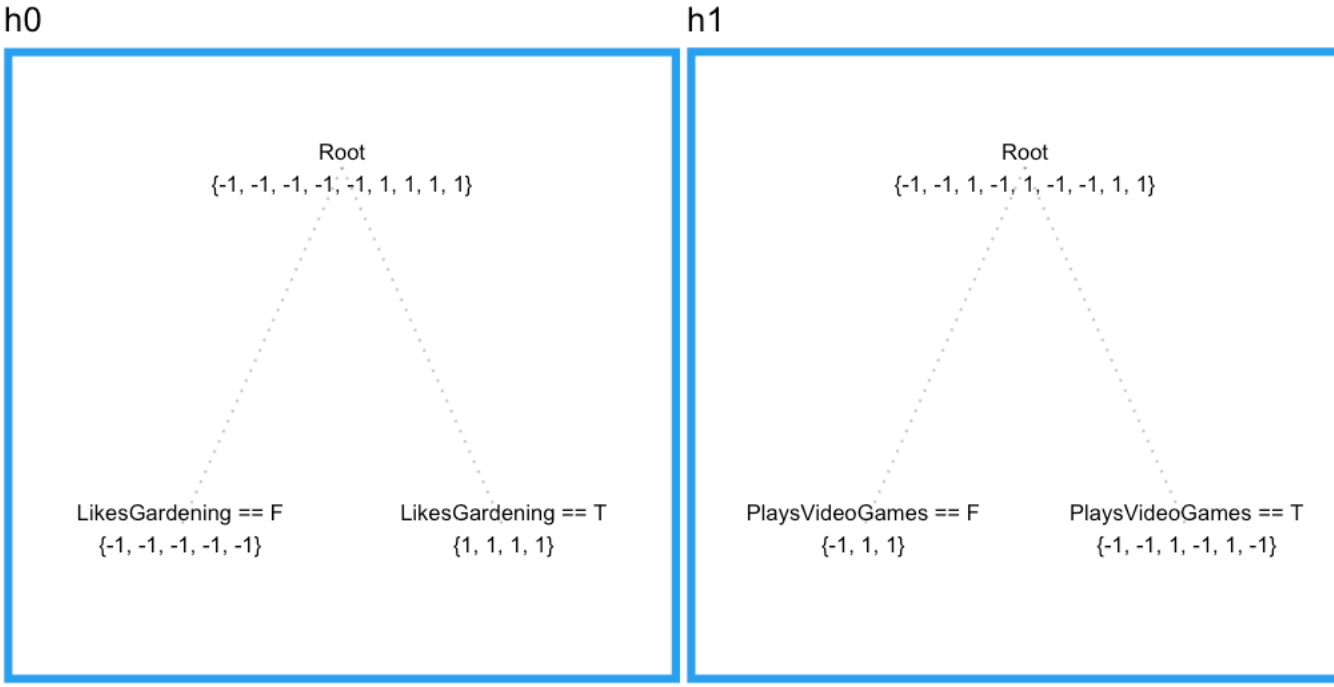

Figure 6: Tree branching with the help of residual error of the second tree

according to:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

It is clear that we approach the true value at each step.

47 Chapter 3:

Tweedy Composite Poisson Model

48 Introduction to Tweedie Composite Poisson Distribution

48.1 Overview

This chapter explores the Tweedie Composite Poisson distribution, a complex amalgamation of mixed distributions as discussed by Jorgensen and Jordan Smith (2002)[20]. The Tweedie compound Poisson model, aligning with the standards of the generalized linear model (GLM), is an essential tool in linear prediction. Its development and applications have been discussed by various researchers including Hall (1999)[14], Murphy et al. (2000), and Sandri and Zoccolotto (2008)[29].

48.2 Challenges and Solutions in Non-linear Prediction

The Tweedy model, with its logarithmic structure, is inherently linearly limited. This limitation necessitates the incorporation of non-linear components, often achieved using splines:

- Low-order splines, while useful, may not sufficiently capture nonlinear coefficients in the data.
- High-order splines pose a risk of overfitting.
- To address these issues, the Generalized Additive Model (GAM) becomes an essential tool. The GAM, as discussed by Hastie and Tibshirani (1990) [16] and Wood (2006)[34], provides effective smoothing for continuous variables.

48.3 Application in Insurance Studies

In insurance studies, datasets frequently exhibit zero clustering, presenting unique modeling challenges:

- Traditionally, zero and non-zero data were modeled together, as suggested by Cragg (1971) and Mullahy (1986).[6]

- Lambert (1992) introduced the Zero-Inflated Poisson (ZIP) model, specifically addressing the issue of zero clustering in Poisson regression models.[21]

48.4 Tweedy Distribution in Statistical Modeling

The Tweedy family of distributions, which includes Gamma, Normal, and Poisson distributions (or their combinations), represents a specialized subset within distribution families:

- The Tweedy distribution is noted for its flexibility and versatility.
- It allows for the representation of data through an overlay of distributions, making it a valuable tool in statistical modeling.

In conclusion, the Tweedy Composite Poisson distribution, with its capacity to model complex data structures and accommodate non-linear patterns, is a significant contributor to the field of statistical modeling, especially in areas like insurance where zero clustering is prevalent.

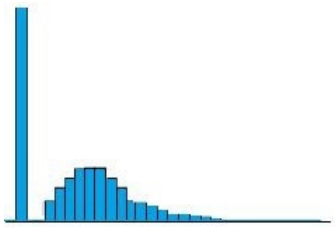

Figure 7: Tweedy Distribution

49 Tweedy Distribution and its Applications

49.1 General Form of Tweedy Distribution

The Tweedy distribution, introduced by Tweedy (1984), is a flexible family of distributions used in statistical modeling. Its general form is given by:[32]

$$f_Y(y, \theta, \phi) = a(y, \theta, \phi) \exp\left(\frac{y\theta - k(\theta)}{\phi}\right), \tag{60}$$

where:

- $\phi \in \mathbf{R}^+$ is the dispersion parameter.
- θ is an indicator in $\mathcal{R}$.
- $k(.)$ is the cumulant function.
- $a(.)$ is the normalizing function.

49.2 Properties of Tweedie Distribution

- **Expectation and Variance:**

$$E(Y) = \mu = \frac{\partial}{\partial\theta} k(\theta) = k'(\theta), \tag{61}$$

$$\mathrm{Var}(Y) = \phi k''(\theta). \tag{62}$$

- **Variance-to-Mean Relationship:**

$$\mathrm{Var}(Y) = \phi\mu^\rho = \phi(E(Y))^\rho. \tag{63}$$

- **Parameterization:** For different values of ρ, the indicator θ varies as follows:

$$\theta = \begin{cases} \frac{\mu^{1-\rho}}{1-\rho}, & \text{if } \rho \neq 1, \\ \log\mu, & \text{if } \rho = 1. \end{cases} \tag{64}$$

49.3 Model Interpretations

Yi Yang, Wei Qian, and Hui Zou have provided further interpretations of the Tweedie distribution in the context of statistical modeling. Smyth and Jorgensen described the relationship between components in Tweedie distribution as follows:[36]

$$\mu = \lambda\alpha\beta, \tag{65}$$

$$\rho = \frac{\alpha + 2}{\alpha + 1}, \tag{66}$$

$$\phi = \frac{\lambda^{1-\rho}(\alpha\beta)^{2-\rho}}{2 - \rho}. \tag{67}$$

An alternate interpretation for ρ in relation to α is given by:

$$\rho = \frac{\alpha - 2}{\alpha - 1}. \tag{68}$$

49.4 Conclusion

The Tweedy distribution provides a versatile framework for statistical modeling, especially in contexts where the variance-to-mean relationship is a crucial factor. Its parameterization allows for a broad range of applications, including actuarial science, ecological modeling, and finance. Understanding its properties and interpretations is key to effectively applying this distribution in various statistical analyses. If ρDifferent features can be converted to Tweed distribution into different distributions:
1: If $\rho = 0$ normal distribution
2: If $\rho = 1$ Poisson distribution
3: If $1 < \rho < 2$ non-negative composite Poisson distribution
4: If $\rho = 2$ Gamma distribution
5: If $2 < \rho < 3$ stable positive distributions
6: If $\rho = 3$ inverse Gaussian distributions
7: If $\rho > 3$stable positive distributions
8: If $\rho = \infty$ extreme stable distributions
to do $\rho from 0 to 1$ is not defined. [19]

50 Composite Poisson-Gamma Distribution:

The intrinsic characteristic of the Poisson distribution, where the mean equals the variance, often leads to the search for alternative models to address over dispersed data. One such alternative for count data is the Poisson-Gamma model, which outperforms the basic Poisson model due to its additional flexibility. Notable variations include the Poisson-Gamma regression model, the Inverse Poisson-Gaussian regression model, and the Generalized Inverse Poisson-Gaussian model (Gardner et al.,1995; Stein and Joritz, 2007; Rigby et al.,2008). In this thesis, we focus on the Poisson-Gamma composite model, a significant statistical distribution widely utilized in problem-solving. Let's consider Y as a variable following a Poisson distribution with a rate $\mu\lambda$, assuming *lambda* experiences multiple occurrences. Under this scenario, the probability density function of the random variable Y, which aligns with the negative binomial distribution, is delineated as follows:

$$P_r(Y = y) = \frac{\Gamma(y + \frac{1}{\sigma})}{\Gamma(y + 1)\Gamma(\frac{1}{\sigma})}(\frac{1}{1 + \sigma\mu})^{\frac{1}{\sigma}} \tag{69}$$

with

$$E(Y) = \mu$$

and:

$$var(Y) = \mu(1 + \sigma\mu)$$

51 Composite Poisson distribution and Tweedy model:

Tweedy compound Poisson distribution is actually a combination of mixed distribution. Jorgensen and Smith In the year 2002 In a paper, they present the compound Tweed Poisson model for insurance claim submissions and present the compound Tweedy Poisson model with *cpois* Showed Jorgensen in the year 1987 had stated that the Tweedy composite Poisson model to the models EDM it's close.EDM Two-line broadcasts are from a special family.[17] In case of accident N_i: The number of claims observed in be classified Z_i: The sum of all claims. (Considered as a serious claim) ω_i: is the number of risk units. Y_i: amount of claim for each unit (which is

here as net insurance premium) The claimed insurance amount has a tweed distribution.[17]

$$Y_i = \frac{Z_i}{\omega_i} \sim TW(\mu_i, \frac{\phi}{\omega_i}, \rho) \tag{70}$$

52 Tweedy Compound Poisson Model in Insurance Applications

52.1 Model Description

The Tweedy Compound Poisson (TCP) model is a versatile tool in insurance modeling, particularly for predicting claim sizes and insurance payments. This model is defined as follows:

- **Claim Size (Z_d^*):** Represents the size of each observed claim, where claims $d_i = 1, \ldots, N_i$ follow a gamma distribution.
- **Independence Assumption:** The number of claims N and the claim sizes Z_d^* are independent.
- **Model Structure:**

$$Z = \begin{cases} 0, & \text{if } N = 0, \\ Z_1^* + \cdots + Z_N^*, & \text{if } N = 1, 2, \ldots \end{cases} \tag{71}$$

- **Distribution:** Z follows a compound Poisson distribution, denoted as $cpois(\mu_i, \phi, \rho)$.

52.2 Probability and Distribution Function

The probability of Z being zero and its distribution function are given by:

$$P_r(Z = 0) = P_r(N = 0) = \exp(-\lambda), \tag{72}$$

$$f_z(Z|\lambda, \alpha, \gamma) = \exp(-\lambda) d_0(z) + \sum_{j=1}^{\infty} \frac{\lambda^j e^{-\lambda} z^{(j\alpha-1)} e^{-\frac{z}{\gamma}}}{j! \gamma^{j\alpha} \Gamma(j\alpha)}. \tag{73}$$

52.3 Insurance Premium Calculation

In insurance applications:

- **Total Claim Payment (Z):** Sum of individual claims for each observation.
- **Premium Issued (Y):** Calculated as the total claim payment divided by the number of risk units or policy duration.

$$Y_i = \frac{Z_i}{\omega_i}, \tag{74}$$

where Z_i is the sum of all insurance claims for category i.

52.4 Tweedy Model Function

The function of the Tweedy model, incorporating the logarithmic transformation, is expressed as:

$$\log f_Z(z|\mu, \phi, \rho) = \left(\frac{1}{\phi} \left(z \frac{\mu^{1-\rho}}{1-\rho} \right) - \frac{\mu^{2-\rho}}{2-\rho} \right) + \log a(z, \rho, \phi), \tag{75}$$

$$a(z, \phi, \rho) = \begin{cases} \frac{1}{z} \sum_{t=1}^{\infty} W_t, & \text{if } z > 0, \\ 1, & \text{if } z = 0. \end{cases} \tag{76}$$

52.5 Model Estimation and Application

The model parameters μ, ρ, and ϕ are typically estimated using Maximum Likelihood Estimation (MLE). The TCP model is used to model the total insurance claims, considering the number of risk units and the duration of the policy.

References:

- Yang et al. (2016) for the formulation of the TCP model.
- Olson and Johansson (2010) for insurance premium calculations.
- Smyth and Jorgensen for relationships within Tweedy distributions.

52.6 Conclusion

The Tweedy Compound Poisson model plays a significant role in insurance modeling, providing a comprehensive framework for predicting claim sizes and insurance payments. Its flexibility and ability to model complex distributions make it an invaluable tool in the actuarial sciences. Since the

53 Tweedy Compound Poisson Model for Insurance Premium Estimation

53.1 The Tweedy Compound Poisson Distribution

The Tweedy Compound Poisson (TCP) model, an essential tool in insurance modeling, is defined as:

$$Z = \begin{cases} 0, & \text{if } N = 0, \\ \hat{Z}_1 + \cdots + \hat{Z}_N, & \text{if } N = 1, 2, \ldots \end{cases} \tag{77}$$

Jorgensen and Sousa (1994) observed the relation:

$$\mu_i = E[Y_i] = \lambda_i \tau_i, \quad \text{Var}(\mu_i) = \mu_i^\rho, \quad P = \frac{\alpha + 2}{\alpha + 1}.$$

53.2 Modeling with Tweedy Distribution

In scenarios where $1 < \rho < 2$:

$$Y_i \sim \text{TW}(\mu_i, \frac{\phi}{\omega_i}, \rho), \quad \text{Var}(Y_i) = \phi_i \mu_i^\rho, \quad \phi_i = \frac{\lambda_i^{1-\rho} \tau_i^{2-\rho}}{2 - \rho}.$$

53.3 Insurance Contract Data

For n insurance contracts represented as (Y_i, x_i, ω_i):

- Y_i: Insurance premium.
- x_i: Risk factors (e.g., property, vehicles).
- ω_i: Policy duration.

- μ_i: Estimated insurance premium using a prediction function F, defined as:

$$\log[\mu_i] = F(x_i).$$

53.4 Maximum Likelihood Estimation

The maximum likelihood function for this model can be written as:

$$l(F(.), \phi, \rho | \{y_i, x_i, \omega_i\}) = \sum_{i=1}^{n} \frac{\omega_i}{\phi} \left(y_i \frac{\mu_i^{1-\rho}}{1-\rho} - \frac{\mu_i^{2-\rho}}{2-\rho} \right) + \log a(y_i, \frac{\phi}{\omega_i}, \rho).$$

53.5 Gradient Boosted Trees in Tweedy Model

Gradient Boosted Trees can be incorporated into the Tweedy model for optimizing the prediction function $F(.)$. The objective is to minimize the loss function ψ with respect to $F(.)$:

$$\hat{F}^{TD}(x) = \exp(\hat{F}^{TD}(x)), \quad \psi(y_i, F(x_i)|\rho) = \omega_i \left(\frac{\mu_i^{1-\rho}}{1-\rho} + \frac{\mu_i^{2-\rho}}{2-\rho} \right).$$

The tree-based approach is used for decision-making and optimization in this context.

53.6 Conclusion

The Tweedy Compound Poisson model, particularly when integrated with Gradient Boosted Trees, provides a robust framework for predicting insurance premiums. It accounts for the variability and risk factors associated with insurance contracts, making it a valuable tool in actuarial science.

54 Least Squares Reduction in Gradient Boosting

54.1 Optimization of Residuals

In the context of gradient boosting, the reduction of residuals is achieved through the least squares method:

$$\hat{\epsilon}^{[m]} = \arg\min_{\epsilon^{[m]}} \sum_{i=1}^{n} [g_i^{[m]} - h(x_i, \epsilon^{[m]})]^2, \tag{78}$$

where $g_i^{[m]}$ is derived from the negative gradient of the loss function $\psi(.,\rho)$ with respect to $F(x_i)$:

$$g_i^{[m]} = -\frac{\partial \psi(y_i, F(x_i))}{\partial F(x_i)} = \omega_i \left[-y_i \exp((1-\rho)\hat{F}(x_i)^{[m-1]}) + \exp((2-\rho)\hat{F}(x_i)^{[m-1]}) \right]. \tag{79}$$

54.2 Optimal Coefficient Estimation

The optimal coefficient $\beta^{[m]}$ is obtained from a linear model:

$$\beta^{[m]} = \arg\min_{\beta} \sum_{i=1}^{n} \left[\psi(y_i, \hat{F}(x_i)) + \beta h(x, \hat{\epsilon}^{[m]}|\rho) \right] \tag{80}$$

$$= \arg\min_{\beta} \sum_{i=1}^{n} \left[\psi(y_i, \hat{F}(x_i)^{[m-1]}) + \beta \sum_{l=1}^{L} \bar{u}_l^{[m]} I(x_i = \hat{R}_l^{[m]}|\rho) \right], \tag{81}$$

where $\bar{U}_l^{[m]} = \text{mean}(u_i^{[m]})$ and $\hat{U}_l^{[m]} = \text{ave}(g_i^{[m]})$ for $l = 1, \ldots, L$.

54.3 Model Update and Shrinkage Factor

Once $\hat{F}(x)^{[m-1]}$ is estimated, the model update is given by:

$$\hat{F}(x)^{[m]} = \hat{F}(x)^{[m-1]} + \nu \sum_{l=1}^{L} \hat{\eta}_l^{[m]} I(x = \hat{R}_l^{[m]}), \tag{82}$$

where $\hat{\eta}_l^{[m]}$ is obtained by minimizing:

$$\hat{\eta}_l^{[m]} = \arg\min_{\eta} \sum_{i} \psi(y_i, \hat{F}(x_i)^{[m-1]} + \eta|\rho). \tag{83}$$

54.4 Empirical Findings and Computational Considerations

Empirical studies by Friedman, along with Belman (2001), Hadorn (2007), and Ridgeway (2007), suggest that a lower shrinkage factor (typically $\nu < 0.1$) improves predictive accuracy. However, this comes at the cost of increased computational time and effort. The value of ν is typically determined by the number of boosting iterations, M, which directly impacts the computational duration.

54.5 Conclusion

The least squares reduction technique in gradient boosting involves optimizing residuals and coefficients iteratively. While a lower shrinkage factor can enhance model accuracy, it requires careful consideration of computational resources and time. This balance is crucial in achieving optimal performance in predictive modeling using gradient boosting.

TDboost Algorithm

1: Initialization of $\hat{F}^{[0]}$

The initial estimate, $\hat{F}^{[0]}$, is computed as follows:

$$\hat{F}^{[0]} = \log\left(\frac{\sum_{i=1}^{n}\omega_i y_i}{\sum_{i=1}^{n}\omega_i}\right)$$

2: Iterative Update for $m = 1, \ldots, M$

A: Calculation of Pseudo-Residuals, $(u_1^{[m]}, \ldots, u_n^{[m]})$**:**

$$\begin{aligned} u_i^{[m]} &= \frac{-\partial\psi(y_i, F(x_i))}{\partial F(x_i)} \\ &= \omega_i\left[-y_i \exp\left((1-\rho)\hat{F}(x_i)^{[m-1]}\right)\right] + \exp\left((2-\rho)\hat{F}(x_i)^{[m-1]}\right), \quad i = 1, \ldots, n \end{aligned}$$

B: Regression Tree Update with Decreasing Averages $(u_1^{[m]}, \ldots, u_n^{[m]})$ **and** $(x_1, \ldots, x_n)^T$**:** For each node $l = 1, \ldots, L$, the region $\hat{R}_l^{[m]}$ is determined, and the update amount $\hat{\eta}_l^{[m]}$ is calculated as:

$$\hat{\eta}_l^{[m]} = \arg\min_{\eta} \sum \psi\left(y_i, \hat{F}(x_i)^{[m-1]} + \eta \mid \rho\right) = \log\left[\frac{\sum \psi_i y_i \exp\left((1-\rho)\hat{F}(x_i)^{[m-1]}\right)}{\sum \psi_i \exp\left((2-\rho)\hat{F}(x_i)^{[m-1]}\right)}\right]$$

C: Update of Current Estimate $\hat{F}(x)^{[m]}$**:** The current estimate is updated as:

$$\hat{F}(x)^{[m]} = \hat{F}(x)^{[m-1]} + \nu\hat{\eta}_l^{[m]} I(x = \hat{R}_l^{[m]})$$

3: Final Estimation $\hat{F}^{[M]}(x)$

The final estimate is computed as:

$$\hat{F}(x)^{[M]} = \hat{F}(x)^{[M-1]} + \nu \sum_{l=1}^{L} \hat{\eta}_l^{[m]} I(x = \hat{R}_l^{[m]}), \quad l = 1, \ldots, L$$

55 Zero Accumulation Models and Their Application in Insurance

Zero accumulation models, initially introduced by Lambert in 1992, are hybrid constructs combining a count component with the prevalence of zero occurrences. These models are particularly pertinent in scenarios like car insurance claims, where the mass of zeros can vary significantly. A notable instance is the high frequency of zero claims in count responses.

These models are indispensable in situations where zero claims are excessively prevalent, rendering conventional statistical models like the Poisson distribution or negative binomials inadequate. For example, the recorded number of insurance claims may be artificially low due to unreported incidents, influenced by concerns such as increased future premiums or specific risk costs for the insurer.

In contexts where counts exhibit a substantial concentration of zeros, they are categorized either as Poisson distributions with zero accumulation (Zero Inflated Poisson, ZIP) or negative binomial distributions with zero accumulation (Zero Inflated Negative Binomial, ZINB). Conversely, if zeros are not predominant, standard Poisson or binomial models may be sufficient.

Poisson Models with Zero Clustering

The ZIP model, popularized by Lambert in 1992, is designed to model count data with an excess of zeros. These models address the surplus zeros stemming from a separate counting process and can be modeled independently. For instance, in a U.S. study on smoking habits, ZIP regression was utilized to analyze smoking frequency in relation to factors like race (Brandes et al., 2010).

Given that Generalized Linear Models (GLMs) extend traditional regression models, the ZIP model integrates both a Poisson and a Logit model. It

is particularly relevant when sample data exhibit significant zero-clustering, leading to overdispersion a scenario where the variance exceeds the mean, contrasting the equal variance-mean relationship in typical Poisson distributions.

Suppose Y is a random number with a mixed distribution in k accumulations. Then, for $y = 0$, we have:

$$P(Y = y) = \omega_1 + \omega_2 e^{-\lambda_2} + \cdots + \omega_k e^{-\lambda_k},$$

and for $y = 1, 2, \ldots$, it follows:

$$P(Y = y) = \omega_2 \frac{\lambda_2 e^{-\lambda_2}}{y!} + \cdots + \omega_k \frac{\lambda_k e^{-\lambda_k}}{y!}.$$

Here, λ_i denotes the mean of the i^{th} component, and ω_i represents the combined weights of the i^{th} component in mode $k = 2$.

The probability mass function for the ZIP model is given as:

$$P_{\text{ZIP}}(Y = 0) = \omega + (1 - \omega) P_{\text{poi}}(K = 0), \tag{84}$$

$$P_{\text{ZIP}}(Y = y) = (1 - \omega) P_{\text{poi}}(K = y), \quad y = 1, 2, \ldots, \tag{85}$$

where $P_{\text{poi}}(K = y)$ is the probability mass function of the Poisson distribution.

Claims are posited to originate from a dual-part process comprising structural zeros and random claims. In two distinct scenarios with zero modes, the first case occurs with probability ω, and the second case with probability $1-\omega$. Here, zeros in the first case are deemed structural, while in the second case, they are considered sampling zeros (Jasakol and John, 2009).
Mohammad Pham and colleagues in the year 2009 in Hamedan of Iran, Poisson regression with zero accumulation for modeling accidents at work was used and showed that based on this model, marital status and accidents play a role in the number of work accidents. Also, technicians are more exposed to accidents than those who work.
Zero inflated $TDboost$: to the model ZIF it is famous. In case of accident N_i the number of claims seen in i th Classification be, that of the distribution $pois(\lambda)$ is and Z_d^* Digital processing of distribution $Gamma(\alpha, \gamma)$ Random car$Z$ define as below:

$$Z = \begin{cases} 0 & N = 0 \\ Z_1^* + \cdots + Z_N^* & N = 1, 2, \ldots \end{cases}$$

From where it produces$Z$ has a probability value of zero,

$$P(Z = 0) = P(N = 0) = exp(-)$$

so this is the compound Poisson distribution Z For each observation, it is defined as follows:

$$f_Z(z|\theta, \phi) = a(z, \phi) \exp(\frac{z\theta - k(\theta)}{\phi})$$

Since in tweed distribution, quantity and variance are for distribution Z is as follows:

$$E(Y) = \mu = k'(\theta)$$

$$\text{var}(Y) = \phi k''(\theta)$$

Variance Ratio Relationship and Its Application

The variance ratio relationship can be represented as:

$$\text{var}(Y) = \phi \mu^{\rho}$$

Clearly, the definitions are articulated as follows:

$$\lambda = \frac{1}{\phi} \frac{\mu^{2-\rho}}{2-\rho}, \tag{86}$$

$$\alpha = \frac{2-\rho}{\rho-1}, \tag{87}$$

$$\gamma = \phi(\rho-1)\mu^{\rho-1} = \phi^{(\rho-1)\mu}. \tag{88}$$

Consequently, we have:

$$f(z \mid \mu, \phi, \rho) = a(z, \phi, \rho) \exp\left(\frac{1}{\phi}\left(z\frac{\mu^{1-\rho}}{1-\rho} - \frac{\mu^{2-\rho}}{2-\rho}\right)\right)$$

where

$$a(z, \phi, \rho) = \begin{cases} \frac{1}{z} \sum_{t=1}^{\infty} W_t & \text{if } z > 0, \\ 1 & \text{if } z = 0. \end{cases}$$

The series $\sum_{t=1}^{\infty} W_t$ is an example of a generalized function [32]

In the context of the composite Poisson model, if $N = 0$, then $Y = 0$ is likely with probability q. The model can be represented as:

$$Y = \begin{cases} 0 & \text{with probability } q, \\ \text{Tweedie}(\mu_i, \phi, \rho) & \text{with probability } 1 - q. \end{cases}$$

The probability mass function for the Zero-Inflated Function (f_{ZIF}) is then given by:

$$f_{ZIF}(y \mid \mu, \phi, \omega, \rho, q) = (1 - q) f_{TW}(y \mid \mu, \phi, \rho) + qI(y = 0) \tag{89}$$

Generalized Functions and Model Implementation

An example of a generalized function is described by Tweedy (1984)[32]. In the context of Twitter, if $N = 0$, then $Y = 0$. However, in the composite Poisson model, if $N = 0$, it is probable that $Y = 0$ with likelihood q. Consequently, we can state:

Gradient tree divisors in a compound Poisson model with zero clustering are complex, especially during the accelerated stages of an algorithm's problem-solving process (Friedman et al. 2001). Considering (y_i, x_i, ω_i) for $i = 1, \ldots, n$, where:

- x_i: A selection of actions considering insurance and the associated risk,
- y_i: The net insurance premium,
- ω_i: The duration of the policy.

The expected premium, utilizing the prediction function F, is given by:

$$\log[\mu_i] = F(x_i)$$

The maximum likelihood function is:

$$l(F(.),\phi,\rho \mid \{y_i,x_i,\omega_i\}) = \sum_{i=1}^{n}\left[\frac{\omega_i}{\phi}\left(\frac{y_i\exp[(1-\rho)F(x_i)]}{1-\rho} - \frac{\exp[(2-\rho)F(x_i)]}{2-\rho}\right) + \log a(y_i,\frac{\phi}{\omega_i},\rho)\right]$$

The present subjunctive F is calculated to minimize:

$$\hat{F}(.) = \arg\min_{F}\frac{1}{n}\sum_{i=1}^{n}\psi(y_i,F(x_i)\mid\rho)$$

With the help of the gradient, the acceleration of tree findings is:

$$F(x) = F^{[0]} + \sum_{m=1}^{M}\beta^{[m]}h(x,\epsilon^{[m]}) \tag{90}$$

Through decision-making, we have:

$$h(x,\epsilon^{[m]}) = h(x\mid u_l^{[m]},R_l^{[m]}),\quad l=1,\ldots,L \tag{91}$$

In the direction of the descending gradient vector, the least squares error function is:

$$\hat{\epsilon}^{[m]} = \arg\min_{\epsilon^{[m]}}\sum_{i=1}^{n}[g_i^{[m]} - h(x_i,\epsilon^{[m]})]^2 \tag{92}$$

The location $(g_1^{[m]},\ldots,g_n^{[m]})$ corresponds to the decreasing gradient vector of ψ:

$$g_i^{[m]} = \frac{-\partial\psi(y_i,F(x_i))}{\partial F(x_i)} \tag{93}$$

That:

$$\hat{u_l}^{[m]} = ave(g_i^{[m]}) \tag{94}$$

so in general:

$$Y = \begin{cases}0 & q\\ Tweedie(\mu_i,\frac{\phi}{\omega_i},\rho) & 1-q\end{cases}$$

Now the probability mass function of can be written as follows:Y_i

$$f_{ZIF}(y_i|x_i,\omega_i,F,\phi,\rho,q) = (1-q)f_{ZIF}(y_i|\mu_i,\frac{\phi}{\omega_i},\rho) + qI(y_i=0) \tag{95}$$

when:

$$\mu_i = exp[F(x_i)]$$

The expected premium is given by the following relation:

$$E(Y_i|x_i) = (1-q)exp[F(x_i)]$$

That is, insurance premium in the form of:

$$\hat{\mu}(x)^{ZIF} = (1-\hat{q})exp[\hat{F}_{ZIF}(x)] \tag{96}$$

That $\hat{q}$ and $\hat{F}_{ZIF}(x)$ must be evaluated.

56 Competitions of programs in the ZIF model:

The goal is to find the maximum likelihood in modelZIF to collect the maximum likelihood for the constructed models, the simplest method is to use the algorithm EM Is. Algorithm EM. It is an iterative calculation method that is used to make maximization. So with hidden introductions to the algorithm EM they are going. [24]
In the algorithm $TDboost$ the doors were placed before the discussion (reciprocal cases and main changes in the model should be specified) is an algorithm for maximization. Now, since the goal is to obtain the maximum accuracy in the model ZIF It should be with the data set $(y_i, x_i, \omega_i), i = 1, \ldots, n$ started.

$$\begin{aligned}(\hat{F}, \hat{\phi}, \hat{q}) &= \arg\max_{(F,\phi,q)} l(F, \phi, q \mid \{(y_i, x_i, \omega_i), \rho\}) \\ &= \arg\max_{(F,\phi,q)} \sum_{l=1}^{L} \log\left[(1-q) f_{Tw}(y_i, \mu_i, \phi \mid \omega_i, \rho) + qI(y_i = 0)\right]\end{aligned}$$

Now with hidden introductions $\pi = (\pi_1, \ldots, \pi_n)$ With the help of algorithm EM: If y_i Examples of $TWeedie(\mu_i, \frac{\phi}{\omega_i}, \rho)$ Then be: $\pi_i = 1$ and if y_i an example of Zero point accumulation then $\pi_i = 0$ Then, it is possible to

correct the maximum log distribution from (y, π) There are as follows:

$$
\begin{aligned}
l(F, \phi, q \mid \{(y_i, x_i, \omega_i, \pi_i)\}) = & \\
& \sum_{i=1}^{n} \left[\pi_i \log a \left(y_i, \frac{\phi}{\omega_i}, \rho \right) + \pi_i \frac{\omega_i}{\phi} \left(y_i, \frac{\mu_i^{1-\rho}}{1-\rho} - \frac{\mu_i^{2-\rho}}{2-\rho} \right) \right] \\
& + \sum_{i=1}^{n} \left[(1-\pi_i) \log I\{y_i = 0\} + (1-\pi_i) \log q + \pi_i \log(1-q) \right]
\end{aligned}
$$

So that:$\mu_i = \exp[F(x_i)]$
Location:

$$\Delta_{0_i} = 1 - \Delta_{1_i}$$

so according toπ

$$
\begin{aligned}
E_{\pi|y,\omega,F,\phi,q} l(F, \phi, q \mid \{(y_i, x_i, \omega_i, \pi_i)\}) = & \sum_{i=1}^{n} \left[\Delta_{1_i} \log a \left(y_i, \frac{\phi}{\omega_i}, \rho \right) + \Delta_{1_i} \frac{\omega_i}{\phi} \left(y_i \frac{\mu_i^{1-\rho}}{1-\rho} - \frac{\mu_i^{2-\rho}}{2-\rho} \right) \right] \\
& + \sum_{i=1}^{n} \left[\Delta_{0_i} \log I\{y_i = 0\} + \Delta_{0_i} \log q + \Delta_{1_i} \log(1-q) \right]
\end{aligned}
$$

So that:

$$
\Delta_{1_i} = P(\phi_i = 1 \mid y_i, \omega_i, F, q) = \begin{cases} 1 & \text{if } y_i > 0; \\ \left(\frac{(1-q) \exp\left(\frac{\omega_i}{\phi} \left(-\frac{\mu_i^{2-\rho}}{2-\rho} \right) \right)}{(1-q) \exp\left(\frac{\omega_i}{\phi} \left(-\frac{\mu_i^{2-\rho}}{2-\rho} \right) \right) + q} \right) & \text{if } y_i = 0. \end{cases}
$$

57 EM Algorithm for Maximum Likelihood Estimation in Mixed Models

57.1 Introduction

In the context of mixed models, the Expectation-Maximization (EM) algorithm is used to obtain the maximum likelihood estimates $(\hat{F}, \hat{\phi}, \hat{q})$. This iterative process involves updating model parameters and hidden variables.

57.2 Initialization

The EM algorithm begins with initial estimates:

$$(\hat{F}^{[0]}, \hat{\phi}^{[0]}, \hat{q}^{[0]}, \Delta_{1_i}^{[0]}, \Delta_{0_i}^{[0]}).$$

To update from step $[t-1]$ to $[t]$, the algorithm transforms $(\Delta_{1_i}, \Delta_{0_i})$ to $(\Delta_{1_i}^{[t-1]}, \Delta_{0_i}^{[t-1]})$.

57.3 E-Step and M-Step

The Expectation (E) step involves estimating hidden variables $(\Delta_{1_i}^{[t-1]}, \Delta_{0_i}^{[t-1]})$. The Maximization (M) step updates $(\hat{F}^{[t-1]}, \hat{\phi}^{[t-1]}, \hat{q}^{[t-1]})$ to $(\hat{F}^{[t]}, \hat{\phi}^{[t]}, \hat{q}^{[t]})$.

57.4 Model Update

For optimal prediction $\hat{F}^{[t]}$, the accelerated gradient method of a tree can be used (Yang et al. 2017). The update equations are given as:

$$\hat{F}^{[t]} = \arg\max_{F} \sum_{i=1}^{n} \Delta_{1_i}^{[t-1]} \omega_i \left(y_i \frac{\exp(F(x_i)(1-\rho))}{1-\rho} - \frac{\exp(F(x_i)(2-\rho))}{2-\rho} \right), \tag{97}$$

$$\phi^{[t]} = \arg\max_{\phi} \sum_{i=1}^{n} \Delta_{1_i}^{[t-1]} \left(\log a(y_i, \frac{\phi}{\omega_i}, \rho) + \frac{\phi}{\omega_i} \left(y_i(\hat{\mu}_i^{[t]}) - \frac{(\hat{\mu}_i^{[t]})^{2-\rho}}{2-\rho} \right) \right). \tag{98}$$

57.5 Tree-Based Accelerated Gradient Method

The tree-based accelerated gradient method in the *TDboost* package can be used for optimal prediction. The loss function ψ with observations (Y_i, x_i, ω_i) is:

$$\psi(y_i, F(x_i)|\rho) = \omega_i \left(\frac{\mu_i^{1-\rho}}{1-\rho} + \frac{\mu_i^{2-\rho}}{2-\rho} \right). \tag{99}$$

57.6 Shrinkage Factor and Computational Considerations

The shrinkage factor ν is typically set to a value less than 0.1, based on empirical findings for optimal predictive accuracy. The choice of ν is deter-

mined by the number of iterations M, with a trade-off between accuracy and computational time.

57.7 Conclusion

The EM algorithm is a powerful tool for estimating parameters in mixed models. By iteratively updating predictions and model parameters, it facilitates the maximum likelihood estimation process. The use of tree-based accelerated gradient methods further enhances the model's predictive power. However, computational efficiency must be balanced with accuracy.

58 Zero-Inflated TDboost

1: Data Collection

The dataset (y_i, x_i, ω_i) includes a zero index set $\tau = \{i \mid y_i = 0\}$. At the T-th processing stage in the Zero-Inflated Function (ZIF) model, the components ρ are placed, then set $\triangle_{1i} = 1$ and $\triangle_{0i} = 0$.

2: Initial Estimation

Estimate $(\hat{F}^{[0]}, \hat{\phi}^{[0]}, \hat{q}^{[0]})$ and $(\triangle_{1i}^{[0]}, \triangle_{0i}^{[0]})$.

3: Iterative Process for $t = 1, \ldots, T$

A: Maximization Step: Utilize $(\hat{F}^{[t-1]}, \hat{\phi}^{[t-1]}, \hat{q}^{[t-1]})$ and $(\triangle_{1i}^{[t-1]}, \triangle_{0i}^{[t-1]})$.
B: Update Step: Update $(\triangle_{1i}^{[t-1]}, \triangle_{0i}^{[t-1]})$ with $(\hat{F}^{[t]}, \hat{\phi}^{[t]}, \hat{q}^{[t]})$. with $(\hat{F}^{[t]}, \hat{\phi}^{[t]}, \hat{q}^{[t]})$

C: Calculation of the Maximum Log Correction in the ZIF Model: Description or content goes here.

$$(\hat{F}^{[t]}, \hat{\phi}^{[t]}, \hat{q}^{[t]}) = \sum_{i=1}^{n} \left[(1 - \hat{q}^{[t]}) f_{TW}(y_i, \hat{\mu}_i^{[t]}, \frac{\hat{\phi}^{[t]}}{\omega_i}, \rho) + \hat{q}^{[t]} I(y_i = 0) \right]$$

D: Calculation of the Maximum Difference Between the Former Groups and the Updated Samples:

$$\sigma^{[t]} = \max\left(|\hat{\phi}^{[t]} - \hat{\phi}^{[t-1]}|, |\hat{q}^{[t]} - \hat{q}^{[t-1]}|, \max_i |\hat{\mu}_i^{[t]} - \hat{\mu}_i^{[t-1]}|\right)$$

If there is a disagreement, revert to $(\hat{F}, \hat{\phi}, \hat{q}) = (\hat{F}^{[t]}, \hat{\phi}^{[t]}, \hat{q}^{[t]})$.

Selection of the Optimal Repetition Stage $\hat{t}$:

$$\hat{t} = \arg\max_{t=1,\ldots,T} l(\hat{F}^{[t]}, \hat{\phi}^{[t]}, \hat{q}^{[t]})$$

Subsequently, set $(\hat{F}, \hat{\phi}, \hat{q}) = (\hat{F}^{[\hat{t}]}, \hat{\phi}^{[\hat{t}]}, \hat{q}^{[\hat{t}]})$.

4: End of the Algorithm:

The algorithm concludes at this stage.

59 Chapter 4:

Numerical Calculations and Simulation Studies in Insurance Premium Prediction:

60 Linear Generalized Models vs. Accelerated Tree Models in Insurance Premium Prediction

60.1 Introduction

This chapter presents a detailed comparison between linear generalized models and accelerated tree models, particularly focusing on their application in predicting insurance premiums. The evaluation criterion for these models is based on prediction error or Mean Squared Error (MSE).

60.2 Key Points

- **Model Implementation and Comparison:** The *TDboost* algorithm, proposed by Yang et al. (2015), is employed for modeling, especially useful in scenarios with a high incidence of zero claims.
- **Algorithm Efficiency:** The efficiency of the *TDboost* algorithm is assessed against Generalized Linear Model (GLM) and Compound Poisson Linear Model (CPLM).
- **Data Utilization:** Simulated data based on the *AutoClaim* model from the CPLM package is used for controlled environment testing.
- **Implementation Details:** The TDboost method, applied to insurance premium data, is implemented using the R software and the TDboost package.
- **Data Types:** The study differentiates between experimental data (researcher-generated) and observational data (naturally occurring data).
- **Data Division:** The data is split into training and testing datasets for model development and performance evaluation.

60.3 Model Building Steps

- Identification of observed data for prediction.
- Determination of the optimal number of trees for the prediction process.

- Comparison of the *TDboost* model forecasts with GLM and CPLM models.
- Simulation Study Outcome: Demonstrating the superiority of *TDboost* in complex scenarios compared to *GLM*.

60.4 Analysis of Educational Dataset for Automobile Insurance Claims

This section focuses on analyzing a dataset comprising automobile insurance claims, with variables such as commercial vehicle use, marital status, and geographical driving locations.

60.4.1 Data Set Overview

The dataset analyzed by Yau et al. (2005) includes 10,296 insurance policies detailing total claim amounts and product types.[38]

60.4.2 Model Variables

- The dataset includes explanatory variables and payment amount variables.
- The payment amount follows a Compound Poisson distribution.

60.4.3 Model Trees and Miniaturization

- A total of 3000 model trees are used.
- Friedman's shrinkage parameter ν is set to 0.005 for optimal balance between accuracy and computational resources.

60.5 Conclusion

The comparative analysis of linear generalized models and accelerated tree models, such as *TDboost*, in insurance premium prediction demonstrates the efficacy of these approaches in different scenarios. The detailed study, encompassing various model variables and extensive data analysis, provides insights into the effectiveness of these models in the realm of insurance. Tree Size and

Interaction Control: The size of the trees in the model controls the extent of interaction among variables. Allowing for J=2 interactions, the model can accommodate up to two variable changes. If $J = 3$, the model may include interactions up to two changes.

This comprehensive analysis of the educational *dataset* model provides a deep understanding of the factors influencing automobile insurance claims. By meticulously examining these variables and employing a robust statistical model, the study aims to accurately predict insurance claim amounts, thereby aiding in more effective policy management and risk assessment.

According to Hesti et al., for effective modeling, the ideal range for the parameter J is between 4 and 8. AJ value of 2 is often insufficient for many applications, while a J greater than 10is rarely necessary, as suggested by Friedman (2001).[8]

In this context, a model featuring multiple interactions is employed, specifically with J set to 3. This approach involves cross-validation using a dataset divided into five subsets, aligning with the definition of a tree structure with three branches per tree. Each of these trees is relatively small and manageable, allowing for efficient processing and analysis using the R software and the TDboost package.

The model developed incorporates an EDM (Empirical Distribution Model) loss function. For the training phase, a subset of 10,266 data points is selected from the entire *dataset*.

The first stage involves identifying and introducing all relevant variables that will effectively contribute to the model. A summary table is prepared to provide an overview of the *dataset*, outlining the key variables and their characteristics. This structured approach ensures a comprehensive understanding and effective utilization of the data in the subsequent modeling and analysis phases.

Table 11: Variables used in the model

Number	Variable	Variable Type	Type	
1	Age	Independent	Numerical	
2	Vehicle Value	Independent	Numerical	
3	Number of Children (Driver)	Independent	Numerical	
4	Number of Children at Home	Independent	Numerical	
5	Recorded Points for Motor Vehicle	Independent	Numerical	
6	Number of Insurance Letters	Independent	Numerical	
7	Commute Time	Independent	Numerical	
8	Traffic Area	Independent	Categorical	U
9	Vehicle Usage	Independent	Categorical	C
10	Type of Vehicle	Independent	Categorical	1=Truck, 2=Van, 3
11	Gender	Independent	Categorical	
12	Job Category	Independent	Categorical	Wo
13	Education	Independent	Categorical	High S
14	Marital Status	Independent	Categorical	
15	Cancellation	Independent	Categorical	

Table 12: a summary of the specification of vehicle data variables

variable	Age	Number of Children at Home	Vehicle Value	Number of Children (Driver)
Min.	16.00	0.0000	1500	0.0000
1st Qu.	39.00	0.0000	9200	0.0000
Median	45.00	0.0000	14405	0.0000
Mean	44.84	0.7199	15666	0.1694
3rd Qu.	51.00	1.0000	20900	0.0000
Max.	: 81.00	5.0000	69740	4.0000

At first saw that:

Table 13: General information table of vehicle data

Min	1st Qu.	Median	Mean	3rd Qu.	Max.
0.000	0.000	0.000	4.112	4.740	57.037

And since it has to be done with this, it is equal to: 0.6066856 Driving table in the city or suburbs: [38]

Table 14: Table of driving in the city or suburbs

Area	
Rural	2072
Urban	8224

Table 15: Table of personal and commerical vehicle

CAR_USE	
Private	6507
Commercial	3789

Table 16: Frequency of car type

CAR_TYPE	
Panel Truck	853
Pickup	1772
Sedan	2693
Sports Car	1176
SUV	2881
Van	921

Table 17: Frequency of gender

GENDER	
F	5540
M	4756

Table 18: Frequency of marital status

MARRIED	
No	4109
Yes	6187

Table 19: Frequency dependent variable total amount of loss cliam

CLM_AMT5	
Min.	0
1st Qu.	0
Median	0
Mean	4032
3rd Qu.	4646
Max.	57037

Table 20: General information

CAR USE	MAX EDUC	CAR TYPE	JOBCLASS
Private:63.2%	<High School:14.6%	Panel Truck:8.3%	Blue Collar:22.2%
Commercial:36.8%	Bachelors: 27.3%	Pickup: 17.3%	Clerical:15.5%
	High School: 28.7%	Sedan: 26.2%	Professional: 13.6%
	Masters: 20.2%	Sports Car:11.4%	Manager: 12.2%
	PhD:9.2%	SUV:27.9%	Lawyer: 10.0%
		Van: 8.9%	Student: 8.7%
			(Other):17.8%

Table 21: Descriptive Statistics of training set data

AREA	MARRIED	REVOKED	GENDER
Rural:20.2%	No:39.9%	No: 87.8%	F:53.8%
Urban: 79.8%	Yes: 60.1%	Yes: 12.2%	M:46.2%

Correlation coefficients:

The interrelation between two quantitative variables is articulated through the computation of a coefficient. This coefficient delineates the inherent relational dynamics and mutual dependencies of the variables under consideration. From this understanding, one can infer the following conclusion: If

Table 22: Correlation coefficients

	CLM_AMT5
AGE	0.009015543
HOMEKIDS	0.007464888
BLUEBOOK	−0.04479927
KIDSDRIV	0.02565067
MVR_PTS	0.2846911
NPOLICY	−0.009077763
TRAVTIME	−0.003670452

the correlation coefficient approaches a value close to or equal to 1, it indicates a direct relationship between the two variables. This implies that an increase in one variable is accompanied by a corresponding increase in the other. Conversely, a correlation coefficient nearing -1 signifies a strong inverse relationship between the variables. Thus, as one variable increases, the other decreases, yet this inverse variation still permits predictable patterns.

Description of variable with diagrams:

Since:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} l(y_i, \gamma)$$
$$= argmin_{\gamma} \sum_{i=1}^{n} (\gamma - y_i)^2$$
$$= \frac{1}{n} \sum_{i=1}^{n} y_i$$

In the analytical framework where Γ signifies the predicted value and y_i represents the observed data, the median of Y is defined as F_0. The residual corresponding to each explanatory variable is computed as $y - F_0$.
The accompanying graphical representations function as instruments for clarification. These diagrams are instrumental in facilitating pairwise comparisons among variables.
The following diagrams help to describe variables:

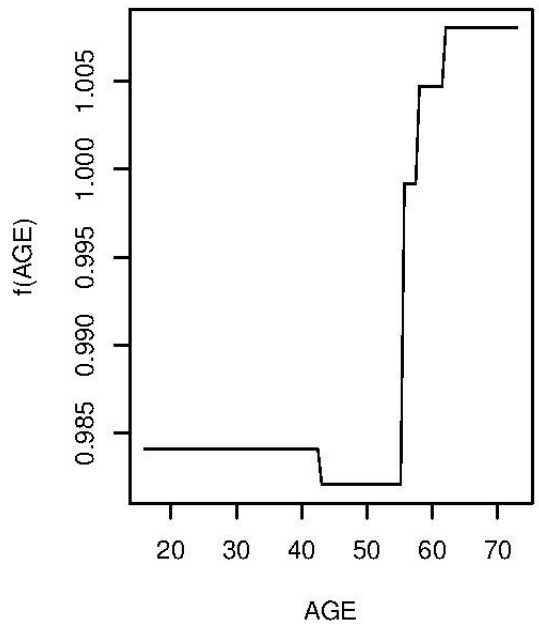

Figure 8: Age

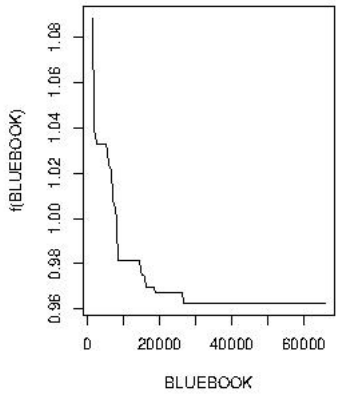

Figure 9: Vehicle Value

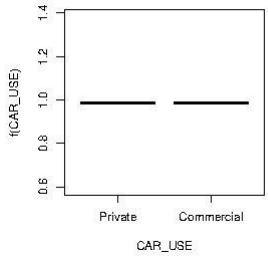

Figure 10: Car USE

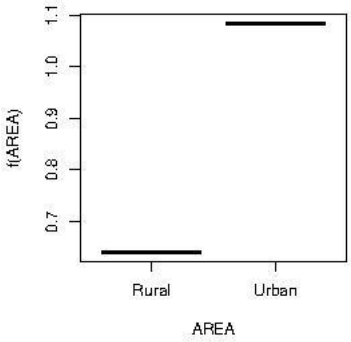

Figure 11: Traffic Area

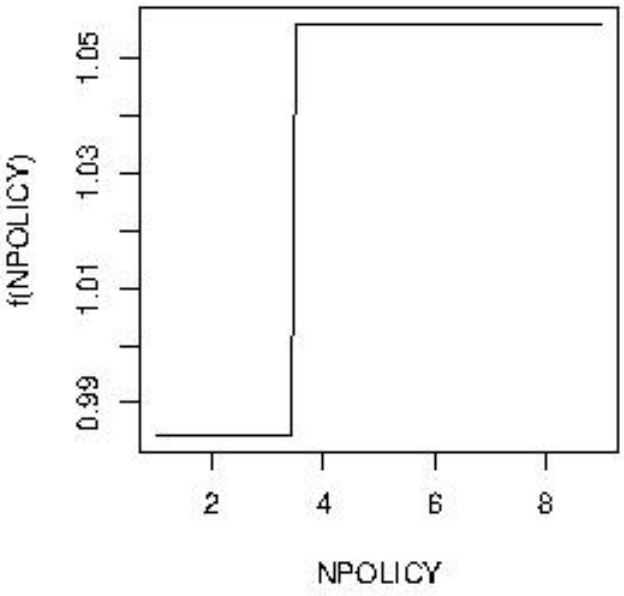

Figure 12: Number of insurance policies

61 Bivariate Comparative Analysis of Variables

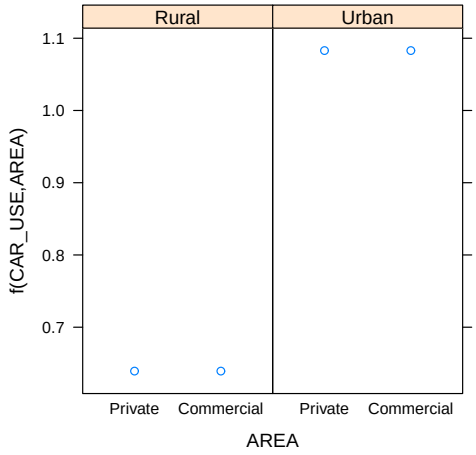

Figure 13: Comparison of traffic area and used car

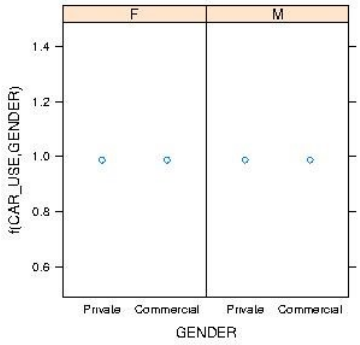

Figure 14: Comparison of gender and car used

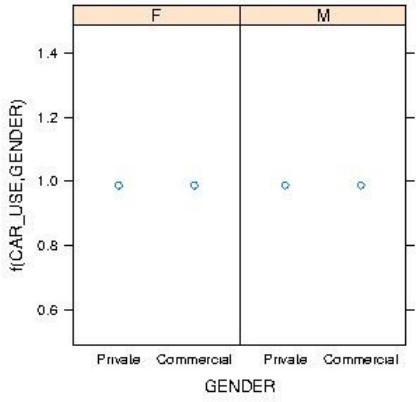

Figure 15: Comparison of gender and commuting time of a person

62 Assessment of Model Efficacy on Training Data

62.1 Evaluation of Predictors

In the initial phase of the analysis, a comprehensive evaluation of fifteen million predictors was conducted. A noteworthy observation was that these predictors collectively amounted to a zero value, highlighting a significant characteristic of the initial data.
newpage

Table 23: Table of Relative Effects of Variables

ID	Variable	Relative Influence (%)
1	REVOKED	25.12
2	MVR_PTS	20.72
3	JOBCLASS	14.77
4	BLUEBOOK	10.68
5	TRAVTIME	7.55
6	AGE	5.61
7	CAR_TYPE	4.93
8	MAX_EDUC	3.01
9	NPOLICY	2.85
10	AREA	2.24
11	MARRIED	0.71
12	HOMEKIDS	0.67
13	KIDSDRIV	0.45
14	GENDER	0.35
15	CAR_USE	0.34

62.1.1 Cross-Validation Evaluation

The analysis employs a quintuple cross-validation approach, generating five distinct evaluation charts. These charts are critical in providing a comparative assessment of the relative values within the automobile dataset. The quintuple cross-validation methodology ensures a comprehensive and robust evaluation, effectively highlighting both similarities and differences in the dataset's inherent characteristics.

62.2 Conclusion

The assessment of model efficacy on training data, involving an extensive analysis of predictors and the application of quintuple cross-validation, offers valuable insights into the predictive capabilities of the model. The detailed evaluation, as represented in the relative effects table and cross-validation charts, serves as a basis for understanding the strengths and weaknesses of the model in accurately predicting outcomes in the automobile *dataset*.

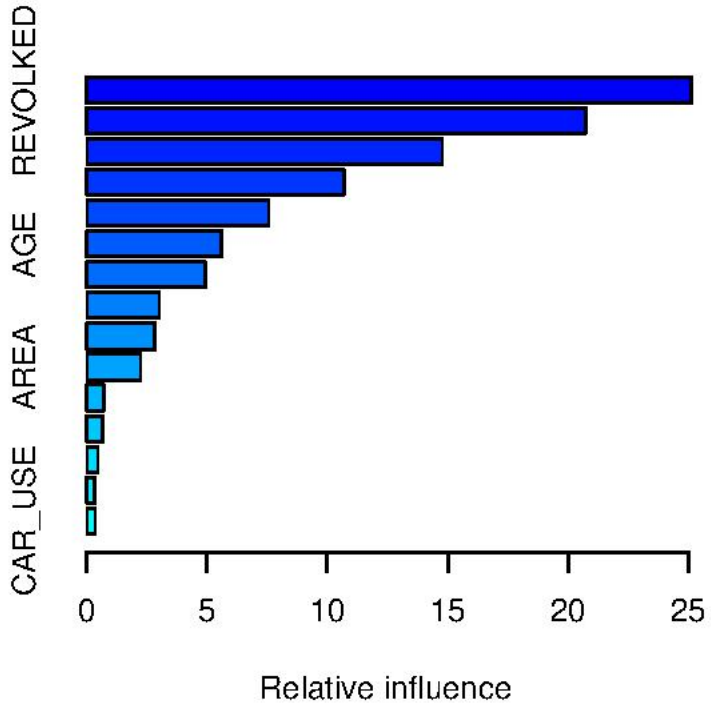

Figure 16: Relative influence of explanatory variables

63 Identification of Critical Predictors in Insurance Policy Modeling

63.1 Analysis of Variable Importance

The comprehensive assessment of variables in the model reveals the most significant factors influencing insurance policy predictions. This analysis is crucial in understanding which settings are pivotal to the model's performance.

63.2 Findings from the Relative Importance of Variables

- **Top Influential Factor:** The analysis indicates that the revocation of the insurance policy in the past seven years holds the greatest impact on the model.
- **Least Influential Factor:** Conversely, the use of the car, whether for personal or commercial purposes, appears to have the least relative importance.

63.3 Negative Impact Analysis

The study also explores functions that potentially exert a negative effect on the model. These functions are essential in refining the model and enhancing its predictive accuracy.

63.4 Cross-Validation (CV) and Testing

- The use of Cross-Validation (CV) and testing methodologies is vital in this context. These approaches help in identifying and addressing any negative impacts on the model.
- The objective is to discern patterns and relationships that might not be immediately apparent but significantly affect the model's predictions.

63.5 Optimizing the Prediction Model

- **Selection of the Best Tree:** The ultimate goal is to determine the most effective tree for prediction. This process involves analyzing various tree structures and their respective contributions to the model's performance.
- **Methodology:** The selection of the optimal tree is based on evaluating different tree configurations and their impacts on the model's MSE. This approach ensures that the chosen tree maximizes prediction accuracy and reliability.

63.6 Conclusion

This chapter underscores the importance of identifying key predictors and their relative impacts on insurance policy modeling. By analyzing variable importance, addressing potential negative effects, and employing robust testing methods like CV, the model can be fine-tuned for optimal performance. The selection of the best predictive tree plays a crucial role in this process, leading to more accurate and reliable insurance premium predictions.

64 Analysis of Accelerated Tree Permutation Algorithms

64.1 Graphical Representation of Algorithm Performance

The graph illustrates the performance of three accelerated tree permutation algorithms as a function of the number of iterations, denoted as M. This is represented by three distinct curves:

- **Black Line:** Represents the loss deviation for the training data.
- **Green Line:** Shows the loss deviation for the experimental data.
- **Red Line:** Indicates the tree minimizing the test error against cross-validation, hence chosen for prediction.

64.2 Optimal Number of Trees and Cross-Validation

Each line in the graph represents the optimal number of trees determined by cross-validation (CV) methodology. Using the coefficient of variation in our estimation helps to ensure the robustness of the model.

64.3 Loss Function as a Model Appropriateness Index

The loss function serves as an index to measure the model's appropriateness in terms of experience in new predictions. It quantifies the loss incurred in an event and is crucial for evaluating model performance.

64.3.1 Error Criteria: Mean Squared Error (MSE)

- The Mean Squared Error (MSE) is calculated as $\text{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$.
- Based on the obtained data, the MSE value is reported to be 1299, indicating the mean square test error for each n tree.

64.3.2 Root Mean Square Error (RMSE)

- RMSE, defined as $\text{RMSE} = \sqrt{\frac{1}{m}\sum_{j=1}^{n}(y_i - \hat{y}_i)^2}$, measures the standard deviation between predictions and observations.
- Alternatively, RMSE can be expressed as $\text{MSE} = \text{mean}(\text{residuals}^2)$.
- The calculated minimum MSE was found to be 2.674243.

64.4 Conclusion

The analysis of accelerated tree permutation algorithms using MSE and RMSE criteria provides a comprehensive understanding of their performance in predicting outcomes. The graphical representation aids in visualizing the optimal number of trees and the effectiveness of each algorithm across different data sets. This approach ensures the selection of the most efficient algorithm for accurate predictions.

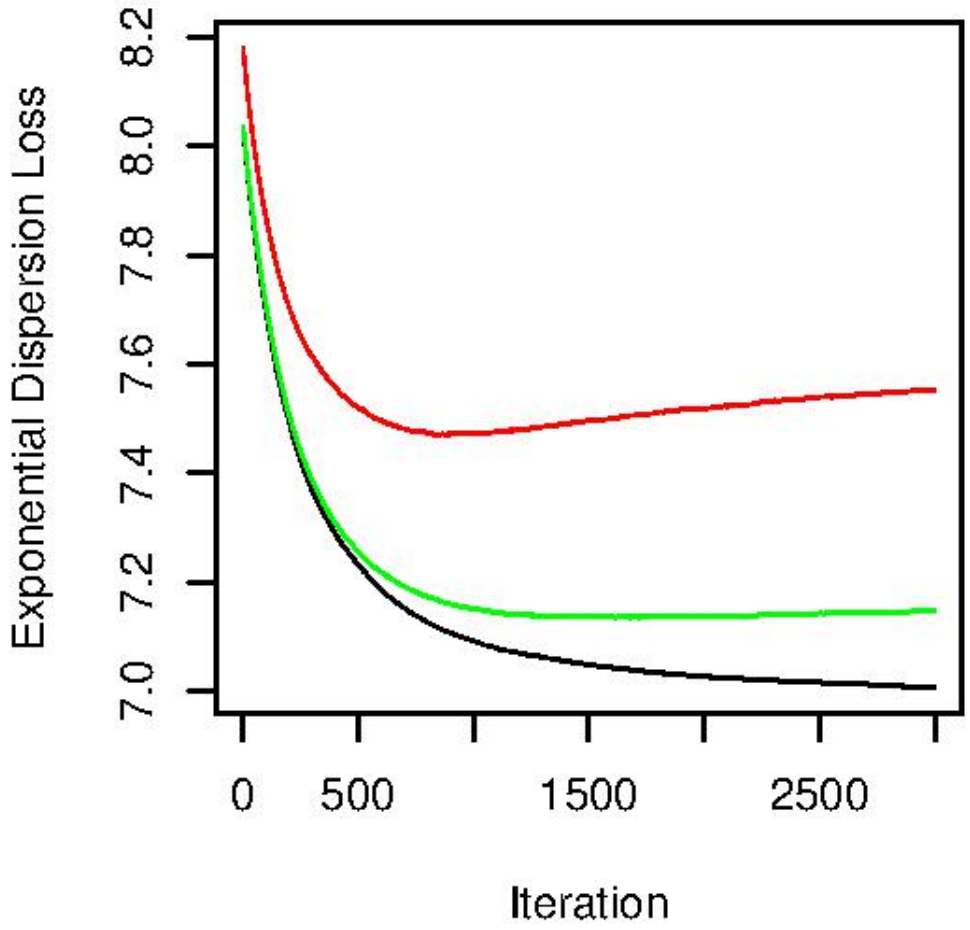

Figure 17: CV diagram

65 Analysis of Model Errors and Over fitting in Training and Test Data

65.1 Graphical Representation of Model Errors

A graph has been plotted to illustrate the variation in model errors for both training and test data as the model complexity evolves. This graph serves as a critical tool for understanding the dynamics between model accuracy and complexity.

65.2 Interpretation of the Graph

- **Black Line (Training Data Error):** Indicates the divergence of loss from the training data. As the model undergoes more iterations, the error rate for training data decreases, signifying model learning and adaptation.
- **Red Line (Test Data Error):** Denotes the error for test data. An increase in error values is observed when the number of model iterations exceeds a certain threshold (e.g., 600 iterations), which is indicative of over fitting.

65.3 Over fitting Phenomenon

Over fitting occurs when the model becomes excessively tailored to the training data, compromising its ability to generalize and perform effectively on new, unseen data. This issue is highlighted by the increase in test data error beyond a certain number of iterations.

65.4 Cross-Validation to Mitigate Over fitting

Cross-validation plays a pivotal role in identifying and mitigating over fitting. It ensures the model maintains a balance between accuracy and generalizability.

65.5 Loss Function: Mean Squared Error (MSE)

- The loss function used is the squared error field, MSE = mean$((y_i - \hat{y}_i)^2)$.
- For the experimental data, MSE is reported as 881, indicating the level of error in model predictions.
- A lower Root Mean Square Error (RMSE) implies better performance, corroborating the effectiveness of cross-validation in model assessment.

65.6 Model Prediction in Training*datasets*

- The prediction in the training *dataset* is evaluated by calculating the test error as a function of the number of trees.
- Clustering large volumes of data and creating a tree model that bifurcates its branches based on distinct changes is a key method in data modeling.
- By generating a prediction board for each tree, the first tree model is created, allowing for detailed analysis of model predictions and adaptations.

65.7 Conclusion

The analysis of model errors in training and test data, along with the assessment of over fitting through cross-validation, provides essential insights into model performance. Understanding these dynamics is crucial for developing robust models that are both accurate and generalizable. The use of MSE and $RMSE$ as error metrics further aids in evaluating and enhancing the model's predictive accuracy.

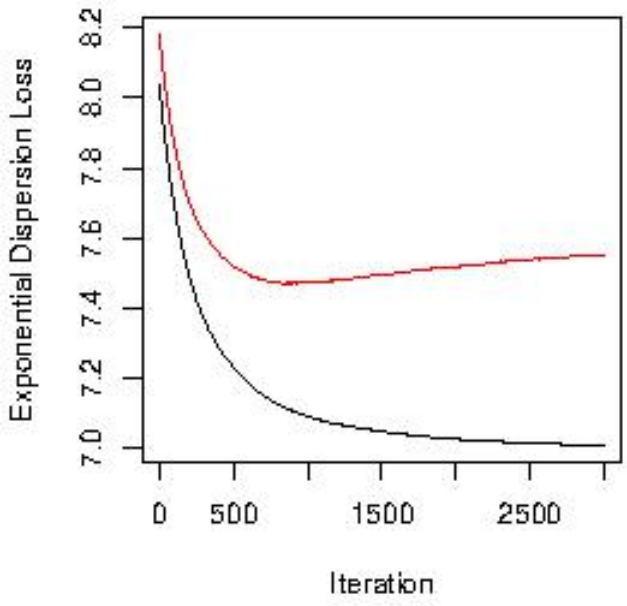

Figure 18: diagram of test data

66 Impact of Variable Changes in the New data set on Model Prediction

66.1 Introduction

The new data set introduces various changes, each influencing the model's predictive capabilities. This section explores how these changes affect the model and the significance of each variable in the prediction process.

66.2 Analysis of Variable Impact

Each variable in the data set contributes differently to the model's prediction.The clustering results help in identifying which variables have a high, moderate, or low impact on the overall prediction accuracy. This analysis is

crucial for understanding the dynamics of the model and for making informed decisions about variable selection and model tuning.

66.3 Model Adaptation to New Data

The introduction of new data necessitates adjustments in the model.The clustering results guide these adaptations by highlighting the variables that require more focus. This iterative process ensures that the model remains relevant and accurate as the data characteristics evolve.

66.4 Conclusion

The analysis of the new*dataset*and the impact of variable changes on model prediction is essential in the field of data science. Understanding how different variables influence the model's predictions allows for more effective model development and optimization. The clustering of the first tree sets the foundation for this analysis, guiding subsequent model adjustments and improvements.

Table 24: Interactions in the first tree

ID	var	rel.inf
1	REVOLKED	66.69582
2	MVR_PTS	33.30418
3	BLUEBOOK	0.00000
4	HOMEKIDS	0.00000
5	AGE	0.00000
6	KIDSDRIV	0.00000
7	MARRIED	0.00000
8	NPOLICY	0.00000
9	AREA	0.00000
11	CAR_USE	0.00000
12	CAR_TYPE	0.00000
13	GENDER	0.00000
14	JOBCLASS	0.00000
15	MAX_EDUC	0.00000

66.5 Branch Development and Stagnation Analysis

Each branch undergoes a unique development process, including a stagnation analysis. This analysis is essential for determining the effectiveness and progression of each branch within the model. The key steps in this stage are as follows:

1. **Branch Creation:** The model diverges into two distinct branches, each representing different aspects or variables of the data set.
2. **Stagnation Analysis:** For each branch, a stagnation analysis is conducted to assess its growth and performance. This involves evaluating whether the branch has reached a point of minimal or no further improvement in predictive accuracy.
3. **Model Enhancement:** Based on the stagnation analysis, necessary adjustments are made to optimize each branch. This may include pruning ineffective parts of the branch or enhancing it with additional variables and parameters.

66.6 Building the Composite Model

The final model is constructed by integrating insights and improvements from both branches. This composite model aims to leverage the strengths of each branch while mitigating any weaknesses identified during the stagnation analysis. The model building process includes:

- **Synthesizing Branches:** Combining the two branches into a cohesive model that encapsulates the comprehensive analysis of the data set.
- **Refinement:** Refining the composite model to ensure that it accurately represents the data and provides reliable predictions.
- **Validation:** Validating the composite model against a set of criteria or benchmarks to confirm its efficacy and readiness for deployment.

66.7 Conclusion

The branching and stagnation analysis phase is a pivotal step in model development. It allows for a detailed examination of different model aspects

and the creation of a robust, composite model. This approach ensures that the final model is well-rounded, accurate, and capable of making reliable predictions based on the data set.

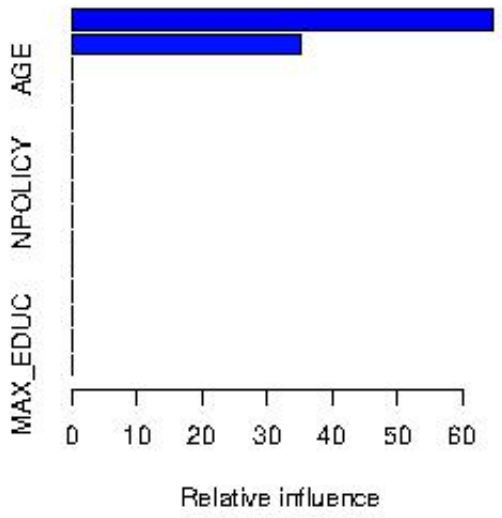

Figure 19: Interactions in the second tree

Table 25: Interactions in the second tree estimate

ID	var	rel.inf
1	REVOLKED	64.87773
2	MVR_PTS	35.12227
3	JOBCLASS	0.00000
4	AREA	0.00000
5	BLUEBOOK	0.00000
6	NPOLICY	0.00000
7	MAX_EDUC	0.00000
8	AGE	0.00000
9	CAR_TYPE	0.00000
10	TRAVTIME	0.00000
11	MARRIED	0.00000
12	HOMEKIDS	0.00000
13	KIDSDRIV	0.00000
14	CAR_USE	0.00000
15	GENDER	0.00000

67 Development of an Optimal Prediction Model with Minimal Loss

67.1 Objective

The primary goal in this stage of model development is to construct a prediction model that achieves the least possible loss for our dataset. This involves building a model that is finely tuned to the specific characteristics and patterns present in the data.

67.2 Building the First Prediction Model

The initial step in this process is the creation of the first prediction model, which is based on our tree structure. This model serves as a foundational framework upon which further refinements and optimizations are made.

67.2.1 Tree-Based Model Assessment

- The first model is constructed by evaluating the relative nature and performance of the tree with n predictors.
- This assessment involves analyzing how each branch and decision point in the tree contributes to the overall prediction accuracy.
- The tree structure allows for a hierarchical and segmented approach to modeling, enabling a detailed understanding of the data and its underlying trends.

67.3 Optimization for Minimal Loss

In pursuit of the least loss, the following steps are undertaken:

1. **Model Refinement:** Iterative refinements are made to the initial model, focusing on areas identified as having significant impact on loss.
2. **Variable Importance:** Variables are evaluated based on their importance and contribution to the prediction accuracy. Less impactful variables may be pruned to streamline the model.
3. **Loss Evaluation:** Continual assessment of the model's loss is conducted to gauge improvements and identify areas for further optimization.

67.4 Finalizing the Prediction Model

Upon achieving satisfactory results in terms of minimized loss, the prediction model is finalized. This model encapsulates the most critical predictors and effectively captures the dynamics of the data set.

67.5 Conclusion

The development of a prediction model with minimal loss requires a structured and analytical approach, starting with a tree-based model and iteratively refining it for optimal performance. The final model is a culmination of careful assessments and optimizations, tailored to provide the most accurate predictions for our data set.

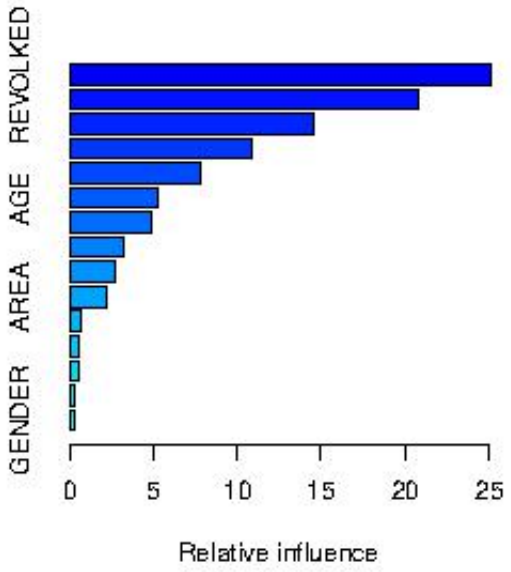

Figure 20: Interactions in the 3000th tree

Table 26: Interactions in the 3000th tree estimate

ID	var	rel.inf
1	REVOLKED	25.1834499
2	MVR_PTS	20.8488540
3	JOBCLASS	14.5475160
4	BLUEBOOK	10.9121453
5	TRAVTIME	7.8068738
6	AGE	5.2469103
7	CAR_TYPE	4.8436488
8	MAX_EDUC	3.2395334
9	NPOLICY	2.7555207
10	AREA	2.1598513
11	HOMEKIDS	0.7108561
12	MARRIED	0.6146097
13	KIDSDRIV	0.5076301
14	CAR_USE	0.3243115
15	GENDER	0.2982892

68 Model Prediction Using Test Data

68.1 Continued Iterations for Prediction Accuracy

In this critical phase, the model undergoes an additional 20 iterations, with a specific focus on making accurate predictions on the test data set. The primary goal of these iterations is to attain a level of prediction accuracy and quality that is on par with the results achieved by the TDboost method.

68.2 Approach with an Extended Tree Model

An alternative strategy considered in this stage involves the deployment of a model configured with 3000 trees. The specifics of this approach are as follows:

- **Creation of a New Tree Model:** A new tree model is constructed for evaluation, differing from the previous iterations in its complexity and depth.
- **Objective:** The expectation is to achieve satisfactory results, compa-

rable in efficacy to those obtained from the preceding models.

- **Rationale:** This approach is based on the hypothesis that an increased number of trees might contribute to enhanced model accuracy and generalization.

68.3 Incorporating a Gamma Model

This phase marks the integration of a Gamma model into the existing predictive framework, aimed at further enhancing the model's capabilities. Key aspects of this integration include:

- **Model Enhancement:** The Gamma model is introduced to improve the handling of diverse data distributions and complexities, particularly beneficial in scenarios with skewed data.
- **Statistical Flexibility:** The Gamma model is renowned for its flexibility and efficacy in statistical modeling, making it a valuable addition to the predictive toolkit.
- **Strategic Inclusion:** The inclusion of the Gamma model is strategically aimed at refining the overall performance and accuracy of the model in predicting outcomes from the test data.
- **Expected Outcome:** The integration of the Gamma model is anticipated to bring an additional layer of robustness and reliability to the model's predictive analysis.

68.4 Conclusion

The extension of the model through additional iterations, the exploration of an alternative approach with an extended tree model, and the incorporation of a Gamma model collectively contribute to the advancement of the model's predictive accuracy. These methodologies, each focusing on different aspects of model enhancement, are instrumental in refining the model's performance and ensuring its effectiveness in accurately predicting outcomes from the test data.

Table 27: Gamma Model fit

Iter	TrainDeviance	StepSize	Improve
1	8.1106	0.0010	0.0010
2	8.1097	0.0010	0.0009
3	8.1089	0.0010	0.0010
4	8.1079	0.0010	0.0008
5	8.1070	0.0010	0.0009
6	8.1060	0.0010	0.0009
7	8.1050	0.0010	0.0009
8	8.1041	0.0010	0.0010
9	8.1032	0.0010	0.0009
10	8.1021	0.0010	0.0008
100	8.0258	0.0010	0.0008

The accompanying table delineates the quantity of iterations as documented by Fisher in the terminal phase of his modeling process. Subsequently, a parallel procedure has been replicated employing the compound Poisson model.

Table 28: Fitting the composite Poisson model

Iter	TrainDeviance	StepSize	Improve
1	8.1105	0.0010	0.0009
2	8.1097	0.0010	0.0010
3	8.1088	0.0010	0.0010
4	8.1078	0.0010	0.0009
5	8.1068	0.0010	0.0009
6	8.1059	0.0010	0.0009
7	8.1050	0.0010	0.0009
8	8.1039	0.0010	0.0008
9	8.1030	0.0010	0.0010
10	8.1022	0.0010	0.0009
100	8.0260	0.0010	0.0007

The tabulations displayed herein expound upon the configurations employed in the Compound Poisson model. Through a meticulous evaluation of the variables' efficacy within this model, it is conjectured that the Mean Squared Error (MSE) loss function is of paramount significance, particularly in light of the experimental data. The MSE quantitatively assesses the squared discrepancies between the forecasted and the observed values. Pertaining to the*dataset* under consideration, the MSE has been computed to be 976. This figure reflects the extent of deviation inherent in the model's prognostications compared to the actual empirical values. Such a metric is indispensable for appraising the precision and dependability of the model, especially in contexts necessitating forecasting and predictive analytical undertakings.

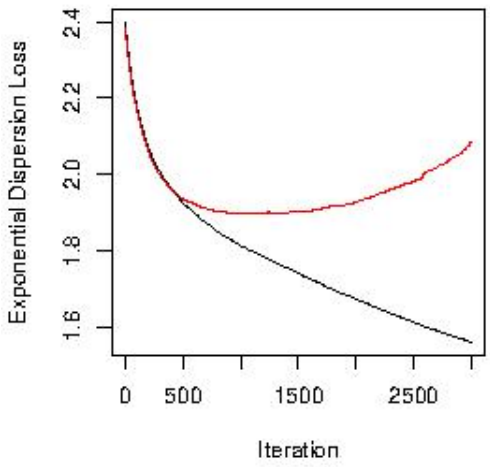

Figure 21: Relative effects based on the gamma model

In the examination of the model's complexity using training and test datasets, a graphical representation employing a Gamma distribution is utilized to delineate the variations in model error. This graphical depiction encompasses a black line, signifying the fluctuation in loss for the training dataset, juxtaposed with a red line that represents the test data. Notably, the graph elucidates a decline in the error rate of the model predicated on training data as the iteration count escalates. However, surpassing a specific iteration threshold, say 500 iterations, manifests an uptick in error rates, suggesting the emergence of potential over-fitting phenomena.
To address this, Cross-Validation (CV) is strategically implemented to ascertain the most judicious number of iterations. Within this framework, the Mean Squared Error (MSE) of the experimental data in the novel model is quantified as 686, shedding light on the models predictive precision.

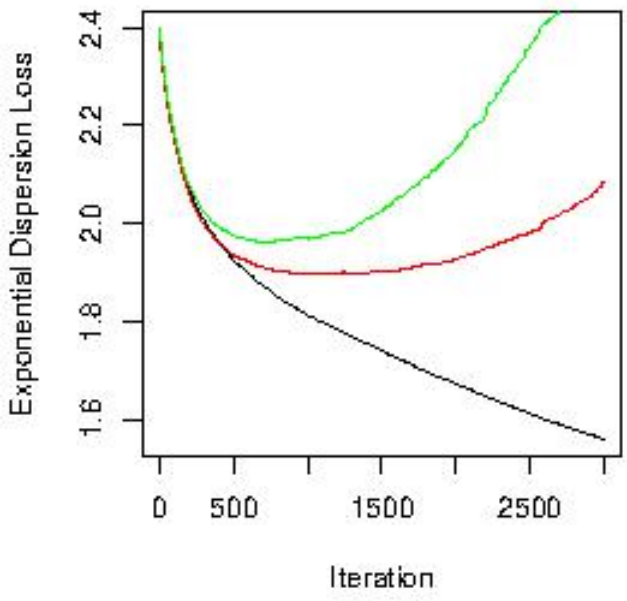

Figure 22: Method of CV

Further scrutiny is undertaken via the application of the Gini coefficient within the TD model. This process entails the evaluation of a spectrum of variables and the computation of their respective Gini coefficients, with the insurance claim functioning as the loss metric. Employing the TDboost package, the inaugural model, designated as model P1, undergoes a thorough assessment to gauge its efficacy and the influence of diverse variables on its predictive accuracy. This methodology is pivotal in appraising the model's overall capacity in forecasting insurance claims and refining it to augment accuracy and reliability for practical applications.

Table 29: Results for the first model in TDboost Method

Iter	TrainDeviance	StepSize	Improve
1	155.4129	0.0010	0.0031
2	155.4045	0.0010	0.0089
3	155.3943	0.0010	0.0099
4	155.3860	0.0010	0.0090
5	155.3762	0.0010	0.0096
6	155.3659	0.0010	0.0100
7	155.3570	0.0010	0.0088
8	155.3467	0.0010	0.0096
9	155.3370	0.0010	0.0094
10	155.3287	0.0010	0.0086
100	154.5791	0.0010	0.0072

Considering an array of explanatory variables, including but not limited to the type of vehicle, insurance policy review, gender, geographical traffic area, marital status, usage of the car, and other relevant factors, the subsequent model, designated as P2, integrates these independent predictors. This approach seeks to meticulously evaluate and assimilate the influence of these multifaceted factors into the models structure, thereby enhancing its predictive robustness and accuracy.

Table 30: Results for the second model in TDboost Method

Iter	TrainDeviance	StepSize	Improve
1	155.4074	0.0010	0.0097
2	155.3969	0.0010	0.0099
3	155.3883	0.0010	0.0091
4	155.3802	0.0010	0.0083
5	155.3709	0.0010	0.0091
6	155.3608	0.0010	0.0095
7	155.3527	0.0010	0.0089
8	155.3431	0.0010	0.0095
9	155.3327	0.0010	0.0100
10	155.3230	0.0010	0.0095
100	154.5940	0.0010	0.0083

Furthermore, in relation to model P3, an analogous methodological approach is employed. This model incorporates a similar spectrum of independent variables, as delineated for P2, with an emphasis on refining and augmenting the model's analytical depth. The aim is to further scrutinize the interactions and correlations among these variables, thereby enriching the predictive capacity and precision of the model:

Table 31: Result for third model in TDboost

Iter	TrainDeviance	StepSize	Improve
1	155.4131	0.0010	0.0030
2	155.4049	0.0010	0.0086
3	155.3938	0.0010	0.0105
4	155.3843	0.0010	0.0095
5	155.3796	0.0010	0.0039
6	155.3697	0.0010	0.0096
7	155.3659	0.0010	0.0034
8	155.3563	0.0010	0.0093
9	155.3459	0.0010	0.0099
10	155.3376	0.0010	0.0087
100	154.7543	0.0010	0.0024

The computation of the Gini coefficient, in the context where the loss function is epitomized by insurance claims, necessitates a specific methodological approach. This process entails a systematic calculation as follows:

Table 32: Gini index for the models presented in the*TDboost* method :

	P1	P2	P3
P1	0.000	−3.648	11.880
P2	3.767	0.000	11.881
P3	3.875	3.874	0.000

Table 33: Standard error for the models presented in *TDboost* method

	P1	P2	P3
P1	0.000	2.185	2.098
P2	2.204	0.000	2.098
P3	2.106	2.106	0.000

The data in the table indicates the standard error values for each pairwise comparison of the models. A lower standard error implies a higher level of precision in the model's predictions. Notably, model P1 demonstrates a zero standard error when compared with itself, which is an expected outcome. Similarly, models P2 and P3 exhibit a zero standard error in self-comparisons, underscoring the internal consistency of each model.

Upon comparing P1 with P2 and P3, it is observed that the standard errors are 2.185 and 2.098, respectively. These values suggest a moderate level of variability in the predictions when P1 is compared against the other two models. Conversely, the comparison of P2 with P3 yields a standard error of 2.098, indicating a similar level of predictive variability.

In conclusion, the analysis of standard errors in this table reveals critical insights into the comparative precision of models P1, P2, and P3 within the TDboost framework. While each model exhibits consistency in its predictions, the variability in their comparative analysis highlights the unique characteristics and performance metrics of each model Standard error for presented models in TDboost model:

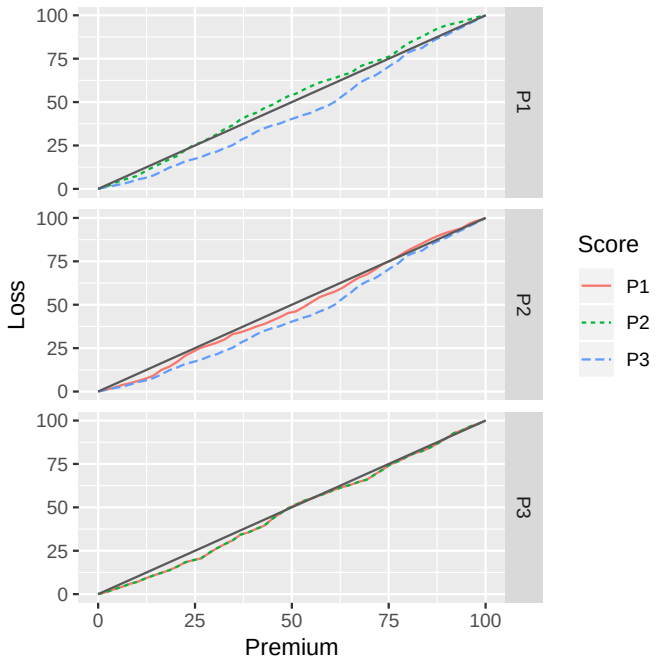

Figure 23: Gini coefficient for the models presented in the *TDboost* method

In this schematic representation, the horizontal axis is methodically arranged to display insurance premiums in an ascending order. Concurrently, the vertical axis quantitatively represents the proportion of total income attributable to varying percentages of the population. The concept of perfect equality is graphically depicted as a line inclined at a 45-degree angle.

The Gini coefficient is determined by calculating the ratio of the area that lies between the Lorenz curve and the 45-degree line to the total area beneath the 45-degree line. This ratio succinctly encapsulates the extent of inequality.

Moreover.

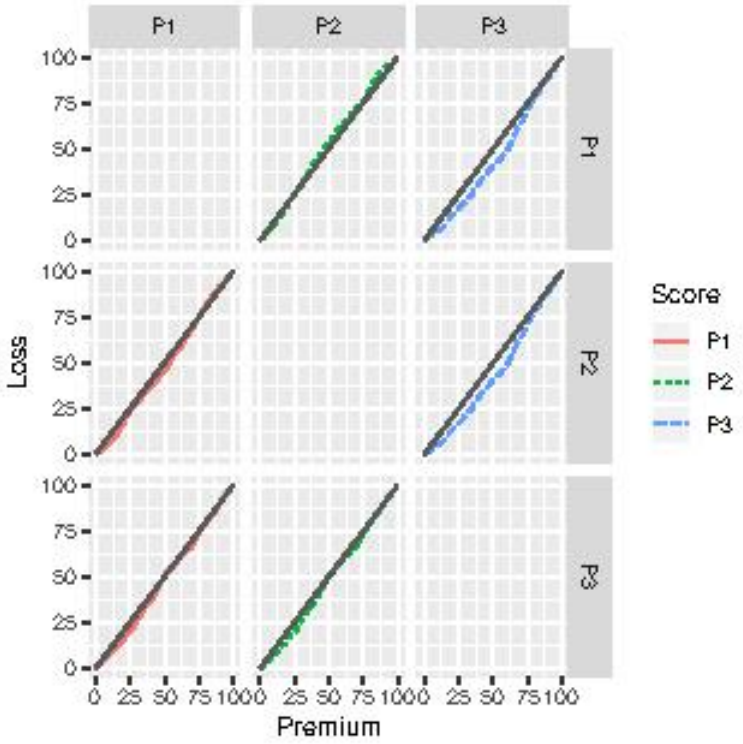

Figure 24: Gini coefficient for models presented in TDboost method

69 Tree Regression :

69.1 Dataset Overview

The dataset, characterized by its substantial volume, provides a robust basis for statistical analysis. The inclusion of 15 key changes further enhances the dataset's complexity and variability, offering a richer field for data exploration and model accuracy assessment:

Table 34: Tree Regression

	CP	nsplit	rel error	xerror	xstd
1	0.17461788	0	1.0000000	1.0002007	0.03400591
2	0.05137114	1	0.8253821	0.8258494	0.02607011
3	0.02627446	2	0.7740110	0.7747244	0.02628361
4	0.01000000	3	0.7477365	0.7485620	0.02568387

Variables and Their Significance:

CP (Complexity Parameter): Represents a threshold for pruning the tree. Lower CP values generally indicate more complex trees.
nsplit: Denotes the number of splits or nodes in the tree. It reflects the complexity of the model; more splits typically mean a more detailed model.
rel error (Relative Error): Indicates the error relative to the tree's complexity. A lower relative error suggests a more accurate model.
$xerror$(Cross-Validation Error): Reflects the model's error estimated via cross-validation. It's crucial for assessing the model's predictive performance.
$xstd$ (Standard Error of Cross-Validation Error): Provides an estimate of the variability in the cross-validation error.
Analysis of Table Entries:

The first row (1) with CP = 0.17461788 and no splits (nsplit = 0) shows the base model, which is the simplest form of the tree.
Subsequent rows reflect more complex models with additional splits. For instance, the second row (2) with CP = 0.05137114 and one split (nsplit = 1) shows a reduction in both relative error and cross-validation error, indicating an improvement in the model's performance.
As we progress to more complex models (rows 3 and 4), we observe further reductions in the relative error and cross-validation error, suggesting increased model accuracy. However, the diminishing returns in error reduction should be balanced against the increased complexity (more splits).

Conclusion and Implications:

The table's data suggests that as the complexity of the regression tree increases (more splits), the model becomes more accurate (lower relative and

cross-validation errors).
However, the choice of the optimal model should consider not only accuracy but also the complexity to avoid over-fitting. The standard error of cross-validation (xstd) can be a useful guide in this decision.
Further Considerations:

The selection of the best tree might involve choosing a balance between simplicity (fewer splits) and accuracy (low error rates).
In practical applications, this balance often necessitates a trade-off, where a slightly higher error might be acceptable for a significantly simpler model.
Variable Importance in Regression Trees:

Table 35: Variable Importance in first Regression Tree

REVOLKED	MVR_PTS
69	31

70 Analysis of Node Number 1 in Regression Tree

70.1 Introduction to Node Number 1

Node number 1 in the regression tree serves as a critical juncture, representing a specific subset of the dataset and playing a significant role in the initial stages of data segmentation.

70.2 Analysis of Observations

This node encompasses a total of $10,296$ observations, indicating a substantial portion of the dataset. This high number of observations suggests that node 1 captures a major segment of the *dataset*, possibly encompassing a wide range of the dependent variable's values.

70.3 Mean Value Interpretation

The mean value at this node is calculated to be 4032.006. This figure represents the average of the dependent variable for the observations falling into node 1, providing a baseline against which the deviations of individual observations can be measured.

70.4 MSE (Mean Squared Error) Analysis

The Mean Squared Error at node number 1 is quantified as 7.625437×10^7. This substantial MSE value indicates the variability of the observations from the mean, suggesting a degree of heterogeneity within this node. A high MSE may imply that the node contains a diverse set of data points with varying characteristics.

70.5 Implications for the Model

The statistics associated with node number 1 have significant implications for the regression tree model. The large number of observations and the high MSE value suggest that further splits or refinements might be necessary to enhance the model's accuracy. Understanding the characteristics of this node is crucial for improving the overall predictive performance of the regression tree model.
Primary divisions:

Table 36: First Regression Tree

REVOLKED splits as LR,	improve=0.174617900,(0$missing$)
MVR_PTS $<$ 3.5 to the left,	improve=0.063791060,(0$missing$)
AREA splits as LR,	improve=0.023841340,(0$missing$)
CAR_TYPE splits as LLLRLR,	improve=0.002160919,(0$missing$)
AGE $<$ 37.5 to the right,	improve=0.002044679, (0$missing$)
Surrogate splits:	
MVR_PTS $<$ 12.5to the left,	agree=0.878, adj=0.002, (0$split$)

71 Analysis of Node Number 2 in Regression Tree

71.1 Introduction to Node Number 2

Node number 2 in the regression tree model is a pivotal component, representing a distinct and specific subset of the dataset, with a total of 9,036 observations.

71.2 Analysis of Observations

This nodes dataset, encompassing 9,036 observations, signifies a particular segment of the total dataset. This segment is characterized by certain features or conditions that have led to its formation at this stage in the tree.

71.3 Mean Value Analysis

The mean value at node number 2 is computed to be 2669.389. This figure represents the average of the dependent variable for the subset of data at this node, offering an insight into the central tendency of this segment of the dataset.

71.4 MSE (Mean Squared Error) Interpretation

The Mean Squared Error at node number 2 is quantified as 3.397663×10^7. This value indicates the extent of variability or dispersion of the observations from the mean at this node, reflecting on the prediction error within this subset of the dataset.

Table 37: second Regression Tree

MVR_PTS $<$ 3.5 to the left,	improve=0.067190920, (0$missing$)
AREA splits as LR,	improve=0.024991830, (0$missing$)
CAR_TYPE splits as RLLRRR,	improve=0.002571936, (0$missing$)
MARRIED splits as RL,	improve=0.001741520, (0$missing$)
JOBCLASS splits as RRLLLLRRR,	improve=0.001321307, (0$missing$)
Surrogate splits: AGE$<$ 20.5to the right,	agree=0.812, adj=0.004, (0$split$)

72 Analysis of Node Number 3 in Regression Tree

72.1 Introduction to Node Number 3

Node number 3 in the regression tree model represents a distinct subset of the dataset, with a focus on a specific group of 1,260 observations.

72.2 Analysis of Observations

The 1,260 observations at node number 3 likely represent a unique segment of the dataset with specific characteristics or conditions. This nodes position in the tree suggests its significance in the models structure and decision-making process.

72.3 Mean Value Analysis

At node number 3, the mean value is calculated as 13803.92. This average indicates the central tendency of the dependent variable within this particular subset, providing a key insight into the data's overall behavior and pattern at this stage in the tree.

72.4 MSE (Mean Squared Error) Interpretation

The Mean Squared Error for node number 3 stands at 2.706405×10^8. This substantial MSE suggests a considerable degree of variability or dispersion in the data points around the mean, indicating the heterogeneity of the observations within this node.

72.5 Model Implications and Insights

The characteristics of node number 3, especially its mean and MSE, offer critical insights into the regression trees modeling dynamics. Understanding these statistics is vital for assessing the nodes role in the trees predictive accuracy and for gaining a deeper understanding of this specific data segment's influence on the overall model.

Table 38: third regression tree:

MVR_PTS $<$ 3.5 to the left,	improve=0.118274000,
AREA splits as LR,	improve=0.017072360,
CAR_TYPE splits as LLLRLL,	improve=0.016294430,
JOBCLASS splits as RRRLRLLLR,	improve=0.008256360,
AGE $<$ 60.5to the left,	improve=0.004325299,
Surrogate splits: AGE$<$ 68 to the left, agree=0.763, adj=0.007, (0*split*)	

72.6 Node Statistical Summary

The regression tree model comprises various nodes, each representing a unique subset of the dataset. The following is a concise summary of the statistical data for nodes 4, 5, 6, and 7:

72.7 Node 4

Observations: 7,328
Mean Value: 1,939.937
MSE (Mean Squared Error): 2.58523×10^7
This node represents a significant segment of the dataset, characterized by a moderate mean value and variability.

72.8 Node 5

Observations: 1,708
Mean Value: 5,799.028
MSE: 5.67557×10^7
Node 5 exhibits a higher degree of variance compared to Node 4, with a substantial mean value.

72.9 Node 6

Observations: 959
Mean Value: 10,634.24
MSE: 2.434114×10^8

This node, with fewer observations, shows a high mean value and considerable variability, indicating a more specific but variable data subset.

72.10 Node 7

Observations: 301
Mean Value: 23,902.65
MSE: 2.233994×10^8
Node 7, with the smallest number of observations, exhibits the highest mean value and a substantial MSE, reflecting its specificity and variability.

72.11 Conclusion

These nodes, with their varying mean values and MSEs, demonstrate the diversity and complexity within the regression tree model. Each node contributes uniquely, reflecting the varying characteristics of different data segments.

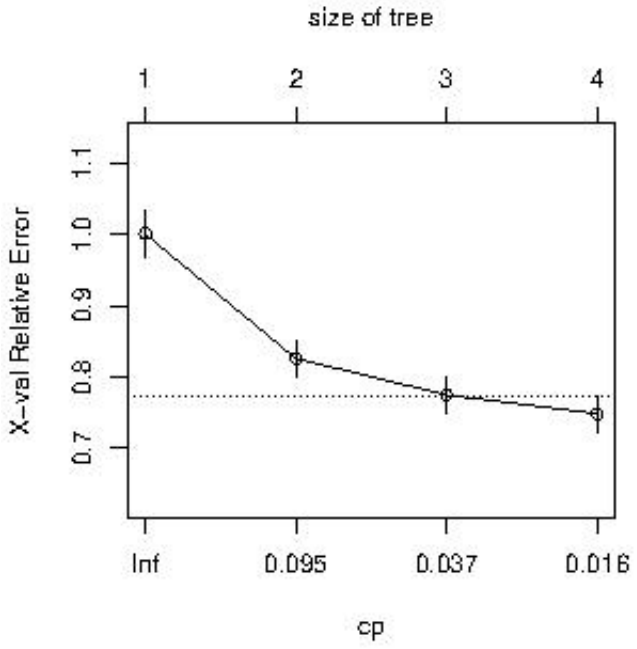

Figure 25: The relative effects of the most appropriate tree

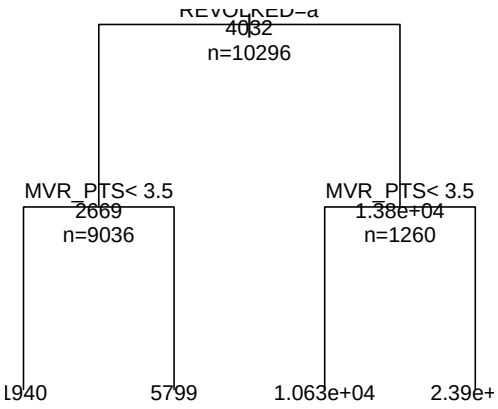

Figure 26: Regression Tree

In light of the evolving methodologies and the adaptations related to regression tree algorithms, the ensuing section delineates these developments:

Table 39: Deviance Residuals

Min	1Q.	Median	3Q.	Max.
−19185	−2700	−1225	778	52537

Table 40: glmboost

	Estimate	Std. Error	z value	Pr(>|z|)
Coefficients:				
(Intercept)	$-5.181e+02$	$8.975e+02$	-0.577	0.56377
AGE	$8.305e+00$	$1.019e+01$	0.815	0.41523
BLUEBOOK	$-6.753e-03$	$7.661e+01$	-0.086	0.93165
MARRIEDYes	$-3.261e+02$	$1.550e+02$	-2.104	0.03540 *
AREAUrban	$2.092e+03$	$2.041e+02$	10.253	$< 2e-16$ ***
MVR_PTS	$9.015e+02$	$3.574e+01$	25.222	$< 2e-16$***
NPOLICY	$-7.431e+01$	$8.038e+01$	-0.925	0.35520
RETAINED	$5.796e+00$	$1.828e+01$	0.317	0.75115
TRAVTIME	$-1.484e+00$	$4.802e+00$	-0.309	0.75731
CAR_USECommercial	$2.905e+02$	$2.454e+02$	1.184	0.23651
CAR_TYPEPickup	$9.505e+01$	$4.062e+02$	0.234	0.81500
CAR_TYPESedan	$3.071e+02$	$4.126e+02$	0.744	0.45669
CAR_TYPESports Car	$1.553e+03$	$5.391e+02$	2.881	0.00398**
CAR_TYPESUV	$5.950e+02$	$5.054e+02$	1.177	0.23911
CAR_TYPEVan	$6.627e+02$	$3.925e+02$	1.688	0.09137.
GENDERM	$3.242e+02$	$2.422e+02$	1.339	0.18062
JOBCLASSBlue Collar	$-5.433e+02$	$4.756e+02$	-1.142	0.25335
JOBCLASSClerical	$-1.649e+02$	$4.973e+02$	-0.332	0.74018
JOBCLASSDoctor	$-1.129e+03$	$6.041e+02$	-1.870	0.06157 .
JOBCLASSHome Maker	$-2.898e+01$	$5.022e+02$	-0.058	0.95398
JOBCLASSLawyer	$-1.415e+03$	$4.374e+02$	-3.236	0.00121**
JOBCLASSManager	$-7.366e+02$	$4.245e+02$	-1.735	0.08275.
JOBCLASSProfessional	$-5.789e+02$	$4.565e+02$	-1.268	0.20478
JOBCLASSStudent	$-3.955e+02$	$5.227e+02$	-0.757	0.44929
MAX_EDUCBachelors	$-1.412e+02$	$2.847e+02$	-0.496	0.61999
MAX_EDUCHigh School	$-1.070e+02$	$2.551e+02$	-0.419	0.67495
MAX_EDUCMasters	$-1.714e+01$	$3.997e+02$	-0.043	0.96579
MAX_EDUCPhD	$-3.563e+02$	$4.717e+02$	-0.755	0.45008
REVOLKEDYes	$1.058e+04$	$2.307e+02$	45.848	$< 2e-16$ ***

Summary Analysis:
Coefficients: The table lists the estimated coefficients for various predictors in the model, including demographic factors like age, marital status, and job

class, as well as vehicle-related variables like car type and usage. The intercept and each predictor have an associated coefficient.

Standard Error: The standard error for each coefficient indicates the level of uncertainty or variability in these estimates. A larger standard error suggests less confidence in the precision of the coefficient estimate.

Z-value: This statistic measures the number of standard deviations each coefficient is away from zero. A higher absolute value of the z-value typically indicates a more significant predictor.

P-value (Pr(>|z|)):
The p-values help in determining the statistical significance of each coefficient. A lower p-value (commonly below 0.05) suggests that the predictor is statistically significant.
For instance, variables like AREA Urban, MVR_PTS , CAR_TYPE Sports Car, JOBCLASS Lawyer, and REVOLKEDYes show high significance with very low p-values.

Significance Indicators:
Asterisks next to the p-values indicate the level of significance, with more asterisks denoting higher significance. For example, three asterisks ('***') indicate a very high level of statistical significance.

Interpretation:

Positive coefficients (e.g., AGE, AREAUrban) suggest that as the predictor increases, the response variable also tends to increase, keeping other factors constant. Negative coefficients (e.g., MARRIEDYes, JOBCLASS Lawyer) imply an inverse relationship with the response variable.
The coefficients with high significance (low p-values) are likely to be more reliable predictors in the model.
Overall Model Insights:
This table provides valuable insights into which factors are most influential in predicting the response variable. The significant predictors identified here could be crucial in understanding the underlying patterns and making informed decisions based on the model.

Signif. codes:0***	0.001 **	0.01 *	0.05.	0.1	1

and:

Residual standard error:	7602 on10266 degrees of freedom
Multiple R-squared:	0.2444, Adjusted R-squared: 0.2423
F-statistic:	114.5 on 29and 10266 DF, p-value:$< 2.2e-16$

72.12 Loss Function in GLM

The foundational component of this estimation process is the loss function, conventionally represented as $(y - f)^2$, where y is the observed value and f is the predicted value from the model. This function, known as the Mean Squared Error (MSE), quantifies the discrepancy between the predicted and actual values, guiding the optimization of the model coefficients.

72.13 Coefficient Estimation Using Tree-Based Methods

In a GLM, coefficients are typically estimated by minimizing the loss function. When integrating tree-based methods, the approach involves segmenting the data into various subsets based on predictor variables, allowing for more nuanced modeling of complex patterns.

72.14 Tree-Based Model Structure

Tree-based models, such as decision trees or boosted trees, divide the dataset into branches and leaves based on certain splitting criteria. Each leaf or segment thus represents a specific subset of the data with its own estimated response, providing a piecewise approximation of the underlying relationship.

72.15 Integration with GLM

Integrating tree-based methods with GLM involves using the structure of the trees to inform the estimation of GLM coefficients. This process can either

involve using the trees to directly estimate the coefficients or to transform the input variables in a way that better captures non-linearities, which are then used in a traditional GLM framework.

73 Advantages and Considerations

The primary advantage of this approach is its ability to model complex and non-linear relationships more effectively than standard GLM. However, it requires careful consideration of model complexity to avoid over fitting, as well as a thorough understanding of the interactions between variables represented in the tree structure.

73.1 Conclusion

The integration of tree-based methods in the coefficient estimation of GLMs offers a powerful tool for advanced statistical modeling. By leveraging the strengths of both GLM and tree-based algorithms, this approach can yield more accurate and robust models, especially in scenarios where the relationships between variables are complex and non-linear.

Table 41: Estimation of Coefficients in the Generalized Linear Model Using Tree Regression Method

Coefficients:

(Intercept)	AGE	BLUEBOOK
−4092.88292	0.00000	0.00000
HOMEKIDS	MARRIEDYes	AREAUrban
0.00000	−85.48305	1697.40875
MVR_PTS	NPOLICY	RETAINED
878.58649	0.00000	0.00000
TRAVTIME	CAR_USECommercial	CAR_TYPEPickup
0.00000	0.00000	0.00000
CAR_TYPESedan	CAR_TYPESports Car	CAR_TYPESUV
0.00000	693.44934	0.00000
CAR_TYPEVan	GENDERM	JOBCLASSBlue Collar
86.47983	0.00000	0.00000
JOBCLASSClerical	JOBCLASSDoctor	JOBCLASSHome Maker
0.00000	−232.95470	0.00000
JOBCLASSLawyer	JOBCLASSManager	JOBCLASSProfessional
−542.31605	0.00000	0.00000
JOBCLASSStudent	MAX_EDUCBachelors	MAX_EDUCHigh School
0.00000	0.00000	0.00000
MAX_EDUCMasters	MAX_EDUCPhD	REVOLKEDYes
0.00000	0.00000	10307.93027

Also:

Selection frequencies::

REVOLKEDYes	MVR_PTS	AREAUrban	CAR_TYPESports Car	
0.30	0.24	0.18		0.11
JOBCLASSLawyer	MARRIEDYes	JOBCLASSDoctor		CAR_TYPEVa
0.09	0.03	0.03		0.02

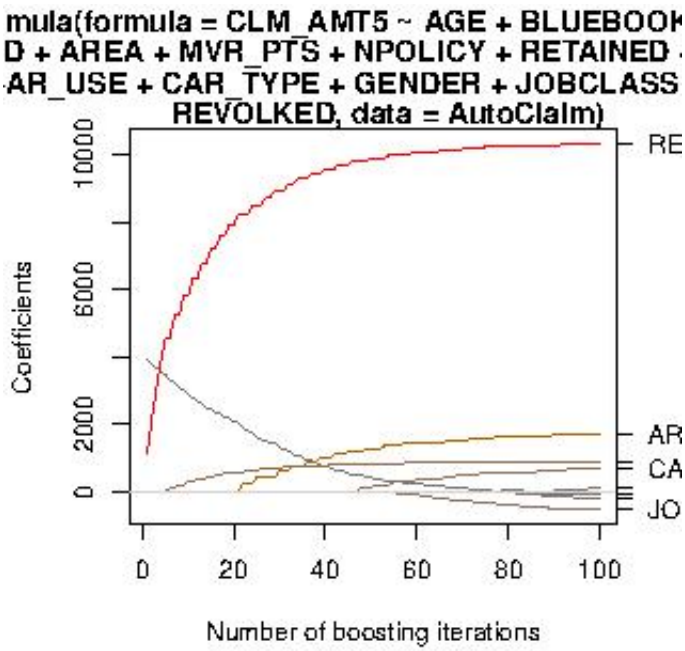

Figure 27: Generalized Linear Model with Tree Method

74 GLM method:

Definition of Residuals: In a GLM, residuals are the differences between the observed data points and the values predicted by the model. They are crucial for diagnosing the model and understanding how well it fits the data.

Types of Residuals in GLM: There are several types of residuals used in the context of GLMs, such as Pearson residuals, deviance residuals, and standardized residuals. Each type provides different insights into the model's fit.

Pearson Residuals: These are the differences between observed and expected outcomes divided by the standard deviation of the expected outcome. They are useful for identifying outliers and data points that the model does not adequately explain.

Deviance Residuals: They are a measure of how much each observation contributes to the overall deviance of the model. The deviance is a generalization

of the residual sum of squares in linear models and provides a way of assessing the goodness of fit of the model.

Analyzing Residual Deviation: A key part of GLM analysis is examining the pattern of residuals. Ideally, residuals should be randomly distributed with no apparent pattern. Patterns or trends in the residuals may indicate issues with the model, such as incorrect specification or the presence of influential outliers.

Graphical Representation: Plotting residuals against fitted values or other variables can help in identifying non-random patterns, suggesting potential modifications or improvements to the model.

Model Improvement: Based on residual analysis, one might consider transformations of variables, adding interaction terms, or using a different link function to improve the model's fit.

Table 42: Deviation of Residuals in the GLM Method

Min	1Q.	Median	3 Q.	Max.
−19185	−2700	−1225	778	52537

Residual deviance: $5.9320e+11$ on10266 degrees of freedom
Null deviance:$7.8512e+11$ on10295 degrees of freedom
AIC:213263
Number of Fisher Scoring iterations:2
So:
Residual Deviation Analysis:
The residual deviation across $10,266$ data points is approximately 5.9320
This metric is critical in understanding how well the model predicts or fits the data. A lower residual deviation typically indicates a better model fit.

Zero Deviation Model:
The deviation for a model with zero deviation is calculated to be around 7.8512
for 10,295 data points. This measure is important for models that assume a zero-inflation or zero-altered process, especially relevant in count data with excess zeros.

Akaike Information Criterion (AIC):
The AIC for the model is found to be $213,263$.
AIC is a widely used criterion for model selection, balancing model complexity and goodness of fit. The lower the AIC, the better the model balances these two aspects.

Fishers Method for Iterative Refinement:
The number of iterations in Fisher's method is noted to be2. This iterative approach is essential for refining the model to reach an optimal solution, especially in complex models where a direct analytical solution is not feasible.

Table 43: Estimation of Model Coefficients in Generalized Linear Model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	$-5.181e+02$	$8.975e+02$	-0.577	0.56377
AGE	$8.305e+00$	$1.019e+01$	0.815	0.41523
BLUEBOOK	$-6.753e-03$	$1.249e-02$	-0.541	0.58869
HOMEKIDS	$-6.571e+00$	$7.661e+01$	-0.086	0.93165
MARRIEDYes	$-3.261e+02$	$1.550e+02$	-2.104	0.03540*
AREAUrban	$2.092e+03$	$2.041e+02$	10.253	$< 2e-16$ ***
MVR_PTS	$9.015e+02$	$3.574e+01$	25.222	$<2e-16$ ***
NPOLICY	$-7.431e+01$	$8.038e+01$	-0.925	0.35520
RETAINED	$5.796e+00$	$1.828e+01$	0.317	0.75115
TRAVTIME	$-1.484e+00$	$4.802e+00$	-0.309	0.75731
CAR_USECommercial	$2.905e+02$	$2.454e+02$	1.184	0.23651
CAR_TYPEPickup	$9.505e+01$	$4.062e+02$	0.234	0.81500
CAR_TYPESedan	$3.071e+02$	$4.126e+02$	0.744	0.45669
CAR_TYPESports Car	$1.553e+03$	$5.391e+02$	2.881	0.00398 **
CAR_TYPESUV	$5.950e+02$	$5.054e+02$	1.177	0.23911
CAR_TYPEVan	$6.627e+02$	$3.925e+02$	1.688	0.09137 .
GENDERM	$3.242e+02$	$2.422e+02$	1.339	0.18062
JOBCLASSBlue Collar	$-5.433e+02$	$4.756e+02$	-1.142	0.25335
JOBCLASSClerical	$-1.649e+02$	$4.973e+02$	-0.332	0.74018
JOBCLASSDoctor	$-1.129e+03$	$6.041e+02$	-1.870	0.06157 .
JOBCLASSHome Maker	$-2.898e+01$	$5.022e+02$	-0.058	0.95398
JOBCLASSLawyer	$-1.415e+03$	$4.374e+02$	-3.236	0.00121**
JOBCLASSManager	$-7.366e+02$	$4.245e+02$	-1.735	0.08275 .
JOBCLASSProfessional	$-5.789e+02$	$4.565e+02$	-1.268	0.20478
JOBCLASSStudent	$-3.955e+02$	$5.227e+02$	-0.757	0.44929
MAX_EDUCBachelors	$-1.412e+02$	$2.847e+02$	-0.496	0.61999
MAX_EDUCHigh School	$-1.070e+02$	$2.551e+02$	-0.419	0.67495
MAX_EDUCMasters	$-1.714e+01$	$3.997e+02$	-0.043	0.96579
MAX_EDUCPhD	$-3.563e+02$	$4.717e+02$	-0.755	0.45008
REVOLKEDYes	$1.058e+04$	$2.307e+02$	45.848	$<2e-16$***

Signif. codes:0***	0.001 **	0.01 *	0.05.	0.1	1

Residual deviance:$5.9320e+11$ on10266 degrees of freedom
Null deviance::$7.8512e+11$ on10295 degrees of freedom
AIC: 213263
Number of Fisher Scoring iterations: 2

Table 44: Coefficients Obtained from Variables in the Generalized Linear Model 95%

	2.5	97.5
(Intercept)	$-2.277251e+03$	$1.241007e+03$
AGE	$-1.167338e+01$	$2.828361e+01$
BLUEBOOK	$-3.122822e-02$	$1.772281e-02$
HOMEKIDS	$-1.567309e+02$	$1.435883e+02$
MARRIEDYes	$-6.299280e+02$	$-2.232400e+01$
AREAUrban	$1.692288e+03$	$2.492198e+03$
MVR_PTS	$8.314406e+02$	$9.715496e+02$
NPOLICY	$-2.318504e+02$	$8.322051e+01$
RETAINED	$-3.002659e+01$	$4.161901e+01$
TRAVTIME	$-1.089507e+01$	$7.927325e+00$
CAR_USECommercial	$-1.904493e+02$	$7.714085e+02$
CAR_TYPEPickup	$-7.011163e+02$	$8.912146e+02$
CAR_TYPESedan	$-5.015767e+02$	$1.115807e+03$
CAR_TYPESports Car	$4.963960e+02$	$2.609715e+03$
CAR_TYPESUV	$-3.955695e+02$	$1.585559e+03$
CAR_TYPEVan	$-1.066077e+02$	$1.431977e+03$
GENDERM	$-1.503847e+02$	$7.988532e+02$
JOBCLASSBlue Collar	$-1.475385e+03$	$3.888529e+02$
JOBCLASSClerical	$-1.139692e+03$	$8.098359e+02$
JOBCLASSDoctor	$-2.313496e+03$	$5.458011e+01$
JOBCLASSHome Maker	$-1.013228e+03$	$9.552622e+02$
JOBCLASSLawyer	$-2.272653e+03$	$-5.582682e+02$
JOBCLASSManager	$-1.568726e+03$	$9.544657e+01$
JOBCLASSProfessional	$-1.473618e+03$	$3.158275e+02$
JOBCLASSStudent	$-1.419858e+03$	$6.289425e+02$
MAX_EDUCBachelors	$-6.991285e+02$	$4.167966e+02$
MAX_EDUCHigh School	$-6.068876e+02$	$3.929534e+02$
MAX_EDUCMasters	$-8.004987e+02$	$7.662110e+02$
MAX_EDUCPhD	$-1.280852e+03$	$5.682603e+02$
REVOLKEDYes	$1.012701e+04$	$1.103152e+04$

Table 45: Coefficients with the Help of Exponential Link Function

(Intercept)	AGE	BLUEBOOK
$9.608095e-226$	$4.044510e+03$	$9.932700e-01$
HOMEKIDS	MARRIEDYes	AREAUrban
$1.399975e-03$	$2.318862e-142$	Inf
MVR_PTS	NPOLICY RETAINED	
Inf	$5.314224e-33$	$3.290489e+02$
TRAVTIME	CAR_USECommercial	CAR_TYPEPickup
$2.267584e-01$	$1.424612e+126$	$1.902434e+41$
CAR_TYPESedan	CAR_TYPESports Car	CAR_TYPESUV
$2.390093e+133$	Inf	$2.529148e+258$
CAR_TYPEVan	GENDERM	JOBCLASSBlue Collar
$6.313648e+287$	$6.503688e+140$	$1.154816e-236$
JOBCLASSClerical	JOBCLASSDoctor	JOBCLASSHome Maker
$2.358157e-72$	$0.000000e+00$	$2.587072e-13$
JOBCLASSLawyer	JOBCLASSManager	JOBCLASSProfessional
$0.000000e+00$	$1.206508e-320$	$3.881174e-252$
JOBCLASSStudent	MAX_EDUCBachelors	MAX_EDUCHigh School
$1.798740e-172$	$4.925140e-62$	$3.505833e-47$
MAX_EDUCMasters	MAX_EDUCPhD	REVOLKEDYes
$3.585227e-08$	$1.831344e-155$	Inf

Table 46: Exponential Coefficients of Variables in the Generalized Linear Model

	2.5	97.5
(Intercept)	$0.000000e+00$	Inf
AGE	$8.517594e-06$	$1.920502e+12$
BLUEBOOK	$9.692543e-01$	$1.017881e+00$
HOMEKIDS	$8.563399e-69$	$2.288728e+62$
MARRIEDYes	$2.665273e-274$	$2.017474e-10$
AREAUrban	Inf	Inf
MVR_PTS	Inf	Inf
NPOLICY	$2.035502e-101$	$1.387421e+36$
RETAINED	$9.112060e-14$	$1.188241e+18$
TRAVTIME	$1.854954e-05$	$2.772002e+03$
CAR_USECommercial	$1.945052e-83$	Inf
CAR_TYPEPickup	$3.228894e-305$	Inf
CAR_TYPESedan	$1.472363e-218$	Inf
CAR_TYPESports Car	$3.819858e+215$	Inf
CAR_TYPESUV	$1.608156e-172$	Inf
CAR_TYPEVan	$5.022072e-47$	Inf
GENDERM	$4.883860e-66$	Inf
JOBCLASSBlue Collar	$0.000000e+00$	$7.527824e+168$
JOBCLASSClerical	$0.000000e+00$	Inf
JOBCLASSDoctor	$0.000000e+00$	$5.056373e+23$
JOBCLASSHome Maker	$0.000000e+00$	Inf
JOBCLASSLawyer	$0.000000e+00$	$3.525375e-243$
JOBCLASSManager	$0.000000e+00$	$2.830852e+41$
JOBCLASSProfessional	$0.000000e+00$	$1.452609e+137$
JOBCLASSStudent	$0.000000e+00$	$1.400404e+273$
MAX_EDUCBachelors	$2.356986e-304$	$1.029154e+181$
MAX_EDUCHigh School	$2.704441e-264$	$4.544696e+170$
MAX_EDUCMasters	$0.000000e+00$	Inf
MAX_EDUCPhD	$0.000000e+00$	$6.198729e+246$
REVOLKEDYes	Inf	Inf

74.1 Gini Coefficient in Generalized Linear Models:

Specifically, in the model labeled 'P1', the glm package is used to estimate the model parameters and subsequently calculate the Gini coefficient. This coefficient then provides insights into the inequality or dispersion of the insurance claims as predicted by the model:

Table 47: The first model with glm method

Min	1Q	Median	3Q	Max
−3.588	−2.084	−1.391	0.634	106.195

Table 48: Estimation of the coefficient in the first model with glm method

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.39085	0.09528	14.6	$< 2e-16$ ***

Signif. codes:	0***	0.001**	0.01 *	0.05.	0.1	1

75 Model Summary

Null deviance:	71764 on 2811 degrees of freedom
Residual deviance:	71764 on 2811 degrees of freedom
AIC:	17094
Number of Fisher Scoring iterations:	2

Table 49: Summary of the Model

Now Start with an Introduction to the Model: Introduce Model 2 (P2) and its significance in the context of your analysis. Mention the variables that the model takes into account.

Discuss the Estimation of Coefficients: Explain how the coefficients in your GLM are estimated. This might involve discussing the fitting process, the link function used, and how each variable contributes to the model.

Present the Results: Include a table or a textual description of the estimated coefficients and what they imply about the relationships between the dependent variable and each explanatory variable.

Table 50: Coefficient estimation in the second model with glm method

Min	1Q	Median	3Q	Max
−4.936	−2.056	−1.039	0.545	105.398
	Estimate	Std. Error	t value	Pr(>|t|)
Intercept)	−0.9653	0.5550	−1.739	0.082059.
factor(CAR_USE)Commercial	1.0478	0.2289	4.578	$4.89e-06$***
factor(REVOLKED)Yes	0.5835	0.2736	2.133	0.032998 *
factor(GENDER)M	0.2084	0.2667	0.781	0.434699
factor(AREA)Urban	1.9494	0.2303	8.465	$< 2e-16$ ***
factor(MARRIED)Yes	−0.7338	0.1919	−3.824	0.000134***
factor(CAR_TYPE)Pickup	0.7477	0.4582	1.632	0.102804
factor(CAR_TYPE)Sedan	0.1017	0.4606	0.221	0.825197
factor(CAR_TYPE)Sports Car	1.5391	0.5562	2.767	0.005696 **
factor(CAR_TYPE)SUV	1.2219	0.5149	2.373	0.017719*
factor(CAR_TYPE)Van	0.6819	0.5190	1.314	0.188958

Signif. codes:	0***	0.001**	0.01 *	0.05.	0.1	1

Null deviance:	71764 on2811 degrees of freedom
Residual deviance:	68458 on2801 degrees of freedom
AIC:	16981
Number of Fisher Scoring iterations:	2

$P3$:

Min	1Q	Median	3Q	Max
-6.393	-1.947	-0.842	0.529	104.019

Table 51: Your Table Caption

Table 52: Coefficient estimation in the third model with glm method

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	$-1.175e+007$	$6.247e-01$	-1.882	0.060002 .
factor(CAR_USE)Commercial	$8.637e-01$	$2.283e-01$	3.784	0.000158 ***
factor(REVOLKED)Yes	$4.876e-01$	$2.713e-01$	1.797	0.072419 .
factor(GENDER)M	$2.187e-01$	$2.644e-01$	0.827	0.408233
factor(AREA)Urban	$1.929e+00$	$2.396e-01$	8.053	$1.19e-15$ ***
factor(MARRIED)Yes	$-7.784e-01$	$1.903e-01$	-4.090	$4.44e-05$ ***
factor(CAR_TYPE)Pickup	$4.370e-01$	$4.597e-01$	0.951	0.341870
factor(CAR_TYPE)Sedan	$-1.090e-01$	$4.596e-01$	-0.237	0.812518
factor(CAR_TYPE)Sports Car	$1.083e+00$	$5.592e-01$	1.936	0.052988.
factor(CAR_TYPE)SUV	$8.473e-01$	$5.177e-01$	1.637	0.101802
factor(CAR_TYPE)Van	$4.780e-01$	$5.149e-01$	0.928	0.353278
TRAVTIME	$1.681e-02$	$5.962e-03$	2.820	0.004841 **
MVR_PTS	$2.566e-01$	$4.385e-02$	5.850	$5.48e-09$ ***
INCOME	$-6.974e-06$	$2.142e-06$	-3.256	0.001142**

Signif. codes:	0***	0.001**	0.01 *	0.05.	0.1	1

Null deviance:	71764 on 2811 degrees of freedom
Residual deviance:	67082 on2798degrees of freedom
AIC:	16930
Number of Fisher Scoring iterations:	2

Assessment of Income Inequality:

Calculating the Gini Coefficient in the Context of Claim Loss Functions: and: The following graph has been plotted:

Table 53: Gini Indices in Generalized Linear Models.

	P1	P2	P3
P1	0.000	37.126	41.335
P2	−12.879	0.000	20.959
P3	−12.183	−3.446	0.000

Table 54: Standard Errors in Generalized Linear Models:

	P1	P2	P3
P1	0.000	2.457	2.588
P2	2.952	0.000	2.951
P3	3.418	3.103	0.000

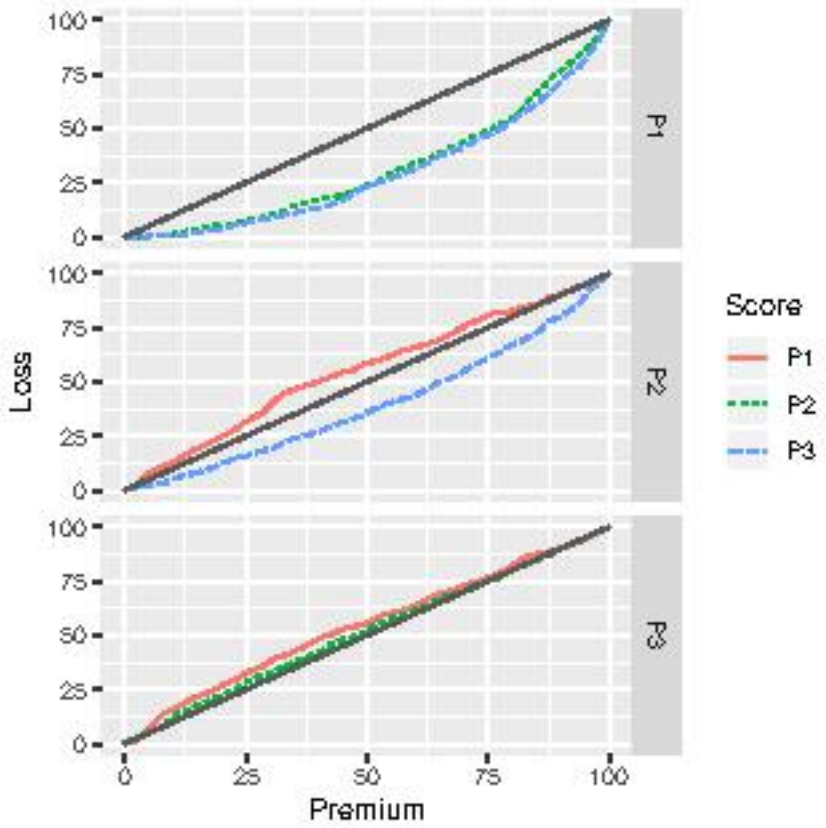

Figure 28: Gini Coefficient Chart in the Generalized Linear Model

And also:

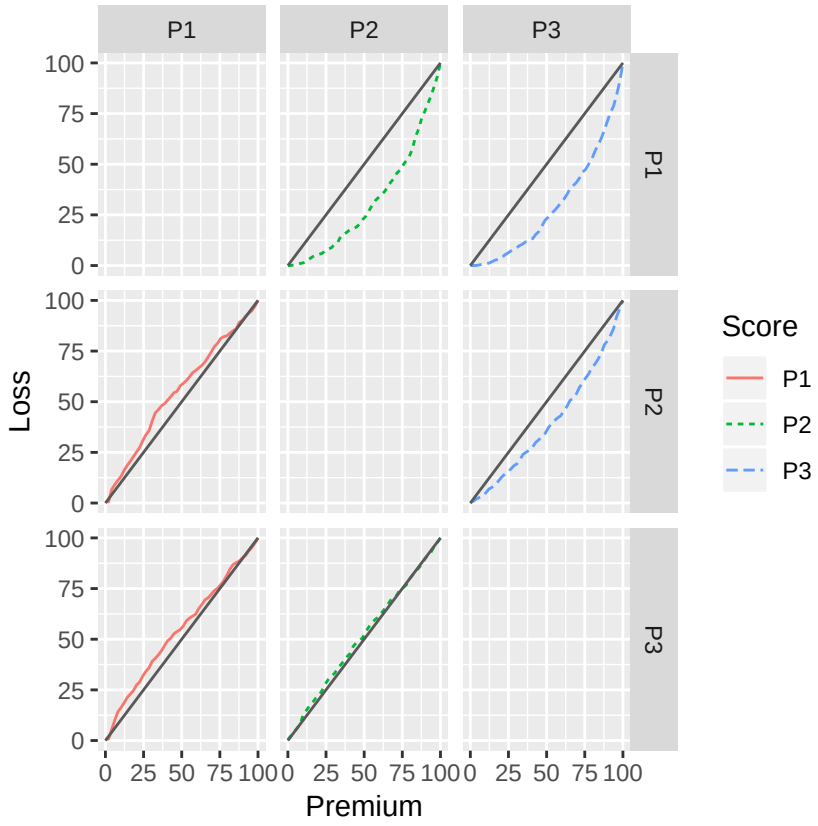

Figure 29: Another graph of the Gini coefficient in the generalized linear model

76 Poisson Regression Analysis

Poisson regression is a fundamental approach in statistical modeling, especially useful for count data or in scenarios where the data represent counts of events. This type of regression falls under the umbrella of Generalized Linear Models (GLMs) and is particularly adept at handling data where the response variable is a count or a rate.

76.1 Coefficient Evaluation in Poisson Regression

In the context of the Poisson regression model, which is part of the Poisson family of distributions, the evaluation of coefficients is a key step. The

coefficients in this model are estimated to explain the relationship between the independent variables and the logarithm of the expected count of the dependent variable. The process typically involves:

- Assuming that the response variable follows a Poisson distribution.
- Utilizing a logarithmic link function to connect the linear predictor to the expected value of the response variable.
- Estimating the coefficients through maximum likelihood estimation, which seeks to find the values that maximize the likelihood of observing the sample data.

The coefficients in a Poisson regression model provide insights into the rate at which the count of the dependent variable changes as the independent variables change. A positive coefficient suggests an increasing trend, whereas a negative coefficient indicates a decreasing relationship.

Table 55: Residual Deviation in Poisson Regression

Min	1Q.	Median	3 Q.	Max.
−279.80	−68.57	−49.92	10.99	557.71

Table 56: Coefficient estimation in Poisson regression model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	$6.704e+00$	$1.909e-03$	3511.642	$< 2e-16$ ***
AGE	$2.606e-03$	$1.995e-05$	130.620	$< 2e-16$ ***
BLUEBOOK	$-2.754e-06$	$2.681e-08$	-102.700	$< 2e-16$ ***
HOMEKIDS	$4.369e-03$	$1.525e-04$	28.639	$< 2e-16$***
MARRIEDYes	$-1.084e-01$	$3.171e-04$	-341.993	$< 2e-16$***
AREAUrban	$9.629e-01$	$6.423e-04$	1499.162	$< 2e-16$***
MVR_PTS	$1.610e-01$	$5.787e-05$	2781.890	$< 2e-16$***
NPOLICY	$-1.332e-02$	$1.769e-04$	-75.318	$< 2e-16$ ***
RETAINED	$1.684e-03$	$3.886e-05$	43.338	$< 2e-16$***
TRAVTIME	$9.098e-05$	$1.014e-05$	8.976	$< 2e-16$ ***
CAR_USECommercial	$5.340e-02$	$5.022e-04$	106.328	$< 2e-16$***
CAR_TYPEPickup	$2.095e-03$	$8.548e-04$	2.451	0.0143*
CAR_TYPESedan	$3.916e-02$	$8.735e-04$	44.837	$< 2e-16$ ***
CAR_TYPESports Car	$3.215e-01$	$1.120e-03$	287.166	$< 2e-16$***
CAR_TYPESUV	$1.306e-01$	$1.074e-03$	121.615	$< 2e-16$***
CAR_TYPEVan	$1.373e-01$	$7.998e-04$	171.657	$< 2e-16$ ***
GENDERM	$6.637e-02$	$5.263e-04$	126.122	$< 2e-16$ ***
JOBCLASSBlue Collar	$-1.537e-01$	$9.629e-04$	-159.660	$< 2e-16$***
JOBCLASSClerical	$-3.876e-02$	$1.003e-03$	-38.633	$< 2e-16$ ***
JOBCLASSDoctor	$-3.509e-01$	$1.336e-03$	-262.561	$< 2e-16$ ***
JOBCLASSHome Maker	$-8.472e-03$	$1.024e-03$	-8.272	$< 2e-16$ ***
JOBCLASSLawyer	$-3.563e-01$	$8.932e-04$	-398.933	$< 2e-16$***
JOBCLASSManager	$-1.733e-01$	$8.553e-04$	-202.560	$< 2e-16$***
JOBCLASSProfessional	$-1.600e-01$	$9.206e-04$	-173.751	$< 2e-16$ ***
JOBCLASSStudent	$-8.577e-02$	$1.053e-03$	-81.468	$< 2e-16$***
MAX_EDUCBachelors	$-2.156e-02$	$5.826e-04$	-37.016	$< 2e-16$ ***
MAX_EDUCHigh School	$-6.271e-03$	$5.201e-044$	-12.057	$< 2e-16$ ***
MAX_EDUCMasters	$5.396e-03$	$8.403e-04$	6.421	$1.35e-10$ ***
MAX_EDUCPhD	$-4.278e-02$	$9.990e-04$	-42.817	$< 2e-16$***
REVOLKEDYes	$1.505e+00$	$3.191e-04$	4718.142	$< 2e-16$ ***

Signif. codes:0***	0.001 **	0.01 *	0.05.	0.1	1

Residual deviance:	81529301 on10266 degrees of freedom
Null deviance::	117047095 on10295 degrees of freedom
AIC:	81571807
Number of Fisher Scoring iterations:	7

77 Analysis of Residual Deviation in Zero-Inflated Models

In the context of zero-inflated statistical models, particularly those pertaining to count data, the evaluation of residual deviations plays a critical role. This section delves into the analysis of such deviations within the framework of Poisson regression, specifically addressing models that accommodate zero accumulation.

Utilizing a set of fifteen predictors, the study employs the 'cpglm' command to facilitate the analysis. This command is instrumental in Poisson regression models, especially when addressing the nuances of zero-inflated data.

The primary focus here is on the deviation of residuals in cumulative models that exhibit a significant presence of zero values. These zero-inflated models require specialized analytical approaches to accurately assess and interpret the residual deviations, which provide insights into model fit and data characteristics.

77.1 Residual Deviation in Zero-Inflated Cumulative Models

The residual deviation in zero-inflated cumulative models is calculated as follows:

This analysis is crucial for understanding the behavior of residuals in models where zero values are prevalent, offering a deeper understanding of the underlying data distribution and model accuracy. therefore:

Table 57: Residual Deviation in Over-dispersed Model.

Min	1Q.	Median	3 Q.	Max.
−5.805	−2.267	−1.774	0.380	13.523

Also:

Table 58: Estimation of Coefficients in Models with Zero Inflation.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	$-1.004e-01$	$4.319e-01$	-0.232	0.8162
AGE	$5.869e-03$	$5.132e-03$	1.144	0.2528
BLUEBOOK	$-1.324e-05$	$6.564e-06$	-2.017	0.0438*
HOMEKIDS	$-2.599e-02$	$4.431e-02$	-0.586	0.5576
MARRIEDYes	$-8.679e-02$	$7.973e-02$	-1.089	0.2765
KIDSDRIV	$5.865e-02$	$7.691e-02$	0.763	0.4458
MVR_PTS	$1.964e-01$	$1.527e-02$	12.861	$< 2e-16$***
NPOLICY	$-1.352e-02$	$4.323e-02$	-0.313	0.7545
TRAVTIME	$1.965e-03$	$2.534e-03$	0.775	0.4382
AREAUrban	$9.006e-01$	$1.303e-01$	6.910	$6e-12$ ***
CAR_USECommercial	$-1.270e-02$	$1.224e-01$	-0.104	0.9174
CAR_TYPEPickup	$3.448e-02$	$2.336e-01$	0.148	0.8827
CAR_TYPESedan	$7.503e-02$	$2.334e-01$	0.321	0.7479
CAR_TYPESports Car	$1.258e-01$	$2.907e-01$	0.433	0.6653
CAR_TYPESUV	$-6.904e-02$	$2.776e-01$	-0.249	0.8036
CAR_TYPEVan	$2.979e-02$	$2.361e-01$	0.126	0.8996
GENDERM	$-1.852e-01$	$1.259e-01$	-1.471	0.1414
JOBCLASSClerical	$-2.107e-01$	$1.410e-01$	-1.495	0.1351
JOBCLASSDoctor	$-5.820e-01$	$3.287e-01$	-1.771	0.0767.
JOBCLASSHome Maker	$-3.470e-01$	$1.781e-01$	-1.948	0.0515.
JOBCLASSLawyer	$-5.168e-01$	$2.360e-01$	-2.189	0.0286 *
JOBCLASSManager	$-6.588e-02$	$1.691e-01$	-0.390	0.6968
JOBCLASSProfessional	$-3.644e-01$	$1.564e-01$	-2.330	0.0199 *
JOBCLASSStudent	$-7.072e-02$	$1.456e-01$	-0.486	0.6273
MAX_EDUCBachelors	$5.577e-02$	$1.389e-01$	0.401	0.6882
MAX_EDUCHigh School	$9.648e-02$	$1.281e-01$	0.753	0.4515
MAX_EDUCMasters	$7.016e-02$	$2.024e-01$	0.347	0.7289
MAX_EDUCPhD	$2.223e-01$	$2.487e-01$	0.894	0.3716
REVOLKEDYes	$1.419e+00$	$8.421e-02$	16.855	$< 2e-16$ ***

Signif. codes:0***	0.001 **	0.01 *	0.05.	0.1	1

77.2 Model Fit and Parameter Estimation

The analysis of zero-adjusted models, particularly focusing on the zero-adjusted Poisson model, reveals a significant fit to the data. The following are the key findings from the model estimation:
Estimated dispersion parameter:6.3011
Residual deviance: 16000 on 2783 degrees of freedom
Estimated index parameter: 1.3459
AIC: 10430
Number of Fisher Scoring iterations: 7
These results underscore the efficacy of zero-adjusted Poisson models in handling datasets characterized by over dispersion and a high prevalence of zero counts. Such models provide a more nuanced and accurate representation of the underlying data distribution compared to standard Poisson models.

Fitting variables in models with zero accumulation allows for more accurate predictions and insights, particularly in datasets where zero values are prevalent. This approach enhances the model's ability to represent the underlying data distribution effectively:

Table 59: Fitting Variables in Zero-Inflated Models

(Intercept)	AGE	BLUEBOOK
$-1.003896e-01$	$5.869352e-03$	$-1.324046e-05$
HOMEKIDS	MARRIEDYes	KIDSDRIV
$-2.598526e-02$	$-8.678879e-02$	$5.864935e-02$
MVR_PTS	NPOLICY	TRAVTIME
$1.964390e-01$	$-1.351721e-02$	$1.964571e-03$
AREAUrban	CAR_USECommercial	CAR_TYPEPickup
$9.006145e-01$	$-1.270410e-02$	$3.447991e-02$
CAR_TYPESedan	CAR_TYPESports Car	CAR_TYPESUV
$7.503423e-02$	$1.257868e-01$	$-6.903604e-02$
CAR_TYPEVan	GENDERM	JOBCLASSClerical
$2.978900e-02$	$-1.851635e-01$	$-2.107461e-01$
JOBCLASSDoctor	JOBCLASSHome Maker	JOBCLASSLawyer
$-5.819723e-01$	$-3.469687e-01$	$-5.167977e-01$
JOBCLASSManager	JOBCLASSProfessional	JOBCLASSStudent
$-6.587701e-02$	$-3.644448e-01$	$-7.071583e-02$
MAX_EDUCBachelors	MAX_EDUCHigh School	MAX_EDUCMasters
$5.576539e-02$	$9.648123e-02$	$7.016109e-02$
MAX_EDUCPhD	REVOLKEDYes	
$2.222647e-01$	$1.419333e+00$	

Gini Coefficient in Zero-Inflated Models

The analysis involves the calculation of the Gini coefficient in zero-inflated models, particularly focusing on insurance claims. The methodology is outlined as follows:

- **Zero-Inflated Probability Model:** The model is defined by the formula:

$$P(y_i = 0) = q_i + (1 - q_i) \exp\left(-\frac{\mu_i^{(2-p)}}{\Phi(2-p)}\right) \tag{100}$$

  where y_i denotes the response variable for the ith observation, q_i is the probability of the ith observation being from the zero-inflated part

of the model, μ_i represents the mean of the non-zero counts, Φ is a dispersion parameter, and p is a power parameter.

- **Gini Coefficient Calculation:** This involves considering a set of variables that influence insurance claims and calculating the Gini coefficient for these variables. The insurance claim is treated as a loss function, and the analysis takes into account the number of insurance policies and the nature of the claims.
- The calculation is facilitated by the `cplm` package, particularly in the model labeled as *P1*.

This approach is aimed at analyzing insurance claim data using a zero-inflated model and assessing the inequality in these claims using the Gini coefficient. It accounts for the high frequency of zero claims in insurance datasets.

Table 60: residual deviation in the first zero-inflated model

Min	1Q	Median	3Q	Max
−3.7278	−2.3362	−1.8835	0.2066	19.5842
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.15666	0.05916	2.648	0.00814**

Signif. codes:	0***	0.001**	0.01 *	0.05.	0.1	1

Estimated dispersion parameter:	6.4117
Estimated index parameter:	1.3786
Residual deviance:	15187 on2811 degrees of freedom
AIC:	8451.2
Number of Fisher Scoring iterations:	5

Analysis of Residual Deviations in the Second Model

In the progression of our statistical analysis, we turn our attention to the second model, incorporating a comprehensive set of explanatory variables. This model meticulously examines the impact of various factors such as the type of car, insurance policy history, gender, geographical location of traffic, marital status, and the primary use of the car.

Assessment of Residual Deviations

The focal point of this section is the deviation of residuals within the second model. This investigation is pivotal for the following reasons:

- **Comprehensive Variable Analysis:** The model integrates a diverse array of variables, providing a holistic view of the factors influencing insurance claims.
- **Model Accuracy and Precision:** The assessment of residual deviations is crucial in evaluating the model's accuracy. It allows us to understand the extent to which the model's predictions align with the actual observed data.
- **Insights into Model's Performance:** This analysis sheds light on the strengths and potential limitations of the model, offering pathways for refinement and enhancement.

The detailed analysis of residual deviations in this context is not just a numerical exercise but a strategic approach to enhancing the reliability and validity of predictive modeling in the domain of insurance data analytics.

Table 61: Residual Deviations in the Second Model with Zero Inflation

Min	1Q	Median	3Q	Max
−4.1831	−2.0806	−1.4971	−0.1391	18.7319
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	−2.0107	0.3776	−5.324	$1.09e-07$***
factor(CAR_USE)Commercial	0.5120	0.1329	3.851	0.000120 ***
factor(REVOLKED)Yes	0.2040	0.1486	1.373	0.169857
factor(GENDER)M	0.1206	0.1673	0.721	0.471019
factor(AREA)Urban	2.0035	0.2217	9.037	$< 2e-16$ ***
factor(MARRIED)Yes	−0.4126	0.1139	−3.622	0.000297 ***
factor(CAR_TYPE)Pickup	0.3629	0.2705	1.342	0.179729
factor(CAR_TYPE)Sedan	−0.1625	0.2820	−0.576	0.564623
factor(CAR_TYPE)Sports Car	0.7258	0.3314	2.190	0.028582*
factor(CAR_TYPE)SUV	0.5452	0.3085	1.768	0.077245.
factor(CAR_TYPE)Van	0.2363	0.3021	0.782	0.434083

Table 62: Significance Codes

Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .	0.1	1

Table 63: Model Parameters

Estimated dispersion parameter	5.6665
Estimated index parameter	1.3416
Residual deviance	13152 on 2801 degrees of freedom
AIC	8136.8
Number of Fisher Scoring iterations	6

For$P3$:

Table 64: Residual Deviations in the Third Model with Zero Inflation

Min	1Q	Median	3Q	Max
-4.971	-1.962	-1.371	-0.334	16.433

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	$-2.079e+00$	$4.076e-01$	-5.100	$3.63e-07$ ***
factor(CAR_USE)Commercial	$4.130e-01$	$1.307e-01$	3.1604	0.00159 **
factor(REVOLKED)Yes	$1.675e-01$	$1.442e-01$	1.161	0.24562
factor(GENDER)M	$1.396e-01$	$1.662e-01$	0.840	0.40105
factor(AREA)Urban	$2.025e+00$	$2.215e-01$	9.143	$< 2e-16$ ***
factor(MARRIED)Yes	$-4.765e-01$	$1.113e-01$	-4.281	$1.93e-05$ ***
factor(CAR_TYPE)Pickup	$1.522e-01$	$2.665e-01$	0.571	0.56802
factor(CAR_TYPE)Sedan	$-2.860e-01$	$2.765e-01$	-1.034	0.30105
factor(CAR_TYPE)Sports Car	$4.743e-01$	$3.290e-01$	1.441	0.14956
factor(CAR_TYPE)SUV	$3.454e-01$	$3.058e-01$	1.129	0.25881
factor(CAR_TYPE)Van	$1.151e-01$	$2.959e-01$	0.389	0.69719
TRAVTIME	$9.744e-03$	$3.582e-03$	2.720	0.00657**
MVR_PTS	$1.174e-01$	$2.254e-02$	5.210	$2.02e-07$ ***
INCOME	$-5.681e-06$	$1.378e-06$	-4.123	$3.85e-05$***

Signif. codes:	0***	0.001**	0.01 *	0.05.	0.1	1

Estimated dispersion parameter:	5.4463
Estimated index parameter:	1.3303
Residual deviance:	12507 on 2798 degrees of freedom
AIC:	8026.6
Number of Fisher Scoring iterations:	6

The Gini Coefficient for Variables is Calculated as Follows When the Loss Function is the Claim of Damage: and: A chart has been drawn below:

Table 65: Gini Indices in the Model with Zero Inflation:

	P1	P2	P3
P1	0.000	36.843	42.239
P2	−1.558	0.000	21.373
P3	1.340	1.636	0.000

Table 66: Standard Errors in Models with Zero Inflation:

	P1	P2	P3
P1	0.000	2.364	2.433
P2	2.882	0.000	3.024
P3	3.035	3.120	0.000

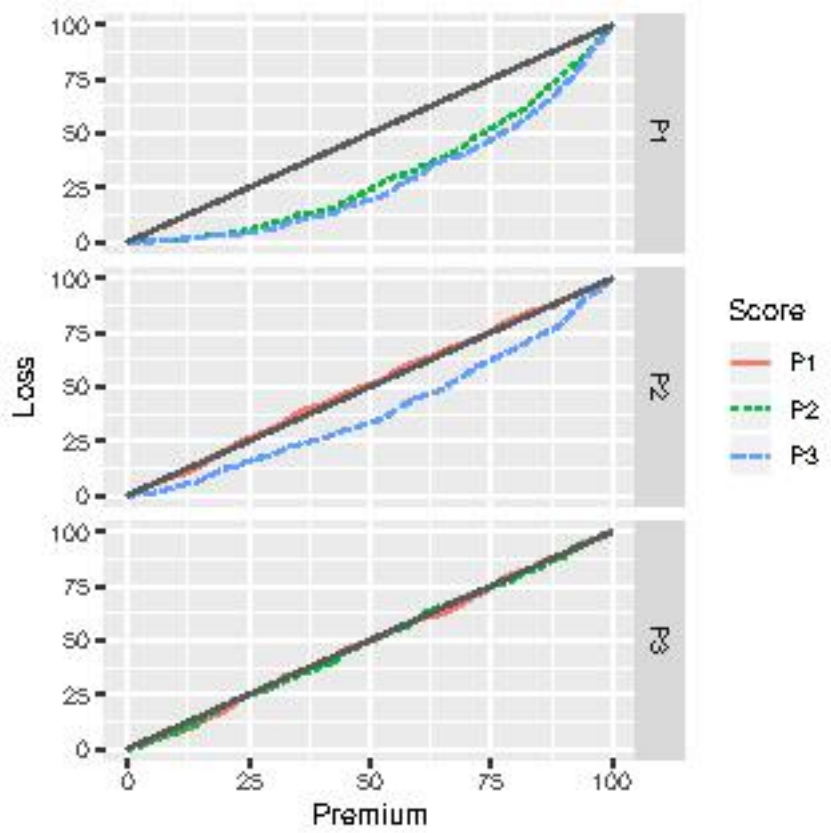

Figure 30: Gini coefficient chart in zero-inflated models

This graph represents the cumulative distribution of the population, arranged in ascending order of insurance premiums. The vertical axis denotes the cumulative income share of the population. The line of perfect equality is depicted as a 45-degree line, representing an equitable distribution.

The Gini coefficient is calculated as the ratio of the area between the Lorenz curve (the curve that represents the actual distribution of income or wealth) and the 45-degree line of equality, to the total area under the equality line. This coefficient serves as a measure of inequality within the population.

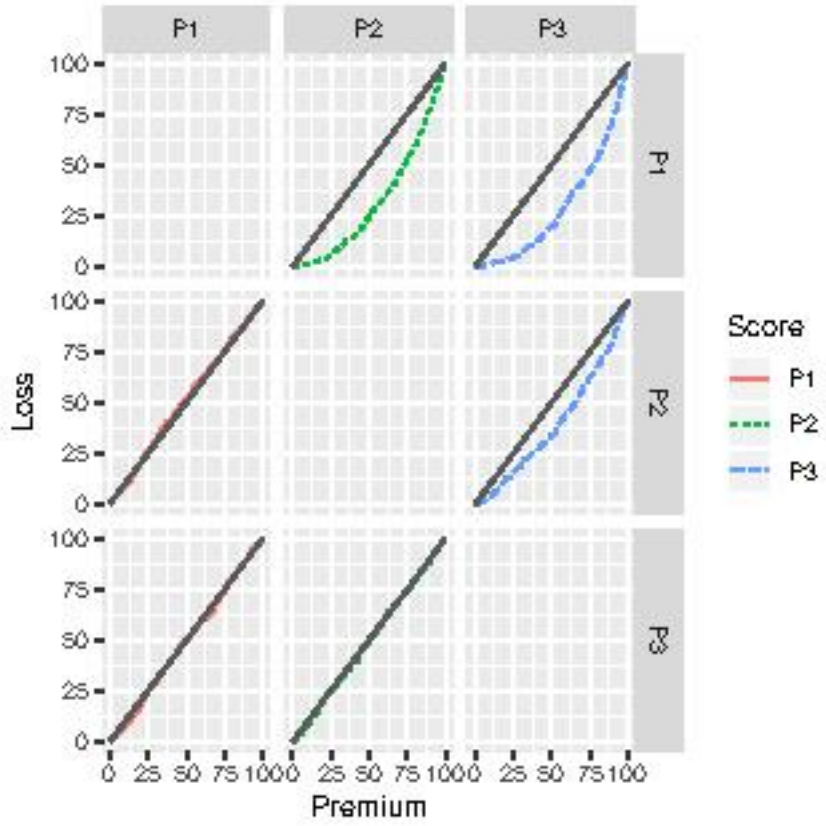

Figure 31: An Additional Graph Depicting Gini Coefficients in Models with Zero-Inflation:

78 Conclusion:

In conclusion, the application of advanced methods and the incorporation of up-to-date statistical knowledge can significantly enhance the performance of insurance companies. This thesis has shed light on the importance of considering not only the characteristics of the insured vehicle but also the specific attributes of the driver. The driver's behavior and features play a crucial role in determining insurance premiums, as highlighted in this treatise.

The key takeaways from this research include:
Comprehensive Risk Assessment: Insurance companies should adopt comprehensive risk assessment models that take into account both vehicle-related factors and driver-specific attributes. This holistic approach leads to more accurate premium calculations.

Data Integration: Access to complete and integrated data, including driver history, vehicle details, and claims records, is essential for precise premium predictions. Collaborations with external entities, such as traffic police, can provide valuable data sources.

Continuous Improvement: The insurance industry should embrace continuous improvement by regularly updating premium prediction models. Incorporating new data and refining algorithms ensures that predictions remain relevant and reliable.

Transparency and Fairness: Maintaining transparency in the premium calculation process and ensuring fairness in pricing are vital for customer satisfaction and trust. Clear communication of the factors influencing premiums is key.

Adoption of Advanced Techniques: Leveraging advanced statistical and machine learning techniques can uncover valuable insights from data. These techniques can identify complex patterns and correlations that influence insurance risk.

Customization: Tailoring premium predictions to individual risk profiles allows insurance companies to provide fair and competitive pricing. Personalized insurance offerings can enhance customer retention.

By implementing these findings and recommendations, insurance companies can not only improve their operational efficiency but also better serve their policyholders. The combination of accurate premium predictions and a customer-centric approach will contribute to the long-term success and sustainability of insurance businesses.

79 Suggestions:

Certainly, here are some suggestions for improving the accuracy of insurance premium predictions:

Complete Information: Ensure that you have access to complete and accurate information from the insurance provider. This includes data on the insured, their vehicle, driving history, and any previous claims. Comprehensive data will lead to more accurate premium predictions.

Traffic Police Records: Collaborate with traffic police authorities to access real-time information about the driver's insurance status and driving history. This information can significantly enhance the accuracy of premium predictions. Regular updates from traffic police can capture changes in the driver's risk profile over time.

Continuous Updates: Implement a system for continuous updates and monitoring of relevant data. As new data becomes available, such as traffic violations or changes in the insured's profile, incorporate this information into the premium prediction model. This ongoing process ensures that premium predictions remain up to date and accurate.

Machine Learning and Data Analysis: Utilize advanced data analysis techniques, including machine learning, to extract valuable insights from the data. Machine learning algorithms can identify patterns and relationships in the data that may not be apparent through traditional methods. This can lead to more precise premium predictions.

Customized Risk Assessment: Develop a risk assessment model that considers individualized factors, such as driving behavior, vehicle type, and lo-

cation. Tailoring premium predictions to specific risk profiles can lead to more accurate pricing.

Customer Feedback: Consider gathering feedback from policyholders to understand their experiences and perceptions of the insurance premium. Customer feedback can provide valuable insights into the fairness and accuracy of premium predictions.

Regular Model Evaluation: Periodically evaluate and refine the premium prediction model. Assess its performance against actual claims data to identify any discrepancies and make necessary adjustments.

Regulatory Compliance: Ensure that the premium prediction process complies with relevant regulations and guidelines. Compliance with legal requirements helps maintain fairness and transparency in pricing.

Transparency: Maintain transparency with policyholders regarding how premiums are calculated. Clearly communicate the factors and data sources used in premium predictions to build trust and credibility.

Benchmarking: Compare premium predictions with industry benchmarks and standards. Benchmarking can help identify areas for improvement and ensure that your predictions are competitive and accurate.

By implementing these suggestions, insurance companies can enhance the accuracy of premium predictions, leading to fairer pricing for policyholders and improved risk management.

80 appendix A:

```
library(lattice)
library(TDboost)
library(coda)
library(Matrix)
library(splines)
library(cplm)
data(AutoClaim)
AutoClaim
head(AutoClaim)

AutoClaim<- as.data.frame(AutoClaim)
da <- subset(AutoClaim, IN_YY == 1) # use data in the Yip and Yau paper
da <- transform(da, CLM_AMT5 = CLM_AMT5/1000,INCOME = INCOME/10000)
summary(da$CLM_AMT5)
sum(da$CLM_AMT5 == 0)/nrow(da)

fit <- TDboost(CLM_AMT5 ~ AGE+BLUEBOOK+ HOMEKIDS+
MARRIED +KIDSDRIV+MVR_PTS+NPOLICY+
TRAVTIME+ AREA+CAR_USE+CAR_TYPE+GENDER+
JOBCLASS+MAX_EDUC+REVOLKED,data = da,
distribution=list(name="EDM",alpha=1.5),
# specify Tweedie index parameter
n.trees=3000 ,                            # number of trees
shrinkage=0.005,                        # shrinkage or learning rate,
# 0.001 to 0.1 usually work
interaction.depth=3,                  # 1: additive model, 2: two-way interactions,
bag.fraction = 0.5,                 # subsampling fraction, 0.5 is probably best
train.fraction = 0.5,                # fraction of data for training,
# first train.fraction*N used for training
n.minobsinnode = 10,               # minimum total weight needed in each node
cv.folds = 5,                       # do 5-fold cross-validation
keep.data=TRUE,                     # keep a copy of the dataset with the object
verbose=TRUE)                        # print out progress

#print out the optimal iteration number M
best.iter <- TDboost.perf(fit,method="test")
```

```
print(best.iter)
summary(TDboost.perf(fit,method="test"))
# check performance using 5-fold cross-validation
best.iter <- TDboost.perf(fit,method="cv")
print(best.iter)
summary(TDboost.perf(fit,method="cv"))

# plot the performance
# plot variable influence
summary(fit,n.trees=1)         # based on the first tree
summary(fit,n.trees=best.iter) # based on the estimated best number of trees

# making prediction on data2
f.predict <- predict.TDboost(fit, AutoClaim, best.iter)   #model prediction / scor
print(sum((AutoClaim$y - f.predict)^2))               #least squares
print(sum((AutoClaim$ CLM_AMT5- f.predict)^2))

# plot loss function as a result of n trees added to the ensemble
TDboost.perf(fit, method = "cv")
TDboost.perf(fit, plot.it = TRUE,
oobag.curve = FALSE, overlay = TRUE, "cv")

par(mar = c(5, 8, 1, 1))
summary(fit,cBars = 16,method = relative.influence)
summary(fit,method = relative.influence)

#find index for n trees with minimum CV error
min_MSE <- which.min(fit$cv.error)
# get MSE and compute RMSE
sqrt(fit$cv.error[min_MSE])

# create marginal plots
# plot variable 1 after "best" iterations
plot.TDboost(fit,1,best.iter)
plot(fit, i.var = 1:3, n.trees = best.iter)

predict=predict.TDboost(fit, AutoClaim, best.iter,
single.tree=FALSE,type=c("response","link"))
```

```
predmatrix<-predict(fit,AutoClaim,n.trees = n.trees)

# do another 20 iterations
TDboost2 <- TDboost.more(fit,20,verbose=FALSE) # stop printing detailed progress
# fit a gamma model (when alpha = 2.0)
data2 <- da[da$CLM_AMT5 !=0,]
TDboost3 <- TDboost(CLM_AMT5 ~AGE+BLUEBOOK+ HOMEKIDS
+ MARRIED +KIDSDRIV  + MVR_PTS+NPOLICY+TRAVTIME
+ AREA+CAR_USE+CAR_TYPE+GENDER+JOBCLASS
+MAX_EDUC+MARRIED+REVOLKED,data = da,
distribution=list(name="EDM",alpha=2),
# specify Tweedie index parameter
n.trees=3000 ,                            # number of trees
shrinkage=0.005,                        # shrinkage or learning rate,
# 0.001 to 0.1 usually work
interaction.depth=3,           # 1: additive model, 2: two-way interactions, etc.
bag.fraction = 0.5,            # subsampling fraction, 0.5 is probably best
train.fraction = 0.5,          # fraction of data for training,
# first train.fraction*N used for training
n.minobsinnode = 10,            # minimum total weight needed in each node
cv.folds = 5,                   # do 5-fold cross-validation
keep.data=TRUE,                 # keep a copy of the dataset with the object
verbose=TRUE)                   # print out progress

best.iter <- TDboost.perf(TDboost3,method="test")
c(coef(fit), p = fit$p, phi = fit$phi)
print(best.iter)
summary(TDboost.perf(TDboost3,method="test"))

# check performance using 5-fold cross-validation
best.iter <- TDboost.perf(TDboost3,method="cv")
print(best.iter)
summary(TDboost.perf(TDboost3,method="cv"))
P1 <- TDboost(CLM_AMT ~ CAR_USE + REVOLKED+AREA+
MARRIED +CAR_TYPE , distribution=list(name="EDM",alpha=1.5) ,data = da)
P2 <- TDboost(CLM_AMT ~ factor(CAR_USE) + factor(REVOLKED) +
factor(GENDER) + factor(AREA) +factor(MARRIED) +
factor(CAR_TYPE), distribution=list(name="EDM",alpha=1.5)
```

```
,data =da)
P3 <- TDboost(CLM_AMT ~ factor(CAR_USE) + factor(REVOLKED) +
factor(GENDER) +factor(AREA) +
factor(MARRIED) + factor(CAR_TYPE) +
TRAVTIME + MVR_PTS + INCOME,
distribution=list(name="EDM",alpha=1.5) ,
data = da)
# compute the Gini indices
gg <- gini(loss = "CLM_AMT", score = paste("P", 1:3, sep = ""),
data = da)
gg
# plot the Lorenz curves theme_set(theme_bw())
plot(gg)
plot(gg, overlay = FALSE)
```

81 appendix B:

Regression Tree:

```
# Regression Tree
library(rpart)

# grow tree
fit <- rpart(CLM_AMT5 ~ AGE+BLUEBOOK+ HOMEKIDS+
MARRIED + AREA +
MVR_PTS+NPOLICY+RETAINED+TRAVTIME+AREA+CAR_USE
+CAR_TYPE+GENDER+JOBCLASS+MAX_EDUC+REVOLKED,
method="anova", data=AutoClaim)

printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits

# create additional plots
par(mfrow=c(1,2)) # two plots on one page
rsq.rpart(fit) # visualize cross-validation results
# plot tree
plot(fit, uniform=TRUE,
```

```
main="Regression Tree ")
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

82 appendix C :

GLM method:

```
AutoClaim<- as.data.frame(AutoClaim)
linear<- glm(CLM_AMT5 ~ AGE+BLUEBOOK+ HOMEKIDS
+ MARRIED + AREA + MVR_PTS+NPOLICY+RETAINED+TRAVTIME
+AREA+CAR_USE+CAR_TYPE+GENDER+JOBCLASS
+MAX_EDUC+REVOLKED,data=AutoClaim)
summary(linear)
pred=predict(linear,data=AutoClaim)
anova(linear)

confint(linear) # 95% CI for the coefficients
exp(coef(linear)) # exponentiated coefficients
exp(confint(linear)) # 95% CI for exponentiated coefficients
predict(linear, type="response") # predicted values
residuals(linear, type="deviance") # residuals

# Poisson Regression
fit <- glm(CLM_AMT5 ~ AGE+BLUEBOOK+ HOMEKIDS
+ MARRIED + AREA + MVR_PTS+NPOLICY+RETAINED
+TRAVTIME+AREA+CAR_USE+CAR_TYPE+GENDER
+JOBCLASS+MAX_EDUC+REVOLKED
,data=AutoClaim,family=poisson())
summary(fit) #display results

P1 <- glm(CLM_AMT ~ 1, data = da,offset = log(NPOLICY))
P2 <- glm(CLM_AMT ~ factor(CAR_USE) + factor(REVOLKED) +
factor(GENDER) + factor(AREA) +
factor(MARRIED) + factor(CAR_TYPE),
data = da, offset = log(NPOLICY))
P3 <- glm(CLM_AMT ~ factor(CAR_USE) + factor(REVOLKED) +
factor(GENDER) + factor(AREA) +
factor(MARRIED) + factor(CAR_TYPE) +
```

```
TRAVTIME + MVR_PTS + INCOME,
data = da, offset = log(NPOLICY))
da <- transform(da, P1 = fitted(P1), P2 = fitted(P2), P3 = fitted(P3))
# compute the Gini indices
gg <- gini(loss = "CLM_AMT", score  = paste("P", 1:3, sep = ""), data = da)
gg
# plot the Lorenz curves theme_set(theme_bw())
plot(gg)
plot(gg, overlay = FALSE)
## End(Not run)
```

83 appendix D:

CPLM method

```
da <- subset(AutoClaim, IN_YY == 1) # use data in the Yip and Yau paper
da <- transform(da, CLM_AMT5 = CLM_AMT5/1000,INCOME = INCOME/10000)
summary(da$CLM_AMT5)
sum(da$CLM_AMT5 == 0)/nrow(da)

P1 <- cpglm(CLM_AMT5 ~AGE+BLUEBOOK+ HOMEKIDS
+ MARRIED +KIDSDRIV+MVR_PTS+NPOLICY
+TRAVTIME+ AREA+CAR_USE+CAR_TYPE+GENDER+
JOBCLASS+MAX_EDUC+MARRIED+REVOLKED,data = da)
summary(P1)
coef(P1)
c(coef(P1), p = P1$p, phi = P1$phi)
print(P1)
---------
#Gini
# Let's fit a series of models and compare them using the Gini index
da <- subset(AutoClaim, IN_YY == 1)
da <- transform(da, CLM_AMT = CLM_AMT / 1000)
P1 <- cpglm(CLM_AMT ~ 1, data = da, offset = log(NPOLICY))
P2 <- cpglm(CLM_AMT ~ factor(CAR_USE) + factor(REVOLKED) +
factor(GENDER) + factor(AREA) +
factor(MARRIED) + factor(CAR_TYPE),
data = da, offset = log(NPOLICY))
```

```
P3 <- cpglm(CLM_AMT ~ factor(CAR_USE) + factor(REVOLKED) +
factor(GENDER) + factor(AREA) +
factor(MARRIED) + factor(CAR_TYPE) +
TRAVTIME + MVR_PTS + INCOME,
data = da, offset = log(NPOLICY))
da <- transform(da, P1 = fitted(P1), P2 = fitted(P2), P3 = fitted(P3))
# compute the Gini indices
gg <- gini(loss = "CLM_AMT", score  = paste("P", 1:3, sep = ""),
data = da)
gg
# plot the Lorenz curves
theme_set(theme_bw())
plot(gg)
plot(gg, overlay = FALSE)
## End(Not run)
```

84 REFERENCES:

References

[1] A. Bazargan -Lari. Applied Linear regression (1384)

[2] Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

[3] Breiman, L., Friedman, J., Olshen, R., Stone, C., Steinberg, D., and Colla, P. (1984),*CART: Classication and regression trees," Wadsworth.*

[4] Breiman, L. (1999),*Prediction games and arcing algorithms, Neural computation,*11*,* 14931517*.* 1

[5] Bühlmann, P. and Hothorn, T., 2007,*Boosting algorithms: Regularization, prediction and model fitting. Statistical Science, 22(4), pp.*477-505*.*

[6] Cragg, J.G., 1971,*Some statistical models for limited dependent variables with application to the demand for durable goods. Econometrica (pre-*1986*), 39(5), p.*829*.*

[7] Dunn, P. K. and Smyth, G. K. (2005),*Series evaluation of Tweedie exponential dispersion model densities, Statistics and Computing, 15,* 267280*.* 3*,*6.4*, F.2*

[8] Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.

[9] Friedman, J., Hastie, T., and Tibshirani, R. (2001),*The elements of statistical learning, vol. 1, Springer series in statistics New York, NY, USA:.*1*,*3

[10] He Zhou, Yi Yang , & Wei Qian. (2018), *Tweedie Gradeint Boosting for Extremely Unbalanced zero-inflated Data.*

[11] Gholahossein djabbari. Insurance terminology (1396)

[12] Gorman, B. (2017),*A Kaggle master explains gradient boosting. Kaggle Blog.*

[13] Green, P. J., and B. W. Silverman (1994) ,*Nonparametric regression and generalized linear models. Number 58 in Monographs on Statistics and Applied Probability. Nonparametric regression and generalized linear models, (58).*

[14] Hall, M. A. (1999),*Correlation-based feature selection for machine learning.*

[15] Hall, D. B. (2000). Zeroinflated Poisson and binomial regression with random effects: a case study. Biometrics, 56(4), 1030-1039.

[16] Hastie, T. J. and Tibshirani,R.J.(1990),*Generalized additive models, volume 43 of Monographs on Statistics and Applied Probability,"*

[17] Jørgensen, B. (1987) ,*Exponential dispersion models, Journal of the Royal Statistical Society. Series B (Methodological),* 127162*.2.1*

[18] Jørgensen, B., & Paes De Souza, M. C. (1994). Fitting Tweedie's compound Poisson model to insurance claims data. Scandinavian Actuarial Journal, 1994(1), 69-93.

[19] Jorgensen, B. (1997) ,*The theory of dispersion models. CRC Press.*

[20] Smyth, G. K., & Jørgensen, B. (2002).*Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. ASTIN Bulletin: The Journal of the IAA,* 32*(1),* 143-157.

[21] Lambert, D(1992),*Zero-inated Poisson regression, with an application to defects in manufacturing, Technometrics, 34,* 114.

[22] Leathwick, J. R., Elith, J., & Hastie, T. (2006),*Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological modelling,* 199*(2),* 188-196.

[23] McCullagh, P. "Nelder. JA (1989),*Generalized linear models* 2 *(85)*

[24] McLachlan, G. and Krishnan, T. (2007),*The EM algorithm and extensions, vol. 382, John Wiley& Sons.*

[25] Mildenhall, S. J. (1999)„*A systematic relationship between minimum bias and generalized linear models, in Proceedings of the Casualty Actuarial Society, vol. 86, pp.*393487*. 1*

[26] Mullahy, John,*"Specification and testing of some modified count data models." Journal of econometrics 33, no.* 3 *(1986):* 341-365*.*

[27] Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. Journal of the Royal Statistical Society: Series A (General), 135(3), 370-384.

[28] Ridgeway, G., 2007,*Generalized Boosted Models: A guide to the gbm package. Update, 1(1), p.2007*

[29] Sandri, M. and Zuccolotto, P., 2008 ,*A bias correction algorithm for the Gini variable importance measure in classification trees. Journal of Computational and Graphical Statistics, 17(3), pp.*611-628*.*

[30] Simon CK Lee,& Katrin Antonio (2015) ,*Why High Dimentional Modeling in Actuarial Science*

[31] Smyth, G. K., & Jørgensen, B. (2002). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. ASTIN Bulletin: The Journal of the IAA, 32(1), 143-157.

[32] Tweedie, M. C. (1984, December). An index which distinguishes between some important exponential families. In Statistics: Applications and new directions: Proc. Indian statistical institute golden Jubilee International conference (Vol. 579, pp. 579-604).

[33] Wood, S. N. (2001). mgcv: GAMs and generalized ridge regression for R. R news, 1(2), 20-25.

[34] Neal, D. T., Wood, W., & Quinn, J. M. (2006),*HabitsA repeat performance. Current directions in psychological science,* 15*(4),* 198-202*.*

[35] *Webster Dictionary*

[36] Yang, Y., Qian, W., & Zou, H. (2018). Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. Journal of Business & Economic Statistics, 36(3), 456-470.

[37] Yang, Z., & Ai, H. (2007),*Demographic classification with local binary patterns. In International Conference on Biometrics (pp. 464-473). Springer, Berlin, Heidelberg.*

[38] Yip, K. C. and Yau, K. K. (2005),*On modeling claim frequency data in general insurance with extra zeros, Insurance: Mathematics and Economics, 36,* 153163. *6.1*

[39] Zhang, T., Yu, B., et al. (2005),*Boosting with early stopping: Convergence and consistency, The Annals of Statistics, 33,* 15381579. *6.1*

[40] Zhang, W. (2011), *cplm: Monte Carlo EM algorithms and Bayesian methods for fitting Tweedie compound Poisson linear models," R package version* 0.2-1*, URL http://CRAN. R-project.org/package= cplm.*