Predicting Insurance Reserve using Varying-Coefficients Model with Machine Learning and Statistical Techniques

Danlei Zhu
Advisor: Dr.Don Hong
Co-advisor: Dr.Vajira Manathunga

Computational and Data Science program
Middle Tennessee State University

Abstract

Since 1980s, actuaries always choose Generalized Linear Models (GLM) serve as a foundation for analysis projects to help insurance companies predict reserves or unearned premium risk at a particular time spot, since industry regulation requires models should be transparent enough to be interpretable. However, classical linear models often fall short in capturing complex relationships present in real-world scenarios, which often leads to a violation of assumptions. With wild usage of Machine learning techniques in many industries, actuaries do not often apply these complicated models as the first choice due to "Black box" performance which causes less interpretability. To improve this problem, this research is the first to introduce decision tree boosted varying-coefficients model (VCM) to actuarial data for analysis. It tries to achieve greater realism by allowing constant coefficients in classical linear models to be dependent on a set of covariates. Through models training and predicting, it has been shown that VCM is worthy to be used, it performs well while keeping interpretability of linear form and allowing complexity of advanced machine learning approach.

Keywords: *reserve; regression model; generalized linear model; robust model; varying-coefficients model; decision tree; gradient boosting decision tree; cubic spline*

1. Introduction

In the health and life insurance sectors, policyholders are monitored for changes in their life status, such as health, illness, or death, to anticipate changes in variables and assess future risks. In the Property & Casualty insurance sectors, the policyholder's insured property's condition, and accident time, season are usually regarded as related factors to the final payment amount. In general, actuaries always choose linear regression models as a foundation, and help insurance companies predict reserves or unearned premium risk at a particular time spot, while Generalized Linear Models (GLM) are frequently utilized for various modeling endeavors. The appeal of linear models comes from linearity, interpretability, and other well-established properties. However, Fomby, Johnson, and Hill (1984) [5] found that when a data set consists of a series of observations over time on a cross-section unit, it often resulted in a violation of the assumptions of the linear regression model. For instance, Fan and Zhang (2008) [4] mentioned that when analyzing infant mortality in China, the relation with many influencing factors like population size were not considered to change over time, which is contrary to reality. In these circumstances, constant coefficient modeling could result in erroneous forecasts or a poor fit of the model to the data. Non-parametric approaches has been proposed to overcome limitations in linear models. In general, non-parametric approaches make no assumptions on the underlying model which is contrary to linearity assumption used by linear model. However, these models also have some limitations. For example, if some prior information cannot be considered, the estimated result of the unknown function will frequently cause a greater variance and is even worse for a dimensional curse.

To address this, statisticians at Stanford University Hastie and Tibshirani (1993) [9] first introduced varying coefficient models (VCM), which allow for regression coefficients to vary smoothly and systematically across multiple dimensions. This is an improvement on the standard approaches to reflect more information about the influence of a certain element and to add it to the model in a logical way to produce forecasts with greater accuracy. Researchers at Princeton University such as Fan and Zhang [3][4], among many others have made significant advancements in the field of VCM, methodologically and theoretically, as well as practically. Then, Runze li and his group at Penn State explored more application of VCM, such as effective method to make feature selection in ultra-high dimension case [2] [13]. Park et al. in there research [18] also mentioned that VCM remains the interpretibility of linear regression with more flexibility by allowing non-linear coefficients, and save model from dimensional curse.

Some traditional statistical methods are popular to be used in VCM, like B-spline in paper[11], polynomial spline[10] and P-spline[14], also a lately paper by Wang, Jiang and Liu in 2024 shows the value of adaptive spline application[20]. VCM can be applied in conjun ction with machine learning models as well. Zhou and Hooke (2022) [22] proposed a varying coefficient model, that can achieve higher model complexity while retaining the structure of underlying parameter models, to generate interpretable predictions. The proposed method is applied to the decision tree, and achieved high training speed, prediction accuracy, and comprehensibility. This also effectively proves that VCM can be applied not only to statistical models but also to machine learning methods. In the study of medical treatment by Nie et al.(2021) [16], VCM applied to the neural network, which greatly improved the performance after combining the structure of the model with a targeted regularization boost.

2. Background

Fomby et al.(1984) [5] found: when a dataset consists of a series of time-varying observations on cross-sectional units, it often leads to a violation of the assumptions of the linear regression model, where the study cannot assume that the parameters of all observations are the same. They proposed varying coefficient-a random variable Y whose distribution depends on a parameter η, and predictors $X_1, X_2, ..., X_p$, $R_0, R_1, R_2, ..., R_p$. A VCM has the form:

$$\eta = \beta_0(R_0) + X_1\beta_1(R_1) + ... + X_p\beta_p(R_p) \tag{1}$$

where $\vec{X} = [X_1, X_2, ..., X_p]$ are predictors for the linear model,$\vec{R} = [R_1, R_2, ..., R_p]$ represents the predictors for coefficients model. Applied GLM, Y function can be expressed via a link function

$$g(\eta) = \theta(R, X) = X^T\beta(R) \tag{2}$$

$g(.)$ is a link function, the common used distributions are normal, Poisson, and NB distribution.

Longitudinal data, repeated measurements of statistics for a same sample over time, is critical to the actuarial field. For example, actuaries collect number of accidents, claim amounts, socio-economic changes for a group of auto insurance policyholders over years to anticipate changes in variables and assess risk for coming years. Common approaches such as fixed effect and random effect models do not allow coefficients to vary over time. Thus, Frees (2009) [6] suggested extension of these models by introducing variable coefficients. He used the following example when explaining why variable coefficients models are better suited than traditional models. Consider Medicare inpatient hospital charges data set for six years for 54 states and assume actuaries are interested about covered claims per discharge (CCPD). Assume that CCPD depends on state, time (YEAR), number of discharge (NUMDCHG), and average hospital stay per discharge in days (AVEDAYS). Ordinary linear regression model:

$$E[CCPD_i] = \eta_i + \beta_1(NUMDCHG)_i + \beta_2(YEAR)_i + \beta_3(AVEDAYS)_i \tag{3}$$

Basic fixed effect model:

$$E[CCPD_{it}] = \eta_i + \beta_1(NUMDCHG)_{it} + \beta_2(YEAR)_{it} + \beta_3(AVEDAYS)_{it} \tag{4}$$

VCM:

$$E[CCPD_{it}] = \eta_i + \beta_1(NUMDCHG)_{it} + \beta_{2i}(YEAR)_t + \beta_3(AVEDAYS)_{it} \tag{5}$$

Here i stands for state and t for time. Frees concluded that when the coefficient associated with YEAR is allowed to vary with state i the model would produce excellent fit.

VCM is more flexible and better fitted than simple linear models since the coefficients can be non-parametric while retaining a linear structure in form, which gives them a reasonably strong explanatory power in practical applications. It is relatively broad, and many classical models can be viewed as special examples of VCM, many common used models are allowed and easily to be developed by varying-coefficient form.

2.1. Generalized additive models

Generalized Additive Models (GAMs) are a flexible class of statistical models that extend the concept of Generalized Linear Models (GLMs) by allowing for non-linear relationships between the response variable and the predictor variables. GAMs were introduced by Trevor Hastie and Robert Tibshirani in 1986 as a way to incorporate smooth functions of the predictors into the model. If two conditions below are satisfied,

1. If $\beta_j(R_j) = \beta_j$, a constant function, usual a linear model or a generalized linear model
2. If $X_j = c$ (say c = 1), the j^{th} term is $\beta_j(R_j)$

then equation(4) has the form of a generalized additive model.

2.2. Dynamic generalized linear model

Dynamic Generalized Linear Model is a statistical approach that extends the concept of GLMs to deal with time series data with time-varying parameters. It allows for modeling dynamic relationships between independent variables and the target variable over time. They are widely used in economics, finance, and environmental sciences fields, where data often shows temporal patterns and changing relationships. If R_j is a factor that such as time or age, use t to represent. In this case, also can modify X_j, we have t in $\{t_1, t_2, ..., t_n\}$:

$$\eta_t = \beta_0(t) + X_1(t)\beta_1(t) + ... + X_p(t)\beta_p(t) \tag{6}$$

It's a dynamic generalized linear model.

2.3. Non-proportional hazards model

A statistical survival analysis method used to examine time-to-event data is the Cox proportional hazards model, sometimes referred to as the Cox regression model or Cox model. The link between survival time and one or more predictor factors is studied using this technique, which was invented by Sir David Cox in 1972 and is frequently utilized in the social and medical sciences. The Cox model's format is:

$$\lambda(t|X_1...X_p) = \lambda_0(t) \exp\{\sum_j X_j\beta\} \tag{7}$$

where $\lambda_0(t)$ is an arbitrary base-line hazard function. When developing it with varying-coefficients model, it's called Non-proportional hazards model, and the formula extend to:

$$\lambda(t|X_1...X_p) = \lambda_0(t)\exp\{\sum_j X_j\beta_j(t)\} \tag{8}$$

2.4. Time-series model

When analyzing and predicting data that changes over time, such as stock prices, temperature, or sales numbers, a time-series model can be a useful statistical tool. Time-series models explicitly take into consideration the temporal dependencies and patterns contained in the data, in contrast to typical regression models. Given time series $\{X_t\}$, applied varying-coefficient form:

$$X_t = \beta_0(X_{t-p}) + \beta_1(X_{t-p})X_{t-1} + ... + \beta_k(X_{t-p})X_{t-k} + \epsilon_t, \tag{9}$$

for some given lags k and p. When considering the differences between it and varying-coefficients models, time series models are designed to analyze data with a temporal structure, where observations are taken at regular intervals over time. VCM is appropriate if suspect that the relationship between the response and predictor variables is not constant but varies with certain conditions or ranges of the predictors, not limited to time-related condition.

2.5. Semi-varying coefficients model

Semi-varying coefficients models are a statistical modeling technique that expands on the idea of VCM to deal with scenarios in which just a portion of the coefficients vary with some covariates while others remain constant (fixed) [4]. It is intended to represent both the regional diversity in particular coefficient effects and the global average link between the response and predictors. When some coefficients do not display considerable fluctuation, the smooth change of all coefficients over time or other covariates in a standard VCM may be excessive and lead to overfitting. By enabling a combination of fixed and varying coefficients, the semi-varying coefficients model provides a more frugal and understandable method.

The general form of a semi-varying coefficients model can be represented as:

$$y = X_1^T\beta_1(R) + X_2^T\beta_2 + \epsilon \tag{10}$$

where $(X_1^T, X_2^T)^T = X$, X_i is a p_i dimensional covariate, i = 1, 2, and $p_1 + p_2 = p$.

Theoretically, many methods can be used to model unspecified function $\beta_j(R_j)$, such as polynomials, Fourier series, piece-wise polynomials, general non-parametric functions, if R_j can be scalar or vector valued, then kernel method, penalization or stochastic Bayesian formulations can be used.

In this project, a statistical and mathematical model will be employed to varying-coefficients model; at the same time, a common used ML technique gradient boosting decision tree is applied.

3. Methodologies

3.1. Varying-coefficients model

3.1.1. Decision tree boosted VCM

Zhou and Hooker (2022) [22] has applied gradient boosting decision tree approach to VCM and developed the source code package called **VCboost**. The package use following methods:

1. Gradient Boosting: this is the most critical numerical method used in implementation. Gradient boosting decision tree is an ensemble technique that builds a number of weak learners sequentially, inside the algorithm each tree is built to correct the residuals between predicted values and real values made by the previous one. It iteratively fits new decision tree to the negative gradients, aims to minimize the loss.

2. Decision Tree: In this model, decision tree is used as base learners in the gradient boosting process. It is a non-linear supervised learning algorithm which can be used for both classification and regression problems. Here, regression trees are applied to approximate the negative gradients for each iteration.

3. Line Search: The code uses line search techniques to adjust the leaf values of the decision trees. The line search is a numerical optimization method to find the step size in the negative gradient direction that minimizes the loss function.

4. Least Squares Regression Tree: The implementation approximates the negative gradient using a least squares regression tree, which is a decision tree built using the mean squared error (MSE) as the splitting criterion.

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right] \tag{11}$$

   This method helps approximate the negative gradients more accurately for updating the coefficients.

3.1.2. Cubic spline VCM

As described in Huang, Zhou and Liu's research of basis functions used in VCM[11], this paper aims to explore the performance of cubic spline applied as varying coefficient models in linear regression, we will explain and show algorithms derivation and proof of mathematical equations. As we said, $\mathbf{X}$ and β are matrices representing independent variables and coefficients respectively, $\mathbf{y}$ means target variables for each observation in data set. $\mathbf{B(t)}$ is a (k by 4*k) matrix, where k is the number of coefficients, here 4 means 4 components in cubic spline: [1, t, t^2, t^3], usually people use t as time series factor, but it's also acceptable to use any numerical variables as varying factor. Thus, we develop another matrix

$$\mathbf{U_i}(\mathbf{t_{ij}})=[\mathbf{X_{ij}}\mathbf{B_i}(\mathbf{t_{ij}})]^T, \tag{12}$$

and $\mathbf{U_i} = (\mathbf{U_i}(\mathbf{t_{i1}}), ..., \mathbf{U_i}(\mathbf{t_{in_i}}))$. Then, we mark

$$\gamma = (\gamma_{00}, \gamma_{01}, \gamma_{02}, \gamma_{03}, ..., \gamma_{k0}, \gamma_{k1}, \gamma_{k2}, \gamma_{k3}) \tag{13}$$

as coefficients in varying coefficient models (cubic spline here). $\mathbf{W_i}$ is diagonal weighted matrix, all factors in diagonal are 1/time_step if we treated varying factor is time. In this way, it's easy to prove

$$l(r) = \sum_{i=1}^{n} (Y_i - U_i\gamma)^T Wi(Y_i - U_i\gamma) \tag{14}$$

and

$$l(r) = \sum_{i=1}^{n} W_i \sum_{j=1}^{n_i} (Y_{ij} - \sum_{l=0}^{k} \sum_{s=0}^{3} X_i(l)(t_{ij}) B_{ks}(t_{ij})\gamma_{ls})^2 \tag{15}$$

are equivalent, which need least square to solve the optimal parameter for γ. To prove that the least squares estimator for y is

$$y = \left(\sum \mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i\right)^{-1} \left(\sum \mathbf{U}_i^T \mathbf{W}_i \mathbf{Y}_i\right),$$

we start by understanding the form of the loss function and then proceed by differentiating it with respect to y and setting the derivative to zero. The loss function is given in equation (14), to find the least squares estimator, we need to minimize this loss function with respect to r. Expand the loss function:

$$l(r) = \sum \left[(\mathbf{Y}_i - \mathbf{U}_i y)^T \mathbf{W}_i (\mathbf{Y}_i - \mathbf{U}_i y)\right]$$

Expanding the quadratic form inside the summation:

$$l(r) = \mathbf{Y}_i^T \mathbf{W}_i \mathbf{Y}_i - \mathbf{Y}_i^T \mathbf{W}_i \mathbf{U}_i y - y^T \mathbf{U}_i^T \mathbf{W}_i \mathbf{Y}_i + y^T \mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i y$$

Since $y^T \mathbf{U}_i^T \mathbf{W}_i \mathbf{Y}_i$ is a scalar, it is equal to its transpose:

$$y^T \mathbf{U}_i^T \mathbf{W}_i \mathbf{Y}_i = (\mathbf{Y}_i^T \mathbf{W}_i \mathbf{U}_i y)^T = y^T \mathbf{U}_i^T \mathbf{W}_i \mathbf{Y}_i$$

So, we can rewrite the expanded form as:

$$l(y) = \sum \left[\mathbf{Y}_i^T \mathbf{W}_i \mathbf{Y}_i - 2\mathbf{Y}_i^T \mathbf{W}_i \mathbf{U}_i y + y^T \mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i y\right]$$

Differentiate the loss function with respect to y: To minimize $l(y)$, we take the derivative with respect to y and set it to zero:

$$\frac{\partial l(y)}{\partial y} = \sum \left[-2\mathbf{U}_i^T \mathbf{W}_i \mathbf{Y}_i + 2\mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i y\right] = 0$$

Solve for y:

$$\sum \left[-2\mathbf{U}_i^T \mathbf{W}_i \mathbf{Y}_i + 2\mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i y\right] = 0$$

$$-2 \sum \mathbf{U}_i^T \mathbf{W}_i \mathbf{Y}_i + 2 \sum \mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i y = 0$$

$$\sum \mathbf{U}_i^T \mathbf{W}_i \mathbf{Y}_i = \sum \mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i y$$

Assuming $\sum \mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i$ is invertible, we solve for y:

$$y = \left(\sum \mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i\right)^{-1} \left(\sum \mathbf{U}_i^T \mathbf{W}_i \mathbf{Y}_i\right)$$

Thus, we have shown that the least squares estimator for y is:

$$y = \left(\sum \mathbf{U}_i^T \mathbf{W}_i \mathbf{U}_i\right)^{-1} \left(\sum \mathbf{U}_i^T \mathbf{W}_i \mathbf{Y}_i\right)$$

3.2. Linear regression model

This research uses the Maximum likelihood estimation (MLE) to fit the constant coefficients for linear regression, it is a widespread statistical technique that maximizes the likelihood function of a collection of observed data x_i as it's name to estimate the parameters θ of a probability distribution, which equals to maximize (minimize) log-likelihood (negative) function while defining the log-likelihood function:

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{16}$$

Estimation for $\hat{\theta}$:

$$\hat{\theta} = \arg\max(\log L(\theta)) = \arg\min(-\log L(\theta)) \tag{17}$$

Set the derivatives to zero:

$$\frac{\partial}{\partial \theta_j} \log L(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \log f(x_i|\theta) = 0 \tag{18}$$

When the underlying data has a known probability distribution, it is a vital tool for statistical inference and is especially helpful.

3.3. Decision tree

Decision trees are steadily developing in the chemical and biological sciences and have a strong base in the literature on machine learning and artificial intelligence [15]. In both classification and regression applications, the decision trees algorithm is a well-liked and adaptable machine learning technique. It works by recursively dividing the feature space into areas according to the values of the input features, resulting in a structure resembling a tree. Each leaf node of the tree represents a prediction or a class label, while each internal node indicates a choice based on a particular trait, leading to one of its child nodes. By attempting to maximize information gain (for classification) or reduce mean squared error (for regression), the method aims to identify the best splits at each node. Decision trees are useful for analyzing feature importance and decision-making processes since they are simple to interpret.

4. Data

In this project, an auto insurance data set was downloaded from the website https://github.com/kasaai/bnn-claims. This data set generated data using the individual claims history simulator developed by Gabrielli and V Wüthrich, information is described in paper [7]. Totally 40023 observations generated, the following columns of information are included for each one: unique claim number, the claimant's line of business, accident year, accident quarter, age of the injured, part of body injured, and the reported delayed years (shown in table 1). Additionally, the insurance development year is 12, from year 1994 to 2005 which means all claims reported in 1994 would be fully settled by 2006. The recorded claims history displays late reporting and negative cash flows for recoveries that match realistic events. Since this dataset contains records for 12 development years, the sum of

Columns	Explanations	Type
ClNr	Claim number (unique)	numeric
LoB	Line of business	categorical
cc	Claim code	categorical
AY	Accident year	numeric
AQ	Accident quarter	numeric
age	Policyholder's age	numeric
inj_part	Injured body part	categorical
RepDel	Reported delayed years	numeric
Pay00-Pay11	Payment in a certain single year after accident	numeric

Table 1. Original data set information

known payment amount in each of future development year can be roughly regarded as unearned premium theoretically, a unearned premium risk will happen if this amount is greater than real unearned premium which decides a loss or a surplus of company. It was inspired to choose to use this set of data to predict insurance company reserves for each year. Thus, each observation will be used to fit 12 reserve models for each development year. Based on the definition of reserve: "Liability set aside to pay future claims" [17], 12 new columns used as dependent variable for each model respectively needs to be created, and some data pre-processing is necessary.

4.1. Data pre-processing

After analysis, the 12 new variables "data_year" are calculated as description in table 2. All categorical variables use one-hot encoded to transfer from string or object to integer form. When analyzing the distribution for each variable and make transformation to those extreme skewed, hence **log** form of 'age', square root of 'inj_part' are used for models. In addition, pay_year (pay00-pay11) are standardized to avoid a large scale. Visually, figure 1 shows the relationship between the newly created variables.

Inside the table 3, Fixed variables contains ['LoB','age','inj_part','cc','RepDel','AQ']. The training data is 80% of total observations, randomly chosen. The rest 20% is test data, cross-validation with kfold (5) is used.

Development Year	1994-2005
Evaluation date	Every Jan 1st
pay_year	0, Pay00, Pay01, Pay02,..., Pay10 The previous year's payment of cutoff date
data_year (Reserves)	Year_x = $\sum pay\_i$ (i = x,x+1,...,11) Total future payment after the evaluation date

Table 2. New variables

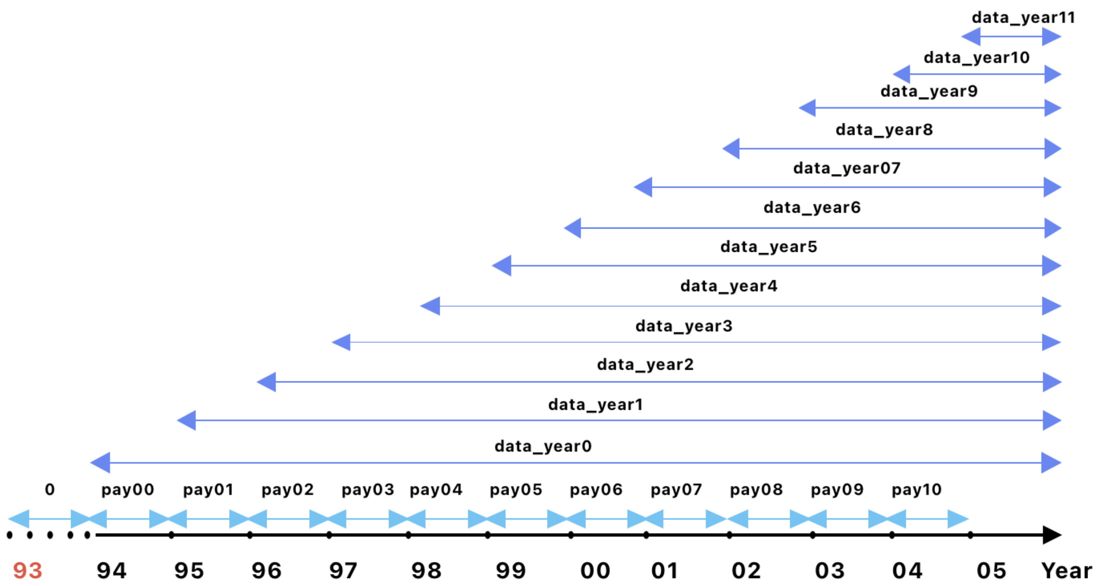

Figure 1. Variable explanation

Model	Evaluation date	Input	#	Output
1	01/01/1994	pay_year[:,0], Fixed	7	data_year[:,0]
2	01/01/1995	pay_year[:,0:1], Fixed	8	data_year[:,1]
3	01/01/1996	pay_year[:,0:2], Fixed	9	data_year[:,2]
...	...	...	...	...
12	01/01/2005	pay_year[:,0:11], Fixed	18	data_year[:,11]

Table 3. Training process

5. Case study

Considering the particularities of the actuarial industry, in general, while pricing insurance, insurers need to be able to explain to regulators, customers and other stakeholders how the model works, and how the model makes premium decisions. Thus, transparent model also promotes trust and fairness. With advanced AI techniques like ML and deep learning developed fast and widely used in many industries, it can be observed that actuarial do not often consider them as the first choice. Actuaries' reluctance to use them may stem from regulatory requirements, such as interpretability. As we know, more complicated means less interpretable, advanced ML or deep learning approaches perform like a complete "Black Box", no one can explain in details how's every step going and clearly see the relationship between features and target variable. Kuo also pointed the importance of interpretability and the fact that complicated methods are not popular in actuarial industry in his paper [12]. After searching and reading a lot of literature, I found that the VCM can improve this problem well, and no actuaries or scholars have applied the it to the field of actuarial data so far. With this in mind, this case study is the first to introduce VCM with decision tree boosted approach to analyze auto insurance reserve.

In this project, after data pre-processing, parametric analysis and model fitting are performed on available insurance data with a 12-year coverage period and containing information on all insured persons, aims to estimate the Reported but Not Settled (RBNS) reserves of insurance company each year. If the outliers represent anomalies that are important in the business and need to be detected, the Mean squared error (MSE):

$$MSE = \frac{1}{n}\sum(y_i - \hat{y}_i)^2 \tag{19}$$

as a loss function should be selected. In contrast, if only outliers are treated as damaged data, the Mean aboslute error (MAE):

$$MAE = \frac{1}{n}\sum|y_i - \hat{y}_i| \tag{20}$$

should be selected. In general, when dealing with outliers, MAE loss function is more stable, but the MSE loss function is more sensitive to outliers. When fitting a reserve model, it is generally not advisable to remove outliers from the data without proper justification and careful consideration. Outliers are data points that lie far away from the majority of other data points, and they can have a significant impact on the statistical analysis and modeling process. Consider the property of actuarial knowledge and insurance company's preferences, preparation for reserve is an essential work. It may lose useful information by remove any outliers and can't stand the model that contains much outliers, they may result in huge estimation error for preparing reserve and cause company financial issue. Therefore, this project choose MSE as evaluation for comparison, MAE as a reference metrics.

Due to the background information of the insurance data simulator, the simulated variables are all generated based on real data, and are all important features for insurance claims. In this way, for models based on linear regression -linear regression model, VCM, and robust Huber regression model, all six fixed variables are considered as essential predictors.

The developed Python package and function **boost.VCBooster** is used to build VCM, the equation can be expressed:

$$y = \beta_0(\mathbf{Z}) + X_1\beta_1(\mathbf{Z}) + ... + X_p\beta_p(\mathbf{Z}) \tag{21}$$

$\beta(\mathbf{Z})$ is a vector contains (p+1)-dimension gradient boosting decision tree and cubic spline models for decision tree boosted VCM and cubic spline VCM respectively. First, the feature selection was conducted, and we assumed a full-varying coefficients is applied. There are two problems: what is the optimal number of predictors in coefficients model, and what is/are the predictor(s) should be input. After training all combinations of 6 predictors ('LoB', 'age', 'inj_part', 'cc', 'RepDel', 'AQ'), table 2 and table 3 show the overall trade of mean MSE an mean MAE for each model with different number of predictors for decision tree VCM. Observation reveals that the mean MSE and mean MAE values are generally lowest and performance is optimal for the reserve model on most of 12 evaluation dates when $\mathbf{Z}$ contains 1 factor. As for cubic spline VCM, the corresponding factor only contain a single one, after examining all of possible predictors, "age" gets the optimization.

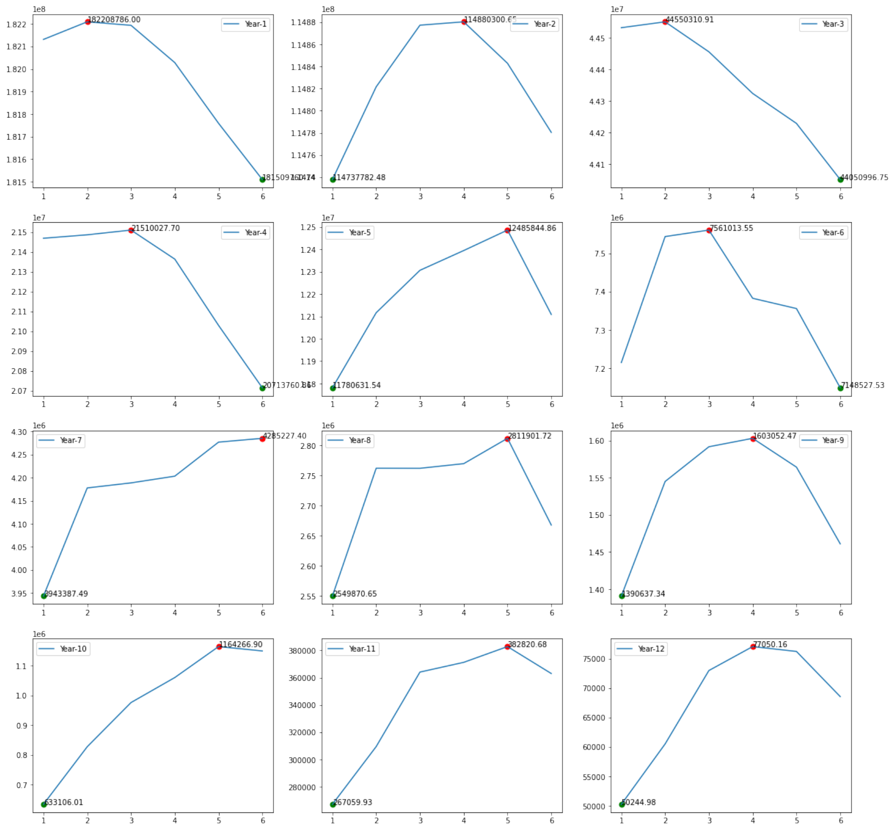

Figure 2. MSE per model

Python function **sklearn.tree.DecisionTreeRegressor.feature_importances_** is used toselect features for the model Decision trees, result shown in figure 4 In all 12 models, only the variable 'RepDel' always has low importance index (< 0.01), hence 'RepDel' was not considered in analysis.

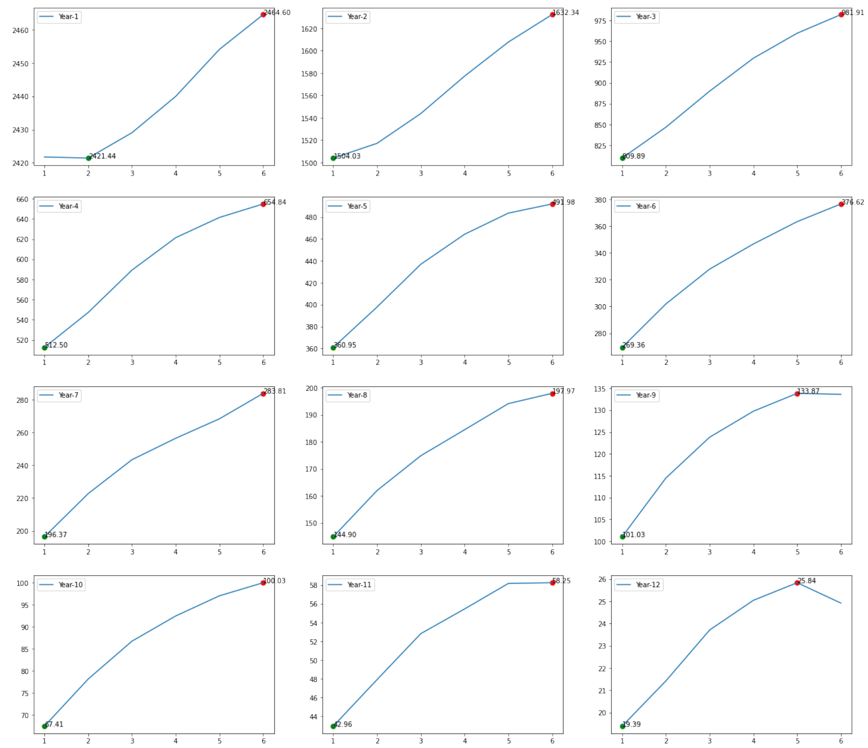

Figure 3. MAE per model

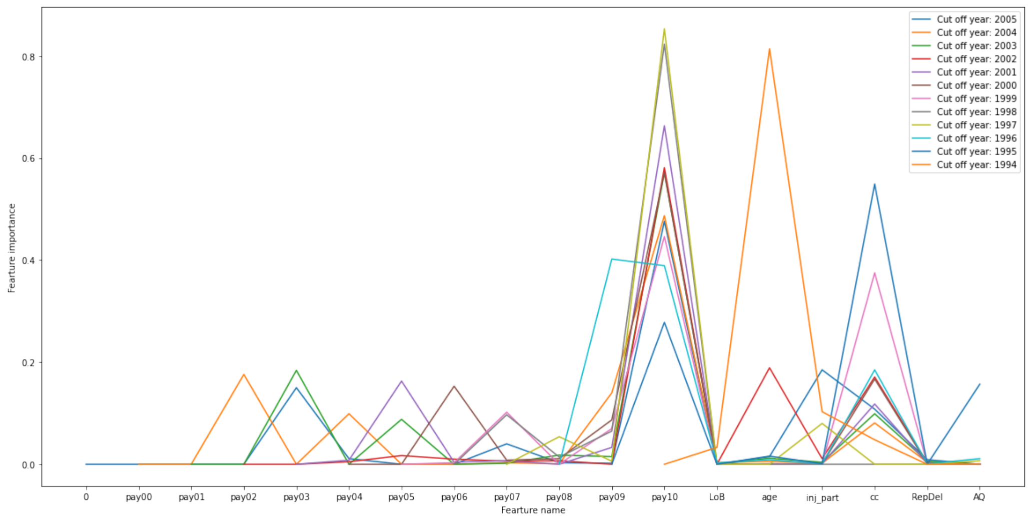

Figure 4. Decision Tree Feature selection

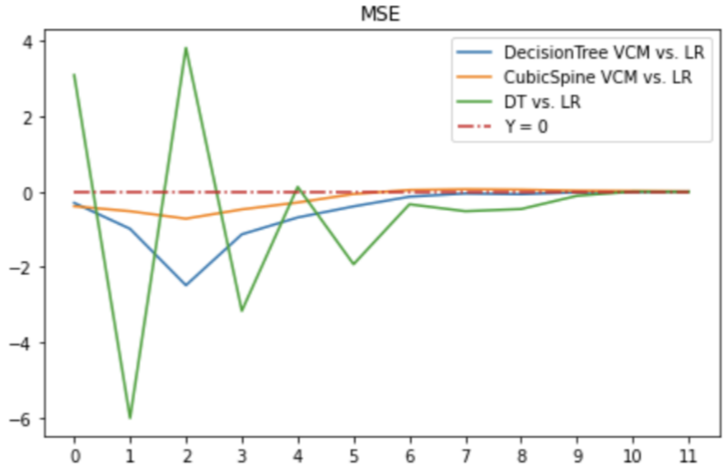

Figure 5. MSE of all methods

6. Results and Conclusion

When cross-validated applied models' fittings was performed on all data for 2014, for MSE, a measure of predictive performance of reserves for each year, it was observed that the normal linear model always had a larger bias than VCM. As for the powerful ML method decision tree, the two VCM also performed better in most of years. Overall, this project primarily prove in this case, two VCM always have less error than other linear regression. Compared to ML techniques, the VCM sometimes has fewer large-error estimated reserve values which would effect the reserve prediction and estimation in a serious way.

The reserve is the capital preparation drawn by the insurance company to guarantee the performance of the future compensation liability as promised. Scientific, reasonable and sufficient reserve withdrawal is the basis for a company to prevent risks, prudent operation and scientific management. Insufficient reserve withdrawal and irregular withdrawal process will cause hidden dangers to the stable operation and risk resistance of the insurance industry. Hence, after model fitting the 2014 data, the project made forecasts for the another data generated by simulation machine contains information who reported claims in 1995. The table 4 shows the average reserve forecasts for each subsequent 12-year insurance period predicted by models. Decision tree boosted and cubic spline VCM's predictions were closest to the real data in 9 of 12 years in total, between VCMs, cubic spline has more significant accurate predictions, and it also performed best when analyzing the absolute value of the total prediction error which means mislead insurance company to prepare payments for total loss in each year, it will help companies be easier to pass reserve exam. It can be observed that the linear regression model and the decision trees also performed not bad in general. In this study, outliers are still kept in dataset, since they are always important in actuarial calculations.

When focusing on the time consumed performance, the VCM models take longer time than traditional linear regression and ML technique we used in this study, especially decision tree boosted VCM needs significant more time. Although it is

Year	Real	DTVCM	LR	DTR	SplineVCM
1995	1788.79	1816.14	1862.73	1781.95	**1793.54**
1996	894.73	908.87	919.77	865.07	**902.61**
1997	420.77	410.86	381.99	445.21	**414.83**
1998	254.89	**251.30**	239.17	259.53	250.44
1999	172.16	**172.63**	167.26	164.73	172.76
2000	124.67	120.79	119.07	96.57	**122.76**
2001	89.40	80.77	**87.88**	64.69	84.70
2002	65.66	**65.38**	58.00	49.33	59.81
2003	46.31	37.26	41.43	34.87	**41.21**
2004	32.48	25.44	**28.07**	25.93	26.54
2005	18.87	15.65	15.65	15.17	**16.04**
2006	7.88	7.53	**7.76**	7.53	8.53
Sum absolute error	-	97.95	185.79	162.19	**50.65**

Table 4. Reserve predictions (per claim) for claims reported in 1995

composed of ML approach and linear regression, its numerical logic is more complex, making its running time even a hundred times more than that of other methods. Generally, more factors set in coefficients models takes longer time as shown in figure 6. It is a non-negligible drawback since this is for insurance companies to consider whether it is worth spending hundreds of times the time to make forecasts. Therefore it is more recommended to use decision tree boosted VCM when the sample size is not too large. As for cubic spline VCM, this more valuable to be used in aspects of more interpretibility of coefficients, a little more time taken but with obvious more accurate predictions.

For many industries, the rapidly developing AI technologies such as machine learning and deep learning can well complete the fitting, analysis and prediction of models. However, some characteristics of the actuarial industry cause a high dependence on probabilistic and statistical models, and inside unseen complex models perform like a 'Black Box' may have the risk of violating the requirements of regulations, the article [1] also was motivated by this fact. Usually more complicated means less interpretive, while facing linear regression model is lack of complexity and machine learning methods cannot satisfy the requirement for interpretibility, VCM handle these two problems well. In this project, when comes to decision tree boosted VCM, it can be explained that the reserve each year is still has linear relationship with factors given, however, the coefficients are varying dependent on 'Reported delayed years', for those observations with same 'Reported delayed years', the coefficients will be the same and follows a linear model. It has to be admitted that when the

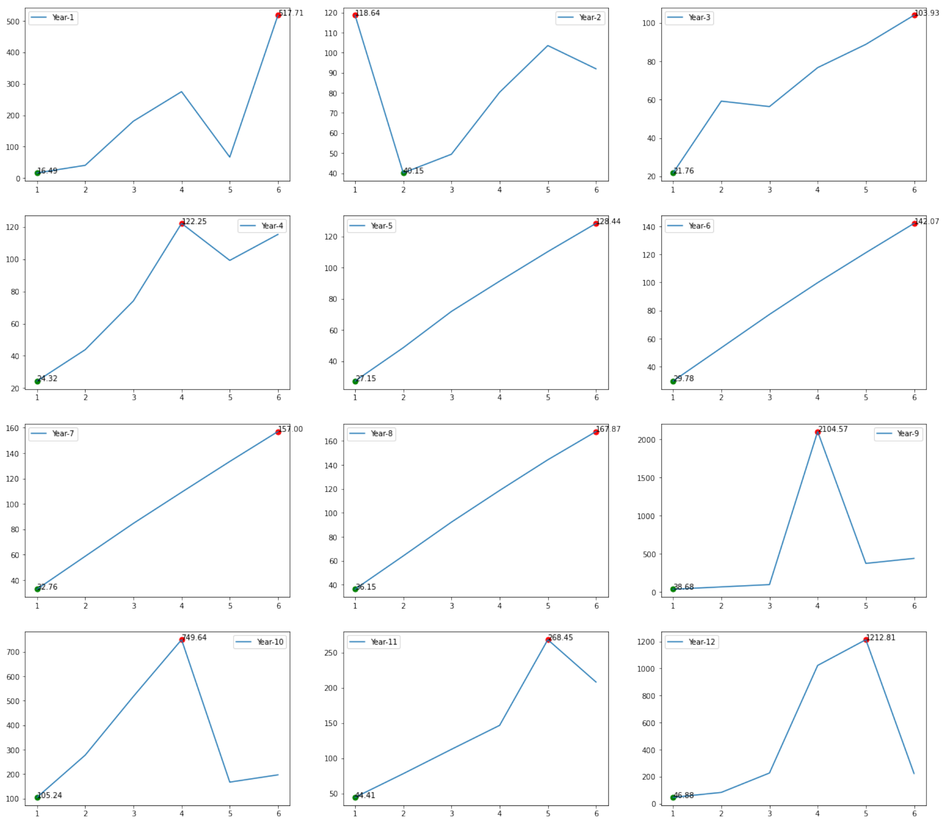

Figure 6. VCM model-Computational speed with different coefficients factors

data scale is huge, the efficiency of decision tree boosted VCM is relatively low. Figure7 displays each coefficient in Cubic spline VCM, when model predicting the first year reserve. We can observe that for each coefficient varying depends on scaled factor "age", not constants. . When we move to analyze cubic spline VCM, it's easy to be understood that coefficients are related to the factor policyholders' "age", it can be regarded as a infinite steps step-wise linear regression.

7. Future Study

The VCM provides a potential powerful model for actuaries to do analysis. According to the above analysis, it can be seen that one of the characteristics of VCM is very time-consuming. The article [8] developed a three-step divide-and-conquer Bayesian approach to improve the efficiency of VCM, this project can be considered to apply Markov chain Monte Carlo (MCMC) method to draw smaller sample and use in data augmentation (DA)-type algorithm to get both of MCMC parameters and predictions in parallel. The authors also developed a combination algorithm named Aggregated Monte Carlo (AMC) posterior, to aggregate estimations based on MCMC method. This method has approved a higher efficiency of VCM, the method applied in this project may be possible to combine with this technique and improve the fitting efficiency.

Besides Decision trees model, other machine learning approches like XGBoost, Ran-

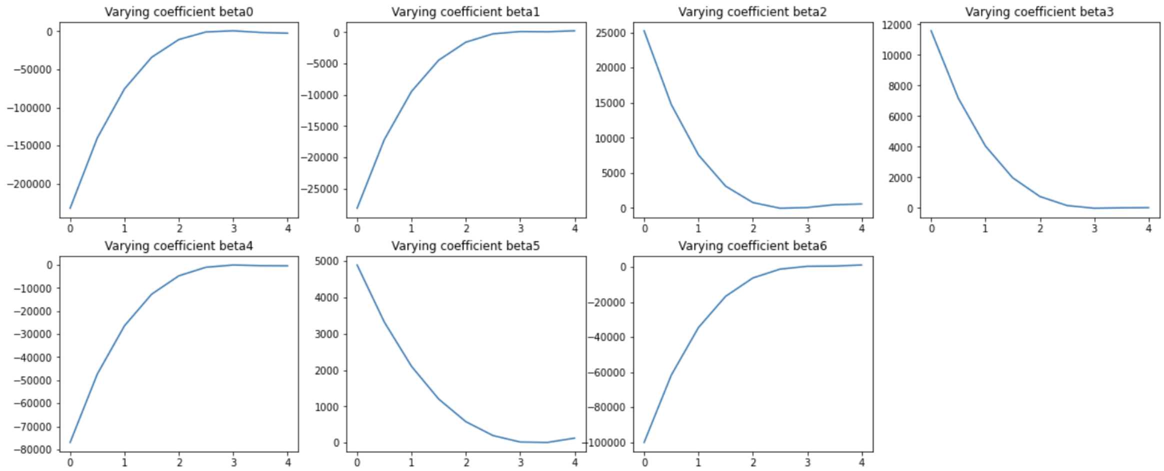

Figure 7. Each varying coefficient in Cubic spline VCM

dom Forrest can be employed to varying-coefficients model. Except for the change and development of model, the target variable can be separated to two things: claim status and reserve for each year, it can avoid the condition that 0 future payment increases the variance of target values which results in a bad fitting efficiency.

In recent years, deep learning has been in a high speed development period and is a rapidly growing field of AI techniques, it could be considered to serve as coefficient model as well. Additionally, Natural language processing (NLP) a powerful text analysis tool, in Xu's research, he confirmed that NLP such as Bert can well help actuaries analyze textual data [21]. Thus, NLP such as BERT technique has potential to be combined with VCM, aims to take use of the contextual semantics that can be captured by BERT and the complicated, non-linear correlations between input data and the target variable that can be captured by VCM. The designed model can be expressed:

$$y = \beta_0(\textbf{text}) + X_1\beta_1(\textbf{text}) + ... + X_p\beta_p(\textbf{text}) \tag{22}$$

What is more, Wang [19] proposed a spatial data distributed over two-dimension domains statistics test procedure based on empirical likelihood approach to asymptotic distribution of VCM function. Although in most cases actuarial data is related to finance, in measures such as Catastrophe Modeling, Health Insurance and Geographical Health Patterns, Spatial Risk Assessment for Insurance, Spatial risk assessment for insurance, spatial data with two-dimension will be used to combine geographical location (characteristics) with loss or risk measurement to do estimation. As long as we can get this type of data, the approach mentioned in this paper can be developed and get a good performance even though the data set is small, and the topic can be extended with VCM combined with ML techniques.

References

[1] Muhammed Taher Al-Mudafer, Benjamin Avanzi, Greg Taylor, and Bernard Wong. Stochastic loss reserving with mixture density neural networks. *Insurance: Mathematics and Economics*, 105:144–174, 2022.

[2] Wanghuan Chu, Runze Li, and Matthew Reimherr. Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data. *The annals of applied statistics*, 10(2):596, 2016.

[3] Jianqing Fan and Wenyang Zhang. Statistical estimation in varying coefficient models. *The annals of Statistics*, 27(5):1491–1518, 1999.

[4] Jianqing Fan and Wenyang Zhang. Statistical methods with varying coefficient models. *Statistics and its Interface*, 1(1):179, 2008.

[5] Thomas B Fomby, Stanley R Johnson, R Carter Hill, Thomas B Fomby, Stanley R Johnson, and R Carter Hill. Varying coefficient models. *Advanced Econometric Methods*, pages 307–323, 1984.

[6] Edward W Frees, Peng Shi, and Emiliano A Valdez. Actuarial applications of a hierarchical insurance claims model. *ASTIN Bulletin: The Journal of the IAA*, 39(1):165–197, 2009.

[7] Andrea Gabrielli and Mario V. Wüthrich. An individual claims history simulation machine. *Risks*, 6(2):29, 2018.

[8] Rajarshi Guhaniyogi, Cheng Li, Terrance D Savitsky, and Sanvesh Srivastava. Distributed bayesian varying coefficient modeling using a gaussian process prior. *The Journal of Machine Learning Research*, 23(1):3642–3700, 2022.

[9] Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 55(4):757–779, 1993.

[10] Jianhua Z Huang and Haipeng Shen. Functional coefficient regression models for non-linear time series: a polynomial spline approach. *Scandinavian journal of statistics*, 31(4):515–534, 2004.

[11] Jianhua Z Huang, Colin O Wu, and Lan Zhou. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1):111–128, 2002.

[12] Kevin Kuo and Daniel Lupton. Towards explainability of machine learning models in insurance pricing. *arXiv preprint arXiv:2003.10674*, 2020.

[13] Jingyuan Liu, Runze Li, and Rongling Wu. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. *Journal of the American Statistical Association*, 109(505):266–274, 2014.

[14] Brian D Marx. P-spline varying coefficient models for complex data. In *Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir*, pages 19–43. Springer, 2009.

[15] Anthony J Myles, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown. An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6):275–285, 2004.

[16] Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. Vcnet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861*, 2021.

[17] Society of Actuaries. Reserve. https://www.soa.org/4a537f/globalassets/assets/files/edu/actuarial-glossary.pdf, 2022. [Online; accessed 16-August-2022].

[18] Byeong U Park, Enno Mammen, Young K Lee, and Eun Ryung Lee. Varying coefficient regression models: a review and new developments. *International Statistical Review*, 83(1):36–64, 2015.

[19] Shuoyang Wang, Honglang Wang, Yichuan Zhao, Guanqun Cao, Yingru Li, et al. Empirical likelihood ratio tests for varying coefficient geo models. *Statistica Sinica*, 33(4), 2021.

[20] Xufei Wang, Bo Jiang, and Jun S Liu. Varying coefficient model via adaptive spline fitting. *Journal of Computational and Graphical Statistics*, 33(2):614–624, 2024.

[21] Shuzhe Xu, Chuanlong Zhang, and Don Hong. Bert-based nlp techniques for classification and severity modeling in basic warranty data study. *Insurance: Mathematics and Economics*, 107:57–67, 2022.

[22] Yichen Zhou and Giles Hooker. Decision tree boosted varying coefficient models. *Data Mining and Knowledge Discovery*, 36(6):2237–2271, 2022.