Predicting Insurance Reserves using Varying Coefficients Model with Machine Learning

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

August, 2023

Content

Introduction

Since 1980s, actuaries always choose Generalized Linear Models (GLM) serve as a foundation for analysis projects to help insurance companies predict premium or unearned premium risk (loss) at a particular time spot. However, classical linear models can often fall short in capturing complex relationships present in real-world scenarios, use of linear regression models in such cases often leads to a violation of assumptions.

- Fan and Zhang (2008) [4] mentioned that when analyzing infant mortality in China, many influencing factors were not considered to change over time, which is contrary to reality.

Introduction

For example: consider the mortality rate of a population over time and attempt to use linear regression to predict mortality rate based on a single predictor variable, such as the population size:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Y: Mortality rate; X: Population size; β_0: Intercept (constant); β_1: Coefficient for the population size (slope); ϵ: Error term

Violation of assumptions: mortality rate may not change linearly with population size. In the early days of New China, a small population may have a high mortality rate due to limited resources, citizens were encouraged to have children, it caused a more infants died since no enough resources. But as time going by, the population grows larger, the mortality rate might decrease due to improved access to healthcare and other resources.

Introduction - motivation

ASOP (Actuarial Standard of Practice) No.41:
Requirements for Actuarial Communications—The performance of a specific actuarial engagement or assignment typically requires significant and ongoing communications between the actuary and the intended users regarding the following: the scope of the requested work; the methods, procedures, assumptions, data, and other information required to complete the work; and the development of the communication of the actuarial findings.

In general, while pricing insurance, insurers need to be able to explain to regulators, customers and other stakeholders how the model makes premium decisions. A transparent model also promotes trust and fairness (Kuo 2020). [11]

Introduction - motivation

Actuaries' reluctance to use machine learning (ML) techniques may stem from regulatory requirements, such as interpretability, More complicated means less interpretability. Advanced ML or deep learning approaches perform like a complete "Black Box".

Improvements:

- 2020 Kuo, Lupton — Conceptualize interpretability in terms of the ability of a model, offer potential model-agnostic techniques to solve it.
- 2022 Al-Mudafer et al. — Mixture density neural networks.
- In this research, varying coefficients regression model tries to achieve greater realism by allowing coefficients in classical linear models to be dependent on a set of covariates, also keep the easy understanding of GLM.

Introduction-Starting example

Programming language: R studio
Package.function: tvReg.tvLM(time-varying coefficients regression models)
Defined variables:

- τ(time): sequence(0.001,1)
- $\beta_0 = \tau^2$, $\beta_1 = log(\tau)$, $\beta_2 = 2^\tau$
- $X_1 = rnorm(1000), X_2 = rchisq(1000, df = 4)$
- $\epsilon = rt(1000, df = 10)$

Linear model(Model1):

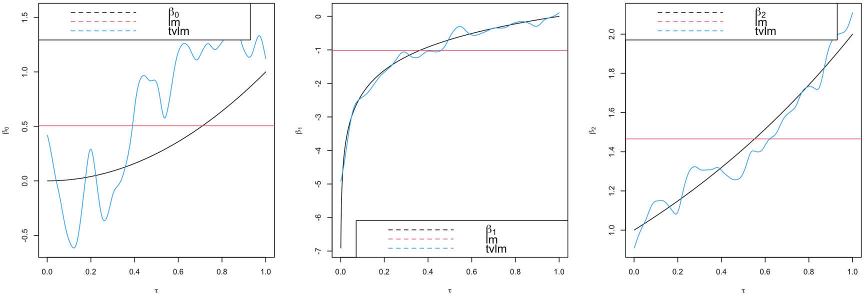
$$y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$$

Time-varying coefficients linear model(Model2):

$$y = \beta_0(\tau) + \beta_1(\tau) * X_1 + \beta_2(\tau) * X_2$$

Introduction-Starting example

Coeffcients(β) plots: real (black) vs. tvLM (blue) vs. lm (red)



Model	MSE (mean squred error)
lm	58.98103
tvLM	25.54048

If coefficients have potential relation to other factors, varying-coefficient models are necessary to be set.

Introduction: Varying-coefficients model

A random variable Y whose distribution depends on

- a parameter η,
- predictors $X_1, X_2, ..., X_p$ and $R_0, R_1, R_2, ..., R_p$

A varying-coefficients model has the form:

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

Applied to generalized linear models, express Y function via link function (common: normal, Poisson, or NB distribution)

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

$g(.)$ is a link function.

Introduction: Varying-coefficients model

Development

- Stanford University statisticians Hastie and Tibshirani (1993) [9] first introduced varying coefficient models (VCM).
- Princeton University researchers Fan and Zhang have made significant advancements in the field of VCM, methodologically, theoretically, and practically.[3][4]
- Penn State Runze li explored more application of VCM.[2] [12]
- University of Chicago, Nie et al.[14] developed VCM applied to the neural network.
- Zhou and Hooker from Cornell University [20] proposed a tree boosted varying-coefficients model.

After learning and reading a lot of papers, I found the VCM can improve interpretability problem well, and no actuaries or scholars have applied it to the actuarial data so far. This research is the first to introduce VCM with ML technique to actuarial science.

Introduction: Varying-coefficients model

- Generalized additive models
  - If $\beta_j(R_j) = \beta_j$, a constant function, usual a linear model or a GLM.
  - If $X_j = c$ (say c $= 1$), the j^{th} term is $\beta_j(R_j)$.
- Dynamic generalized linear model
  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 $\{t_1, t_2, ..., t_n\}$:

$$\eta_t = \beta_0(t) + X_1(t)\beta_1(t) + ... + X_p(t)\beta_p(t) \tag{3}$$

- Non-proportional hazards model
  Proportional hazards model(Cox model):

$$\lambda(t|X_1...X_p) = \lambda_0(t)\exp\{\sum_j X_j\beta\} \tag{4}$$

$\lambda_0(t)$ is an arbitrary base-line hazard function.

$$\text{Develop to } \lambda(t|X_1...X_p) = \lambda_0(t)\exp\{\sum_j X_j\beta_j(t)\} \tag{5}$$

- Time-series model
  Given time series $\{X_t\}$, applied varying-coefficient form:

$$X_t = \beta_0(X_{t-p}) + \beta_1(X_{t-p})X_{t-1} + ... + \beta_k(X_{t-p})X_{t-k} + \epsilon_t, \quad (6)$$

  for some given lags k and p.
- Semi-varying coefficients model

$$y = X_1^T \beta_1(R) + X_2^T \beta_2 + \epsilon \quad (7)$$

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

Introduction: Varying-coefficients model

For these cases, many methods can be used to model unspecified function $\beta_j(R_j)$:

- Polynomials
- Fourier series
- Piece-wise polynomials
- More 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, decision tree boosted is applied to varying-coefficients model.

Algorithm

Reference: Zhou, Y., Hooker, G. Decision tree boosted varying coefficient models. Data Min Knowl Disc 36, 2237–2271 (2022). [20]
Source code: https://github.com/sply88/vcboost/tree/master

- Numerical method:
  - Gradient Boosting: The core numerical method used in this implementation. It is an ensemble technique that builds multiple learners (decision tree) sequentially, where each tree corrects the errors made by the previous ones. It involves iteratively fitting new models to the negative gradients of the loss function to minimize the loss.
  - Decision Trees: It is used as base learners. Decision trees is a non-linear supervised learning algorithm used for both classification and regression tasks. Here, regression trees are employed to approximate the negative gradients for each coefficient.

Algorithm

- - Least Squared Regression Tree: It is a decision tree built using the mean squared error (MSE) as the splitting criterion. This method helps approximate the negative gradients more accurately for updating the coefficients.

$$\min_{j,s}\left[\min_{c_1}\sum_{x_i\in R_1(j,s)}(y_i-c_1)^2+\min_{c_2}\sum_{x_i\in R_2(j,s)}(y_i-c_2)^2\right] \tag{8}$$

  - Line Search: This technique is used 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.

Algorithm: VCBooster

Input	Description
X	Design matrix for parametric coefficients
Z	Design matrix of effect modifiers
y	Dependent variable
`n_stages`	Number of boosting stages
`learning_rate`	Learning rate for updating coefficients
`max_depth`	Max depth of base learners
`min_samples_leaf`	Minimum number of samples required at a leaf node
`splitter`	Splitting strategy for internal nodes
`mini_updates`	Option to update predictions, residuals, and gradients
`verbose`	Option to print progress

Table: Input

Algorithm: VCBooster

Output: Trained `VCBooster` model with ensembles of decision trees.

1. Initialize `VCBooster` with provided hyper-parameters and attributes.
2. Validate and preprocess input data.
3. Initialize variables and lists to store ensembles and training losses.
4. Start the training loop for `n_stages` boosting stages.
    - For each boosting stage, iterate through each coefficient in a randomized order.
    - Update current prediction, residuals, and negative gradient if required.
    - Approximate negative gradient using a least squares regression tree.
    - Adjust the leaf values based on the `line_search_strategy`.
    - Add the trained decision tree to the ensemble of the current coefficient.
    - Update the coefficient values based on the trained decision trees.
    - Repeat the process for `n_stages` iterations.
5. After training is completed, the trained `VCBooster` model is returned.

Data

Auto insurance claims data

Observations: 40023 policy claims

Resource from: https://github.com/kasaai/bnn-claims

This data set generated data using the individual claims history simulator (Gabrielli and V Wüthrich 2018)

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 year after accident	numeric

Table: Original data set information

Data pre-processing

Programming language: Python.
Data new variables created:

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

- All categorical variables use one-hot encoded.
- Transformation:
  - age: log(age)
  - inj_part: sqrt(inj_part)
  - pay_year(pay00-pay11): normalization scaled.

Data pre-processing

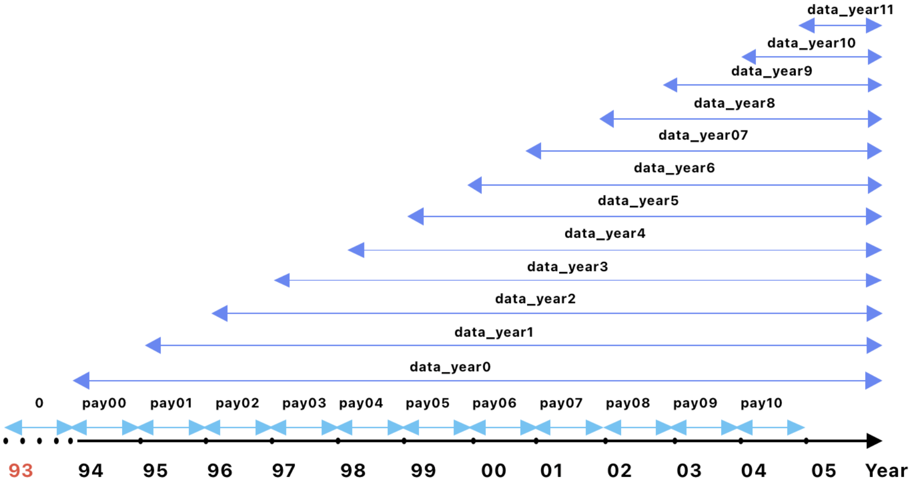


Figure: Variable explanation

Data pre-processing

Fixed variables:'LoB','age','inj_part','cc','RepDel','AQ'

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: Training process

Cross-validation: kfold = 5

- Training data: 80%, randomly chosen
- Test data: 20%
- For each observation of training data, 12 models are fitted. Then test data will do prediction for 12 models respectively.

Case study

The purpose of this project is to train the insurance loss reserve model by selecting factors in the data, and to see whether the performance of the varying-coefficients model in the actuarial data and its application make sense through the ordinary constant coefficient linear model, the variable coefficient model, and the common machine learning model.

- ***Reserve:** Liability set aside to pay future claims. (SOA glossary [15])
- ***Outliers:** For insurance company, there is always a risk to accept huge-payment claim, the outliers which are rejected may effect the reserve in a serious way, and may result in huge estimation error for preparing reserve and cause company financial issue. This project keeps all observations.

Case study: Maximum Likelihood Estimation (MLE)

Define the log-likelihood function, where θ are parameters:

$$\log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta) \tag{9}$$

Set the derivatives to zero:

$$\frac{\partial}{\partial \theta_j} \log L(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta_j} \log f(x_i|\theta) = 0 \tag{10}$$

Estimation for $\hat{\theta}$:

$$\hat{\theta} = \arg\max(\log L(\theta)) = \arg\min(-\log L(\theta)) \tag{11}$$

Case study: varying-coefficients model

$$\text{Mean squared error(MSE)} = \frac{1}{n}\sum(y_i - \hat{y}_i)^2$$
$$\text{Mean aboslute error(MAE)} = \frac{1}{n}\sum|y_i - \hat{y}_i|$$

- If the outliers represent anomalies that are important in the business and need to be detected, the MSE loss function should be selected. In contrast, if only outliers are treated as damaged data, the MAE loss function 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.
- This project choose MSE as evaluation for comparison, MAE as a reference metrics.

Case study

Package.function: **boost.VCBooster**

$$y = \beta_0(\mathbf{Z}) + X_1\beta_1(\mathbf{Z}) + ... + X_p\beta_p(\mathbf{Z}) \tag{12}$$

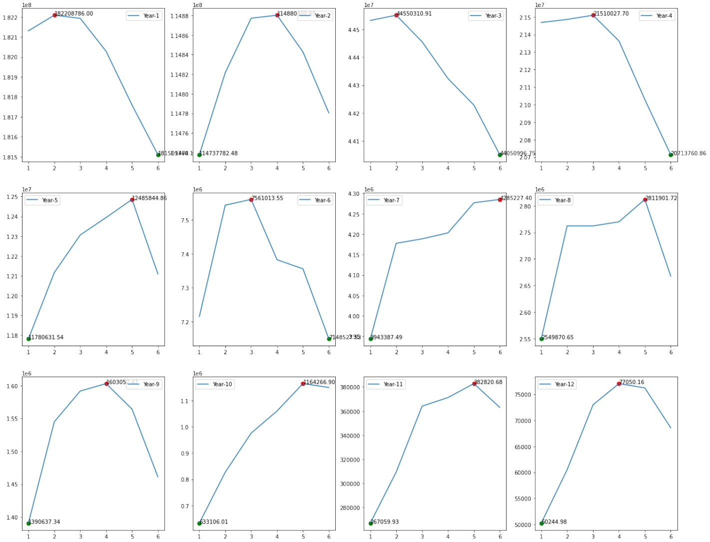
$\beta(\mathbf{Z})$ is a vector contains p+1 gradient boosting decision tree models.

Steps to select varying-coefficients model's factors ($\mathbf{Z}$):

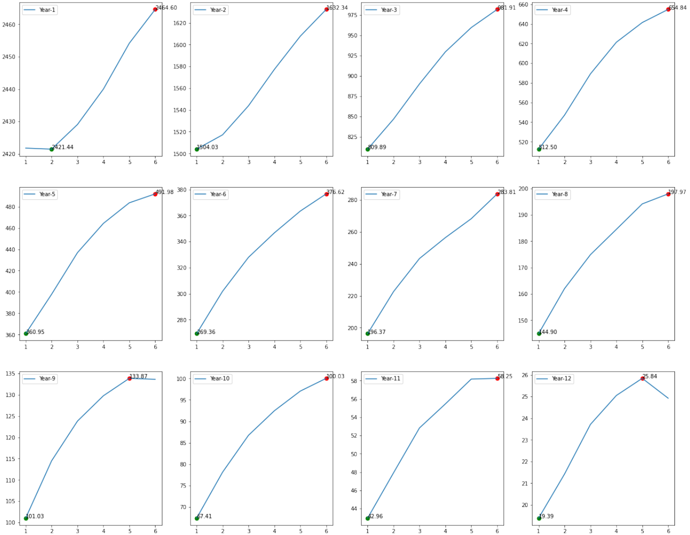
- Totally there are 6 factors, n_col=[0:6].
- Fit models of all possible Z vector:
  Example: when factor number is 3, try:
  ['LoB','age','inj_part'],['LoB','inj_part','cc'],['LoB',' age','cc']...
- Choose the combination with the lowest MSE (and MAE if possible).

The next two pages show the MSE and MAE plot respectively.

Case study: varying-coefficients model (MSE)



Case study: varying-coefficients model (MAE)



Case study: varying-coefficients model

- X: 'LoB',' age', 'inj part', 'cc', 'RepDel', 'AQ', 'pay_year[:,0:i](i = 1, 2,...,12)'
- Z: 'RepDel'
- y: 'data_year[:,i]'(i = 1, 2,...,12)
- Parameter setting:

Parameter	Settings
learning_rate	0.1
min_samples_leaf	1
max_depth	5
n_stages	1000
others	default

Table: VCBooster optimal parameter

Case study: Robust (Huber) Regression model

Since the outliers are not removed from the models, compare to Robust Huber Regression model which can handle outliers better, and see if the outliers worsen the linear model.

- **sklearn.linear_model.HuberRegressor**
- Inside Huber regression model, Huber's loss take place of least squares loss to find the parameter θ. Huber loss:

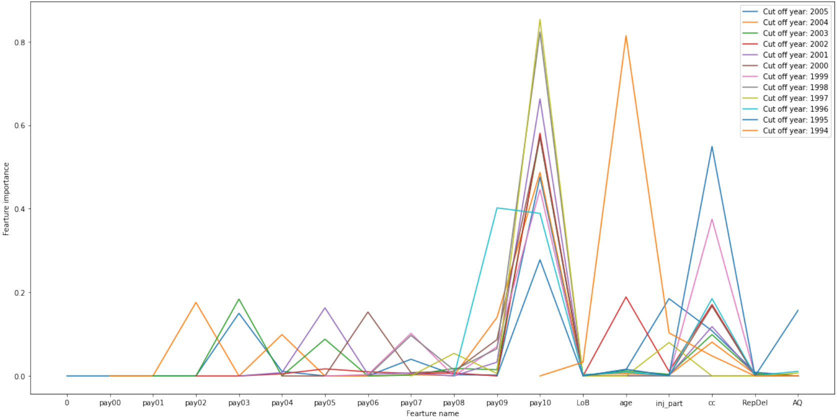
$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{if } |y - f(x)| \leq \delta \\ \delta \cdot |y - f(x)| - \frac{1}{2}\delta^2 & \text{if } |y - f(x)| > \delta \end{cases} \tag{13}$$

  $\delta > 0$, smaller value means less sensitive to outliers. It can handle outliers through paying less penalty to large residuals (residuals greater than δ: $|y - f(x)| \leq \delta$)

Case study: Decision trees

sklearn.tree.DecisionTreeRegressor

Calculate feature importances:



In all 12 models, only the variable **"RepDel"** always has low importance index(< 0.01), delete **"RepDel"**.

Comparison

Metrics difference = Metrics(Other methods) - Metrics(VC)

- Red line: no difference

Above the red line: worse performance than VC model.
Below the red line: better performance than VC model.

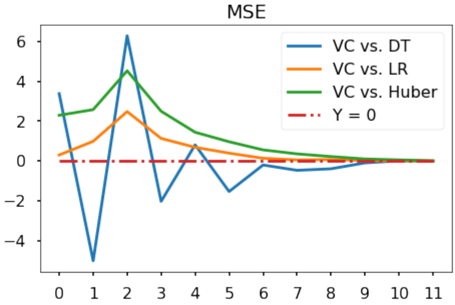


Figure: MSE difference

Comparison

The reserve is the capital preparation drawn by the insurance company to guarantee the performance of the future compensation liability as promised.

- 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.

After model fitting the 2014 data, the project made forecasts for the another data generated by simulation ma- chine contains information who reported claims in 1995.

Comparison: prediction on claims in 1995

Test on data set: claims reported in year 1995.

- Training data: reported claims in 1994.
- Test data: reported claims in 1995.
- Compare: Estimated reserves in each of following 12 develop years.

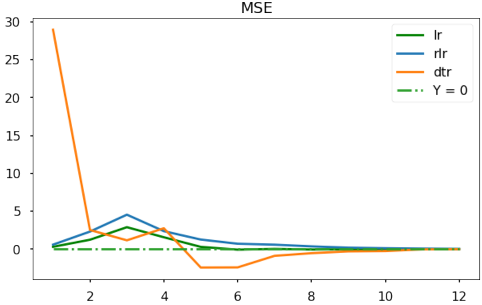


Figure: Prediction mse

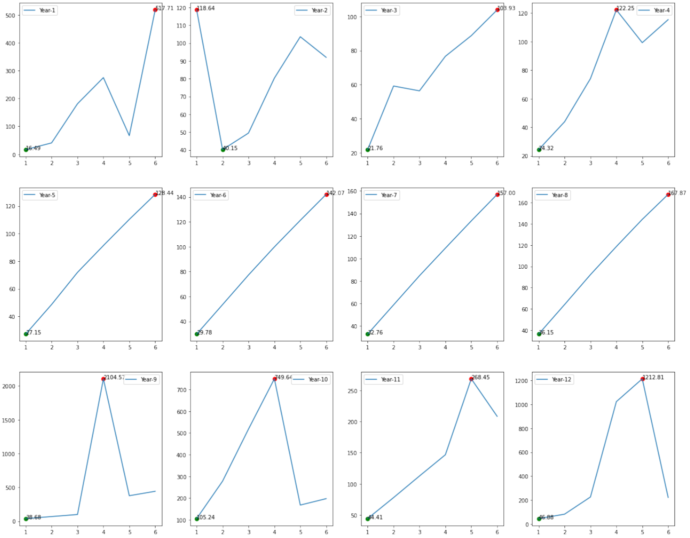
Comparison: prediction on claims in 1995

Year	Real	VCM	LR	DTR	RLR
1995	1788.79	1816.14	1862.73	**1781.95**	1322.37
1996	894.73	**908.87**	919.77	865.07	91.75
1997	420.77	**410.86**	381.99	445.21	0.0003
1998	254.89	**251.30**	239.17	259.53	7.29
1999	172.16	**172.63**	167.26	164.73	7.31
2000	124.67	**120.79**	119.07	96.57	-6.36
2001	89.40	80.77	**87.88**	64.69	-1.95
2002	65.66	**65.38**	58.00	49.33	4.93
2003	46.31	37.26	**41.43**	34.87	3.06
2004	32.48	25.44	**28.07**	25.93	-7.17
2005	18.87	**15.65**	**15.65**	15.17	-0.0001
2006	7.88	7.53	**7.76**	7.53	-3.455
Sum absolute error	-	**97.95**	185.79	162.19	2502.53

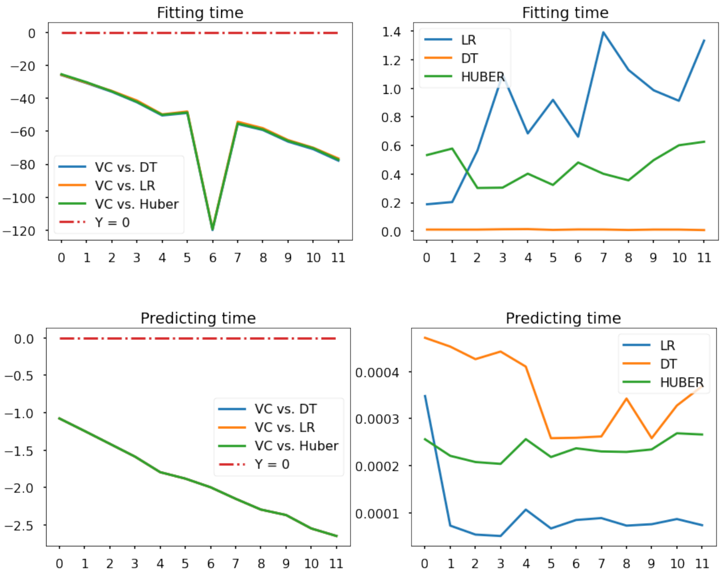
Table: Reserve (per claim) predictions for test data

Comparison: computational speed

Mean computational speed for 6 number of factors per year model.



Comparison: computational speed



Conclusion

- **Evaluation metrics:**
  For the reserve estimation on 12 evaluation date, linear regression and robust regression models always performed relatively worse than VCM, some of years, VCM is even better than Decision tree.
- **Prediction:**
  - During 12 development years, in 7 years, VCM gives the closet estimated reserve per claim to the real data. Also gives the smallest sum of absolute error for 12 years' prediction.
  - Robust linear regression performed worst with large amount of prediction residual. It approves the outliers contains important information in actuarial data.

Conclusion

- **Interpretability**:
  Loss reserve in 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.
- **Computational speed:**
  - Generally, more factors in coefficients model takes longer time.
  - ML technique takes least fitting time, linear regression model takes least predicting time.
  - Although the varying-coefficients model is composed of decision tree model and linear regression, its numerical logic is more complex, making its running time even a hundred times more than that of other methods.
  - It is recommended to use when the sample size is not too large.

Conclusion

- For many industries, the rapidly developing ML and DL can well complete the fitting, analysis and prediction of models well. However, some characteristics of the actuarial industry cause a high dependence on probabilistic and statistical models, and inside unseen complex models may have the risk of violating the requirements of regulations. Thus, the VCM has the use value by freeing actuaries from limited GLM.

  It has to be admitted that when the data scale is huge, its efficiency is relatively low. It could be the next study stage of improvement of VCM.

Future study

- Improve time-consuming. Article [8] developed a three-step divide-and-conquer Bayesian approach:
  1. Apply Markov chain Monte Carlo (MCMC) method to draw smaller sample.
  2. Use in data augmentation (DA)-type algorithm to get both of MCMC parameters and predictions in parallel.
  3. Use 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.
- Other approches like XGBoost, Random Forrest, neural networks can be employed to varying-coefficients model.

Future study

- Combine BERT technique with varying-coefficient model, 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:

$$y = \beta_0(\textbf{text}) + X_1\beta_1(\textbf{text}) + ... + X_p\beta_p(\textbf{text}) \tag{14}$$

- Wang [17] proposed a spatial data distributed over two-dimension domains statistics test procedure based on empirical likelihood approach to asymptotic distribution of VCM function. When we get this type of data, then extend the topic with VCM combined with ML techniques.

[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] Peter J Huber.
Robust estimation of a location parameter: Annals mathematics statistics, 35.
Ji, S., Xue, Y. and Carin, L.(2008),'Bayesian compressive sensing', IEEE Transactions on signal processing, 56(6):2346–2356, 1964.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] Shuzhe Xu, Salvador E Barbosa, and Don Hong.
Bert feature based model for predicting the helpfulness scores of online customers reviews.
In *Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC), Volume 2*, pages 270–281. Springer, 2020.

[19] 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.

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