2022 Actuarial Colloquia

A New Framework of Prediction Error Decomposition for the Machine Learning Era

June 21st, 2022

Kazuki Kuriyama (The Institutes of Actuaries of Japan)
Masafumi Suzuki (Munich Re, Japan Branch)
Hirokazu Iwasawa (Waseda University)

Agenda

Ⅰ. Introduction

Ⅱ. A New Framework of Prediction Error Decomposition

Ⅲ. Development of prediction error decomposition tool

Ⅳ. Summary and next steps

References

Ⅰ. Introduction

Initiatives of the Institute of Actuaries of Japan on data science

- Since 2017, IAJ (The Institutes of Actuaries of Japan) have conducted a full-scale study of data science.

- In 2019, IAJ established Data Science Related Basic Research Working Group to enhance IAJ members‘ knowledge and skills on data science.

- This presentation is based on the work of the WG “prediction errors in predictive modeling“.

- The scope of this research are as follows

  ■ Study of general methodology of error decomposition (regardless of model type)
  ■ Development of a tool for error decomposition (R package, etc.)

Ⅰ. Introduction

Prediction Error Overview

Prediction Error
= Process Error
+ Parameter Error (*)
+ Model Error

(*) does not include the risk of future changes of the probability law (e.g. ones caused by regulation changes)

Process Error

Error due to the contingencies even if the predictive model is true

Parameter Error

Error in parameter estimation using limited data

Model Error

Errors caused by the difference from the true model

<Illustration of each error>

① Training Data

No.	Age	Vehicle type	Region	Color of driver license	Claim (USD)
1	20	Private	City A	Green	0
2	50	Private Small	City B	Gold	5,000
3	40	Truck	City C	Blue	100

② Predictive Model

Machine Learning
GLM

Parameter Error

Model Error

③Risk Factor

40 years old, Private Car, City A, Gold

Process Error

④ Precited Value

USD 300

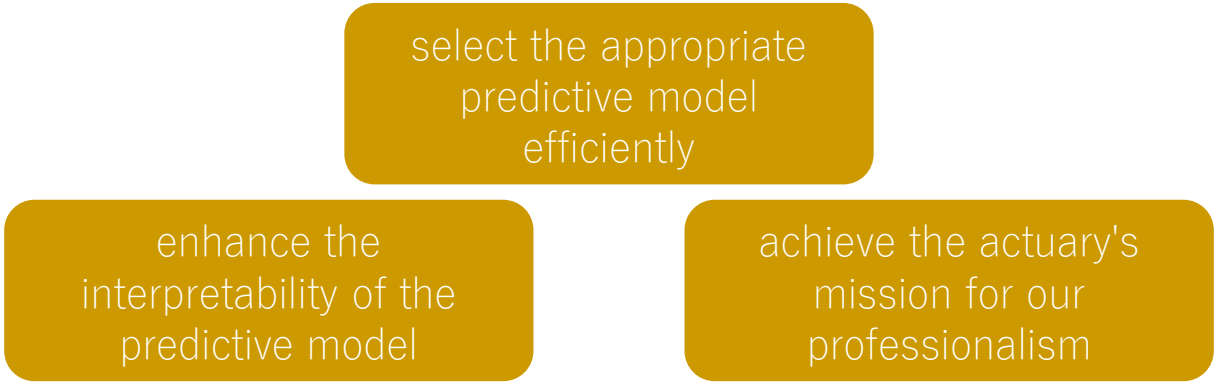
⑤ Actual

(USD)
0
100
10,000
0

I . Introduction

Motivation of our study

- The WG believes that it is worthwhile to construct the general framework of the prediction error decomposition and develop a general tool to …



- On the other hand,
  - there seems to be little research of general error decomposition methodology applicable even to machine learning models.
  - packages such as R and Python do not currently provide general tools to decompose the prediction errors

I . Introduction

Previous research and our study

- Previous research: Casualty Actuarial Society (2015), McGuire et al. (2021), Taylor et al. (2016), and others for more details.

$$\mathrm{MSEP}[\hat{y}] = \mathrm{E}[(\hat{y} - y)^2] = \boxed{(\mathrm{E}[\hat{y}] - \mathrm{E}[y])^2} + \boxed{\mathrm{E}[(\hat{y} - \mathrm{E}[\hat{y}])^2]} + \boxed{\mathrm{E}[(\mathrm{E}[y] - y)^2]}$$

Model error　　Parameter error　　Process error

- In order to achieve the goal, we seek to develop a "framework of the prediction error decomposition for the machine learning era“

- In this study, we
  - propose a new framework for prediction error decomposition by
    - proposing a "general concept of prediction error decomposition”
    - showing examples of tentative estimation methods for each error
  - prototype a tool of prediction error decomposition based on the framework

II . A New Framework of Prediction Error Decomposition

Process to build a new framework

- We intend to provide the following features for the framework of the general prediction error decomposition.

	Feature of the framework	Condition to be assumed	Approach
In Formulation	#1: Applicable to various predictive models, including machine learning ones	Assume a prediction function whose argument is a feature vector	Formulate the predictive model as a function of feature vectors
	#2: Estimating process errors without presuming the error distribution implied by the predictive model	Not presume the error distribution implied by the predictive model	Formulate it in a data-driven manner instead
	#3: Clarifying the independence between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$	Consider that $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ depend on the common feature vector	Formulate it as conditional independence under the condition of the feature vector
In Estimation	#4: Model error estimation method applicable to various predictive modeling method	Not presume specific predictive models to estimate model errors	Consider estimating model errors based on data

II . A New Framework of Prediction Error Decomposition

Definitions

- Definitions of variables and functions are as follows.
  - Feature vector: $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$
  - Target variable: $y_i = f(\mathbf{x}_i) + \varepsilon_i$
    Here, $\mathrm{E}[\varepsilon_i|\mathbf{x}_i] = 0$, $\mathrm{V}[\varepsilon_i|\mathbf{x}_i] = \sigma_\varepsilon^2(\mathbf{x}_i)$ ($y_1, y_2, \ldots$: mutually independent), and $\mathrm{E}[y_i|\mathbf{x}_i] = f(\mathbf{x}_i)$
  - Training data: $\mathcal{T}_n = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ (Stationarity Assumption)
  - Feature vector of the object to be predicted: $\mathbf{x}_{\mathbf{new}}$
  - Target variable of the object to be predicted: $y_{\mathbf{new}}$
  - Error in prediction: $\varepsilon_{\mathbf{new}} = y_{\mathbf{new}} - f(\mathbf{x}_{\mathbf{new}})$ Here, $\varepsilon_{\mathbf{new}}|\mathbf{x}_{\mathbf{new}} \perp\!\!\!\perp \mathcal{T}_n$
  - $\widehat{f_n}$: prediction function (returns the predicted value when a feature vector is input).
    Prediction here is based on the model created by $\mathcal{T}_n$.
    Mathematically, $\widehat{f_n}(\mathbf{x})$ is a $\sigma(\mathcal{T}_n)$-measurable random variable for fixed $\mathbf{x} \in \mathbb{R}^p$.
  - Conditional expectation of the predicted value under the condition of a feature vector $\mathbf{x}_{\mathbf{new}}$:
    $\mathcal{E}_n(\mathbf{x}_{\mathbf{new}}) \coloneqq \mathrm{E}[\hat{f}_n(\mathbf{x}_{\mathbf{new}})|\mathbf{x}_{\mathbf{new}}]$
    Note that it depends on the sample size n but is independent of the training data $\mathcal{T}_n$. It is the same as $\mathrm{E}[\hat{f}_n(\mathbf{x}_{\mathbf{new}})]$ if $\mathbf{x}_{\mathbf{new}}$ is fixed.

II . A New Framework of Prediction Error Decomposition

General formula of prediction error decomposition

- General formula is derived as follows. Here, Loss function is square loss.

$$\mathrm{Err}_n(\mathbf{x}_{\mathrm{new}}) \coloneqq \mathrm{E}\left[\left(y_{\mathrm{new}} - \hat{f}_n(\mathbf{x}_{\mathrm{new}})\right)^2 \middle| \mathbf{x}_{\mathrm{new}}\right]$$

$$= \mathrm{E}\left[\left\{(y_{\mathrm{new}} - f(\mathbf{x}_{\mathrm{new}})) + (f(\mathbf{x}_{\mathrm{new}}) - \mathcal{E}_n(\mathbf{x}_{\mathrm{new}})) + (\mathcal{E}_n(\mathbf{x}_{\mathrm{new}}) - \hat{f}_n(\mathbf{x}_{\mathrm{new}}))\right\}^2 \middle| \mathbf{x}_{\mathrm{new}}\right]$$

$$= \mathrm{E}\left[\left(y_{\mathrm{new}} - f(\mathbf{x}_{\mathrm{new}})\right)^2 \middle| \mathbf{x}_{\mathrm{new}}\right] + \left(f(\mathbf{x}_{\mathrm{new}}) - \mathcal{E}_n(\mathbf{x}_{\mathrm{new}})\right)^2 + \mathrm{E}\left[\left(\mathcal{E}_n(\mathbf{x}_{\mathrm{new}}) - \hat{f}_n(\mathbf{x}_{\mathrm{new}})\right)^2 \middle| \mathbf{x}_{\mathrm{new}}\right]$$

$$= \sigma_\varepsilon^2(\mathbf{x}_{\mathrm{new}}) + \left(\mathcal{E}_n(\mathbf{x}_{\mathrm{new}}) - f(\mathbf{x}_{\mathrm{new}})\right)^2 + \mathrm{V}\left[\hat{f}_n(\mathbf{x}_{\mathrm{new}}) \middle| \mathbf{x}_{\mathrm{new}}\right]$$

$$= \mathrm{Err}_n^{\mathrm{proc}}(\mathbf{x}_{\mathrm{new}}) + \mathrm{Err}_n^{\mathrm{model}}(\mathbf{x}_{\mathrm{new}}) + \mathrm{Err}_n^{\mathrm{param}}(\mathbf{x}_{\mathrm{new}})$$

Error	Symbol	Formula
Process error	$\mathrm{Err}_n^{\mathrm{proc}}(\mathbf{x}_{\mathrm{new}})$	$\sigma_\varepsilon^2(\mathbf{x}_{\mathrm{new}})$
Parameter error	$\mathrm{Err}_n^{\mathrm{param}}(\mathbf{x}_{\mathrm{new}})$	$\mathrm{V}\left[\hat{f}_n(\mathbf{x}_{\mathrm{new}}) \middle| \mathbf{x}_{\mathrm{new}}\right]$
Model error	$\mathrm{Err}_n^{\mathrm{model}}(\mathbf{x}_{\mathrm{new}})$	$\left(\mathcal{E}_n(\mathbf{x}_{\mathrm{new}}) - f(\mathbf{x}_{\mathrm{new}})\right)^2$

II . A New Framework of Prediction Error Decomposition

Parameter error estimation under general formula

Error	Symbol	Formula
Parameter error	$\mathrm{Err}_n^{\mathrm{param}}(\mathbf{x}_{\mathrm{new}})$	$\mathrm{V}\left[\hat{f}_n(\mathbf{x}_{\mathrm{new}})\middle\vert\mathbf{x}_{\mathrm{new}}\right]$

- Parameter error : the variance of the predicted value
- It can be estimated using the bootstrap method (Efron (1979)).

- The parameter error for the new data is estimated as:

$$\widehat{\mathrm{Err}_n^{\mathrm{param}}}(\mathbf{x}_{\mathrm{new}}) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{f}_n^{*b}(\mathbf{x}_{\mathrm{new}}) - \frac{1}{B}\sum_{b=1}^{B}\hat{f}_n^{*b}(\mathbf{x}_{\mathrm{new}})\right)^2$$

Here,

- B denotes the number of bootstrap samples
- $\mathcal{T}_n^{*1}, \cdots, \mathcal{T}_n^{*B}$ are B bootstrap samples taken from the training data $\mathcal{T}_n$
- $\hat{f}_n^{*1}, \cdots, \hat{f}_n^{*B}$ are created based on these bootstrap samples

II . A New Framework of Prediction Error Decomposition

Process error estimation under general formula

Error	Symbol	Formula
Process error	$\mathrm{Err}_n^{\mathrm{proc}}(\mathbf{x}_{\mathrm{new}})$	$\sigma_\varepsilon^2(\mathbf{x}_{\mathrm{new}})$

- Process error : intrinsic stochastic variability that would remain even if the resulting predictive model were true.
- It cannot be estimated directly since the true model itself cannot be identified.

- Here, we give <u>a very tentative</u> estimation method:

$$\widehat{\mathrm{Err}_n^{\mathrm{proc}}}(\mathbf{x}_{\mathrm{new}}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_n(\mathbf{x}_i)\right)^2 =: \widehat{\sigma^2}$$

- This method…
  - has <u>significant shortcomings</u> for practical use (mentioned later), but
  - is valuable in that it can be applied to any predictive model

II . A New Framework of Prediction Error Decomposition

Model error estimation under general formula

Error	Symbol	Formula
Model error	$\mathrm{Err}_n^{\mathrm{model}}(\mathbf{x}_{\mathrm{new}})$	$\left(\mathcal{E}_n(\mathbf{x}_{\mathrm{new}}) - f(\mathbf{x}_{\mathrm{new}})\right)^2$

- The model error cannot be estimated directly since we cannot identify the true model $f(\boldsymbol{x})$
- The average model error with respect to the training data can be estimated as

$$\widehat{\mathrm{AVErr}}_n^{\mathrm{model}} = \widehat{\mathrm{AVErr}}_n - \left[\frac{1}{n}\sum_i \widehat{\mathrm{Err}}_n^{\mathrm{proc}}(\mathbf{x}_i) - \frac{1}{n}\sum_i \widehat{\mathrm{Err}}_n^{\mathrm{param}}(\mathbf{x}_i)\right]$$

Can be estimated (e.g. by k-fold CV)

Can be estimated (as mentioned already)

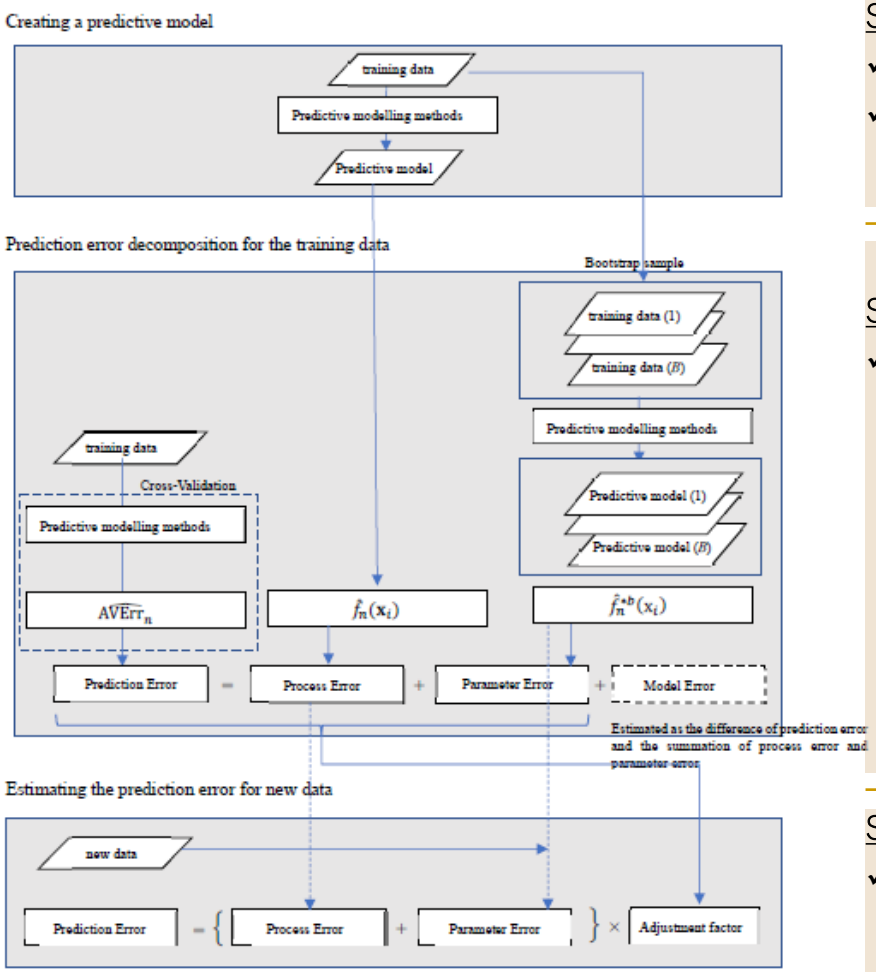
- The estimator of the model error for the new data is as:

$$\widehat{\mathrm{Err}}_n^{\mathrm{model}}(\mathbf{x}_{\mathrm{new}}) = (\gamma - 1) \times \left(\widehat{\mathrm{Err}}_n^{\mathrm{proc}}(\mathbf{x}_{\mathrm{new}}) + \widehat{\mathrm{Err}}_n^{\mathrm{param}}(\mathbf{x}_{\mathrm{new}})\right)$$

- Here, we assume that **"model error is proportional to the sum of process error and parameter error"** and define the adjustment factor $\gamma := \dfrac{n\widehat{\mathrm{AVErr}}_n}{\sum_i \widehat{\mathrm{Err}}_n^{\mathrm{proc}}(\mathbf{x}_i) + \sum_i \widehat{\mathrm{Err}}_n^{\mathrm{param}}(\mathbf{x}_i)}$

III. Development of prediction error decomposition tool

Flow chart



Step 1: Create a predictive model

- ✓ Create a predictive model using a predictive modeling method
- ✓ Here, we select a method that is implemented so that the model can be created automatically

Step 2: Error decomposition for the training data

- ✓ Each error for the training data is estimated as follows:
  - ✓ Process error (-> p. 11)
    estimated under the equal variance assumption based on the prediction model used
  - ✓ Parameter error (-> p. 10)
    estimated by bootstrap method
  - ✓ Model error (-> p. 12)
    estimated by subtracting the process and parameter errors from the total prediction error calculated by k-fold CV

Step 3: Estimating Prediction Error for New Data

- ✓ Estimated each error using error decomposition results for the training data (adjustment factor (p. 12) is used for model error calculation)

III. Development of prediction error decomposition tool

Numerical experiment (introduction)

- A numerical example of applying the tool to Boston data (housing values in suburbs of Boston) in R's MASS package

  - Predictive modeling methods:
    - Linear model without interaction
    - Linear model with interaction
    - Random Forest

  - Out of the total 506 records of the Boston data, 10 records are used as holdout data

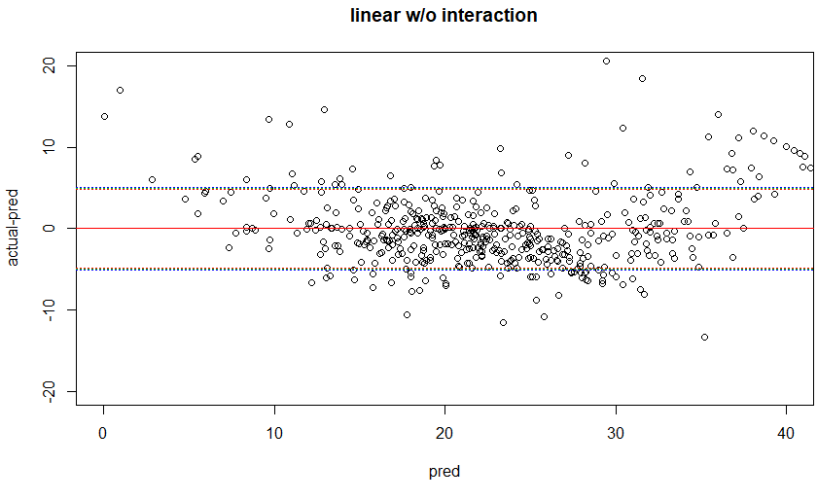
  - The target variable y : `medv` (housing price)
  - The feature vector `X` : be composed of features other than `medv`*
  - If hyperparameters are used in the predictive modeling method, they are regarded as fixed values (that is, errors related to the hyperparameter are not included in the parameter error)

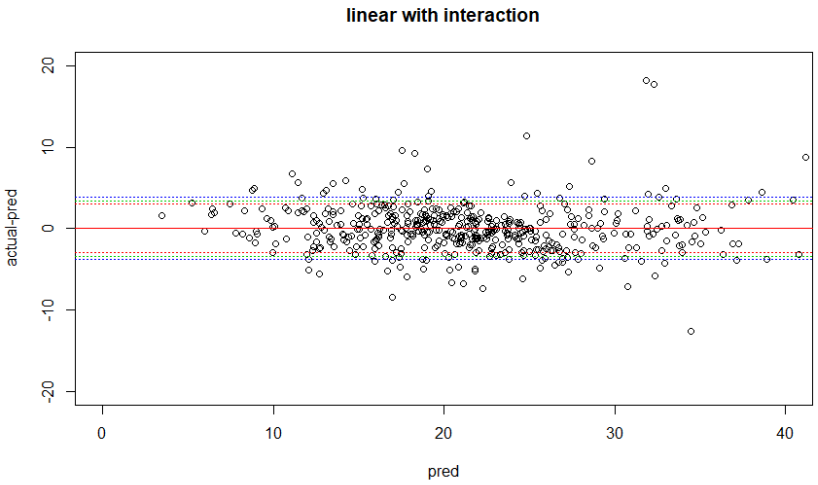
* The variables `chas` and `rad` are not used to prevent overlearning when interaction terms are considered.

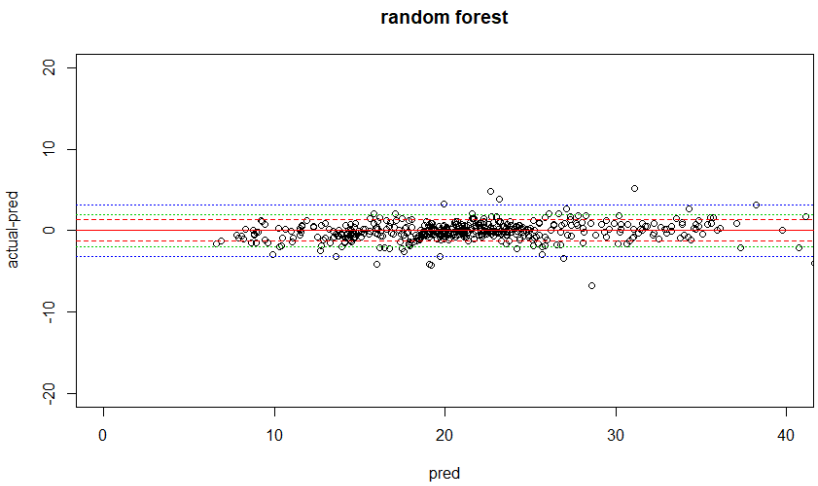
III. Development of prediction error decomposition tool

Results of application to Boston data (Training data 1)

- Results of fitting to training data







Red dotted line	Square root of process error
Green dotted line	Square root of (process error + parameter error)
Blue dotted line	Square root of prediction error (process error + parameter error + model error)

III. Development of prediction error decomposition tool

Results of application to Boston data (Training data 2)

- Composition of prediction error based on training data

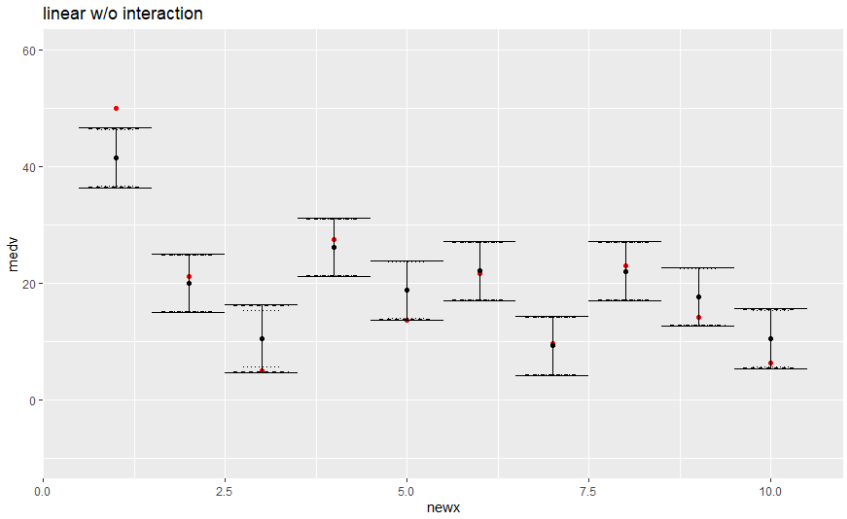
Predictive modeling method	Prediction error	Breakdown		
		Process error	Parameter error	Model error
Linear model without interaction	25.7811	23.58959	0.84544	1.34607
	100.0%	91.5%	3.3%	5.2%
Linear model with interaction	14.4787	8.87786	2.99085	2.61000
	100.0%	61.3%	20.7%	18.0%
Random Forest	10.1715	1.74307	2.09529	6.33313
	100.0%	17.1%	20.6%	62.3%

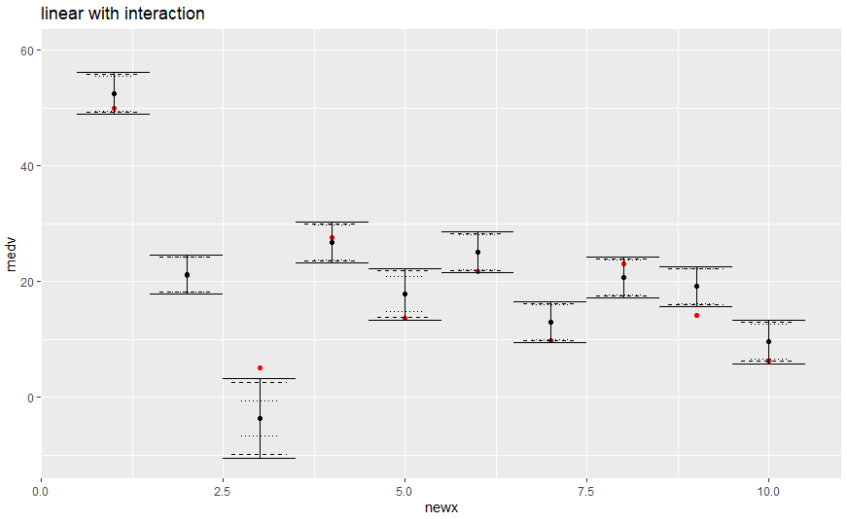
- Linear model without interaction:
  - the process error is large
  - there are many points that are not within the range of the prediction error
- Linear model with interaction:
  - each error is not as large as in the linear model without interaction
  - the differences between the actual and predicted values are often within the range of the prediction error
- Random forest:
  - many points are within the range of process error
  - most other points are still within the range of prediction error.

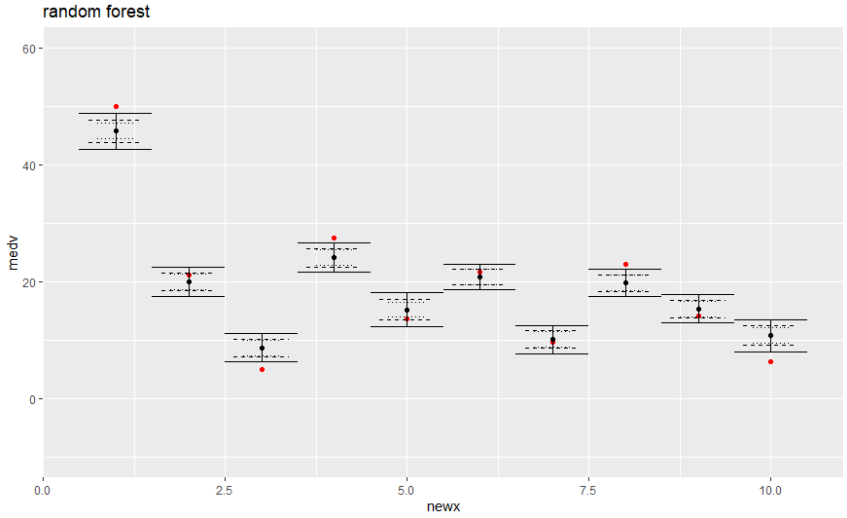
III. Development of prediction error decomposition tool

Results of application to Boston data (Holdout data)

- Error estimation results for new data (holdout data)







Dotted line	Square root of process error
Dashed line	Square root of (process error + parameter error)
Solid line	Square root of prediction error (process error + parameter error + model error)

IV. Summary and next steps

Summary and next steps

- In this presentation, we have …
  - presented a general framework of the prediction error decomposition
  - illustrated, albeit tentatively, how to estimate process, parameter, and model errors, respectively, and
  - provided specific R code and numerical examples.

- We believe that the proposed framework is appropriate and applicable to actuarial practice.

- On the other hand, the tentative estimation methods we have presented are open to various discussion, such as:
  - Estimation method of process errors: explained in the next slide
  - Estimation method of model error: The concept of adjustment factor needs to be considered based on actual data, etc.

IV. Summary and next steps

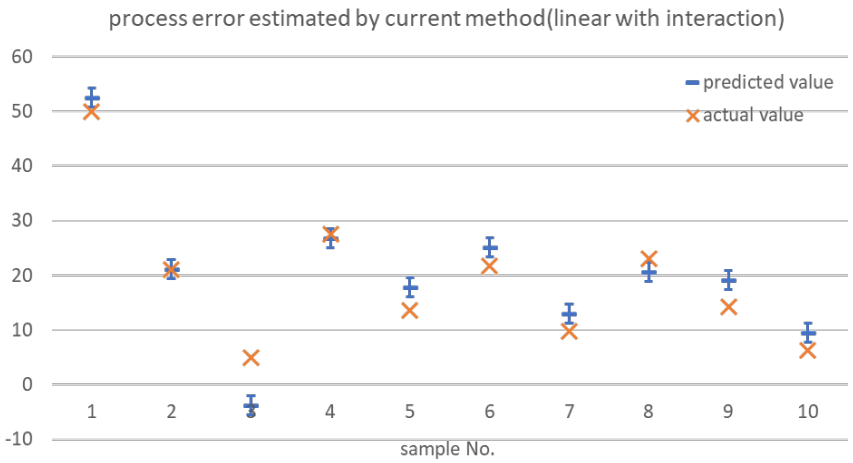
Reconsideration about process error estimation method

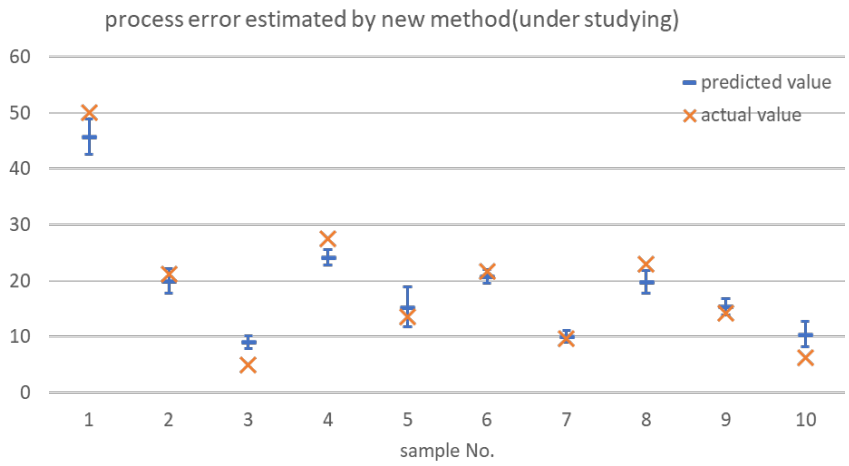
Tentative estimation method of process error	$\widehat{\mathrm{Err}_n^{\mathrm{proc}}}(\mathbf{x}_{\mathbf{new}}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}_n(\mathbf{x}_i)\right)^2 =: \widehat{\sigma^2}$	
Assumption needed to give the tentative method	✓ That the predictive model created based on the training data $\mathcal{T}_n$ is "**a model that can predict accurately enough**“ ✓ That **the process error is estimated based on the actual and predicted values** by the model	✓ That the value of **the process error does not change** depending on the feature vector $\mathbf{x}_i$
Problem	✓ The process error is intrinsic and therefore should not differ depending on the predictive model. ✓ If the predictive model is poorly fitted, the errors caused by the inaccuracy of the model are also included in the process error.	✓ In our framework, the process error of course should depend on the feature vector.

IV. Summary and next steps

Next steps on process error estimation

- We are developing a new method for consistently estimating process errors using a random forest.





- In the new method, the widths for each error bar are all different.

- We are more confident about the feasibility of developing a more appropriate error estimation method based on the general formula.

Reference

[1] Actuarial Standards Board. (2007). Actuarial Standard of Practice No. 43. (ASOP 43)

[2] Cairns, A. J. (2000). A discussion of parameter and model uncertainty in insurance. *Insurance: Mathematics and Economics*, *27*(3), 313-330.

[3] Casualty Actuarial Society. (2015). Incorporating Model Error into the Actuary's Estimate of Uncertainty. *Casualty Actuarial Society E-Forum*, *Summer 2015*.

[4] Data Science Related Basic Research Working Group of The Institute of Actuaries of Japan. (2020). Yosokumoderingu ni okeru gosa hyouka ni kansuru kenkyuu houkoku (Report on Error Evaluation in Predictive Modeling). *Akuchuarii jaanaru(Actuary Journal), 110.*

[5] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics, 7*(1), 1-26.

[6] Efron, B. & Hastie, T. (2016). *Computer Age Statistical Inference:Algorithms, Evidence, and Data Science*. Cambridge University Press.

[7] Gutterman, S. (2017). IAA Risk Book Chapter 17—Risk and Uncertainty.

[8] Hindley, D. (2018). *Claims Reserving in General Insurance*. Cambridge University Press.

[9] McGuire, G., Taylor, G., & Miller, H. (2021). Self-Assembling Insurance Claim Models Using Regularized Regression and Machine Learning. *Variance, 14 (1)*.

[10] O'Dowd, C., Smith, A., & Hardy, P. (2005). A framework for estimating uncertainty in insurance claims cost. In *Institute of Actuaries of Australia (Ed.), XVth General Insurance Seminar*.

[11] Richards, S. J., & Currie, I. D. (2009). Longevity risk and annuity pricing with the Lee-Carter model. *British Actuarial Journal*, 317-365.

[12] Richman, R., von Rummell, N., & Wuthrich, M. V. (2019). Believing the Bot— Model Risk in the Era of Deep Learning. *Available at SSRN 3444833*.

[13] Risk Margins Task Force (2008). A Framework for Assessing Risk Margins. In *Institute of Actuaries Australia, 16th General Insurance Seminar*.

[14] Taylor, G. (2019). Loss reserving models: Granular and machine learning forms. *Risks*, *7*(3), 82.

[15] Taylor, G. (2020). Loss reserving prediction error with special reference to a Tweedie sub-family. *Available at SSRN 3642378*.

[16] Taylor, G., & McGuire, G. (2016). Stochastic Loss Reserving Using Generalized Linear Models. *CAS Monograph, 3.*

- Thank you for your listening!