North American Actuarial Journal

ISSN: 1092-0277 (Print) 2325-0453 (Online) Journal homepage: www.tandfonline.com/journals/uaaj20

Regression Tree Credibility Model

Liqun Diao & Chengguo Weng

To cite this article: Liqun Diao & Chengguo Weng (2019) Regression Tree Credibility Model, North American Actuarial Journal, 23:2, 169-196, DOI: 10.1080/10920277.2018.1554497

To link to this article: https://doi.org/10.1080/10920277.2018.1554497

Published online: 18 Mar 2019.

North American Actuarial Journal, 23(2), 169–196, 2019
© 2019 Society of Actuaries
ISSN: 1092-0277 print / 2325-0453 online
DOI: 10.1080/10920277.2018.1554497

Routledge
Taylor & Francis Group

Regression Tree Credibility Model

Liqun Diao and Chengguo Weng
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada

This article applies machine learning techniques to credibility theory and proposes a regression-tree-based algorithm to integrate covariate information into credibility premium prediction. The recursive binary algorithm partitions a collective of individual risks into mutually exclusive subcollectives and applies the classical Bühlmann-Straub credibility formula for the prediction of individual net premiums. The algorithm provides a flexible way to integrate covariate information into individual net premiums prediction. It is appealing for capturing nonlinear and/or interaction covariate effects. It automatically selects influential covariate variables for premium prediction and requires no additional ex ante variable selection procedure. The superiority in prediction accuracy of the proposed algorithm is demonstrated by extensive simulation studies. The proposed method is applied to the U.S. Medicare data for illustration purposes.

1. INTRODUCTION

Risk classification, as a key ingredient of insurance pricing, exploits observable characteristics to classify insureds with similar expected claims and to build a tariffication system to conduct price discrimination on insurance products. An *a priori classification* scheme is a premium rating system to correctly express a priori information about a new policyholder or an insured without any claims experience. An a priori classification scheme is often unable to identify all the important factors for premium rating because some of the factors are either unmeasurable or unobservable. A *posteriori classification* (or *experience rating*) system is therefore necessary to rerate risks by integrating claims experience into the rating system and to obtain a more fair and reasonable price discrimination scheme.

Credibility theory, viewed as a cornerstone in the field of actuarial science (Hickman and Heacox 1999), has become a paradigm of insurance experience rating. Its advent dates back to Whitney (1918). The idea behind the theory is to view the net premium of an individual risk as a function $\mu(\Theta)$ of a random element Θ, which represents the unobservable characteristics of the individual risk, and calculate the (credibility) premium as a linear combination of the average rate of individual claims experience and the collective net premium. Credibility theory has been developed into two streams: the *limited fluctuation credibility theory* and the *greatest accuracy credibility theory*. The former is a stability-oriented theory. Its main objective is to incorporate individual claims experience into the premium calculation as much as possible and to keep the premium sufficiently stable. Most modern applications of credibility theory mainly rely on the greatest accuracy credibility theory. This theory adopts the best linear unbiased estimator to approximate individual net premiums, using both individual claims experience and collective claims experience, to achieve the minimum mean squared prediction error.

The theoretical foundation of credibility theory was established by Bühlmann (1967, 1969) and subsequently extended in many important directions. Bühlmann and Straub (1970) developed a model that allows individual risks to associate with distinct volume measures (e.g., risk exposure). Hachemeister (1975) generalized the model to a regression context, and Jewell (1973) extended its applications to multidimensional risks. The hierarchical credibility model was developed to utilize the hierarchical structure of typical insurance data (Jewell 1975; Taylor 1979; Sundt1980; Norberg 1986; Bühlmann and Jewell 1987). Credibility theory in the context of generalized linear model can be found in Nelder and Verrall (1997), Ohlsson (2008), Garrido and Zhou (2009), Christiansen and Schizinger (2016), and Quijano Xacur and Garrido (2018).

Relevant covariate information (e.g., age, gender, annual household income, driving history) is usually available to partially reflect, if not fully, the risk profile of individual risks in insurance practice. Various credibility models have been proposed to

Address correspondence to Chengguo Weng, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada. E-mail: chengguo.weng@uwaterloo.ca

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uaaj.

incorporate useful covariate information into premium prediction (e.g., Hachemeister 1975; Nelder and Verrall 1997; Ohlsson 2008; Garrido and Zhou 2009). These models all adopt a parametric regression method and prespecify a concrete regression form for the relationship between covariate variables and expected claims. As a result, they lack the flexibility of capturing intricate nonlinear and/or interaction covariate effects.

In this paper we propose a regression tree credibility (RTC) model, which applies machine learning techniques to credibility theory and yields an improved performance in terms of premium prediction accuracy. Regression trees (Breiman et al. 1984) are popular machine learning algorithms to construct prediction models by partitioning the data space into small regions in which a simple model is sufficient to achieve a good fit. Many interesting extensions and alterations of regression trees have been proposed in the literature, and a thorough review can be found in Loh (2014). Regression trees are powerful nonparametric alternatives to parametric regression models. They are particularly appealing for risk classification and prediction. Our RTC model provides a flexible way to utilize covariate information for premium prediction. It takes two steps. First, a regression-tree-based algorithm is proposed to utilize covariate information for partitioning a collective of risks into mutually exclusive subcollectives such that the risk profiles tend to be homogeneous within each subcollective and heterogeneous across subcollectives. The resulting regression tree is called a *credibility regression tree* because it is built to minimize the total credibility loss in the prediction of individual net premiums. Second, the classical Bühlmann-Straub credibility premium formula is applied to each subcollective. Since the eventual computation of individual net premiums comes from the implementation of a regression-tree-based algorithm, we call our credibility model a *regression tree credibility model*.

Our RTC model possesses three major advantages compared to the existing credibility models. First, in our model covariate variables are exploited to classify the unobservable risk profiles in a prioritized manner, whereas they are typically used to characterize conditional net premiums of individual risks, for example, in the linear regression credibility model of Hachemeister (1975). The well-known hierarchical credibility model (Bühlmann and Jewell 1987) also utilizes covariate information to classify individual risks for an enhanced performance of premium prediction; however, the classification is conducted in an ad hoc manner. The classification in the hierarchical credibility model is created according to a certain inherent structure defined by the involved covariate variables such as a geographic or demographic variable. In contrast, our RTC model classifies risks in a data-driven and prioritized manner to achieve the best prediction accuracy. Second, our credibility model provides an effective framework to capture the possible nonlinear and/or interaction covariate effects on individual net premiums. No explicit regression form needs to be assumed for either the individual net premiums or the underlying risk distribution in the implementation of our RTC model, whereas the existing credibility models typically require to prespecify a concrete regression form to integrate covariate information into a prediction model (e.g., Hachemeister 1975; Nelder and Verrall 1997). Last but not least, our RTC model automatically selects influential covariate variables for premium prediction during the course of building a credibility regression tree. Our model requires no ex ante variable selection procedure.

The rest of the article proceeds as follows. Section 2 presents our model assumptions, reviews the Bülhmann-Straub credibility model, studies the benefit of partitioning the data space, and describes a general partitioning-based credibility model. Section 3 describes the procedure of establishing a credibility regression tree and calculating premium prediction. Sections 4 and 5 contain simulation results for balanced and unbalanced claims data, respectively. Section 6 presents an application of our regression-tree–based prediction model to the U.S. Medicare data. Section 7 concludes the article. Some technical proofs are given in Appendix A, and selected graphics from the simulation studies are presented in Appendix B.

2. CREDIBILITY MODEL

2.1. Model Setup

The model we consider in this article is as follows.

Model 1. *Consider a portfolio of I risks numbered with $1, ..., I$. Let $\mathbf{Y}_i = (Y_{i,1}, ..., Y_{i,n_i})^T$ be the vector of claim ratios, $\mathbf{m}_i = (m_{i,1}, ..., m_{i,n_i})$ be the corresponding weight (also called volume measure) vector, and $\mathbf{X}_i = (X_{i,1}, ..., X_{i,p})^T$ be the covariate vector associated with individual risk i, $i = 1, ..., I$. The risk profile of individual risk i is characterized by a scalar θ_i, which is a realization of a random element Θ_i, $i = 1, ..., I$. The following two conditions are further assumed:*

1. *The triplets $(\Theta_1, \mathbf{Y}_1, \mathbf{X}_1), ..., (\Theta_I, \mathbf{Y}_I, \mathbf{X}_I)$ are mutually independent;*
2. *Given $\Theta_i = \theta_i$ and $\mathbf{X}_i = \mathbf{x}_i$, entries $Y_{i,j}, j = 1, ..., n$, are independent with*

$$\mathbb{E}\left[Y_{i,j} | \mathbf{X}_i = \mathbf{x}_i, \Theta_i = \theta_i\right] = \mu(\mathbf{x}_i, \theta_i) \text{ and } \mathrm{Var}\left[Y_{i,j} | \mathbf{X}_i = \mathbf{x}_i, \Theta_i = \theta_i\right] = \frac{\sigma^2(\mathbf{x}_i, \theta_i)}{m_{i,j}}$$

for some unknown but deterministic functions $\mu(\cdot, \cdot)$ and $\sigma^2(\cdot, \cdot)$.

We call the above model an *unbalanced claims model* because it allows distinct lengths of claims experience and distinct volume measures across the individual risks. The model is referred to as a *balanced claims model* if $n_i = n$ and $m_{i,j} = 1, j = 1, ..., n_i, i = 1, ..., I$, for a positive integer n.

Model 1 is rather general compared to the existing credibility models which without exception assume some specific functional forms for $\mu(\mathbf{X}_i, \Theta_i)$ and $\sigma^2(\mathbf{X}_i, \Theta_i)$. For example, the well-known Hachemeister model assumes $\mu(\mathbf{X}_i, \Theta_i) = \mathbf{X}_i^T \boldsymbol{\beta}(\Theta_i)$, a linear regression form with random coefficient $\boldsymbol{\beta}(\Theta_i)$ linked to the underlying risk profile Θ_i of individual risk i. In contrast, the form of the individual net premiums $\mu(\mathbf{X}_i, \Theta_i)$ is not specified in our model.

2.2. Bühlmann-Straub Credibility Model

For Model 1 described in the preceding subsection, if we discard the information from covariate variables $\{\mathbf{X}_i, i = 1, 2, ..., I\}$ and predict individual net premiums by claims data $\mathbf{Y}_i$ only, the model reduces to the well-known *Bühlmann-Straub credibility model*, where a linear predictor is applied for individual net premiums (e.g., Bühlmann and Straub 1970). Below we review the Bühlmann-Straub credibility premium formula and the corresponding minimum credibility losses. They will be used in the regression tree credibility model, which we will propose later.

With a slight abuse of notation, we let $\mu(\cdot)$ and $\sigma^2(\cdot)$ be two deterministic functions such that

$$\mathbb{E}\left[Y_{i,j}|\Theta_i = \theta_i\right] = \mu(\theta_i) \text{ and } \mathrm{Var}\left[Y_{i,j}|\Theta_i = \theta_i\right] = \frac{\sigma^2(\theta_i)}{m_{i,j}}, \; j = 1, ..., n_i \text{ and } i = 1, ..., I.$$

Define structural parameters $\sigma^2 = \mathbb{E}[\sigma^2(\Theta_i)]$ and $\tau^2 = \mathrm{Var}[\mu(\Theta_i)]$. Put $m_i = \sum_{j=1}^{n_i} m_{i,j}$, $m = \sum_{i=1}^{I} m_i$, and $\mu = \mathbb{E}[Y_{i,j}]$. μ is the collective net premium. We further denote

$$\overline{Y_i} = \sum_{j=1}^{n_i} \frac{m_{i,j}}{m_i} Y_{i,j} \text{ and } \bar{Y} = \sum_{i=1}^{I} \frac{m_i}{m} \overline{Y_i},$$

which are the sample mean of the claims experience of individual risk i and that of the collective $\mathcal{I}$, respectively.

In the literature two credibility premium predictors were proposed for individual premiums $\mu(\Theta_i)$: the *inhomogeneous credibility premium* and the *homogeneous credibility premium*. The inhomogeneous credibility premium is the best premium predictor $P_i^{(\mathrm{I})}$ among the class

$$\left\{ P_i = a_0 + \sum_{i=1}^{I} \sum_{j=1}^{n_i} a_{i,j} Y_{i,j}, a_0 \in \mathbb{R}, a_{i,j} \in \mathbb{R} \right\}$$

in terms of minimizing the quadratic loss $\mathbb{E}[(\mu(\Theta_i) - P_i)^2]$ or equivalently minimizing $\mathbb{E}[(Y_{i,n_i+1} - P_i)^2]$. The inhomogeneous credibility premium formula is given by

$$P_i^{(\mathrm{I})} = \alpha_i \overline{Y_i} + (1 - \alpha_i)\mu,$$

where

$$\alpha_i = \frac{m_i}{m_i + \sigma^2/\tau^2} \tag{2.1}$$

is known as the *credibility factor* of individual risk i (Bühlmann and Gisler 2005). The minimum values of the quadratic losses (also referred to as *credibility losses*) $\mathbb{E}[(\mu(\Theta_i) - P_i)^2]$ and $\mathbb{E}[(Y_{i,n_i+1} - P_i)^2]$ are, respectively, given by

$$L_{1,i}^{(\mathrm{I})} = \frac{\sigma^2}{m_i + \sigma^2/\tau^2} \tag{2.2}$$

and

$$L_{2,i}^{(\mathrm{I})} = \sigma^2 + \frac{\sigma^2}{m_i + \sigma^2/\tau^2}. \tag{2.3}$$

The homogeneous credibility premium for individual risk i is defined as the best premium predictor $P_i^{(\mathrm{H})}$ among the class

$$\left\{ P_i : \ P_i = \sum_{i=1}^{I} \sum_{j=1}^{n_i} a_{i,j} Y_{i,j}, \ \mathbb{E}[P_i] = \mathbb{E}[\mu(\Theta_i)], \ a_{i,j} \in \mathbb{R} \right\}$$

in terms of minimizing the quadratic loss $\mathbb{E}[(\mu(\Theta_i) - P_i)^2]$ or equivalently minimizing $\mathbb{E}[(Y_{i,n_i+1} - P_i)^2]$. The homogeneous credibility premium formula is given by

$$P_i^{(\mathrm{H})} = \alpha_i \overline{Y_i} + (1-\alpha_i)\bar{Y}, \tag{2.4}$$

where the credibility factor α_i is the same as in (2.1). The minimum values of the quadratic losses $\mathbb{E}[(\mu(\Theta_i) - P_i)^2]$ and $\mathbb{E}[(Y_{i,n_i+1} - P_i)^2]$ are, respectively, given by

$$L_{1,i}^{(\mathrm{H})} = \tau^2(1-\alpha_i)\left(1 + \frac{1-\alpha_i}{\alpha_\bullet}\right) \tag{2.5}$$

and

$$L_{2,i}^{(\mathrm{H})} = \sigma^2 + \tau^2(1-\alpha_i)\left(1 + \frac{1-\alpha_i}{\alpha_\bullet}\right), \tag{2.6}$$

where $\alpha_\bullet = \sum_{i}^{I} \alpha_i$.

In the implementation of the above credibility premium formulas, parameters σ^2 and τ^2 are, respectively, estimated by the following plug-in estimators:

$$\hat{\sigma}^2 = \frac{1}{I} \sum_{i}^{I} \frac{1}{n_i - 1} \sum_{j=1}^{n_i} m_{i,j}(Y_{i,j} - \overline{Y_i})^2 \ \text{ and } \ \hat{\tau}^2 = \max(\hat{\hat{\tau}}^2, 0), \tag{2.7}$$

where

$$\hat{\hat{\tau}}^2 = c \cdot \left[\frac{I}{I-1} \sum_{i}^{I} \frac{m_i}{m} (\overline{Y_i} - \bar{Y})^2 - \frac{I\hat{\sigma}^2}{m} \right] \tag{2.8}$$

and

$$c = \frac{I-1}{I} \left[\sum_{i}^{I} \frac{m_i}{m} \left(1 - \frac{m_i}{m}\right) \right]^{-1}.$$

Here $\hat{\sigma}_g^2$ and $\hat{\hat{\tau}}_g^2$ are unbiased and consistent estimators of σ^2 and τ^2, respectively, but $\hat{\hat{\tau}}^2$ can possibly be negative (Bühlmann and Gisler 2005, 62); thus, $\hat{\tau}^2$ is used to estimate τ^2.

2.3. Benefit of Partitioning

The regression tree credibility model, which we will propose later, recursively partitions a collective of individual risks into mutually exclusive subcollectives and consequently applies the Bühlmann-Straub credibility premium formula to each subcollective for premium prediction. In this section we investigate the benefits of such partitioning-based prediction method. We focus on the balance claims model, because it is challenging to study with the general unbalanced claims model 1. In the balance claims model, the credibility factor is the same for all the individual risks, that is,

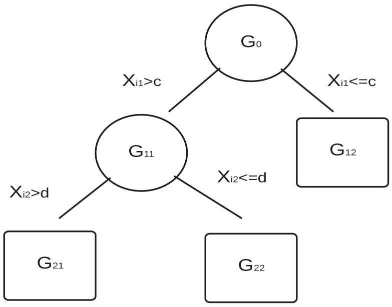

FIGURE 1. Example of Covariate-Dependent Partitioning.

$$\alpha := \alpha_i = \frac{n}{n + \sigma^2/\tau^2}, \;\; i = 1, ..., I,$$

and the four credibility loss functions in (2.2), (2.3), (2.5), and (2.6), respectively, reduce to

$$\begin{aligned}
L_{1,i}^{(\mathrm{I})} &= \frac{\sigma^2}{n + \sigma^2/\tau^2} = \tau^2(1-\alpha),\\
L_{2,i}^{(\mathrm{I})} &= \sigma^2 + \frac{\sigma^2}{n + \sigma^2/\tau^2} = \sigma^2 + \tau^2(1-\alpha),\\
L_{1,i}^{(\mathrm{H})} &= \tau^2(1-\alpha) + \frac{1}{I}\frac{\tau^2(1-\alpha)^2}{\alpha},\\
L_{2,i}^{(\mathrm{H})} &= \sigma^2 + \tau^2(1-\alpha) + \frac{1}{I}\frac{\tau^2(1-\alpha)^2}{\alpha}.
\end{aligned}$$

In most real insurance applications, I is usually a large number; thus, the second item in the expression of $L_{1,i}^{(\mathrm{H})}$ and the third term in that of $L_{2,i}^{(\mathrm{H})}$ are ignorable, and $L_{1,i}^{(\mathrm{H})}$ and $L_{2,i}^{(\mathrm{H})}$ are close to $L_{1,i}^{(\mathrm{I})}$ and $L_{2,i}^{(\mathrm{I})}$, respectively. Therefore, only $L_{1,i}^{(\mathrm{I})}$ and $L_{2,i}^{(\mathrm{I})}$ will be studied in the rest of the subsection.

We will show that the overall credibility loss for a collective $\mathcal{I} = \{1, 2, ..., I\}$ is not worsen off if the collective is arbitrarily split into two mutually exclusive subcollectives $\mathcal{I}_1$ and $\mathcal{I}_2$, where $\mathcal{I}_1 \cup \mathcal{I}_2 = \mathcal{I}$ and $\mathcal{I}_1 \cap \mathcal{I}_2 = \emptyset$. For the loss function $L_{1,i}^{(\mathrm{I})}$, the total credibility loss for the collective $\mathcal{I}$ is given by

$$L = \sum_{i=1}^{I} L_{1,i}^{(I)} = \frac{I}{n/\sigma^2 + 1/\tau^2}, \tag{2.9}$$

where σ^2 and τ^2 are the structural parameters of the collective $\mathcal{I}$. Next we consider the summation of the credibility losses from the two subcollectives $\mathcal{I}_1$ and $\mathcal{I}_2$. Let p be the probability that a randomly selected individual risk from the collective $\mathcal{I}$ falls into the subcollective $\mathcal{I}_1$ so that $q = 1-p$ is the probability for it to fall into the subcollective $\mathcal{I}_2$. On average there are pI individual risks lying in the subcollective $\mathcal{I}_1$ and qI individual risks in the subcollective $\mathcal{I}_2$. For $k = 1$, 2, let $\mu_{(k)}, \sigma^2_{(k)}$ and $\tau^2_{(k)}$ be structural parameters of the subcollective $\mathcal{I}_k$, defined in the same manner as μ, σ^2 and τ^2 shown in Section 2.2. The expected total credibility losses for subcollectives $\mathcal{I}_1$ and $\mathcal{I}_2$ are, respectively, given by

$$L_1 = \frac{Ip}{n/\sigma^2_{(1)} + 1/\tau^2_{(1)}} \text{ and } L_2 = \frac{Iq}{n/\sigma^2_{(2)} + 1/\tau^2_{(2)}}. \tag{2.10}$$

Further, it is easy to check that

$$\begin{cases}
\sigma^2 = p\sigma^2_{(1)} + q\sigma^2_{(2)},\\
\tau^2 = p\tau^2_{(1)} + q\tau^2_{(2)} + pq(\mu_{(1)}-\mu_{(2)})^2 = \tilde{\tau}^2 + pq(\mu_{(1)}-\mu_{(2)})^2,
\end{cases} \tag{2.11}$$

where $\tilde{\tau}^2 = p\tau^2_{(1)} + q\tau^2_{(2)}$.

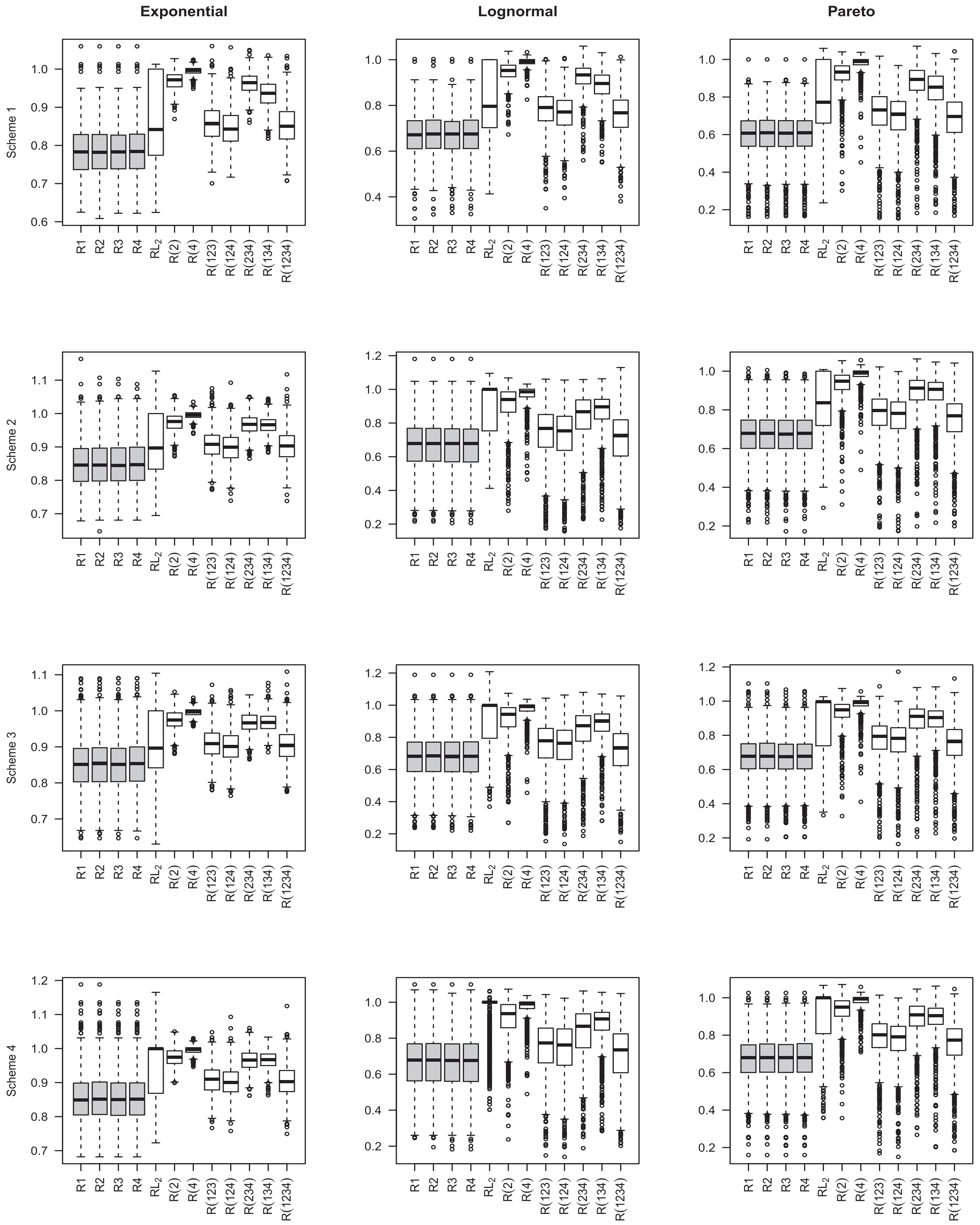

FIGURE 2. Boxplots of RPEs for Balanced Claims Model: $n = 5$ and $p = 10$.

Proposition 2.1. *The quadratic losses L, L_1, and L_2 defined in (2.9) and (2.10) satisfy*

$$L_1 + L_2 \le L. \tag{2.12}$$

Proof. See Appendix A. □

For the loss function $L_{2,i}^{(\mathrm{I})}$, the total credibility loss of the collective $\mathcal{I}$ is given by

$$L' = I\sigma^2 + \frac{I}{n/\sigma^2 + 1/\tau^2} = I\sigma^2 + L.$$

The expected total credibility losses for subcollectives $\mathcal{I}_1$ and $\mathcal{I}_2$ are, respectively, given by

$$L_1' = Ip\sigma_{(1)}^2 + \frac{Ip}{n/\sigma_{(1)}^2 + 1/\tau_{(1)}^2} = Ip\sigma_{(1)}^2 + L_1 \text{ and } L_2' = Ip\sigma_{(2)}^2 + \frac{Iq}{n/\sigma_{(2)}^2 + 1/\tau_{(2)}^2} = Ip\sigma_{(2)}^2 + L_2.$$

It is easy to show that $L_1' + L_2' \le L'$ by applying Proposition 2.1 in conjunction with the fact that $\sigma^2 = p\sigma_{(1)}^2 + q\sigma_{(2)}^2$.

Remark 2.1. Proposition 2.1 suggests that any partitioning of a collective of individual risks does not deteriorate the prediction accuracy of the Bühlmann-Straub credibility premium formula, given that the structural parameters of each resulting subcollective can be computed without any statistical errors. This result provides a theoretical foundation for the use of partitioning-based premium prediction methods. As in many other statistical models, however, statistical estimation error is inevitable, the structural parameters are unknown in general, and the estimation of these parameters using available data naturally brings about statistical errors. The existence of estimation error restrains us from excessively partitioning the overall collective, and we need to strike a tradeoff between the benefit from partitioning and the adverse effect from estimation errors.

2.4. Partitioning-Based Credibility Premium

Since partitioning of the data space can potentially improve premium prediction accuracy as demonstrated in Section 2.3, we approximate $\mu(\mathbf{X}_i, \Theta_i)$ and $\sigma^2(\mathbf{X}_i, \Theta_i)$, respectively, by piecewise constant forms as

$$\sum_{k=1}^{K} \mathrm{I}\{\mathbf{X}_i \in A_k\}\mu_{(k)}(\Theta_i) \text{ and } \sum_{k=1}^{K} \mathrm{I}\{\mathbf{X}_i \in A_k\}\frac{\sigma_{(k)}^2(\Theta_i)}{m_{i,j}},$$

where $\mathrm{I}\{\cdot\}$ is the indicator function, K is a positive integer, $\{A_1, A_2, ..., A_K\}$ is a partition of the covariate space, and $\mu_{(k)}(\Theta_i)$ and $\sigma_{(k)}^2(\Theta_i)$, respectively, represent the net premium and the variance of an individual risk i from the kth subcollective, which has a risk profile Θ_i:

$$\mu_{(k)}(\theta_i) = \mathbb{E}\big(Y_{i,j}|\mathbf{X}_i \in A_k,\ \ \Theta_i = \theta_i\big) \text{ and } \sigma_{(k)}^2(\theta_i) = \mathrm{Var}\big(Y_{i,j}|\mathbf{X}_i \in A_k,\ \ \Theta_i = \theta_i\big). \tag{2.13}$$

The condition "$\mathbf{X}_i \in A_k$" means that the individual risk i is classified into the kth subcollective.

We apply the Bühlmann-Straub credibility premium formula to each subcollective. For an individual i classified into the kth subcollective, the inhomogeneous credibility premium to predict $\mu_{(k)}(\Theta_i)$ is computed as

$$\pi_i^{(\mathrm{I})(k)} = \alpha_i^{(k)}\overline{Y}_i + \left(1 - \alpha_i^{(k)}\right)\mu_{(k)}, \tag{2.14}$$

where $\mu_{(k)} = \mathbb{E}[Y_{i,j}|\mathbf{X}_i \in A_k] = \mathbb{E}[\mu_{(k)}(\Theta_i)]$ is the net premium of the kth subcollective, and

$$\alpha_i^{(k)} = \frac{m_i}{m_i + \sigma_{(k)}^2/\tau_{(k)}^2}.$$

Here $\tau^2_{(k)}$ and $\sigma^2_{(k)}$ are the structural parameters of the kth subcollective: $\tau^2_{(k)} = \mathrm{Var}[\mu_{(k)}(\Theta_i)]$ and $\sigma^2_{(k)} = \mathbb{E}[\sigma^2_{(k)}(\Theta_i)]$. For a general individual i, the partitioning-based inhomogeneous credibility premium is computed as

$$\pi_i^{(\mathrm{I})} = \sum_{k=1}^{K} \mathrm{I}\{\mathbf{X}_i \in A_k\}\pi_i^{(\mathrm{I})(k)}. \tag{2.15}$$

The credibility losses $\mathbb{E}[(\mu_{(k)}(\Theta_i) - \pi_i^{(\mathrm{I})(k)})^2]$ and $\mathbb{E}[(Y_{i,n_i+1} - \pi_i^{(\mathrm{I})(k)})^2]$ are, respectively, given by

$$L^{(\mathrm{I})}_{1,(k)} = \sum_{i=1}^{I} \mathrm{I}\{\mathbf{X}_i \in A_k\} \frac{\sigma^2_{(k)}}{m_i + \sigma^2_{(k)}/\tau^2_{(k)}} \tag{2.16}$$

and

$$L^{(\mathrm{I})}_{2,(k)} = \sum_{i=1}^{I} \mathrm{I}\{\mathbf{X}_i \in A_k\} \left(\sigma^2_{(k)} + \frac{\sigma^2_{(g)}}{m_i + \sigma^2_{(k)}/\tau^2_{(k)}}\right). \tag{2.17}$$

We similarly have the following partitioning-based homogeneous credibility prediction rule:

$$\pi_i^{(\mathrm{H})} = \sum_{k=1}^{K} \mathrm{I}\{\mathbf{X}_i \in A_k\}\pi_i^{(\mathrm{H})(k)}, \tag{2.18}$$

where

$$\pi_i^{(\mathrm{H})(k)} = \alpha_i^{(k)}\overline{Y_i} + \left(1 - \alpha_i^{(k)}\right)\bar{Y}_{(k)}, \tag{2.19}$$

and $\bar{Y}_{(k)}$ is the average individual claims experience of the kth subcollective. For the kth subcollective, the two homogeneous credibility losses are, respectively, given by

$$L^{(\mathrm{H})}_{1,(k)} = \sum_{i=1}^{I} \mathrm{I}\{\mathbf{X}_i \in A_k\}\tau^2_{(k)}\left(1-\alpha_i^{(k)}\right)\left(1 + \frac{1-\alpha_i^{(k)}}{\alpha_\bullet^{(k)}}\right) \tag{2.20}$$

and

$$L^{(\mathrm{H})}_{2,(k)} = \sum_{i=1}^{I} \mathrm{I}\{\mathbf{X}_i \in A_k\}\left[\sigma^2_{(k)} + \tau^2_{(k)}\left(1 - \alpha_i^{(k)}\right)\left(1 + \frac{1-\alpha_i^{(k)}}{\alpha_\bullet^{(k)}}\right)\right], \tag{2.21}$$

where $\alpha_\bullet^{(k)} = \sum_{i=1}^{I} \mathrm{I}\{\mathbf{X}_i \in A_k\}\alpha_i^{(k)}$.

Figure 1 illustrates one example of a covariate-dependent partitioning, in which covariate variable X_1 and a threshold c are selected to partition the collective into two subcollectives G_{11} and G_{12}. Consequently, the profile of individual risks in subcollective G_{11} is characterized by the conditional distribution of $(\Theta_i|X_{i,1}>c)$, whereas that in subcollective G_{12} is governed by the conditional distribution of $(\Theta_i|X_{i,1} \le c)$. If the risk profile variable Θ_i is independent of the covariate variable $X_{i,1}$, the two subcollectives are homogeneous in the sense that the risk profile variables from the two subcollectives follow the same distribution. In this case, any partitioning guided by the covariate variable X_1 does not lead to an improvement of prediction accuracy, because the total credibility loss remains the same. Thus, a good partitioning scheme should avoid choosing such a covariate variable for a meaningful split.

In Figure 1, subcollective G_{11} is selected for a further split into two further subcollectives G_{21} and G_{22}. While the partitioning procedure may go further to reach a finer partition of the data space, Figure 1 demonstrates the case where the data space is partitioned into three subcollective G_{12}, $G_{21,}$ and G_{22}, which are, respectively, defined by conditions $\{X_{i,1} \le c\}, \{X_{i,1}>c \;\&\; X_{i,2}>d\}$, and $\{X_{i,1}>c \;\&\; X_{i,2} \le d\}$. In an auto insurance application, X_1 may represent gender (male if $X_1 = 1$, and female if $X_1 = 0$) and X_2 may represent age. If we choose $c = 0$ and $d = 21$, then G_{21} denotes the group of male drivers aged more than 21, G_{22} is the group of male drivers with an age of 21 or less, and G_{12} represents the group of female drivers. Given a certain partition of the data space, the insurer may consequently offer discriminated price to each insured group according to their disparate risk profiles. Thus, the choices of covariate variables and thresholds used for partitioning are crucial to gain the benefits of partitioning, because if they are not properly selected, the benefit of partitioning can be minor and the statistical error from estimating structural parameters can substantially deteriorate the premium prediction accuracy.

3. REGRESSION TREE CREDIBILITY PREMIUM

In this section we propose a credibility regression tree algorithm to partition the data space so as to achieve the minimum total credibility loss for predicting individual net premiums. The algorithm chooses covariate variables, cutting thresholds, and order of partitioning in a data-driven manner. It is altered from the algorithm of Classification and Regression Trees (CART; Breiman et al. 1984). A brief review of CART is given in Section 3.1. The details of our proposed algorithm are described in Section 3.2.

3.1. General Procedure of Regression Tree Methods

Recursive partitioning methods are machine learning techniques to quantify relationship between two sets of variables, the responses and the covariates (e.g., Zhang and Singer 2010). These methods recursively partition the data space into small regions so that a simple regression model is sufficient for a good fit in each region. These methods are powerful alternatives to parametric regression methods and possess three inherent advantages. First, recursive partitioning methods do not require one to prespecify any mathematical form of the regression relationship between the responses and the covariates, and they therefore enjoy the flexibility to capture nonlinear and/or interaction covariate effects on the responses. Second, these methods do not require a priori variable selection procedure to establish a regression model. Variable selection is automatically achieved during the course of model building. Third, the decision structure can be utilized for the purpose of prediction and is easily visualized by a hierarchical graphical model. Recursive partitioning methods have become popular in many scientific fields.

CART is perhaps the most well-known recursive partitioning algorithm (e.g., Breiman et al. 1984). Many extensions and alterations of CART have been developed in the literature; see Loh (2014) for a thorough review. CART is formulated in three steps: (1) tree growing, (2) tree pruning, and (3) tree selection. Below we briefly describe each step.

1. *Tree Growing*: This step is to grow a large tree. Consider a data set with a response W and covariates $\mathbf{X}_i = (X_{i1}, ..., X_{ip})^T$. The partition algorithm begins with the top node containing the entire data set and proceeds through binary splits of the form $\{X_{ij} \leq c\}$ versus $\{X_{ij} > c\}$ (where c is a constant) for ordinal or continuous covariate X_{ij}, and the form $X_{ij} \in S$ (where S is a subset of the values that X_{ij} may take) for categorical covariate X_{ij}. This leads to two mutually exclusive descendant nodes. The principle behind the splitting procedure is heterogeneity reduction in the response distribution, and it is realized by separating the data set into two exclusive descendant nodes, where the responses are more homogeneous within each subset than when they were combined together in the parent node. The heterogeneity of each node can be measured in various ways and is usually quantified by a loss function such as the L_2 loss function (for a node g):

$$L_{2,(g)} = \sum_{i=1}^{I} \mathrm{I}\{i \in g\} \Big(W_i - \bar{W}_{(g)} \Big)^2, \tag{3.22}$$

   where W_i is the ith response from the node g and $\bar{W}_{(g)}$ is the sample mean of responses in the node g. During the course of tree building, the first split is selected as the one that yields the largest reduction in loss. The second split is chosen as the one that leads to the largest loss reduction among all possible splits in the two descendant nodes. The tree keeps splitting until some prespecified criteria, such as the minimum number of observations in each terminal node and the depth of tree, are achieved, and the splitting process produces a large tree ψ_0. The derived large tree usually overfits the data, and a less complex model is needed. This is obtained through steps (2) and (3) below.
2. *Tree Pruning*: This step cuts (prunes) the large tree ψ_0 built in the last step to create a sequence of nested subtrees. The creation of subtrees relies on a cost-complexity pruning algorithm. Let ψ_g denote a subtree of ψ_0 rooted at node g containing g and all of its descendant nodes. Further denote the set of terminal nodes of a tree ψ by $\tilde{\psi}$. The *cost complexity* of the subtree ψ is defined as

$$L_\nu(\psi) = L(\psi) + \nu|\psi|,$$

where $L(\psi) = \sum_{g \in \tilde{\psi}} L_{2,(g)}$ is the total loss of the subtree ψ, $|\psi|$ denotes the size of ψ (i.e., the number of terminal nodes in ψ), and ν is called the complexity parameter of ψ. Whether the descendants of the node g should be pruned is determined by comparing the following two quantities:

$$L_\nu(\psi_g) = L(\psi_g) + \nu|\psi_g| \text{ and } L_\nu(g) = L(g) + \nu,$$

where $L_\nu(\psi_g)$ is the cost complexity of the tree ψ_g rooted at node g containing g and all its descendant nodes, and $L_\nu(g)$ is that of the tree containing the node g only. The quantity

$$\nu_g = \frac{L(g) - L(\psi_g)}{|\psi_g| - 1}$$

is the critical point at which $L_\nu(\psi_g)$ and $L_\nu(g)$ are equal. For $\nu > \nu_g$, we have $L_\nu(\psi_g) > L_\nu(g)$, which suggests that a further split on node g is not worthwhile and all the descendant nodes of g should be collapsed into the node g. The pruning process starts with the weakest-linked node, the nonterminal node with the smallest value of ν_g. In other words, we calculate ν_g for all nonterminal nodes in the large tree ψ_0 and prune the descendants of the one with the smallest ν_g. Then those ν_g of nonterminal nodes in the resulting subtree are recalculated and the weakest-linked node is pruned off. The pruning procedure continues until only the top node remains. The pruning procedure yields a sequence of subtrees of ψ_0 corresponding to a sequence of critical points of the weakest-linked nodes of the subtrees. We denote the sequence of critical points by $\{\nu_1, \nu_2, ..., \nu_M\}$, where M is the number of subtrees. Each point in the sequence corresponds to a subtree.

3. *Tree Selection*: This step selects the best model among the sequence of nested subtrees that were established in the last step. The selection is usually achieved by a cross-validation procedure (e.g., Breiman et al. 1984, ch. 8.5). An L-fold cross-validation procedure divides the data set into L mutually exclusive subsets $\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_L$, and computes the cross-validation error of the subtree corresponding to a critical point ν_m as

$$\sum_{\ell=1}^{L} \sum_{i=1}^{I} \mathrm{I}\{i \in \mathcal{I}_\ell\} \Big(W_i - \hat{\psi}_\ell(\mathbf{X}_i; \nu_m) \Big)^2,$$

where $\hat{\psi}_\ell(\mathbf{X}_i; \nu_m)$ is the predicted value of the ith response using the tree which is grown from the entire data set excluding the subset $\mathcal{I}_\ell$ and pruned with a complexity parameter value of ν_m. The best tree is selected as the one for which the smallest cross-validation error is attained. The L_2 loss function is used in the computation of the above cross-validation error. The L_2 loss function may be replaced by other loss functions. A typical cross-validation procedure applies the same loss function as the one used for tree growing and tree pruning.

As described above, the loss function plays a fundamental role in building the large tree, creating the sequence of candidate trees, and selecting the best tree. The CART algorithm can be implemented by the `rpart` package (Therneau, Atkiuson, and Ripley 2018) of the statistical software `R` (`R` Core Team 2014).

3.2. Credibility Regression Tree

We propose a credibility regression tree (CRT) algorithm by altering the loss function and the cross-validation procedure in the default CART algorithm. The details are as follows:

1. *Loss Function*: The loss function is the key component of the CART algorithm because it dictates the result of the algorithm. In our algorithm we adopt one of the four credibility loss functions given in (2.16), (2.17), (2.20), and (2.21). We also apply the default L_2 loss function (3.22) for comparison purposes.
2. *Longitudinal Cross-validation*: The cross-validation procedure in the CART algorithm divides the collective $\mathcal{I} = \{1, 2, ..., I\}$ into L subsets by a random sampling scheme, and it is applicable only to new policyholders who do not have any claims experience yet. In the context of credibility theory, however, the primary goal is to predict the individual net premiums for policyholders who have already had some claims experience. So we propose a new cross-validation procedure, which duplicates the claim ratios for some individual risks so that all the individual risks have the same

length of claim ratios. The duplication procedure follows the longitudinal order, and therefore we refer to our validation procedure as *longitudinal cross-validation*.

Recall that we observe $(\mathbf{Y}_i, \mathbf{X}_i)$ for individual risk i in the collective, where $\mathbf{Y}_i = (Y_{i,1}, ..., Y_{i,n_i})^T$ is the observed claims experience, and $\mathbf{X}_i$ is the associated covariates, $i = 1, 2, ..., I$. Our longitudinal cross-validation procedure takes four steps:

1. *Step 1*: For each $i = 1, ..., I$, denote $b_i = \lfloor n_i/L \rfloor$ and $\ell_i = (n_i \mod L)$, and define a sequence

$$\mathcal{T}_i = \left\{ \underbrace{\mathcal{B}, ..., \mathcal{B}}_{b_i \ \text{blocks}}, 1, 2, ..., \ell_i \right\},$$

where $\mathcal{B} = \{1, 2, ..., L\}$. For example, $\mathcal{T}_i = \{1, 2, 3\}$ for $n_i = 3$, and $\mathcal{T}_i = \{1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 1, 2, 3\}$ for $n_i = 13$ in a five-fold cross-validation. For each $i = 1, ..., I$, independently draw a sample $\{V_{i,j}, j = 1, ..., n_i\}$ from the sequence $\mathcal{T}_i$ without replacement.

2. *Step 2*: Denote

$$\mathcal{J}_{i,\ell} := \left\{ j \in \{1, ..., n_i\} : \ V_{i,j} = \ell \right\} \ \text{and} \ {}^{\ell}\mathbf{Y}_i = \{Y_{i,j} : j \in \mathcal{J}_{i,\ell}\}, \ \ell = 1, ..., 5.$$

We divide the whole data set into five subsets denoted by

$$\left\{ \left({}^{\ell}\mathbf{Y}_i, \mathbf{X}_i \right) : \ {}^{\ell}\mathbf{Y}_i \neq \emptyset, \ i \in \mathcal{I} \right\}, \ \ell = 1, ..., 5.$$

3. *Step 3*: As seen in the CART algorithm that we perviously reviewed, the pruning step in our credibility regression tree also results in a sequence of nested candidate trees associated with a sequence of complexity parameters, $\nu_1, ..., \nu_M$. The cross-validation error for the candidate tree corresponding to ν_m is computed as

$$\sum_{\ell=1}^{L} \sum_{i=1}^{I} \sum_{j=1}^{n_i} \mathrm{I}\{j \in \mathcal{J}_{i,\ell}\} \Big(Y_{i,j} - \hat{\psi}_{\ell}(\mathbf{X}_i; \nu_m) \Big)^2,$$

where $\hat{\psi}_{\ell}(\mathbf{X}_i; \nu_m)$ is the predicted value of the ith response using the tree grown from the entire data excluding the set $\{{}^{\ell}\mathbf{Y}_i, i \in \mathcal{I}\}$ and pruned with complexity parameter ν_m.

4. *Step 4*: Select the tree with the smallest cross-validation error.

3.3. Premium Calculation

We follow the procedure described in Section 3.2 to build a regression tree. We use one of the four loss functions in (2.16), (2.17), (2.20), and (2.21), as well as the default L_2 loss function in the procedure. The terminal nodes of the resulting regression tree form a partition of the sample space, say, with K subcollectives:

$$\mathcal{I}_k = \{ i \in \mathcal{I} : \ \mathbf{X}_i \in A_k \}, \ k = 1, ..., K,$$

where $\{A_k, \ k = 1, ..., K\}$ is a partition of the covariate space. For each subcollective, the individual premium prediction is computed by formula (2.19).

4. SIMULATION STUDIES FOR BALANCED CLAIMS MODEL

In this section we conduct extensive simulation studies to compare the prediction performance between our regression-tree-based methods and some ad hoc partitioning methods. The simulation studies are conducted with the balanced claims model. We consider a collective of $I = 300$ individual risks and assume that each individual has n claim ratios, i.e., $n_i = n$ for $i = 1, ..., I$ in Model 1. We further assume that the volume measures $m_{i,j} = 1$ for $i = 1, ..., I$, and $j = 1, 2, ..., n$, that is, the balanced model. Three distinct values (5, 10, and 20) are considered for n.

4.1. Simulation Schemes

Four distinct schemes are set up to simulate claim ratios data. These schemes are designed to include disparate complex structures of covariate effects because we intend to demonstrate the performance of our proposed regression-tree-based algorithm in the presence of complex covariate effects.

The first scheme is the baseline and the other three are its modifications with certain extra covariate effect terms. All four simulation schemes link the individual claim ratios to the four covariate variables $X_1, ..., X_4$. Additional noise variables $\{X_5, ..., X_p\}$ are included in the implementation of our CRT algorithm to reflect common insurance practice where insignificant covariate variables are often mingled with significant ones. For each individual risk i, we independently simulate its covariate vector $\mathbf{X}_i = (X_{i,1}, ..., X_{i,p})$ from the discrete uniform distribution over the set $\{1, ..., 100\}$. We respectively consider $p = 10$ and 50 to reflect two different scales of noise variables contained in the data set.

Scheme 1. *(Base simulation scheme)*

For each $i = 1, ..., I$, independently simulate $\{\varepsilon_{i,j}, j = 1, ..., n\}$ from a given distribution function $F(\cdot)$, and generate n claims for individual risk i by

$$Y_{i,j} = \mathrm{e}^{f(\mathbf{X}_i)} + \varepsilon_{i,j},\ \ j = 1, ..., n, \tag{4.23}$$

where

$$f(\mathbf{X}_i) = 0.01\left(X_{i,1} + 2X_{i,2} - X_{i,3} + 2\sqrt{X_{i,1}X_{i,3}} - \sqrt{X_{i,2}X_{i,4}}\right). \tag{4.24}$$

In our simulation, F takes one of the following three distributions:

1. EXP(1.6487): *An exponential distribution with mean 1.6487, that is,* $F(x) = 1-\mathrm{e}^{-x/1.6487}, x \geq 0$
2. LN (0, 1): *A log-normal distribution with parameters 0 and 1, that is,* $\ln \varepsilon_{i,j} \sim N(0, 1)$
3. PAR(3, 3.2974): *A Pareto distribution* $F(x) = 1-(\frac{3.2974}{x+3.2974})^3$.

The above three distributions are selected to represent three different levels of heavy tailedness. Their parameter values are set to have an equal expected value. The two terms $2\sqrt{X_{i,1}X_{i,3}}$ and $\sqrt{X_{i,2}X_{i,4}}$ are included in the expression of $f(\mathbf{X}_i)$ to represent nonlinear and interaction covariate effects. The response variable $Y_{i,j}$ depends only on the first four covariate variables, and the rest of the covariates are included in the regression-tree-based model as noise to mimic the scenario where some covariate variables may not be influential for the claims $Y_{i,j}$. The mean and the variance of individual risk i generated by Scheme 1 are, respectively, given by

$$\mu(\mathbf{X}_i, \Theta_i) = \mathrm{e}^{f(\mathbf{X}_i)} + \mathbb{E}[\varepsilon] \approx \mathrm{e}^{f(\mathbf{X}_i)} + 1.65 \tag{4.25}$$

and

$$\sigma^2(\mathbf{X}_i, \Theta_i) = \mathrm{Var}[\varepsilon] \approx \begin{cases} 1.6487, & \text{for EXP}\,(1.6487), \\ 4.6708, & \text{for LN}\,(0, 1), \\ 8.1181, & \text{for PAR}(3, 3.2974), \end{cases} \tag{4.26}$$

for $i = 1, ..., I$, where $\mathbb{E}[\varepsilon]$ and $\mathrm{Var}[\varepsilon]$ denote the mean and the variance of the distribution function F, respectively.

As indicated by (4.26), the variance of the claim random variable from Scheme 1 are independent of the covariate variables, and it has the same value across the collective of individual risks. To illustrate the scenario where the variance of the claim random variable differs across individual risks, we design Scheme 2 below.

Scheme 2. *For each $i = 1, ..., I$, independently simulate $\{\varepsilon_{i,j}, j = 1, ..., n\}$ from a distribution function $F(\cdot; \mathbf{X}_i)$ which depends on $\mathbf{X}_i$. Generate n claims for individual risk i by*

$$Y_{i,j} = \mathrm{e}^{f(\mathbf{X}_i)} + \varepsilon_{i,j},\ \ j = 1, ..., n, \tag{4.27}$$

where $f(\mathbf{X}_i)$ is given in (4.24). We respectively consider three distinct distributions for $F(\cdot; \mathbf{X}_i)$:

$$(1)\ \ \mathrm{EXP}\left(\mathrm{e}^{\gamma(\mathbf{X}_i)/2}\right),\ \ (2)\quad \mathrm{LN}\left(0,\gamma(\mathbf{X}_i)\right),\ \ (3)\ \ \mathrm{PAR}\left(3,2\mathrm{e}^{\gamma(\mathbf{X}_i)/2}\right), \tag{4.28}$$

where

$$\gamma(\mathbf{X}_i)=\frac{1}{102}\left|2X_{i,1}-X_{i,2}+\sqrt{X_{i,1}X_{i,2}}\right|.$$

The function $\gamma(\mathbf{X}_i)$ is normalized by 102 to maintain $\mathbb{E}[\gamma(\mathbf{X}_i)]=1$, the same value as the scale parameter in the log-normal distribution that we assigned in Scheme 1. For $i=1,...,I$, the mean and the variance of individual risk i can be, respectively, computed as

$$\mu(\mathbf{X}_i,\Theta_i)=\mathrm{e}^{f(\mathbf{X}_i)}+\mathbb{E}[\varepsilon_{i,1}]=\begin{cases}\mathrm{e}^{f(\mathbf{X}_i)}+\mathrm{e}^{\gamma(\mathbf{X}_i)/2}, & \text{for EXP }\left(\mathrm{e}^{\gamma(\mathbf{X}_i)/2}\right),\\ \mathrm{e}^{f(\mathbf{X}_i)}+\mathrm{e}^{\gamma(\mathbf{X}_i)/2}, & \text{for LN }\left(0,\gamma(\mathbf{X}_i)\right),\\ \mathrm{e}^{f(\mathbf{X}_i)}+2\mathrm{e}^{\gamma(\mathbf{X}_i)/2}, & \text{for PAR }\left(3,2\mathrm{e}^{\gamma(\mathbf{X}_i)/2}\right)\end{cases} \tag{4.29}$$

and

$$\sigma^2(\mathbf{X}_i,\Theta_i)=\mathrm{Var}[\varepsilon_{i,1}]=\begin{cases}\mathrm{e}^{\gamma(\mathbf{X}_i)/2}, & \text{for EXP}\left(\mathrm{e}^{\gamma(\mathbf{X}_i)/2}\right),\\ \left(\mathrm{e}^{\gamma(\mathbf{X}_i)}-1\right)\mathrm{e}^{\gamma(\mathbf{X}_i)}, & \text{for LN}\left(0,\gamma(\mathbf{X}_i)\right),\\ 3\mathrm{e}^{\gamma(\mathbf{X}_i)}, & \text{for PAR}\left(3,\ \ 2\mathrm{e}^{\gamma(\mathbf{X}_i)/2}\right).\end{cases} \tag{4.30}$$

In Schemes 1 and 2, the claim distribution of each individual risk is fully determined by its covariate information, and thus there is no random effect on the claim distribution. In insurance practice, however, two insureds may have distinct risk profiles even though the collected covariate information is the same for both. To reflect this scenario, we add a random effect component to the claim distribution and design simulation Schemes 3 and 4 as below.

Scheme 3. *For each $i=1,...,I$, this scheme generates n claims by three steps:*

1. *Independently simulate random effect variable Θ_i from the uniform distribution $U(0.9,\ 1.1)$*
2. *Independently simulate $\{\varepsilon_{i,j},j=1,...,n\}$ from $F(\cdot;\mathbf{X}_i)$, which depends on $\mathbf{X}_i$ and takes one of the three distributions specified in Scheme 2 and*
3. *Generate n claims for individual risk i by*

$$Y_{i,j}=\Theta_i\left[\mathrm{e}^{f(\mathbf{X}_i)}+\varepsilon_{i,j}\right],\ j=1,...,n, \tag{4.31}$$

where $f(\mathbf{X}_i)$ is given in (4.24).

In Scheme 3, a random variable Θ_i is multiplied to the term $[\mathrm{e}^{f(\mathbf{X}_i)}+\varepsilon_{i,j}],\ j=1,...,n$, for the generation of claims $Y_{i,j}$. As a result, the claim distributions of two risks may no longer be the same even when the covariate information is the same for both.

It is easy to check that the mean and the variance of individual risk i simulated from Scheme 3 are respectively given by

$$\mu(\mathbf{X}_i,\Theta_i)=\Theta_i\left(\mathrm{e}^{f(\mathbf{X}_i)}+\mathbb{E}[\varepsilon_{i,1}]\right)=\begin{cases}\Theta_i\left(\mathrm{e}^{f(\mathbf{X}_i)}+\mathrm{e}^{\gamma(\mathbf{X}_i)/2}\right), & \text{for EXP}\left(\mathrm{e}^{\gamma(\mathbf{X}_i)/2}\right),\\ \Theta_i\left(\mathrm{e}^{f(\mathbf{X}_i)}+\mathrm{e}^{\gamma(\mathbf{X}_i)/2}\right), & \text{for LN}\left(0,\gamma(\mathbf{X}_i)\right),\\ \Theta_i\left(\mathrm{e}^{f(\mathbf{X}_i)}+2\mathrm{e}^{\gamma(\mathbf{X}_i)/2}\right), & \text{for PAR}\left(3,2\mathrm{e}^{\gamma(\mathbf{X}_i)/2}\right)\end{cases} \tag{4.32}$$

and

$$\sigma^2(\mathbf{X}_i,\Theta_i)=\begin{cases}\Theta_i^2\mathrm{e}^{\gamma(\mathbf{X}_i)/2}, & \text{for EXP}\left(\mathrm{e}^{\gamma(\mathbf{X}_i)/2}\right),\\ \Theta_i^2\left(\mathrm{e}^{\gamma(\mathbf{X}_i)}-1\right)\mathrm{e}^{\gamma(\mathbf{X}_i)}, & \text{for LN}\left(0,\gamma(\mathbf{X}_i)\right),\\ 3\Theta_i^2\mathrm{e}^{\gamma(\mathbf{X}_i)}, & \text{for PAR}\left(3,\ \ 2\mathrm{e}^{\gamma(\mathbf{X}_i)/2}\right).\end{cases} \tag{4.33}$$

While Scheme 3 contains a multiplicative random effect, Scheme 4 below includes a different shape of random effect.

Scheme 4. *For each $i = 1, ..., I$, the risk profile of individual i is characterized by the vector $\Theta_i = (\xi_{i,1}, \xi_{i,2})^T$, where $\xi_{i,1}$ and $\xi_{i,2}$ are two random variables. For $i = 1, ..., I$, this scheme generates n claims by the following steps:*

1. *Independently simulate random effect variables $\xi_{i,1}$ and $\xi_{i,2}$ from the uniform distribution $U(0.9,\ 1.1)$*
2. *Independently simulate $\{\varepsilon_{i,j}, j = 1, ..., n\}$ from a distribution function $F(\cdot; \mathbf{X}_i, \xi_{i,2})$ which will be specified later and*
3. *Generate n claims for individual risk i by*

$$Y_{i,j} = \mathrm{e}^{\xi_{i,1}\cdot f(\mathbf{X}_i)} + \varepsilon_{i,j},\ \ j = 1, ..., n, \tag{4.34}$$

where $f(\mathbf{X}_i)$ is defined in (4.24).

For $F(\cdot; \mathbf{X}_i, \xi_{i,2})$ in the above, we take each of the three distributions in (28) with $\gamma(\mathbf{X}_i)$ modified to $\xi_{i,2} \cdot \gamma(\mathbf{X}_i)$.

For $i = 1, ..., I$, the mean and the variance of individual risk i simulated from Scheme 4 are, respectively, given by

$$\mu(\mathbf{X}_i, \Theta_i) = \mathrm{e}^{\xi_{i,1}\cdot f(\mathbf{X}_i)} + \mathbb{E}[\varepsilon_{i,1}] = \begin{cases} \mathrm{e}^{\xi_{i,1}\cdot f(\mathbf{X}_i)} + \mathrm{e}^{[\xi_{i,2}\cdot\gamma(\mathbf{X}_i)]/2}, & \text{for EXP}\left(\mathrm{e}^{[\xi_{\mathrm{i},2}\cdot\gamma(\mathbf{X}_\mathrm{i})]/2}\right), \\ \mathrm{e}^{\xi_{i,1}\cdot f(\mathbf{X}_i)} + \mathrm{e}^{[\xi_{i,2}\cdot\gamma(\mathbf{X}_i)]/2}, & \text{for LN}\left(0,\ \ \xi_{\mathrm{i},2}\cdot\gamma(\mathbf{X}_\mathrm{i})\right), \\ \mathrm{e}^{\xi_{i,1}\cdot f(\mathbf{X}_i)} + 2\mathrm{e}^{[\xi_{i,2}\cdot\gamma(\mathbf{X}_i)]/2}, & \text{for PAR}\ \left(3,\ \ 2\mathrm{e}^{[\gamma(\mathbf{X}_i)]/2}\right) \end{cases} \tag{4.35}$$

and

$$\sigma^2(\mathbf{X}_i, \Theta_i) = \begin{cases} \mathrm{e}^{[\xi_{i,2}\cdot\gamma(\mathbf{X}_i)]/2}, & \text{for EXP}\left(\mathrm{e}^{[\xi_{\mathrm{i},2}\cdot\gamma(\mathbf{X}_\mathrm{i})]/2}\right), \\ (\mathrm{e}^{\xi_{i,2}\cdot\gamma(\mathbf{X}_i)} - 1)\mathrm{e}^{\xi_{i,2}\cdot\gamma(\mathbf{X}_i)}, & \text{for LN}\left(0,\ \ \xi_{\mathrm{i},2}\cdot\gamma(\mathbf{X}_\mathrm{i})\right), \\ 3\mathrm{e}^{\xi_{i,2}\cdot\gamma(\mathbf{X}_i)}, & \text{for PAR}\left(3,\ \ 2\mathrm{e}^{[\xi_{\mathrm{i},2}\cdot\gamma(\mathbf{X}_\mathrm{i})]/2}\right). \end{cases} \tag{4.36}$$

For each of the above four simulation schemes and each of the three involved distributions (exponential, log-normal, and Pareto), we independently simulate 1000 samples for the claims experience of $I = 300$ individual risks with $n = 5$, 10, and 20, respectively. For each simulated sample, we, respectively, use the L_2 loss function in (3.22) and the four credibility loss functions in (2.16), (2.17), (2.20), and (2.21) to grow and prune regression trees and to select the best tree using the longitudinal cross-validation described in Section 3.2. In addition, we also consider some ad hoc covariate-dependent partitionings, which are defined via the following notations:

$$\mathcal{P}(X_k) = \{\{i \in \mathcal{I} : X_{i,k} \le 50\},\ \ \{i \in \mathcal{I} : X_{i,k} > 50\}\},\ \ k = 1, ..., 5.$$

and

$$\mathcal{P}(X_{j_1}, X_{j_2}) = \{\{i \in \mathcal{I}\colon X_{i,j_1} \le 50,\ \ X_{i,j_2} \le 50\},\ \ \{i \in \mathcal{I}\colon X_{i,j_1} \le 50,\ \ X_{i,j_2} > 50\},\ \ \{i \in \mathcal{I}\colon X_{i,j_1} > 50,\ \ X_{i,j_2} \le 50\},\ \ \{i \in \mathcal{I}\colon X_{i,j_1} > 50,\ \ X_{i,j_2} > 50\}\},$$

for $j_1,\ j_2 = 1, ..., 4$ with $j_1 \neq j_2$. $\mathcal{P}(X_k)$ represents an equally binary partitioning based on a single covariate variable X_k and $\mathcal{P}(X_{j_1}, X_{j_2})$ denotes a partition based on the two covariate variables X_{j_1} and X_{j_2}. Those based on three or four covariates are, respectively, denoted by $\mathcal{P}(X_{j_1}, X_{j_2}, X_{j_3})$ and $\mathcal{P}(X_{j_1}, X_{j_2}, X_{j_3}, X_{j_4})$, and defined in the same manner. Our simulation studies include these partitionings: $\mathcal{P}(X_2), \mathcal{P}(X_4), \mathcal{P}(X_1, X_2, X_3), \mathcal{P}(X_1, X_2, X_4), \mathcal{P}(X_2, X_3, X_4), \mathcal{P}(X_1, X_3, X_4)$, and $\mathcal{P}(X_1, X_2, X_3, X_4)$.

4.2. Prediction Error

To compare the prediction performance of various partitioning (ad hoc or data-driven) methods, we compute the prediction error for a given partitioning $\{\mathcal{I}_1, ..., \mathcal{I}_K\}$ as

$$\mathrm{PE} = \frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K}\mathrm{I}\{\mathbf{X}_i \in \mathcal{I}_k\}\left(\pi_i^{(\mathrm{H})(k)} - \mu(\mathbf{X}_i, \Theta_i)\right)^2, \tag{4.37}$$

where $\pi_i^{(\mathrm{H})(k)}$ is defined in (2.19), and the net premium $\mu(\mathbf{X}_i, \Theta_i)$ is respectively given in (4.25), (4.29), (4.32), and (4.35) for Schemes 1–4, respectively.

In addition, we also directly apply the credibility premium $P_i^{(\mathrm{H})}$ given in (2.4) for prediction. The resulting collective prediction error is computed as

$$\mathrm{PE}_0 = \frac{1}{I}\sum_{i=1}^{I}\left(P_i^{(\mathrm{H})} - \mu(\mathbf{X}_i, \Theta_i)\right)^2. \tag{4.38}$$

This prediction does not use any covariate information and thus is anticipated to underperform these covariate-dependent partitioning procedures. We define the relative prediction error (RPE) for each covariate-dependent partitioning as the ratio of its prediction error to the collective prediction error, that is, the RPE is computed as $\mathrm{R} = \mathrm{PE}/\mathrm{PE}_0$. By definition, a smaller relative prediction error suggests a better prediction performance.

4.3. Simulation Results

The RPEs of our simulation studies for $n = 5$ and $p = 10$ are demonstrated in Figure 2, which contains 12 plots with each showing simulation results for a combination of one distribution and one simulation scheme. The 12 boxplots in each plot respectively correspond to the four credibility regression tree premiums (highlighted in gray), the L_2 regression tree premium (RL$_2$), and the seven ad hoc partitions which we, respectively, labeled by $\mathcal{P}(X_2)$ (R$_{(2)}$), $\mathcal{P}(X_4)$ (R$_{(4)}$), $\mathcal{P}(X_1, X_2, X_3)$ (R$_{(123)}$), $\mathcal{P}(X_1, X_2, X_4)$ (R$_{(124)}$), $\mathcal{P}(X_2, X_3, X_4)$ (R$_{(234)}$), $\mathcal{P}(X_1, X_3, X_4)$ (R$_{(134)}$), and $\mathcal{P}(X_1, X_2, X_3, X_4)$ (R$_{(1234)}$).

We have the following observations from Figure 2:

1. Almost all the boxes in Figure 2 are below the level of 1 except the one labeled by R$_{(4)}$. A RPE smaller than 1 suggests a better prediction accuracy compared to the prediction with no partitioning performed. Thus, Figure 2 indicates that all the considered covariate-dependent partitioning methods except $\mathcal{P}(X_4)$ are helpful to enhance the prediction accuracy of credibility premiums. Revisiting the form of $f(\mathbf{X}_i)$ in (4.24), we notice that $X_{i,4}$ appears only in the interaction term, and thus, it may not be sufficiently informative to be used solely for an improved prediction accuracy.
2. The first four boxplots in each of the 12 plots represent the PRE's from applying the four proposed credibility regression trees to 1000 simulated samples. The first four boxplots have similar shapes across these 12 plots in Figure 2. This observation implies that the four credibility loss functions perform equally well.
3. For each of the 12 plots, the first four boxplots are constantly positioned lower than the others in Figure 2. This means that our credibility regression-tree-based prediction rules constantly outperform the others regardless of the simulation scheme and the noise distribution used in simulations.
4. The L_2 regression-tree-based prediction performs consistently worse than the four credibility regression-tree-based methods, but it outperforms the seven ad hoc partitioning methods.
5. Comparing the columns in Figure 2, one can see that the relative prediction errors tend to be smaller for the Pareto distribution and larger for the exponential distribution. The Pareto distribution is commonly perceived as a heavy-tailed distribution, the exponential distribution is light-tailed, and the log-normal is in between. Therefore, such a comparison suggests that our regression-tree-based methods can improve relative prediction accuracy to a larger extent for a heavy-tailed claim distribution than for a light-tailed one.
6. The last two rows in Figure 2 respectively contain the results from simulation Schemes 3 and 4. Compared with the first two rows of plots in the figure, the first four boxplots in these two rows do not display an obvious shift in their positions and shapes. We recall that random effects are not assumed in simulation Schemes 1 and 2 while they are in Schemes 3 and 4. So this comparative observation implies that the presence of random effects does not deteriorate the prediction performance of our regression tree credibility premium.
7. The rightmost seven boxplots in each plot are, respectively, resulted from the seven ad hoc partitions $\mathcal{P}(X_2), \mathcal{P}(X_4), \mathcal{P}(X_1, X_2, X_3), \mathcal{P}(X_1, X_2, X_4), \mathcal{P}(X_2, X_3, X_4), \mathcal{P}(X_1, X_3, X_4)$, and $\mathcal{P}(X_1, X_2, X_3, X_4)$. These partitions are all based on some or all of the four influential covariate variables $X_{i,1}, X_{i,2}, X_{i,3}$, and $X_{i,4}$, and do not involve the six noise covariates. In contrast, we did not preclude these noise covariate variables in the course of building the credibility regression trees and the L_2 regression tree. The boxplots in Figure 2, however, show that the regression-tree-based methods yield significantly smaller RPEs than those ad hoc partitioning methods even though the former are disadvantaged by the inclusion of the noise variables.

TABLE 1
Average Prediction Error from 1000 Simulations for Balanced Claims Model ($p = 10$)

		Scheme 1			Scheme 2			Scheme 3			Scheme 4		
		EXP	LN	PAR	EXP	LN	PAR	EXP	LN	PAR	EXP	LN	PAR
$n = 5$	PE_0	0.528	0.890	1.463	0.646	2.101	1.737	0.653	2.012	1.793	0.650	2.162	1.776
	PE_3	0.414	0.586	0.819	0.548	1.299	1.100	0.556	1.291	1.131	0.554	1.325	1.124
	$\text{PE}L_2$	0.457	0.728	1.161	0.582	1.896	1.470	0.590	1.832	1.563	0.611	2.048	1.602
	$\text{PE}_{(1234)}$	0.450	0.667	0.923	0.582	1.298	1.214	0.590	1.311	1.240	0.587	1.373	1.250
$n = 10$	PE_0	0.268	0.459	0.761	0.331	1.091	0.965	0.327	1.082	0.955	0.330	1.115	0.962
	PE_3	0.226	0.342	0.490	0.297	0.757	0.665	0.298	0.768	0.681	0.300	0.784	0.673
	$\text{PE}L_2$	0.241	0.397	0.623	0.308	0.957	0.845	0.309	0.961	0.837	0.316	1.050	0.893
	$\text{PE}_{(1234)}$	0.243	0.386	0.562	0.311	0.792	0.752	0.309	0.796	0.742	0.312	0.821	0.745
$n = 20$	PE_0	0.135	0.232	0.404	0.166	0.577	0.496	0.166	0.591	0.480	0.166	0.617	0.494
	PE_3	0.121	0.190	0.287	0.157	0.449	0.385	0.158	0.465	0.387	0.158	0.469	0.390
	$\text{PE}L_2$	0.126	0.212	0.344	0.159	0.513	0.444	0.161	0.535	0.435	0.162	0.584	0.466
	$\text{PE}_{(1234)}$	0.128	0.210	0.326	0.161	0.463	0.419	0.161	0.481	0.417	0.162	0.488	0.422

8. The relative prediction errors of partitions $\mathcal{P}(X_1, X_2, X_3)$ and $\mathcal{P}(X_1, X_2, X_4)$ are in general smaller than those of partitions $\mathcal{P}(X_2, X_3, X_4)$ and $\mathcal{P}(X_1, X_3, X_4)$ even though the two pairs contain the same number of subcollectives. Therefore, for a good prediction accuracy, it is not the number of subcollectives into which the collective is partitioned that matters, but the form of the partition that matters. A partition formed in an ad hoc manner does not guarantee a promising prediction result, and, thus, it is necessary to adopt data-driven partitioning algorithms, such as our credibility regression-tree-based algorithm, to achieve an enhanced premium prediction accuracy.

We also examine the prediction performance of the various considered rules by increasing the number of individual claims to $n = 10$ and $n = 20$, respectively. The resulting RPEs are reported by Figures B.1 and B.2 in Appendix B, respectively. From these two figures, we see that all the comments we previously made for $n = 5$ also apply to the cases of $n = 10$ and $n = 20$.

For all the prediction rules in the study, we further explore how their prediction accuracy may change with the length of individual claims experience. To this end, we compute the average prediction error from 1000 simulations for each combination of a value of n (5, 10, or 20), a distribution (exponential, log-normal, or Pareto) and one simulation scheme from those four which we have previously defined. Table 1 are the results for the base prediction rule (PE_0), the CRT premium using credibility loss function (2.20) (PE_3), the L_2 regression tree premium using loss function (3.22) ($\text{PE}L_2$), and the ad hoc partitioning method $\mathcal{P}(X_1, X_2X_3, X_4)$ using all the four influential covariates ($\text{PE}_{(1234)}$).

PE_3 is chosen as a representative of the four credibility-loss-based prediction rules in Table 1, because these four rules perform similarly as one can observe from Figures 2, B.1, and B.2. Further, we select only $\text{PE}_{(1234)}$ from the seven ad hoc partitioning methods in Table 1, because it performs the best among the group. Table 1 demonstrates that the average prediction error reduces as we increase the length of individual claims experience, regardless of the prediction rule, the simulation scheme, and the noise distribution adopted in the simulation.

To investigate the performance of our regression-tree-based prediction methods applied to data containing a large number of noise variables, we increase p to 50 in the aforementioned four simulation schemes. This means that as many as 46 noise variables are included in the analysis. The simulation results are reported in Figures B.3–B.5 for $n = 5$, 10, and 20, respectively. The increase in the number of noise variables changes neither the collective prediction error PE_0 nor the performance of the seven ad hoc partitioning methods, since their prediction rules are based on influential variables only. Comparing Figures B.3–B.5 with Figures 2, B.1 and B.2, one can see that the boxplots remain the same for the seven ad hoc partitioning methods. The first four boxplots in each plot of these figures represent the PRE's resulted from our credibility regression-tree-based prediction rules. These boxplots in Figures B.3–B.5 are slightly shifted upwards compared to those in Figures 2, B.1 and B.2. This means that an increase in the number of noise variables yields an adverse effect on the performance of our regression-tree-based prediction methods, but the effect is quite small. This observation reinforces our previous assertion that there is no

TABLE 2
Distribution $D(n_0)$ for Number of Years of Claims Experience

k	n_0-2	$n_0-1,$	$n_0,$	$n_0+1,$	n_0+2
$\Pr(N=k)$	1/16	1/8	5/8	1/8	1/16

need to conduct any ex ante variable selection procedure to implement the regression-tree-based algorithms for premium prediction.

5. SIMULATION STUDIES FOR UNBALANCED CLAIMS MODEL

The simulation studies in the preceding section were conducted for the balanced claims model. In this section we consider an unbalanced claims model with $I = 300$.

The length of individual claims experience, n_i, usually varies from one risk to another. We generate n_i for individual risk i according to the distribution $D(n_0)$ exhibited in Table 2. The parameter n_0 measures the mean of the distribution. We independently simulate the risk exposure variables $m_{i,j}$ from a random variable M valued at either 0.5 or 1 with probabilities 0.2 and 0.8, respectively. We consider an independent covariate vector $\{X_i, i=1,...,10\}$ with each component following the discrete uniform distribution over the set $\{1,...,100\}, i=1,...,300$.

We consider four simulation schemes (labeled by Schemes 5–8 respectively) for the generation of claim ratios. These schemes are modified from Schemes 1–4 as follows:

For each $i=1,...,I$ and $j=1,...,n_i$, independently simulate random variables $Y_{i,j,1}$ and $Y_{i,j,2}$ from one of the following four schemes:

Scheme 5: *the claim random variable in (4.23) from Scheme 1*

Scheme 6: *the claim random variable in (4.27) from Scheme 2*

Scheme 7: *the claim random variable in (4.31) from Scheme 3*

Scheme 8: *the claim random variable in (4.34) from Scheme 4.*

The claim ratios are given by

$$Y_{i,j} = \begin{cases} 2Y_{i,j,1}, & \text{if } m_{i,j} = 0.5, \\ Y_{i,j,1} + Y_{i,j,2}, & \text{if } m_{i,j} = 1. \end{cases}$$

For each simulation scheme and each distribution type (exponential, log-normal, and Pareto), we independently simulate 1000 samples and calculate the relative prediction errors for each of the thirteen prediction rules which we applied in the preceding section. We choose $n_0 = 5$ in the simulation study. The performance of these prediction rules is summarized in Figure 3, where each row corresponds to one simulation scheme, and each column is for a different noise distribution (exponential, log-normal, and Pareto). The figure exhibits the same pattern as what we have observed in Figure 2. Therefore, all the comments we have previously made for the balanced claims model also apply to the unbalanced claims model. However, the relative prediction errors tend to be larger for the unbalanced claims model. Furthermore, as we have previously observed from the results of the balanced claims model, our credibility regression-tree-based prediction rules are the best among all the 13 methods in the study.

6. REAL DATA ANALYSIS

The Centers for Medicare and Medicaid Services (CMS), part of the Department of Health and Human Services (HHS) of the United States, administers programs including Medicare, Medicaid, the Children's Health Insurance Program (CHIP), and the Health Insurance Marketplace. CMS has released to the public a series of data files that summarize the utilization and payments for procedures, services, and prescription drugs provided to Medicare

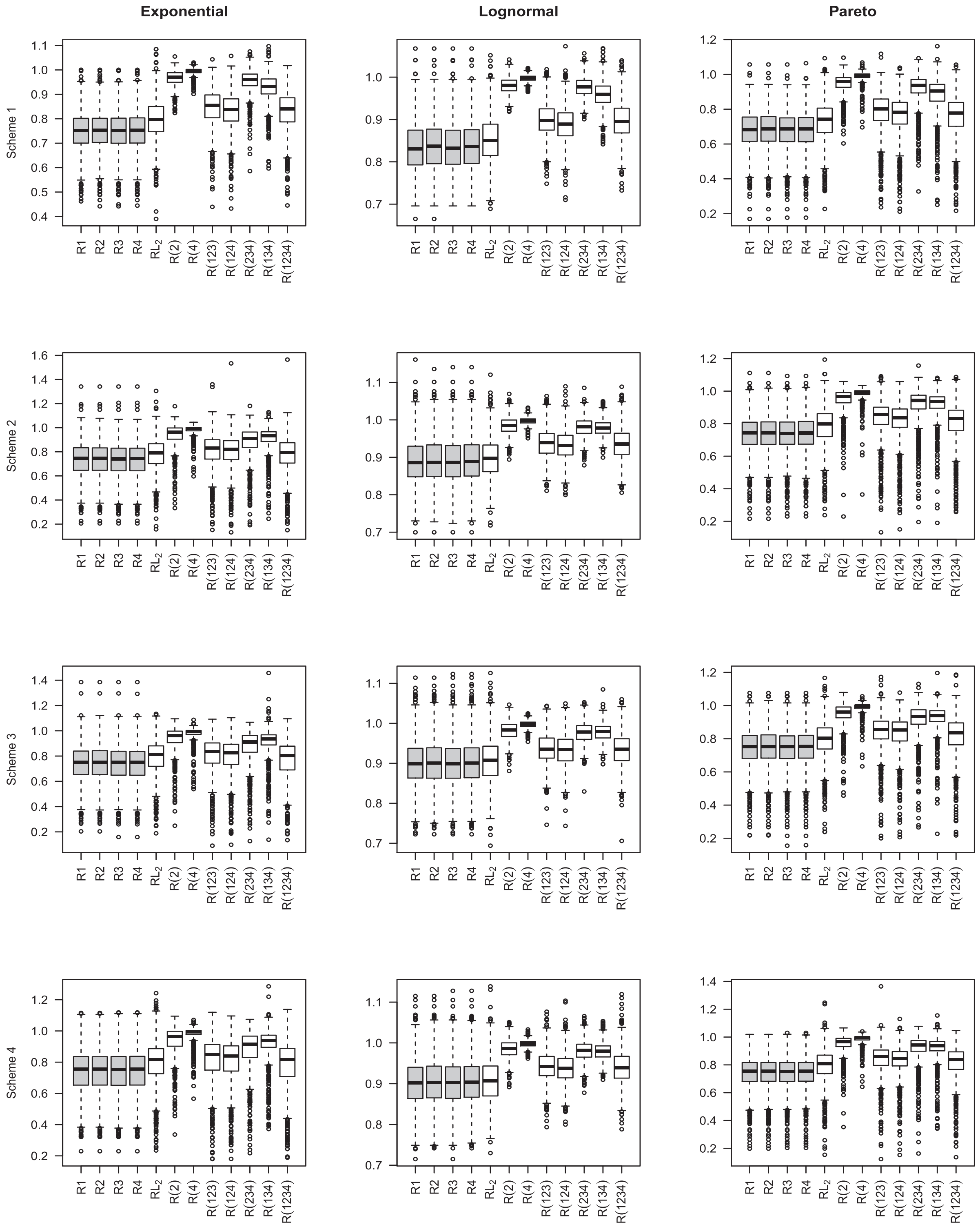

FIGURE 3. Boxplots of RPEs for Unbalanced Claims Model: Distinct n_i and $p = 10$.

beneficiaries by specific inpatient and outpatient hospitals, physicians, and other suppliers. In this section we analyze the Physician and Other Supplier Public Use File (PUF), CMS administrative claims data for calendar years 2012–2015, which provides information about services and procedures provided to Medicare beneficiaries by physicians and other health care professionals. Details of the CMS programs and the descriptions of the PUFs are

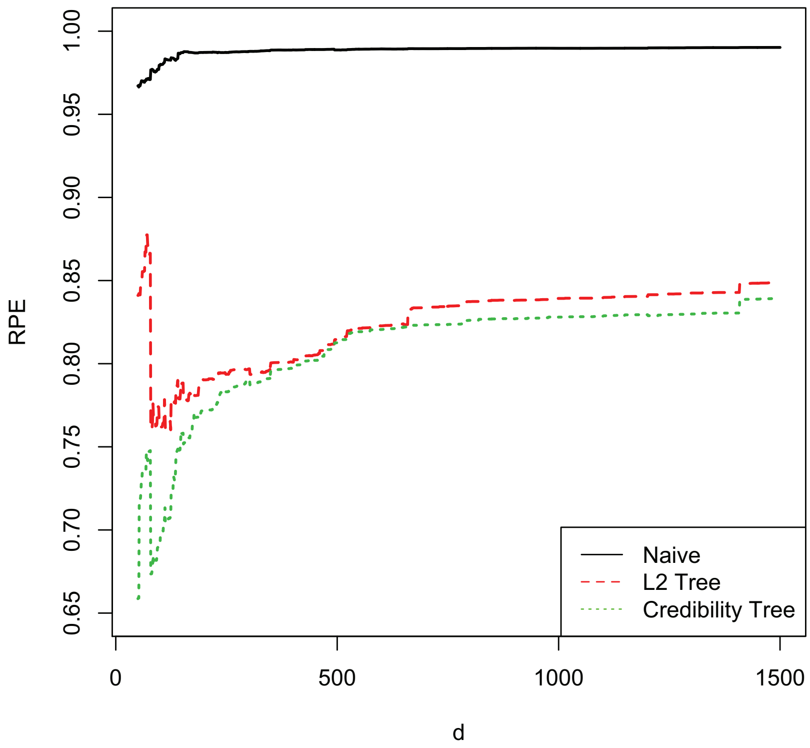

FIGURE 4. Relative Prediction Errors (RPE) of "Naive Model," L_2-Loss-Based-Tree Model, and Credibility-Loss-Based Model for Providers with Number of Services Less than d.

available on the CMS website (https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/index.html).

We are interested in predicting the Medicare payments for procedures, services, and prescription drugs provided by physicians and other supplies for future years. The "Medicare Physician and Other Supplier Aggregate Tables" in the PUF contain data of aggregate payments for each provider throughout years 2012–2015. The variable "total medical medicare payment amount" records the total amount paid by Medicare after deductible and coinsurance amounts for all medical non–Average Sales Price (ASP) services, which are not listed in the Medicare ASP File. The "total number of medical services" is the total number of medical non-ASP services that a provider provides for each year. For each health care provider i, we, respectively, use $Y_{i,j}, j = 1, ..., 4$, to denote its "average medical medicare payment amount" for each of years 2012–2015, where the "average medical medicare payment amount" is computed as the "total medical medicare payment amount" divided by the "total number of medical services." We further use $m_{i,j}, j = 1, ..., 4$, to denote the "total number of medical services" provided by the ith provider in the calendar years 2012–2015, respectively. We are interested in predicting the medical medicare payment amount per service provided by each provider. The total number of medical services should be taken as the volume measure to apply the credibility Model 1 defined in the beginning of Section 2.

We consider the variables "provider entity type," "provider gender," "provider credentials," "specialty type of the provider," and "beneficiary average risk score in 2012" as predictors. The "provider entity type" indicates whether the provider is registered in the National Plan and Provider Enumeration System (NPPES) as an organization or an individual. If the provider is registered as an organization, the "provider gender" is blank. We combine the two variables "provider entity type" and "provider gender" into a new categorical variable, "gender-entity," which is valued at 0 for female individuals, 1 for male individuals, and 2 for organizations. The providers are divided into two groups by the variable "credentials" valued at "MD" and "Others." There are 13 distinct values for the variable "specialty type" after carrying out certain necessary data-cleaning procedure. "Beneficiary average risk score" is a continuous variable, a score to reflect beneficiaries' Medicare fee-for-service spending. The value of the variable "beneficiary average risk score" may change over the years but is rather stable for most providers. So we use its value for 2012 to reflect the spending behavior of beneficiaries who receive medical care with a provider.

The raw data set contains 6835 providers with complete information of "average medical Medicare payment amount" for years 2012 – 2015 and unchanged demographics. We randomly select 20% among the 6835 providers and take their average medical Medicare payment amounts for year 2015 as the test set. There are 1367 average medical Medicare payment amounts

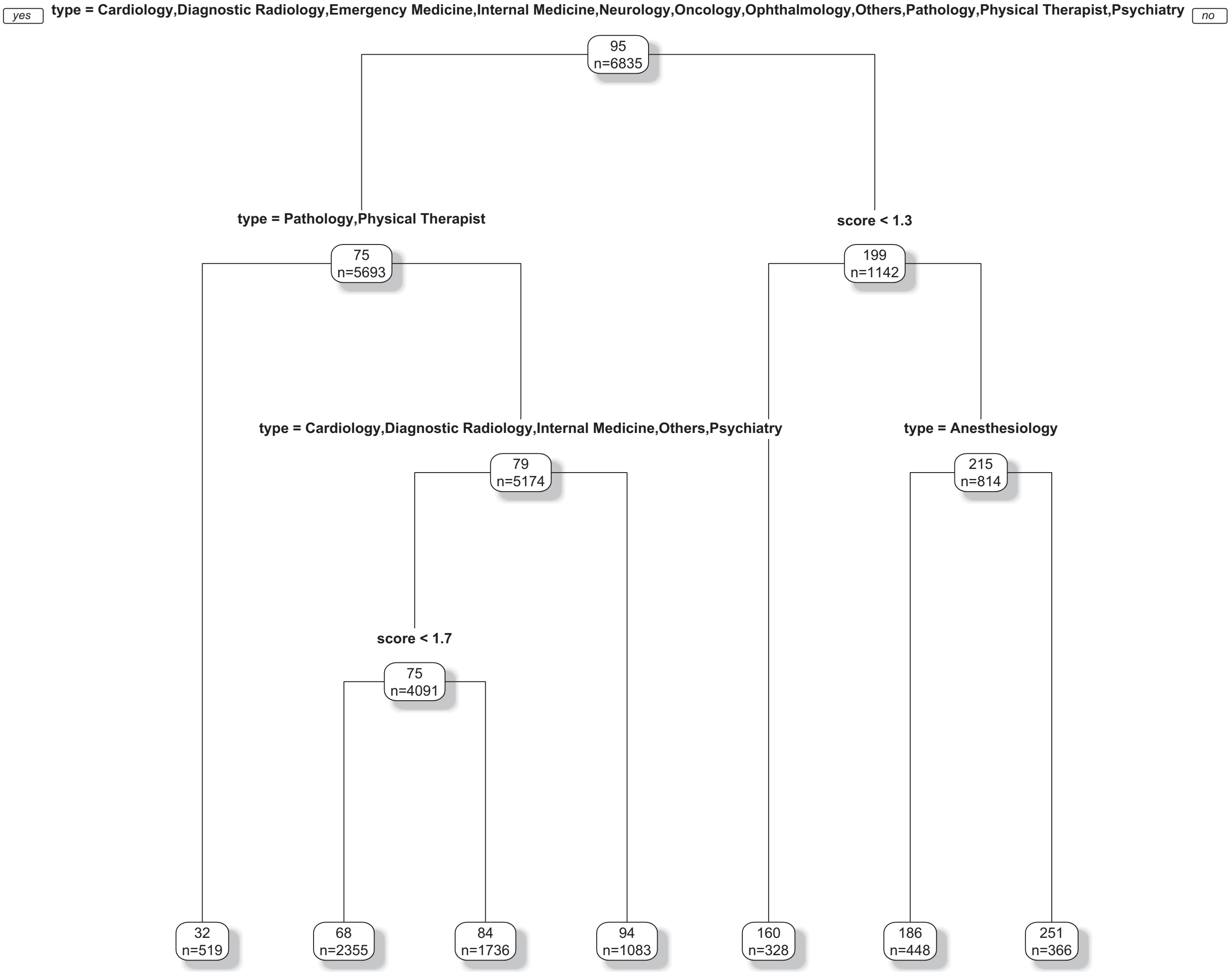
yes
type = Cardiology,Diagnostic Radiology,Emergency Medicine,Internal Medicine,Neurology,Oncology,Ophthalmology,Others,Pathology,Physical Therapist,Psychiatry
no
95
n=6835
type = Pathology,Physical Therapist
75
n=5693
score < 1.3
199
n=1142
type = Cardiology,Diagnostic Radiology,Internal Medicine,Others,Psychiatry
79
n=5174
type = Anesthesiology
215
n=814
score < 1.7
75
n=4091
32
n=519
68
n=2355
84
n=1736
94
n=1083
160
n=328
186
n=448
251
n=366

FIGURE 5. Tree Structure Built from L_2-Loss-Based Algorithm.

in the test set. The learning set contains information from all the providers for years 2012–2014 and the remaining 80% of providers for year 2015. We apply our credibility regression tree algorithm to form premium prediction rules by using the L_2 loss function and the four credibility loss functions which we proposed, respectively. For comparison purposes, we also consider the Bühlmann-Straub credibility premium formula directly applied to the data set as the baseline model. We also consider the ad hoc partitioning where the data set is partitioned into three subcollectives by the variable "gender-entity" variable. We refer to the resulting prediction rule as the "naive model".

We use the baseline model as the benchmark and compute the RPE for each prediction model. We find that the regression tree credibility model leads to the same tree structure and thus the same prediction results for all the four credibility loss functions. The PRE from our regression tree credibility model is 0.855, which is the least among all the prediction models that we considered. The PRE is 0.991 for the "naive model" and 0.862 for the L_2-loss-based model.

We recall Equation (2.1) for the formula of credibility factor:

$$\alpha_i = \frac{m_i}{m_i + \sigma^2/\tau^2}. \tag{6.39}$$

Its value is close to one for a large volume measure m_i. Therefore, the prediction for those providers with a large number of services does not benefit much from a partitioning, since their credibility premium formula yields a similar prediction value. In

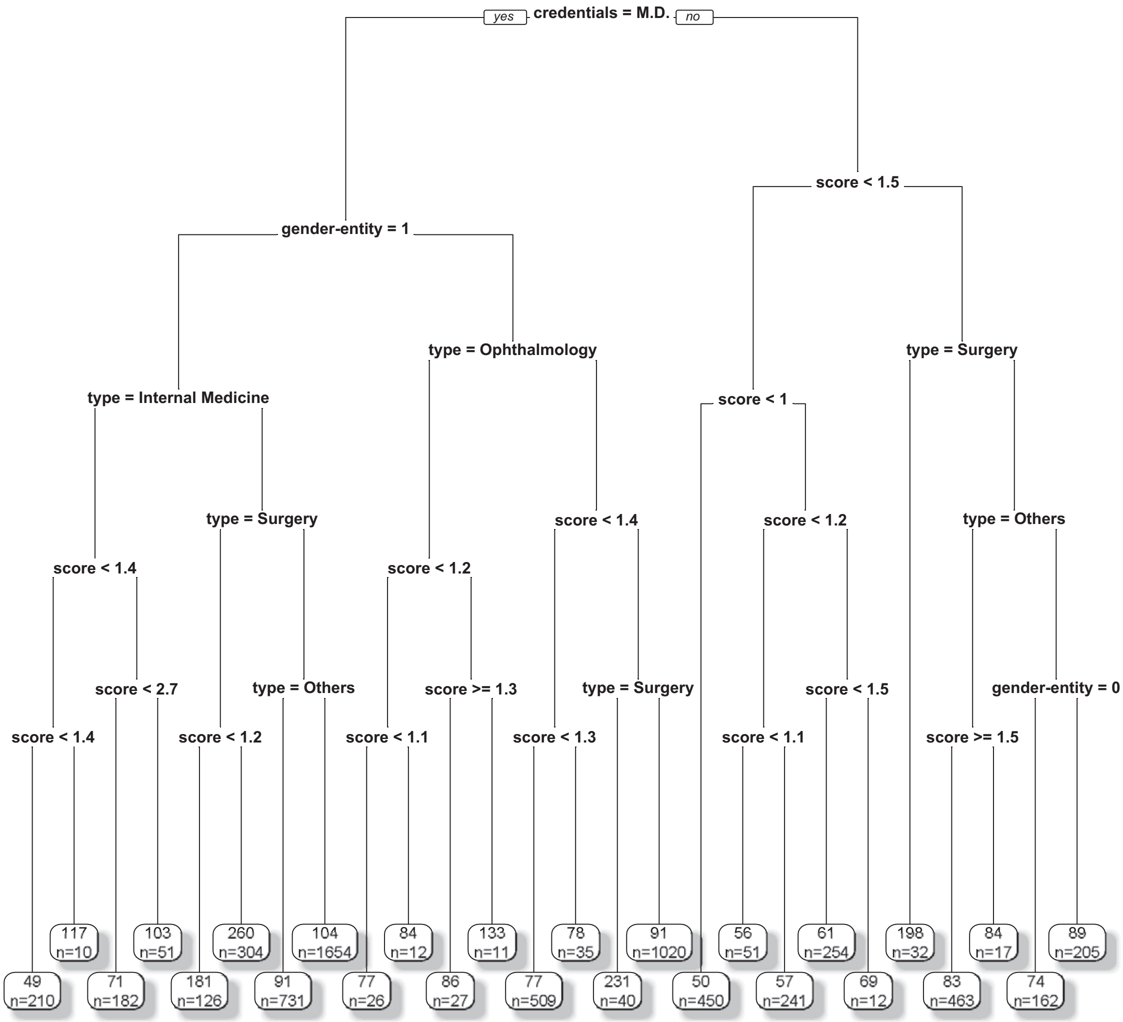

FIGURE 6. Tree Structure Built from Credibility-Loss-Based Algorithm.

contrast, the prediction for those providers with a relatively small volume measure potentially gains more from a partitioning. So we consider the PRE for the subset of providers whose number of services is smaller than a threshold d, and compute the PREs of each prediction rule with an increasing d values.

The results are demonstrated in Figure 4. From the figure, we have the following observations:

1. The black solid curve locates at the top of the plot. It reaches its lowest value 0.967 for d around 50 and stays around 0.980 for a large d. This suggests that the "naive" and the "baseline" models have similar prediction power.
2. Both the green dot and the red dash curves are far underneath the black one in Figure 4. This suggests that the regression-tree-based models significantly improve the prediction accuracy. Furthermore, as indicated by the relatively lower position of the green dot curve compared to the red dash one, the credibility-loss-based method performs better than the L_2-loss-based method.
3. Both the green dot and the red dash curves tend upwards as the threshold d increases. For instance, the PRE of the credibility regression tree method is 0.692 for $d = 100$ and increases to 0.83 for $d = 600$. This increasing trend of PRE can be explained as follows. As we have previously mentioned, a larger number of services implies less gain in prediction accuracy from a partitioning. The inclusion of more providers with a large number of services yields a larger RPE.

The tree structures obtained from the L_2 loss and the credibility loss based algorithms are shown in Figures 5 and 6, respectively. Each node in these two figures displays the average medical medicare payment amount of the providers falling into the node and the size of the terminal node. The L_2-loss-based algorithm selects the two variables "specialty type of the provider" and “beneficiary average risk score in 2012” and partitions the data space into seven subsets. The credibility-loss-based algorithm selects all the four covariate variables for partitioning and the data space is partitioned into 26 subsets.

7. CONCLUDING REMARKS

Credibility theory has been widely applied by actuaries for insurance experience rating, and relevant covariate variables have been incorporated into various credibility models to improve the prediction performance of credibility premiums. Most existing methods lack the flexibility to capture nonlinear and/or interaction covariate effects, because they require one to prespecify a parametric regression form in their implementation. The regression tree credibility (RTC) model proposed in the present article applies machine learning techniques to credibility theory for an enhanced premium prediction accuracy. It provides a nonparametric alternative to those parametric-regression-based models in the literature. For our RTC model, there is no need to conduct any ex ante analysis on the relationship between individual net premiums and covariate variables, and the proposed algorithm automatically selects influential covariate variables and informative cutting points to form a partition of the data space, upon which a well-performed premium prediction rule can be consequently established. Although only the Classification and Regression Trees algorithm is introduced in this article, it will be fruitful to pursue further research by considering other recursive partitioning methods, for example, the partDSA (partitioning deletion/substitution/addition algorithm) and MARS (multivariate adaptive regression splines) algorithms. It will be promising to consider ensemble algorithms such as bagging, boosting, and random forests. These machine learning algorithms constitute a large set of quantitative tools for the development of effective predictive models.

ACKNOWLEDGMENT

The authors are grateful to the two anonymous reviewers for their valuable suggestions which have led to a significant improvement of the article in both the contents and the organization.

FUNDING

This article is an outcome of the research project “Recursive Partitioning Methods for Credibility Theory” funded by the 2016 Individual Grants program run by the Casualty Actuarial Society (CAS) and the Society of Actuaries’s (SOA) Committee on Knowledge Extension Research (CKER). Both authors acknowledge financial support from the SOA and the Actuarial Foundation.

REFERENCES

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and regression trees*. Belmont, CA: Wadsworth.

Bühlmann, H. 1969. Experience rating and credibility. *ASTIN Bulletin* 5 (2): 157–65.

Bühlmann, H., and A. Gisler. 2005. *A course in credibility theory and its applications*. Berlin: Springer.

Bühlmann, H., and W. S. Jewell. 1987. Hierarchical credibility revisited. *Bulletin of the Swiss Association of Actuaries* 87: 35–54.

Bühlmann, H., and E. Straub. 1970. Glaubwürdigkeit für Schadensätze. *Bulletin of the Swiss Association of Actuaries* 70 (1): 111–33.

Bühlmann, H. 1967. Experience rating and credibility. *ASTIN Bulletin* 4 (3): 199–207.

Christiansen, M. C., and E. Schizinger. 2016. A credibility approach for combining likelihoods of generalized linear models. *ASTIN Bulletin* 46 (3): 531–69.

Garrido, J., and J. Zhou. 2009. Full credibility with generalized linear and mixed models. *ASTIN Bulletin* 39 (1): 61–80.

Hachemeister, C. A. 1975. Credibility for regression models with application to trend. In *Credibility: Theory and applications*, ed. P. M. Kahn, 307–48. New York, NY: Academic Press.

Hickman, J. C., and L. Heacox. 1999. Credibility theory: The cornerstone of actuarial science. *North American Actuarial Journal* 3 (2): 1–8.

Jewell, W. S. 1973. Multi-dimensional credibility. No. ORC-73-7. University of California at Berkeley Operation Research Center.

Jewell, W. S. 1975. The use of collateral data in credibility theory: A hierarchical model. *Giornale dell’Istituto Italiano degli Attuari* 38: 1–16.

Loh, W.-Y. 2014. Fifty years of classification and regression trees. *International Statistical Review* 82 (3): 329–48.

Nelder, J. A., and R. J. Verrall. 1997. Credibility theory and generalized linear models. *ASTIN Bulletin* 27 (1): 71–82.

Norberg, R. 1986. Hierarchical credibility: Analysis of a random effect linear model with nested classification. *Scandinavian Actuarial Journal* 1986 (3-4): 204–22.

Ohlsson, E. 2008. Combining generalized linear models and credibility models in practice. *Scandinavian Actuarial Journal* 2008 (4): 301–14.

Quijano Xacur, O. A., and J. Garrido. 2018. Bayesian credibility for GLMs. *Insurance: Mathematics and Economics* 83: 180–9.

R Core Team. 2014. R: A language and environment for statistical computing, Vienna: R Foundation for Statistical Computing.

Sundt, B. 1980. A multi-level hierarchical credibility regression model. *Scandinavian Actuarial Journal* 1980 (1): 25–32.

Taylor, G. C. 1979. Credibility analysis of a general hierarchical model. *Scandinavian Actuarial Journal* 1979 (1): 1–12.

Therneau, T. M., B. Atkinson, and B. Ripley. 2018. `rpart`: Recursive partitioning and regression trees. `R` package version 4.1-13. http://CRAN.R-project.org/package=rpart.

Whitney, A. W. 1918. The theory of experience rating. *Proceedings of the Casualty Actuarial Society* 4: 274–92.

Zhang, H., and B. H. Singer. 2010. *Recursive partitioning and applications*. 2nd ed. New York, NY: Springer.

Discussions on this article can be submitted until January 1, 2020. The authors reserve the right to reply to any discussion. Please see the Instructions for Authors found online at http://www.tandfonline.com/uaaj for submission instructions.

APPENDIX A

The proof of Proposition 2.1 relies on the concavity of the following function:

$$h(x,y)=\frac{1}{\frac{1}{x}+\frac{1}{y}}:(0,\infty)\times(0,\infty)\mapsto(0,\infty). \tag{A.1}$$

Lemma 7.1. *h(x, y) defined in (A.1) is a concave function.*

Proof. For $(x,y)\in(0,\infty)\times(0,\infty)$, function $h(x, y)$ is differentiable with derivatives as follows:

$$\begin{aligned}
h''_{xx}(x,y)&=\left(1+xy^{-1}\right)^{-2}=-2y^{-1}\left(1+xy^{-1}\right)^{-3}=-2y^2(y+x)^{-3},\\
h''_{xy}(x,y)&=2xy^{-2}\left(1+xy^{-1}\right)^{-3}=2xy(y+x)^{-3},\\
h''_{yy}(x,y)&=-2x^2(x+y)^{-3}.
\end{aligned}$$

Let $\mathbf{H}$ be the Hessian matrix of function h. Then, for any vector $\mathbf{a}=(a_1,a_2)'\in\mathbb{R}^2$ and for any $(x,y)\in(0,\infty)^2$,

$$\begin{aligned}
\mathbf{a}'\mathbf{H}\mathbf{a}&=(a_1,a_2)\begin{pmatrix}-2y^2(y+x)^{-3}, & 2xy(y+x)^{-3}\\ 2xy(y+x)^{-3}, & -2x^2(x+y)^{-3}\end{pmatrix}\begin{pmatrix}a_1\\ a_2\end{pmatrix}\\
&=-2(x+y)^{-3}(a_1y-a_2x)^2\le 0.
\end{aligned}$$

This means that $h(\cdot,\cdot)$ is concave on $(0,\infty)^2$. □

Proof of Proposition 2.1. We use the function h to rewrite the loss L defined in (2.9) into $L=I\cdot h(\sigma^2/n,\tau^2)$. Similarly, we have $L_1=pI\cdot h(\sigma^2_{(1)}/n,\tau^2_{(1)})$ and $L_2=qI\cdot h(\sigma^2_{(2)}/n,\tau^2_{(2)})$. Since $h(x, y)$ is a concave function (see Lemma 7.1), we use Equation (2.11) to obtain

$$\begin{aligned}
L_1+L_2&=pI\cdot h\left(\sigma^2_{(1)}/n,\tau^2_{(1)}\right)+qI\cdot h\left(\sigma^2_{(2)}/n,\tau^2_{(2)}\right)\\
&\le I\cdot h\left(p\sigma^2_{(1)}/n+q\sigma^2_{(2)}/n,\ p\tau^2_{(1)}+q\tau^2_{(2)}\right)\\
&=I\cdot h\left(\sigma^2/n,\tilde{\tau}^2\right)\\
&=\frac{I}{n/\sigma^2+1/\tilde{\tau}^2}\\
&\le\frac{I}{n/\sigma^2+1/\tau^2}\\
&=L,
\end{aligned}$$

where the first inequality follows from the convexity of h, and the second inequality is due to the fact that $\tilde{\tau}^2\le\tau^2$ which is in turn implied by (2.11). □

APPENDIX B

This section presents boxplots Figs. B.1–B.5 of relative prediction errors from a variety of simulation settings.

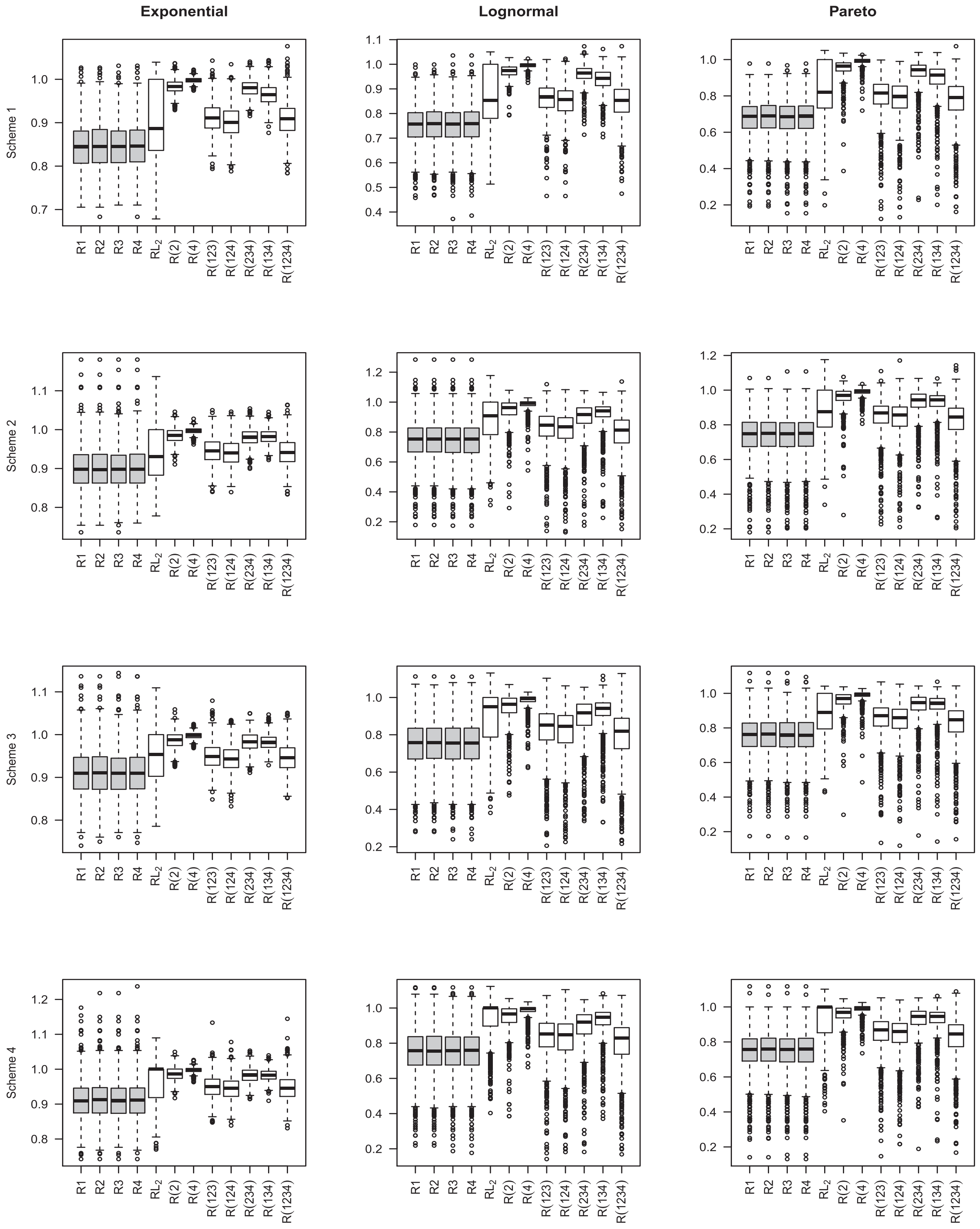

FIGURE B.1. Boxplots of RPEs for Balanced Claims Model: $n=10$ and $p=10$.

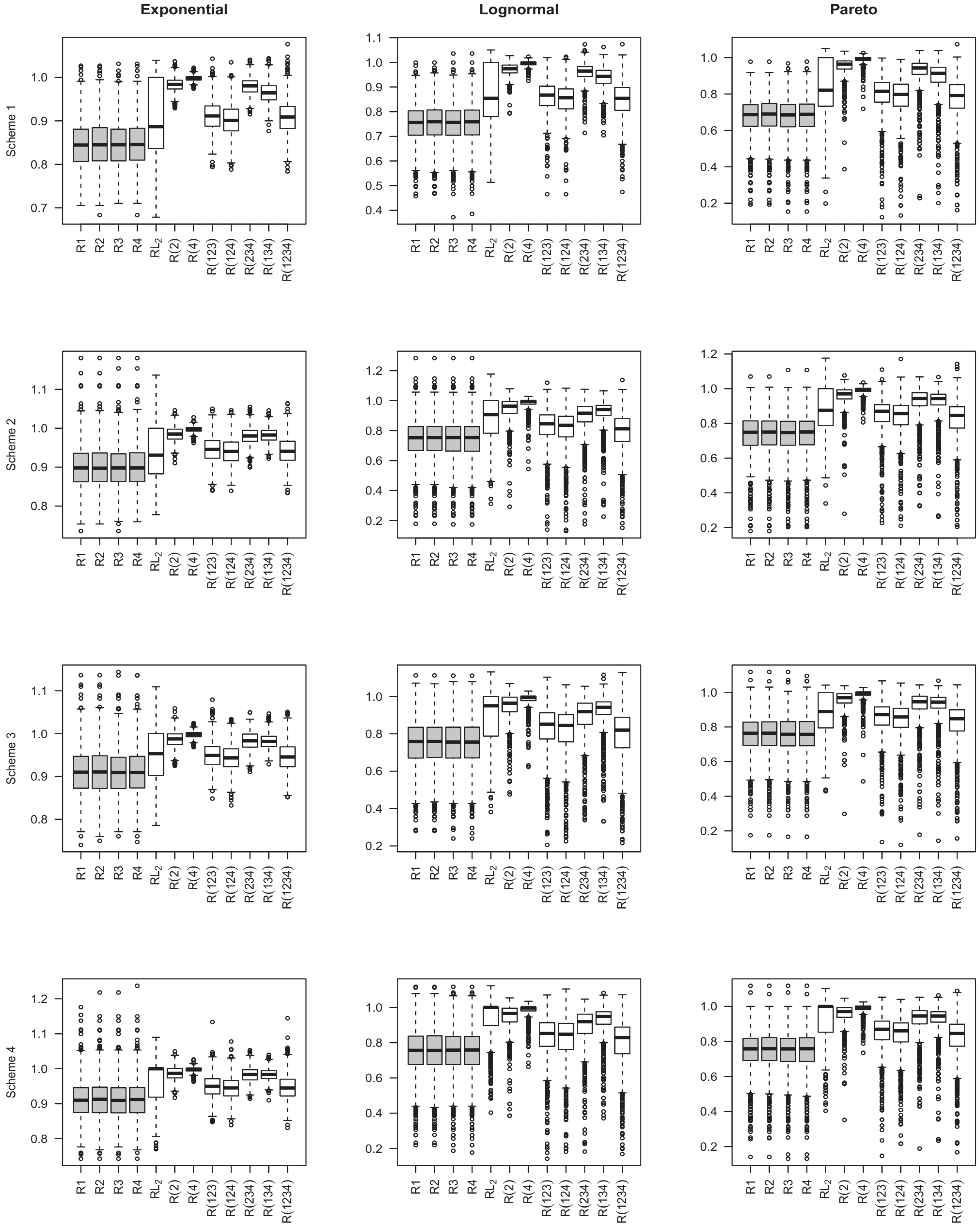

FIGURE B.2. Boxplots of RPEs for Balanced Claims Model: $n = 20$ and $p = 10$.

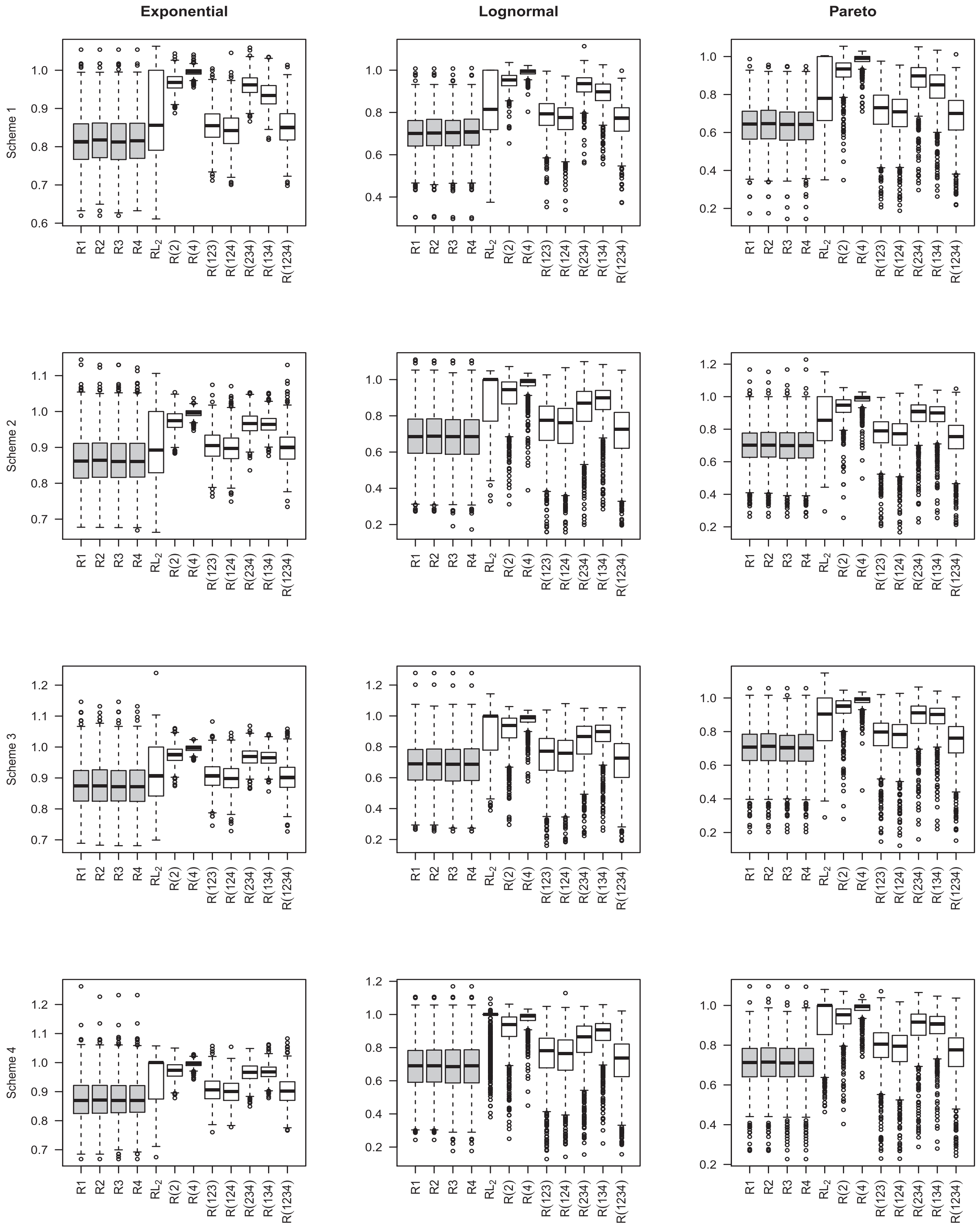

FIGURE B.3. Boxplots of RPEs for Balanced Claims Model: $n = 5$ and $p = 50$.

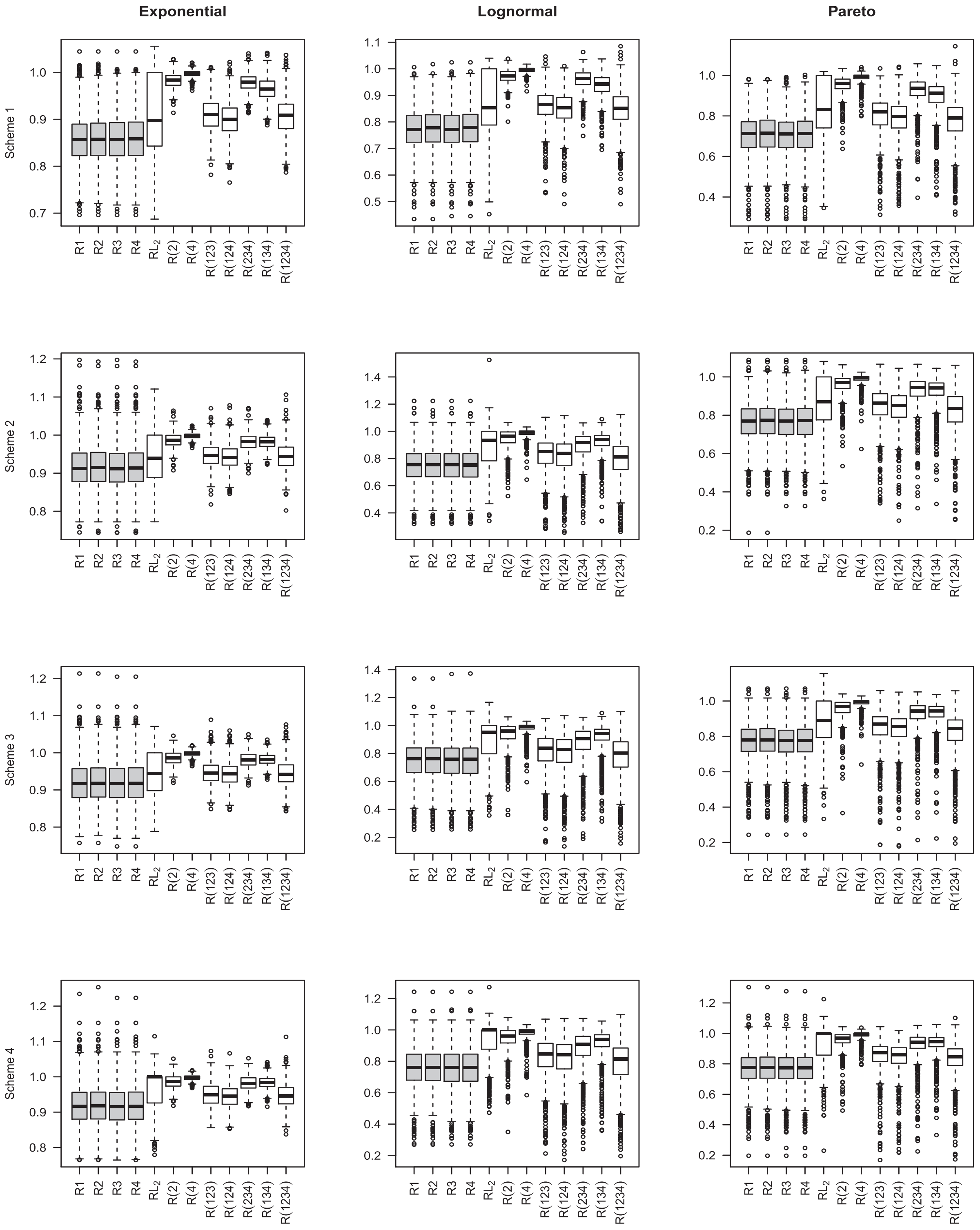

FIGURE B.4. Boxplots of RPEs for Balanced Claims Model: $n = 10$ and $p = 50$.

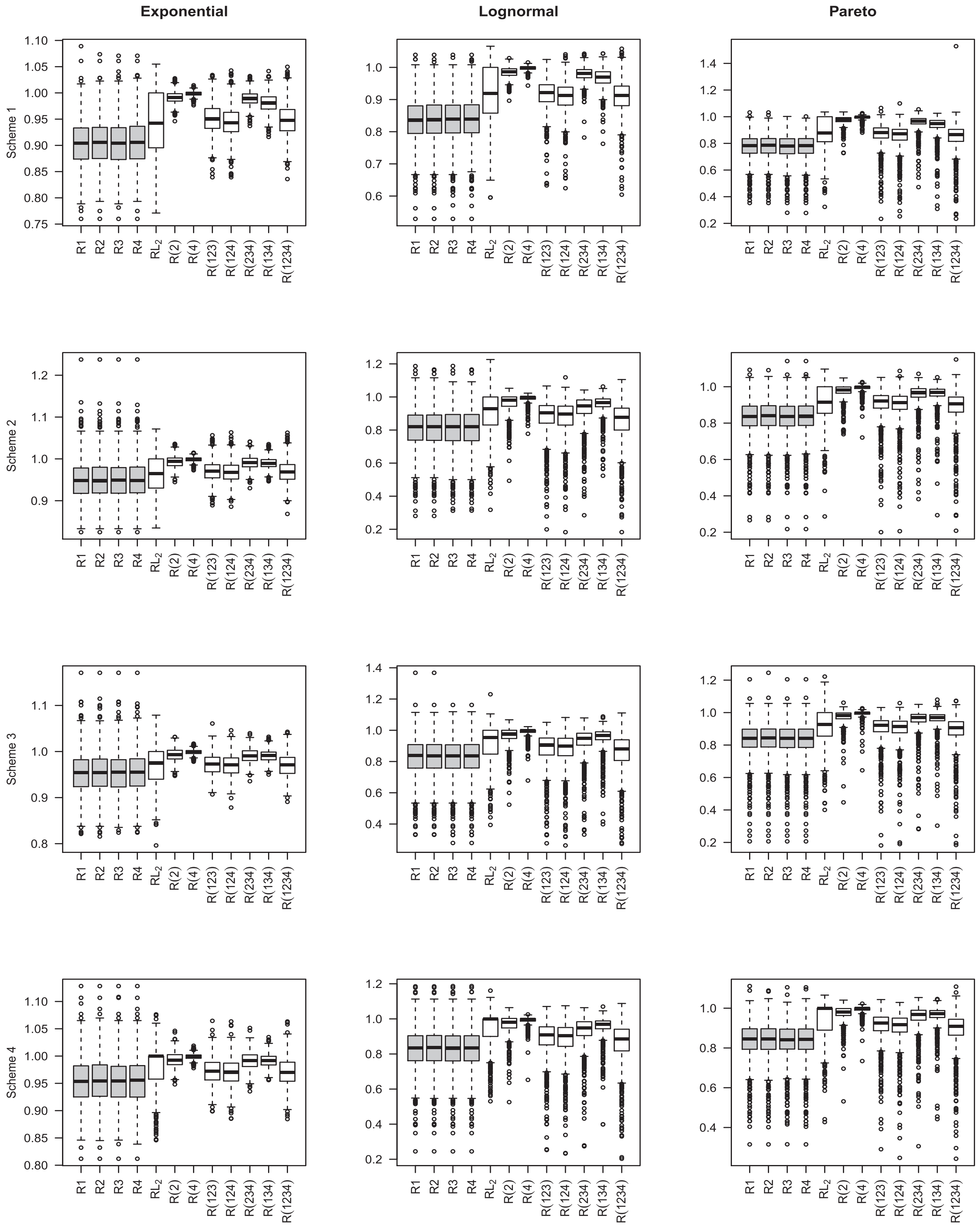

FIGURE B.5. Boxplots of RPEs for Balanced Claims Model: $n = 20$ and $p = 50$.