North American Actuarial Journal

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

Fat-Tailed Regression Modeling with Spliced Distributions

Guojun Gan & Emiliano A. Valdez

North American Actuarial Journal, 22(4), 554–573, 2018
© 2018 Society of Actuaries
ISSN: 1092-0277 print / 2325-0453 online
DOI: 10.1080/10920277.2018.1462718

Routledge
Taylor & Francis Group

Fat-Tailed Regression Modeling with Spliced Distributions

Guojun Gan and Emiliano A. Valdez
Department of Mathematics, University of Connecticut, Storrs, Connecticut

Insurance claims data usually contain a large number of zeros and exhibits fat-tail behavior. Misestimation of one end of the tail impacts the other end of the tail of the claims distribution and can affect both the adequacy of premiums and needed reserves to hold. In addition, insured policyholders in a portfolio are naturally non-homogeneous. It is an ongoing challenge for actuaries to be able to build a predictive model that will simultaneously capture these peculiar characteristics of claims data and policyholder heterogeneity. Such models can help make improved predictions and thereby ease the decision-making process. This article proposes the use of spliced regression models for fitting insurance loss data. A primary advantage of spliced distributions is their flexibility to accommodate modeling different segments of the claims distribution with different parametric models. The threshold that breaks the segments is assumed to be a parameter, and this presents an additional challenge in the estimation. Our simulation study demonstrates the effectiveness of using multistage optimization for likelihood inference and at the same time the repercussions of model misspecification. For purposes of illustration, we consider three-component spliced regression models: the first component contains zeros, the second component models the middle segment of the loss data, and the third component models the tail segment of the loss data. We calibrate these proposed models and evaluate their performance using a Singapore auto insurance claims dataset. The estimation results show that the spliced regression model performs better than the Tweedie regression model in terms of tail fitting and prediction accuracy.

1. INTRODUCTION

For many lines of insurance business, actuarial data often exhibit more extreme tail behavior than normally distributed data. Existing techniques to deal with fat-tailed data include (Shi 2014) transformation, generalized linear models (GLMs), regression models based on generalized distributions, quantile regression, and mixture models. Transformation involves creating a new distribution based on a function of an underlying random variable. For example, the logarithmic transformation is commonly applied to claim amounts to obtain a more symmetric distribution (Klugman et al. 2012). The GLMs extend the linear models based on the normal distribution to a large family of models based on distributions that can be discrete, continuous, or both (McCullagh and Nelder 1989; de Jong and Heller 2008). To apply a GLM to model fat-tailed data, people commonly use fat-tailed distributions such as the gamma distribution and the inverse Gaussian distribution (Frees 2009). Existing regression models based on generalized distributions often involve transforming random variables. The GB2 regression model is an example based on generalized distributions (Frees and Valdez 2008; Gan and Valdez 2017). Regression models based on extreme value distributions have been developed to deal with highly skewed data (Wang and Dey 2010). Quantile regression models are different from traditional regression models in that the former focus on the quantiles, instead of the mean, of the response variable (Koenker 2005; Kudryavtsev 2009). Mixture models are based on a weighted sum of distributions and have been used to model fat-tailed loss data (Lee and Lin 2010; Miljkovic and Grün 2016).

However, these techniques have limitations. A drawback of the transformation method is that it changes the variance structure of the data and may magnify the error of the prediction. GLMs and regression based on generalized distributions model the data using a single distribution, which might not be suitable when tail behavior is inconsistent with the behavior of the entire loss distribution. Quantile regression is generally unsuitable for small datasets since it relies on empirical quantiles. Calibrating existing models with such data can lead to biases that tend to either underestimate or overestimate the tail of the distribution. Finally, while mixture models are becoming increasingly popular, estimating the parameters within such framework is often based on the Expectation-Maximization algorithm which poses significant challenges, especially on the initialization of parameter estimates (McLachlan and Peel 2000; Yin and Lin 2016).

Address correspondence to Emiliano A. Valdez, Department of Mathematics, University of Connecticut, 341 Mansfield Road, Storrs, CT 06269-1009. E-mail: emiliano.valdez@uconn.edu

In this article, we explore the potential use of spliced distributions to better capture fat-tail characteristics of insurance claims data. A spliced distribution involves using different distributions in subdivided intervals to describe the behavior of a loss random variable. An advantage of spliced distributions is that they allow us to model different parts of a response variable with different distributions. Splicing is classified as a method for creating new distributions, and this method provides the flexibility of decomposing a distribution into several parts to better capture behavior within distinct regions. For each part, smooth functions of covariates will be incorporated into the parameters of the distribution. Spliced distributions have the potential to better capture tails of a loss distribution with wide ranging applications in general insurance, health insurance, and life insurance.

The method of splicing has appeared in Klugman et al. (2012) as a method for creating new distributions, and it has been proposed for modeling heavy tails for operational risks (Peters and Shevchenko 2015). Two-component spliced distributions have been used to fit skewed loss data. See, for example, Cooray and Ananda (2005), Scollnik (2007), Vernic et al. (2009), Pigeon and Denuit (2011), and Nadarajah and Bakar (2014). An R package was developed for composite lognormal distributions (Nadarajah and Bakar 2013). However, little work has been done on using spliced distributions with covariate information to fit data and draw inference from the results. The familiar two-part frequency-severity regression model (Frees 2009, Chap. 16) for insurance claims can be considered as a special case of spliced distribution models. Another example that used spliced distributions is Fang and Ma (2013), where a three-part regression model was proposed to analyze the quality of health insurance coverage in China.

To demonstrate the promise of using spliced distributions to model fat-tailed data, we consider a dataset of auto insurance claims from a Singapore company. A classic actuarial problem is ratemaking, the process of determining the price of insurance products. For auto and other types of general insurance products, the premium rate is usually determined in advance of knowing the ultimate cost of the claims. As a result, a major step of the ratemaking process is to predict as accurately as possible the expected claims based on historical data (Ohlsson and Johansson 2010; Gray and Pitts 2012; Friedland 2014; Parodi 2014). Regression models are typically used to model the relationships between the claims and the underlying explanatory variables to capture heterogeneity and make improved predictions. It is a challenge to build regression models for the Singapore auto claims data; the data exhibit fat-tail behavior where extreme values are more likely to occur than in normally distributed data. In this article, we address this challenge by using spliced regression models.

This article is organized as follows. In Section 2 we give a description of the spliced distributions. In Section 3 we present regression models based on spliced distributions with covariates. In particular, we describe two specific spliced regression models with three components. In Section 4 we demonstrate the effectiveness of the spliced regression models by calibrating them to the Singapore auto claims data. Some concluding remarks are presented in Section 5.

2. SPLICED DISTRIBUTIONS

The density function of an n-component spliced distribution is defined as follows (Klugman et al. 2012):

$$f(x) = \begin{cases} a_1 f_1(x), & \text{if } x \in C_1, \\ a_2 f_2(x), & \text{if } x \in C_2, \\ \vdots & \\ a_n f_n(x), & \text{if } x \in C_n. \end{cases} \tag{1}$$

Here $a_1, a_2, ..., a_n$ are positive weights that add up to one:

$$\sum_{i=1}^{n} a_i = 1.$$

For $i = 1, 2, ..., f_i(x)$ is a legitimate density function with all probability on the interval $\mathrm{C_i}$:

$$\int_{C_i} f_i(x)\ \mathrm{d}x = 1.$$

The intervals $C_1, C_2, ..., C_n$ are mutually exclusive:

$$C_i \cap C_j = \emptyset, \quad \forall i \neq j.$$

The intervals are also sequentially ordered: $x <$ y if $x \in C_i$ and $y \in C_j$ for all i < j. For example, the intervals can be formed by $C_1 = [c_0, c_1], C_2 = (c_1, c_2], ..., C_n = (c_{n-1}, \infty)$, where $c_0, c_1, ..., c_{n-1}$ are break points or thresholds. An advantage of spliced

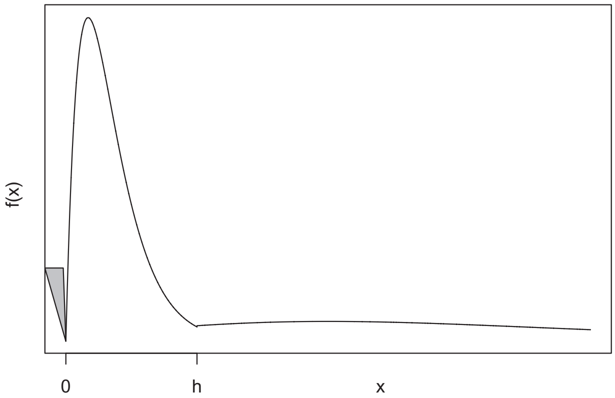

FIGURE 1. Hypothetical Density Function of Spliced Distribution with Three Components. *Note:* The shaded triangle represents the point mass at zero, and *h* is the threshold that separates the middle and tail segments.

distributions is that they allow us to model different parts of a response variable with different distributions. In addition, the spliced distribution allows the inclusion of point mass distributions.

The density function given in Equation (1) can be written compactly as:

$$f(x) = \sum_{i=1}^{n} I_{C_i}(x) a_i f_i(x), \tag{2}$$

where I is an indicator function. The cumulative distribution function can be expressed as

$$F(x) = \sum\nolimits_{i=1}^{n} I_{C_i}(x)\left(\sum\nolimits_{j=1}^{i-1} a_j + a_i F_i(x)\right), \tag{3}$$

where F_i is the corresponding cumulative distribution function of f_i in the interval C_i. Figure 1 shows the shape of a hypothetical density function of a three-part model with clear marks of the thresholds.

We now consider a spliced distribution with three components. Auto insurance policies often include deductibles. If a claim amount is less than the deductible, the claim amount is not observed. Let n be the number of policies in the dataset that are used to build the regression model. For $i = 1, 2, ..., n$, let Y_i^* be the unobserved, or latent, variable that represents the total claim amount of the ith policy. Let $\mathrm{Y_i}$ by the observed variable that represents the total claim payment made by the insurance company for the ith policy. Suppose that the deductible of the ith policy is $\mathrm{d_i}$. Then we have

$$Y_i = \begin{cases} 0, & \text{if } Y_i^* \le d_i, \\ Y_i^* - d_i, & \text{if } Y_i^* > d_i. \end{cases}$$

Suppose that Y_i^* follows a three-component spliced distribution whose density is given by

$$f^*(y_i^*) = \begin{cases} p_1 \dfrac{f_0(y_i^*; \eta_i)}{F_0(d_i; \eta_i)}, & y_i^* \in [0, d_i], \\ p_2 \dfrac{f_1(y_i^*; \alpha_i)}{F_1(c + d_i; \alpha_i) - F_1(d_i; \alpha_i)}, & y_i^* \in (d_i, c + d_i], \\ (1 - p_1 - p_2) \dfrac{f_2(y_i^*; \gamma_i)}{1 - F_2(c + d_i; \gamma_i)}, & y_i^* \in (c + d_i, \infty), \end{cases} \tag{4}$$

where $p_1 > 0, p_2 > 0$, $c > 0$, and f_j and F_j are the probability density and cumulative distribution functions of the distribution for the $(j+1)$th component for $j = 0$, 1, 2. Here η_i, α_i, and γ_i are parameters associated with the distributions for the three components, respectively. The sum of the weights p_1 and p_2 is less than 1: $p_1 + p_2 < 1$. The threshold c is a parameter to be estimated from the data.

Let f be the probability density function for the observed variable Y_i. Noting that

$$f(0) = P(Y_i = 0) = P\left(Y_i^* \le d_i\right) = p_1,$$

we have from Equation (4)

$$f(y_i) = \begin{cases} p_1, & y_i = 0, \\ p_2 \dfrac{f_1(y_i + d_i; \alpha_i)}{F_1(c + d_i; \alpha_i) - F_1(d_i; \alpha_i)}, & y_i \in (0, c], \\ (1 - p_1 - p_2) \dfrac{f_2(y_i + d_i; \gamma_i)}{1 - F_2(c + d_i; \gamma_i)}, & y_i \in (c, \infty). \end{cases} \tag{5}$$

To estimate the parameters of the spliced regression model, we can use the method of maximum likelihood by maximizing the following log-likelihood function:

$$\begin{aligned} L(\theta) = {} & \log p_1 \sum_{i=1}^{n} I_{\{0\}}(y_i) + \sum_{i=1}^{n} I_{(0,c]}(y_i) \left[\log p_2 + \log f_1(y_i + d_i; \alpha_i) - \log(F_1(c + d_i; \alpha_i) - F_1(d_i; \alpha_i))\right] \\ & + \sum_{i=1}^{n} I_{(c,\infty)}(y_i) \left[\log(1 - p_1 - p_2) + \log f_2(y_i + d_i; \gamma_i) - \log(1 - F_2(c + d_i; \gamma_i))\right]. \end{aligned} \tag{6}$$

Note that the weights p_1 and p_2 are estimated from the proportions of points falling in the intervals $[0, 0]$ and $(0, c]$, respectively. This can be shown by solving the following estimating equations:

$$\frac{\partial L}{\partial p_1} = \frac{1}{p_1} \sum_{i=1}^{n} I_{\{0\}}(y_i) - \frac{1}{1 - p_1 - p_2} \sum_{i=1}^{n} I_{(c,\infty)}(y_i) = 0$$

and

$$\frac{\partial L}{\partial p_2} = \frac{1}{p_2} \sum_{i=1}^{n} I_{(0,c]}(y_i) - \frac{1}{1 - p_1 - p_2} \sum_{i=1}^{n} I_{(c,\infty)}(y_i) = 0.$$

In general, we assume that the threshold c is unknown and estimated from the data. We use the method of maximum likelihood to estimate the threshold and the rest of the parameters in the model. We anticipate challenges because the likelihood at the threshold will be unstable leading us to possible difficulties of solving the optimization problem. To overcome these problems, we propose to use a multistage optimization approach (Gan and Valdez 2017) to estimate the parameters. The subsequent section shows the effectiveness of this approach.

3. SPLICED REGRESSION MODELS WITH COVARIATES

In this section we propose two specific spliced regression models. There are many ways to specify the spliced distribution given in Equation (5) because we have many choices of distributions for the second and the third components. To specify the distributions for the second and the third components, we consider distributions that are medium-tailed, heavy-tailed, and extreme value distributions (Coles 2001; Kotz and Nadarajah 2000; Panjer 2006; Foss et al. 2013; Klugman et al. 2014). In this article, we consider two specifications of the spliced distribution. The first specification uses a medium-tailed distribution for the second component and a heavy-tailed distribution for the third component. The second specification uses a heavy-tailed distribution for the second component and an extreme value distribution for the third component. In particular, we consider the gamma, Pareto, and Type I Gumbel distributions. The Pareto distribution is typically used to model the heavy tail of a distribution. However, it can also be used to model the central values of a variable by truncation (Aban et al. 2006). The gamma distribution is considered as a medium-tailed distribution and is commonly used to model claim severity (Frees et al. 2014). The Pareto distribution is considered as a heavy-tailed distribution (Foss et al. 2013). The Type I Gumbel distribution is an example of extreme value distributions (Kotz and Nadarajah 2000; Coles 2001).

3.1. Spliced Model 1

In the first specification, we use the gamma distribution and the Pareto distribution to model the second and the third components of the spliced distribution. In this case, we have

$$f_1(y;k,\theta_1) = \frac{y^{k-1}\exp\left(-\frac{y}{\theta_1}\right)}{\theta_1^k\Gamma(k)}, \quad y>0, \tag{7}$$

where $\Gamma(k)$ is the complete gamma function, $k>0$ is a shape parameter, and $\theta_1>0$ is a scale parameter. The corresponding cumulative distribution function is given by

$$F_1(y;k,\theta_1) = \frac{\gamma\left(k,\frac{y}{\theta_1}\right)}{\Gamma(k)}, \tag{8}$$

where $\gamma(k,y)$ is the lower incomplete gamma function defined as

$$\gamma(k,y) = \int_0^y t^{k-1}e^{-t}\ dt.$$

The expectation of a gamma variable Y with density f_1 is

$$E[Y] = k\theta_1. \tag{9}$$

In this specification, we use the Pareto distribution for the third component. The probability density and cumulative distribution functions of the Pareto distribution are given by

$$f_2(y;\alpha,\theta_2) = \frac{\alpha\theta_2^\alpha}{(y+\theta_2)^{\alpha+1}}, \quad y>0 \tag{10}$$

and

$$F_2(y;\alpha,\theta_2) = 1-\left(\frac{\theta_2}{y+\theta_2}\right)^\alpha, \tag{11}$$

respectively, where $\alpha>0$ is the shape parameter and $\theta_2>0$ is the scale parameter. When $\alpha>1$, the expectation of a Pareto variable Y with density f_2 is

$$E[Y] = \frac{\theta_2}{\alpha-1}. \tag{12}$$

Plugging Equations (7), (8), (10), and (11) into Equation (5), we have

$$f(y_i) = \begin{cases} p_1, & y_i = 0, \\ p_2\dfrac{(y_i+d_i)^{k-1}\exp\left(-\dfrac{y_i+d_i}{\theta_1}\right)}{\theta_1^k\left[\gamma\left(k,\dfrac{c+d_i}{\theta_1}\right)-\gamma\left(k,\dfrac{d_i}{\theta_1}\right)\right]}, & y_i\in(0,c], \\ (1-p_1-p_2)\dfrac{\alpha(c+d_i+\theta_2)^\alpha}{(y_i+d_i+\theta_2)^{\alpha+1}}, & y_i\in(c,\infty). \end{cases} \tag{13}$$

One challenge of using the method of maximum likelihood to fit the above probability density function to the claims data is to calculate the logarithm of the term $\gamma(k,\frac{c+d_i}{\theta_1})-\gamma(k,\frac{d_i}{\theta_1})$ because this term can be close to zero during the optimization process. To address this numerical problem, we calculate the logarithm of this term as follows:

$$\begin{aligned}\ln\left[\gamma\left(k,\frac{c+d_i}{\theta_1}\right)-\gamma\left(k,\frac{d_i}{\theta_1}\right)\right] &= \ln\int_{\frac{d_i}{\theta_1}}^{\frac{c+d_i}{\theta_1}} t^{k-1}e^{-t}\ dt \\ &= -\frac{d_i}{\theta_1}+\ln\int_0^{\frac{c}{\theta_1}}\left(t+\frac{d_i}{\theta_1}\right)^{k-1}e^{-t}\ dt.\end{aligned}$$

The corresponding cumulative distribution function is given by

$$F(y_i)=\begin{cases} p_1, & y_i=0,\\ p_1+p_2\dfrac{\gamma\left(k,\dfrac{y_i+d_i}{\theta_1}\right)-\gamma\left(k,\dfrac{d_i}{\theta_1}\right)}{\gamma\left(k,\dfrac{c+d_i}{\theta_1}\right)-\gamma\left(k,\dfrac{d_i}{\theta_1}\right)}, & y_i\in(0,c],\\ 1-(1-p_1-p_2)\left(\frac{c+d_i+\theta_2}{y_i+d_i+\theta_2}\right)^{\alpha}, & y_i\in(c,\infty). \end{cases} \tag{14}$$

We can use the inverse of the cumulative distribution function to simulate claim payments from the fitted model. When the difference $\gamma(k,\frac{c+d_i}{\theta_1})-\gamma(k,\frac{d_i}{\theta_1})$ is close to zero, we need to calculate the underlying fraction:

$$\frac{\gamma\left(k,\dfrac{y_i+d_i}{\theta_1}\right)-\gamma\left(k,\dfrac{d_i}{\theta_1}\right)}{\gamma\left(k,\dfrac{c+d_i}{\theta_1}\right)-\gamma\left(k,\dfrac{d_i}{\theta_1}\right)}=\frac{\displaystyle\int_{\frac{d_i}{\theta_1}}^{\frac{y_i+d_i}{\theta_1}} t^{k-1}e^{-t}\ dt}{\displaystyle\int_{\frac{d_i}{\theta_1}}^{\frac{c+d_i}{\theta_1}} t^{k-1}e^{-t}\ dt}$$

$$=\frac{\displaystyle\int_0^{\frac{y_i}{\theta_1}}\left(t+\frac{d_i}{\theta_1}\right)^{k-1}e^{-t}\ dt}{\displaystyle\int_0^{\frac{c}{\theta_1}}\left(t+\frac{d_i}{\theta_1}\right)^{k-1}e^{-t}\ dt}.$$

To estimate the expected claim payments for a policy, we can calculate the expectation from the model. When $\alpha>1$, the expectation of Y_i is given by

$$E[Y_i]=p_2\left(\theta_1\frac{\gamma\left(k+1,\dfrac{c+d_i}{\theta_1}\right)-\gamma\left(k+1,\dfrac{d_i}{\theta_1}\right)}{\gamma\left(k,\dfrac{c+d_i}{\theta_1}\right)-\gamma\left(k,\dfrac{d_i}{\theta_1}\right)}-d_i\right) + (1-p_1-p_2)\frac{\alpha c+d_i+\theta_2}{\alpha-1}. \tag{15}$$

To ensure numerical stability, we also need to calculate the above expectation carefully when the difference $\gamma(k,\frac{c+d_i}{\theta_1})-\gamma(k,\frac{d_i}{\theta_1})$ is close to zero. In such cases, we can calculate the fraction as follows:

$$\frac{\gamma\left(k+1,\dfrac{c+d_i}{\theta_1}\right)-\gamma\left(k+1,\dfrac{d_i}{\theta_1}\right)}{\gamma\left(k,\dfrac{c+d_i}{\theta_1}\right)-\gamma\left(k,\dfrac{d_i}{\theta_1}\right)}=\frac{\int_{\frac{d_i}{\theta_1}}^{\frac{c+d_i}{\theta_1}} t^{k}e^{-t}\ dt}{\int_{\frac{d_i}{\theta_1}}^{\frac{c+d_i}{\theta_1}} t^{k-1}e^{-t}\ dt}$$

$$=k+\frac{\left(\frac{d_i}{\theta_1}\right)^k-\left(\frac{c+d_i}{\theta_1}\right)^k\exp\left(-\dfrac{c}{\theta_1}\right)}{\int_0^{\frac{c}{\theta_1}}\left(t+\frac{d_i}{\theta_1}\right)^{k-1}e^{-t}\ dt}.$$

Covariates and exposures can be incorporated into the spliced regression model through the scale parameters θ_1 and θ_2. Let $\mathbf{x}_i$ be a numerical vector representing the covariate values and E_i be the exposure of the ith policy. Then we can let the scale parameters to be dependent on $\mathbf{x}_i$ as

$$\theta_1=E_i\exp\left(\beta_1'\mathbf{x}_i\right),\quad \theta_2=E_i\exp\left(\beta_2'\mathbf{x}_i\right), \tag{16}$$

where β_1 and β_2 are vectors of regression parameters.

3.2. Spliced Model 2

In the second specification of the spliced model, we use the Pareto distribution for the second component and the Type-I Gumbel distribution for the third component. In this case, we have

$$f_1(y; \alpha, \theta_1) = \frac{\alpha\theta_1^\alpha}{(x+\theta_1)^{\alpha+1}}, \quad y>0 \tag{17}$$

and

$$F_1(y; \alpha, \theta_1) = 1 - \left(\frac{\theta_1}{y+\theta_1}\right)^\alpha, \tag{18}$$

respectively, where $\alpha>0$ is the shape parameter and $\theta_1>0$ is the scale parameter.

The probability density function of the Type I Gumbel distribution is defined as (Kotz and Nadarajah 2000)

$$f_2(y; \mu, \theta_2) = \frac{1}{\theta_2}\exp\left[-\frac{y-\mu}{\theta_2} - \exp\left(-\frac{y-\mu}{\theta_2}\right)\right], \quad y \in (-\infty, \infty), \tag{19}$$

where μ is the location parameter and $\theta_2>0$ is the scale parameter. The corresponding cumulative distribution function of the Type I Gumbel distribution is defined as

$$F_2(y; \mu, \theta_2) = \exp\left[-\exp\left(-\frac{y-\mu}{\theta_2}\right)\right], \quad y \in (-\infty, \infty). \tag{20}$$

The expectation of a Type I Gumbel variable Y with density f_2 is given by

$$E[Y] = \mu + \gamma\theta_2,$$

where γ is the Euler-Mascheroni constant.

Plugging Equations (17), (18), (19), and (20) into Equation (5), we get the probability density function of the spliced distribution as:

$$f(y_i) = \begin{cases} p_1, & y_i = 0, \\ p_2 \dfrac{\alpha(y_i+d_i+\theta_1)^{-\alpha-1}}{(d_i+\theta_1)^{-\alpha} - (c+d_i+\theta_1)^{-\alpha}}, & y_i \in (0, c], \\ p_3 \dfrac{\frac{1}{\theta_2} e^{-\frac{y_i+d_i-\mu}{\theta_2} - \exp\left(-\frac{y_i+d_i-\mu}{\theta_2}\right)}}{1 - \exp\left[-\exp\left(-\dfrac{c+d_i-\mu}{\theta_2}\right)\right]}, & y_i \in (c, \infty), \end{cases} \tag{21}$$

where $p_3 = 1 - p_1 - p_2$. Similar as in the first specification, calculating the log-likelihood function naively leads to numerical instability because the term $1 - \exp\left[-\exp\left(-\frac{c+d_i-\mu}{\theta_2}\right)\right]$ can be close to zero. To calculate the logarithm of this term, we notice that

$$\lim_{x\to 0}\frac{1-e^{-x}}{x} = 1, \tag{22}$$

which gives the approximation

$$\ln(1-e^{-x}) \approx \ln(x),$$

where x is close to zero. Then we calculate the logarithm of this term as follows:

$$\begin{cases} \ln\left(1-\exp\left[-\exp\left(-\dfrac{c+d_i-\mu}{\theta_2}\right)\right]\right), & \text{if } \dfrac{c+d_i-\mu}{\theta_2}<20, \\ -\dfrac{c+d_i-\mu}{\theta_2}, & \text{if } \dfrac{c+d_i-\mu}{\theta_2}\geq 20. \end{cases}$$

The cumulative distribution function of the above spliced distribution is given by

$$F(y_i)=\begin{cases} p_1, & y_i=0, \\ p_1+p_2\dfrac{(d_i+\theta_1)^{-\alpha}-(y_i+d_i+\theta_1)^{-\alpha}}{(d_i+\theta_1)^{-\alpha}-(c+d_i+\theta_1)^{-\alpha}}, & y_i\in(0,c], \\ 1-p_3\dfrac{1-\exp\left[-\exp\left(-\dfrac{y_i+d_i-\mu}{\theta_2}\right)\right]}{1-\exp\left[-\exp\left(-\dfrac{c+d_i-\mu}{\theta_2}\right)\right]}, & y_i\in(c,\infty). \end{cases} \tag{23}$$

When the term $\exp\left(-\frac{c+d_i-\mu}{\theta_2}\right)$ is close to zero, the term $\exp\left(-\frac{y_i+d_i-\mu}{\theta_2}\right)$ will also be close to zero because $y_i\in(c,\infty)$. In such cases, we can use Equation (22) to calculate the fraction:

$$\frac{1-\exp\left[-\exp\left(-\frac{y_i+d_i-\mu}{\theta_2}\right)\right]}{1-\exp\left[-\exp\left(-\frac{c+d_i-\mu}{\theta_2}\right)\right]}\approx\exp\left(\frac{c-y_i}{\theta_2}\right).$$

When $\alpha\neq 1$, the expectation of $\mathrm{Y_i}$ is given by

$$\begin{aligned} E[Y_i]=p_2&\left(\frac{\alpha}{\alpha-1}\cdot\frac{(d_i+\theta_1)^{-\alpha+1}-(c+d_i+\theta_1)^{-\alpha+1}}{(d_i+\theta_1)^{-\alpha}-(c+d_i+\theta_1)^{-\alpha}}-d_i-\theta_1\right) \\ &+(1-p_1-p_2)\left(\mu-d_i-\frac{\theta_2\int_0^{\exp\left(-\frac{c+d_i-\mu}{\theta_2}\right)}e^{-y}\ln y\ dy}{1-\exp\left[-\exp\left(-\dfrac{c+d_i-\mu}{\theta_2}\right)\right]}\right). \end{aligned} \tag{24}$$

Calculating the last fraction in the above expectation naively can also lead to numerical problems because the term $\exp(-\frac{c+d_i-\mu}{\theta_2})$ can be close to zero or very large. To calculate the fraction correctly when the term is close to zero, we use that

$$\lim_{z\to 0}\frac{\int_0^z e^{-y}\ln y\ dy}{(1-e^{-z})\ln z}=1,$$

which gives the approximation

$$\frac{\displaystyle\int_0^{\exp\left(-\frac{c+d_i-\mu}{\theta_2}\right)}e^{-y}\ln y\ dy}{1-\exp\left[-\exp\left(-\frac{c+d_i-\mu}{\theta_2}\right)\right]}\approx-\frac{c+d_i-\mu}{\theta_2}$$

when $\frac{c+d_i-\mu}{\theta_2}\geq 20$. To calculate the fraction correctly when $\exp(-\frac{c+d_i-\mu}{\theta_2})$ is very large, and so we use the following:

$$\lim_{z\to\infty}\int_0^z e^{-y}\ln y\ dy=-\gamma\approx-0.5772156649,$$

where γ is the the Euler-Mascheroni constant.

TABLE 1
Summary of Categorical Variables of the Training Dataset

Variable	Category	Zeros	Nonzeros	Total Count
vehicleAge	0–2	4989 (91.4%)	468 (8.6%)	5457
	3–5	8310 (91.6%)	766 (8.4%)	9076
	6–10	5533 (92.8%)	428 (7.2%)	5961
	11–15	2773 (96.4%)	103 (3.6%)	2876
	15+	1130 (97.3%)	31 (2.7%)	1161
gender	F	4144 (93.2%)	304 (6.8%)	4448
	M	18591 (92.6%)	1492 (7.4%)	20083
age	0–22	32 (94.1%)	2 (5.9%)	34
	23–25	322 (92.5%)	26 (7.5%)	348
	26–35	7232 (92.0%)	628 (8.0%)	7860
	36–45	8247 (92.9%)	633 (7.1%)	8880
	46–55	4561 (92.9%)	348 (7.1%)	4909
	56–65	1981 (93.6%)	135 (6.4%)	2116
	65+	360 (93.8%)	24 (6.3%)	384
NCD	0	4140 (91.4%)	388 (8.6%)	4528
	10	2251 (92.0%)	197 (8.0%)	2448
	20	2192 (91.3%)	210 (8.7%)	2402
	30	2276 (92.6%)	181 (7.4%)	2457
	40	1925 (92.5%)	157 (7.5%)	2082
	50	9951 (93.8%)	663 (6.2%)	10614

TABLE 2
Summary of Continuous Variables of the Training Dataset

Variable	Min	1st Q	Mean	Median	3rd Q	Max
payment	0	0	319.2635	0	0	183572.5
exposure	0.0027	0.5041	0.7637	1	1	1
deductible	0	0	43.2926	0	0	5000

Similarly as in the first specification, we can incorporate covariates and exposures through the scale parameters as shown in Equation (16).

4. APPLICATION IN MODELING AUTO CLAIMS

4.1. Data Description

To demonstrate the performance of the spliced regression model, we obtained an auto dataset from a Singapore insurance company. The auto dataset contains information about the vehicle age, the gender and age of the policyholder, and the NCD (No Claim Discount) of the policy. We converted the vehicle age and the age of the policyholder into several buckets and treated them as factors.

We organized the dataset in calendar years and used the data in calendar year 1994 as the training data and the data in calendar year 1995 as the test data. The training dataset contains 24,531 records, and the test dataset contains 22,684 records. We will use the training dataset for fitting the models and use the test dataset for out-of-sample validation.

Tables 1 and 2 show the summaries of the categorical variables and the continuous variables of the training dataset, respectively. The payment is the sum of all payments made by the insurer to a policy. From Table 1 we see that the number of positive payments decreases when the vehicle age increases. The relationship between the number of positive payments and the

TABLE 3
Summary of Categorical Variables of the Test Dataset

Variable	Category	Zeros	Nonzeros	Total Count
vehicleAge	0–2	4024 (91.2%)	386 (8.8%)	4410
	3–5	6726 (91.4%)	635 (8.6%)	7361
	6–10	6597 (91.8%)	588 (8.2%)	7185
	11–15	2154 (97.1%)	65 (2.9%)	2219
	15+	1489 (98.7%)	20 (1.3%)	1509
gender	F	4101 (92.8%)	318 (7.2%)	4419
	M	16889 (92.5%)	1376 (7.5%)	18265
age	0–22	10 (90.9%)	1 (9.1%)	11
	23–25	107 (93.0%)	8 (7.0%)	115
	26–35	4999 (91.7%)	454 (8.3%)	5453
	36–45	7582 (92.6%)	605 (7.4%)	8187
	46–55	5181 (92.6%)	415 (7.4%)	5596
	56–65	2548 (93.4%)	181 (6.6%)	2729
	65+	563 (94.9%)	30 (5.1%)	593
NCD	0	2114 (92.3%)	177 (7.7%)	2291
	10	1896 (91.9%)	168 (8.1%)	2064
	20	1895 (89.0%)	235 (11.0%)	2130
	30	1665 (92.6%)	133 (7.4%)	1798
	40	1699 (93.3%)	122 (6.7%)	1821
	50	11721 (93.2%)	859 (6.8%)	12580

TABLE 4
Summary of Continuous Variables of the Test Dataset

Variable	Min	1st Q	Mean	Median	3rd Q	Max
payment	0	0	308.0030	0	0	131,545.1
exposure	0.0027	0.5836	0.7935	1	1	1
deductible	0	0	39.9731	0	0	7000

NCD is negative in general. From Table 2 we see that most payments and deductibles are zero. The minimum exposure is 0.0027 years or one day. The maximum exposure is one year.

Tables 3 and 4 show the summaries of the categorical variables and the continuous variables of the test dataset, respectively. We see similar patterns as in the training dataset.

Figure 2 shows the histograms of the positive payments of the training dataset and the test dataset. Since more than 90% of the payments are zero, we omitted zero payments in the histograms so that we can see the distribution of positive payments in detail. From the histograms we see that the distributions of the payments have long tails. Most of the payments are small, but there are a few quite large payments.

4.2. Estimation Results

In this section we evaluate the spliced regression models empirically using the Singapore auto datasets. In particular, we compare the spliced regression models and the Tweedie regression model (Smyth and Jørgensen 2002; Frees et al. 2016). Unlike the frequency-severity model (Frees 2009), both the spliced regression models and the Tweedie regression model directly model the loss costs.

In recent years, there has been an increase in interest of the use of Tweedie exponential family models to fit loss models. See, for example, Frees et al. (2016). The Tweedie family of distributions belong to the exponential family with a variance

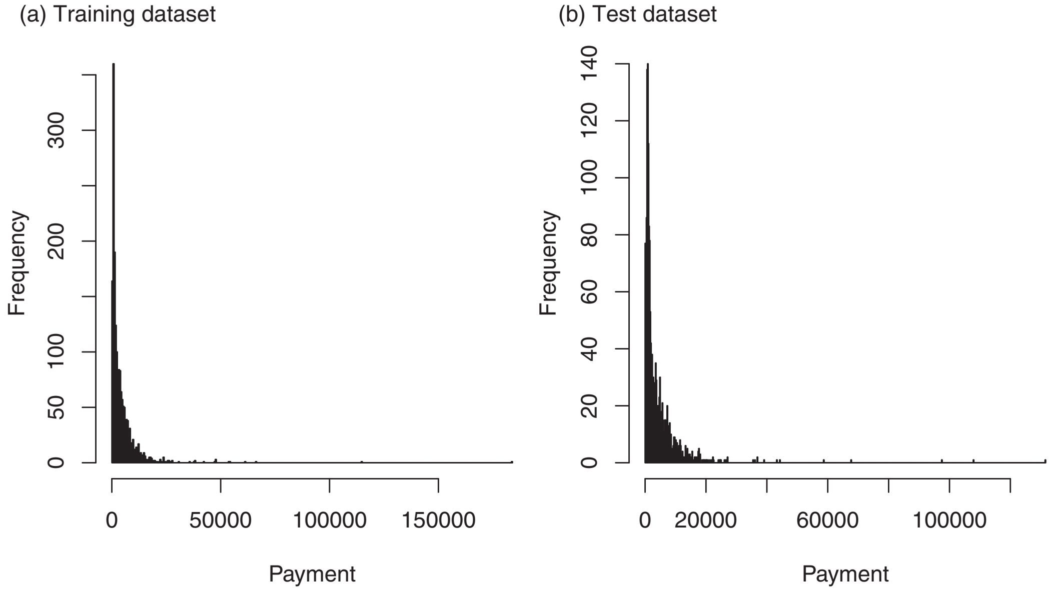

FIGURE 2. Histograms of Positive Payments.

TABLE 5
Values of Log-Likelihood Function, AIC, and BIC of Spliced Models

Model	Log-likelihood	AIC	BIC
Spliced Model 1	−23,212.96	46,499.91	46,799.90
Spliced Model 2	−23,124.27	46,322.53	46,622.52
Tweedie Model	−29,313.89	58,663.78	58,809.72

function of the power form as $V(\mu) = \tau\mu^p$, for p not in (0, 1). However, when $1<p<2$, the Tweedie distribution has a compound Poisson-gamma interpretation with a probability mass at zero. Although in this case, there is no explicit expression for the density function, the primary advantage of fitting such Tweedie models is to avoid the two-part model of fitting the frequency and then the severity.

We used the R function optim to fit the spliced regression model to the training dataset by maximizing the log-likelihood function given in Equation (6). To make the parameters have similar magnitudes, we reparameterize the threshold c by using a logarithmic transformation: $s = \ln(c)$. Instead of inputing c into the optimization algorithm directly, we input s into the optim function. For the second spliced regression model, we did the same for the parameter μ by using $r = \ln(\mu)$.

To fit the Tweedie regression model to the training dataset, we followed a two-step process. First, we used the R function tweedie.profile to estimate the Tweedie index parameter p. Second, we used the R function glm with the Tweedie family to fit the Tweedie regression model to the training dataset.

Table 5 shows the values of the log-likelihood function, the AIC, and the BIC of the two spliced regression models as well as the Tweedie model. These values are calculated based on the training dataset. From the table, we see that the second spliced regression model provides the best fit among the three models. The two spliced regression models are much better than the Tweedie model, which fits the data the worst among the three models.

Once we fitted the models to the training dataset, we used the models to calculate the expected payments for each policies. Table 6 shows the total payments aggregated from the expected payments of individual policies. The table also shows the actual aggregate payments from the datasets. From Table 6 we see that the total payments obtained from the first spliced model match the actual total payments the best among the three models. The second spliced regression model overestimated the total payments. This is reasonable because the distributions used by the second spliced model have heavier tails than those used by the first spliced model. As a result, the total payment estimated by the second spliced model is higher than that estimated by the first spliced model. From Table 6 we also see that the Tweedie regression model underestimated the total payments. The results show that the spliced regression model fits the tail better than the Tweedie model does.

TABLE 6
Total Payments Predicted by Models and Actual Aggregate Payments

	Training Data	Std Error	Test Data	Std Error
Actual amount	7,831,853.44	na	6,986,738.99	na
Spliced model 1	7,546,625.74	523,296.10	7,045,004.06	480,257.90
Spliced model 2	9,544,723.60	315,216.90	8,671,783.07	310,155.40
Tweedie model	7,121,984.37	401,990.40	6,309,039.75	351,389.50

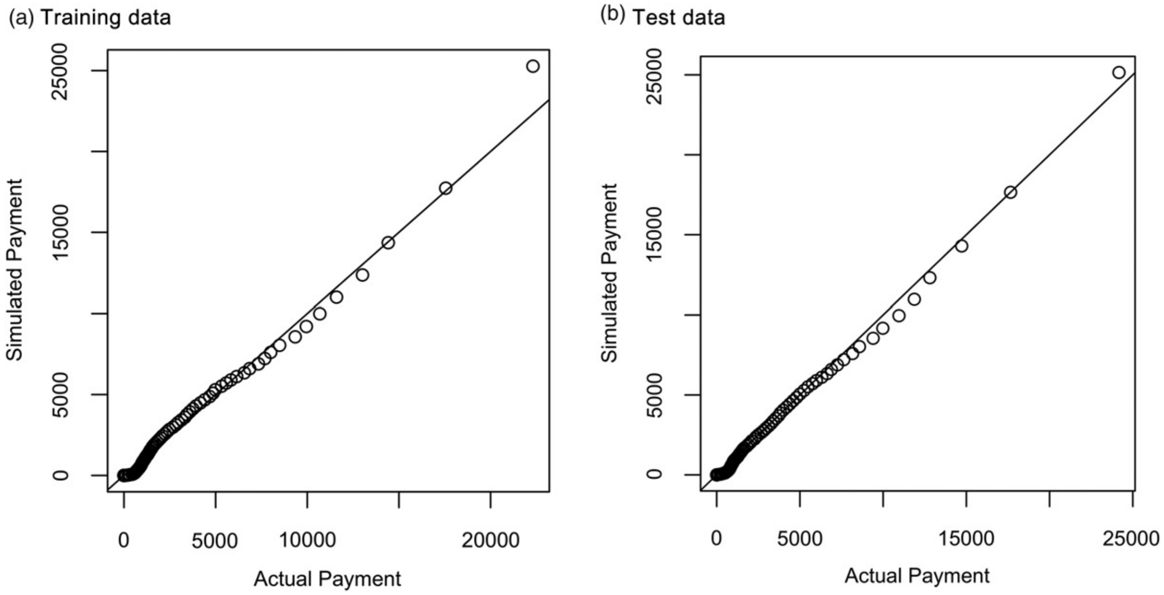

FIGURE 3. QQ Plots of Actual Payments and Those Simulated from Spliced Model 1.

In addition to the aggregate expected payments, Table 6 shows the standard errors of the aggregated payments. The standard errors were obtained from 1000 sets of simulations from the fitted models. Let n be the number of policies in the dataset. For $j = 1, 2, ..., 1000$, the jth set of simulated payments, $\{\tilde{y}_{1j}, \tilde{y}_{2j}, ..., \tilde{y}_{nj}\}$ was generated as follows. We first generated n uniform random numbers $r_{1j}, r_{2j}, ...,$ $\mathrm{r_{nj}}$. Then we obtained the simulated payments by using the inverse method,

$$\tilde{y}_{ij} = F^{-1}(r_{ij}), \quad i = 1, 2, ..., n,$$

where F is the cumulative distribution function of the corresponding model. For the ith simulated payment, the covariate information of the ith policy was used. The total payment of the jth set of simulated payments is calculated as

$$S_j = \sum_{i=1}^{n} \tilde{y}_{ij}, \quad j = 1, 2, ..., 1000.$$

The standard error is calculated as the standard deviation of the 1000 total payments $\mathrm{S_1}$, $\mathrm{S_2}$, ..., $\mathrm{S_{1000}}$.

From Table 6 we see that the second spliced model has the lowest standard errors. However, the actual amount is not within two standard deviations of the predicted amount.

To see the performance of the models in terms of fitting the data, we simulated payments from the fitted models and compared the QQ plots between the simulated payments and the actual payments from the data. Since covariates were incorporated into the models, we simulated 10 payments for each policy by using the covariate values of the policy. Then we created QQ plots using the simulated payments and the actual payments from the data.

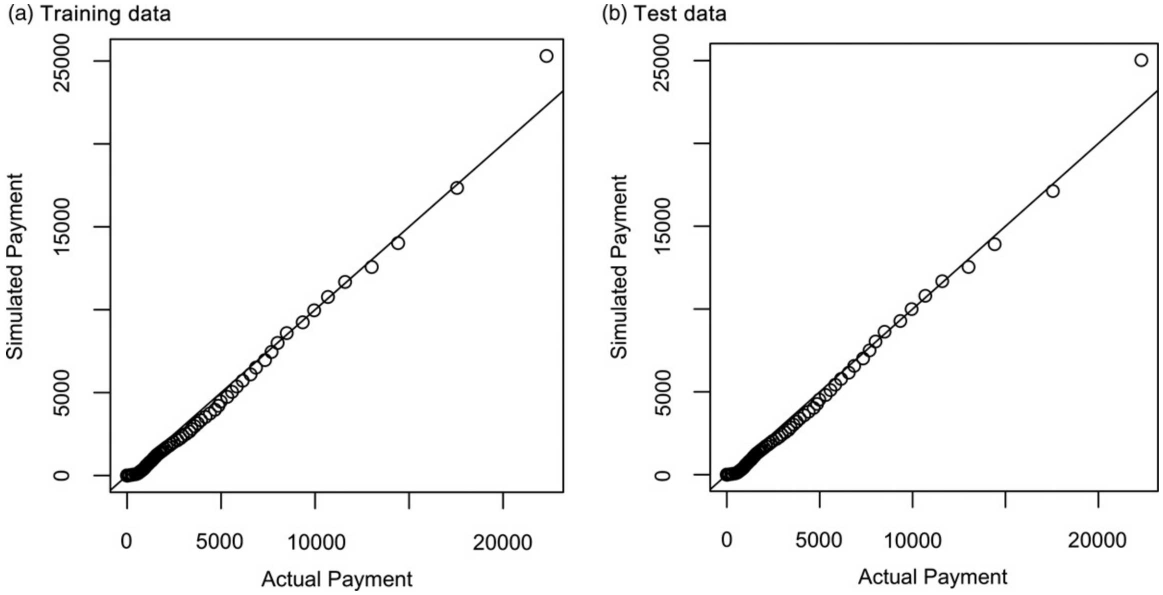

FIGURE 4. QQ Plots of Actual Payments and Those Simulated from Spliced Model 2.

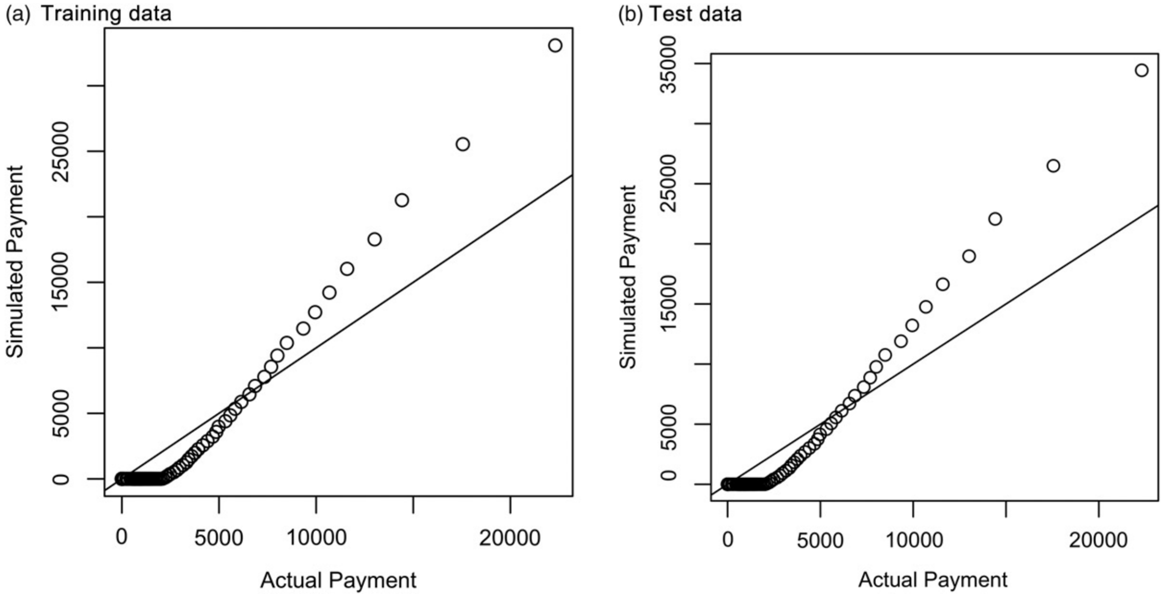

FIGURE 5. QQ Plots of Actual Payments and Those Simulated from Tweedie Model.

Figure 3 shows the QQ plots of the actual payments and those simulated from the first spliced model. The QQ plots show that the first spliced model fits the data well. Similarly Figure 4 shows the QQ plots of the actual payments and those simulated from the second spliced model. The QQ plots in Figure 4 show that the second spliced model also fits the data well. However, the first spliced model is better than the second spliced model in terms of predicting the total payments (see Table 6). Figure 5 shows the QQ plots of the actual payments and those simulated from the Tweedie model. The QQ plots clearly show that the Tweedie model does not fit the data well.

Table 7 shows the estimates of the parameters of the spliced regression model. The covariates corresponding to the regression coefficients are given in Table 8. From Table 7 we see that the threshold estimated by the second spliced model is much larger than that estimated by the first spliced model. In the first spliced model, the parameters k and α represent the shape parameters of the gamma distribution and the Pareto distribution, respectively. In the second spliced model, the parameters α and μ represent the shape parameter of the Pareto distribution and the location parameter of the Type I Gumbel distribution, respectively. The parameters β_1 and β_2 represent the regression coefficients of the covariates. Since the covariates are

TABLE 7
Estimates of Parameters of Spliced Models

Spliced Model 1			Spliced Model 2		
Parameter	Estimate	Std Error	Parameter	Estimate	Std Error
$s(\ln c)$	8.6295	0.0083	$s(\ln c)$	9.4213	0.0030
k	0.6803	0.0000	α	0.6501	0.1229
$\beta_{1,0}$	8.3411	0.0002	$\beta_{1,0}$	9.1800	5.8871
$\beta_{1,1}$	1.0461	0.4533	$\beta_{1,1}$	−0.2065	0.1886
$\beta_{1,2}$	0.9252	0.2904	$\beta_{1,2}$	−0.1848	0.2120
$\beta_{1,3}$	−0.6725	0.2582	$\beta_{1,3}$	−0.7603	0.2750
$\beta_{1,4}$	5.3341	4.2176	$\beta_{1,4}$	−0.5618	0.4174
$\beta_{1,5}$	0.1049	0.0000	$\beta_{1,5}$	−0.2709	0.1917
$\beta_{1,6}$	1.7755	3.3111	$\beta_{1,6}$	−0.2716	5.8128
$\beta_{1,7}$	0.5277	0.0000	$\beta_{1,7}$	−0.6667	5.8652
$\beta_{1,8}$	0.8395	0.2587	$\beta_{1,8}$	−0.5569	5.8648
$\beta_{1,9}$	1.0093	0.4251	$\beta_{1,9}$	−0.3384	5.8633
$\beta_{1,10}$	0.2623	0.4438	$\beta_{1,10}$	−0.6206	5.8564
$\beta_{1,11}$	3.9462	4.8294	$\beta_{1,11}$	−0.1005	5.8278
$\beta_{1,12}$	−1.3588	0.0000	$\beta_{1,12}$	−0.6487	0.2712
$\beta_{1,13}$	−0.5086	0.3513	$\beta_{1,13}$	−0.6493	0.2616
$\beta_{1,14}$	0.3718	0.7775	$\beta_{1,14}$	−0.2805	0.2840
$\beta_{1,15}$	−0.5154	0.4369	$\beta_{1,15}$	−0.6575	0.2952
$\beta_{1,16}$	−0.3042	0.3633	$\beta_{1,16}$	−0.9450	0.2193
α	2.0481	0.1311	$r\ (\ln \mu)$	7.0983	1.9193
$\beta_{2,0}$	6.9416	5.8761	$\beta_{2,0}$	7.4586	229.8717
$\beta_{2,1}$	−0.5493	1.4543	$\beta_{2,1}$	0.1635	0.2716
$\beta_{2,2}$	0.5422	1.9854	$\beta_{2,2}$	−0.3263	0.3127
$\beta_{2,3}$	1.0285	1.8312	$\beta_{2,3}$	−1.5175	0.8096
$\beta_{2,4}$	0.0450	4.9411	$\beta_{2,4}$	−2.6449	1.8361
$\beta_{2,5}$	0.1242	2.7053	$\beta_{2,5}$	0.6958	0.2813
$\beta_{2,6}$	0.9743	5.6850	$\beta_{2,6}$	0.4559	229.8785
$\beta_{2,7}$	−1.0308	6.4777	$\beta_{2,7}$	0.9574	229.8526
$\beta_{2,8}$	−0.4936	5.4869	$\beta_{2,8}$	1.5983	229.8462
$\beta_{2,9}$	−0.5263	5.5260	$\beta_{2,9}$	1.4133	229.8528
$\beta_{2,10}$	−0.1779	6.1527	$\beta_{2,10}$	1.9811	229.8601
$\beta_{2,11}$	−1.0659	5.9968	$\beta_{2,11}$	1.2883	229.8640
$\beta_{2,12}$	−0.3517	3.1792	$\beta_{2,12}$	0.9232	0.3875
$\beta_{2,13}$	1.8602	4.1704	$\beta_{2,13}$	−0.5996	0.3493
$\beta_{2,14}$	−0.7991	4.0653	$\beta_{2,14}$	−0.1985	0.3679
$\beta_{2,15}$	0.4104	3.8923	$\beta_{2,15}$	−0.0058	0.4113
$\beta_{2,16}$	−0.5568	3.6407	$\beta_{2,16}$	−0.1548	0.2597

incorporated into the spliced models through the scale parameters and the expectations of the spliced distributions are not related to the scale parameters proportionally, it is not straightforward to interpret the regression coefficients.

Table 9 shows the estimates of the parameters of the Tweedie regression model. The parameters p and ϕ represent the Tweedie index parameter and the dispersion parameter, respectively. The parameter β contains the regression coefficients of the covariates. In our Tweedie model, the covariates are incorporated into the model through the mean as

$$\mu_i = E_i \exp\left(\mathbf{x}_i'\beta\right).$$

TABLE 8
Regression Coefficients and Corresponding Covariates

Regression Parameter	Covariate
β_0	Intercept
β_1	vehicleAge2
β_2	vehicleAge3
β_3	vehicleAge4
β_4	vehicleAge5
β_5	genderM
β_6	age2
β_7	age3
β_8	age4
β_9	age5
β_{10}	age6
β_{11}	age7
β_{12}	NCD10
β_{13}	NCD20
β_{14}	NCD30
β_{15}	NCD40
β_{16}	NCD50

TABLE 9
Estimates of Parameters of Tweedie Model with the Power Parameter and the Dispersion Parameter Estimated to be 1.5612 and 657.4602, Respectively

Parameter	Estimate	Std Error
β_0 (Intercept)	5.7226	1.5487
β_1 (vehicleAge2)	0.0895	0.1212
β_2 (vehicleAge3)	−0.2224	0.1372
β_3 (vehicleAge4)	−1.3622	0.2045
β_4 (vehicleAge5)	−1.8829	0.3245
β_5 (genderM)	0.2632	0.1259
β_6 (age2)	1.0020	1.5899
β_7 (age3)	0.6709	1.5456
β_8 (age4)	0.7667	1.5462
β_9 (age5)	0.8349	1.5482
β_{10} (age6)	0.7512	1.5537
β_{11} (age7)	0.8343	1.5944
β_{12} (NCD10)	−0.2913	0.1826
β_{13} (NCD20)	−0.3927	0.1863
β_{14} (NCD30)	−0.4318	0.1840
β_{15} (NCD40)	−0.4894	0.1953
β_{16} (NCD50)	−0.7776	0.1404

Interpreting the regression coefficients of the Tweeide model is straightforward. For example, all the estimates of the NCD categories are negative. This means that a policyholder with a positive NCD is expected to have a lower payment than a policyholder with a zero NCD.

In summary, the numerical results demonstrate that the spliced regression models can be used to fit skewed auto claims data. Using a medium-tailed distribution and a heavy-tailed distribution can produce satisfactory results. Using extreme value distributions in the spliced model can overestimate the claim payments. In addition, the spliced regression models outperform the Tweedie regression model in term of fitting the tails.

5. CONCLUDING REMARKS

Frequency-severity models have been widely used to model auto claims data. These models involve modeling separately the claim count and the claim severity. For simplicity, studies have also been conducted to model the aggregate loss data directly; the Tweedie model is an example of such models. In this article, we proposed the spliced regression model for fitting the aggregate loss data directly. In particular, we considered a spliced distribution that consists of three components: the first component contains zeros, the second component models the middle segment of the loss data, and the third component models the tail segment of the loss data. We fitted two spliced regression models to a Singapore auto claims dataset. In the first spliced model, we used a medium-tailed distribution for the second component and a heavy-tailed distribution for the third component. In the second spliced model, we used a heavy-tailed distribution for the second component and an extreme value distribution for the third component.

The numerical results show that both spliced models are superior to the Tweedie model in terms of fitting the tail of the loss distribution. However, spliced models suffer from some drawbacks. One drawback of the spliced regression models is that interpreting the regression coefficients of the spliced models is not as straightforward as in the Tweedie model. Another drawback is that estimating the parameters of the spliced model can be challenging as the log-likelihood function resulted from the spliced distribution is highly nonlinear and non-continuous.

ACKNOWLEDGMENTS

Guojun Gan and Emiliano A. Valdez wish to thank the anonymous reviewers and the co-editor for their helpful comments and feedback which helped improve the final version of this article.

FUNDING

Guojun Gan and Emiliano A. Valdez would like to acknowledge the Committee on Knowledge Extension Research (CKER) of the Society of Actuaries for funding this research.

ORCID

Guojun Gan http://orcid.org/0000-0003-3285-7116

REFERENCES

Aban, I. B., M. M. Meerschaert, and A. K. Panorska. 2006. Parameter estimation for the truncated Pareto distribution. *Journal of the American Statistical Association* 101 (473):270–7.

Coles, S. 2001. *An introduction to statistical modeling of extreme values*. London, England: Springer-Verlag.

Cooray, K., and M. M. Ananda. 2005. Modeling actuarial data with a composite lognormal-Pareto model. *Scandinavian Actuarial Journal* 2005 (5):321–34.

de Jong, P., and G. Z. Heller. 2008. *Generalized linear models for insurance data*. Cambridge, England: Cambridge University Press.

Fang, K., and S. Ma. 2013. Three-part model for fractional response variables with application to Chinese household health insurance coverage. *Journal of Applied Statistics* 40 (5):925–40.

Foss, S., D. Korshunov, and S. Zachary. 2013. *An introduction to heavy-tailed and subexponential distributions*. 2nd ed. New York, NY: Springer-Verlag.

Frees, E. W. 2009. *Regression modeling with actuarial and financial applications*. Cambridge, England: Cambridge University Press.

Frees, E. W., and E. A Valdez. 2008. Hierarchical insurance claims modeling. *Journal of the American Statistical Association* 103 (484):1457–69.

Frees, E. W., G. Lee, and L. Yang. 2016. Multivariate frequency-severity regression models in insurance. *Risks* 4 (1):1–36.

Frees, E. W., R. A. Derrig, and G. Meyers, editors. 2014. *Predictive modeling applications in actuarial science: Volume 1, predictive modeling techniques*. Cambridge, England: Cambridge University Press.

Friedland, J. 2014. *Fundamentals of general insurance actuarial analysis*. Schaumburg, IL: Society of Actuaries.

Gan, G., and E. A. Valdez. 2017. Regression modeling for the valuation of large variable annuity portfolios. *North American Actuarial Journal* 22 (1):40–54.

Gray, R. J., and S. M Pitts. 2012. *Risk modelling in general insurance: From principles to practice*. Cambridge, England: Cambridge University Press.
Klugman, S., H. Panjer, and G. Willmot. 2012. *Loss models: From data to decisions*. 4th ed. Hoboken, NJ: Wiley.
Klugman, S. A., H. H. Panjer, and G. E. Willmot. 2014. *Loss models: Further topics*. Hoboken, NJ: Wiley.
Koenker, R. 2005. *Quantile regression*. Cambridge, England: Cambridge University Press.
Kotz, S., and S. Nadarajah. 2000. *Extreme value distributions: Theory and applications*. London, England: Imperial College Press.
Kudryavtsev, A. A. 2009. Using quantile regression for rate-making. *Insurance: Mathematics and Economics* 45 (2):296–304.
Lee, S. C. K., and X. S. Lin. 2010. Modeling and evaluating insurance losses via mixtures of Erlang distributions. *North American Actuarial Journal* 14 (1):107–30.
McCullagh, P., and J. A. Nelder. 1989. *Generalized linear models*. 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.
McLachlan, G., and D. Peel. 2000. *Finite mixture models*. Hoboken, NJ: Wiley.
Miljkovic, T., and B. Grün. 2016. Modeling loss data using mixtures of distributions. *Insurance: Mathematics and Economics* 70:387–96.
Millar, R. B. 2011. *Maximum likelihood estimation and inference: With examples in R, SAS and ADMB*. West Sussex, England: Wiley.
Nadarajah, S., and S. Bakar. 2014. New composite models for the Danish fire insurance data. *Scandinavian Actuarial Journal* 2014 (2):180–7.
Nadarajah, S., and S. A. A. Bakar. 2013. CompLognormal: An R package for composite lognormal distributions. *R Journal* 5 (2):98–104.
Ohlsson, E., and B. Johansson. 2010. *Non-life insurance pricing with generalized linear models*. Berlin, Germany: Springer.
Panjer, H. H. 2006. *Operational risk: Modeling analytics*. Hoboken, NJ: Wiley.
Parodi, P. 2014. *Pricing in general insurance*. Boca Raton, FL: CRC Press.
Peters, G. W., and P. V. Shevchenko. 2015. *Advances in heavy tailed risk modeling: A handbook of operational risk. Wiley handbooks in financial engineering and econometrics*. Hoboken, NJ: Wiley.
Pigeon, M., and M. Denuit. 2011. Composite lognormal-Pareto model with random threshold. *Scandinavian Actuarial Journal* 2011 (3):177–92.
Scollnik, D. P. M. 2007. On composite lognormal-Pareto models. *Scandinavian Actuarial Journal* 2007 (1):20–33.
Shi, P. 2014. Fat-tailed regression models. In *Predictive modeling applications in actuarial science. Volume I: Predictive modeling techniques*, ed. E. W. Frees, R. A. Derrig, and G. Meyers, 236–59. Cambridge, England: Cambridge University Press.
Smyth, G. K., and B. Jørgensen. 2002. Fitting Tweedie's compound Poisson model to insurance claims data: Dispersion modelling. *ASTIN Bulletin* 32 (1):143–57.
Vernic, R., S. Teodorescu, and E. Pelican. 2009. Two lognormal models for real data. *Annals of Ovidius University, Series Mathematics* 17 (3):263–77.
Wang, X., and D. K. Dey. 2010. Generalized extreme value regression for binary response data: An application to b2b electronic payments system adoption. *Annals of Applied Statistics* 4 (4):2000–23.
Yin, C., and X. S. Lin. 2016. Efficient estimation of Erlang mixtures using Iscad penalty with insurance application. *ASTIN Bulletins* 46 (3):779–99.

Discussions on this article can be submitted until July 1, 2019. 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. PARAMETER ESTIMATION

The spliced regression models have many parameters, which include the regression coefficients for the covariates. As a result, estimating the parameters of the spliced models is quite challenging. We used the R function optim to maximize the log-likelihood function of the spliced models. This function requires initial values of the parameters to be estimated. Supplying good initial values to optim is critical to obtain plausible estimates of the parameters. In this paper, we adopted a two-stage optimization procedure (Millar, 2011) to estimate the parameters. The multistage optimization procedure has been successfully applied in estimating parameters of complex regression models (Gan and Valdez, 2017). In this appendix, we describe the two-stage optimization procedure used to estimate the parameters of the spliced regression models.

In the first stage, we fix the threshold $c = c_0$ and fit the probability distributions to the second and the third components separately. In particular, we set c_0 to be the 90th quantile of the positive payments of the training dataset. To fit the probability distribution to the second component, we use the optim function to maximize the following log-likelihood function:

$$\sum_{i=1}^{n} I_{(0,c_0]}(y_i)\left[\log f_1(y_i + d_i; \alpha_i) - \log(F_1(c + d_i; \alpha_i) - F_1(d_i; \alpha_i))\right].$$

To fit the probability distribution to the third component, we use the optim function to maximize the following log-likelihood function:

$$\sum_{i=1}^{n} I_{(c_0,\infty)}(y_i)[\log f_2(y_i + d_i; \gamma_i) - \log(1 - F_1(c + d_i; \gamma_i))].$$

In the first stage, we use a moment-matching method to set the initial values of the parameters for the optim function. In the first spliced model, we set the shape parameters to be $k = 2$ and $\alpha = 3$. Then we figure out the scale parameters θ_1 and θ_2 by solving the following equations:

TABLE A.1
Initial Values for Fitting the First Spliced Model in the First Stage
(a) The Second Component

Parameter	Initial Values	Optimized Values
k	2	0.5672
$\beta_{1,0}$ (Intercept)	7.4069	8.2747
$\beta_{1,1}$ (vehicleAge2)	0	0.8079
$\beta_{1,2}$ (vehicleAge3)	0	1.1520
$\beta_{1,3}$ (vehicleAge4)	0	−0.3368
$\beta_{1,4}$ (vehicleAge5)	0	4.8966
$\beta_{1,5}$ (genderM)	0	0.2599
$\beta_{1,6}$ (age2)	0	1.6106
$\beta_{1,7}$ (age3)	0	0.4998
$\beta_{1,8}$ (age4)	0	1.5016
$\beta_{1,9}$ (age5)	0	1.7227
$\beta_{1,10}$ (age6)	0	0.7906
$\beta_{1,11}$ (age7)	0	0.1515
$\beta_{1,12}$ (NCD10)	0	−1.4035
$\beta_{1,13}$ (NCD20)	0	−0.7931
$\beta_{1,14}$ (NCD30)	0	−0.1653
$\beta_{1,15}$ (NCD40)	0	−0.4184
$\beta_{1,16}$ (NCD50)	0	−0.6056

(b) The Third Component

Parameter	Initial Values	Optimized Values
α	3	2.0878
$\beta_{2,0}$ (Intercept)	8.7888	9.1814
$\beta_{2,1}$ (vehicleAge2)	0	0.0974
$\beta_{2,2}$ (vehicleAge3)	0	0.1710
$\beta_{2,3}$ (vehicleAge4)	0	0.0337
$\beta_{2,4}$ (vehicleAge5)	0	−0.1755
$\beta_{2,5}$ (genderM)	0	−0.0054
$\beta_{2,6}$ (age2)	0	−0.3913
$\beta_{2,7}$ (age3)	0	−0.5664
$\beta_{2,8}$ (age4)	0	−0.6221
$\beta_{2,9}$ (age5)	0	−0.6399
$\beta_{2,10}$ (age6)	0	−0.6937
$\beta_{2,11}$ (age7)	0	−0.4227
$\beta_{2,12}$ (NCD10)	0	−0.0556
$\beta_{2,13}$ (NCD20)	0	0.0456
$\beta_{2,14}$ (NCD30)	0	0.0442
$\beta_{2,15}$ (NCD40)	0	0.0112
$\beta_{2,16}$ (NCD50)	0	−0.0657

TABLE A.2
Initial Values for Fitting the Second Spliced Model in the First Stage
(a) The Second Component

Parameter	Initial Values	Optimized Values
α	2	0.6026025
$\beta_{1,0}$ (Intercept)	8.689295	9.24843245
$\beta_{1,1}$ (vehicleAge2)	0	−0.19414535
$\beta_{1,2}$ (vehicleAge3)	0	−0.21879452
$\beta_{1,3}$ (vehicleAge4)	0	−0.73420871
$\beta_{1,4}$ (vehicleAge5)	0	−0.50590499
$\beta_{1,5}$ (genderM)	0	−0.31415013
$\beta_{1,6}$ (age2)	0	−0.19595894
$\beta_{1,7}$ (age3)	0	−0.70125607
$\beta_{1,8}$ (age4)	0	−0.53241048
$\beta_{1,9}$ (age5)	0	−0.44496368
$\beta_{1,10}$ (age6)	0	−0.67809813
$\beta_{1,11}$ (age7)	0	0.05108412
$\beta_{1,12}$ (NCD10)	0	−0.65339639
$\beta_{1,13}$ (NCD20)	0	−0.62254226
$\beta_{1,14}$ (NCD30)	0	−0.25293234
$\beta_{1,15}$ (NCD40)	0	−0.64208372
$\beta_{1,16}$ (NCD50)	0	−0.89410846

(b) The Third Component

Parameter	Initial Values	Optimized Values
$r(\ln \mu)$	9.371422	7.09129194
$\beta_{2,0}$ (Intercept)	8.814632	7.43252237
$\beta_{2,1}$ (vehicleAge2)	0	0.08305781
$\beta_{2,2}$ (vehicleAge3)	0	−0.32173068
$\beta_{2,3}$ (vehicleAge4)	0	−1.30299914
$\beta_{2,4}$ (vehicleAge5)	0	−2.3196715
$\beta_{2,5}$ (genderM)	0	0.80742237
$\beta_{2,6}$ (age2)	0	0.35909548
$\beta_{2,7}$ (age3)	0	1.04870881
$\beta_{2,8}$ (age4)	0	1.44567675
$\beta_{2,9}$ (age5)	0	1.35360982
$\beta_{2,10}$ (age6)	0	1.95801864
$\beta_{2,11}$ (age7)	0	0.83477819
$\beta_{2,12}$ (NCD10)	0	0.52773722
$\beta_{2,13}$ (NCD20)	0	−0.54048957
$\beta_{2,14}$ (NCD30)	0	−0.30918772
$\beta_{2,15}$ (NCD40)	0	−0.15583413
$\beta_{2,16}$ (NCD50)	0	−0.25830768

$$\theta_1 \frac{\gamma\left(k+1, \frac{c+d_i}{\theta_1}\right) - \gamma\left(k+1, \frac{d_i}{\theta_1}\right)}{\gamma\left(k, \frac{c+d_i}{\theta_1}\right) - \gamma\left(k, \frac{d_i}{\theta_1}\right)} - d_i = \bar{y}_1, \quad \frac{\alpha c + d_i + \theta_2}{\alpha - 1} = \bar{y}_2,$$

where $\bar{y}_1$ is the average of payments in the interval $(0, c_0]$ and $\bar{y}_2$ is the average of payments in the interval (c_0, ∞). Table A.1 shows the initial values for the optim function as well as the optimized parameter values produced by the optim function. In this stage, we set the initial regression coefficients of all covariates to be zero. The initial intercept coefficients are obtained by taking logarithms of the scale parameters obtained from solving the above equations.

In the second spliced model, we set the shape parameter to be $\alpha = 2$ for the second component and the location parameter $\mu = \frac{\bar{y}_2}{1.5772}$ for the third component. Then we figure out the scale parameters θ_1 and θ_2 by solving the following equations:

$$\frac{\alpha}{\alpha - 1} \cdot \frac{(d_i + \theta_1)^{-\alpha+1} - (c + d_i + \theta_1)^{-\alpha+1}}{(d_i + \theta_1)^{-\alpha} - (c + d_i + \theta_1)^{-\alpha}} - d_i - \theta_1 = \bar{y}_1,$$

$$\mu - d_i - \frac{\theta_2 \int_0^{\exp\left(-\frac{c+d_i-\mu}{\theta_2}\right)} e^{-y} \ln y \, dy}{1 - \exp\left[-\exp\left(-\frac{c + d_i - \mu}{\theta_2}\right)\right]} = \bar{y}_2,$$

where $\bar{y}_1$ is the average of payments in the interval $(0, c_0]$ and $\bar{y}_2$ is the average of payments in the interval (c_0, ∞). Table A.2 shows the initial values for the optim function as well as the optimized parameter values produced by the optim function.

In the second stage, we use the optimized parameters from the first stage as initial values for the optim function and do a full optimization. Note that the weights p_1 and p_2 in the log-likelihood function given in Equation (6) depend on the threshold c. Given the threshold c, the weights can be determined by counting the number of payments in the intervals $(0, c]$ and (c, ∞).