浙江科技学院学报，第 35 卷第 6 期，2023 年 12 月
Journal of Zhejiang University of Science and Technology
Vol. 35 No. 6, Dec. 2023

doi：10.3969/j.issn.1671-8798.2023.06.002

基于半连续两部模型的保险损失预测

鲁亚会[1]，刘爱义[2]

（1.浙江科技学院 经济与管理学院，杭州 310023；2.美国国立卫生研究院，美国 贝塞斯达 20817）

摘　要：【目的】提高保险领域中保单累积损失预测的准确率。传统的 Tweedie 回归模型只能对非零均值建立回归模型，却不能对零概率建立回归模型，从而导致该模型的拟合效果并不理想。**【方法】**考虑到保单损失数据中往往包含着大量的零索赔，此时可视其为一种半连续型数据。因此，基于半连续两部模型，并考虑到累积损失中非零连续部分的分布类型，提出 3 种不同的累积损失预测模型，并结合一组实际损失数据进行模型对比分析。**【结果】**与 Tweedie 回归模型相比，本研究所提出的半连续两部回归模型的赤池信息准则值（Akaike information criterion，AIC）和贝叶斯信息量准则值（Bayesian information criterion，BIC）更小，具有较好的拟合效果。**【结论】**本研究结果可为保险领域中的保单累积损失预测提供参考。

关键词：累积损失预测；半连续数据；Tweedie 回归模型；两部回归模型

中图分类号：F842；O212.1　　**文献标志码：**A　　**文章编号：**1671-8798(2023)06-0467-08

Prediction of insurance loss based on semicontinuous two-part model

LU Yahui[1], LIU Aiyi[2]

(1. School of Economics and Management, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China; 2. National Institutes of Health, Bethesda 20817, Maryland, USA)

Abstract: [**Objective**] It is imperative to improve the prediction accuracy of policy accumulated loss in the insurance field. The traditional Tweedie regression model can only establish a regression model for non-zero mean value, but not for zero probability, so the fitting effect of the model is not ideal. [**Method**] Considering that the policy loss data often contains a large number of zero claims, it can be regarded as semicontinuous data. Therefore, based on the semicontinuous two-part model and considering the distribution type of non-zero continuous part of accumulated loss, three different accumulated loss prediction models were proposed, and a set of actual loss data was combined to make a comparative analysis of the models. [**Result**] The results show that, compared with the Tweedie regression model the AIC (Akaike information criterion) and BIC (Bayesian information criterion) values of the two semicontinuous

收稿日期：2022-11-19

基金项目：杭州市哲学社会科学规划课题(Z23JC042)；国家自然科学基金项目(11971433)

通信作者：鲁亚会(1990—　)，女，河南省商丘人，讲师，博士，主要从事应用统计研究。E-mail：luyahui92@163.com。

regression models are smaller, which have better fitling effects. [**Conclusion**] The results can provide a reference for the prediction of policy accumulated loss in the insurance field.

Keywords: accumulated loss prediction; semicontinuous data; Tweedie regression model; two-part regression models

保单持有人和保险公司的风险防范意识不断增强,使得大部分保单并不会发生风险,或者保险公司与被保险人签订免赔偿或无赔偿折扣等条约,也使得在发生较轻的事故时被保险人不提出索赔[1]。这会导致一份保单或一个风险类别的累积损失数据具有下述2个特点:一是出现零过多现象,因为大部分保单在保险期间并未产生索赔,即在零点产生一个较大的概率堆积;二是非零部分可假设服从连续分布。此时,若直接采用传统的Tweedie回归模型[2-3]进行累积损失预测,虽然该方法不要求损失次数与损失金额满足相互独立,但其可能会在预测零概率值时产生较大偏差,即由Tweedie分布得到的零概率值远小于累积损失观察值的零概率值。虽然目前也有研究者对Tweedie回归模型进行了改进[4],但是也只能对均值引入协变量,而不能对零概率值引入协变量。可见,Tweedie回归模型在预测累积损失时具有一定的局限性。上述累积损失数据的2个特点其实质上就是一种半连续型数据,这是由于计量数据中包含过多零值时,除零以外的非零观测值往往是连续的,所以被称为半连续数据。对于半连续数据,Madden[5]指出此类型数据可以看作由混合分布产生,即可假设由零值数据(退化分布)和非零连续数据(连续分布)各占一定比例所构成的混合分布所产生[6],目前两部模型是最常用的拟合分析方法[7-8]。对于半连续数据的两部模型,其模型构建的基本思路是将数据看作由2个不同的随机过程产生。第一个过程考虑零值是否出现,即表示某种行为是否发生,此过程通常被称为数据的二元部分,此部分可假设服从伯努利分布[9];第二个过程考虑非零值的产生,此过程通常被称为数据的连续部分,此部分可假设服从一般的连续分布,如正态分布、伽马分布等[10]。为了进一步分析半连续数据中自变量对因变量的影响,需对二元部分参数和连续部分参数分别引入协变量,从而构造半连续两部回归模型[11-12]。因此,基于半连续两部模型,本研究将提出3种不同的累积损失预测模型。即将累积损失看作2个过程进行分别处理:一是损失是否发生,假设服从伯努利分布;二是在损失发生情况下累积损失金额的分布,分别考虑正态分布、伽马分布和逆高斯分布。对累积损失的2个过程分别引入相关的协变量进行解释,从而对累积损失预测建立相应的伯努利-正态(Bernoulli-Normal)回归模型、伯努利-伽马(Bernoulli-Gamma)回归模型和伯努利-逆高斯(Bernoulli-Inverse Gaussian)回归模型。

1 半连续两部回归累积损失模型

基于累积损失数据所具有的特点,其实质上就是一种半连续型数据。在半连续两部模型的框架下,下面将提出3种不同的半连续两部回归累积损失模型。

1.1 模型构建

在一个保险期间,假设$X=\{X_1,X_2,\cdots,X_n\}$为保单的累积损失金额,其中$X_i(i=1,2,\cdots,n)$表示第i份保单的累积损失金额,n为保单总份数[13]。此时,可将累积损失X看作2个过程分别进行处理:1)损失是否发生,假设服从伯努利分布;2)损失发生情况下,假设累积损失金额服从不同的分布。由此可对X构建半连续两部模型:

$$f(x_i)=(1-\pi)I(x_i=0)+[\pi g(x_i\mid x_i>0;\mu,\sigma,\kappa)]I(x_i>0),x_i\geqslant 0,i=1,2,\cdots,n。\tag{1}$$

式(1)中:$\pi=\Pr(X>0)$为非零概率值,且$0\leqslant\pi\leqslant 1$;$I(\cdot)$为示性函数;$g(X\mid X>0)$为$X>0$部分选定的连续分布函数;μ为位置参数;$\sigma>0$为尺度参数;$\kappa\in\mathbb{R}$为形状/偏度参数。

另外,在累积损失预测问题中,研究者往往更关注零概率值。因此,记$\nu=1-\pi$,并将其代入式(1)中。经过整理,则式(1)转换为

$$f(x_i)=\nu I(x_i=0)+[(1-\nu)g(x_i\mid x_i>0;\mu,\sigma,\kappa)]I(x_i>0),x_i\geqslant 0,i=1,2,\cdots,n。\tag{2}$$

式(2)中：$\nu=\Pr(X=0)$为零概率值。对非零累积损失数据的连续分布函数 $g(X\mid X>0)$，下面将分别采用正态分布(一般需进行对数转换)、伽马分布和逆高斯分布进行拟合分析。

1.2　伯努利-正态回归累积损失模型

在半连续两部模型(2)中，假设 $X>0$ 部分服从正态分布 $N(\mu,\sigma^2)$，且考虑到 $X>0$ 部分具有一定的偏态性，在实际应用中，一般需对 $X>0$ 进行对数转换。此时，$g(X\mid X>0)$分布的密度函数

$$g(x_i\mid x_i>0,\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(x_i-\mu)^2\right]。\tag{3}$$

将式(3)代入式(2)中，对累积损失 X 构建伯努利-正态两部模型，即构建由零点的退化分布和非零的正态分布组合的混合分布，其密度函数

$$f_{\mathrm{BN}}(x_i)=\nu I(x_i=0)+[(1-\nu)N(x_i;\mu,\sigma^2)]I(x_i>0)=$$
$$\begin{cases}\nu,x_i=0;\\(1-\nu)\dfrac{1}{\sqrt{2\pi}\sigma}\exp\left[-\dfrac{1}{2\sigma^2}(x_i-\mu)^2\right],x_i>0。\end{cases}\tag{4}$$

式(4)中：μ 为正态分布的均值，是位置参数；$\sigma>0$ 为正态分布的标准方差，是尺度参数。

为了进一步识别风险，在伯努利-正态两部模型式(4)中，分别对 ν 和 μ 引入相关的协变量，从而能够分析不同因素对 ν 和 μ 所产生的影响。另外，结合逻辑连接函数和对数连接函数，得到预测累积损失的伯努利-正态回归模型：

$$\begin{cases}X_i\sim f_{\mathrm{BN}}(x_i;\nu_i,\mu_i,\sigma);\\\mathrm{logit}(\nu_i)=\boldsymbol{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1;\\\ln(\mu_i)=\boldsymbol{z}_{2i}^{\mathrm{T}}\boldsymbol{\beta}_2。\end{cases}\tag{5}$$

式(5)中：$\boldsymbol{z}_{1i}=(z_{1i0},z_{1i1},\cdots,z_{1iq_1})^{\mathrm{T}}$ 为零概率 ν_i 的 q_1+1 维协变量向量；$\boldsymbol{\beta}_1=(\beta_{10},\beta_{11},\cdots,\beta_{1q_1})^{\mathrm{T}}$ 为所对应的 q_1+1 维回归系数向量。$\boldsymbol{z}_{2i}=(z_{2i0},z_{2i1},\cdots,z_{2iq_2})^{\mathrm{T}}$ 为均值参数 μ_i 的 q_2+1 维协变量向量；$\boldsymbol{\beta}_2=(\beta_{20},\beta_{21},\cdots,\beta_{2q_2})^{\mathrm{T}}$ 为其所对应的 q_2+1 维回归系数向量。设定 $z_{1i0}=z_{2i0}=1$，则 β_{10} 和 β_{20} 分别表示 2 个子回归部分的截距项。另外，在实际应用中，混合比例 ν_i 的协变量 $\boldsymbol{z}_{1i}$ 和均值参数 μ_i 的协变量 $\boldsymbol{z}_{2i}$ 可以相同，也可以不同。

1.3　伯努利-伽马回归累积损失模型

在半连续两部模型式(2)中，假设 $X>0$ 部分服从伽马分布 $G(\mu,\sigma^2)$。此时，$g(X\mid X>0)$分布的密度函数[14]

$$g(x_i\mid x_i>0,\mu,\sigma)=\frac{1}{(\sigma^2\mu)^{\frac{1}{\sigma^2}}}\times\frac{x_i^{\frac{1}{\sigma^2}-1}\exp\left(-\dfrac{x_i}{\sigma^2\mu}\right)}{\Gamma\left(\dfrac{1}{\sigma^2}\right)}。\tag{6}$$

将式(6)代入式(2)中，对累积损失 X 构建伯努利-伽马两部模型，即构建由零点的退化分布和非零的伽马分布组合的混合分布，其密度函数

$$f_{\mathrm{BG}}(x_i)=\nu I(x_i=0)+[(1-\nu)G(x_i;\mu,\sigma^2)]I(x_i>0)=$$
$$\begin{cases}\nu,x_i=0;\\(1-\nu)\dfrac{1}{(\sigma^2\mu)^{\frac{1}{\sigma^2}}}\times\dfrac{x_i^{\frac{1}{\sigma^2}-1}\exp\left(-\dfrac{x_i}{\sigma^2\mu}\right)}{\Gamma\left(\dfrac{1}{\sigma^2}\right)},x_i>0。\end{cases}\tag{7}$$

式(7)中：μ 为伽马分布的均值，是位置参数。

类似于伯努利-正态回归累积损失模型，在伯努利-伽马两部模型式(7)中，对 ν 和 μ 分别引入相关的协变量，并结合逻辑连接函数和对数连接函数，得到预测累积损失的伯努利-伽马回归模型：

$$\begin{cases} X_i \sim f_{\mathrm{BG}}(x_i;\nu_i,\mu_i,\sigma); \\ \mathrm{logit}(\nu_i)=\boldsymbol{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1; \\ \ln(\mu_i)=\boldsymbol{z}_{2i}^{\mathrm{T}}\boldsymbol{\beta}_2。 \end{cases}$$

1.4 伯努利-逆高斯回归累积损失模型

在半连续两部模型式(2)中,假设 $X>0$ 部分服从逆高斯分布 $N(\mu,\sigma^2)$。此时,$g(X\mid X>0)$分布的密度函数[15]

$$g(x_i\mid x_i>0,\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2x_i^3}}\exp\left[-\frac{(x_i-\mu)^2}{2\mu^2\sigma^2x_i}\right]。\tag{8}$$

将式(8)代入式(2)中,对累积损失 X 构建伯努利-逆高斯两部模型,即构建由零点的退化分布和非零的逆高斯分布组合的混合分布,其密度函数

$$f_{\mathrm{BIG}}(x_i)=\nu I(x_i=0)+[(1-\nu)IG(x_i;\mu,\sigma^2)]I(x_i>0)=$$
$$\begin{cases} \nu, x_i=0; \\ (1-\nu)\dfrac{1}{\sqrt{2\pi\sigma^2x_i^3}}\exp\left[-\dfrac{(x_i-\mu)^2}{2\mu^2\sigma^2x_i}\right], x_i>0。 \end{cases}\tag{9}$$

式(9)中:μ 为逆高斯分布的均值,是位置参数。

同样,在伯努利-逆高斯两部模型(9)中,对 ν 和 μ 分别引入相关的协变量,并结合逻辑连接函数和对数连接函数,即得到预测累积损失的伯努利-逆高斯回归模型:

$$\begin{cases} X_i \sim f_{\mathrm{BIG}}(x_i;\nu_i,\mu_i,\sigma); \\ \mathrm{logit}(\nu_i)=\boldsymbol{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1; \\ \ln(\mu_i)=\boldsymbol{z}_{2i}^{\mathrm{T}}\boldsymbol{\beta}_2。 \end{cases}$$

2 参数估计

目前,针对半连续两部回归模型的参数估计方法较多,而在实际应用中,具体的参数估计方法需根据调查目的及所选用的模型而定。极大似然法是一种最常用的参数估计方法,其基本算法就是高斯-牛顿(Gauss-Newton)迭代法[16-17]。由于伯努利-伽马和伯努利-逆高斯回归累积损失模型的参数估计过程类似于伯努利-正态回归模型,因此,下面只给出伯努利-正态回归模型的高斯-牛顿迭代估计过程。

基于伯努利-正态回归模型式(5),得到模型的似然函数

$$L(v_i,\mu_i\mid x)=\prod_{i=1}^{n}v_i^{1-y_i}\left\{\frac{1-v_i}{\sqrt{2\pi}\sigma}\exp\left[-\frac{1}{2\sigma^2}(x_i-\mu_i)^2\right]\right\}^{y_i}。\tag{10}$$

式(10)中:$y_i=I(x_i>0)$。

又由 $\mu_i=\boldsymbol{z}_{2i}^{\mathrm{T}}\boldsymbol{\beta}_2$,$v_i=\dfrac{\exp(\boldsymbol{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1)}{1+\exp(\boldsymbol{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1)}$,容易推导得到:

$$1-v_i=\frac{1}{1+\exp(\boldsymbol{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1)};\tag{11}$$

$$\ln v_i=\boldsymbol{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1-\ln[1+\exp(\boldsymbol{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1)];\tag{12}$$

$$\ln(1-v_i)=-\ln[1+\exp(\boldsymbol{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1)]。\tag{13}$$

将式(11)~(13)代入伯努利-正态似然函数(10)中,得到其对数似然函数

$$l(\boldsymbol{\theta}\mid x)=\sum_{i=1}^{n}\left\{(1-y_i)\ln v_i+y_i\left[\ln(1-v_i)-\frac{1}{2\sigma^2}(x_i-\mu_i)^2-\ln\sigma-\frac{1}{2}\ln(2\pi)\right]\right\}=l_1(\boldsymbol{\beta}_1)+l_2(\boldsymbol{\beta}_2,\sigma)。\tag{14}$$

式(14)中:

$$\boldsymbol{\theta}=(\boldsymbol{\beta}_1^{\mathrm{T}},\boldsymbol{\beta}_2^{\mathrm{T}},\sigma)^{\mathrm{T}};$$

$$l_1(\boldsymbol{\beta}_1)=\sum_{i=1}^{n}\{y_i\boldsymbol{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1-\ln[1+\exp(\boldsymbol{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1)]\};$$

$$l_2(\boldsymbol{\beta}_2,\sigma)=\sum_{i=1}^{n}\left\{I(y_i=1)\left[-\frac{1}{2\sigma^2}(x_i-\boldsymbol{z}_{2i}^{\mathrm{T}}\boldsymbol{\beta}_2)^2-\ln\sigma-\frac{1}{2}\ln(2\pi)\right]\right\}。$$

根据伯努利-正态对数似然函数式(14)，显然 $l(\boldsymbol{\theta}|x)$可分为 2 个独立的部分：1) 取值是否为零的二项部分，其对应的逻辑回归对数似然函数为 $l_1(\boldsymbol{\beta}_1)$，此部分可以对参数 $\boldsymbol{\beta}_1$ 进行估计；2) 非零连续值部分，其对应的正态分布对数似然函数为 $l_2(\boldsymbol{\beta}_2,\sigma)$，此部分可以对参数 $\boldsymbol{\beta}_2$ 和 σ 进行估计。此时，极大化对数似然函数 $l(\boldsymbol{\theta}|\boldsymbol{x})$等价于分别极大化 $l_1(\boldsymbol{\beta}_1)$和 $l_2(\boldsymbol{\beta}_2,\sigma)$，据此得到参数 $\boldsymbol{\theta}$ 的极大似然估计值$\hat{\boldsymbol{\theta}}=(\hat{\boldsymbol{\beta}}_1^{\mathrm{T}},\hat{\boldsymbol{\beta}}_2^{\mathrm{T}},\hat{\sigma})^{\mathrm{T}}$，即

$$\hat{\boldsymbol{\beta}}_1=\underset{\boldsymbol{\beta}_1}{\operatorname{argmax}}\,l_1(\boldsymbol{\beta}_1)；$$

$$(\hat{\boldsymbol{\beta}}_2,\hat{\sigma})=\underset{\boldsymbol{\beta}_2,\sigma}{\operatorname{argmax}}\,l_2(\boldsymbol{\beta}_2,\sigma)。$$

接下来采用高斯-牛顿迭代法分别对 $l_1(\boldsymbol{\beta}_1)$和 $l_2(\boldsymbol{\beta}_2,\sigma)$进行参数估计。

2.1　用高斯-牛顿迭代法估计对数似然函数 $l_1(\boldsymbol{\beta}_1)$

记参数 $\boldsymbol{\beta}_1$ 的得分函数

$$\boldsymbol{U}(\boldsymbol{\beta}_1)=\frac{\partial l_1(\boldsymbol{\beta}_1)}{\partial\boldsymbol{\beta}_1},$$

从而得到：

$$\boldsymbol{U}(\boldsymbol{\beta}_1)=\sum_{i=1}^{n}\left\{-\frac{1}{1+\exp(\boldsymbol{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1)}+y_i\right\}\boldsymbol{z}_{1i}。\tag{15}$$

记参数 $\boldsymbol{\beta}_1$ 的观测信息阵

$$\boldsymbol{I}(\boldsymbol{\beta}_1)=-\frac{\partial^2 l_1(\boldsymbol{\beta}_1)}{\partial\boldsymbol{\beta}_1\partial\boldsymbol{\beta}_1^{\mathrm{T}}},$$

从而得到：

$$\boldsymbol{I}(\boldsymbol{\beta}_1)=\sum_{i=1}^{n}\left\{-\frac{1}{[1+\exp(\boldsymbol{z}_{1i}^{\mathrm{T}}\boldsymbol{\beta}_1)]^2}\right\}\boldsymbol{z}_{1i}\boldsymbol{z}_{1i}^{\mathrm{T}}。\tag{16}$$

因此，基于式(15)和式(16)，根据迭代方程$\hat{\boldsymbol{\beta}}_1^{(t+1)}=\hat{\boldsymbol{\beta}}_1^{(t)}+\boldsymbol{I}^{-1}(\hat{\boldsymbol{\beta}}_1^{(t)})U(\hat{\boldsymbol{\beta}}_1^{(t)})$可以获得参数 $\boldsymbol{\beta}_1$ 的极大似然估计值$\hat{\boldsymbol{\beta}}_1$，其中$\hat{\boldsymbol{\beta}}_1^{(t)}(t=1,2,\cdots)$为第 t 步获得的参数迭代值。

2.2　用高斯-牛顿迭代法估计对数似然函数 $l_2(\boldsymbol{\beta}_2,\sigma)$

记参数 $\boldsymbol{\omega}=(\boldsymbol{\beta}_2^{T},\sigma)^{\mathrm{T}}$ 的得分函数 $\boldsymbol{U}(\boldsymbol{\omega})=(\boldsymbol{U}_{\boldsymbol{\beta}_2}^{\mathrm{T}},U_\sigma)^{\mathrm{T}}=\dfrac{\partial l_2(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}}$，从而得到：

$$\boldsymbol{U}_{\boldsymbol{\beta}_2}=\sum_{i=1}^{n}\{I(y_i=1)\frac{1}{\sigma^2}(x_i-\boldsymbol{z}_{2i}^{\mathrm{T}}\boldsymbol{\beta}_2)\boldsymbol{z}_{2i}\};\tag{17}$$

$$U_\sigma=\sum_{i=1}^{n}\{I(y_i=1)\frac{1}{\sigma^3}[(x_i-\boldsymbol{z}_{2i}^{\mathrm{T}}\boldsymbol{\beta}_2)^2-\sigma^2]\}。\tag{18}$$

又记参数 $\boldsymbol{\omega}=(\boldsymbol{\beta}_2^{\mathrm{T}},\sigma)^{\mathrm{T}}$ 的观测信息阵

$$\boldsymbol{I}(\boldsymbol{\omega})=-\frac{\partial^2 l_1(\boldsymbol{\omega})}{\partial\boldsymbol{\omega}\partial\boldsymbol{\omega}^{\mathrm{T}}},$$

通过计算得到：

$$\boldsymbol{I}(\boldsymbol{\omega})=\begin{bmatrix}\boldsymbol{I}_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2} & \boldsymbol{I}_{\boldsymbol{\beta}_2\sigma}\\ \boldsymbol{I}_{\boldsymbol{\beta}_2\sigma}^{\mathrm{T}} & \boldsymbol{I}_{\sigma\sigma}\end{bmatrix}。\tag{19}$$

式(19)中：

$$-\boldsymbol{I}_{\boldsymbol{\beta}_2\boldsymbol{\beta}_2}=\sum_{i=1}^{n}\{I(y_i=1)\frac{1}{\sigma^2}\boldsymbol{z}_{2i}\boldsymbol{z}_{2i}^{\mathrm{T}}\};$$

$$-\boldsymbol{I}_{\boldsymbol{\beta}_2\sigma}=\sum_{i=1}^{n}\{-I(y_i=1)\frac{1}{2\sigma^3}(x_i-\boldsymbol{z}_{2i}^{\mathrm{T}}\boldsymbol{\beta}_2)\}\boldsymbol{z}_{2i};$$

$$-\boldsymbol{I}_{\sigma\sigma}=\sum_{i=1}^{n}\left\{I(y_i=1)\left[-\frac{3}{\sigma^4}(x_i-\boldsymbol{z}_{2i}^{\mathrm{T}}\boldsymbol{\beta}_2)^2+\frac{1}{\sigma^2}\right]\right\}。$$

类似地，根据迭代方程$\hat{\boldsymbol{\omega}}^{(k+1)}=\hat{\boldsymbol{\omega}}^{(k)}+\boldsymbol{I}^{-1}(\hat{\boldsymbol{\omega}}^{(k)})\boldsymbol{U}(\hat{\boldsymbol{\omega}}^{(k)})$可得参数$\boldsymbol{\omega}=(\boldsymbol{\beta}_2^{\mathrm{T}},\sigma)^{\mathrm{T}}$的极大似然估计值[17]$\hat{\boldsymbol{\omega}}=(\hat{\boldsymbol{\beta}}_2^{\mathrm{T}},\hat{\sigma})^{\mathrm{T}}$，其中$\hat{\boldsymbol{\omega}}^{(k)}(k=1,2,\cdots)$表示第$k$步获得的参数迭代值。

3 实证研究

下面将本研究所提出的3种半连续两部回归模型和Tweedie回归模型，在一组机动车辆第三者责任险的累积损失数据[18]中进行拟合，以比较4种回归模型的拟合效果。

3.1 数据描述

原始数据集来源于R语言程序包“CASdatasets”，它是一组经典的保险精算数据集，共记录着429 350条损失信息。由于多次损失会发生在同一份保单中，通过累加同一份保单的多次损失，即能够得到累积损失数据集。此外，考虑到预测模型的稳健性，仅将累积损失小于15 000元的保单保留下来，由此共得到412 990份保单作为最终的累积损失数据来源。在这些数据中，共包含397 779份零累积损失保单，因此数据中含有大量的零值，也导致一个很大的零概率堆积。又考虑到数据的偏态性，对累积损失数据中非零值进行对数转换。此时，分别采用Tweedie模型、伯努利-正态两部模型、伯努利-伽马两部模型和伯努利-逆高斯两部模型对累积损失数据进行拟合，并使用AIC(Akaike information criterion，赤池信息量准则)来比较它们的拟合效果。4种模型的AIC值分别为180 652、174 964、177 893和180 483，结果表明：相较于传统的Tweedie模型，3种半连续两部模型具有较好的拟合效果，其中伯努利-正态两部模型又比其他2种两部模型的拟合效果更好。

3.2 模型选择

原始数据中包含着一些连续型和分类型解释变量，其中连续型变量有车龄、驾驶人车龄、人口密度，分类型变量有发动机功率、汽车品牌、汽车油耗类型。各分类解释变量的取值见表1。为了分析不同因素对累积损失产生的影响，对于3种半连续两部模型，将数据中所有解释变量分别引入零概率回归模型和均值回归模型，建立相应的伯努利-正态回归模型、伯努利-伽马回归模型和伯努利-逆高斯回归模型。但是对于Tweedie模型，只能将解释变量引入均值回归模型中，建立Tweedie回归模型。对于本研究所构建的4种回归模型，记l为对数似然函数值，p为模型的参数个数，n为样本容量。采用AIC和BIC(Bayesian information criterion，贝叶斯信息准则)进行模型比较和选择，其中AIC值$C_{\mathrm{AIC}}=-2l+2p$，BIC值$C_{\mathrm{BIC}}=-2l+p\ln l$，且AIC值和BIC值越小，表明模型具有越好的拟合效果。4种回归模型的拟合优度统计量见表2。

表1 分类解释变量的取值

Table 1 Values of classification explanatory variables

解释变量	符号	变量取值	代码
发动机功率	Power	d或e或f	power=“DEF”
		g或h	power=“GH”
		其他	power=“other”
汽车品牌	Brand	日本(尼桑除外)或韩国	brand=“F”
		其他	brand=“other”
汽车油耗类型	Gas	柴油	gas=“Diesel”
		普通油	gas=“Regular”

表2 4种回归模型的拟合优度统计量

Table 2 Goodness-of-fit statistics of four regression models

模型	$-2l$	AIC值	BIC值
Tweedie回归模型	179 954	179 976	180 097
伯努利-正态回归模型	174 019	174 057	174 264
伯努利-伽马回归模型	176 973	177 011	177 218
伯努利-逆高斯回归模型	179 480	179 518	179 725

由表2可知，3种半连续两部回归模型的AIC值和BIC值都小于Tweedie回归模型，表明半连续两部回归模型对累积损失的拟合效果优于Tweedie回归模型。该结果可能是由于Tweedie回归模型只能对均值建立回归模型，而无法对零概率建立回归模型；半连续两部回归模型能够同时对均值和零概率建立相应的回归模型。另外，在半连续两部回归模型中，伯努利-正态回归模型具有的AIC值和BIC值最小，表明伯努利-正态回归模型的拟合效果优于其他2种回归模型，该结果可能是由于所使用的损失数据

并不具有明显的尖峰厚尾特征。

3.3　结果分析

根据 4 种回归模型的 AIC 值和 BIC 值可知，伯努利-正态回归模型对本例的损失数据具有最优的拟合效果。因此，对于本例的累积损失数据，本节将最终建立伯努利-正态回归模型，其中对零概率建立逻辑回归模型，对均值建立对数回归模型，且将原始数据中的解释变量作为 2 个子回归模型中的协变量集。考虑到连续型变量对零概率和均值产生的影响并不一定是线性的，在伯努利-正态回归模型中，将车龄平方项和驾驶人车龄平方项作为 2 个子回归模型的协变量，采用高斯-牛顿迭代法进行参数估计。伯努利-正态回归模型的参数估计值见表 3。

表 3　伯努利-正态回归模型的参数估计值

Table 3　Parameter estimates for Bernoulli-Normal regression model

变量	零概率回归参数		均值回归参数	
	估计值	p 值	估计值	p 值
截距	2.544 0	0.001 2	7.105 0	0.000 0
发动机功率 EF	0.118 0	0.000 2	−0.108 9	0.678 9
发动机功率 GH	0.113 2	0.004 6	−0.006 8	0.811 3
汽车品牌 F	0.545 8	0.000 0	0.228 4	0.000 0
普通油耗	0.096 5	0.001 9	−0.044 5	0.011 2
人口密度	−0.000 1	0.000 0	0.000 0	0.469 3
车龄	−0.038 3	0.000 0	−0.002 9	0.296 1
车龄^2	0.002 4	0.000 0	0.000 0	0.476 3
驾驶人车龄	0.026 4	0.000 0	−0.013 8	0.000 0
驾驶人车龄^2	−0.000 3	0.000 0	0.000 2	0.000 0

由表 3 可知，对于零概率回归参数，在显著性水平为 5%的情况下，发动机功率、汽车品牌、油耗类型、人口密度、车龄和驾驶人车龄都对零概率具有显著影响，即这些变量都显著影响着损失发生的概率。其中人口密度的估计系数为负值，表明它与零概率存在着负相关，即人口密度值越大，损失发生的可能性也就越大。另外，车龄平方项和驾驶人车龄平方项对零概率也具有显著性影响，但它们的估计符号分别与车龄、驾驶人车龄变量相反，该现象表明车龄和驾驶人车龄对损失发生概率存在非线性影响。对于均值回归参数，在显著性水平为 5%的情况下，汽车品牌、油耗类型和驾驶人车龄都对均值具有显著影响，即这些变量都显著影响着累积损失的大小。油耗类型的估计系数为负值，表明它与均值存在着负相关，即汽车油耗类型为普通油时，会减少累积损失的金额。另外，驾驶人车龄平方项对均值也具有显著性影响，且该项的估计符号与驾驶人车龄变量相反，该现象也表明驾驶人车龄对累积损失金额存在非线性影响。

4　结　语

在保险研究中，累积损失预测是厘定纯保费的关键工作，目前最常用的模型就是 Tweedie 回归模型。但是该模型只能对非零均值建立回归模型，却不能对零概率建立回归模型，从而导致其对累积损失的拟合结果产生偏差。由于累积损失数据往往会出现零过多现象，本研究将该数据视作半连续数据构建模型，并考虑到数据中非零连续部分的分布类型，分别提出伯努利-正态两部模型，伯努利-伽马两部模型和伯努利-逆高斯两部模型。在这 3 种不同的模型中，对零概率参数和均值参数分别引入相关的协变量，从而建立预测累积损失的伯努利-正态回归模型，伯努利-伽马回归模型和伯努利-逆高斯回归模型。此外，本研究将所提出模型应用于一组机动车辆第三者责任保险的损失数据拟合中，并与传统的 Tweedie 回归模型进行对比。实证结果表明：相较于 Tweedie 回归模型，3 种半连续两部回归模型具有较好的拟合效果；其中伯努利-正态回归预测模型又优于其他 2 种模型。

本研究仍存在着一些待进一步探讨的问题。例如，随着信息技术的发展，获取累积损失数据中往往包含大量的候选解释变量，如何在保证模型的准确性和解释性的前提下，更好地选择出更重要的变量子集，这在模型构建中就会产生一个变量选择的问题。因此，半连续两部回归损失预测模型的变量选择将是我们后续研究的重点之一。

参考文献：

[1]　BIGNOZZI V, PUCCETTI G, RUSCHENDORF L. Reducing model risk via positive and negative dependence

assumptions[J]. Insurance Mathematics & Economics,2015,61(1):17.

[2] JORGENSEN B, PAES D S. Fitting Tweedie's compound Poisson model to insurance claims data[J]. Scandinavian Actuarial Journal,1994,1994(1):70.

[3] SMYTH G K. Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling[J]. Astin Bulletin,2002,32(1):145.

[4] 孟生旺,李政宵.基于随机效应零调整回归模型的保险损失预测[J].统计与信息论坛,2015,30(12):7.

[5] MADDEN D. Sample selection versus two-part models revisited: the case of female smoking and drinking[J]. Journal of Health Economics,2008,27(2):303.

[6] OLSEN M K, SCHAFER J L. A two-part random-effects model for semicontinuous longitudinal data[J]. Journal of the American Statistical Association,2001,96(454):732.

[7] SU L, TOM B D M, FAREWELL V T. Bias in 2-part mixed models for longitudinal semicontinuous data[J]. Biostatistics,2009,10(2):378.

[8] BOCK H M, BRENNER H. Inequalities in out of pocket payments for health care services among elderly Germans: results of a population-based cross-sectional study[J]. International Journal for Equity Health,2014,13(1):3.

[9] RUSTAND D, BRIOLLAISI L, TOURNIGAND C, et al. Two-part joint model for a longitudinal semicontinuous marker and a terminal event with application to metastatic colorectal cancer data[J]. Biostatistics,2022,23(1):50.

[10] 鲁亚会,刘爱义,江涛.多变量半连续数据的似然比检验[J].系统科学与数学,2021,41(11):3258.

[11] OLSEN M K, SCHAFER J L. A two-part random-effects model for semicontinuous longitudinal data[J]. Journal of the American Statistical Association,2001,96(454),730.

[12] JAFFA M A, GEBREGZIABHER M, JAFFA A A. Shared parameter and copula models for analysis of semicontinuous longitudinal data with nonrandom dropout and informative censoring[J]. Statistical Methods in Medical Research,2022,31(3):468.

[13] 段星德,张实,罗露璐,等.Tweedie复合泊松回归模型的Bayes估计和影响分析[J].高校应用数学学报A辑,2020,35(4):393.

[14] 李康,史宪铭,李广宁,等.基于正态-逆伽马分布的反巡航导弹命中概率估计方法[J].系统工程与电子技术,2022,44(8):2621.

[15] 赵远英,徐登可,冉庆.泊松逆高斯回归模型的贝叶斯统计推断[J].应用数学,2021,34(2):253.

[16] WANG T, CHENG J, GENG J. Wave equation reflection traveltime inversion using Gauss-Newton optimization [J]. IEEE Geoscience and Remote Sensing Letters,2022,19:19.

[17] 韦博成.参数统计教程[M].北京:高等教育出版社,2006.

[18] CHARPENTIER A. Computational actuarial science with R[M]. London: CRC Press,2014.