Statistics and Application 统计学与应用, 2024, 13(3), 864-871
Published Online June 2024 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2024.133088

Hans汉斯

零膨胀贝叶斯非参数模型对车险索赔频率的估计

董思彤

天津商业大学理学院，天津

收稿日期：2024年6月1日；录用日期：2024年6月21日；发布日期：2024年6月30日

摘　要

在非寿险精算领域，往往数据中会出现大量的零次索赔情况，这一零聚集现象称为零膨胀现象。在保险实务中，导致零膨胀现象的原因有多方面：比如某些保险产品在设计时就设定了较高的理赔门槛，导致很多小额理赔无法触发，从而产生大量的零值数据；或是在保险期限内被保险人没有出险因而没有产生索赔等。为了拟合数据中过多的零值，用零膨胀模型是一种很有效的方法。目前精算中领域中解决零膨胀问题大多使用的零膨胀模型，都用传统的参数估计方法进行参数估计，都局限在有限维的参数空间中。本研究使用贝叶斯非参数模型，它是一种定义在无限维参数空间上的贝叶斯模型，其大小可以随着模型内数据的增大或减小而自适应模型的变化。因此，将贝叶斯非参数方法引入零膨胀问题中，使得模型综合了贝叶斯方法和非参数方法的诸多优点，具有更大的灵活性。对解决保险精算领域中的问题具有重要的理论意义与实际应用价值。

关键词

贝叶斯非参数，零膨胀泊松分布，狄利克雷过程混合模型，车险索赔频率估计

Estimation of Auto Insurance Claim Frequency by a Zero-Expansion Bayesian Nonparametric Model

Sitong Dong

Faculty of Science, Tianjin University of Commerce, Tianjin

Received: Jun. 1st, 2024; accepted: Jun. 21st, 2024; published: Jun. 30th, 2024

Abstract

In the field of non-life actuarial science, there are often a large number of zero claims in the data,

文章引用: 董思彤. 零膨胀贝叶斯非参数模型对车险索赔频率的估计[J]. 统计学与应用, 2024, 13(3): 864-871.
DOI: 10.12677/sa.2024.133088

and this zero aggregation phenomenon is called zero inflation. In insurance practice, there are many reasons for the phenomenon of zero inflation: for example, some insurance products are designed with a high claim threshold, resulting in many small claims that cannot be triggered, resulting in a large amount of zero-value data, or the insured does not have an insurance during the insurance period and therefore does not generate a claim. In order to fit too many zeros in the data, a zero-inflation model is an effective method. At present, most of the zero-dilation models used to solve the zero-dilation problem in the actuarial field use traditional parameter estimation methods for parameter estimation, which are limited to the finite-dimensional parameter space. In this study, we use a Bayesian nonparametric model, which is a Bayesian model defined on an infinite-dimensional parametric space, the size of which can adapt to the change of the model as the data within the model increases or decreases. Therefore, the Bayesian nonparametric method is introduced into the zero-expansion problem, which makes the model combine many advantages of Bayesian method and non-parametric method, and has greater flexibility. It has important theoretical significance and practical application value for solving problems in the field of actuarial science.

Keywords

Bayesian Nonparametric, Zero-Inflated Poisson Distribution, Dirichlet Process Mixture Model, Estimated Frequency of Auto Insurance Claims

1. 引言

随着保险行业的快速发展和市场竞争的加剧，保险公司对于风险管理和定价的精确度的要求日益提高。然而，在实际操作中，保险公司经常面临数据中存在大量零值的情况，这给风险评估和保费定价带来了不小的挑战。零膨胀问题在保险精算中主要体现在以下几个方面：首先，在保险索赔数据中，由于免赔额、风险防范意识提高或保险条款的特定设计，许多保单在观察期内并未发生索赔，导致索赔次数数据中存在大量的零值。其次，在保险需求或保险购买行为的研究中，也可能出现零值过多的情况，例如某些客户在一段时间内未购买任何保险产品。这些零值的存在不仅影响了传统计数模型在保险精算中的适用性，还可能导致风险评估结果偏差，进而影响到保费的定价准确性和保险公司的盈利能力。因此，对于零膨胀问题的深入研究，对于提升保险精算的准确性和有效性具有重要意义。

索赔次数是衡量风险大小和保险费率的关键因素之一，它直接影响了保险公司的风险评估、定价策略以及经营决策。首先，通过对索赔次数进行准确的估计，保险公司可以更好地了解被保险人的风险状况，从而为其提供更个性化的保险服务。这包括设定合适的保费、制定针对性的保险条款以及提供及时的理赔服务等。其次，索赔次数的估计有助于保险公司进行风险管理和控制。通过对索赔数据的分析，保险公司可以识别出高风险群体或高风险行为，从而采取相应的措施来降低风险，比如调整保费、加强风险教育或是优化产品设计等。此外，索赔次数的估计还对保险公司的财务稳定性具有重要影响。通过对索赔次数的合理预测，保险公司可以更好地规划资金运用和风险管理，确保在面临大量索赔时能够保持充足的偿付能力。

在保险精算中，对索赔次数进行估计具有重要的意义。通过对零膨胀问题的深入研究，我们可以更准确地了解保险市场的真实情况，为保险公司的经营决策提供有力支持。其次，零膨胀问题对于保险精算的

准确性和可靠性有着重要影响。如果我们不能正确处理零值数据，那么精算结果可能会出现偏差，进而影响保险公司的风险评估和定价策略。因此，研究零膨胀问题有助于提高保险精算的准确性和科学性。

2. 零膨胀泊松模型(ZIP 模型)

1992 年，Lambert [1]提出了零膨胀泊松分布回归模型(zero-inflatedpoisson, ZIP)，它是一种混合分布模型，将零计数分布与泊松分布进行混合，同时考虑了协变量的影响。ZIP 模型的提出被应用到各个领域中。模型形式如下：

$$P_Y\left(Y=y;\phi,\lambda\right)=\begin{cases}\phi+\left(1-\phi\right)e^{-\lambda}, & y=0\\ \left(1-\phi\right)\dfrac{\lambda^y e^{-\lambda}}{y!}, & y=1,2,\cdots\end{cases} \tag{1}$$

其中，参数 ϕ 为结构零的比例，通常将其设为常数。当 $\phi=0$ 时，退化为泊松分布，当 $0<\phi<1$ 时，ϕ 越大，结构零的比例越大，零膨胀的现象越明显。

随机变量 Y 的均值和方差分别为：

$$\begin{aligned}E\left(Y\right)&=\left(1-\phi\right)E\left(X\right)\\ Var\left(Y\right)&=\left(1-\pi\right)\left\{Var\left(X\right)+\pi\left[E\left(X\right)\right]^2\right\}\end{aligned} \tag{2}$$

3. 零膨胀狄利克雷过程混合模型(BNP-ZIP 模型)

3.1. 生成模型

观测数据 $D=\left(D_i\right)_{i=1:n}=\left(Y_i,A_i,L_i\right)_{i=1:n}$ 来自 n 个独立的采样对象，L_i 为 $q\times1$ 的协变量向量，包含分类协变量和连续协变量。$A_i\in\{0,1\}$ 为逻辑变量。Y_i 为标量结果，其经验分布可以表现为多零、偏态或多模态。首先定义协变量向量 $x_i=\left(1,A_i,L_i\right)'$ 和 $m_i=\left(1,L_i\right)'$。我们指定了一个生成模型[2]，该模型的联合分布如下：

$$p\left(D_i\middle|\omega_i\right)=p\left(Y_i\middle|A_i,L_i,\omega_i\right)p\left(A_i\middle|L_i,\omega_i\right)p\left(L_i\middle|\omega_i\right) \tag{3}$$

也可以表示成如下的层次模型：

$$\begin{aligned}Y_i\middle|A_i,L_i,\beta_i,\gamma_i,\phi_i&\sim\pi\left(x_i'\gamma_i\right)\delta_0\left(y_i\right)+\left(1-\pi\left(x_i'\gamma_i\right)\right)\cdot N\left(y_i\middle|x_i'\beta_i,\phi_i\right)\\ A_i\middle|L_i,\eta_i&\sim Ber\left(\exp it\left(m_i'\eta_i\right)\right)\\ L_i\middle|\theta_i&\sim p\left(l_i\middle|\theta_i\right)\\ \omega_i\middle|G&\sim G\\ G\middle|\alpha,G_0&\sim DP\left(\alpha G_0\right)\end{aligned} \tag{4}$$

其中，令 $\omega_i=\left(\beta_i,\gamma_i,f_i,\eta_i,\theta_i\right)$，$Y_i$ 的条件分布为两部分的混合：在 0 点处的质量 $d_0\left(y_i\right)=I\left(y_i=0\right)$ 以及均值为 $x_i'b_i$ 方差为 f_i 的高斯分布。这允许结果为 0 的正概率 $P\left(Y_i=0\right)=\pi\left(x_i'\gamma_i\right)=\exp it\left(x_i'\gamma_i\right)$。

假设参数是从狄利克雷分布 G 中提取的。狄利克雷分布是“分布的分布”，由两个参数 α 和 G_0 确定，记为 $G_0\sim DP\left(\alpha G_0\right)$。参数 α 为分布参数，G_0 为基分布[3]。

3.2. 后验预测分布推导

令 $\tilde{y}$ 表示后验的预测结果，$\tilde{y}$ 的后验预测分布为：

$$p\left(\tilde{y}|D\right)=\int_{\omega_{1:n}}\iint_{\tilde{l}\ \tilde{\omega}}p\left(\tilde{y}|\tilde{l},\tilde{\omega},\omega_{1:n},D\right)p\left(\tilde{l}|\tilde{\omega},\omega_{1:n},D\right)p\left(\tilde{\omega}|\omega_{1:n}\right)p\left(\omega_{1:n}|D\right)d\tilde{\omega}d\tilde{l}d\omega_{1:n} \tag{5}$$

通常假设在新的参数的条件下，新的估计结果独立于之前的观测值和之前的参数，因此，

$$p\left(\tilde{y}|D\right)=\int_{\omega_{1:n}}\iint_{\tilde{l}\ \tilde{\omega}}p\left(\tilde{y}|\tilde{l},\tilde{\omega}\right)p\left(\tilde{l}|\tilde{\omega}\right)p\left(\tilde{\omega}|\omega_{1:n}\right)p\left(\omega_{1:n}|D\right)d\tilde{\omega}d\tilde{l}d\omega_{1:n} \tag{6}$$

假设可忽略性和一致性成立，

$$p\left(\tilde{y}|D\right)=\int_{\omega_{1:n}}\iint_{\tilde{l}\ \tilde{\omega}}p\left(\tilde{y}|A=a,\tilde{l},\tilde{\omega}\right)p\left(\tilde{l}|\tilde{\omega}\right)p\left(\tilde{\omega}|\omega_{1:n}\right)p\left(\omega_{1:n}|D\right)d\tilde{\omega}d\tilde{l}d\omega_{1:n} \tag{7}$$

根据 Polya Urn Blackwell and MacQueen，1973 的结论[4]：

$$\omega_i\left|\omega_{1:(i-1)}\right.\propto\frac{\alpha}{\alpha+i-1}G_0\left(\omega_i\right)+\frac{\alpha}{\alpha+i-1}\sum_{j=1}^{i=1}I\left(\omega_i=\omega_j\right)$$

代入 $i=n+1$ 可得

$$\omega_i\left|\omega_{1:n}\right.\propto\frac{\alpha}{\alpha+n}G_0\left(\tilde{\omega}\right)+\frac{\alpha}{\alpha+n}\sum_{j=1}^{n}I\left(\tilde{\omega}=\omega_j\right)$$

代入式(5)中可得

$$p\left(\tilde{y}|D\right)=\int_{\omega_{1:n}}\iint_{\tilde{l}\ \tilde{\omega}}p\left(\tilde{y}|A=a,\tilde{l},\tilde{\omega}\right)\left[\frac{\alpha}{\alpha+n}G_0\left(\tilde{\omega}\right)+\frac{1}{\alpha+n}\sum_{j=1}^{n}I\left(\tilde{\omega}=\omega_j\right)\right]p\left(\omega_{1:n}|D\right)d\tilde{\omega}d\tilde{l}d\omega_{1:n}$$

$$=\int_{\omega_{1:n}}\int_{\tilde{l}}\left[\begin{array}{l}\int_{\tilde{\omega}}p\left(\tilde{y}|A=a,\tilde{l},\tilde{\omega}\right)p\left(\tilde{l}|\tilde{\omega}\right)\frac{\alpha}{\alpha+n}G_0\left(\tilde{\omega}\right)d\tilde{\omega}+\\ \frac{1}{\alpha+n}\sum_{j=1}^{n}\int_{\tilde{\omega}}p\left(\tilde{y}|A=a,\tilde{l},\tilde{\omega}\right)p\left(\tilde{l}|\tilde{\omega}\right)I\left(\tilde{\omega}=\omega_j\right)d\tilde{\omega}\end{array}\right]p\left(\omega_{1:n}|D\right)d\tilde{l}d\omega_{1:n}$$

$$=\int_{\omega_{1:n}}\int_{\tilde{l}}\left[\begin{array}{l}\frac{\alpha}{\alpha+n}\int_{\tilde{\omega}}p\left(\tilde{y}|A=a,\tilde{l},\tilde{\omega}\right)p\left(\tilde{l}|\tilde{\omega}\right)G_0\left(\tilde{\omega}\right)d\tilde{\omega}+\\ \frac{1}{\alpha+n}\sum_{j=1}^{n}\int_{\tilde{\omega}}p\left(\tilde{y}|A=a,\tilde{l},\tilde{\omega}\right)p\left(\tilde{l}|\tilde{\omega}\right)\end{array}\right]p\left(\omega_{1:n}|D\right)d\tilde{l}d\omega_{1:n}$$

对上式的积分除以 y 可得出后验预测均值

$$E\left(\tilde{y}|D\right)=\int_{\omega_{1:n}}\int_{\tilde{l}}\left[\begin{array}{l}\frac{\alpha}{\alpha+n}\int_{\tilde{\omega}}E\left(\tilde{y}|A=a,\tilde{l},\tilde{\omega}\right)p\left(\tilde{l}|\tilde{\omega}\right)G_0\left(\tilde{\omega}\right)d\tilde{\omega}\\ +\frac{1}{\alpha+n}\sum_{j=1}^{n}E\left(\tilde{y}|A=a,\tilde{l},\tilde{\omega}\right)p\left(\tilde{l}|\omega_j\right)\end{array}\right]p\left(\omega_{1:n}|D\right)d\tilde{l}d\omega_{1:n}$$

可以通过蒙特卡洛算法计算上述积分[5]。

4. 模型的选择与比较

4.1. 布里尔分数

布里尔分数是一种计算预测值和真实值差异的指标，其计算公式为：

$$Brier_{(q)}=\frac{1}{m}\sum_{i=1}^{m}\left(f_i^{(q)}-y_i^{(q)}\right)^2 \tag{8}$$

其中，m 表示总共检测的样本数目，$f_{(i)}$ 表示模型预测的概率，$y_{(i)}$ 表示真实值。Brier Score 计算出来的

值在 0 到 1 之间，数值越小代表模型的准确率越高。

4.2. Bais 偏差和均方误差

Bais 偏差和 MSE 都是用来衡量模型预测值和实际值之间的差异，Bias 表示观测值和预测值的平均误差，MSE 为观测值和预测值之间的均方误差。其计算公式分别为：

$$\text{Bais} = \sum_{i=1}^{n} \frac{x_i - y_i}{n} \tag{9}$$

$$\text{MSE} = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{n} \tag{10}$$

4.3. Vuong 检验

Vuong 在 1989 年提出了非巢式模型和其检验统计量[6]，用于确定是否应该使用零膨胀模型，也可以用来比较两个非嵌套模型的拟合优度。令，

$$m_i = LN\left(\frac{f_1(y_i \mid X_i)}{f_2(y_i \mid X_i)}\right) \tag{9}$$

其中，$f_1(y_i|X_i)$为模型一的概率密度，$f_2(y_i|X_i)$为模型二的概率密度。取 m_i 的均值为，

$$\bar{m} = \left((1/n)\sum_{i=1}^{n} m_i\right) \tag{10}$$

Vuong 统计量的计算公式如下，

$$V = \frac{\sqrt{\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n} m_i\right]}}{\sqrt{\left[\left(\frac{1}{n}\right)\sum_{i=1}^{n}(m_i - \bar{m})^2\right]}} = \frac{\sqrt{n}(\bar{m})}{S_m} \tag{11}$$

其中 S_m 为标准差，n 为样本数量。V 服从极限正态分布，当 $V \geq 1.96$ 时，认为模型一优于模型二；当 $V \leq 1.96$ 时，认为模型二优于模型一；当 $|V| < 1.96$，无法判断两个模型哪个更优，则需要借助其他指标来进行评估，进而选择模型。

4.4. AIC BIC DIC 准则

AIC，BIC 和 DIC 准则是统计学常用的三种模型选择准则，用于给定一组模型的情况下，选择一个最优的模型。

AIC 准则由日本统计学家赤池弘次提出，又称赤池信息准则[7]。其计算公式为：

$$\text{AIC} = -2\log(L) + 2k \tag{12}$$

其中，L 为模型的拟合优度，k 为模型参数的个数。AIC 准则在评估模型拟合优度的同时还考虑了模型的复杂度，因此可以在一定程度上避免过拟合的情况。AIC 的值越小，则模型越优。

BIC 准则由纳维亚统计学家施瓦茨提出，又称贝叶斯信息准则[8]。其计算公式为：

$$\text{BIC} = -2\log(L) + k\log(n) \tag{13}$$

其中，L 为模型的拟合优度，k 为模型参数的个数，n 为样本量。BIC 相比于 AIC 对模型复杂度更加严格地惩罚，因此当样本量较大时，BIC 准则更倾向于选择更简单的模型。

DIC 准则由 Spiegelhalter 等人在 1998 年提出，又称偏差信息准则[9]。它基于对数后验概率的估计值，结合了模型拟合优度和复杂度。其计算公式为：

$$DIC = Dbar + pD \tag{12}$$

其中， Dbar 为后验均值下的拟合优度， pD 为模型有效参数的个数，DIC 准则考虑了贝叶斯估计的不确定性，通常用于贝叶斯模型选择。

5. 实例分析

5.1. 数据介绍与数据处理

本文研究的数据是机动车辆第三者责任险的保单索赔数据，数据中包含的解释变量有：驾驶员年龄、车龄、发动机年龄、汽车行驶时区域、汽车品牌、奖惩系数、油耗类型、保单持有人所在地区、居住人口密度。共纳入了 413169 份保单数据[10]。

Table 1. The value of the categorical explanatory variable
表 1. 分类解释变量的取值

解释变量	变量取值	代码
驾驶员年龄	17~22	Driver Age = “(17, 22]”
	23~26	Driver Age = “(23, 26]”
	27~42	Driver Age = “(27, 42]”
	43~74	Driver Age = “(43, 74]”
	74~99	Driver Age = “(74, 99]”
车龄	0~15	Car Age = “(0, 15]”
	16~100	Car Age = “(16, 100]”
居住人口密度	0~40	Density = “(0, 40]”
	41~200	Density = “(41, 200]”
	201~500	Density = “(201, 500]”
	501~4500	Density = “(501, 4500]”
汽车品牌	日本(尼桑除外)或韩国	Brand = “F”
	其他	Brand = “other”
汽车油耗类型	柴油	Gas = “Diesel”
	普通油	Gas = “regular”

5.2. 描述性统计

本研究共提取了 10000 个索赔数据，下面是对索赔次数的描述性统计。

Table 2. Descriptive statistical analysis table for number of claims
表 2. 索赔次数的描述性统计分析表

特征变量	变量取值	索赔次数
驾驶员年龄	17~22	316
	23~26	355
	27~42	1670
	43~74	2450
	74~99	162

续表

车龄	0~15	4643
	16~100	310
居住人口密度	0~40	671
	41~200	1496
	201~500	850
	501~4500	1646
	4501~27000	290

通过表 2 可以看出驾驶员年龄位于 43~74 岁最容易发生理赔；车龄小于 15 年的驾驶员理赔频率远远高于车龄大于 15 年的驾驶员；居住人口密度介于 501 到 4500 的被保险人最容易发生理赔。

5.3. 结果分析

Table 3. Comparison of goodness-of-fit indexes of models
表 3. 模型的拟合优度指标比较

模型	Bais	MSE	AIC	BIC
零膨胀泊松模型 (ZIP)	0.1527	0.4368	−309.24	−238.65
零膨胀狄利克雷过程混合模型 (BNP-ZIP 模型)	0.0962	0.2874	−732.46	−693.31

由表 3 结果可以看出零膨胀狄利克雷混合模型与传统的零膨胀泊松模型相比，其 Bais 值、MSE 值、AIC 值和 BIC 值均更小，由此可知，零膨胀狄利克雷混合模型具有更好的拟合与预测能力。

6. 小结

零膨胀问题在保险理赔中很常见，选择合适的模型进行预测具有很重要的现实意义。本文运用贝叶斯非参数模型进行预测，相比于传统的零膨胀模型，贝叶斯非参数模型具有更强的灵活性，不需要对参数的个数预先进行设定，其大小可以随着模型内数据的增大或减小而自适应模型的变化。本文运用零膨胀狄利克雷过程混合模型对零膨胀数据进行拟合和预测，狄利克雷过程混合模型是一种贝叶斯非参数模型。通过对比可知，零膨胀狄利克雷混合模型具有很好的拟合效果。

参考文献

[1] Lambert, D. (1992) Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, **34**, 1-14. https://doi.org/10.2307/1269547

[2] Barcella, W., De Iorio, M., Baio, G. and Malone-Lee, J. (2016) A Bayesian Nonparametric Model for White Blood Cells in Patients with Lower Urinary Tract Symptoms. *Electronic Journal of Statistics*, **10**, 3287-3309. https://doi.org/10.1214/16-ejs1177

[3] Ng, K.W., Tian, G.-L. and Tang, M.-L. (2011) Dirichlet and Related Distributions: Theory, Methods and Applications. John Wiley & Sons, 37-95, 97-98, 247.

[4] Blackwell, D. and MacQueen, J.B. (1973) Ferguson Distributions via Polya Urn Schemes. *The Annals of Statistics*, **1**, 353-355. https://doi.org/10.1214/aos/1176342372

[5] 张艳. 利用蒙特卡罗方法求解数值积分[J]. 高等数学研究, 2023, 26(1): 44-46, 61.

[6] Vuong, Q.H. (1989) Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. *Econometrica*, **57**, 307-333. https://doi.org/10.2307/1912557

[7] Akaike, H. (1974) A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, **19**, 716-723. https://doi.org/10.1109/tac.1974.1100705

[8] Schwarz, G. (1978) Estimating the Dimension of a Model. *The Annals of Statistics*, **6**, 461-464. https://doi.org/10.1214/aos/1176344136

[9] Wellwood, J., Sculpher, M.J., Stoker, D., Nicholls, G.J., Geddes, C., Whitehead, A., *et al*. (1998) Randomised Controlled Trial of Laparoscopic versus Open Mesh Repair for Inguinal Hernia: Outcome and Cost. *The BMJ*, **317**, 103-110. https://doi.org/10.1136/bmj.317.7151.103

[10] 徐昕, 袁卫, 孟生旺. 零膨胀广义泊松回归模型与保险费率厘定[J]. 数学的实践与认识, 2009, 39(24): 99-107.