Statistics and Application 统计学与应用, 2023, 12(2), 327-331
Published Online April 2023 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2023.122034

Hans汉斯

零膨胀泊松模型的变量选择及应用

马元凯

云南师范大学数学学院，云南 昆明

收稿日期：2023年3月13日；录用日期：2023年4月3日；发布日期：2023年4月18日

摘　要

零膨胀泊松模型是研究零过多的计数数据的方法。在实际应用中，为避免遗漏，常选择过多变量进行分析，因此需要对模型进行变量选择，本文在零膨胀泊松模型的基础上构造模型进行变量选择，本文通过研究影响德国卫生保健需求的因素，对比岭回归、Lasso、自适应Lasso和加权弹性网方法的效果。

关键词

零膨胀泊松模型，变量选择，EM算法

Variable Selection and Application of Zero Inflated Poisson Model

Yuankai Ma

School of Mathematics, Yunnan Normal University, Kunming Yunnan

Received: Mar. 13th, 2023; accepted: Apr. 3rd, 2023; published: Apr. 18th, 2023

Abstract

The zero inflated Poisson model is a method to study zero and excessive counting data. In practical application, in order to avoid omission, many variables are often selected for statistical analysis, so it is necessary to select variables for the model. This paper compares the effects of ridge regression, Lasso, adaptive Lasso and weighted elastic network by studying the factors affecting the demand for health care in Germany.

Keywords

Zero Inflated Poisson Model, Variable Selection, EM Algorithm

文章引用: 马元凯. 零膨胀泊松模型的变量选择及应用[J]. 统计学与应用, 2023, 12(2): 327-331.
DOI: 10.12677/sa.2023.122034

1. 引言

泊松分布是研究计数数据中最常用的模型，计数但是面对散度过大的数据时，拟合效果往往不够理想，而这类数据的特点是存在大量的零值，我们把这类数据称为零膨胀数据。对这类数据的研究最早可见于20世纪60年代中关于零膨胀现象的探索，1992年Lamber [1]提出了零膨胀泊松分布模型，并用于电子制造业的质量控制当中；近些年零膨胀模型得到充分的扩展，Ghosh [2]等研究了零膨胀模型的贝叶斯方法，Fahrmeir [3]等提出了一类零膨胀可加模型。当数据变量较多时就需要进行变量选择，本文基于惩罚似然的EM算法[4] [5]，用岭回归、Lasso、自适应Lasso和加权弹性网对德国卫生保健数据进行分析。

2. 模型假定和方法分析

2.1. 零膨胀泊松模型

零膨胀泊松分布模型的思想是取值为0的部分和取值为泊松分布的部分按一定比例构成的混合分布，由此我们可以得到零膨胀泊松模型的具体表达：

$$y_i \sim \begin{cases} 0, & \varphi_i \\ \text{Poisson}(\lambda), & 1-\varphi_i \end{cases} \tag{2.1}$$

其中φ_i表示0占的比例，$g(y)$来自泊松分布，且每一部分的概率都在$(0,\ 1)$之间。由(2.1)式可知观测值0由两部分构成，这里把来自额外的零的部分叫做结构零，来自离散分布的零的部分记为分布零，则$Y=y_i$的概率密度为：

$$P(Y=y_i \mid \varphi_i) = \begin{cases} \varphi_i + (1-\varphi_i)\mathrm{e}^{-\lambda_i}, & y_i = 0 \\ (1-\varphi_i)\dfrac{\lambda_i^{y_i}}{y_i!}\mathrm{e}^{-\lambda_i}, & y_i \geq 1 \end{cases} \tag{2.2}$$

φ_i为结构零的比例，当$\varphi_i=0$时模型(2.2)退化为泊松分布。为了考虑零膨胀计数数据中响应变量与协变量之间的关系，Lambert对零膨胀参数φ_i和泊松分布参数λ_i引入协变量，二者分别用logistic回归和对数线性回归模型建模，由此得到零膨胀泊松回归模型，连接函数如下所示：

$$\begin{cases} \log(\lambda_i) = x_i^{\mathrm{T}}\beta \\ \text{logit}(\varphi_i) = \log\left(\dfrac{\varphi_i}{1-\varphi_i}\right) = z_i^{\mathrm{T}}\gamma \end{cases} \tag{2.3}$$

其中$\beta=(\beta_1,\beta_2,\cdots,\beta_p)^{\mathrm{T}}$是协变量$x_i^{\mathrm{T}}$的回归系数，$\gamma=(\gamma_1,\gamma_2,\cdots,\gamma_q)^{\mathrm{T}}$是协变量$z_i^{\mathrm{T}}$的回归系数，且$x_i, z_i$分别是$p$维和$q$维向量。

2.2. ZIP模型似然函数

引入潜在数据ω_i，如果y_i来自额外零，记$\omega_i=1$，否则$\omega_i=0$，故ω_i有如下分布：

$$\omega_i = \begin{cases} 1, & \varphi_i \\ 0, & 1-\varphi_i \end{cases} \tag{2.4}$$

令 $Y=\left(Y_0,\omega_i\right)$，其中 $Y_0=\left(y_i,x_i,z_i\right)$ 为观测数据，则 Y 为完全数据集，取 $\phi=\left(\beta^{\mathrm{T}},\gamma^{\mathrm{T}}\right)^{\mathrm{T}}$，则基于完全数据得到的 ZIP 模型的似然函数为：

$$L\left(\phi\mid Y\right)=\varphi_i^{\sum_{i=1}^{n}\omega_i}\left(\left(1-\varphi_i\right)\mathrm{e}^{-\lambda_i}\right)^{\sum_{i=1}^{n}I_{(y_i=0)}-\sum_{k=1}^{n}\omega_i}\left[\left(1-\varphi_i\right)\frac{\lambda_i^{y_i}\mathrm{e}^{-\lambda_i}}{y_i!}\right]^{\sum_{i=1}^{n}I_{(y_i>0)}} \tag{2.5}$$

则对数似然函数为：

$$\begin{aligned}l\left(\phi\mid Y,\omega_i\right)&=\left(\sum_{i=1}^{n}\omega_i\right)\log\left(\varphi_i\right)+\left(\sum_{i=1}^{n}I_{(y_i=0)}-\sum_{i=1}^{n}\omega_i\right)\log\left(\left(1-\varphi_i\right)\mathrm{e}^{-\lambda_i}\right)+\left(\sum_{i=1}^{n}I_{(y_i>0)}\right)\log\left(\left(1-\varphi_i\right)\frac{\lambda_i^{y_i}\mathrm{e}^{-\lambda_i}}{y_i!}\right)\\&=\sum_{i=1}^{n}\left[\omega_i z^{\mathrm{T}}\gamma-\log\left(1+\mathrm{e}^{z^{\mathrm{T}}\gamma}\right)\right]+\sum_{i=1}^{n}\left(1-\omega_i\right)\left(y_i x^{\mathrm{T}}\beta-\mathrm{e}^{x_i^{\mathrm{T}}\beta}\right)-\sum_{i=1}^{n}\log\left(y_i!\right)\\&=l_{c,1}\left(\gamma\mid Y,\omega_i\right)+l_{c,2}\left(\beta\mid Y,\omega_i\right)\end{aligned} \tag{2.6}$$

3. 惩罚似然方法

假设线性回归模型为 $y=X\beta+\varepsilon$，其中 $y=\left(y_1,\cdots,y_n\right)^{\mathrm{T}},X=\left(x_1,\cdots,x_n\right)^{\mathrm{T}},\mathcal{A}=\left\{j,\hat{\beta}_j\neq 0\right\}$，参数 $\hat{\beta}$ 有如下表达：

$$\hat{\beta}=\arg\min_{\beta}\left\{\left\|y-X\beta\right\|_2^2+\lambda P\left(\beta\right)\right\} \tag{3.1}$$

其中上式 $P\left(\beta\right)$ 控制了模型的复杂性，在变量选择时 $P\left(\beta\right)$ 可以是岭惩罚、Lasso 惩罚、或弹性网惩罚，λ 为非负的调整参数。其中岭惩罚可以定义为 $P\left(\beta\right)=\sum_{j=1}^{p}\beta_j^2$；Lasso 惩罚[6]作为选择变量的正则化方法，可以定义为 $P\left(\beta\right)=\left\|\beta\right\|_1=\sum_{j=1}^{p}\left|\beta_j\right|$ 估计，即：

$$\hat{\beta}=\arg\min_{\beta}\left\|Y-X\beta\right\|_2^2+\lambda_n\sum_{j=1}^{p}\hat{\omega}_j\left|\beta_j\right| \tag{3.2}$$

其中 $\hat{\omega}_j$ 表示惩罚项中参数系数的权重。本节从弹性网[7]的思想出发，将惩罚项中的权重用于 L_1 惩罚和 L_2 惩罚中，利用 EM 算法对加权弹性网的对数似然函数获得最优解，记 $l\left(\phi;x_i,z_i\right)$ 作为对数似然函数，且 $\phi=\left(\beta,\gamma\right)$，则相应的零膨胀计数模型的惩罚似然函数为：

$$\hat{\phi}=\arg\min_{\theta}\left\{-l\left(\theta;x_i,z_i\right)+\lambda_1\sum_{j=1}^{p}\left(\alpha_1\hat{\omega}_{1j}\left|\beta_j\right|+\frac{1}{2}\left(1-\alpha_1\right)\hat{\omega}_{1j}\beta_j^2\right)+\lambda_2\sum_{j=1}^{p}\left(\alpha_2\hat{\omega}_{2j}\left|\gamma_j\right|+\frac{1}{2}\left(1-\alpha_2\right)\hat{\omega}_{2j}\gamma_j^2\right)\right\} \tag{3.3}$$

其中 $\hat{\omega}_{1j}=SE\left(\hat{\beta}_j\right)\Big/\hat{\beta}_{ridge,j},\hat{\omega}_{2j}=SE\left(\hat{\gamma}_j\right)\Big/\hat{\gamma}_{ridge,j}$，参数 $\hat{\beta}_{ridge,j},\hat{\gamma}_{ridge}$ 都由岭回归估计得到，调整参数 $\lambda_1,\lambda_2\geq 0$，$SE\left(\hat{\beta}\right),SE\left(\hat{\gamma}\right)$ 是相关系数的标准误。

3.1. EM 算法

令 $W_i=I\left(y_i=0\right)$，根据零膨胀泊松分布模型的条件期望有：

$$P\left(Y_i=y_i,W_i=\omega_i\mid x_i,z_i,\beta,\gamma\right)=\varphi_i^{\omega_i}\left(1-\varphi_i\right)^{1-\omega_i}\left(\frac{\lambda^{y_i}\mathrm{e}^{-\lambda_i}}{y_i!}\right)^{1-\omega_i}=\frac{\left(\mathrm{e}^{z_i^{\mathrm{T}}\gamma}\right)^{\omega_I}}{1+\mathrm{e}^{z_i^{\mathrm{T}}\gamma}}\left(\frac{\mathrm{e}^{y_i x_i^{\mathrm{T}}\beta-\mathrm{e}^{z_i^{\mathrm{T}}\gamma}}}{y_i!}\right)^{1-\omega_i} \tag{3.4}$$

基于 EM 算法，首先用无惩罚最大似然估计所得 $\left(\beta,\gamma\right)$ 作为初始化参数 $\hat{\phi}^0=\left(\beta^0,\gamma^0\right)$，首先进行 E 步，根据完全数据和更新得到参数估计值的条件期望更新 ω_i^t，假设参数值为 $\hat{\beta}^{(t)},\hat{\gamma}^{(t)}$，那么：

$$\hat{\omega}_i^t = E\left(\omega_i \mid Y_0, \hat{\phi}^{(t)}\right) = P\left(\omega_i \mid Y_0 = y_i, \hat{\phi}^{(t)}\right) = \begin{cases} \left[1+\exp\left\{-z_i^{\mathrm{T}}\gamma^{(t)}\right\} f\left(y_i;\hat{\phi}^{(t)}\right)\right]^{-1}, & y_i = 0 \\ 0, & y_i = 1,2,\cdots \end{cases} \tag{3.5}$$

其中 $f\left(y_i;\hat{\phi}^{(t)}\right)$ 为零膨胀泊松分布，即 $f\left(y_i;\hat{\phi}^{(t)}\right) = \mathrm{e}^{-\lambda_i} = \mathrm{e}^{-\mathrm{e}^{x_i^{\mathrm{T}}\beta^{(t)}}}$；

$$\begin{aligned} Q\left(\phi \mid \hat{\phi}^{(t)}\right) &= \sum_{i=1}^{n}\left[1-E\left(\omega_i \mid Y_0, \hat{\phi}^{(t)}\right)\right]\left(\log\left(f\left(y_i;\hat{\beta}^{(t)},\hat{\gamma}^{(t)}\right)\right)\right) + E\left(\omega_i \mid Y_0, \hat{\phi}^{(t)}\right) z_i^{\mathrm{T}}\gamma^{(t)} - \log\left(1+\mathrm{e}^{z_i^{\mathrm{T}}\gamma^{(t)}}\right) \\ &\quad + \lambda_1\sum_{j=1}^{p}\left(\alpha_1\hat{\omega}_{1j}\left|\beta_j^{(t)}\right| + \frac{1}{2}(1-\alpha_1)\hat{\omega}_{1j}\left(\beta_j^{(t)}\right)^2\right) + \lambda_2\sum_{j=1}^{p}\left(\alpha_2\hat{\omega}_{2j}\left|\gamma_j^{(t)}\right| + \frac{1}{2}(1-\alpha_2)\hat{\omega}_{2j}\left(\gamma_j^{(t)}\right)^2\right) \\ &= Q_1\left(\beta \mid \hat{\phi}^{(t)}\right) + Q_2\left(\gamma \mid \hat{\phi}^{(t)}\right) \end{aligned} \tag{3.6}$$

M 步：对于给定的 $\hat{\omega}_i^t$ 将 Q 函数最小化，即分别最小化 $Q_1\left(\beta \mid \hat{\phi}^{(t)}\right)$ 和 $Q_2\left(\gamma \mid \hat{\phi}^{(t)}\right)$，则：

$$\begin{aligned} \beta^{(t+1)} &= \arg\min_{\beta} Q_1\left(\beta \mid \hat{\phi}^{(t)}\right), \\ \gamma^{(t+1)} &= \arg\min_{\gamma} Q_2\left(\gamma \mid \hat{\phi}^{(t)}\right), \end{aligned} \tag{3.7}$$

其中 $\hat{\beta}^{(t)}, \hat{\gamma}^{(t)}$ 分别表示 t 步迭代时参数 β,γ 的估计值；按照上述 E 步和 M 步进行迭代最后收敛得到 $\hat{\beta}^{(n)}, \hat{\gamma}^{(n)}$。

3.2. 调整参数选择

本文根据最小 BIC 原则确定调整参数 λ_1,λ_2，BIC 准则定义如下：

$$\mathrm{BIC} = -2l\left(\hat{\theta},\lambda_1,\lambda_2\right) + \log(n)k \tag{3.8}$$

其中 $l\left(\hat{\theta},\lambda_1,\lambda_2\right)$ 为对数似然函数，n 为样本个数，$k = \sum_{j=1}^{p} I\left\{\beta_j \neq 0\right\} + \sum_{j=1}^{p} I\left\{\gamma_j \neq 0\right\}$ 为零膨胀泊松分布有效参数个数。

4. 实例分析

患者的医疗需求一直是医疗研究中的重要问题之一，本文选择德国卫生需求数据，该用于研究就诊次数与患者状况之间的关系，数据包括响应变量就诊次数数据和 13 个与患者状况相关数据，共有 1812 个观测值。本文用岭回归、Lasso、自适应 Lasso、和加权弹性网分别对影响就诊次数的相关因素进行变量选择，其中自适应 Lasso 的权重参数为 $1/\hat{\theta}$，加权弹性网的权重参数为 $se\left(\hat{\theta}\right)/\hat{\theta}$，$\hat{\theta}$ 是由岭回归估计得到的参数。

Table 1. Coefficient estimation of data model fitting
表 1. 数据模型拟合的系数估计

	岭回归		Lasso		Adaptive Lasso		加权弹性网	
	计数部分	零部分	计数部分	零部分	计数部分	零部分	计数部分	零部分
常数	2.1161	−2.2898	2.2095	−2.0280	2.1368	−2.4808	2.8120	−2.3068
age	0.0104	−0.0032	0.0058	−0.0062	0.0098	0	0.0067	0
health	−0.1639	0.2870	−0.1596	0.2546	−0.1633	0.2873	−0.1656	0.2569
handicap	0.3514	−0.2988	0.1000	−0.0717	0.2477	−0.3323	0.1123	−0.4211
hdegree	−0.0050	−0.0005	0	−0.0176	−0.0027	0	0	0

Continued

married	−0.1443	−0.2862	0	0	−0.1313	−0.3300	0	0
schooling	−0.0002	0	0	0	0	0	0	0
hhincome	0.0026	−0.0136	0	0	0	0	0	0
children	0.0748	0.4773	0	0.2228	0.0702	0.5053	0	0.2916
self	−0.1837	0.2019	−0.0840	0	−0.1873	0.1783	−0.1596	0
civil	−0.1220	0.1892	−0.0513	0	−0.1237	0.1826	−0.0922	0
bluec	0.0774	0	0	0	0.0370	0	0	0
employed	−0.1464	0.0658	−0.0719	0	−0.1297	−0.0599	−0.0528	0
public	0.1252	−0.0284	0.1055	0	0.1176	0	0.1120	0
addon	0.2550	0.0563	0.0530	0	0.2553	0	0.1590	0
BIC	9110.89		9076.56		9063.43		9048.96	

根据表 1 的结果可以看出加权弹性网相比于其他方法的 BIC 值更小，说明该方法比岭回归、Lasso 和自适应 Lasso 拟合效果更好；分析加权弹性网的参数估计可知，残疾程度、结婚与否、受教育年限、收入、是否蓝领与就诊次数无关；健康满意度、是否有孩子与就诊次数为零时呈正相关，残疾程度与就诊次数为零时呈负相从泊松部分拟合参数可知，增加就诊次数的影响因素有年龄、是否残疾、是否有公共健康保险和是否有附加保险，其中是否残疾和是否有保险对增加就诊次数的起主要影响，说明身体情况和参保对就诊次数有较强的影响；对就诊次数呈负相关的影响因素为健康满意度、是否结婚、是否自营和是否公务员，其中主要因素为健康满意度、是否自营、是否公务员和是否被雇佣，说明生活和工作越繁忙，就诊次数越少。

5. 结论

本文基于就诊次数数据特征，构建了零膨胀泊松模型，并分别用岭回归、Lasso、自适应 Lasso 和加权弹性网进行变量选择，得到影响就诊次数的主要变量。但是，目前本文中关注的仅是截面数据，在纵向数据等方面仍有较大的发展空间。

参考文献

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

[2] Ghosh, S.K., Mukhopadhyay, P. and Lu, J.C. (2006) Bayesian Analysis of Zero Inflated Regression Models. *Journal of Statistical Planning and Inference*, **134**, 1360-1375. https://doi.org/10.1016/j.jspi.2004.10.008

[3] Fahrmeir, L. and Echavarria, L.O. (2006) Structured Additive Regression for over Dispersed and Zero-Inflated Count Data. *Applied Stochastic Models in Business and Industry*, **22**, 351-369. https://doi.org/10.1002/asmb.631

[4] Tang, Y.L., Xiang, L.Y. and Zhu, Z.Y. (2014) Risk Factor Selection in Rate Making: EM Adaptive LASSO for Zero-Infated Poisson Regression Models. *Risk Analysis*, **34**, 1112-1127. https://doi.org/10.1111/risa.12162

[5] Mallick, H., Tiwari, H.K. (2016) EM Adaptive LASSO—A Multilocus Modeling Strategy for Detecting SNPs Associated with Zero-Inflated Count Phenotypes. *Frontiers in Genetics*, **7**, Article No. 32. https://doi.org/10.3389/fgene.2016.00032

[6] Tibshirani, R. (1996) Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society*: *Series B*, **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[7] Hong, D. and Zhang, F. (2010) Weighted Elastic Net Model for Mass Spectrometry Imaging Processing. *Mathematical Modelling of Natural Phenomena*, **5**, 115-133. https://doi.org/10.1051/mmnp/20105308