

零膨胀泊松回归模型及其在交通事故中的应用

陈 异,戴 琳,寇 鹏

(昆明理工大学 理学院,云南 昆明 650093)

摘 要:零点膨胀泊松回归模型是利用零处的概率质量退化分布和一个泊松回归模型进行混合所得到的,该模型是分析零点膨胀计数数据的有效工具。文中采用零膨胀泊松回归模型对某高速公路数据进行拟合并对该数据采用 score 检验统计量就是否存在零膨胀进行了检验,研究结果表明当车流量达到 1.8 辆万/日以上交通事故发生的频率会明显增加,有关部门应采取相应的措施,如限制收费站进口车辆或者采取分流等,特别是在极端天气出现时,更应及时控制好车流量并提醒驾驶员保持车距。

关键词:零膨胀;ZIP 模型;score 检验;交通事故

中图分类号:O212.4

文献标识码:A

文章编号:1673-629X(2013)10-0163-04

doi:10.3969/j.issn.1673-629X.2013.10.041

Zero-inflated Poisson Regression Model and Its Application in Traffic Accident

CHEN Yi, DAI Lin, KOU Peng

(College of Science, Kunming University of Science and Technology, Kunming 650093, China)

Abstract: Zero-inflated Poisson regression model uses the zero probability mass distribution and a Poisson regression model for mixing. The model is very effective to analyze and research the problem of excessive containing zero. In this paper, adopt the zero-inflated Poisson regression model to research and analyze the relationship about highway traffic flow and traffic accident frequency, and use score test statistic for test whether there is zero-inflation existed in this count data, obtaining the zero-inflated model is valid. Research results show that when the traffic flow reached 1.8 million/day or more traffic accident frequencies will increase obviously, the relevant departments should take the corresponding measures, such as restrictions on imported vehicle toll station or adopting billabong, especially in extreme weather occurs, it should prompt control car traffic and remind the driver to keep distance.

Key words: zero-inflation; ZIP model; score test; traffic accident

0 引言

计数数据广泛地存在于金融、保险、社会科学和生物医学等研究领域。目前,对计数数据的研究已成为统计学的一大热点问题。传统的,拟合计数数据的常用分布主要有泊松分布、二项分布、负二项分布、广义泊松分布等,但在实际问题中计数数据常常会呈现出零观测值过多的情形,若仍然采用上述几类模型进行分析将导致有偏的统计推断结果。故为更好地解决此类问题,零膨胀回归模型被提出并用来分析和研究数据中含零过多的问题^[1-3]。

随着近年来社会经济的发展以及出行方式的改变,交通问题日益成为政府急需解决的一大问题,高速

公路不但可以提高物流效率,同时也体现了一个国家或地区的综合实力。然而随着汽车保有量和里程的快速增长,高速公路交通事故也在逐年提高,严重威胁了人们的生命和财产安全^[4]。以往,在对交通事故的研究中人们通常采用普通的广义线性回归模型进行分析,如 2006 年阚伟生详细介绍了泊松分布、负二项分布、零堆积泊松分布、零堆积负二项分布在交通事故^[5]中的运用;2009 年钟连德提出了基于负二项分布回归的高速公路事故预测模型,并得出环境变量和交通流变量对事故的发生有较大影响^[6];2012 年陈敏等提出多元统计回归的方法^[7];2012 年崔立志提出灰色预测模型^[8]。事实上,早在 1997 年,Shankar、Milton 和

收稿日期:2013-01-10

修回日期:2013-04-16

网络出版时间:2013-07-24

基金项目:国家自然科学基金青年基金(11201200);国家自然科学基金天元基金(11126310);云南省基金(2010CZ059)

作者简介:陈 异(1986-),男,湖南株洲人,硕士生,研究方向为多元统计诊断;戴 琳,教授,硕士生导师,研究方向为多元统计分析。

网络出版地址: <http://www.cnki.net/kcms/detail/61.1450.TP.20130724.1012.060.html>

Manning 就指出传统的泊松和负二项模型并不能说明零过多的情形,需要对过多的零值观测进行有效处理^[9]。在实际应用中,传统的广义线性回归模型对于高速路段平日车流量相对较小的情形表现良好,而对于节假日出现井喷现象并偶尔伴随极端的天气的高速路段则表现不佳。由于高速路段事故发生频数呈现出明显的零点膨胀特征,故文中拟采用零膨胀泊松回归模型对某高速段交通事故进行拟合分析,并对该模型进行 score 检验^[10-11]。

1 零膨胀泊松回归模型

1.1 零膨胀模型

对于含零特别多的一类计数数据,一般采用零膨胀回归模型(zero inflated regression models)来拟合。零膨胀回归模型的基本原理是利用在零处具有概率质量的退化分布和一个普通的回归模型进行混合。即:

$$P(Y_i = y_i) = \begin{cases} \omega + (1 - \omega)f(0) & y_i = 0 \\ (1 - \omega)f(y_i) & y_i > 0 \end{cases} \quad (1)$$

式中, ω 称为膨胀系数,且 $0 \leq \omega < 1$; $f(y_i)$ 为生成分布。具体的,如果生成分布分别为泊松分布、二项分布、混合泊松分布、广义泊松分布、负二项分布时,则可以相应得到零膨胀泊松分布回归模型(ZIP)、零膨胀二项分布回归模型(ZIB)、零膨胀混合泊松分布回归模型(ZIMP)、零膨胀广义泊松分布回归模型(ZIGP)以及零膨胀负二项分布回归模型(ZINB)。

1.2 零膨胀泊松回归模型

当(1)式中的 $f(y_i)$ 为泊松分布时,则该零膨胀混合回归模型称为零膨胀泊松回归模型(zero-inflated Poisson regression model),简称为ZIP回归模型。其模型定义如下:

$$f(y_i, \lambda_i, \omega_i) = \begin{cases} \omega_i + (1 - \omega_i)e^{-\lambda_i}, y_i = 0 \\ \frac{(1 - \omega_i)e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, y_i = 1, 2, \dots, n \end{cases} \quad (2)$$

根据上述模型,可以得到其均值和方差如下:

$$E(Y_i) = (1 - \omega) \lambda_i$$

$$\text{Var}(Y_i) = (1 - \omega) \lambda_i (1 + \omega \lambda_i)$$

一般的情况下,由于所观测结果之间并不是相互独立的,故可以假定零膨胀泊松混合回归模型中的参数 λ_i 和 ω_i 分别可以由相应的协变量来决定。

1.3 参数估计

在模型(2)框架的基础之上,可以选择其连接函数为:

$$\begin{cases} \eta_i = \log(\lambda_i) = X_i \beta \\ \xi_i = \log\left(\frac{\omega_i}{1 - \omega_i}\right) = W_i \alpha \end{cases} \quad i = 1, 2, \dots, n \quad (3)$$

其中, X_i' 和 W_i' 均是协变量向量; β 和 α 分别为相应的 $p \times 1$ 和 $q \times 1$ 回归系数向量。

为了简单起见而又不失一般性,可以将连接函数 ξ_i 设为一固定值,即: $\xi_i = \log\left(\frac{\omega_i}{1 - \omega_i}\right) = W_i \alpha = \gamma_0$ 。

实际中常把 ω_i 设为一个固定值 ω 以简化模型,从而得到

$$f(y_i, \lambda_i, \omega) = \begin{cases} \omega + (1 - \omega)e^{-\lambda_i}, y_i = 0 \\ \frac{(1 - \omega)e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}, y_i = 1, 2, \dots, n \end{cases} \quad (4)$$

故可以得到该分布模型的对数似然如下:

$$\begin{aligned} l = \sum_{i=1}^n \{ I_{\zeta_{y_i}=0} \log[\omega + (1 - \omega)e^{-\lambda_i}] + \\ I_{\zeta_{y_i}>0} [\log(1 - \omega) + y_i \log \lambda_i - \lambda_i - \log y_i!] \} = \\ \sum_{i=1}^n \{ \log(1 - \omega) + y_i \log \lambda_i - \lambda_i - \log y_i! + \\ I_{\zeta_{y_i}=0} \left[\log\left(\frac{\omega}{1 - \omega}\right) + \lambda_i \right] \} \end{aligned} \quad (5)$$

在计算参数 $\hat{\beta}$ 的估计问题上,采用极大似然估计程序得到模型参数的极大似然估计^[12]。

2 零膨胀回归模型的 score 统计检验

2.1 score 统计检验

对于计数数据是否存在零膨胀,必须对其进行检验,此检验对于模型选择具有重要意义。为了得到有效的参数空间,在零膨胀模型检验中常常设 $\gamma = \frac{\omega}{1 - \omega}$,由于 $0 \leq \omega < 1$,因此,对原假设 $H_0: \omega = 0$ 与备择假设 $H_1: \omega \neq 0$ 的检验就等价于原假设 $H_0^*: \gamma = 0$ 和备择假设 $H_1^*: \gamma \neq 0$ 的检验。于是,模型的对数似然也可表示为:

$$l = \sum_{i=1}^n \{ -\log(1 + \gamma) + y_i \log \lambda_i - \lambda_i - \log y_i! + I_{\zeta_{y_i}>0} [\log(\gamma + e^{-\lambda_i}) + \lambda_i] \} \quad (6)$$

为了得到 score 检验统计量,需要分别对对数似然函数 l 关于 β 和 γ 求一阶和二阶导数并令 $\gamma = 0$:

$$\begin{aligned} \frac{\partial l}{\partial \beta} \Big|_{\gamma=0} &= \sum_{i=1}^n \left\{ y_i - \lambda_i + I_{(y_i=0)} \left[\frac{-\lambda_i e^{-\lambda_i}}{\gamma + e^{-\lambda_i}} + \lambda_i \right] \right\} X_i \Big|_{\gamma=0} \\ &= \sum_{i=1}^n \{ y_i - \lambda_i \} X_i = 0 \\ \frac{\partial l}{\partial \gamma} \Big|_{\gamma=0} &= \sum_{i=1}^n \left\{ -\frac{1}{1 + \gamma} + I_{(y_i=0)} \frac{1}{\gamma + e^{-\lambda_i}} \right\} \Big|_{\gamma=0} = \\ &\quad \sum_{i=1}^n \{ I_{\zeta_{y_i}=0} e^{\lambda_i} - 1 \} \\ \frac{\partial^2 l}{\partial \beta \partial \beta} \Big|_{\gamma=0} &= \sum_{i=1}^n \left\{ -\lambda_i + \right. \end{aligned}$$

$$I_{(y_i=0)} \left[\frac{\gamma \lambda_i^2 e^{-\lambda_i} - \gamma \lambda_i e^{-\lambda_i} - \lambda_i e^{-2\lambda_i}}{(\gamma + e^{-\lambda_i})^2} + \lambda_i \right] \Bigg|_{\gamma=0}$$
$$X_i X_i' = \sum_{i=1}^n \{ -\lambda_i \} X_i X_i'$$
$$\frac{\partial^2 l}{\partial \beta \partial \gamma} \Bigg|_{\gamma=0} = \sum_{i=1}^n \left\{ I_{(y_i=0)} \left[\frac{\lambda_i e^{-\lambda_i}}{(\gamma + e^{-\lambda_i})^2} + \lambda_i \right] \right\} \Bigg|_{\gamma=0} X_i =$$
$$\sum_{i=1}^n \{ I_{(y_i=0)} [\lambda_i e^{\lambda_i} + \lambda_i] \} X_i$$
$$\frac{\partial^2 l}{\partial \gamma^2} \Bigg|_{\gamma=0} = \sum_{i=1}^n \left\{ \frac{1}{(1 + \gamma)^2} - I_{(y_i=0)} \frac{1}{(\gamma + e^{-\lambda_i})^2} \right\} \Bigg|_{\gamma=0} =$$
$$\sum_{i=1}^n \{ 1 - I_{(y_i=0)} e^{2\lambda_i} \}$$

从而可以得到 Fisher 信息阵,因此,在原假设的 $H_0^*:\gamma=0$ 成立下的 score 检验统计量为:

$$S(\tilde{\beta}, \tilde{\gamma}) = S(\tilde{\beta}, 0) =$$
$$\left(\frac{\partial l}{\partial \beta} \right)' \left(\frac{\partial^2 l}{\partial \beta \partial \beta} \quad \frac{\partial^2 l}{\partial \beta \partial \gamma} \right)^{-1} \left(\frac{\partial l}{\partial \beta} \right) \Bigg|_{\gamma=0} =$$
$$\left(\frac{\partial l}{\partial \gamma} \right)' \left(\frac{\partial^2 l}{\partial \beta \partial \gamma} \quad \frac{\partial^2 l}{\partial \gamma^2} \right) \left(\frac{\partial l}{\partial \gamma} \right) \Bigg|_{\gamma=0}$$
$$\frac{\left(\sum_{i=1}^n (I_{(y_i=0)} \cdot e^{\lambda_i} - 1) \right)^2}{\sum_{i=1}^n (e^{\lambda_i} - 1) - n\bar{y}} \tag{7}$$

式中, $\tilde{\lambda}_i$ 的值可以由前面提到的估计值 $\tilde{\beta}$ 相应得到。

2.2 score 检验的抽样分布

如果在有限的样本情况下,该模型的 score 检验统计量的抽样分布可以运用模拟研究得到,考虑如下模型为:

$$\log(\lambda_i) = a + b * x_i, i = 1, 2, \cdots, n \tag{8}$$

假设协变量 x_i 定义为服从(0,1)上的均匀分布。对其进行 1 000 次重复试验计算可得到检验统计量 S ,在此基础上将 S 的经验统计量与相应的自由度为 1 的卡方分布的分位点进行对比。在上述模型中,分别考虑 $a=0.1, b=1$ 和 $a=1, b=0.5$ 两种情况并可以得到如下 Q-Q 图(见图 1 和图 2)。

由图 1 和图 2 可以看出 score 检验统计量 S 在原假设 H_0^* 下服从自由度为 1 的卡方分布。

3 零膨胀回归模型在高速公路事故分析中的应用

在高速公路的交通事故模型中,车流量和天气是影响交通事故的两大主要因素,掌握这两大因素与交通事故之间的关系能够有助于预测交通事故的发生情况,及时做好防范措施以保证人民的生命财产安全。文中针对高速公路上频发交通事故且伤亡人数也在急剧上升这一现状,有代表性选取了中国中部某省某高

速路段 2011 年 1 月 1 日至 2011 年 12 月 31 日一年内每天的事故次数数据(见表 1),同时为了简单起见且不失一般性只选取了每天的车流量来作为唯一的协变量。

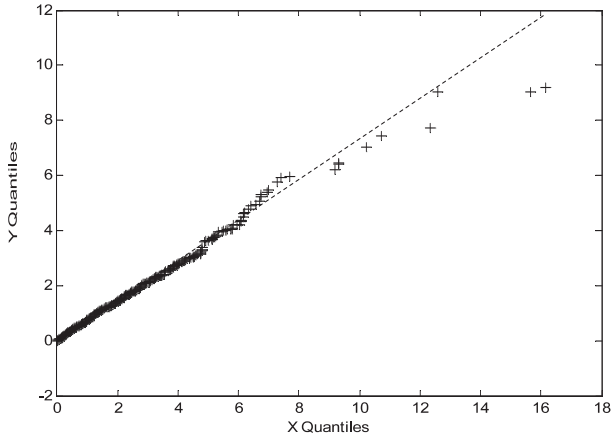


图 1 当 $a=0.1, b=1$ 时在 $H_0^*:\gamma=0$ 下的统计量 S 与 χ_1^2 分布的分位点的 Q-Q 图

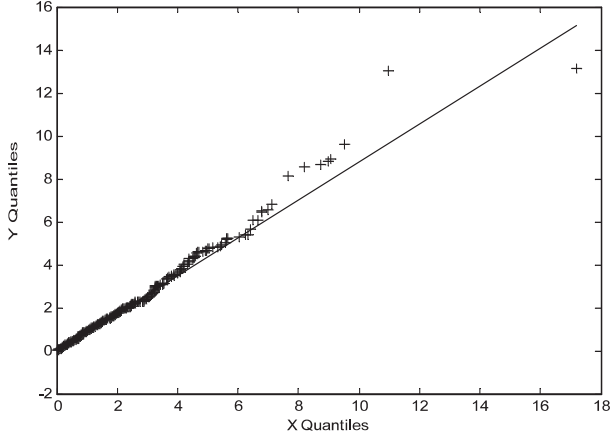


图 2 当 $a=1, b=0.5$ 时在 $H_0^*:\gamma=0$ 下的统计量 S 与 χ_1^2 分布的分位点的 Q-Q 图

表 1 该高速路段一年内各天发生次数及其频数

事故次数	0	1	2	3	4	5	6	7
频数	300	28	23	8	2	2	1	1

首先对事故次数做统计直方图,如图 3。

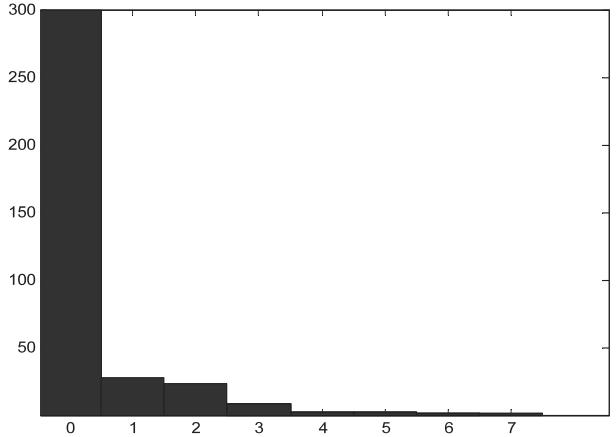


图 3 该高速路段一年内各天事故统计直方图

从该直方图可以看出,零出现的次数非常多,故有理由认为该计数数据出现了零膨胀现象。同时从图 3 可以看出图像类似于零膨胀的泊松分布统计直方图。在上述分析的基础上,文中对该实际数据考虑零膨胀泊松回归模型并对其进行分析,具体的,用极大似然估计算法得到相应的参数估计值 $a = -3.34, b = 1.84, \omega = 0.10$, 同时根据(7)式可得 score 检验统计量 $S = 448.67$, 对应 $\chi^2_{1,0.01} = 6.637$, 则拒绝原假设认为该数据存在明显的零点膨胀。因此,对上述实际数据考虑 ZIP 回归模型是合理的。

根据参数估计值可知当车流量控制在 1.8 万辆/日以内交通事故发生的概率几乎为零,当车流量达到 1.8 万辆/日以上交通事故发生的频率会明显增加,并会随着车流量的继续增加而导致交通事故的发生频数不断增大。所以当车流量即将超过 1.8 万辆/日时有关部门就应采取相应的措施,在极端天气出现时,更应及时控制好车流量并提醒驾驶员保持车距。

此外,文中还从贝叶斯方法的角度对该数据进行了分析,得到参数估计结果见表 2,该结果和前述极大似然估计结果是一致的。

表 2 各参数的估计值和相应的标准差以及 MC 误差

node	mean	sd	MC error
a	-3.375	0.219 3	0.007 368
b	1.838	0.117 4	0.003 94
ω	0.102	0.009 918	1.339E-4

4 结束语

文章针对实际生活中所研究的存在过多零的计数数据的普遍情况,先回顾了该类情况下常使用的零膨胀泊松混合分布模型,并提出 score 检验方法来检验是否存在零膨胀。在精确的参数估计问题上采用了极大似然估计和贝叶斯方法。实例说明零膨胀回归模型对高速公路交通事故的分析和预测有很好的可靠性和实

用性,是控制交通事故发生所采用的措施和决策的有力根据。另外,对于二维数据的分析,同样也可以将该方法推广到三维甚至多维的情形。

参考文献:

[1] Lambert D. Zero-inflated Poission regression with an application to defects in manufacturing[J]. Technometrics, 1992, 34(1):1-14.

[2] Ridout M, Demetrio C G B, Hinde J. Models for count data with many zeros[C]//Proc of the Nineteenth International Biometrics Conference. Cape Town: [s. n.], 1998:179-192.

[3] Bohning D. Zero-inflated Poisson models and C. A. Man;a tutorial collection of evidence[J]. Biometrical Journal, 1998, 40(7):833-843.

[4] 马昌喜. 高速公路交通安全对策研究[J]. 中国公共安全, 2008(3):168-170.

[5] 阙伟生. 路侧事故预测模型的统计分析方法研究[J]. 道路交通与安全, 2006(12):18-21.

[6] 钟连德, 孙小端, 陈永胜, 等. 高速公路事故预测模型[J]. 北京工业大学学报, 2009, 35(7):966-971.

[7] 陈 敏, 于静涛, 陆 建. 道路交通事故多元回归预测模型研究[J]. 公路交通科技(应用技术版), 2012(1):175-179.

[8] 崔立志. 高速公路交通事故的灰色预测模型[J]. 科学技术与工程, 2012, 12(19):4843-4846.

[9] Shankar V, Milton J, Mannering F. modeling accident frequencies as zeroaltered probability processes; an empirical inquiry [J]. Accident Analysis and Prevention, 1997, 29(6):829-837.

[10] van den B J. A score test for zero-inflation in a Poission distribution[J]. Biometrics, 1995, 51(2):738-743.

[11] Jansakul N, Hinde J P. Score test for zero-inflated Poission models[J]. Computational Statistics and Data Analysis, 2002, 40(1):75-96.

[12] 孙大飞, 陈志国, 刘文举. 基于 EM 算法的极大似然参数估计探讨[J]. 河南大学学报(自然科学版), 2002, 32(4):35-41.

(上接第 133 页)

[8] 张 茜, 朱艳琴, 罗喜召. OCSP 协议的改进和实现[J]. 计算机工程, 2007, 34(23):167-169.

[9] Micali S. Novomodo:scalable certificate validation and simplified PKI management[C]//Proc of 1st Annual PKI Research Workshop. [s. l.]:[s. n.], 2002.

[10] 王 政, 赵 明, 斯雪明, 等. 基于局部签名 Hash 表的证书撤销列表方案[J]. 计算机工程, 2009, 35(1):36-39.

[11] 李景峰, 潘 恒, 祝跃飞. 基于单向散列链的公钥证书撤销机制[J]. 小型微型计算机系统, 2006, 27(4):642-645.

[12] Wikipedia. Common access card[EB/OL]. 2013-08. http://en.wikipedia.org/wiki/Common_Access_Card.

零膨胀泊松回归模型及其在交通事故中的应用

作者：[陈异](#)，[戴琳](#)，[寇鹏](#)，[CHEN Yi](#)，[DAI Lin](#)，[KOU Peng](#)
作者单位：[昆明理工大学 理学院, 云南 昆明, 650093](#)
刊名：[计算机技术与发展](#)

英文刊名：[Computer Technology and Development](#)

年，卷(期)：2013(10)

本文链接：http://d.g.wanfangdata.com.cn/Periodical_wjfz201310041.aspx