零膨胀 Poisson 分布模型回归分析

胡良平[1,2*]

（1. 军事科学院研究生院，北京 100850；
2. 世界中医药学会联合会临床科研统计学专业委员会，北京 100029
* 通信作者：胡良平，E - mail：lphu812@sina.com）

【摘要】 本文目的是介绍零膨胀 Poisson 分布模型回归分析。首先，介绍零膨胀计数资料及其零膨胀 Poisson 分布回归模型构建原理，包括“零膨胀 Poisson 分布回归模型的形式”和“零膨胀 Poisson 分布回归模型的求解”；其次，介绍“零膨胀 Poisson 分布回归模型的 SAS 实现”，包括“创建 SAS 数据集”“呈现因变量 Y 的频数分布”“求出因变量 Y 的均值和方差”和“基于全部自变量对因变量 Y 构建多重零膨胀 Poisson 分布回归模型”。本文结果提示，当计数资料为非严重过离散的零膨胀计数资料时，拟合“多重零膨胀 Poisson 分布回归模型”，可获得满意的拟合效果。

【关键词】 零膨胀计数资料；过离散；零膨胀 Poisson 分布回归模型；概率函数；极大似然估计

中图分类号：R195.1 文献标识码：A doi：10.11886/j.issn.1007-3256.2018.05.004

The regression analysis of the zero – inflated Poisson distribution model

Hu Liangping[1,2*]

（1. *Graduate School，Academy of Military Sciences PLA China，Beijing* 100850，*China*；
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies，Beijing* 100029，*China*
* *Corresponding author：Hu Liangping，E – mail：lphu812@sina.com*）

【Abstract】 The purpose of this paper was to introduce the regression analysis of the zero – inflated Poisson distribution model. Firstly，the concepts of the zero – inflated count data and the building principle of the zero – inflated Poisson distribution regression model were given，which included the following two aspects：①the form of the zero – inflated Poisson distribution regression model；②the solution for the model mentioned before. Secondly，the SAS realization of this kind of model for the zero – inflated count data was presented. The contents were as follows：①creating SAS data set；②displaying the frequency of the count dependent variable Y；③calculating the arithmetic mean and variance of the count dependent variable Y；④building a multiple zero – inflated Poisson distribution regression model based on all independent variables. The results of the article showed that the satisfaction of the fitted effects could be gotten by building the zero – inflated Poisson distribution regression model to the zero – inflated count data with the not severe over – dispersion.

【Keywords】 Zero – inflated count data；Over – dispersion；Zero – inflated Poisson distribution regression model；Probability function；Maximum likelihood estimation

1 零膨胀计数资料及其零膨胀 Poisson 分布回归模型构建原理

1.1 零膨胀计数资料 Poisson 分布回归模型的形式

1.1.1 适于零膨胀计数资料 Poisson 分布回归模型的数据结构

适于零膨胀计数资料 Poisson 分布模型的数据结构见表 1。

【对数据结构的分析】因变量 count 为“计数变量”，它是“鱼的条数”。不难发现，没有钓到鱼的人数很多，或者说，count = 0 出现的频数很多，这种现象被称为“零膨胀”或“零堆积”。显然，钓鱼者的“性别”“年龄”可被视为两个自变量。

1.1.2 零膨胀计数资料 Poisson 分布回归模型的表达式

在零膨胀 Poisson（简称 ZIP）分布回归模型中，早期定义的数据产生过程被称为“过程 2”，即 Poisson分布概率函数，由下面的式（1）给出：

$$g(y_i) = \frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \tag{1}$$

在式（1）中，令 $\mu_i = e^{x_i'\beta}$。因此，ZIP 回归模型被定义为下面的式（2）[1]：

$$\left\{\begin{aligned} P(y_i = 0 \mid x_i, z_i) &= F_i + (1 - F_i)\exp(-\mu_i), y_i = 0 \\ P(y_i \mid x_i, z_i) &= (1 - F_i)\frac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}, y_i > 0 \end{aligned}\right\} \tag{2}$$

在式（2）的第 1 个式子中，F_i 和 $(1 - F_i)$ 分别为 y_i 取值为“0”与“非 0 正整数”的概率。若用“$e^{x_i'\beta}$ 取

项目基金：国家高技术研究发展计划课题资助（2015AA020102）

代式(2)中的“μ_i”,就可获得在给定“x_i,z_i”条件下,y_i 依赖自变量向量“x_i'”的“零膨胀 Poisson 分布回归模型”的完整公式。

表 1 研究者随机调查在某公园钓鱼的 52 人的性别、年龄和过去半年内钓鱼条数

sex	age	count	sex	age	count	sex	age	count	sex	age	count
F	54	18	M	37	0	F	48	12	M	27	0
M	55	0	M	32	0	F	49	12	F	45	11
M	39	0	F	34	1	F	50	0	M	52	4
M	33	0	M	32	0	F	23	1	F	17	0
F	44	5	M	44	0	F	26	0	F	30	0
F	38	0	F	38	0	F	52	18	M	23	1
F	23	0	M	32	0	F	33	3	M	26	0
F	46	8	M	45	5	M	51	10	F	48	5
F	31	2	F	25	1	M	22	0	M	41	0
M	19	0	M	23	0	M	31	1	M	17	0
F	21	0	F	44	7	M	28	0	M	47	3
M	23	0	F	29	3	F	24	0	M	34	1
F	19	0	F	35	2	M	39	0	M	43	6

注:sex,性别;age,年龄(岁);count,鱼的条数;F,女;M,男;此资料取自 SAS 9.3 的“帮助信息”,具体来说,取自 SAS/STAT 的“FMM 过程”中的“getting started”中的第 2 个例子

y_i 的条件期望(或均值)与条件方差分别见下面的式(3)与式(4):

$$E(y_i|x_i,z_i)=\mu_i(1-F_i) \tag{3}$$

$$V(y_i|x_i,z_i)=E(y_i|x_i,z_i)(1+\mu_iF_i) \tag{4}$$

值得注意的是,比较式(4)与式(3)可知,方差大于均值,故 ZIP 回归模型(也包括 ZINB 回归模型,即零膨胀负二项分布回归模型)刻划的是具有“零堆积”且具有“过离散”现象的计数因变量随自变量变化而变化的依赖关系。

1.2 零膨胀计数资料 Poisson 分布回归模型的求解

1.2.1 求解参数的思路

如何求解出式(2)中的概率“F_i”和回归系数“β_i”呢? 与大多数“广义线性模型”[1-2] 求解思路大同小异,其常规做法如下:第 1 步,基于“最大似然原理”并利用“概率函数(对离散型随机变量而言)或概率密度函数(对连续型随机变量而言)”构建“目标函数”,即所谓的“似然函数”;第 2 步,求取似然函数的自然对数 L;第 3 步,求取 L 关于参数的一阶和二阶偏导数;第 4 步,令二阶偏导数为“0”,得到方程组;第 5 步,采用某种迭代计算方法,求出方程组的解,即求得参数的估计值。

1.2.2 零膨胀 Poisson 分布回归模型的自然对数似然函数

概括地说,ZIP 回归模型的自然对数似然函数具有下面的形式,见式(5):

$$L=\sum_{i=1}^{N}w_i\ln[P(y_i\mid x_i,z_i)] \tag{5}$$

在一个被指定的连接函数(即 Logistic 分布函数或标准正态分布函数)用于概率 F_i 之后,就可以写出自然对数似然函数与其导数的精确表达式。

1.2.3 用 Logistic 分布函数作为连接函数的 ZIP 回归模型及参数估计

首先,考虑在 ZIP 回归模型中用 Logistic 分布函数作为连接函数来表达概率 F_i,这个函数就是下面的式(6):

$$F_i=\frac{\exp(z_i'\gamma)}{1+\exp(z_i'\gamma)} \tag{6}$$

此时,ZIP 回归模型的自然对数似然函数见下面的式(7):

$$L=\sum_{(i:y_i=0)}w_i\ln[\exp(z_i'\gamma)+\exp(-\exp(x_i'\beta))]$$
$$+\sum_{(i:y_i>0)}w_i[y_ix_i'\beta-\exp(x_i'\beta)-\sum_{k=2}^{y_i}\ln(k)]$$
$$-\sum_{i=1}^{N}w_i\ln[1+\exp(z_i'\gamma)] \tag{7}$$

式(7)中的“w_i”为“权重系数”，视具体情形而定，大体可分为以下 6 种情形：

(1)$w_i = 1$

当数据集中没有可作为“权重”变量或“频数”变量及其取值时，就令“$w_i = 1$”。

(2)$w_i = W_i$

当用户在“WEIGHT 语句”中指定了“非正态化”选项，而且，在“WEIGHT 语句”中指定了变量及其非正态化取值时，就令“$w_i = W_i$”。

(3)$w_i = \dfrac{n}{\sum_{i=1}^{n} w_i} W_i$

在“WEIGHT 语句”中指定了变量及其非正态化取值时，就令“$w_i = \dfrac{n}{\sum_{i=1}^{n} w_i} W_i$”。

(4)$w_i = F_i$

在“FREQ 语句”中，指定了变量及其取值 F_i 时，就令“$w_i = F_i$”。

(5)$w_i = W_i F_i$

当用户在“WEIGHT 语句”中没有使用“非正态化”选项，而且，同时使用了“WEIGHT 语句”和“FREQ 语句”，当然，W_i 和 F_i 分别为前面提及的两个语句中变量的取值，就令“$w_i = W_i F_i$”。

(6)$w_i = \dfrac{\sum_{i=1}^{n} F_i}{\sum_{i=1}^{n} F_i W_i} W_i F_i$

当用户同时使用了“WEIGHT 语句”和“FREQ 语句”，当然，W_i 和 F_i 分别为前面提及的两个语句中变量的取值，就令“$w_i = \dfrac{\sum_{i=1}^{n} F_i}{\sum_{i=1}^{n} F_i W_i} W_i F_i$”。

对于 ZIP 回归模型的偏导数分别见下面的式(8)和式(9)：

$$\frac{\partial L}{\partial \gamma} = \sum_{(i:y_i=0)} w_i \left[\frac{\exp(z_i'\gamma)}{\exp(z_i'\gamma) + \exp(-\exp(x_i'\beta))} \right] z_i - \sum_{i=1}^{N} w_i \left[\frac{\exp(z_i'\gamma)}{1 + \exp(z_i'\gamma)} \right] z_i \tag{8}$$

$$\frac{\partial L}{\partial \beta} = \sum_{(i:y_i=0)} w_i \left[\frac{-\exp(x_i'\beta)\exp(-\exp(x_i'\beta))}{\exp(z_i'\gamma) + \exp(-\exp(x_i'\beta))} \right] x_i + \sum_{(i:y_i>0)} w_i [y_i - \exp(x_i'\beta)] x_i \tag{9}$$

1.2.4 用标准正态分布函数作为连接函数的 ZIP 回归模型及参数估计

首先，考虑在 ZIP 回归模型中用标准正态分布函数作为连接函数来表达概率，这个函数就是下面的式(10)：

$$F_i = \Phi(z_i'\gamma) \tag{10}$$

此时，ZIP 回归模型的自然对数似然函数见下面的式(11)：

$$L = \sum_{(i:y_i=0)} w_i \ln\{\Phi(z_i'\gamma) + [-\Phi(z_i'\gamma)] \exp(-\exp(x_i'\beta))\} + \sum_{(i:y_i=0)} w_i \left\{ \ln [1 - \Phi(z_i'\gamma))] - \exp(x_i'\beta) + y_i x_i'\beta - \sum_{k=2}^{y_i} \ln(k) \right\} \tag{11}$$

式(11)中对“w$_i$ 的定义同式(7)，此处从略。

对 ZIP 回归模型的偏导数分别见下面的式(12)和式(13)：

$$\frac{\partial L}{\partial \gamma} = \sum_{(i:y_i=0)} w_i \frac{\varphi(z_i'\gamma)\ [1 - \exp(-\exp(x_i'\beta)]}{\Phi(z_i'\gamma + [1 - \Phi(z_i'\gamma)] \exp(-\exp(x_i'\beta))} z_i - \sum_{(i:y_i>0)} w_i \frac{\varphi(z_i'\gamma)}{[1 - \Phi(z_i'\gamma)]} z_i \tag{12}$$

$$\frac{\partial L}{\partial \gamma} = \sum_{(i:y_i=0)} w_i \frac{-[1 - \Phi(z_i'\gamma)] \exp(x_i'\beta)\exp(-\exp(x_i'\beta))}{\Phi(z_i'\gamma) + [1 - \Phi(z_i'\gamma)] \exp(-\exp(x_i'\beta))} x_i + \sum_{(i:y_i>0)} w_i [y_i - \exp(x_i'\beta)] x_i \tag{13}$$

【说明】以上计算公式来源于 SAS/ETS 模块中“COUNTREG 过程”的“details”部分。

1.2.5 二阶偏导数及迭代计算

通常情况下，需要在一阶偏导数的基础上，求取二阶偏导数；进而，令各参数的二阶偏导数及二阶混合偏导数为“0”，形成方程组。再利用某种迭代计算方法，从而求出各参数的估计值。因篇幅所限，详细做法此处从略。

2 零膨胀 Poisson 分布回归模型的 SAS 实现

2.1 创建 SAS 数据集

利用下面的 SAS 数据步程序，创建名为“catch”的临时 SAS 数据集：

```
data catch;
    input gender $ age count @@;
    if gender =´F´then sex =0;
    else if gender =´M´then sex =1;
datalines;
```

F	54	18	M	37	0	F	48	12	M	27	0
M	55	0	M	32	0	F	49	12	F	45	11

M	39	0	F	34	1	F	50	0	M	52	4
M	33	0	M	32	0	F	23	1	F	17	0
F	44	5	M	44	0	F	26	0	F	30	0
F	38	0	F	38	0	F	52	18	M	23	1
F	23	0	M	32	0	F	33	3	M	26	0
F	46	8	M	45	5	M	51	10	F	48	5
F	31	2	F	25	1	M	22	0	M	41	0
M	19	0	M	23	0	M	31	1	M	17	0
F	21	0	F	44	7	M	28	0	M	47	3
M	23	0	F	29	3	F	24	0	M	34	1
F	19	0	F	35	2	M	39	0	M	43	6

```
;
run;
```

2.2 编制出因变量 count 的频数分布表并绘制出其频数分布直条图

利用下面的一个 SAS 过程步程序，编制出因变量 count 的频数分布表，并绘制出其频数分布直条图（因为 count 是离散型随机变量，只取"0"和"非 0 正整数"值）：

```
proc freq data = catch;
    tables count/out = aaa plots = freqplot;
run;
```

【SAS 输出结果】

count	频数	百分比	累积频数	累积百分比
0	28	53.85	28	53.85
1	6	11.54	34	65.38
2	2	3.85	36	69.23
3	3	5.77	39	75.00
4	1	1.92	40	76.92
5	3	5.77	43	82.69
6	1	1.92	44	84.62
7	1	1.92	45	86.54
8	1	1.92	46	88.46
10	1	1.92	47	90.38
11	1	1.92	48	92.31
12	2	3.85	50	96.15
18	2	3.85	52	100.00

以上是计数因变量 count 的频数分布表，count = 0 出现的频数为 28，占总频数 52 的 53.85%。

以下是"计数因变量 count 的频数分布直条图"，见图 1。

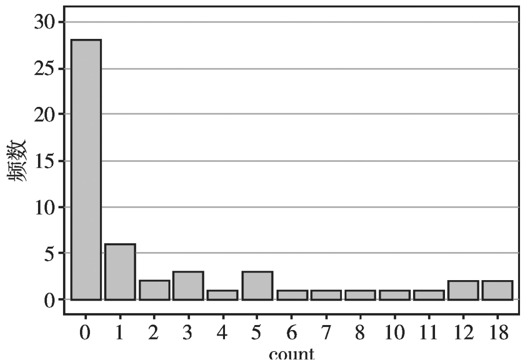

图 1 计数因变量 count 的频数分布直条图

图 1 直观、形象地呈现出计数因变量 count 取"0"的频数非常多，这种现象被称为"零膨胀"或"零堆积"。

2.3 求出因变量 count 的均值和方差

利用下面的两个 SAS 过程步程序，求出除"0"之外的其他计数资料的方差和均值：

```
proc univariate data = catch(where = (count ne 0))
noprint;
    var count;
    output out = aaa mean = mfish var = vfish;
run;
proc print data = aaa;
run;
```

【SAS 输出结果】

Obs	mfish	vfish
1	5.83333	27.0145

求得除"0"之外的其他计数因变量的均值为 5.83333、方差为 27.0145，方差约为均值的 4.63 倍。

Obs	mfish	vfish
1	2.69231	20.8054

以上是基于计数因变量 count 的全部 52 个观测值计算所得到的均值和方差，方差约为均值的 7.73 倍。

由此可知，此计数资料属于“过离散计数资料”。

2.4 基于全部自变量对因变量 count 构建多重零膨胀 Poisson 分布回归模型

2.4.1 利用下面的 SAS 过程步程序尝试性构建多重 Poisson 分布回归模型

```
/* 仅拟合多重 Poisson 分布回归模型* /
proc fmm data = catch;
    class gender;
    model count = gender*  age/dist = Poisson;
run;
```

【SAS 输出结果】

Fit Statistics

-2 Log Likelihood	182.7
AIC (smaller is better)	188.7
AICC (smaller is better)	189.2
BIC (smaller is better)	194.6
Pearson Statistic	85.9573

Parameter Estimates for 'Poisson' Model

效应	gender	估计值	标准误差	z 值	Pr > |z|
Intercept		-3.9811	0.5439	-7.32	<0.0001
age* gender	F	0.1278	0.01149	11.12	<0.0001
age* gender	M	0.1044	0.01224	8.53	<0.0001

关于模型对资料的拟合效果，暂不评价，需要与其他同类模型的拟合结果进行比较，才能得出有说服力的判定。

2.4.2 利用下面的 SAS 过程步程序尝试性构建多重负二项分布回归模型

```
/* 仅拟合多重负二项分布回归模型* /
proc fmm data = catch;
    class gender;
    model count = gender*  age/dist = nb;
run;
```

【SAS 输出结果】

Fit Statistics

-2 Log Likelihood	162.1
AIC (smaller is better)	170.1
AICC (smaller is better)	171.0
BIC (smaller is better)	177.9
Pearson Statistic	38.3597

Parameter Estimates for 'Negative Binomial' Model

效应	gender	估计值	标准误差	z 值	Pr > |z|
Intercept		-4.3816	0.8160	-5.37	<0.0001
age* gender	F	0.1378	0.01957	7.04	<0.0001
age* gender	M	0.1130	0.01971	5.73	<0.0001
Scale Parameter		0.6688	0.3263		

比较以上两个统计模型给出的“拟合统计量”部分结果可知，“多重负二项分布回归模型”好于“多重 Poisson 分布回归模型”，因为 AIC、AICC、BIC 和“Pearson Statistic（皮尔逊卡方统计量）”数值均有所下降。

2.4.3 利用下面的 SAS 过程步程序构建多重零膨胀 Poisson 分布回归模型

（1）按性别分层且基于常数与 Poisson 分布分别拟合零膨胀与非零膨胀两部分的零膨胀 Poisson 分布回归模型

```
proc fmm data = catch;
    class gender;
    model count  =  gender*  age  /  dist = Poisson;
    model        +             /  dist = Constant;
run;
```

【SAS 主要输出结果】

Fit Statistics

-2 Log Likelihood	145.6

AIC (smaller is better)	153.6	Pearson Statistic	43.4467
AICC (smaller is better)	154.5	Effective Parameters	4
BIC (smaller is better)	161.4	Effective Components	2

Parameter Estimates for 'Poisson' Model

成分	效应	gender	估计值	标准误差	z 值	$Pr > \lvert z \rvert$
1	Intercept		−3.5215	0.6448	−5.46	<0.0001
1	age* gender	F	0.1216	0.01344	9.04	<0.0001
1	age* gender	M	0.1056	0.01394	7.58	<0.0001

Parameter Estimates for Mixing Probabilities

效应	链接尺度				概率
	估计值	标准误差	z 值	$Pr > \lvert z \rvert$	
Intercept	0.8342	0.4768	1.75	0.0802	0.6972

关于模型对资料的拟合效果，暂不评价，需要与其他同类模型的拟合结果进行比较，才能得出有理由的判定。若仅依据"AIC、AICC、BIC"的数值来判断，要好于前面的"多重负二项分布回归模型"的拟合效果。

（2）不按性别分层且基于常数与 Poisson 分布分别拟合零膨胀与非零膨胀两部分的零膨胀 Poisson 分布回归模型

```
proc fmm data = catch;
    class gender;
    model count = gender age / dist = Poisson;
    model + / dist = Constant;
run;
```

【SAS 主要输出结果】

Fit Statistics

−2 Log Likelihood	144.4
AIC (smaller is better)	152.4
AICC (smaller is better)	153.3
BIC (smaller is better)	160.2
Pearson Statistic	43.7210
Effective Parameters	4
Effective Components	2

Parameter Estimates for 'Poisson' Model

成分	效应	gender	估计值	标准误差	z 值	$Pr > \lvert z \rvert$
1	Intercept		−4.1261	0.6624	−6.23	<0.0001
1	gender	F	0.7922	0.2193	3.61	0.0003
1	gender	M	0			
1	age		0.1178	0.0133	8.86	<0.0001

Parameter Estimates for Mixing Probabilities

效应	链接尺度				概率
	估计值	标准误差	z 值	$Pr > \lvert z \rvert$	
Intercept	0.9190	0.4934	1.86	0.0625	0.7148

若仅依据"AIC、AICC、BIC"的数值来判断，要略好于前面"按性别分层"的拟合效果，但二者无差异。

2.5 最终的回归模型与专业结论

2.5.1 最终的回归模型

依据上面最后一次选定的回归模型拟合资料所得到的计算结果写出"多重零膨胀 Poisson 分布回归模型"，此模型由两个表达式组成，一个对应于"$y_i = 0$"；另一个对应于"$y_i > 0$"。模型表达式见前面的式（2），其中，$F_i = 0.7148$、$\mu_i = e^{-4.1261 + 0.7922(sex = F) + 0.1178age}$。将这两个参数的估计结果代入式（2），就可获得与表 1 资料对应的"零膨胀 Poisson 分布回归模型"。

2.5.2 专业结论

任何一人钓不到鱼的概率约为 71.5%（因上面第一式"加号"后面与钓鱼者性别、年龄有关，但其数值并不大）；女性较男性、年龄长者较年龄小者钓鱼数目多。

参考文献

[1] SAS Institute Inc. STAT SAS 9.3 User's Guide [M]. Cary, NC: SAS Institute Inc, 2011: 2437 − 2548, 2605 − 2804.

[2] 胡良平. 有限混合模型回归分析 [J]. 四川精神卫生, 2018, 31(4): 307 − 312.

（收稿日期：2018 − 10 − 10）

（本文编辑：唐雪莉）