一般计数资料 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】 本文目的是介绍一般计数资料 Poisson 分布模型回归分析。首先,介绍一般计数资料及其 Poisson 分布模型构建原理,包括“一般计数资料 Poisson 分布回归模型的形式”和“一般计数资料 Poisson 分布回归模型的求解”;其次,介绍“一般计数资料 Poisson 分布回归模型的 SAS 实现”,包括“创建 SAS 数据集”“求出因变量 Y 的均值和方差”“检验因变量是否存在过离散现象”“对过离散进行校正”和“基于全部自变量对因变量 Y 构建多重 Poisson 分布回归模型”。本文结果提示,在“过离散”不十分严重的情况下,通过在 GENMOD 过程的“model 语句”中增加选项“dist = poisson”和“scale = deviance”,可以较好地校正“过离散”导致的不良后果。

【关键词】 计数资料;Poisson 分布;负二项分布;过离散;拉格朗日乘子
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.05.002

The regression analysis of the Poisson distribution model for the general count data

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 Poisson distribution model for the general count data. Firstly, the concepts of the general count data and the building principle of the Poisson distribution regression model were introduced, which included the following two aspects: ①the form of the Poisson distribution regression model of count data; ②the solution for the model mentioned before. Secondly, the SAS realization of the Poisson distribution regression model of count data was presented. The contents were as follows: ①creating SAS data set; ②calculating the arithmetic mean and variance of the dependent variable Y; ③checking whether there was the overdispersion in the dependent variable Y; ④adjusting the overdispersion; ⑤building a multiple Poisson distribution regression model for the dependent variable Y based on all independent variables. The results of the article showed that, under the situation of not severe overdispersion, the harmful results came from the overdispersion could be adjusted preferably through the following measures, such as using options of "dist = poisson" and "scale = deviance" in the model statement in the GENMOD procedure in SAS software.

【**Keywords**】 Count data; Poisson distribution; Negative binomial distribution; Overdispersion; Lagrange multiplier

1 一般计数资料及其 Poisson 分布模型构建原理

1.1 一般计数资料 Poisson 分布回归模型的形式

1.1.1 适于一般计数资料 Poisson 分布回归模型的数据结构

适于一般计数资料 Poisson 分布模型的数据结构见表 1[1]。

【对数据结构的分析】因变量 Y 为“计数变量”,其均值为 7.467、方差为 13.016,方差稍大于均值,只能将 Y 近似视为服从 Poisson 分布的离散型随机变量(理论上,均值应等于方差);拟考察的三个自变量均是“二值变量”,在回归分析中是可以接受的,但在统计学理论中,默认自变量为“计量变量”。

1.1.2 Poisson 分布概率函数[2]

设 Y 是一个服从 Poisson 分布的随机变量,$X=(1,x_1,x_2,\cdots,x_q)'$是一个协变量向量,$\beta=(\beta_0,\beta_1,\beta_2,\cdots,\beta_q)'$是参数向量。如果 Y 的数学期望的对数可以表示为协变量的线性表达式,即 $E(Y|X)=\exp(X'\beta)$ 或 $\ln[E(Y|X)]=X'\beta$,则称(Y,X)服从 Poisson 分布回归模型。对应的表达式见下面的式(1):

$$P(Y=k|X)=\frac{\lambda^k(X)e^{-\lambda(X)}}{k!},k=0,1,2,3,\cdots \quad (1)$$

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 1 三个可能的影响因素(X_1 – X_3)对 30 例非气质性心脏病且仅有胸闷症状就诊者 24 小时早搏数 Y 的观察结果

编号	X_1	X_2	X_3	Y	编号	X_1	X_2	X_3	Y
1	0	1	1	11	16	1	0	0	11
2	0	0	0	7	17	0	1	1	8
3	0	0	0	3	18	1	0	1	9
4	1	0	1	5	19	0	0	0	8
5	0	0	0	2	20	1	0	0	5
6	1	1	1	13	21	0	1	1	5
7	0	1	0	6	22	1	1	0	8
8	1	0	1	10	23	1	1	0	13
9	0	0	0	4	24	0	0	1	8
10	1	0	1	7	25	1	0	0	6
11	0	0	0	1	26	0	0	1	4
12	0	0	1	9	27	0	0	0	6
13	0	0	1	6	28	1	1	1	13
14	1	1	1	17	29	1	1	0	9
15	0	0	0	5	30	0	0	1	5

注:Y,24 小时中早搏数;X_1,是否吸烟(1 – 吸烟、0 – 不吸烟);X_2,是否喝咖啡(1 – 喝、0 – 不喝);X_3,性别(1 – 男、0 – 女)

1.1.3 一般计数资料 Poisson 分布回归模型的表达式

令 $\lambda(X)=e^{x_i'\beta}$,将其代入式(1),便可得到多重 Poisson 分布回归模型,见式(2):

$$P(Y=y_i|x_i')=\frac{(e^{x_i'\beta})^{yi}e^{-e^{x_i'\beta}}}{yi!} \quad (2)$$

值得注意的是:统计软件给出的计算结果是下面式(3)等号右边回归系数的估计值。也就是说,Poisson 分布回归模型的“核心部分”是其“期望值(或称均值)”,用自变量的线性组合来表达计数因变量 Y 的期望的自然对数。

$$\ln[\lambda(X)]=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta mx_m+\varepsilon \quad (3)$$

式(3)中的“ln”代表“自然对数函数”,$\lambda(X)$代表服从 Poisson 分布的离散型随机变量 Y 的均值,X 代表所有自变量构成的“向量”,ε 代表模型的误差。有些统计学教科书将式(3)称为“Poisson 分布回归模型”,这是不够准确的。它应被视为“Poisson 分布回归模型”的“内核”,即“核心内容”,外文文献中称其为“Poisson 分布回归模型”的“连接函数”。

1.2 一般计数资料 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(3)中的回归系数呢?首先,应该明白:$\lambda(X)$是 X 的函数,在实际资料中是无法观测到的,它随着 X 取值的改变而变化,是一个“隐变量”;回忆求解多重线性回归模型中的参数时所面临的“处境”,在那里,计量因变量 Y 的取值是可以观测到的,是“显变量”[3-6]。

显然,直接根据式(3)无法求出式中的回归系数。需要基于式(2)并采用“最大似然法”[7]构造“目标函数”,运用高等数学方法,将“目标函数”转变成“方程组”,再运用“计算数学”或“统计计算”等技术方法求解“方程组”。由于方程组为“非线性方程组”,一般没有直接解,需用到迭代计算法,如加权最小二乘迭代或牛顿 – 纳法生迭代法。从而获得式(3)的“参数估计”结果。因篇幅所限,参数估计公式的详细推导过程从略。

1.2.2 Possion 回归系数的假设检验

1.2.2.1 似然比检验

通过比较两个相嵌套模型(如模型 P 嵌套于模型 K 内)的对数似然函数统计量 G(又称 Deviance)来进行,其统计量见式(4):

$$G=G_P-G_K=-2\times(\text{模型 }P\text{ 的对数似然函数}-\text{模型 }K\text{ 的对数似然函数}) \quad (4)$$

其中,模型 P 中的变量是模型 K 中变量的一部分,另一部分就是需要检验的变量。这里,G 服从自由度为 $K-P$ 的 χ^2 分布。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1.2.2.2 回归系数的 Wald 检验

比较估计系数与“0”的差别是否有统计学意义，其检验统计量见式(5)：

$$z = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} \quad (5)$$

这里 z 为标准正态变量。参数的置信区间是基于 Wald 统计量导出的。β 的 95% 置信区间见式(6)：

$$\hat{\beta} - 1.96 \times SE(\hat{\beta}) \sim \hat{\beta} + 1.96 \times SE(\hat{\beta}) \quad (6)$$

1.2.3 Possion 拟合优度检验

1.2.3.1 Deviance 偏差统计量

设 $L(b_{max};y)$ 为全模型的似然函数，$L(b;y)$ 为选模型(含部分自变量的模型)的似然函数，则定义 λ 统计量见下面的式(7)：

$$\lambda = L(b_{max};y)/L(b;y) \quad (7)$$

计算 Deviance 偏差(D)统计量，采用公式(8)：

$$D = 2\ln\lambda = 2[\ln L(b_{max};y) - \ln L(b;y)] \quad (8)$$

且 D 服从 χ^2 分布。很显然，D 越大，模型的估计值与观测值的偏差就越大，拟合效果就越差。

1.2.3.2 广义 χ^2 统计量

广义 χ^2 统计量见式(9)：

$$\chi^2_G = \sum \frac{(y - \hat{\mu})^2}{V(\hat{\mu})} \quad (9)$$

显然，χ^2_G 越大，估计值与观测值的差别就越大，模型的拟合效果就越差。在正态分布中，χ^2_G 就是离差平方和；在 Poisson 分布或二项分布中，χ^2_G 就是一般的 Pearson χ^2。

Deviance 偏差具有可加性，而 χ^2_G 不具有这种性质。但 χ^2_G 比 Deviance 偏差更易解释。

2 一般计数资料 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“maop1006”的 SAS 数据集：

```
data maop1006;
    input obs X1 - X3 Y @@;
cards;
 1  0  1  1  11  16  1  0  0  11
 2  0  0  0   7  17  0  1  1   8
 3  0  0  0   3  18  1  0  1   9
 4  1  0  1   5  19  0  0  0   8
 5  0  0  0   2  20  1  0  0   5
 6  1  1  1  13  21  0  1  1   5
 7  0  1  0   6  22  1  1  0   8
 8  1  0  1  10  23  1  1  0  13
 9  0  0  0   4  24  0  0  1   8
10  1  0  1   7  25  1  0  0   6
11  0  0  0   1  26  0  0  1   4
12  0  0  1   9  27  0  0  0   6
13  0  0  1   6  28  1  1  1  13
14  1  1  1  17  29  1  1  0   9
15  0  0  0   5  30  0  0  1   5
;
run;
```

2.2 求出因变量 Y 的均值与方差

利用下面的两个 SAS 过程步程序，求出因变量 Y 的均值与方差。

```
proc univariate data = maop1006 noprint;
    var Y;
    output out = aaa mean = ybar var = yvar;
run;
proc print data = aaa;
    var ybar yvar;
run;
```

【SAS 输出结果】

Obs	ybar	yvar
1	7.46667	13.0161

求出 Y 的均值和方差分别为 7.467、13.016。

2.3 基于全部自变量对因变量 Y 构建多重 Poisson 分布回归模型

利用下面的 SAS 过程步构建 Poisson 分布回归模型：

```
proc genmod data = maop1006;
    model Y = X1 - X3/link = log dist = poisson;
run;
```

【SAS 程序说明】“link = log”表明采用“自然对数”作为连接函数，即对计数因变量取自然对数变换；“dist = poisson”表明要基于 Poisson 分布构建 Poisson 分布回归模型。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

【SAS 主要输出结果】

最大似然参数估计值的分析

参数	自由度	估计值	标准误差	Wald 95%	置信限	Wald 卡方	Pr > 卡方
Intercept	1	1.5066	0.1323	1.2473	1.7659	129.70	<.0001
X1	1	0.4162	0.1381	0.1455	0.6869	9.08	0.0026
X2	1	0.4012	0.1382	0.1304	0.6720	8.43	0.0037
X3	1	0.2546	0.1362	-0.0123	0.5214	3.50	0.0615
尺度	0	1.0000	0.0000	1.0000	1.0000		

以上结果表明:X3(性别)对因变量的影响无统计学意义。此结果是否正确,有待进一步研究。

2.4 在构建多重 Poisson 分布回归模型时检验因变量是否存在“过离散”现象

利用下面的 SAS 过程步程序构建 Poisson 分布回归模型,重点在于检验因变量是否存在“过离散”现象:

```
proc genmod data = maop1006;
    model Y = X1 - X3 /link = log dist = nb noscale;
run;
```

【SAS 程序说明】“dist = nb”说明指定资料中因变量(误差项)的分布为“负二项分布”;关键问题在于:只有指定为此分布时,才能发挥选项“noscale”的作用,它是将冗余参数 k 的取值固定为“0”,相当于进行 Poisson 分布回归分析,同时,采用拉格朗日(Lagrange)乘子检验“因变量是否存在‘过离散’现象”。

【SAS 主要输出结果】

第 1 部分输出结果同上,从略。

拉格朗日乘数统计量

参数	卡方	Pr > 卡方
离散度	8.6315	0.0017*

注:* 单侧 P 值

以上结果表明:因变量确实存在“过离散”现象,构建 Poisson 分布回归模型时应对其进行校正。

2.5 在构建多重 Poisson 分布回归模型时对“过离散”现象进行校正

利用下面的 SAS 过程步程序构建 Poisson 分布回归模型,对“过离散”现象进行校正:

```
proc genmod data = maop1006;
    model Y = X1 - X3 /link = log dist = poisson scale
 = deviance;
run;
```

【SAS 程序说明】选项“scale = deviance”是要求对“过离散”进行校正后,再构建 Poisson 分布回归模型。

【SAS 主要输出结果】

最大似然参数估计值的分析

参数	自由度	估计值	标准误差	Wald 95%	置信限	Wald 卡方	Pr > 卡方
Intercept	1	1.5066	0.1258	1.2600	1.7532	143.41	<.0001
X1	1	0.4162	0.1313	0.1588	0.6737	10.04	0.0015
X2	1	0.4012	0.1314	0.1436	0.6588	9.32	0.0023
X3	1	0.2546	0.1295	0.0008	0.5084	3.86	0.0493
尺度	0	0.9510	0.0000	0.9510	0.9510		

以上结果表明,尺度参数由“1”调整为“0.9510”,由此,模型中各参数的“标准误”均得到“校正”。故最后两列的数值也都得到校正。经校正后,原来无统计学意义的“X3”变为有统计学意义。由以上结果可以写出 Poisson 分布回归模型中的“内核”:

$$\ln[\lambda(X)] = 1.5066 + 0.4162X1 + 0.4012X2 + 0.2546X3$$

$$\lambda(X) = e^{1.5066 + 0.4162X1 + 0.4012X2 + 0.2546X3}$$

若将上面的 $\lambda(X)$ 代入本文第 1.1 节中的式(2),就可获得完整的“Poisson 分布回归模型”。

【专业结论】因 X1、X2 和 X3 前的系数分别为 $\exp(0.4162) = 1.516189$、$\exp(0.4012) = 1.493616$ 和 $\exp(0.2546) = 1.289946$,又因为 X1 代表“是否吸烟(1 - 吸烟、0 - 不吸烟)”、X2 代表“是否喝咖啡

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

(1 – 喝、0 – 不喝)”、X3 代表“性别(1 – 男、0 – 女)”,说明吸烟者出现早搏的概率是不吸烟者的1.516倍、喝咖啡者出现早搏的概率是不喝咖啡者的 1.494 倍,而男性受试者出现早搏的概率是女性受试者的 1.290 倍。

参考文献

[1] 茆诗松. 统计手册[M]. 北京: 科学出版社, 2003: 1004 – 1007.

[2] 胡良平. 面向问题的统计学——(2)多因素设计与线性模型分析[M]. 北京: 人民卫生出版社, 2012: 283 – 290.

[3] 胡良平. 多重线性回归分析的核心内容与关键技术概述[J]. 四川精神卫生, 2018, 31(1): 1 – 6.

[4] 谷恒明, 胡良平. 基于经典统计思想实现多重线性回归分析[J]. 四川精神卫生, 2018, 31(1): 7 – 11.

[5] 谷恒明, 胡良平. 基于贝叶斯统计思想实现多重线性回归分析[J]. 四川精神卫生, 2018, 31(1): 12 – 14.

[6] 谷恒明, 胡良平. 基于机器学习统计思想实现多重线性回归分析[J]. 四川精神卫生, 2018, 31(1): 15 – 18.

[7] 柳青. 中国医学统计百科全书——多元统计分册[M]. 北京: 人民卫生出版社, 2004: 231 – 233.

(收稿日期:2018 – 10 – 10)

(本文编辑:唐雪莉)

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net