过离散计数资料负二项分布模型回归分析

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】 本文目的是介绍过离散(即方差明显大于均值)计数资料负二项分布模型回归分析。首先,介绍了过离散计数资料及其负二项分布回归模型构建原理,包括"过离散计数资料负二项分布回归模型的形式"和"过离散计数资料负二项分布回归模型的求解";第二,介绍了"过离散计数资料负二项分布回归模型的 SAS 实现",包括:①"创建 SAS 数据集";②"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"和"基于全部自变量对因变量 Y 构建多重负二项分布回归模型"。本文结果提示,在"过离散"非常严重的情况下,应使用"负二项分布回归模型"取代"Poisson 分布回归模型"。否则,易得出不正确的结果和结论。

【关键词】 计数资料;过离散;Poisson 分布回归模型;负二项分布回归模型;拉格朗日乘子
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.05.003

The regression analysis of the negative binomial distribution model for the over-dispersion 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 negative binomial distribution model for the over-dispersion count data. Firstly, the concepts of the over-dispersion count data and the building principle of the negative binomial distribution regression model were given, which included the following two aspects: ①the form of the negative binomial distribution regression model; ②the solution for the model mentioned before. Secondly, the SAS realization of the negative binomial distribution regression model was presented. The contents were as follows: ①creating SAS data set; ②calculating the arithmetic mean and variance of the dependent variable Y; ③checking if there was the over-dispersion in the dependent variable Y; ④building a multiple negative binomial distribution regression model for the dependent variable Y based on all independent variables. The results of the article showed that, under the situation of the severe over-dispersion, the harmful results came from the over-dispersion could be adjusted perfectly by using the negative binomial distribution regression model instead of the Poisson distribution regression model. Otherwise, incorrect results and conclusions could be gotten.

【**Keywords**】 Count data; Overdi-spersion; Poisson distribution regression model; Negative binomial distribution regression model; Lagrange multiplier

1 过离散计数资料及其负二项分布模型构建原理

1.1 适于过离散计数资料负二项分布回归模型的数据结构

适于过离散计数资料负二项分布模型的数据结构见表 1[1]。

表 1 数据来自德国的卫生改革项目。改革的目的是减少患者到医生处的就诊次数。数据共包括 2 227例观测,分别属于改革之前的 1996 年和改革之后的 1998 年。响应变量为患者在 3 个月之内的就诊次数,自变量为改革前后(0 = 改革前,1 = 改革后)、健康状况(0 = 良好,1 = 不良)、年龄、受教育时间和家庭收入的对数值。由于数据量较大,表 1 中仅列出部分结果。求出"就诊次数"的均值与方差如下。

	均值	方差
1	2.58913	16.1299

【对数据结构的分析】因变量 Y"就诊次数"为"计数变量",其均值为 2.589、方差为 16.130,方差明显大于均值,只能将 Y 视为服从负二项分布的离散型随机变量;拟考察的四个自变量中,有两个是"二值变量",另外两个是"计量变量"。虽然,在统

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

计学的理论上，默认自变量全为“计量变量”，但从实用性角度出发，“二值”自变量也是可以接受的。

表 1 患者就诊次数及其影响因素数据

观测号	就诊次数	改革与否	健康状况	年龄	受教育时间	家庭收入(取对数后)
1	1	0	0	45	10.5	7.6368
2	9	0	1	53	9.0	7.6992
3	40	0	1	48	10.5	7.0574
4	0	1	0	52	18.0	7.6886
5	1	0	0	40	10.5	7.5415
6	1	1	0	42	10.5	7.3319
7	0	0	1	57	10.5	7.4289
8	0	1	0	23	8.5	7.0978
9	3	0	0	55	10.0	7.8458
…	…	…	…	…	…	…
2 227	0	1	0	38	7.0	7.7232

1.2 过离散计数资料负二项分布回归模型的构建与求解[2]

1.2.1 Poisson 分布回归模型的推广

在对第 i 个个体进行观测时，通过引入一个观测不到的非齐性项“τ_i”，从而使 Poisson 分布回归模型得到推广。因此，假定个体之间以随机的方式呈现彼此的差异，原因在于可观测的协变量不能完全解释不同个体之间实际存在的差异。此种情况可用如下的式子来呈现，见式(1)：

$$E(y_i \mid x_i,\tau_i)=\mu_i\tau_i=e^{x_i\beta+\varepsilon_i} \tag{1}$$

在式(1)中，观测不到的非齐性项 $\tau_i=e^{\varepsilon_i}$ 独立于自变量向量 x_i。于是，y_i 关于 x_i 和 τ_i 的条件分布是关于条件均值和条件方差均为 $\mu_i\tau_i$ 的 Poisson 分布，其概率函数见式(2)：

$$f(y_i \mid x_i,\tau_i)=\frac{\exp(-\mu_i\tau_i)(\mu_i\tau_i)^{y_i}}{y_i!} \tag{2}$$

令 $g(\tau_i)$ 为 τ_i 的概率密度函数，于是，通过对 $f(y_i \mid x_i,\tau_i)$ 求关于 τ_i 的积分，便可求得 $f(y_i \mid x_i)$ 的概率分布，见式(3)：

$$f(y_i \mid x_i)=\int_0^{\infty} f(y_i \mid x_i,\tau_i)g(\tau_i)\,d\tau_i \tag{3}$$

在式(3)中，当假定 τ_i 服从伽玛分布时，其积分的结果存在解析解，这个解就是负二项分布。

为了确定分布的均值，当模型中包含常数项时，必需假定 $E(e^{\varepsilon_i})=E(\tau_i)=1$。因此，假定 τ_i 为服从 $gamma(\theta,\theta)$ 分布且其均值和方差分别为 $E(\tau_i)=1$ 和 $V(\tau_i)=1/\theta$，其概率密度函数见式(4)：

$$g(\tau_i)=\frac{\theta^{\theta}}{\Gamma(\theta)}\tau_i^{\theta-1}\exp(-\theta\tau_i) \tag{4}$$

在式(4)中，分母中的伽玛函数见式(5)：

$$\Gamma(x)=\int_0^{\infty} z^{x-1}\exp(-z)\,dz \tag{5}$$

而式(4)中的 θ 是一个正的参数。

1.2.2 负二项分布概率函数形式之一

于是，在给定了 x_i 的条件下，推导出 y_i 的概率函数见式(6)：

$$\begin{aligned} f(y_i \mid x_i) &= \int_0^{\infty} f(y_i \mid x_i,\tau_i)g(\tau_i)\,d\tau_i \\ &= \frac{\theta^{\theta}\mu_i^{y_i}}{y_i!\Gamma(\theta)}\int_0^{\infty} e^{-(\mu_i+\theta)\tau_i}\tau_i^{\theta+y_i-1}\,d\tau_i \\ &= \frac{\theta^{\theta}\mu_i^{y_i}\Gamma(y_i+\theta)}{y_i!\Gamma(\theta)(\theta+\mu_i)^{\theta+y_i}} \\ &= \frac{\Gamma(y_i+\theta)}{y_i!\Gamma(\theta)}\left(\frac{\theta}{\theta+\mu_i}\right)^{\theta}\left(\frac{\mu_i}{\theta+\mu_i}\right)^{y_i} \end{aligned} \tag{6}$$

1.2.3 负二项分布概率函数形式之二

令 $\alpha=\frac{1}{\theta}(\alpha>0)$，将其取代式(6)中的 θ，于是，负二项分布概率函数可以被写成下面式(7)的形式：

$$f(y_i \mid x_i)=\frac{\Gamma(y_i+\alpha^{-1})}{y_i!\,\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu_i}\right)^{y_i},\quad y_i=0,1,2\cdots \tag{7}$$

因此，负二项分布概率函数是由服从 Poisson 分布与伽玛分布的两类随机变量混合而成的。其条件均值和条件方差分别见式(8)和式(9)：

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

$$E(y_i|x_i)=\mu_i=e^{x_i'\beta} \quad (8)$$

$$V(y_i|x_i)=\mu_i\left[1+\frac{1}{\theta}\mu_i\right]=\mu_i\ [1+\alpha\mu_i]>E(y_i|x_i) \quad (9)$$

由式(9)可知,负二项分布的条件方差明显大于其条件均值,这种“过离散”的结果是忽视未观测到的非齐性项所致。

式(9)显示出:负二项分布的方差函数是其均值的二次方,这个结果可被称为第 2 种定义下的负二项分布回归模型(即“Cameron and Trivedi 1986”),简称为“2 型负二项分布回归模型”。当 $\alpha=0$ 时,负二项分布就退化成为 Poisson 分布了。检验“$\alpha=0$”这个假设是否成立,可使用 WALD 检验。

1.2.4 2 型负二项分布回归模型的自然对数似然函数及一阶偏导数

2 型负二项分布回归模型的自然对数似然函数可由下面的式(10)给出:

$$L=\sum_{i=1}^{n}w_i\left\{\sum_{j=0}^{y_i-1}\ln(j+\alpha^{-1})-\ln(y_i!)-(y_i+\alpha^{-1})\ln(1+\exp(x_i'\beta))+y_i\ln(\alpha)+y_ix_i'\beta\right\} \quad (10)$$

在式(10)中,w_i 的定义很繁琐,参见有关文献,此处从略;在式(10)等号右边第二个求和号之后的第一项是怎么得到的?它是由如下的原因所致:在式(7)中,两个伽玛函数相除可以改写成乘积形式,见下面的式(11):

$$\Gamma(y+a)/\Gamma(a)=\prod_{j=0}^{y-1}(j+a) \quad (11)$$

基于式(10),求 L 关于参数 β 的偏导数,见下面的式(12):

$$\frac{\partial L}{\partial\beta}=\sum_{i=1}^{N}w_i\frac{y_i-\mu_i}{1+\alpha\mu_i}X_i \quad (12)$$

基于式(10),求 L 关于参数 α 的偏导数,见下面的式(13):

$$\frac{\partial L}{\partial\alpha}=\sum_{i=1}^{N}w_i\left\{-\alpha^{-2}\sum_{j=0}^{y_i-1}\frac{1}{(j+\alpha^{-1})}+\alpha^{-2}\ln(1+\alpha\mu_i)+\frac{y_i-\mu_i}{\alpha(1+\alpha\mu_i)}\right\} \quad (13)$$

1.2.5 1 型负二项分布回归模型的自然对数似然函数及一阶偏导数

Cameron and Trivedi(1986)考虑负二项分布模型的一般类型,其均值为 μ_i、方差为 $\mu_i+\alpha\mu_i^p$。$p=2$ 的 2 型负二项分布模型是负二项分布模型的标准公式。具有其他 p 值(即 $-\infty<p<\infty$)的模型具有相同的概率函数 $f(y_i|x_i)$,但需要用 $\alpha^{-1}\mu^{2-p}$ 取代 α^{-1}。若令 $p=1$,就可获得 1 型负二项分布模型,此时,其方差函数见式(14):

$$V(y_i|x_i)=\mu_i+\alpha\mu_i \quad (14)$$

为了估计这样的模型,在“model 语句”中指定“DIST=NEGBIN($p=1$)”这样的选项。

于是,1 型负二项分布回归模型的自然对数似然函数由下面的式(15)给出:

$$L=\sum_{i=1}^{N}w_i\left\{\sum_{j=0}^{y_i-1}\ln(j+\alpha^{-1}\exp(x_i'\beta))-\ln(y_i!)-(y_i+\alpha^{-1}\exp(x_i'\beta))\ln(1+\alpha)+y_i\ln(\alpha)\right\} \quad (15)$$

在式(15)中,w_i 为权重系数,其定义视具体情况而定,因篇幅所限,此处从略。

基于式(15),求 L 关于参数 β 的偏导数,见下面的式(16):

$$\frac{\partial L}{\partial\beta}=\sum_{i=1}^{N}w_i\left\{\left(\sum_{j=0}^{y_i-1}\frac{\mu_i}{(j\alpha+\mu_i)}\right)x_i-\alpha^{-1}\ln(1+\alpha)\mu_ix_i\right\} \quad (16)$$

基于式(15),求 L 关于参数 β 的偏导数,见下面的式(17):

$$\frac{\partial L}{\partial\alpha}=\sum_{i=1}^{N}w_i\left\{-\left(\sum_{j=0}^{y_i-1}\frac{\alpha^{-1}\mu_i}{(j\alpha+\mu_i)}\right)-\alpha^{-2}\mu_i\ln(1+\alpha)-\frac{y_i+\alpha^{-1}\mu_i}{1+\alpha}+\frac{y_i}{\alpha}\right\} \quad (17)$$

1.2.6 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 过离散计数资料负二项分布回归模型的 SAS 实现[3]

2.1 创建 SAS 数据集

利用下面的 SAS 数据步程序,创建名为“nbd1”的 SAS 数据集:

```
data nbd1;
infile 'd:\SASTJFX\mdvisit.dat';
input id numvisit reform badh age edu loginc;
run;
```

【SAS 程序说明】将本例中全部 2 227 行数据(注意:表 1 中仅列出了前 9 行和最后一行)录入计算机,以“文本格式”且以“mdvisit.dat”为文件名,存

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

储在名为“SASTJFX”的 D 盘文件夹内。每行有 7 个数据，分别为“编号(id)”“三个月内的就诊次数(numvisit)”“改革前后(reform)(0 = 改革前，1 = 改革后)”“健康状况(badh)(0 = 良好，1 = 不良)”“年龄(age)”“受教育时间(edu)”和“家庭收入的对数值(loginc)”。

2.2 求出因变量 numvisit 的均值和方差

利用下面的两个 SAS 过程步程序，求出因变量 Y 的均值和方差：

```
proc univariate data = nbd1 noprint;
    var numvisit;
    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	2.58913	16.1299

求出 3 个月内就诊次数(numvisit)的均值和方差分别为 2.589 和 16.130。此结果表明，方差约为均值的 6.23 倍，属于较严重的“过离散”。

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，尝试构建 Poisson 分布回归模型，目的是为了使用选项“noscale”进行拉格朗日乘子检验，以明确是否存在明显的“过离散”现象。

```
proc genmod data = nbd1;
model numvisit = reform badh age edu loginc/link = log dist = nb noscale;
run;
```

【SAS 程序说明】“link = log”表明采用“自然对数”作为联接函数，即对计数因变量取自然对数变换；“dist = nb”表明要基于负二项分布构建负二项分布回归模型。但是，选项“noscale”的作用是将冗余参数 k 的取值固定为“0”，此时，相当于构建 Poisson 回归模型。与此同时，该选项的主要用途是进行拉格朗日乘子检验，以明确因变量是否存在明显的“过离散”现象。

【SAS 主要输出结果】基于 Poisson 分布回归模型的参数估计结果没有实用价值，此处从略。

拉格朗日乘数统计量

参数	卡方	Pr > 卡方
离散度	551.7077	<.0001*

注：* 单侧 P 值

以上结果表明，因变量存在极严重的“过离散”现象。

2.4 尝试对“过离散”现象进行校正后再构建 Poisson分布回归模型

利用下面的 SAS 过程步程序对“过离散”现象进行校正后再构建 Poisson 分布回归模型：

```
proc genmod data = nbd1;
model numvisit = reform badh age edu loginc/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】

评估拟合优度的准则

准则	自由度	值	值/自由度
偏差	2221	7422.1244	3.3418
调整后的偏差	2221	2221.0000	1.0000
Pearson 卡方	2221	9681.6920	4.3592
调整后的 Pearson X^2	2221	2897.1541	1.3044
对数似然		129.4790	
完全对数似然		-5943.8280	
AIC(越小越好)		11899.6561	
AICC(越小越好)		11899.6939	
BIC(越小越好)		11933.9066	

以上是采用 Poisson 分布回归模型拟合此资料所对应的“拟合优度”评价指标及其取值，这些结果需要与其他同类模型比较，才有价值。

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

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	-0.4215	0.4917	-1.3852	0.5422	0.73	0.3913
reform	1	-0.1399	0.0485	-0.2350	-0.0448	8.31	0.0039
badh	1	1.1326	0.0554	1.0241	1.2412	418.17	<.0001
age	1	0.0049	0.0023	0.0004	0.0094	4.55	0.0329
edu	1	-0.0118	0.0109	-0.0332	0.0095	1.18	0.2781
loginc	1	0.1520	0.0658	0.0231	0.2810	5.34	0.0208
尺度	0	1.8281	0.0000	1.8281	1.8281		

Note: The scale parameter was estimated by the square root of DEVIANCE/DOF

最后一行“注释”的含义是:尺度参数是“离差/自由度”的平方根,即(7422.124/2221)的平方根,其数值为 1.8281。

以上结果表明:尺度参数由原先的“1”调整为“1.8281”,由此,模型中各参数的“标准误”都得到“校正”。从而,最后两列的数值也都得到校正。

2.5 在“过离散”现象严重的情况下构建多重负二项分布回归模型

```
proc genmod data = nbd1;
model numvisit = reform badh age edu loginc/link = log dist = nb;
run;
```

【SAS 主要输出结果】

评估拟合优度的准则

准则	自由度	值	值/自由度
偏差	2221	2412.3464	1.0862
调整后的偏差	2221	2412.3464	1.0862
Pearson 卡方	2221	2693.5514	1.2128
调整后的 Pearson X^2	2221	2693.5514	1.2128
对数似然		1813.1299	
完全对数似然		-4563.3905	
AIC(越小越好)		9140.7809	
AICC(越小越好)		9140.8314	
BIC(越小越好)		9180.7398	

以上“拟合优度评价指标的取值”与前面基于“Poisson 分布回归模型”的相应评价指标的取值相比,现在的模型,即“负二项分布回归模型”对此资料的拟合效果好多了(偏差变小了,AIC、AICC 和 BIC 都变小了很多)。

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	-0.4075	0.5336	-1.4533	0.6384	0.58	0.4451
reform	1	-0.1374	0.0511	-0.2376	-0.0372	7.22	0.0072
badh	1	1.1312	0.0748	0.9847	1.2778	228.84	<0.0001
age	1	0.0056	0.0024	0.0009	0.0103	5.39	0.0203
edu	1	-0.0053	0.0115	-0.0278	0.0172	0.21	0.6430
loginc	1	0.1371	0.0712	-0.0024	0.2767	3.71	0.0541
离散度	1	1.0021	0.0476	0.9130	1.1000		

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

以上结果表明，截距项无统计学意义，去掉截距项后，可使模型精简一些。

2.6 构建不含截距项且删除无统计学意义的自变量 edu 的负二项分布回归模型

```
proc genmod data = nbd1;
model numvisit = reform badh age loginc/noint link = log dist = nb;
run;
```

【SAS 主要输出结果】

评估拟合优度的准则

准则	自由度	值	值/自由度
偏差	2223	2412.1691	1.0851
调整后的偏差	2223	2412.1691	1.0851
Pearson 卡方	2223	2687.2382	1.2088
调整后的 Pearson X^2	2223	2687.2382	1.2088
对数似然		1812.7232	
完全对数似然		-4563.7972	
AIC(越小越好)		9137.5943	
AICC(越小越好)		9137.6213	
BIC(越小越好)		9166.1364	

与前面的结果相比，拟合优度评价指标的取值又有所改善。

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	0	0.0000	0.0000	0.0000	0.0000	.	.
reform	1	-0.1397	0.0510	-0.2398	-0.0397	7.49	0.0062
badh	1	1.1309	0.0745	0.9848	1.2769	230.29	<0.0001
age	1	0.0056	0.0024	0.0010	0.0103	5.71	0.0169
loginc	1	0.0764	0.0119	0.0531	0.0997	41.17	<0.0001
离散度	1	1.0029	0.0477	0.9137	1.1008		

基于以上的结果可以写出最终的负二项分布回归模型的“内核”，如下：

$$\ln[\mu(X)] = -0.1397\text{reform} + 1.1309\text{badh} + 0.0056\text{age} + 0.0764\text{loginc}$$

$$\mu(X) = e^{-0.1397\text{reform} + 1.1309\text{badh} + 0.0056\text{age} + 0.0764\text{loginc}}$$

若将上面“$\mu(X)$”等号后面的内容代入式(6)或式(7)，便可获得多重负二项分布回归模型的完整表达式。

【专业结论】改革与否、健康状况和年龄这三个因素对患者就诊次数的影响具有统计学意义，结合参数估计的结果来看：改革与否的系数为 -0.1397，说明改革之后，患者的就诊次数下降了。具体地说，因 $\exp(-0.1397) = 0.869619$，即就诊次数下降了约 $1 - 0.869619 = 13.04\%$；相对于良好健康状况、年龄小的患者，具有不良健康状况、年龄大均会使就诊次数增加(因为它们的回归系数为正值)。

参考文献

[1] 胡良平. 医学统计学——运用三型理论进行现代回归分析[M]. 北京：人民军医出版社，2010：148-157.

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

[3] 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，2012：291-297.

(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