合理进行均值比较——泊松分布回归模型

胡纯严[1]，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：lphu927@163.com）

【摘要】 本文目的是介绍与泊松分布回归模型有关的6个基本概念、计算方法、一个临床调查实例及其SAS实现。基本概念包括泊松分布、泊松分布回归模型、偏移量、标准化死亡比（SMR）、偏差信息准则和最高后验密度区间。计算方法涉及泊松分布回归参数的经典估算方法和贝叶斯估算方法。临床调查实例涉及1975年-1980年苏格兰56个县的唇癌观察和预期病例的数据。本文给出了采用SAS处理实例中计数资料的全过程，包括基于bglimm过程构建5个泊松分布回归模型和展示预测的SMR与观测的SMR之间的吻合程度。对输出结果作出了解释，并基于模型拟合效果评价统计量，对所构建的多个泊松分布回归模型进行比较，得出了适合本文资料的最优泊松分布回归模型。

【关键词】 泊松分布回归模型；偏移量；标准化死亡比；偏差信息准则；最高后验密度区间
中图分类号：R195.1 文献标识码：A doi：10.11886/scjsws20230201003

Reasonably carry out mean value comparison: Poisson distribution regression models

Hu Chunyan[1], *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*: *lphu927@163.com*）

【Abstract】 The purpose of this paper was to introduce 6 basic concepts, calculation methods, a clinical investigation example and its SAS implementation related to the Poisson distribution regression model. The basic concepts included the Poisson distribution, Poisson distribution regression models, offsets, standardized mortality ratio (SMR), deviation information criteria and the highest posterior density intervals. The calculation method involved the classical estimation method and the Bayesian estimation method of the Poisson distribution regression parameters. The clinical investigation example involved the data on observed and expected cases of lip cancer in 56 Scottish counties from 1975 to 1980. This article presented the whole process of using SAS software to deal with the count data in the example, including constructing five Poisson distribution regression models based on the bglimm procedure and showing the degree of agreement between the predicted SMR and the observed SMR. The output results were explained, and based on the evaluation statistics of the model fitting effect, the multiple constructed Poisson distribution regression models were compared, and finally the optimal Poisson distribution regression model suitable for the data in the paper was obtained.

【Keywords】 Poisson distribution regression model; Offset; Standardized mortality ratio; Deviation information criterion; Highest posterior density interval

在单因素k（设k=2）水平下收集的两个服从泊松分布的计数结果，可以采用Z检验进行均值之间的比较[1-2]。然而，在多个协变量影响下，且当k>10时，收集的k个服从泊松分布的计数结果，就不适合采用Z检验了。此时，需要构建泊松分布回归模型。根据问题的复杂程度，回归模型可能是普通的广义线性回归模型[3]，也可能需要采用广义混合效应回归模型[4-5]。本文将结合一个临床调查资料，展示如何合理选择拟合效果好的泊松分布回归模型[1,5]。

1 基本概念

1.1 泊松分布

定义：若离散型随机变量X的取值为非负整数，且相应的概率函数由式（1）给出，则称随机变量X服从泊松分布，记作X~P(k；λ)。

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

式（1）中，k为随机变量X的具体取值，λ为随机变量X的总体平均值。

1.2 泊松分布回归模型

定义：设 Y 是一个服从泊松分布的随机变量，$X=(1,x_1,x_2,\cdots,x_m)'$ 是一个协变量向量，$\beta=(\beta_0,\beta_1,\beta_2,\cdots,\beta_m)'$ 是参数向量。若 Y 的数学期望的对数可以表示为协变量的线性表达式：$E(Y|X)=\exp(X'\beta)$，则称(X,Y)服从泊松分布回归模型[1]。对应的表达式见式(2)。

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

式(2)中，均值$\lambda(X)$的表达式见式(3)：

$$\lambda(X)=\exp(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_mx_m)=\exp(X'\beta) \quad (3)$$

1.3 偏移量

文献[6]给出了地区特定相对风险的扩展模型，见式(4)。

$$\lambda_i=\exp\{\beta_0+\beta_1x+\gamma_i\},i=1,\cdots,56 \quad (4)$$

式(4)中，β_0和β_1是固定效应参数，γ_i是第i县的随机效应参数，$x=\text{Employment}_i/10$是从事农业、渔业和林业工作的人口比例。与第i个县对应的泊松变量的平均值(以随机效应为条件)可用下式表达，见式(5)。

$$\mu_i=E_i\lambda_i \quad (5)$$

将式(4)代入式(5)等号右边，再对该式两边取对数，得到式(6)。

$$\log(\mu_i)=\log(E_i)+\beta_0+\beta_1\text{Employment}_i/10+\gamma_i \quad (6)$$

式(6)中，$\log(E_i)$是一个偏移量，一个回归变量，已知其系数为1。注意，假设E_i是已知的，它们是与各县对应的一个比例常数。

1.4 标准化死亡比

标准化死亡比(standardized mortality ratio，SMR)是指某人群实际死亡数与预期死亡数之比。预期死亡数是某人群(通常为某一特殊职业人群)各年龄组人口数与标准人口的各年龄组死亡率乘积的总和。

1.5 偏差信息准则

偏差信息准则(deviation information criterion，DIC)是评价模型对资料拟合效果的统计量，它是通过使用模型中参数的后验均值估计值来计算的。在SAS输出的“偏差信息准则表”中，还包括其他3个相关的统计量，即偏差的后验均值(Dbar)、后验均值评估的偏差(Dmean)和有效参数个数(pD)[5]。

1.6 最高后验密度(HPD)区间

100(1−α)%最高后验密度(HPD)区间是满足以下两个条件的区域：①该区域的后验概率为100(1−α)%；②该区域内任何点的最小密度大于或等于该区域外任何点的密度。HPD是所估计参数的大部分分布所在的区间。一些统计学家更喜欢这个区间，因为它是最小的区间。

2 计算方法

2.1 泊松回归参数估计

基于最大似然法求泊松分布回归模型中参数估计值的步骤如下[1]：第一步，基于泊松分布回归模型构建对数似然函数$l(\beta)$；第二步，对对数似然函数求各参数的二阶偏导数，并形成估计方程组；第三步，求解估计方程组，得到各参数的估计值。值得一提的是，采用Newton-Raphson迭代法对$l(\beta)$求最大值，同样可得到参数β的最大似然估计$\hat{\beta}_{\text{mle}}$。

基于贝叶斯理论和马尔科夫链蒙特卡罗(MCMC)方法求泊松分布回归模型中参数估计值的方法非常复杂[7-8]，它主要基于不同的回归模型，选择不同的随机抽样算法，从已知样本中产生与各参数对应的马尔科夫链随机样本，通过大规模随机抽样，以各参数的大样本随机抽样结果的均值作为各参数的估计值，并构造各参数的95% HPD[5]。

2.2 偏差信息准则的计算

偏差信息准则(DIC)是一种模型评估工具，它是Akaike信息准则(AIC)和贝叶斯信息准则(BIC，也称为Schwarz准则)的贝叶斯替代方法[9]。DIC使用后验密度，这意味着它考虑了先验信息。DIC可应用于非嵌套模型和具有非独立同分布数据的模型。MCMC中DIC的计算是微不足道的——它不需要参数空间的最大化，如AIC和BIC。较小的DIC表示所拟合的模型更适合数据集。

让θ代表模型的参数，DIC的公式见式(7)。

$$DIC=\overline{D(\theta)}+p_D=D(\bar{\theta})+2p_D \quad (7)$$

式(7)中，$D(\theta)=2\{\log[f(y)]-\log[P(y|\theta)]\}$，其中，$P(y|\theta)$代表具有归一化常数的似然函数；$f(y)$是一个标准化项，是数据的唯一函数，该项相对于参数是常数。由于该项在DIC比较中被抵消，故通常省略其计算。

3 实例与SAS实现

3.1 问题与数据结构

3.1.1 一个临床调查问题及数据

【例1】文献[10]提供了1975年-1980年苏格兰56个县的唇癌观察和预期病例的数据。预期病例数是由一个单独的乘法模型确定的，该模型考虑了各县人口的年龄分布。原作者收集到的数据(共56行)形式见表1。试完成以下3项任务：构建由协变量x=Employment/10预测患唇癌人数的回归模型；基于DIC评价不同模型对资料的拟合效果；展示预测的SMR与观测的SMR之间的吻合程度。

表1 1975年-1980年苏格兰56个县的唇癌观察和预期病例的数据

Table 1 Data of observed and expected cases of lip cancer in 56 counties of Scotland from 1975 to 1980

County	Observed	Expected	Employment	SMR
1	9	1.4	16	652.2
2	39	8.7	16	450.3
3	11	3.0	10	361.8
…	…	…	…	…
54	1	7.0	1	14.2
55	0	4.2	16	0.0
56	0	1.8	10	0.0

注：County是“县”编号；Observed为观察的唇癌患者人数；Expected为预期的唇癌患者人数；Employment为从事农业、渔业和林业工作的人口比例；SMR为标准化死亡率

3.1.2 对数据结构的分析

数据集中的“县(County)”是观察单位，相当于普通统计资料中的“受试对象”；观察的唇癌患者人数(Observed)是一个的结果变量；期望的唇癌患者人数(Expected)是一个计量的结果变量；从事农业、渔业和林业工作的人口比例(Employment)是一个计量的自变量；标准化死亡率(SMR)是一个计量的结果变量。

这是一个非常特殊的数据结构，真正可以用于建模的变量为“Observed”和“Employment”。假设Observed服从泊松分布，基于此分布的理论，在泊松分布回归模型中需要引入一个偏移量(logN)。logN的计算见式(8)。

$$logN = log(100*Observed/SMR) \quad (8)$$

值得一提的是，服从泊松分布的随机变量有一个重要特性，即每个取值可以被视为一个“均值”。也就是说，在本例中，每个县Observed的取值都是一个均值。直接比较56个县Observed值的意义并不大，人们关注的是Observed随Employment变化的依赖关系，即需要构建带偏移量的泊松分布回归模型，并将其用于预测。

3.1.3 创建SAS数据集

设所需要的SAS程序如下：

```
data LipCancer;
input County Observed Expected Employment SMR;
if (Observed>0) then ExpCount= 100*Observed/SMR;
else ExpCount = Expected;
x=Employment/10;
LogN=log(ExpCount);
datalines;
1  9  1.4  16  652.2
2  39  8.7  16  450.3
3  11  3.0  10  361.8
55  0  4.2  16  0.0
56  0  1.8  10  0.0
;
run;
```

【变量说明】County代表“县”；Observed和Expected分别代表观察的和预期的唇癌患者人数；Employment代表从事农业、渔业和林业工作的人口比例；SMR代表标准化死亡率；x代表Employment/10；ExpCount代表100*Observed/SMR；LogN代表log(ExpCount)

3.2 用SAS实现统计分析

3.2.1 呈现观测的唇癌患者人数的频数分布

设所需要的SAS程序如下：

```
data abc;
set LipCancer;
proc sort;
by Observed;
run;
proc univariate data=abc;
var Observed;
histogram Observed/ vscale=count
endpoints=0 to 39 by 1 barlabel=count ;
run;
```

【SAS输出结果及解释】56个县患唇癌人数的频数分布见图1。56个县患唇癌人数呈正偏态分布，与标准的泊松分布比较接近[11]。可基于泊松分布理论，构建因变量关于自变量的泊松分布回归模型。

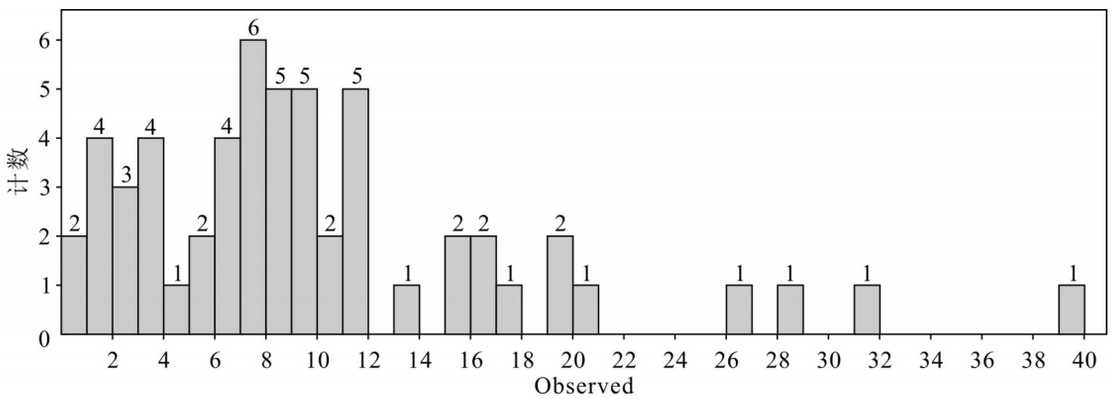

图1 56个县患唇癌人数的频数直方图

Figure 1 Frequency histogram of the number of lip cancer patients in 56 counties

3.2.2 拟合5种泊松分布回归模型

模型1：拟合既不带偏移量也不带随机截距的泊松分布回归模型，DIC=536.497。模型2：拟合带偏移量但不带随机截距的泊松分布回归模型，DIC=451.053。模型3：拟合不带偏移量但带随机截距的泊松分布回归模型，DIC=314.917。模型4：拟合既带偏移量也带随机截距的泊松分布回归模型，DIC=309.501。模型5：基于模型4，引入自变量x的平方项，DIC=308.802。

由于在待估参数相同或接近的条件下，DIC值越小，模型对资料的拟合效果越好，故本例资料选择模型5为宜。拟合模型5所需要的SAS程序如下：

```
proc bglimm data=LipCancer seed=10571042 nmc=10000
outpost=LipCancer_Out DIC;
class County;
model Observed =x x*x/dist=poisson offset=LogN;
random int / sub=County;
run;
```

【SAS主要输出结果及解释】模型中各参数的后验汇总和区间的计算结果见表2。由表2可知，截距、x的一次方和二次方的回归系数以及截距的随机效应方差与0之间的差异均有统计学意义（因为各行上最后两个数均不包含0）。

表2 后验汇总和区间的计算结果

Table 2 Calculation results of the posterior summaries and intervals

参 数	数 目	均 值	标准差	95% HPD	
Intercept	10 000	−0.793	0.219	−1.202	−0.346
x	10 000	1.616	0.455	0.729	2.502
x*x	10 000	−0.433	0.201	−0.844	−0.054
Random Var	10 000	0.438	0.114	0.245	0.671

评价模型对资料拟合效果的拟合统计量DIC的计算结果见表3。输出结果为评价模型对资料拟合效果的4个偏差统计量。

表3 偏差信息准则的计算结果

Table 3 Calculation results of deviation information criteria

统计量	估计值
Dbar（偏差的后验均值）	267.697
Dmean（后验均值评估的偏差）	226.592
pD（有效参数个数）	41.105
DIC（偏差信息准则）	308.802

3.2.3 展示预测的SMR与观测的SMR之间的吻合程度

设所需要的SAS程序如下（在运行前面模型4之后运行以下程序）：

```
data SMR_PRED;
array gamma [56] Intercept__County_1-Intercept__County_56;
array SMR_pred[56];
set LipCancer_Out;
do i = 1 to 56;
set LipCancer(rename=(x=data_x)) point=i;
SMR_pred [i] =100*exp (Intercept+x*data_x+gamma[i]);
end;
keep smr_pred:;
run;
% sumint (data=SMR_PRED, var= _numeric_, print=NO, out=SMR_SI)
data combine;
merge LipCancer SMR_SI;
run;
proc sgplot data=combine noautolegend aspect=1;
```

```
yaxis label="Predicted SMR" max=700;
xaxis label="Observed SMR" max=700;
text x=SMR y=mean text=employment;
lineparm x=0 y=0 slope=1;
run;
```

【SAS 主要输出结果及解释】基于模型预测的 SMR 与观测的 SMR 之间吻合程度较好。若基于前面的模型 1、模型 2、模型 3 和模型 5，得到结果相似，其吻合程度均较差，因篇幅所限，此部分图形从略。

3.3 结论

针对本例资料，采用 5 种泊松分布回归模型拟合资料，模型 5 的拟合效果最佳，该模型引入了自变量 x 的平方项、偏移量 logN 和随机截距；而与观测的 SMR 吻合度最好的是基于模型 4 计算得到的预测的 SMR，该模型中只包含自变量 x 的一次项、偏移量 logN 和随机截距。

4 讨论与小结

4.1 讨论

在构建泊松分布回归模型时，选取适当的偏移量，对提高模型的拟合效果至关重要；此外，当不同受试对象的计数观测结果之间的变异度较大时，在回归模型中引入随机截距是很有必要的；在调用 proc bglimm 过程构建泊松分布回归模型时，在过程语句中增加选项 DIC，可以输出 4 种评价模型对资料拟合效果的偏差信息统计量的计算结果。

4.2 小结

本文介绍了与泊松分布回归模型有关的 6 个基本概念，介绍了泊松分布回归参数估计方法和 DIC 的计算方法；针对一个临床调查实例，拟合了 5 个不同的泊松分布回归模型；还展示了预测的 SMR 与观测的 SMR 之间的吻合程度。

参考文献

[1] 茆诗松. 统计手册[M]. 北京：科学出版社，2003：120-121，1004-1007.
Mao SS. Statistical manual[M]. Beijing：Science Press，2003：120-121，1004-1007.

[2] 胡良平. 如何正确运用Z检验：两Poisson均值比较一般差异性Z检验及SAS实现[J]. 四川精神卫生，2020，33(5)：427-430.
Hu LP. How to use Z test correctly：comparison of two Poisson mean values for the general difference Z test and the SAS implementation[J]. Sichuan Mental Health，2020，33(5)：427-430.

[3] 陈希孺. 广义线性模型的拟似然法[M]. 合肥：中国科学技术大学出版社，2011：43-130.
Chen XR. Quasi-likelihood method for generalized linear model[M]. Hefei：University of Science and Technology of China Press，2011：43-130.

[4] Littell RC，Milliken GA，Stroup WW，et al. SAS system for mixed models[M]. Cary，NC：SAS Institute Inc，1996：423-460.

[5] SAS Institute Inc. SAS/STAT®15.1 user's guide[M]. Cary，NC：SAS Institute Inc，2018：129-166，1205-1306，6533-6728.

[6] Breslow NE，Clayton DG. Approximate Inference in generalized linear mixed models[J]. J Am Stat Assoc，1993，88(421)：9-25.

[7] 刘金山，夏强. 基于MCMC算法的贝叶斯统计方法[M]. 北京：科学出版社，2016：4-117.
Liu JS，Xia Q. Bayesian statistical method based on MCMC algorithm[M]. Beijing：Science Press，2016：4-117.

[8] 康崇禄. 蒙特卡罗方法理论和应用[M]. 北京：科学出版社，2015：86-149.
Kang CL. Monte Carlo method theory and application[M]. Beijing：Science Press，2015：86-149.

[9] Spiegelhalter DJ，Best NG，Carlin BP，et al. Bayesian measures of model complexity and fit[J]. J R Stat Soc Series B Stat Methodol，2002，64(4)：583-616.

[10] Clayton D，Kaldor J. Empirical Bayes estimates of age-standardized relative risks for use in disease mapping[J]. Biometrics，1987，43(3)：671-681.

[11] 方开泰，许建伦. 统计分布[M]. 北京：科学出版社，1987：81-90.
Fang KT，Xu JL. Statistical distribution[M]. Beijing：Science Press，1987：81-90.

（收稿日期：2023-02-01）
（本文编辑：陈 霞）