第 32 卷 第 2 期 华侨大学学报(自然科学版) Vol. 32 No. 2
2011 年 3 月 Journal of Huaqiao University (Natural Science) Mar. 2011

文章编号: 1000-5013(2011)02-0226-05

负二项分布的广义线性模型及其应用

陈卓恒

(华侨大学 数学科学学院, 福建 泉州 362021)

摘要: 讨论一类散度偏大的分布负二项分布的相关性质,以服从负二项分布的索赔次数为响应变量,引入风险分级变量和对数联结函数,建立广义线性模型.采用极大似然估计法进行参数估计,并用 Wald 检验法进行检验.最后,利用 SAS 软件包对一组保险索赔数据进行实证分析.

关键词: 负二项分布; 广义线性模型; Wald 检验; 风险分级

中图分类号: O 212; F 84 **文献标志码:** A

在风险理论中,总索赔次数的分布的研究一直是个中心论题.一般情况下,常采用均值等于方差的 Poisson 分布来描述索赔次数分布,但这与实际情况是不符的.事实上,索赔次数的分布规律往往偏离实际出事故次数的分布规律.鉴于此,本文引入方差大于均值的分布负二项(NB)分布,并在此基础上建立相应的广义线性模型.

1 负二项分布的性质

对于风险非同质性保单组合而言,索赔次数往往可用混合 Poisson 分布来拟合 . 即索赔次数满足

$$P(N=y)=\int_a^b \frac{\lambda^y}{y\,!}\exp(-\lambda)u(\lambda)\mathrm{d}\lambda, \qquad y=0,1,2,\cdots.$$

其中:$u(\lambda)$是某个区间$[a,b]$上某连续分布的密度函数,$0\leqslant a<b\leqslant+\infty$.若取$u(\lambda)$为参数$\alpha$和$\beta$的 Gamma$(\alpha,\beta)$分布,那么有

$$P(N=y)=\int_0^{+\infty}\frac{\lambda^y}{y\,!}\exp(-\lambda)\frac{\beta^\alpha\exp(-\beta\lambda)\lambda^{\alpha-1}}{\tau(\alpha)}\mathrm{d}\lambda=$$

$$\frac{\beta^\alpha}{y\,!\tau(\alpha)(1+\beta)^{y+\alpha}}\int_0^{+\infty}\exp(-\lambda(1+\beta))[\lambda(1+\beta)]^{y+\alpha-1}\mathrm{d}[\lambda(1+\beta)]=$$

$$\frac{\tau(y+\alpha)\cdot\beta^\alpha}{y\,!\tau(\alpha)\cdot(1+\beta)^{y+\alpha}}=\frac{\tau(y+\alpha)}{\tau(\alpha)\cdot y\,!}\cdot\left(\frac{\beta}{1+\beta}\right)^\alpha\cdot\left(\frac{1}{1+\beta}\right)^y, \qquad y=0,1,2,\cdots,$$

此时的索赔次数 N 服从负二项分布.若记$\alpha=r$,$\beta=r/\mu$,则带参数 r 和$\mu(r>0)$的负二项分布的概率函数又可表示为

$$P(N=y\mid\mu,r)=\frac{\tau(y+r)}{\tau(r)\cdot y\,!}\left(\frac{r}{\mu+r}\right)^r\left(\frac{\mu}{\mu+r}\right)^y, \qquad y=0,1,2,\cdots. \tag{1}$$

由负二项分布的性质,易得

$$E(N\mid\mu,r)=\mu, \qquad \mathrm{Var}(N\mid\mu,r)=\mu\left(1+\frac{\mu}{r}\right). \tag{2}$$

由于负二项分布的散度参数$\Phi=\dfrac{\mathrm{Var}(N)}{E(N)}=1+\dfrac{\mu}{r}>1$,因此负二项分布是个散度偏大的分布.若令$r\to\infty$,那么负二项分布 NB$(\mu,r)$收敛于参数为$\mu$的 Poisson 分布[1-2].

收稿日期: 2009-04-14

通信作者: 陈卓恒(1980-),女,讲师,主要从事金融统计和保险精算方向的研究. E-mail:ranic@163.com.

基金项目: 华侨大学科研基金资助项目(07HZR04)

2　广义线性模型

2.1　模型的建立

广义线性模型是由 Nelder 提出的，十分适合离散的，厚尾的保险数据. 它对于传统线性模型有以下 3 个方面的推广 .

(1) 响应变量 Y 的分布，可以取自于指数型分布族中的任一种分布 .

(2) 自变量的线性组合为 $\eta=\beta_1 x_1+\cdots+\beta_k x_k=\boldsymbol{X}'\boldsymbol{\beta}$. 这与多元线性回归模型没有什么区别，$Y,X$ 可取连续或离散值，但在应用上更多的是取离散值 .

(3) 响应变量的均值 $E(Y)=\mu=h(\boldsymbol{X}'\boldsymbol{\beta})$，$h$ 单调且可导，其反函数 $g=h^{-1}$ 称为联结函数.

设某险种的保单按其属性分为 n 类风险组，Y_i 表示第 i 类保单的索赔次数，且 Y_i 服从于 $\mathrm{NB}(\mu_i,r)$，$i=1,2,\cdots,n$，Y_i 之间相互独立. 这里的 r 可看成冗余参数，在各次观察中不变. $\boldsymbol{X}$ 为风险分级变量，采用对数联结函数 $g(x)=\log(x)$ 建立广义线性模型，有

$$E(Y_i)=\mu_i=\exp(\boldsymbol{X}_i{}'\boldsymbol{\beta}),\qquad i=1,2,\cdots,n, \tag{3}$$

式中：$\boldsymbol{X}_i=(x_{i,1},x_{i,2},\cdots,x_{i,k})'$，$\boldsymbol{\beta}$ 为 k 维待估参数.

2.2　极大似然估计

考虑用极大似然估计法进行参数估计. 对于独立样本 $(\boldsymbol{X}_i,Y_i{}')$，$i=1,2,\cdots,n$，把 Y 的分布写成指数标准型，有

$$P(Y_i=y_i\mid\mu_i,r)=\frac{\tau(y_i+r)}{\tau(r)\cdot y_i!}\exp\left(y_i\log(\frac{\mu_i}{\mu_i+r})+r\log(\frac{r}{\mu_i+r})\right). \tag{4}$$

若令 $\theta_i=\log(\frac{\mu_i}{\mu_i+r})$，式(4)又可表示为

$$P(Y_i=y_i\mid\theta_i,r)=\frac{\tau(y_i+r)}{\tau(r)\cdot y_i!}\exp\{y_i\theta_i-r\log(\frac{1}{1-\exp(\theta_i)})\},\qquad i=0,1,2,\cdots,n. \tag{5}$$

由于一维指数型分布的分布密度(概率函数)的标准形式为

$$f(y;\theta)=c(y)\exp(\theta y-b(\theta)), \tag{5}$$

将式(5)对应于式(6)，即有

$$c(y)=\frac{\tau(y+r)}{\tau(r)\cdot y!},\qquad b(\theta)=r\log\left(\frac{1}{1-\exp(\theta)}\right). \tag{7}$$

由于 $\mu_i=\exp(\boldsymbol{X}_i{}'\boldsymbol{\beta})$，因此式(5)中的 θ_i 与 $\boldsymbol{X}_i$ 有关，与参数 $\boldsymbol{\beta}$ 也有关. 似然函数为

$$L(\theta_i,r)=\prod_{i=1}^{n}\frac{\tau(y_i+r)}{\tau(r)\cdot y_i!}\exp\left\{y_i\theta_i-r\log\left(\frac{1}{1-\exp(\theta_i)}\right)\right\}; \tag{8}$$

而对数似然为

$$\log L=\sum_{i=1}^{n}\log\left(\frac{\tau(y_i+r)}{\tau(r)\cdot y_i!}\right)+\sum_{i=1}^{n}(y_i\theta_i-r\log\left(\frac{1}{1-\exp(\theta_i)}\right)), \tag{9}$$

记 $l_i(\theta_i)=y_i\theta_i-r\log\left(\frac{1}{1-\exp(\theta_i)}\right)$，即要找 $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}$，使得 $l(\hat{\boldsymbol{\beta}})=\sum\limits_{i=1}^{n}l_i(\theta_i(\boldsymbol{\beta}))$ 最大即可. 令似然方程

$$\frac{\partial l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{n}(y_i-r\frac{\exp(\theta_i)}{1-\exp(\theta_i)})\cdot\frac{\partial\theta_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=0. \tag{10}$$

由式(3)可得

$$\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}=\exp\{\boldsymbol{X}_i{}'\boldsymbol{\beta}\}\cdot X_i. \tag{11}$$

由指数标准型分布的常用结论，易知

$$\mu_i=\tilde{b}(\theta_i)=r\frac{\exp(\theta_i)}{1-\exp(\theta_i)},\qquad\frac{\mathrm{d}\mu_i}{\mathrm{d}\theta_i}=\tilde{\tilde{b}}(\theta_i)=\mathrm{Var}(Y_i)\triangleq\sigma_i^2(\boldsymbol{\beta}). \tag{12}$$

其中：$\tilde{b}(\theta_i)$ 和 $\tilde{\tilde{b}}(\theta_i)$ 分别指 $b(\theta_i)$ 对 θ_i 的一阶和二阶导数 . 由式(2)可得

$$\mathrm{Var}(Y_i)=\mu_i(1+\frac{\mu_i}{r})=\exp(\boldsymbol{X}_i{}'\boldsymbol{\beta})\left(1+\frac{\exp(\boldsymbol{X}_i{}'\boldsymbol{\beta})}{r}\right). \tag{13}$$

于是,由式(11),(12),(13)可得

$$\frac{\partial\theta_i}{\partial\boldsymbol{\beta}}=\frac{\mathrm{d}\theta_i}{\mathrm{d}\mu_i}\cdot\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}=\exp(-\boldsymbol{X}_i{}'\boldsymbol{\beta})\cdot\frac{r}{r+\exp(\boldsymbol{X}_i{}'\boldsymbol{\beta})}\cdot\exp(\boldsymbol{X}_i{}'\boldsymbol{\beta})\cdot\boldsymbol{X}_i=\frac{r\cdot\boldsymbol{X}_i}{r+\exp(\boldsymbol{X}_i{}'\boldsymbol{\beta})}. \tag{14}$$

因此,似然方程(10)又可写成

$$\sum_{i=1}^{n}(y_i-\mu_i(\boldsymbol{\beta}))\cdot\frac{r\cdot\boldsymbol{X}_i}{r+\exp(\boldsymbol{X}_i{}'\boldsymbol{\beta})}=0. \tag{15}$$

记 $s_i(\boldsymbol{\beta})=(y_i-\mu_i(\boldsymbol{\beta}))\cdot\dfrac{r\cdot\boldsymbol{X}_i}{r+\exp(\boldsymbol{X}_i{}'\boldsymbol{\beta})}$. 那么,式(15)又可表示为

$$s(\boldsymbol{\beta})\triangleq\sum_{i=1}^{n}s_i(\boldsymbol{\beta})=0. \tag{16}$$

一般来说,在一定条件下,满足似然方程 $\sum\limits_{i=1}^{n}s_i(\boldsymbol{\beta})=0$ 的极大似然解(MLE)$\hat{\boldsymbol{\beta}}$ 是存在且唯一的[3].

2.3　MLE的迭代计算

通常通过求方程 $\sum\limits_{i=1}^{n}s_i(\boldsymbol{\beta})=0$ 的MLE,得到 $\boldsymbol{\beta}$ 的MLE $\hat{\boldsymbol{\beta}}$,其具体计算采用牛顿法. 若以 $J_n\triangleq\sum\limits_{i=1}^{n}s_i(\boldsymbol{\beta}^0)$($\boldsymbol{\beta}^0$ 表示 $\boldsymbol{\beta}$ 的真值),那么,$E(J_n)=0$,其协方差阵为

$$\Lambda_n\triangleq\mathrm{cov}(J_n)=\sum_{i=1}^{n}\mathrm{cov}(s_i(\boldsymbol{\beta}^0))=\sum_{i=1}^{n}D_i^2(\boldsymbol{\beta}^0)\sigma_i^2(\boldsymbol{\beta}^0)\boldsymbol{X}_i\boldsymbol{X}_i{}', \tag{17}$$

其中:$D_i(\boldsymbol{\beta})=\dfrac{\partial h(\boldsymbol{X}_i{}'\boldsymbol{\beta})}{\partial(\boldsymbol{X}_i{}'\boldsymbol{\beta})}=\exp\{\boldsymbol{X}_i{}'\boldsymbol{\beta}\}$. 由式(12),(13)得

$$\sigma_i^2(\boldsymbol{\beta})=\mathrm{Var}(Y_i)=\exp(\boldsymbol{X}_i{}'\boldsymbol{\beta})+\frac{1}{r}\exp(2\boldsymbol{X}_i{}'\boldsymbol{\beta}).$$

记

$$H(\boldsymbol{\beta})=\sum_{i=1}^{n}D_i^2(\boldsymbol{\beta})\sigma_i^{-2}(\boldsymbol{\beta})\boldsymbol{X}_i\boldsymbol{X}_i{}'=\sum_{i=1}^{n}\frac{\mathrm{e}\boldsymbol{X}p(\boldsymbol{X}_i{}'\boldsymbol{\beta})\cdot r}{r+\exp(\boldsymbol{X}_i{}'\boldsymbol{\beta})}\boldsymbol{X}_i\boldsymbol{X}_i{}', \tag{18}$$

则从初始值 $\hat{\boldsymbol{\beta}}^{(0)}$ 开始,第 k 步算到 $\hat{\boldsymbol{\beta}}^{(k)}$,并采用迭代式

$$\hat{\boldsymbol{\beta}}^{(k+1)}=\hat{\boldsymbol{\beta}}^{(k)}+H^{-1}(\hat{\boldsymbol{\beta}}^{(k)})s(\hat{\boldsymbol{\beta}}^{(k)}). \tag{19}$$

式中:$H^{-1}(\boldsymbol{\beta})$表示式(18)的 $H(\boldsymbol{\beta})$逆矩阵. 对于设定的 ε,当进行到 $\|\hat{\boldsymbol{\beta}}^{(k+1)}-\hat{\boldsymbol{\beta}}^{(k)}\|/\|\hat{\boldsymbol{\beta}}^{(k)}\|<\varepsilon$ 时,即停止迭代. 初始条件 $\hat{\boldsymbol{\beta}}^{(0)}$ 可取为$\{(g(y_i),\boldsymbol{X}_i),1\leqslant i\leqslant n\}$下线性回归系数的LS估计,即

$$\hat{\boldsymbol{\beta}}^{(0)}=(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i{}')^{-1}\sum_{i=1}^{n}\boldsymbol{X}_iy_i.$$

式(19)最终可化为一个加权最小二乘估计的形式,并通过采用SAS软件包中线性回归的程序实现.

2.4　假设检验

检验部分采用Wald检验. 由于模型的选择和解释变量的显著性检验问题可化为线性假设检验,原假设 H_0:$\boldsymbol{C}\boldsymbol{\beta}=\xi$,备则假设 H_1:$\boldsymbol{C}\boldsymbol{\beta}\neq\xi$. 其中,$\boldsymbol{C}$ 为 $s\times k$ 的行满秩矩阵,$\boldsymbol{\beta}$ 为 k 维向量.

引入Wald统计量,即 $W=(\boldsymbol{C}\hat{\boldsymbol{\beta}}-\xi)'[\boldsymbol{C}H^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C}']^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}}-\xi)$. 这里的 $\hat{\boldsymbol{\beta}}$ 表示 β 的MLE,当原假设成立时,W^d 服从于 $\chi^2(s)$,即统计量 W 渐进服从自由度为 s 的 χ^2 分布. 所以,对于给定检验水平 $\alpha(0<\alpha<1)$,若 $W>\chi^2_\alpha(s)$,则拒绝原假设 H_0;否则,接受 H_0.

3　实例分析

数据来源于某保险公司关于机动车辆的保险索赔资料[4-5],如表1所示. 考虑有如下3种风险因素影响着索赔额(N)和索赔次数(Y):(1) 保单持有者的年龄(PA). 有17~20岁,21~24岁,25~29岁,30~34岁,35~39岁,40~49岁,50~59岁,60岁以上8个水平,分别用 $\alpha_1\sim\alpha_8$ 表示;(2) 车型(CG). 有 A,B,C,D 共4个水平,分别用 $\beta_1\sim\beta_4$ 表示;(3) 车龄(VA). 有0~3 a,4~7 a,8~9 a,10 a以上4个水平,分别用 $\gamma_1\sim\gamma_4$ 表示.

按照这3种风险因素,可以将保单持有人分为128个风险单元,用 Y_i 表示第 i 个风险单元的索赔次数. 设 Y_i 服从于 $\mathrm{NB}(\mu_i,r)$,$i=1,2,\cdots,128$. Y_i 之间相互独立,采用对数联结函数建立的广义线性模

型为

$$E(Y_i) = \mu_i = \exp(\boldsymbol{X}_i{}'\boldsymbol{\beta}), \qquad i = 1,2,\cdots,128. \tag{20}$$

式中：$\boldsymbol{X}_i=(x_{i,1},x_{i,2},\cdots,x_{i,18})'$；$\boldsymbol{\beta}=(\mu_0;\mu_1;\alpha_1,\cdots,\alpha_8;\beta_1,\cdots,\beta_4;\gamma_1,\cdots,\gamma_4)'$；$\mu_0$ 为截距；μ_1 为索赔额的对数值. 采用 SAS 软件的 GENMOD 过程编程计算，所得结果算法收敛.

表 1 3 种风险因素影响下的索赔额和索赔次数

Tab. 1 Number and amount of claims in three risk factors

PA	CG	VA							
		γ_1		γ_2		γ_3		γ_4	
		N/欧元	Y/次	N/欧元	Y/次	N/欧元	Y/次	N/欧元	Y/次
α_1	β_1	289	8	282	8	133	4	160	1
	β_2	372	10	249	28	288	1	11	1
	β_3	189	9	288	13	179	1	—	0
	β_4	763	3	850	2	—	0	—	0
α_2	β_1	302	18	194	31	135	10	166	4
	β_2	420	59	243	96	196	13	135	3
	β_3	268	44	343	39	193	7	104	2
	β_4	407	24	320	18	205	2	—	0
α_3	β_1	268	56	285	55	181	17	110	12
	β_2	275	125	243	172	179	36	264	10
	β_3	334	163	274	129	208	18	150	8
	β_4	383	72	305	50	116	6	636	1
α_4	β_1	236	43	270	53	160	15	110	12
	β_2	259	179	226	211	161	39	107	19
	β_3	340	197	260	125	189	30	104	9
	β_4	400	104	349	55	147	8	65	2
α_5	β_1	207	43	129	73	157	21	113	14
	β_2	208	191	214	219	149	46	137	23
	β_3	251	210	232	131	204	32	141	8
	β_4	233	119	325	43	207	4	—	0
α_6	β_1	254	90	213	98	149	35	98	22
	β_2	218	380	209	434	172	97	110	59
	β_3	239	401	250	253	174	50	129	15
	β_4	387	199	299	88	325	8	137	9
α_7	β_1	251	69	227	120	172	42	98	35
	β_2	196	366	229	353	164	95	132	45
	β_3	268	310	250	148	175	33	152	13
	β_4	391	105	228	46	346	10	167	1
α_8	β_1	264	64	198	100	167	43	114	53
	β_2	224	228	193	233	178	73	101	44
	β_3	269	183	258	103	227	20	119	6
	β_4	385	62	324	22	192	6	123	6

GENMOD 过程的第 3 型分析表明，PA，CG，VA 的自由度分别是 7，3，3，卡方值分别为 150.07，107.64，119.32，而其显著性水平 p 值均小于 0.000 1. 由此可知，在决定投保人的索赔行为时，3 种因素都是非常显著的.

从评价拟合优度的标准可知，3 种因素自由度都为 109，但 NB 分布相对于 Poisson 分布具有更小的离差和平均离差(NB 分布分别为 131.732 7，1.208 6，而 Poisson 分布分别为 1 107.793 5，10.163 2). 因此，与 Poisson 分布相比，用 NB 分布来拟合数据，其拟合程度更好. 参数估计的结果，如表 2 所示.

从表 2 可以看出，从保单持有者年龄来说，年龄在 40～49 岁的人风险最大；而处于 17～20 岁年龄阶段的人风险最小. 这里的风险是相对于索赔次数而言. 由此可以说明，17～20 岁年龄阶段的投保人索赔频率不高，当然，索赔额则不一定了. 从车型来看，A，B，C，D 等 4 种车型中，车型 B 风险最大，而车

型 D 风险最小. 从车龄来看,0～3 a 的车风险最大,而 10 a 以上的车风险最小. 分析原因应该是与新车驾驶员的车技和心理等因素有关,而老车驾驶员相对更重视安全. 此外,从表中的 Wald 卡方统计量和显著性水平 p 值可以看出,所检验的参数对于模型基本都是显著的.

表 2 参数估计的结果

Tab. 2 Analysis of parameter estimates

参数	分类水平	自由度	估计值	标准差	卡方值	p
μ_0	—	1	−2.900 3	0.181 3	255.87	<0.000 1
PA	α_1	1	−2.999 6	0.225 0	177.69	<0.000 1
PA	α_2	1	−1.545 6	0.188 8	67.04	<0.000 1
PA	α_3	1	−0.611 3	0.177 5	11.86	0.000 6
PA	α_4	1	−0.289 0	0.177 3	2.66	0.103 1
PA	α_5	1	−0.116 7	0.178 5	0.43	0.513 2
PA	α_6	1	0.407 7	0.173 0	5.55	0.018 4
PA	α_7	1	0.145 5	0.173 4	0.70	0.401 4
PA	α_8	0	0	0	—	—
CG	β_1	1	1.083 8	0.164 2	43.54	<0.000 1
CG	β_2	1	1.815 7	0.135 0	180.79	<0.000 1
CG	β_3	1	1.152 2	0.132 9	75.15	<0.000 1
CG	β_4	0	0	0	—	—
VA	γ_1	1	1.348 9	0.130 7	106.59	<0.000 1
VA	γ_2	1	1.248 5	0.129 4	93.15	<0.000 1
VA	γ_3	0	0	0	—	—
VA	γ_4	1	−0.236 0	0.144 3	2.68	0.101 9
μ_1	—	1	0.205 3	0.039 6	—	—

参考文献:

[1] SUSANNE G, CLAUDIA C. Modelling count data with overdispersion and spatial effects[J]. Statistical Papers, 2008, 49(3): 531-552.

[2] 田霆,刘次华. 定时截尾缺失数据下指数分布的参数 AMLE[J]. 华侨大学学报:自然科学版,2006,27(4):351-353.

[3] FAHRMEIR L, THTZ G. Multivariate statistical modelling based on generalized linear models[M]. 2nd ed. New York: Springer-Verlag, 1996.

[4] 毛泽春,刘锦萼. 一类索赔次数的回归模型及其在风险分级中的应用[J]. 应用概率统计,2004,20(4):359-367.

[5] MCCULLAGH P, NELDER J A. Generalized linear models[M]. 2nd ed. London: Chapman and Hall, 1989.

Generalized Linear Model Based on Negative Binomial Distribution and Its Application

CHEN Zhuo-heng

(School of Mathematical Sciences, Huaqiao University, Quanzhou 362021, China)

Abstract: The properties of the negative binomial distribution which is over-dispersion is discussed in the paper. A generalized linear model which based on the distribution is intruduced. The maximum likelihood estimates and wald test for the model are considered. At last the model is applied to a real data set of aggregate claims for automobile insurance using SAS package.

Keywords: negative binomial distribution; generalized linear model; wald test; risk classification

(**责任编辑**: 陈志贤 **英文审校**: 张金顺,黄心中)