有限混合模型回归分析

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】 本文目的是介绍有限混合模型回归分析的概念、作用以及如何用软件实现计算的方法。先介绍有关的基本概念,再介绍基本原理,最后通过一个实例并基于 SAS 软件演示如何实施有限混合模型回归分析。结果表明:有限混合模型回归分析最适用于"具有两个或有限的多个频数分布资料进行频数分布曲线拟合"的场合。

【关键词】 有限混合模型;回归分析;概率密度函数;频数分布曲线
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.04.004

Finite mixture model regression analysis

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 concepts and functions and the calculation methods by using the statistical software of the finite mixture model regression analysis. Firstly, the basic concepts of the regression analysis were introduced. Secondly, the basic principles of the regression analysis were given. Finally, the finite mixture model regression analysis was demonstrated through one example by using the SAS software. The results showed that it was suitable for fitting the frequency distributional curve for the data with two or multiple samples.

【**Keywords**】 Finite mixture model; Regression analysis; Probability density function; Frequency distributional curve

1 概 述

1.1 有限混合模型

有限混合模型[1]见式(1):

$$f(y;\alpha,\beta)=\sum_{j=1}^{k}P_r(S=j)\,f(y;\alpha_j,\beta_j\mid S=j)=\sum_{j=1}^{k}\pi_j\,f(y;\alpha_j,\beta_j\mid S=j) \quad (1)$$

在式(1)中,假定 y 是一个可观测的随机变量,其分布取决于一个不可观测的隐变量 S,它有 k 种可能的状态;k 的取值是未知的,但至少是有限的。

令 π_j 表示 S 取值 j 的概率,在 $S=j$ 的条件下,假定反应变量 Y 的分布是 $f_j(y;\alpha_j,\beta_j\mid S=j)$(注:它是一个概率密度函数)。

1.2 有限混合模型回归分析应用的场合

根据研究目的确定了一个具有同质性的研究总体,若从该总体中随机抽取样本含量为 n 的个体,从每个个体身上测量某计量指标的数值,依此法收集到的 n 个计量数据被称为"单组设计一元计量资料"。描述这组计量数据的方法有如下几种:第一,编制频数分布表;第二,绘制直方图;第三,拟合频数分布曲线。

当直方图显示只有一个高峰时,就称其为"单峰分布"。此时,可以根据高峰所处的位置,分为"正偏态分布(高峰位于左侧)""对称分布(高峰居中,其特例为正态分布)"和"负偏态分布(高峰位于右侧)"。此时,若希望拟合频数分布曲线,可以利用 SAS 中 CAPABILITY 过程,见文献[2]。

当直方图显示有两个或多个峰时,就称其为"多峰分布"。它们很可能是由多个"单峰分布"叠加而形成,被称为"有限混合分布样本"。在实际问题中,这种混合分布样本很可能来自一个"不同质"的总体,例如正常人样本、某种疾病不同严重程度(轻、中、重度)的患者样本。若希望拟合频数分布曲线,可以利用 SAS 中 FMM 过程,见文献[1]。

1.3 有限混合模型回归分析的计算原理

1.3.1 概率分布

对于样本资料而言,描述单组设计一元计量资料的频数分布情况,所采用的方法被称为"拟合频数分布曲线";而对于总体资料而言,描述其一元连

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

续性变量的变化规律，所采用的方法被称为“呈现概率分布密度函数”。在运用统计学解决实际问题时，通常都是通过样本信息去推论总体的规律，故通常都是以“拟合频数分布曲线”取代“呈现概率分布密度函数”。换言之，就是以“精确的概率分布密度函数”作为理论依据，描述“频数分布曲线”的变化规律。

1.3.2 计算原理

式(1)描述的“频数分布曲线”仅适用于“单组设计一元计量资料”，然而，当资料中还有影响计量结果变量的自变量 z 和 x 时，需要将式(1)修改成式(2)的形式：

$$f(y) = \sum_{j=1}^{k} \pi_j(z,\alpha_j)\, p_j(y; x'\beta_j, \phi_j) \qquad (2)$$

在式(2)中，有 k 个样本混合在一起，k 值需要结合专业知识事先给定；求和符号之后的第 1 项为样本来自第 j 个总体的概率，其自变量为 z；而其后的那一项为第 j 个样本的“频数分布(对样本而言)”或“概率密度函数(对总体而言)”，其自变量为 x。式(2)中求和符号之后的第 1 项还应满足下式要求，见式(3)：

$$\sum_{j=1}^{k} \pi_j(z,\alpha_j) = 1 \qquad (3)$$

在式(3)中，$\pi_j \geqslant 0$，对于所有的 j($j=1$ 到 k)都成立。

要想在式(3)的约束条件下求解式(2)，涉及到“一般混合模型”的求解理论和技术方法，涉及到基于多种常用概率密度函数构造“对数似然函数”并求解，还涉及到贝叶斯统计模拟算法[1,3-4]等深入的内容，因篇幅所限，此处从略。

2 基于有限混合模型回归分析解决实际问题

2.1 问题与数据结构

【例 1】下面是一个关于牛的喂食时间间隔的数据(计量数据)。按两种情形对时间进行划分：第 1 种情形：分为 3 类不同的时间段(相当于前面所说的混合样本个数 k=3)，即：①各次进食之间的时间间隔很短；②各次进食之间的时间间隔稍长一点，间隙让牛饮水；③每两次进食之间的时间间隔很长。第 2 种情形：在上面的 3 类时间间隔组中的每一种时间间隔内，不同牛的进食时间是不尽相同的。

测量 141 414 头牛中每头牛每两次进食之间的“时间间隔数据(int)”，再取其对数，记为“logint”。以其为“计量结果变量(特别说明：选取这样的指标作为结果变量，只有在特定专业领域内才有意义，在一般的实际问题中，‘时间间隔’是不可能用作结果变量的)”，由于原始数据的精确度很高，保留其精确度为“0.05”，这就导致了相同的数据很多，于是，可用类似“频数表”简化地呈现原始数据，其数据格式见表 1。

表 1 141 414 头牛中每头牛每两次进食之间的“时间间隔数据(logint)”

logint	f	logint	f	logint	f	logint	f
0.70	195	1.10	233	1.40	355	1.60	563
1.80	822	1.95	926	2.10	1018	2.20	1712
2.30	3190	2.40	2212	2.50	1692	2.55	1558

注：表中仅列出了极少部分数据，详细数据见后面 SAS 程序

【对数据结构的分析】以上采用“频数分布表”形式呈现了资料，而资料的原始形式很简单，即一个计量变量“logint”，它有 141 414 个取值。通常这样的数据被称为“单组设计一元计量资料”；而在本例中，因所有数据分别属于“3 类不同的时间段”。也就是说，若用一个变量 k 代表“3 类不同的时间段”，将此变量“k”及其具体取值也体现在每个“logint”数据之前，相当于多了一个“分组变量”。此时，全部数据就可被视为“单因素三水平设计一元计量资料”了。

【统计分析方法的选择】若希望采用“单因素三水平设计一元计量资料”方差分析处理此资料，就必须在资料中全面反映出变量 k 及其取值；而在本例中，统计分析目的是“拟合频数分布曲线”，就只需要告知有“k 个样本”(注意：k 必须是一个具体的正整数)，并且还需要告知这 k 个样本所代表的“分布”分别是什么。与特定分布对应的“参数”可以告知，也可以不告知。例如“dist = normal k = 2 parms (3 1, 5 1)”，这就是告知：有 2 个正态分布的样本，它们的均值分别为 3 和 5、方差分别为 1 和 1；又例如“dist = normal k = 2”，这就是告知：有 2 个正态分布的样本，它们的均值和方差都没有指定，由统计软件根据实际数据去估算。

2.2 创建 SAS 数据集

创建一个名为“cattle”的临时 SAS 数据集的 SAS 数据步程序:

```
data cattle;
input LogInt Count @@;
datalines;
```

0.70	195	1.10	233	1.40	355	1.60	563
1.80	822	1.95	926	2.10	1018	2.20	1712
2.30	3190	2.40	2212	2.50	1692	2.55	1558
2.65	1622	2.70	1637	2.75	1568	2.85	1599
2.90	1575	2.95	1526	3.00	1537	3.05	1561
3.10	1555	3.15	1427	3.20	2852	3.25	1396
3.30	1343	3.35	2473	3.40	1310	3.45	2453
3.50	1168	3.55	2300	3.60	2174	3.65	2050
3.70	1926	3.75	1849	3.80	1687	3.85	2416
3.90	1449	3.95	2095	4.00	1278	4.05	1864
4.10	1672	4.15	2104	4.20	1443	4.25	1341
4.30	1685	4.35	1445	4.40	1369	4.45	1284
4.50	1523	4.55	1367	4.60	1027	4.65	1491
4.70	1057	4.75	1155	4.80	1095	4.85	1019
4.90	1158	4.95	1088	5.00	1075	5.05	912
5.10	1073	5.15	803	5.20	924	5.25	916
5.30	784	5.35	751	5.40	766	5.45	833
5.50	748	5.55	725	5.60	674	5.65	690
5.70	659	5.75	695	5.80	529	5.85	639
5.90	580	5.95	557	6.00	524	6.05	473
6.10	538	6.15	444	6.20	456	6.25	453
6.30	374	6.35	406	6.40	409	6.45	371
6.50	320	6.55	334	6.60	353	6.65	305
6.70	302	6.75	301	6.80	263	6.85	218
6.90	255	6.95	240	7.00	219	7.05	202
7.10	192	7.15	180	7.20	162	7.25	126
7.30	148	7.35	173	7.40	142	7.45	163
7.50	152	7.55	149	7.60	139	7.65	161
7.70	174	7.75	179	7.80	188	7.85	239
7.90	225	7.95	213	8.00	235	8.05	256
8.10	272	8.15	290	8.20	320	8.25	355
8.30	307	8.35	311	8.40	317	8.45	335
8.50	369	8.55	365	8.60	365	8.65	396
8.70	419	8.75	467	8.80	468	8.85	515
8.90	558	8.95	623	9.00	712	9.05	716
9.10	829	9.15	803	9.20	834	9.25	856
9.30	838	9.35	842	9.40	826	9.45	834
9.50	798	9.55	801	9.60	780	9.65	849
9.70	779	9.75	737	9.80	683	9.85	686
9.90	626	9.95	582	10.00	522	10.05	450
10.10	443	10.15	375	10.20	342	10.25	285
10.30	254	10.35	231	10.40	195	10.45	186
10.50	143	10.55	100	10.60	73	10.65	49
10.70	28	10.75	36	10.80	16	10.85	9
10.90	5	10.95	6	11.00	4	11.05	1
11.15	1	11.25	4	11.30	2	11.35	5
11.40	4	11.45	3	11.50	1		

```
;
run;
```

2.3 利用 SAS/STAT 中 KDE 过程绘制资料的频数分布曲线

利用下面的 SAS 过程步程序,可以绘制出反映计量变量 logint 的频数分布直方图和频数分布曲线图。

```
ods graphics on;
proc kde data = cattle;
univar LogInt / bwm = 4;
freq count;
run;
```

【SAS 主要输出结果】

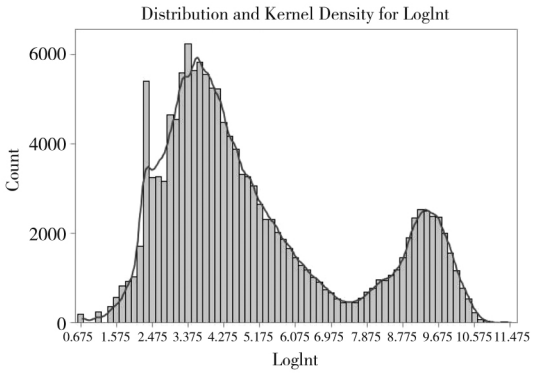

图 1 本例资料的频数分布直方图与频数分布曲线图

图 1 给人的印象是由两个频数分布混合而成的,然而,由专业知识可知,它实际上是由三个频数分布混合而成的。

2.4 利用 SAS/STAT 中 FMM 过程拟合“三分量频数分布曲线”

“三分量频数分布曲线”就是拟合“由三个不同分布样本混合而成的一个总样本”的频数分布曲线。程序将给出三条各自的频数分布曲线以及一条混合的频数分布曲线。

领域专家给出的三个分布分别为:①正态分布, $N(3,1^2)$;**②正态分布,** $N(5,1^2)$;**③威布尔分布。利用下面的 SAS 过程步程序,可以拟合并绘制出反映计量变量 logint 的频数分布曲线图。**

```
proc fmm data = cattle gconv = 0;
model LogInt = / dist = normal k = 2 parms(3 1, 5 1);
  model + / dist = weibull;
  freq count;
run;
```

【SAS 主要输出结果及解释】

Fit Statistics

-2 Log Likelihood	563153
AIC (smaller is better)	563169
AICC (smaller is better)	563169
BIC (smaller is better)	563248
Pearson Statistic	141458
Effective Parameters	8
Effective Components	3

以上是拟合统计量的有关计算结果：前 5 行都是关于拟合效果的评价指标及其取值，这些数值只有在两个或多个模型比较时才有参考的价值。

Parameter Estimates for Normal Model

成分	参数	估计值	标准误差	z 值	Pr > |z|
1	Intercept	3.3415	0.01260	265.16	<.0001
2	Intercept	4.8940	0.05447	89.84	<.0001
1	Variance	0.6718	0.01287		
2	Variance	1.4497	0.05247		

以上是基于“正态分布”假定条件下，计算出两个正态分布对应的“参数估计”结果。因为对于每一个特定的正态分布而言，只要给定了“均值”与“标准差（或方差）”，该正态分布也就唯一被确定了。

第 1 个正态分布为： $N(3.3415, 0.6718) = N(3.3415, 0.8193^2)$；

第 2 个正态分布为： $N(4.8940, 1.4497) = N(4.8940, 1.2040^2)$。

Parameter Estimates for Weibull Model

成分	参数	估计值	标准误差	z 值	Pr > |z|	逆关联估计
3	Intercept	2.2531	0.000506	4452.11	<.0001	9.5174
3	Scale	0.06848	0.000427			

以上是基于“威布尔分布”假定条件下，计算出对应的“参数估计”结果。

第 3 个威布尔分布为： $W(\alpha, \beta, \delta)$，**其中，**$\alpha > 0$ **为形状参数，**$\beta > 0$ **为尺度参数，**$\delta \geqslant 0$ **为位置参数。上面计算的结果为：** $\alpha = \exp(2.2531) = 9.5174$、$\beta = 0.0685$、$\delta = 0$。

Parameter Estimates for Mixing Probabilities

成分	参数	链接尺度				概率
		估计值	标准误差	z 值	Pr > |z|	
1	Probability	0.8106	0.03409	23.78	<.0001	0.4545
2	Probability	0.5305	0.04640	11.43	<.0001	0.3435

以上输出的是各分布在混合分布中出现的概率，第 1 个正态分布出现的概率为 0.4545，第 2 个正态分布出现的概率为 0.3435，而第 3 个威布尔分布出现的概率为 $1 - (0.4545 + 0.3435) = 0.2020$。**于是，就可以写出混合样本的概率密度函数如下：**

$$\hat{y} = 0.4545\ N(3.3415, 0.8193^2) + 0.3435\ N(4.8940, 1.2040^2) + 0.2020\ W(9.5174, 0.0685, 0)$$

说明：上式中的 $\hat{y}$ **代表图 2 中“混合样本”频数曲线上纵坐标的估计值，仅当给定了横坐标上变量的一个确定的取值，**$\hat{y}$ **才有一个具体的数值与其对应，下同，不再赘述。**

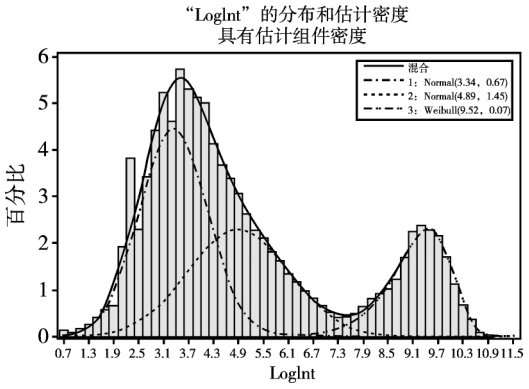

图 2 本例资料的频数分布直方图与拟合的频数分布曲线图之一

若利用下面的 SAS 程序，可以获得与上面类似的结果，但会有较明显的变化：

```
proc fmm data = cattle gconv = 0;
model LogInt = / dist = normal k = 2;
    model + / dist = weibull;
    freq count;
run;
ods graphics off;
```

【SAS 程序说明】这段 SAS 程序与前面那段 SAS 程序非常相似，其主要区别在于：前面指定了两个正态分布的“均值”与“方差”，而现在这段 SAS 程序没有指定参数的具体数值，完全由实际的样本数据计算而得。

【SAS 主要输出结果如下】

Fit Statistics

-2 Log Likelihood	564431
AIC (smaller is better)	564447
AICC (smaller is better)	564447
BIC (smaller is better)	564526
Pearson Statistic	141228
Effective Parameters	8
Effective Components	3

以上是拟合统计量的有关计算结果：前 5 行都是关于拟合效果的评价指标及其取值，与前面相同内容作比较，AIC、AICC 和 BIC 的数值（说明：这些数值越小越好）都变大了，说明现在的模型对资料的拟合效果有所下降。

Parameter Estimates for 'Normal' Model

成分	参数	估计值	标准误差	z 值	Pr > |z|
1	Intercept	9.2883	0.005031	1846.28	<.0001
2	Intercept	4.9106	0.02604	188.56	<.0001
1	Variance	0.4158	0.005086		
2	Variance	1.7410	0.02753		

第 1 个正态分布为：$N(9.2883, 0.4158) = N(9.2883, 0.6448^2)$；

第 2 个正态分布为：$N(4.9106, 1.7410) = N(4.9106, 1.3195^2)$。

Parameter Estimates for 'Weibull' Model

成分	参数	估计值	标准误差	z 值	Pr > |z|	逆关联估计
3	Intercept	1.2908	0.002790	462.71	<.0001	3.6358
3	Scale	0.2093	0.001311			

第 3 个威布尔分布为：$W(\alpha, \beta, \delta)$，其中，$\alpha > 0$ 为形状参数，$\beta > 0$ 为尺度参数，$\delta \geqslant 0$ 为位置参数。

上面计算的结果为：$\alpha = \exp(1.2908) = 3.6358$、$\beta = 0.2093$、$\delta = 0$。

Parameter Estimates for Mixing Probabilities

成分	参数	链接尺度				概率
		估计值	标准误差	z 值	Pr > |z|	
1	Probability	-0.8280	0.01922	-43.08	<.0001	0.1902
2	Probability	-0.1505	0.03678	-4.09	<.0001	0.3745

以上输出的是各分布在混合分布中出现的概率，第 1 个正态分布出现的概率为 0.1902、第 2 个正态分布出现的概率为 0.3745，而第 3 个威布尔分布出现的概率为 $1 - (0.1902 + 0.3745) = 0.4353$。

于是，就可以写出混合样本的概率密度函数如下：

$$\hat{y} = 0.1902\ N(9.2883, 0.6448^2) + 0.3745\ N(4.9106, 1.3195^2) + 0.4353 W(3.6358, 0.2093, 0)$$

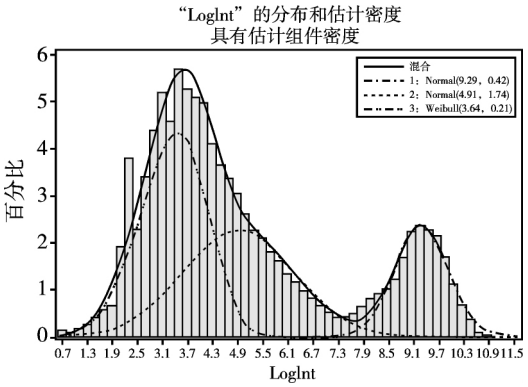

图 3 本例资料的频数分布直方图与拟合的频数分布曲线图之二

文献[5]中有一个关于"1 000 例血清谷丙转氨酶(SGPT)的资料"，请感兴趣的读者运用本文提供的SAS程序对"混杂样本(包括'非肝病患者'与'肝病患者')"进行剖分，并与此文献中基于"G-C级数"剖分的结果进行比较。

参考文献

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

[2] 胡良平. 面向问题的统计学——(1)科研设计与统计基础[M]. 北京: 人民卫生出版社, 2012: 300-311.

[3] 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.

[4] 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, 2017: 4-90.

[5] 杨树勤. 中国医学百科全书医学统计学[M]. 上海: 上海科学技术出版社, 1985: 193-196.

(收稿日期: 2018-08-10)
(本文编辑: 陈 霞)