泊松回归在生育率研究中的应用*

郭志刚　巫锡炜

【摘　要】 泊松回归是专门分析因变量为计数变量的回归模型。文章通过对2001年全国计划生育/生殖健康调查数据的泊松回归分析来介绍其在生育率研究中的应用。泊松回归除了可以接受虚拟编码方式的年龄、城乡等常规人口学分类自变量外，还可以直接接受支出、收入等连续型自变量，因此可以更深入地进行生育率的测量、比较与分析。

【关键词】 泊松回归　计数变量　发生率比　生育率

【作　者】 郭志刚　北京大学中国社会与发展研究中心、北京大学社会学系，教授；巫锡炜　北京大学社会学系，硕士研究生。

一、研究背景和研究目的

虽然生育率的统计指标很多，但年龄别生育率及总和生育率最为重要，应用最普遍。在各种正式公布的统计数据中，一般也都包括年龄别生育率和总和生育率，用来反映全国或各地的生育水平。年龄别生育率是对某一年份某一年龄组妇女生育水平的具体测量指标，总和生育率则是建立在某一年份系列年龄别生育率基础之上的概括性指标。总和生育率表达了时期的生育水平，而该时期的年龄别生育率系列则反映了生育的年龄模式。这些指标的含义简单明了，既容易理解又容易计算。

在实际应用中，总和生育率要比一般生育率的可比性强，是一种更“单纯”的生育水平测量，因而更适用于不同时间和不同地域之间生育水平的比较。就方法论而言，因为总和生育率是系列年龄别相对数（即生育率）的合计，因此已经控制了育龄妇女年龄结构的影响。

换用一般的回归分析建模语言，生育水平是因变量，而年龄则是自变量（或称为协变量、控制变量）。这种基本关系其实与一个回归方程没有什么不同。而就认识生育问题而言，仅仅知道生育率指标的计算、在控制年龄结构的条件下比较生育率差别和变化是远远不够的，还需要进一步对生育率差别和变化做出解释。因为生育本身是一个受到生理、人口、社会、经济、政策和文化观念等诸多因素共同作用的过程，这些因素都直接或间接地对生育发生着影响。

比如，城乡二元结构的社会环境显然对生育率有很大影响，因此我们经常分别计算城乡生育率。这也是统计控制的途径之一，但它的代价是必须先将育龄妇女生育数据划分为城乡两

* 本研究为国家社会科学基金资助课题“人口学方法论研究”（05BRK007）的成果之一。

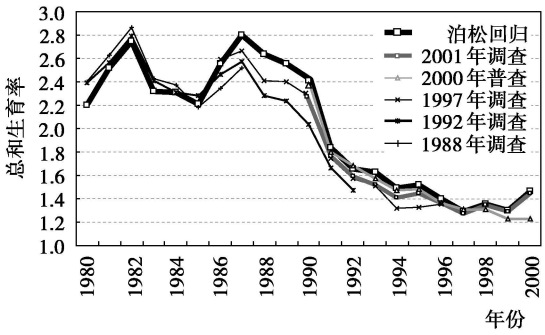

图 **3** 泊松回归估计与正式公布统计及其他来源生育率变化趋势的比较

注：1988 年全国 2‰人口生育节育抽样调查引自于景元、袁建华，1996；1997 年全国人口与生殖健康调查结果引自郭志刚，2000；2000 年全国人口普查结果引自郭志刚，2004b；2001 年全国计划生育与生殖健康调查结果引自潘贵玉等，2003。

从图 3 可以看出，采用泊松回归估计得到的 1980～1986 年的总和生育率要略低于其他来源的统计水平，而 1986 年以后却正好相反，泊松回归估计比其他来源统计偏高。这除了数据来源不同的原因外，还因为泊松回归简单地假定这段时期中生育模式不变，因而在一定程度上脱离了实际。尽管如此，从反映生育率长期变化趋势的角度来看，泊松估计重建的 20 年生育趋势和现有各类统计数据所反映的情况仍然做到了基本吻合。从图 3 还可以看出，由于使用同一数据，泊松回归估计的生育趋势与“2001 年计划生育/生殖健康调查”公布值极为接近，两条线在 1990 年中后期基本上重合了。

（四）对 **2000** 年生育率影响因素的分析

生育行为既会受到生理因素的影响，也会受到社会、经济、文化等方面多因素的影响。下面将应用泊松回归对 2000 年生育率的有关社会变量的影响进行检验。图 4 提供了解释性研究的理论框架。

泊松回归要求因变量数据分布等离散，表 3 中关于 2000 年所有人年记录生育数的描述性统计表明，生育子女数的平均值等于0.04，而方差也接近于0.040①，二者几乎相等。因此，因变量分布可以认为满足等离散假定。

采用自变量分步纳入模型的方式，得到 3 个模型，分析结果见表 4。下面针对各模型进行讨论。

模型一只考虑了年龄的影响，它为后面其他模型的评价建立了一个基线模型，这一模型实际上等价于上述报告的以 5 岁年龄组做泊松回归的模型。不同的是，这

表 **3** **2000** 年人年变量测量的相关说明及描述性统计

变量名称	含义与测量	平均值	标准差
因变量	2000 年的人年生育孩子数：n= 0，1，2（2 为双胞胎）	0.040	0.199
协变量			
人年年龄	按 5 岁组虚拟编码：1= 15～19；…；7= 45～49	32.692[a]	9.149
民　　族	1= 少数民族；0= 汉族	0.095	0.294
居住地类型	1= 农村；0= 城镇	0.745	0.436
初婚年龄	根据妇女申报的初婚年月和出生年月计算[b]	21.670	3.041
受教育程度	1= 文盲；2= 小学；3= 初中；4= 初中以上		
生育意愿	妇女申报的家庭最理想的子女数[c]	1.702	0.658
地　　区	1= 东部；2= 中部；3= 西部		

注：a 年龄的平均值和标准差均按人年年龄直接计算。b 有少量未婚人年案例的初婚年龄缺失。对于这些未婚案例采用 2000 年年底时该案例的确切年龄替补。c 调查问卷中该问题的编码 9 代表无所谓要几个孩子。一些案例回答中选择了此项。为了将该变量在回归中作为定距自变量，同时不损失这些案例，用 2000 年时其他人年案例的理想子女数平均值作为替补。

① 表 3 给出了生育数的平均值和标准差，将标准差 0.199 平方就得到了方差 0.0396。

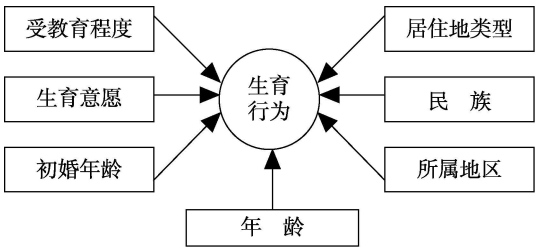

图 **4**　解释性研究理论假设框图

里回归的目的并不是取得生育率，而是分析各年龄组之间的差异，因此这里不再将年龄组虚拟变量饱和纳入，并且允许模型出现常数项。在这种情况下，仍然可以根据输出结果间接地计算各年龄组的生育率，但此种输出更方便于分析与比较。常数项系数的幂 exp(－5.34)＝0.005表示参照年龄组 15～19 岁的生育率，而其他各年龄组回归系数的幂表示与参照组生育率之间的倍数差别。比如，20～24 岁组回归系数幂 28.336 表示该年龄组生育率是 15～19 岁组的 28.3 倍(也可得出该年龄组生育率为 0.136)。由于尚未在模型中纳入其他解释因素，模型一中各年龄组生育率之间的倍数差异其实还包含着其他各种影响因素的共同作用。换句话说，此时的生育率没有得到其他方面足够的统计控制。但当在模型中纳入其他社会变量时，便可以检验各自变量的净影响是否显著。

表 **4**　解释性泊松回归模型的统计结果

自变量	模型一		模型二		模型三	
	$\hat{\beta}$	$\exp\hat{\beta}$	$\hat{\beta}$	$\exp\hat{\beta}$	$\hat{\beta}$	$\exp\hat{\beta}$
常数项	－5.340	—	－8.317	—	－8.720	—
年龄组(岁)						
15～19(参照类)						
20～24	3.344	28.336**	2.561	12.945**	2.546	12.751**
25～29	3.108	22.376**	2.046	7.740**	2.045	7.726**
30～34	1.991	7.322**	0.883	2.419**	0.902	2.464**
35～39	0.174	1.191	－0.958	0.384**	－0.937	0.392**
40～44	－1.002	0.367*	－2.315	0.099**	－2.275	0.103**
45～49	－3.171	0.042**	－4.590	0.010**	－4.539	0.011**
初婚年龄			0.176	1.192**	0.184	1.202**
理想子女数			0.253	1.288**	0.230	1.258**
受教育程度						
文盲(参照类)						
小学			－0.210	0.810*	－0.183	0.833*
初中			－0.366	0.694**	－0.285	0.752**
高中及以上			－0.953	0.386**	－0.711	0.491**
民族						
汉族(参照类)						
少数民族					0.069	1.071
居住地类型						
城镇(参照类)						
农村					0.311	1.365**
地区						
西部(参照类)						
中部					－0.022	0.978
东部					－0.190	0.827**
Log likelihood	－5471.05		－5303.12		－5288.96	
自由度	6		11		15	

注：(1)在 Stata 中进行泊松回归时，如果要求给出常数项，则不能直接得到 $\exp\hat{\beta}$ 或 IRR。因此，表中的 $\exp\hat{\beta}$ 值是通过重新计算得到的。(2) * 表示在 0.05 水平上显著，** 表示在 0.01 水平上显著。

模型二在模型一的基础上又纳入了妇女的初婚年龄、受教育程度和理想子女数 3 个影响因素，

其中的初婚年龄是个定距变量，且近似于连续。整体上，模型二的对数似然值比模型一大167.93，由此，可以得到模型拟合优度统计指标值 G^2 为 335.86，远远大于临界值 $\chi^2_{0.01}(5)=15.086$，表明在0.01的显著性水平上肯定了模型整体拟合优度的提高。换句话说，模型二新加的3个因素（5个自变量）贡献了很大的解释能力，模型二对数据的拟合要显著优于模型一。

从表4提供的模型二单个变量的效应来看，3个社会因素（理想子女数、初婚年龄和受教育程度）都对育龄妇女的生育行为具有统计性显著的影响。正如理论假设的那样，在控制了其他变量的条件下，理想子女数每增加1个，生育率将会提高29%。而初婚年龄晚的育龄妇女的生育率要高于初婚年龄早的育龄妇女。初婚年龄每推迟1岁，生育率将是原来的1.19倍。这一结果乍看起来似乎与“晚婚导致少育”的通识有违。但应当指出，上述这种通识在表达时并没有任何限制条件，而这里的回归结果所描述的则是“在相同年龄、相同教育、相同生育意愿条件下”的情况。这一结果实际上意味着，晚婚会将同样的意愿生育数量压缩在更晚且更短的生育期内完成。也就是说，在同样条件下初婚较早的更可能已经在2000年前生育完了，而晚婚者则更可能仍处于实际生育阶段。模型二的结果还显示出，受教育越多生育率越低的反向关系。在其他条件相同时，小学程度妇女的生育率只有文盲类的81%，初中程度为文盲类的69%，而高中及以上程度育龄妇女的生育率则相对更低，还不到文盲类40%的水平。

在生育行为研究中，经常需要根据民族、居住地类型、地区等群组变量对育龄妇女进行分组，讨论不同群组特征对妇女生育水平的影响。因此，模型三在模型二的基础上又纳入了分别表示民族、城乡、地区的虚拟变量。

这些变量的引入又带来了模型拟合优度上的改进。反映模型拟合优度改进的差异统计量 G^2 为28.32，大于相应的检验临界值 $\chi^2_{0.01}(4)=13.277$，仍然可以在显著度为0.01的水平上认为模型三对数据的拟合又要优于模型二。

表4的模型三输出表明，在控制了更多自变量的情况下，初婚年龄、受教育程度和理想子女数与模型二结果稍有变化，但基本结论仍维持不变。新加的民族变量的发生率比虽然是正值，表示少数民族生育率高于汉族，但差别比例很小，并没有按照理论预期那样对生育水平存在统计性显著的影响。在施加了统计控制条件下，不同地区的生育率仍存在着差异。与作为参照类的西部相比，中部、东部的生育率水平依次递减。中部生育率约为西部的98%，这种差异并未达到统计性显著；而东部生育率显著区别于西部，只有西部的83%。

前面我们曾经做过假定城乡生育模式相同、仅控制年龄组影响的模型，得到的结论是，农村生育率是城镇的1.512倍，而这里的分析结果显示出，在控制了很多社会自变量的条件下，农村生育率仅为城镇的1.365倍。相比前后两个模型关于城乡生育率差别的不同结论，可以得知，前面所做的模型其实将很多与城乡有关的其他因素的影响全部算到了城乡差别之上，夸大了城乡生育水平差别。在这里，由于模型中纳入了很多社会变量，于是部分地将这种张冠李戴的影响加以正名，分划到其他社会变量身上，因此在具有更多统计控制条件下的城乡生育率差别便明显缩小。

四、总　结

泊松回归模型，在处理生育史信息方面具有以下的优越性。

第一，泊松回归可以方便地估计年龄别生育率及总和生育率，它得到的估计值与常规人口统计方法得到的结果基本相同。由于这种方法的数据处理不同，因而保持了数据的整体性，可

以便捷地计算各种不同的模型。

第二，泊松回归模型可以同时容纳很多自变量，这是常规人口统计方法所不具备的。这一特点十分有利于加强研究分析中的统计控制能力，也适于开发样本规模并不太大的抽样调查数据。它除了可以很方便地服务于描述性分析和解释性分析，还可以实施统计检验。

第三，泊松回归模型既可以容纳代表分类信息的虚拟变量，也可以容纳定距的连续变量（如本文解释性模型中纳入了较精确测量的初婚年龄）。本文还示范了将年龄作为分类的虚拟变量使用，其优点是摆脱了对年龄只存在线性作用的简单化假设，可以根据数据信息拟合更复杂的年龄别生育模式。当研究要求必须考虑不同类别（如城乡）存在不同年龄模式时，也可以通过加入城乡与年龄的交互作用自变量的方法加以估计。

第四，本文主要示范了个体数据的分析，但实际上泊松回归也可以应用于群组数据。比如，只要具有各交互分组中的事件发生总数和人期总数（与人口统计在计算各年龄组生育率时必须有各组的分子和分母数据相同），也可以直接用泊松回归来计算各组生育率。

第五，泊松回归是一种更为一般化的统计分析工具，它还可以计算其他方面的事件发生率，因此在社会科学领域具有更为广阔的应用空间。

虽然泊松回归具有种种优点，但应用上仍存在一定难度，因为其数据处理工作比较复杂。泊松回归和其他高级社会统计方法（如事件史分析）一样，其分析虽然可以用统计软件来轻易完成，但要求研究者必须具有很强的数据处理能力，能够事先将原始数据改造成具有统计软件可接受的特定格式。

参考文献：

1. 陈卫、吴丽丽（2006）：《中国人口迁移与生育率关系研究》，《人口研究》，第 1 期。
2. 郭申阳（1999）：《使用 SPSS 软件对事件史原始数据进行预处理》，载于郭志刚主编：《社会统计分析方法——SPSS 软件应用》，中国人民大学出版社。
3. 郭志刚（1999）：《社会统计分析方法——SPSS 软件应用》，中国人民大学出版社。
4. 郭志刚（2000）：《从近年来的时期生育行为看终身生育水平》，《人口研究》，第 1 期。
5. 郭志刚（2001）：《历时研究与事件史分析》，《中国人口科学》，第 1 期。
6. 郭志刚（2004a）：《分析单位、分层结构、分层模型》，《北大社会学》（第 1 期），北京大学出版社。
7. 郭志刚（2004b）：《对中国 1990 年代生育水平的研究与讨论》，《人口研究》，第 2 期。
8. 靳小怡、李树茁、马库斯·费尔德曼（2004）：《婚姻形式与生育水平：对中国农村三个县的考察》，《人口与经济》，第 5 期。
9. 梁在（1999）：《事件史分析》，载于郭志刚主编：《社会统计分析方法——SPSS 软件应用》，中国人民大学出版社。
10. 李克、余顺章（1997）：《Poisson 模型与队列研究——原理与应用》，《中国卫生统计》，第 1 期。
11. 李树茁、马库斯·费尔德曼、朱楚珠（1998）：《中国农村妇女就业状态与生育行为比较研究》，《人口与经济》，第 1 期。
12. 林富德、刘金塘（1998）：《中国生育率转变中的发展因素》，《南方人口》，第 1 期。
13. 潘贵玉等（2003）：《2001 年全国计划生育/生殖健康调查数据集》，中国人口出版社。
14. 沈其君等（1999）：《基于 Poisson 回归模型归因比例的极大似然估计》，《中华预防医学杂志》，第 1 期。
15. 孙全富、邹剑明（1998）：《放射流行病学群组研究资料 Poisson 回归分析及其进展》，《中华放射医学与防护杂志》，第 6 期。
16. 王金营等（2004）：《中国省级 2000 年育龄妇女总和生育率评估》，《人口研究》，第 2 期。
17. 夏结来、徐雷（2003）：《计数资料的统计分析模型》，《疾病控制杂志》，第 2 期。

18. 项永兵等(1995):《Poisson回归和Cox回归模型在队列随访资料分析中的应用与对比》,《中国慢性病预防与控制》,第2期。
19. 杨玲等(2005):《中国2000年及2005年恶性肿瘤发病率死亡的估计与预测》,《中国卫生统计》,第4期。
20. 宇传华等(1996):《Poisson回归在职业队列研究中的应用》,《中国卫生统计》,第1期。
21. 于浩等(1996):《Poisson回归模型的应用》,《江苏预防医学》,第3期。
22. 于景元、袁建华(1996):《近年来中国妇女生育状况分析》,载于蒋正华主编:《1992年中国生育率抽样调查论文集》,中国人口出版社。
23. 查瑞传(1991):《人口普查资料分析技术》,中国人口出版社。
24. Allison, Paul(1985), Survival Analysis of Backward Recurrence Time Data. *Journal of the American Statistical Association*, 80, pp. 315-322.
25. Cameron, A. Colin and Pravin K. Trivedi(1998), *Regression Analysis of Count Data*. Cambridge; New York: Cambridge University Press.
26. Cleland, J. and G. Rodriguez(1988), The Effect of Parental Education on Marital Fertility in Developing Countries. *Population Studies*, 42, pp. 419-442.
27. Healy, M.J.R.(1988), *Glim: An Introduction*. Oxford: Clarendon Press.
28. Lindsey, James K.(1995), *Modelling Frequency and Count Data*. Oxford: Clarendon Press; New York: Oxford University Press.
29. Long, Scott J.(1997), *Regression Models for Categorical and Limited Dependent Variables*, Thousand Oaks: Sage Publications.
30. Long, Scott J. and Freese, Jeremy(2001), *Regression Models for Categorical Dependent Variables Using Stata*, College Station, Tex.: Stata Press.
31. Poston, Dudley L. and Baochang Gu(1987), Socioeconomic Development, Family Planning, and Fertility in China. *Demography*, 24, pp. 531-551.
32. Powers, Daniel A. and Yu Xie(2000), *Statistical Methods for Categorical Data Analysis*, San Diego: Academic Press.
33. Rodriguez G. and J. Cleland(1988), Modelling Marital Fertility by Age and Duration: an Empirical Appraisal of the Page Model. *Population Studies*, 42, pp. 241-257.
34. Schoumaker, Bruno(2004), A Person Period Approach to Analyzing Birth Histories. *Population E*, 59, pp. 689-702.
35. Trussell J. and G. Rodriguez(1990), Heterogeneity in Demographic Research. In J. Adams, D. Lam, A. Hermalin, and P. Smouse(eds.), *Convergent Issues in Genetics and Demography*. New York: Oxford University Press, pp. 111-132.
36. Tuma, Nancy Brandon, Michael T. Hannan, and Lyle P. Groeneveld(1979), Dynamic Analysis of Event Histories. *The American Journal of Sociology*, 84, pp. 820-854.
37. White, Michael J., et al.(2005), Urbanization and the fertility transition in Ghana. *Population Research and Policy Review*, 24, pp. 59-83.
38. Winkelmann, Rainer(1995), Duration Dependence and Dispersion in Count Data Models. *Journal of Business & Economic Statistics*, 13, pp. 467-474.
39. Winkelmann, Rainer(2000), Econometric Analysis of Count Data. Berlin, New York: Springer.

(责任编辑:朱 犁)

Chinese Journal of Population Science *No. 4*
A bimonthly August *1, 2006*

ABSTRACTS

Application of Poisson Regression in Fertility Study *Guo Zhigang Wu Xiwei* · 2 ·

Poisson regression is a regression model for analyzing the dependent variable of count data. This paper illustrates its application to fertility study with the data from 2001 national family planning and reproductive health survey. Poisson regression not only accepts dummy independent variables standing for age, sex, and other concepts commonly used in demography, but also takes continuous variables as covariates such as income and expense. Therefore, it facilitates fertility study in estimating, comparing, and analyzing.

Management Costs and Efficiencies of Rural Medical Financial Assistance Programs *Zhu Ling* · 16 ·

Financial programs for rural medical assistance are complicated in their implementation, for they involve numerous actors, including various government departments, health service providers, self-governing village organizations, rural households, and individuals. The coordination among these stakeholders is difficult, and operational costs of these programs are rather high. This paper shows that a considerable number of management offices tried to pass some management costs on to other participating institutions due to the shortage of program funds and management outlays. This led to various distortions of the management system in practice and reduced the effectiveness of the programs. Therefore, it is necessary for both central and provincial governments to intensify the transfer of medical financial assistance funds to poor counties. At the same time, institutional innovation should be encouraged in order to create a simpler and more effective management system. In addition, these programs should be constantly monitored with the Public Expenditure Tracking Survey.

Life Cycle Model and Its Application to Research in Aging China *Li Hongxin Bai Xuemei* · 28 ·

This paper analyzes the changes of consumption, saving and asset per capita in aging process in a two period life cycle model. According to Chinese population data, a computable OLG model is established under Walras equilibrium condition with production function, government revenues and pension payments. A basic relationship and several alternative pension reform schemes are simulated under an ageing context.

Estimation and Analysis on Total Fertility Rate *Zhang Qing* · 35 ·

Paying attention to the properties and defects of existing sample data of population, this paper estimated China's total fertility rates from 1994 to 2004 by revising total fertility rate index, and then discussed causal factors of the index outlined. The result shows that total fertility rate in the country dropped from about 1.80 in 1994-1996 to about 1.62 in 2001-2004, and it was 1.66 in 2000. The important factors affecting total fertility rate are economic development level, the general fertility rate, the child-bearing age and the level of urbanization.

Migrant Workers' Employment Substitution for Urban Labors *Ding Renchuan Wu Ruijun* · 43 ·

This paper aims to establish a new quantitative model on the basis of microeconomic theories to explain the substitution of migrant workers for urban labors. An empirical study on 2000 Census data finds that 7.2% of urban labor employment is substituted by migrant workers, which account for 33.6% of the total migrants employed in urban areas. Therefore, the relationship between these two groups of labor force is supplementary, rather than substitutive. However, the degrees of substitution differ among genders, educational levels and occupations, thus the labor market is still segregated.