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。

个数据再分别计算生育率。而在回归分析中,要达到同样的统计控制,只需要在模型中加入一个表示案例城乡属性的自变量即可。如果需要在生育率研究中加入更多自变量时,常规生育率计算方法的局限性便越发凸显出来,因为这需要将原始数据分成更多类型的子样本。而当生育率研究必须将一些连续变量(如精确测量的收入)作为自变量时,就更困难了。通常有两种方法来实现操作化。

第一种方法是将连续变量分段,形成少数类别,然后加以使用。这种做法有两大代价:一是会极大损失原有连续变量中的信息;二是分隔点的选择不可避免主观任意性,而分隔点的不同选择还可能造成完全不同的统计结论。

第二种方法是提高分析单位层次。比如,以省、县为单位,用社会、经济、人口等自变量对总和生育率做回归分析(Poston 等,1987;林富德、刘金塘,1998;王金营等,2004)。尽管这类研究对于理解中国生育率转变非常重要,但生育毕竟是妇女的个体行为,因此上述宏观研究结果很难解释个体生育行为的发生机理。如果简单将这些宏观层次的结论推论到微观的家庭、个人层次上,便可能导致一种与分析单位相连的方法论谬误,即生态学谬误①。

另有一些研究坚持致力于微观分析,采用个体为分析单位,用各种回归模型来解释个体生育行为发生的原因。比如,李树茁等(1998)分别以期望孩子数和曾生子女数作为因变量,应用多元回归研究妇女就业的水平和形式如何影响其生育行为。与此类似,靳小怡等(2004)应用多元回归分析了农村婚姻形式与妇女的终身子女数的关系。陈卫、吴丽丽(2006)应用 Logistic 回归分析迁移行为与生育行为的关系,其模型的因变量为代表"普查前一年是否生育了孩子"的二分变量。然而,这些研究的因变量并不是生育率,而是与生育率有关的其他测量指标。

我们注意到,国内研究文献中几乎没有对生育率的微观回归分析,其实是因为缺乏一种恰当的回归模型可用于个体数据的生育率研究。常规回归方法的因变量必须是连续变量。但生育率是一种平均事件发生率,通常只能是对集合数据的统计描述,而对个人则无法计算生育率。这就造成个人层面的回归模型不可能有生育率指标来作为因变量。因此,微观层面的回归分析只能采用其他测量指标。

曾生子女数实际上记录的是一位妇女已经生育的次数,其取值范围很小,并且只能为非负整数值,在统计中称为计数变量。严格地说,它的分布既不是连续的,也不是正态的(生育较少的人很多,而生育很多的人很少)。将此类变量作为因变量进行常规回归分析便会违反这种方法本身所要求的假定条件。然而在缺乏更好回归模型时,将常规回归应用于计数因变量是一种常见的做法(White 等,2005)。但是,我们知道在违反其假定的条件下,常规回归的估计是有严重偏差的,并且相应的统计检验都是无效的(郭志刚,1999:177)。而泊松回归恰好可以解决上述问题。首先,泊松回归是专门为分析因变量为诸如生育次数、迁移次数等计数变量发展出来的统计模型(Lindsey,1995;Winkelmann,2000)。它不仅可以应用于生育率研究,也可以应用于其他更为广阔的研究领域。其次,泊松回归一方面可以纳入年龄、性别等常用的人口自变量,用于估计年龄别生育率、性别年龄别迁移率等经典描述性人口学指标以外,也可以纳入更多社会经济变量。在数据信息具备的情况下,还可以在估计有关发生率的同时进行解释性研究的探索(Powers 等,2000;Schoumaker,2004)。第三,泊松回归中的自变量不仅可以是代表年龄组和城乡分组的虚拟变量,也可以是连续变量。所以,不必再像常规人口统计中那样,

① 参见郭志刚(2004a)对与分析单位相连的方法论谬误的评论。

非要先将连续变量转换为分类变量后才可应用。

本文的主要目的：一是通过实际数据分析来演示应用泊松回归来估计生育率；二是通过在泊松回归中引入更多的解释变量来展示该方法在测量生育率变化趋势与分析生育率影响因素等方面的功效和灵活性。这种方法十分有利于深入开发实际调查的生育史、迁移史等数据，从而为国内人口学、社会学界的研究人员提供一种新方法的选择。

二、泊松回归模型

在社会科学量化研究中，如果研究者试图进行解释性研究或者对某一理论进行检验，那么，回归模型很可能是最基本的工具。近些年，回归模型已经从最基础的正态线性回归模型发展出更多的类型。回归模型的选择在很大程度上取决于因变量的类型。

在社会科学研究中，因变量常常是计数类型的变量，诸如一定时期内的生育孩子数、迁移次数、犯罪次数、某类疾病的发病次数、看病次数等，它们都是某种事件发生数。计数变量的特征非常鲜明，它们取值为 0、1、2、3……离散的非负整数，且通常最大值并不是很大。如妇女终身生育数在理论上小于 20，在实际数据中可见到的曾生子女数就更小了。在统计文献中，这类变量被称为计数变量，且经常被作为分类变量的一种形态（Powers 等，2000；Long，1997；Long 等，2001）。由于计数变量不是连续的，并且分布又呈明显偏态，因而不可以作为常规回归的因变量。从 1980 年开始，在计量经济学研究和流行病学研究中就开始发展出一类专门用于对计数变量数据进行分析的模型，被称为计数变量模型或事件——计数分析[①]（Tuma 等，1979）。计数变量的标准模型为泊松分布，也就是说泊松回归模型是建立在泊松分布基础上的回归模型（Cameron 等，1998：9），它构成了对计数变量进行多元量化分析的起点。

（一）几个基本概念

1. 事件

泊松回归模型的因变量是一定时期内事件的发生次数。和事件史分析一样，这里所谓的事件是一个宽泛的概念。它可以看做是地位的变化或者是性质状态的转换，比如结婚、生育、死亡、失业或就业、迁移等。

2. 比率

在统计学上，比率属于相对数，又可称为率。很多人口统计指标都与比率有关，因此，我们对于率的概念似乎比较熟悉。但由于习惯用法上的不严谨，人口统计中有些被称作率的指标实际上应该分别称为“比”或“比例”，但却错误地被称作“比率”（查瑞传，1991：55）。实际上，比率在研究中具有其特定的定义。简单地讲，比率是指单位时期内某一事件的发生数与该时期暴露在可能发生该事件风险中的人期总数的比。比如，1990 年的一般生育率等于 1990 年育龄妇女的生育总数除以所有育龄妇女在当年所存活的人年总数。与比例不同，比率是对某一事件发生的瞬时概率的测量，属于动态概念。而比例表达的是成功的试验次数与试验总数之比，是一个静态概念。

3. 暴露期

对暴露期的考虑是计算事件发生率的关键。暴露期指个体或观察案例在转入下一状态之

① 与事件计数相比，也许人们对持续期模型更为熟悉。实际上持续期模型和事件计数模型密切相关，因为累计等待时间的分布决定着计数的分布（Winkelmann，2000：52～64）。

前的初始状态上所持续的时间长度，这又被称为个体或观测案例在初始状态中的持续期或风险期或等待时间。在标准的泊松回归模型中，假定处于暴露期内的所有的观察案例具有同质性，即视他们所有个人特征对事件发生率没有影响。这种标准泊松模型也称无条件泊松模型。但实际上暴露期会随着观察案例特征的不同而不同(Winkelmann，2000：73)。一般而言，妇女在结婚以后才会有生育行为，不同的妇女婚后多久才会生育第一个孩子存在很大差异。对于农村妇女，大多在婚后一年生育，而城镇妇女可能会等待更长的时间才会有第一次生育。然而，一旦事件发生，个体的风险持续期或等待时间就可以观测得到，并用于计算暴露期总数。

(二) 泊松回归模型原理

国内流行病学领域的研究者视泊松回归为队列随访资料分析中常见的多变量统计分析方法之一(李克、余顺章，1997；孙全富、邹剑明，1998；沈其君等，1999；夏结来、徐雷，2003)，也有不少应用泊松回归进行研究的成果(项永兵等，1995；宇传华等，1996；于浩等，1996；杨玲等，2005)。但是，国内其他社会科学领域还很少有介绍和实际应用该模型。对于这一模型，有不同的叫法。Allison(1985)称其为恒定风险模型；Long 等则称其为泊松回归(Long，1997；Long 等，2001；Cameron 等，1998)；而 Powers 和 Xie(2000)称其为对数率模型。但他们都是以事件发生次数作为研究对象，研究风险暴露期和其他协变量对事件发生率的影响①。更为重要的是，这一模型是假定事件发生遵循著名的泊松分布的基础上推导出来的(Cameron 等，1998：9)。

1. 模型设定

以 y 表示对某一事件发生数的观测，假定随机变量 Y 等于 y 的概率，并遵循均值为 μ 的泊松分布，则该泊松分布的密度函数为：

$$\Pr(Y=y|\mu)=\frac{e^{-\mu}\mu^{y}}{y!} \qquad y=0,1,2\cdots \tag{1}$$

在式(1)中，$\mu>0$ 且 μ 它是定义分布时的唯一参数。当然，这是针对单变量泊松分布的情况。也可以通过允许每一观测具有不同的 μ 值将泊松分布扩展为泊松回归模型(Long 等，2001：229)。在更一般的情况下，泊松回归模型假定，表示对个体 i 某一事件发生数的观测 y_i 遵循均值为 μ_i 的泊松分布，那么，该分布的密度函数为：

$$\Pr(Y_i=y_i|\mu_i)=\frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \qquad y_i=0,1,2\cdots \tag{2}$$

μ_i 可根据一些可观察的特征估计得到，这就有以下结构方程：

$$\mu_i=E(y_i|X_i)=\exp(X_i'\beta')=\prod_{j=1}^{k}\exp(\beta_j x_{ji}) \tag{3}$$

实际上，式(1)和(3)联合起来才定义了一个完整的泊松回归模型(Cameron 等，1998：10)，对 $X_i'\beta'$ 取指数是为了保证参数 μ_i 为非负数。这时，均值 μ_i 也是一个条件均值，反映的是在一系列因素作用下事件的平均发生数，只不过作用被表达为乘法形式。将式(3)两边取对数，可以得到该条件均值的一种加法形式表达：

$$\ln\mu_i=X_i'\beta'=\sum_{j=1}^{k}\beta_j x_{ji} \tag{4}$$

通过式(4)对事件发生数的平均值的对数转换，方程左侧的对数条件均值(或称对数率)已经表达为 k 个自变量的线性函数。

① 美国密歇根大学谢宇教授对文章评阅后指出，泊松回归与对数率模型有所不同；泊松回归模型并不涉及暴露期，而对数率模型包含暴露期。因此，本文实际上是利用泊松回归形式求解对数率模型。

泊松分布有一个重要的特征,就是均值和方差相等。在泊松回归模型中,这成为了一个非常关键的假定条件,即等离散假定。违背等离散假定的情况既可能是过离散(即方差大于均值),也可能是欠离散(即方差小于均值)。对等离散假定的违背足以造成对泊松假定的违背(Winkelmann, 2000: 11)。

2. 参数估计方法

根据函数关系表达的形式,上述式(3)和(4)分别被称作乘法模型与加法模型。其中,都只有 β_j 是未知参数,可以采用最大似然法进行估计,或者采用迭代再加权最小二乘法求解(Powers 等, 2000; Cameron 等, 1998)。

3. 模型拟合优度评价与模型选择

模型拟合的输出结果一般都会给出对数似然值,由于该值会受到样本量大小的影响,因而不能单独用作对模型拟合优度评价的指标。对同一数据拟合不同的模型就可以得到不同的对数似然值,如果这些模型之间存在嵌套关系,那么我们可以采用似然比指标 G^2 对不同模型的拟合优度做出评价,从而对模型进行选择。

以 L_c 表示当前模型的似然值,在当前模型中继续纳入协变量,得到限制模型的似然值 L_r。那么,$G^2=2(L_r-L_c)\sim\chi^2_k$。其中,$k$ 为限制模型与当前模型协变量数目的差值。这里零假设为限制模型和当前模型无差异。统计软件很可能只会给出每一模型的对数似然值,在这种情况下,我们需要计算 χ^2_k 的值,如果 $\chi^2_k>\chi^2_{\alpha,k}$,那么我们就拒绝零假设,认为限制模型对数据的拟合优于当前模型①。当然,反过来,这也可以用来作为判断某一因素是否纳入到模型中加以分析。

4. 回归系数解释

对泊松回归模型进行解释有多种不同的方式,这取决于研究者是对计数变量的期望值还是对计数的分布感兴趣(Long 等, 2001: 231)。如果对期望值感兴趣的话,有多种方法可以用于计算某一自变量一定程度的变化量所带来的计数变量期望值的变化量,既可以用期望值的倍数变化来表达,也可以用百分比变化来表达,甚至还可以用期望值的边际变化来表达。其中,最常用的解释方法是计算倍数变化。这一解释方法非常直观、非常容易理解。

泊松回归系数 β_j 可以被解释为:在控制其他变量的条件下,x_j 变化 1 个单位,将带来对数均值上的变化量。然而研究人员真正关心的并不是取对数的均值,而是期望计数(即率)本身。因此,可以用 $\exp(\beta_j)$ 来反映 x_j 变化 1 个单位时期望计数的倍数变化。$\exp(\beta_j)$ 又称为发生率比(标为 IRR)。当然,这是针对连续自变量而言。当自变量为代表分类的虚拟变量时,$\exp(\beta_j)$ 表示在控制其他变量的条件下,某一类别的期望计数为参照类期望计数的相应倍数。这其实与 Logistic 回归系数的解释类似。

5. 实现手段

泊松回归模型的参数估计采用最大似然法或者迭代重复加权最小二乘法求解。以前,这些计算一般是通过专门用于对广义线性模型进行统计分析的 GLIM 软件包来进行(Trussell 等, 1990; Rodríguez 等, 1988; Healy, 1988)。现在,SAS 和 Stata 等许多常见的统计分析软件也都可以对泊松回归模型进行估计。本文采用 Stata 8.0 软件进行分析。

(三) 泊松回归模型在生育研究中的具体化

① 关于模型拟合优度评价的详细介绍,可参见 Long(1997) 或 Powers 等(2000)。

Rodríguez 和 Cleland(1988)指出,如果将妇女的生育数视为独立的泊松随机变量①,那么,其均值可以表达为暴露期和理论已婚生育率两者的乘积。据此,如果以 y_j 表示一定时期内育龄妇女 i 的生育数,那么,均值 μ_i 反映了妇女 i 在某一时期的平均生育数。该均值可被分解成生育率 $_i$ 和风险长度 t_i 两者的乘积:$\mu_i = t_i$ $_i$,因而,均值 μ_i 的对数就等于风险长度 t_i 与生育率 $_i$ 的对数和,即 $\ln\mu_i = \ln t_i + \ln$ $_i$。式中 $\ln t_i$ 被称为偏移量,它是系数固定为 1 的自变量(Trussell 等,1990)。纳入它意在对每一位妇女的风险长度进行控制(Powers 等,2000:156),同时意味着假定风险随着持续期的延长按比例增加。

进一步,生育率($_i$)的对数可以被表达成 k 个解释变量的线性函数:$\ln$ $_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki}$,因此,$\ln\mu_i = \ln t_i + \beta_1 x_{1i} + \beta_2 x_{1i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki}$。

这样,就把一定时期育龄妇女 i 的生育数的对数表达成了其风险长度的对数和 k 个解释变量的线性函数。泊松回归系数的符号反映了各解释变量对生育率的影响方向,而系数的大小则反映影响强度。回归系数的幂表达了不同妇女群体的生育率或类别之间的生育率差异,这取决于研究人员如何将代表各类别的虚拟变量纳入模型。

泊松回归在生育数据分析中存在的优势:(1)它可以凭借调整偏移量 offset 来控制模型中每一人年中的实际风险长度。比如,在某人年中生育可以发生于任一日期,而生育以后距该年底的时间实际上并不属于暴露期。(2)泊松回归既可以计算经典的生育水平指标又可以对生育率进行解释研究,因此可以把生育分析的描述性研究和解释性研究结合在一起。

(四) 数据来源和数据处理

泊松模型既可以处理分组数据也可以处理个体数据(Powers 等,2000;Rodríguez 等,1988)。特别是 Schoumaker(2004)又进一步提出,可以先将个体数据改造为人期数据,然后应用泊松回归模型来估计生育率②。在这种人期数据中,分析单位不再是作为个体的社会行动者而是人期。也就是说,此时分析单位不再是个人案例而是由个人生育史转换的若干单位时间。一旦将原始生育史调查数据转换为人期数据之后,便可以简单应用泊松回归来估计和分析生育率及其影响因素。并且,这种方式还可以将"随时间变动的变量"或称"动态变量"纳入到解释模型之中(梁在,1999;郭志刚,2001)。沿着 Schoumaker 发展出来的数据使用方式,本文以人年作为分析单位对 2001 年全国计划生育/生殖健康调查的个人问卷原始数据进行类似地改造,以便利用这一调查的有关信息开展对生育率的泊松回归分析。

与 Schoumaker 使用前 5 年的回顾性调查数据不同,"2001 年全国计划生育/生殖健康调查"对当年为 15~45 岁的育龄妇女的历次怀孕情况均进行了回顾性调查,问卷表格中最高设计了 12 次怀孕事件。这也就意味着,调查时每一位妇女已经度过的生育期是不一样长的。而且,这几乎是对每一位妇女全部生育史的回顾。因此,每一个妇女所能提供的人年信息记录的数量是不同的。下面将按照原始数据中提供的个人生育史信息来形成人年格式的数据。表 1 显示了原始的个体数据格式。

① 他们同时指出,妇女的生育数遵循泊松分布的观点可以由关于生育子女数的随机性质的自然假设得到证明。另据 Allison(1985)的研究,只要事件是可重复的而且是无差异的,就可以把回顾性事件时间问题作为一个随机过程加以处理。

② 在 Schoumaker 的文章中,以调查时间之前 5 年作为人期单位,这对应着他计算生育率按 5 岁组划分。但实际上根据数据和研究需要,可以选用不同人期单位,比如人年或人月,甚至人天或人时。

表 1　以妇女为记录单位的原始数据格式示意

妇女编号	固定变量				随时间发生变化的怀孕事件变量		
	城乡	出生时间	民族	略	每次怀孕结束的年月和结果		略
1	1	198007	2	……	222222 7		……
2	1	196411	2	……	198707 2	198811 1	……
……	……	……		……	……		
2605	1	197803	1	……	200005 1		
		……			……		

注: 本表只列 3 段, 实际最多 12 次。

根据本次调查规则, 编号为 1 的妇女只怀孕了一次, 并处于现孕状态(结果编码为 7), 所以此次怀孕结束年月编码全部为 2。编号为 2 的妇女于 1987 年 7 月活产 1 个女婴(结果编码为 2), 又于 1988 年 11 月又生了 1 个男孩(结果编码为 1)。我们将主要以此为例, 说明数据改造的操作。

人年数据的形成需要根据原始数据中的生育史(而不是所有怀孕史)的事件信息, 将妇女育龄阶段的所有人年均建立单独的一条记录。由于原始数据格式提供的是每一次怀孕事件的明确时间信息, 因此需要先将每名妇女横排列的怀孕史数据转换为多行的人年生育记录数据(郭申阳, 1999: 428)。但是, 由于原始数据只对每一妇女有怀孕事件的年份进行记录, 因此这就需要对每一位进入调查的 15~49 岁的育龄妇女从 15 岁至调查年份(2001 年)之间的每一年产生一条记录, 这样才能重建每一位妇女在 2001 年之前完整的人年生育数据。这一数据处理是一个较复杂的工作, 可以采用 SPSS 软件来完成, Stata 软件也有类似的数据处理功能, 或者可以用 VB 等其他编程软件来完成。

若以表 1 中编号为 2 的育龄妇女为例, 表 2 显示了数据改造之后的人年数据格式。由于我们在数据改造时, 是以人年为分析单位, 因此在最后的人年数据中, 并没有明确出现风险持续期这个变量。这实际上意味着每条人年记录的风险持续期默认为 1。对于那些没有发生生育的人年这是恰如其分的, 但对于那些当年有生育的人年, 这种数据处理就意味着假定生育发生在年底。从后面的分析结果可以看到, 这种忽略对分析结果并未造成什么明显损失。

先来看一下表 1 中编号为 2 妇女的原始数据。该妇女生于 1964 年, 所以在 1979 年时为 15 岁, 进入育龄, 此后即暴露于生育风险下, 到 2001 年调查时经历了 22 年, 因此该妇女有 22 条人年记录。但根据表 1 提供的生育信息, 她只在 1987 年(23 岁)和 1988 年(24 岁)有生育, 因此只有这两年的人年记录的生育数为 1(如果是多胞胎, 生育数就是相应的整数)。本来还可以针对这两个有生育的人年进行偏移量 offset 的调整, 但本研究省略了这一步。

经过改造之后, 得到的分析数据的规模与原始数据发生了巨大变化, 样本量由原始数据的 39 586 条记录增加到 769 966 条记录。这主要是由于时间因素的引入, 分析单位不再是每一个妇女, 而是根据每一个妇女已经在育龄期内经历的年数建立了多条对应每年生育经历的人年记录。改造后的人年生育数据的汇总结果表明, 各年的生育总数和年龄别生育数分布与这次调查后国家人口和计划生育委员会正式发表的数据集(潘贵玉等, 2003)完全相同, 而从人年数据得到的年龄别统计数则与公布数据集中的育龄妇女分布高度吻合。

三、应用泊松回归研究生育率

人年生育数据建立以后, 用 Stata 软件做泊松回归只需要一条命令:“poisson *depvar indvars*”。

表 2　改造后人年数据的格式示意

妇女编码	固定变量				与生育事件相关的人年变量		
	城乡	出生时间	民族	略	人年年份	人年年龄	生育数
2	1	196411	2	……	1979	15	0
2	1	196411	2	……	1980	16	0
……	……	……	……	……	……	……	……
2	1	196411	2	……	1985	21	0
2	1	196411	2	……	1986	22	0
2	1	196411	2	……	*1987*	*23*	*1*
2	1	196411	2	……	*1988*	*24*	*1*
2	1	196411	2	……	1989	25	0
2	1	196411	2	……	1990	26	0
……	……	……	……	……	……	……	……
2	1	196411	2	……	2001	37	0

其中，poisson 启动泊松回归，*depvar* 代表因变量名，*indevars* 代表自变量名。自变量如果有多个可以用空格分隔依次列出。

（一）估计某一年的生育率

对于上述整理好的人年数据，采用泊松回归来估计生育率在操作上简单易行。以妇女在每一人年的生育数作为因变量，将各人年口径的年龄组对应的虚拟变量作为自变量，即可完成各年龄别生育率的回归估计。要做某一年（如 2000 年）生育率估计，便可在命令中加上 if 语句的附加选项，特定选择 2000 年的所有人年记录来进行泊松回归。还可以再增加一些其他选项命令来满足特殊需要。比如，如果希望直接输出年龄别生育率，就需要饱和输入代表 15～49 岁各年龄对应的全部 35 个虚拟变量（如下面命令中的 dage1-dage35）[①]，并且注明不设立作为参照类的常数项，以及直接输出回归系数的幂（irr）来取代默认的回归系数输出。即："poisson kidnum dage1-dage35 if fertyear==2000, noconstant nolog irr"。这里 irr 即 $\exp(\beta_j)$，是年龄别生育率。

泊松回归对 2000 年各年龄别生育率的输出结果与发表数据集中的结果十分接近。泊松回归的年龄别生育率合计出来的总和生育率为 1.455，而相应的公布统计值为 1.448，两者水平极为接近。从图 1 可以看出，泊松估计的生育率在峰值以前均低于公布值，但在峰值以后又基本上高于公布值。并且泊松估计值曲线在 27 岁又出了一个小尖，而在公布生育率的曲线中，并没有出现这一特征。我们认为，这是因为常规生育率计算方法其实具有年龄组间修匀的功能，而泊松回归估计则没有这一功能。

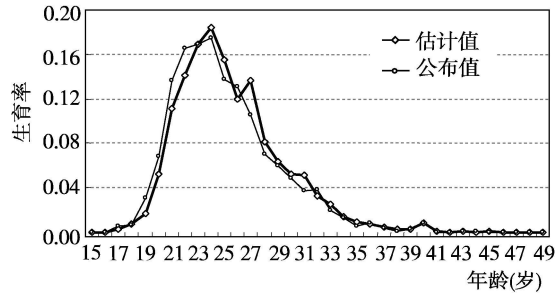

图 1　2000 年中国育龄妇女生育率曲线的比较

如前所述，泊松回归估计生育率时既可以

① 这种做法与通常回归时将年龄作为连续变量纳入模型不同。作为连续变量使用，只能得到一个回归系数，实际上是假定年龄具有线性作用。但用 35 个年龄虚拟变量纳入模型，意味着放弃年龄只存在线性作用的假定，以便为各个年龄组分别估计出最佳拟合的生育率。

按单岁分组，也可以按 5 岁分组。各回归系数的幂仍然是年龄组的生育率。在用 5 岁组生育率计算总和生育率时，需要将合计值扩大 5 倍。采用 5 岁分组在回归估计生育率时有一个明显的好处，就是自变量数量大大减少了，只需要 7 个对应年龄组的虚拟变量。随之而来，输出结果也精简了许多。按 5 岁分组的泊松回归估计计算的总和生育率仍然是 1. 455。

(二) 估计某年份分城乡的生育率

中国城乡之间在社会、经济发展上存在较大差距，因此城乡之间生育率水平存在明显差异。常规计算城乡生育率时必须分别对城乡育龄妇女及其生育数加以汇总，然而在应用泊松回归估计时只需要做一个方程便可以完成。也就是说，只要在定义自变量时，除了表示年龄组的虚拟变量外，再加上一个表示城乡的虚拟变量就行。根据泊松回归估计，得到的城镇 2000 年总和生育率为 1. 051。而农村虚拟变量的发生率比(irr)为 1. 512，表示乡村生育率为城镇妇女的 1. 512 倍，于是得出农村总和生育率为 1. 589(1. 051× 1. 512)。而调查数据集公布的城乡总和生育率分别为 0. 974 和1. 610，可见泊松回归估计与其差别并不大。

上述这种考虑城乡差别的泊松回归只是一种简化分析，由于只加了一个识别城乡的虚拟变量，因此并未考虑城镇的生育模式与乡村生育模式很有可能存在很大差别。实际上这一模型假定对于每一个年龄组，城乡生育率水平都呈同样的比例关系。所以尽管城乡总和生育率水平的估计很接近正式发表的统计，但实际上关于城乡生育模式相同的模型设置却是不符合实际的。对此，我们可以通过在模型中引入城乡变量和各年龄组变量两两相乘得到的交互项变量进行调整，交互项的引入也就意味着回归模型允许城乡的生育模式完全根据数据来计算，不再强制城乡有相同的生育模式。

在这种带交互项的泊松回归模型估计基础上计算出来的城乡总和生育率分别为 0. 991 和 1. 613，与公布的 0. 974 和 1. 610 结果变得更为接近。图 2 提供了泊松回归估计的城乡生育率曲线与公布统计水平之间的比较。

(三) 应用泊松回归重构生育趋势

采用人年数据还使“随时间变动的变量”或者“动态变量”纳入到模型之中成为可能。而人年所属年份和人年所属年龄则是动态变量的一种简单情况。如果考虑将时期因素纳入泊松回归模型时，可以根据人年数据的生育史信息重构过去若干年内的生育趋势(Schoumaker，2004)。

为了模型简单，我们假定不同年份的生育模式不变①。将年份和年龄组作为虚拟变量纳入泊松回归模型，由此可以得到每一年龄组的回归系数和特定年份的回归系数。根据年龄组的回归系数幂，可以计算出参照年份的总和生育率。而特定年份回归系数幂表达的是该年份总和生育率与参照年份总和生育率之间的倍数。于是，就能计算出各年总和生育率估计。由此，通过一个泊松回归，可以方便地计算很多年的生育水平变化趋势。

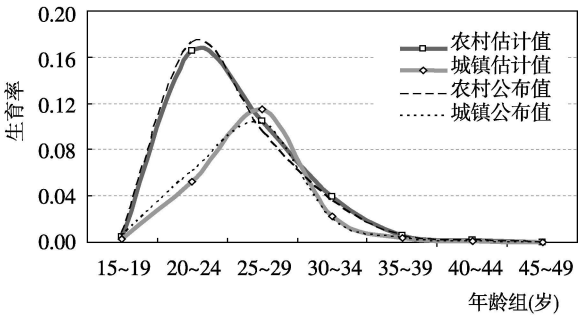

图 2　2000 年城乡生育率曲线：估计值与公布值

注：公布值没有直接提供 5 岁组的计算结果，这里根据潘贵玉等(2003) 主编的《2001 年全国计划生育/生殖健康调查数据集》(上册)，第 8~9、226~227 页提供的数据，采用人口统计方法计算得到。

① 实际上，通过将年份与年龄组变量之间建立交互效应自变量，也可以考虑生育模式随年份变动的情况。

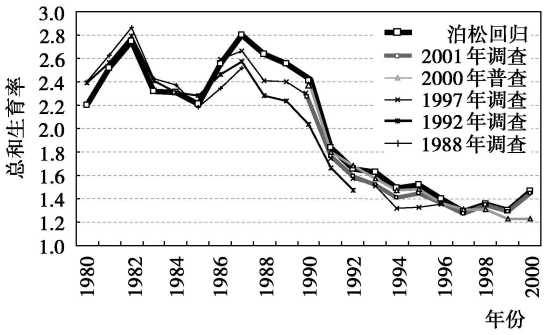

图 **3** 泊松回归估计与正式公布统计及其他来源生育率变化趋势的比较

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。

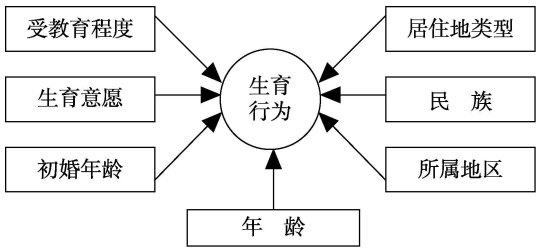

图 **4**　解释性研究理论假设框图

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.