零膨胀模型在社会科学实证研究中的应用

——以中国人工流产影响因素的分析为例

王存同

提要:文章以零膨胀模型(zero-inflated modeling)对中国已婚育龄妇女人工流产影响因素的分析为例,介绍了零膨胀模型在社会科学领域的具体应用,并比较了零膨胀模型与泊松模型(Poisson modeling)、负二项模型(negative binomial modeling)等一般计数模型的分析结果,发现零膨胀模型能较好地处理计数资料中零值过多的问题,其参数估计更为精确,得出的结论更符合实际。

关键词:计数资料　零膨胀模型　人工流产　影响因素

一、引　言

在社会科学计数资料(count data)的实际研究中,经常发现观察事件发生数[①]中含有大量的零值,即许多观察个体在观察单位时间、空间、面积内没有发生相应的随机事件,如一年内的住院次数、离婚次数、坐牢次数、生育子女数、人工流产次数等。这样一种特殊的离散(discrete)和受限因变量(limited dependent variable)数据,超出了泊松/负二项模型等一般计数模型的预测能力(一般模型中零发生概率常被低估),在多学科领域中引起了广泛关注。由于计数资料中的零值过多,且取相同的零值反映了不同的情况,常常会导致计数资料表现出较大的变异,这类现象被称为计数资料的零膨胀(zero-inflated)。

* 课题来源:中央财经大学"211 工程"三期资助、2010 年度国家社会科学基金项目《中国避孕行为的定量与定性研究:1960 - 2008》(10CRK012)。感谢美国科学院院士、密西根大学教授谢宇及美国伊利诺大学香槟分校教授廖福挺(Tim Futing Liao)的帮助。

① 事件数(event count)是指单位时间、空间内事件发生次数,变量取值为 0,1,2,3 等非负整数,表现为事件发生次数的离散型随机变量。

20 世纪 60 年代，就有学者注意到零膨胀现象（Johnson & Kotz，1969）。1986 年，有学者提出了一种解决零膨胀现象的 Hurdle 模型，应用于经济学领域的研究（Mullahy，1986）。① 1992 年，兰伯特提出了另外一种处理零膨胀现象的零膨胀泊松模型（zero-inflated Poisson，ZIP），即引入协变量，对零计数和非零计数建立混合概率分布，建立有协变量的零膨胀泊松模型，应用于电子制造业中的质量控制（Lambert，1992）。1994 年，格瑞因将零膨胀泊松模型扩展到零膨胀负二项模型（zero-inflated negative binomial，ZINB），并采用 BHHH 方法估计模型参数的标准误，应用到消费者银行信用卡不良记录的研究（Greene，1994）。这种零膨胀负二项模型是对泊松模型与负二项模型技术的发展，弥补了泊松模型或负二项模型技术在分析零膨胀结构数据时的不足，能解释计数资料中过多的零值，使因变量中真实零值的鉴别成为可能，同时也使估计结果更为有效与无偏（efficient and unbiased estimates），从而获得可靠的假设检验和参数估计，以帮助研究者解答一系列具有实际意义而传统模型无法回答的问题。

一般而言，处理零膨胀现象的模型包括 Hurdle 模型、零膨胀泊松模型及零膨胀负二项模型等。由于 Hurdle 模型在经济学中的特殊性与争议性（Dalrymple et al.，2003），本研究中的零膨胀模型（zero inflated model，ZIM）特指零膨胀泊松模型及零膨胀负二项模型。本研究旨在介绍零膨胀模型在社会科学领域的应用，通过零膨胀负二项模型对中国人工流产影响因素的分析实例，比较零膨胀模型与一般计数模型的分析结果，试图说明零膨胀模型是处理零值过多的计数资料的适宜工具。

二、零膨胀模型的技术原理

零膨胀模型的基本思想是把事件数的发生看成两种可能的过程：第一种过程对应零事件的发生，假定服从伯努里（Bernoulli）分布，个体

① Hurdle 模型是用来分析计数资料中存在过多零值的另一种统计方法，和零膨胀模型认为数据来源于非同质总体的思想不同，Hurdle 模型是先将样本中零值从数据集中分离出来，并在零处设定一个函数，再对所有正的计数过程用另一个函数确定。

取值只可能为零，且这个过程产生的零解释了数据中可能存在的过多零的原因；第二种过程对应事件数的发生过程，假定服从泊松或负二项分布，在这个过程中个体的取值可以为零或正的事件数。该模型将计数资料中的零看成“过多的零”（extra zero）和“真实的零”（true zero），并从零分段，对零计数和非零计数建立混合概率分布，对零部分和非零部分分别建立 logit 模型和一般计数模型（泊松或负二项），从而处理资料中过多零的问题。logit 部分主要回答协变量影响事件发生与否的问题，泊松或负二项模型部分主要回答协变量影响事件发生次数多寡的问题。例如，若对育龄妇女一年内的人工流产次数进行考察，则会发现人工流产次数为零值的比重较大，这种取值为零的情况可分为两组，一组为该时间内没有性生活或患不孕症的妇女（组 A），另一组为有性生活但没有人工流产经历的妇女（组 $\bar{A}$），这两组人工流产数都为零，但取零值的原因明显不同。在调查中通常并不知道谁属于组 A、谁属于组 $\bar{A}$。若符合组 A 的案例较多或组 $\bar{A}$ 中零值存在，则计数中会出现过多的零值，存在零膨胀现象。因此，可以根据具体问题的脉络（context）把原始数据集看成是由一个全零数据集和一个服从泊松分布/负二项分布的数据集组成的混合数据集，用一些特征变量对案例是否属于组 A 进行预测，然后将组 A 的案例排除掉，只对组 $\bar{A}$ 的案例进行泊松/负二项模型的计数建模，这就形成了零膨胀泊松模型或零膨胀负二项模型。换句话来说，零膨胀模型是一种针对零值较多且符合泊松分布/负二项分布的等离散（方差等于均值）或过离散（方差明显大于均值）数据进行的复合计数模型，[①]其中包含二分类的 logit 模型（对零值进行鉴别）及泊松模型/负二项模型。

（一）零膨胀计数模型的混合概率分布

零膨胀计数模型中，由零计数和非零计数集建立的混合概率分布为：

① 当数据存在过离散时，利用泊松模型，其估计仍能保持一致性，但估计的效率会有所下降，标准误会有所偏低，同时伴有较大的 Z 值（这种 Z 值往往是虚假的），采用负二项模型则可以校正泊松模型所导致的偏倚（bias）。在大部分估计值上，负二项模型与泊松模型的系数估计较为接近，系数的解释也与泊松模型相同。

$$y_i \sim \begin{cases} 0, p_i \\ g(y_i), 1 - p_i \end{cases} \tag{1}$$

p_i 表示个体来源于第一个过程伯努里分布的概率，表示数据中过多"0"的概率；$g(y_i)$ 表示个体来源于第二个过程，服从泊松或负二项分布。数据中的零一部分来源于那些从不可能发生事件的个体，概率为 p_i；另一部分来源于在泊松或负二项理论分布下没有发生事件的个体，概率为 $1 - p_i$，因此 $Y = y_i$ 的概率密度为：

$$\begin{cases} P(y_i = 0 \mid x_i) = p_i + (1 - p_i) g(0) \\ P(y_i \mid x_i) = (1 - p_i) g(y_i), y_i > 0 \end{cases} \tag{2}$$

若 p_i 的取值受个体自身协变量的影响，则 $p_i = F(w_i'\gamma)$，$F(\cdot)$ 称为零膨胀连接函数（zero-inflated link function），可选择 logit 或 probit：

$$p_i = \Lambda(w_i'\gamma) = \frac{\exp(w_i'\gamma)}{1 + \exp(w_i'\gamma)} \tag{3}$$

$$p_i = \Phi(w_i'\gamma) = \int_0^{w_i'\gamma} \frac{1}{\sqrt{2\pi}} \exp(-\mu^2/2)\, d\mu \tag{4}$$

式中 w' 为 $1 \times q$ 零膨胀自变量向量，γ 为 $q \times 1$ 零膨胀参数。

（二）零膨胀泊松模型（ZIP）

当 $g\left(y_i = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}\right)$ 时，称为零膨胀泊松模型，记作：

$$\begin{cases} P(y_i = 0 \mid x_i, w_i) = p_i + (1 - p_i) \dfrac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!} \\ = p_i + (1 - p_i)\exp(-\mu_i) \\ P(y_i \mid x_i) = (1 - p_i) \dfrac{\exp(-\mu_i)\mu_i^{y_i}}{y_i!}, y_i > 0 \end{cases} \tag{5}$$

x_i 和 w_i 可一致，也可不同。当在 ZIP 模型第一个过程中个体事件数取值为零的概率并不受个体自身因素影响，即零膨胀协变量 w_i 只包含常数项时，ZIP 模型比泊松模型多估计一个参数；当影响两个过程的协变量向量 x_i 和 w_i 相同时，整个 ZIP 模型需要估计的参数系数是泊松模型的两倍。

ZIP 模型 y_i 的期望和方差分别为 $E(Y) = \mu(1 - p)$，$Var(Y) = E(Y)(1 + \mu p)$。

(三) 零膨胀负二项模型(ZINB)

按照同样的思想,可将零膨胀泊松模型扩展到零膨胀负二项模型,当

$$g(y_i) = \frac{\Gamma(y_i + \alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i} \quad (6)$$

时,称为零膨胀负二项模型,记作:

$$\begin{cases} P(y_i = 0 \mid x_i, w_i) = p_i + (1 - p_i)(1 + \alpha\mu_i)^{-\alpha^{-1}} \\ P(y_i \mid x_i) = (1 - p_i)\dfrac{\Gamma(y_i + \alpha^{-1})}{y_i!\Gamma(\alpha^{-1})}\left(\dfrac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right)^{\alpha^{-1}}\left(\dfrac{\mu_i}{\alpha^{-1} + \mu_i}\right)^{y_i}, y_i > 0 \end{cases} \quad (7)$$

ZINB 模型 y_i 的期望和方差分别为:

$$E(Y) = \mu(1 - p),$$

$$Var(Y) = E(Y)[1 + \mu(p + \alpha)]。$$

当 $\alpha = 0$ 时,ZINB 模型等同于 ZIP 模型。

三、零膨胀模型实例分析

在实例分析中,笔者将针对 1988 – 2001 年"全国计划生育/生殖健康抽样调查"的原始数据,利用零膨胀模型分析中国已婚育龄妇女人工流产行为的影响因素,以此案例来说明该模型在社会科学实证研究中的具体应用。

(一) 研究目的、数据、变量

人工流产[①]作为衡量生殖健康的客观指标之一,不仅是健康问题,也是重要的人口社会问题,对其影响因素的探讨已成为人类行为研究中较为复杂且富有挑战性的领域之一。本实例分析的目的是考察近

① 1988 年全国生育节育调查方案中,将怀孕 3 个月以下采用人工方法终止妊娠的行为称为人工流产,4 – 6 个月采用人工方法终止妊娠的称为中期人工流产,怀孕 7 个月以上用人工方法终止妊娠的称为晚期人工流产。政府发布的统计数据中则将这 3 种人工流产合并,统称为人工流产。1997、2001 年的数据公报也类同。本案例效行之。

30 年间中国已婚育龄妇女人工流产的影响因素。所用数据为 1988、1997 与 2001 年国家人口计生委进行的"全国计划生育/生殖健康抽样调查"的原始数据，其调查内容主要涉及生育、避孕与人工流产等。在数据初期处理中，以 1997 年为基底，每波（wave）中随机选取已婚育龄妇女样本 14000 人（20－49 岁），剔除了不孕症、未婚及西藏（因问卷不一）的样本。研究中的因变量为已婚育龄妇女的人工流产次数，自变量为个体特征、社会经济及计划生育政策等变量。自变量的选取是根据文献回顾、社会理论支持及单因素 Wald test 所确定的，最终模型的选择是根据简约原则（parsimony principle）[①]及"嵌套模型"（nested models）的似然比检验（LR test）所确定的。[②]

（二）模型选择

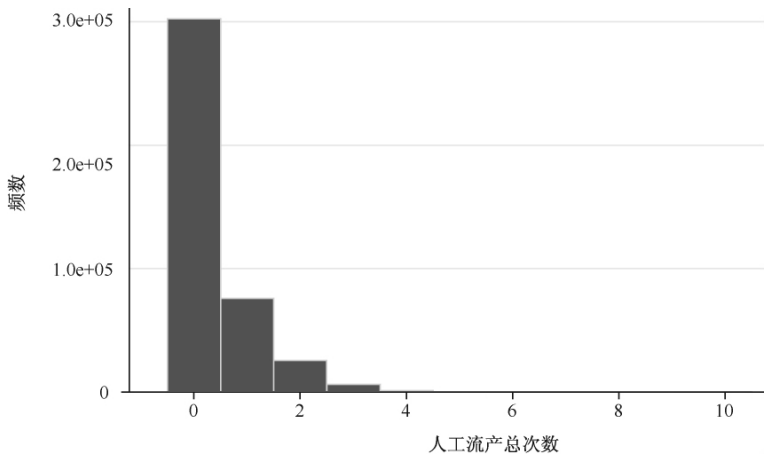
采用零膨胀模型的基本思路主要包括三个步骤：第一，观察计数资料是否存在零膨胀情况。第二，判断计数资料是否存在过离散。在判断 0 值较多的基础上，考察计数变量的均值与方差是否相等及 alpha 检验是否显著。若基本相等且 alpha 检验不显著（$p > 0.05$），则为等离散，服从泊松分布，宜采用零膨胀泊松模型；若均值与方差不等，且方差明显大于均值，alpha 检验也显著（$p < 0.05$），则为过离散，服从负二项分布，宜采用零膨胀负二项模型。第三，以 Vuong 检验决定模型的选择，并以图形比较人工流产观测数的实际分布与截距回归、泊松模型、负二项模型及零膨胀负二项模型拟合曲线的差异。

经过初步数据探索，发现中国已婚育龄妇女人工流产次数值为零的百分比明显多于其他取值，占 73.32%，取值为 1 的占 18.43%，呈偏态分布，存在零膨胀情况（图 1）。人工流产次数的均值为 0.3764938，方差为 0.5287292，可见方差明显大于均值，为过离散数据，且 alpha 检

① 模型选择的简约原则，即奥卡姆剃刀理论（Occam's razor principles），主要是说自然界的规律呈简约性。简约模型也意味着消除了一些冗余变量的干扰。冗余变量会导致重大的统计错误，如研究中会错过理论上有意义的发现、自由度的浪费、估计精度的下降等。

② 在普通非分层模型中，若含有高层级变量（如政策、地域等），因违背误差项独立同分布的假定，易导致估计有偏，理应采用分层泊松或 ZINB 模型（multilevel models）。但本案例对此数据做二层模型带随机效应的单因素方差分析时，发现组间的变异占总变异的比例（intra－class correlation，组内相关系数）不足 5%，意味着层二所能解释总方差的功能过小。当组间方差解释比例小于 5% 时，分层模型与普通单水平模型的估计结果基本一致（Raudenbush & Bryk，2001），故本案例未采用分层模型。

验呈现统计学显著意义($p < 0.05$),可认为人工流产次数的分布符合负二项分布,宜选用零膨胀负二项模型。



数据来源:1988 – 2001 年全国计划生育/生殖健康抽样原始数据。

图 1 1988 – 2001 年中国已婚育龄妇女人工流产次数的分布

以下将分别进行泊松截距回归、泊松回归、负二项回归及零膨胀负二项回归,并比较其结果。

1. 泊松截距回归

先进行截距回归,以便取得一个平均值与变量人工流产次数均值相等(即 0.3764938)的单变量泊松分布。但截距回归并没有考虑不同妇女在人工流产期望均值上的差异,需要进一步将模型扩展为包含自变量的泊松模型。

2. 泊松模型

根据单变量检验、社会理论支持及"嵌套模型"的似然比检验,最终将民族、存活子女数量、受教育程度、户口、地域、年龄、最小子女性别、政策(知情选择)等自变量纳入泊松模型,发现模型整体检验显著($p = 0.000$),并通过拟合优度检验(goodness of fit Chi-square)($p > 0.05$),说明回归拟合尚好,即加入自变量后的泊松模型可以用来拟合中国已婚育龄妇女的人工流产次数的实际分布,且自变量均呈显著统计学差异($p < 0.05$)。泊松模型已对各案例的人工流产数进行了估计,得出人工流产预测数的均值为 0.3764938 ± 0.2215334 及泊松模型

对每一种计数的平均预测概率。① 但泊松模型所得的均值与方差并不相等，方差大于均值（均值为 0.3764938，方差为 0.5287292），这提示本案例要进行过离散检验。

3. 负二项模型

利用 alpha 检验对人工流产数据的离散程度进行检验（likelihood - ratio test of alpha）时，发现 alpha 显著（p = 0.000），这说明该数据存在过离散。在过离散情况下，虽然泊松模型仍能保持估计的一致性，但估计效率会相应下降，标准误偏低，易得出虚假的 Z 分值（偏大）。此时，宜采用负二项模型来校正泊松估计。

对该数据进行负二项回归，可以发现模型整体检验显著（p = 0.000），且各自变量均呈显著统计学意义（$p < 0.05$）（见表 1）。

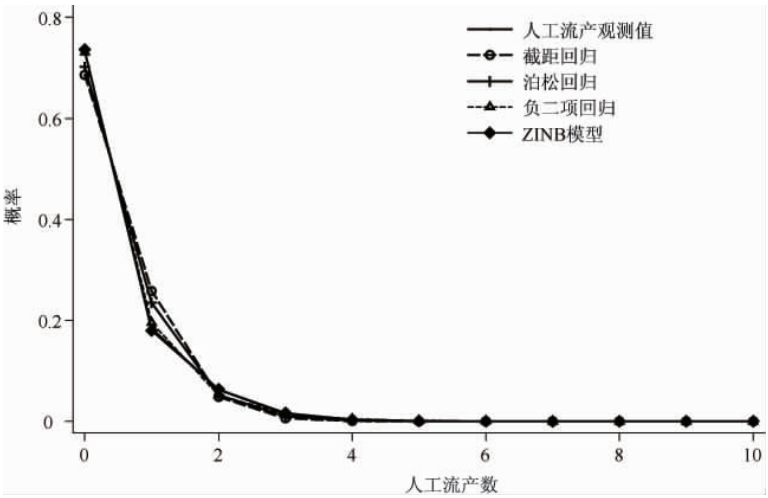
对负二项模型系数的解释与泊松模型相同，即采用发生率之比（e^b）来解释。通过负二项模型对各案例的人工流产数进行估计，可以得到人工流产预测数的均值为 0.377691 ± 0.2280975，更为接近于实际观测值均值。应该说，无论是模型估计还是理论解释，负二项模型对人工流产的考察都优于泊松模型，但零值左右的拟合与实际观测值还存在着较大差异（可从图形比较上得到验证，参见图 2），主要表现在 0 值的概率被低估、1 值的概率被高估。这是由于人工流产数据中零值比重较大（73.32%）所引起的（Powers，2009）。针对这种情况，宜采用零膨胀负二项模型。

4. 零膨胀负二项模型

将各自变量纳入零膨胀负二项模型，发现模型整体检验显著（p = 0.000），民族、子女数量、年龄、受教育程度、所在地域（经济状况）、户口、最小子女性别、政策（知情选择）等变量均呈显著统计学意义（$p < 0.05$），而膨胀因子中民族、子女数量、年龄、受教育程度、所在地域、户口、最小子女性别等变量也均呈显著统计学意义（$p < 0.05$）（表 2）。

同时，利用图形对比 ZINB 模型与泊松截距模型、泊松模型、负二项模型与实际观测数据拟合的差异（图 2），可以发现零膨胀负二项模型拟合曲线比其他三种模型更为逼近于实际观测值的分布曲线，呈基本重合态势，这说明利用 ZINB 模型所得的预测值与实际观测值更为接近。

① 需要说明的是，在泊松模型中，回归系数是期望人工流产次数的对数（log-rate）。



数据来源:1988－2001 年全国计划生育/生殖健康抽样原始数据。

图 2 人工流产次数的分布、截距回归、泊松模型、负二项模型与 ZINB 模型比较

除了以图形观察 ZINB 模型预测值与实测值间的拟合程度,还可以通过 Vuong 值检验比较 ZINB 与负二项模型的优劣。若 $p < 0.05$,则说明 ZINB 优于负二项模型。本案例中 Vuong 值检验是显著的($p < 0.05$),说明零膨胀负二项模型优于负二项模型。同时,也把拟合优度作为参考指标,并对原假设"实测数据分布与零膨胀模型预测分布之间无差异"进行检验。若接受原假设($p > 0.05$),说明模型拟合较好;若拒绝原假设($p < 0.05$),则说明模型拟合较差。本案例对拟合优度进行卡方检验时,发现极不显著(goodness-of-fit Chi-square = 87862.43, $p = 1.000$),说明 ZINB 模型拟合较好。

从上述一系列指标或拟合图形上来看,ZINB 模型不仅较适合本案例数据,且明显优于泊松模型与负二项模型。

值得说明的是,对零膨胀负二项模型系数的解释包含两个部分,一部分是负二项回归(组$\overline{A}$的回归系数),一部分是 logit 回归(组 A 的回归系数)。零膨胀模型中负二项回归部分系数的解释与一般负二项模型相同,即回归系数是平均(期望)人工流产次数的对数(log-rate),一般以发生率之比(incidence rate ratio, IRR)即 e^b 来解释。若针对连续自变量,可以用 IRR 系数的百分比变化(percent change in the IRR)来

表 1　　中国已婚育龄妇女人工流产次数负二项模型结果

变量		回归系数（b）	发生率之比（e^b）	%	% StdX	p 值
民族（以汉族为参照组）	非汉族	-.2035（.3029）	.8159（.9402）	-18.4	-6.0	.000
存活子女数量（以无子女为参照组）	一个子女	1.4758（.4622）	4.3747（1.9781）	337.5	97.8	.000
	两个	1.2475（.4487）	3.4816（1.7502）	248.2	75.0	.000
	多个	.7824（.4816）	2.1868（1.4576）	118.7	45.8	.000
受教育程度（以小学及以下为参照组）	初中	.2140（.4335）	1.2386（1.0972）	23.9	9.7	.000
	高中	.2027（.3179）	1.2247（1.0666）	22.5	6.7	.000
	大学及以上	.1497（.1289）	1.1614（1.0195）	16.1	10.9	.000
户口（以农村为参照组）	城市	.5049（.4308）	1.6567（1.2430）	65.7	24.3	.000
地域（以东部为参照组）	中部地区	-.1524（.4583）	.8587（.9326）	-14.1	-6.7	.000
	西部地区	-.0268（.4600）	.9736（.9878）	-2.6	-1.2	.000
年龄	20-49	.0488（.0383）	1.0500（1.4511）	5.0	45.1	.000
最小子女性别（以男孩为参照组）	女孩	-.0230（.4984）	.9772（.9886）	-2.3	-1.1	.000
知情选择（以未知情为参照组）	部分知情选择	-.0076（.1786）	.9925（.9987）	-.8	-.1	.637
	全面知情选择	-.2582（.2800）	.7725（.9303）	-22.8	-7.0	.000
截距项		-3.9737（.0308）				

注：（1）模型整体显著性 p=0.000（log likelihood = -315116.06）。（2）案例数为 412291。（3）括号内为标准误。（4）其中% 为其他自变量不变时，X 每变化一个单位而引起计数因变量的百分比变化；% stdX 则是指其他不变时，X 每增加一个标准差而引起计数因变量变化的百分比。（5）alpha 显著（lnalpha = -0.4227564，alpha = 0.6552382，LR test of alpha=0：p=0.000）。

数据来源：1988-2001 年全国计划生育/生殖健康抽样原始数据。

表 2 已婚育龄妇女人工流产影响因素的 ZINB 模型分析结果

变量		回归系数(b)	发生率之比(e^b)	%	% StdX	p 值
民族(以汉族为参照组)	非汉族	-.0506 (.3029)	.9507 (.9848)	-4.9	-1.5	.001
存活子女数量(以无子女为参照组)	一个子女	.9866 (.4622)	2.6822 (1.5778)	168.2	57.8	.000
	两个	.8057 (.4487)	2.2383 (1.4355)	123.8	43.5	.000
	多个	.4362 (.4816)	1.5469 (1.2338)	54.7	23.4	.000
受教育程度(以小学及以下为参照组)	初中	.0477 (.4335)	1.0489 (1.0209)	4.9	2.1	.000
	高中	-.0228 (.3179)	.9775 (.9928)	-2.3	-.7	.064
	大学及以上	-.0330 (.1289)	.9675 (.9958)	-3.2	-.4	.112
户口(以农村为参照组)	城市	-.0446 (.4308)	.9564 (.9810)	-4.4	-1.9	.000
地域(以东部为参照组)	中部	.1402 (.4583)	1.1505 (1.0663)	15.0	6.6	.000
	西部	.2239 (.4600)	1.2510 (1.1085)	25.1	10.9	.000
年龄	20-49	.0407 (.0383)	1.0416 (1.3651)	4.2	36.5	.000
最小子女性别(以男孩为参照组)	女孩	-.0146 (.4984)	.9855 (.9927)	-1.5	-.7	.011
知情选择(以未知情为参照组)	部分知情选择	-.0397 (.1786)	.9611 (.9929)	-3.9	-.7	.009
	全面知情选择	-.2768 (.2800)	.7582 (.9254)	-24.2	-7.5	.000
截距项		1.9293 (.1083)				
膨胀因子(inflate)						
民族(以汉族为参照组)	非汉族	.3082 (.3029)	1.3610 (1.0978)	36.1	9.8	.000
存活子女数量(以无子女为参照组)	一个	-1.5166 (.4622)	.2195 (.4961)	-78.1	-5.4	.000
	两个	-1.2860 (.4487)	.2764 (.5616)	-72.4	-43.8	.000
	多个	-1.0429 (.4816)	.3524 (.6052)	-64.8	-39.5	.000

续表 2

变量		回归系数(b)	发生率之比(e^b)	%	% StdX	p 值
受教育程度(以小学及以下为参照组)	初中	-.4495 (.4335)	.6379 (.8229)	-36.2	-17.7	.000
	高中	-.9033 (.3179)	.4052 (.7504)	-59.5	-25.0	.000
	大学及以上	-1.5138 (.1289)	.2201 (.8228)	-78.0	-17.7	.000
户口(以农村为参照组)	城市	-2.1996 (.4308)	.1108 (.3877)	-88.9	-61.2	.000
地域(以东部为参照组)	中部	.8544 (.4583)	2.3500 (1.4793)	135.0	47.9	.000
	西部	.7296 (.4600)	2.0743 (1.3988)	107.4	39.9	.000
年龄	20-49	-.0264 (.0383)	.9740 (.8177)	-2.6	-18.2	.000
截距项		1.9293 (.1083)				

注:(1)模型整体显著性 p=0.000(log likelihood=-310169.30)。(2)案例数为412291,其中0值案例数302303。(3)括号内为标准误。(4)alpha显著(lnalpha=-0.4227564,alpha=0.6552382,LR test of alpha=0:p=0.000),说明过离散严重。(5)Vuong值检验(Vuong test of ZINB vs. standard Negative Binomial Poison Regression)显著(p=0.000),说明ZINB模型比负二项模型更好。

数据来源:1988-2001年全国计划生育/生殖健康抽样原始数据。

考察自变量的影响,即在控制其他变量时,考察自变量每增加一个单位或每增加一个标准差给因变量所带来的百分比变化(表2中%一栏)。这种方法得到的值是一种标准化系数,可以直接测量与比较所有自变量对人工流产次数影响的相对重要性,"% StdX"的绝对值越大,其影响就越大。① 表2中显示的"膨胀因子(inflate)"栏,为零膨胀模型中logit模型中的回归系数,其解释与普通二分类变量的logit模型相同。

在控制其他变量的条件下,通过零膨胀负二项模型对1988-2001年间已婚育龄妇女人工流产行为分析的结果,可以看出:人工流产的选择存在整体性民族差异,非汉族妇女人工流产的发生率之比是汉族妇

① 这种标准化比较曾受到质疑,被认为只能针对连续自变量,并不适用于分类变量,因为分类变量的标准差意义不明确(Long & Freese,2005)。

女的0.9507倍,即非汉族妇女发生人工流产的期望值比汉族妇女低4.93%(4.93% =1 -0.9507);伴随受教育程度的提高,育龄妇女期望人工流产次数的发生率之比在降低,初中、高中、大学及以上受教育程度的育龄妇女,其期望人工流产次数分别是小学及以下文化妇女的1.0489倍、0.9775倍、0.9675倍;城市育龄妇女的期望人工流产次数是农村育龄妇女的0.9564倍,即城市妇女的人工流产发生率之比要比农村妇女低4.36%(0.0436 =1 -0.9564);西部妇女的期望人工流产次数最高,为东部妇女的1.2510倍,其次是中部,为东部妇女的1.1505倍,表明随着地区由东向西的渐移,人工流产发生的可能性在逐步提高;已婚育龄妇女每增加一岁,其人工流产的发生可能性就会增加4.16%(4.16% =1.0416 -1),表明人工流产发生的可能性随育龄妇女年龄的增加而增加;一孩、二孩与多孩妇女的期望人工流产次数分别是零孩妇女的2.6822倍、2.2383倍、1.5469倍,表明已婚育龄妇女人工流产发生的可能性随着子女数量的增加而有所降低;"最小子女为女孩"的育龄妇女,其期望人工流产次数是"最小子女为男孩"妇女的0.9855倍,即当最小子女是女孩时,育龄妇女人工流产的发生率之比低于"最小子女为男孩"的妇女1.45%(1.45% =1 -0.9855);"部分知情选择"时期与"全面知情选择"时期的育龄妇女人工流产发生率之比分别是"未知情选择"时期的0.9611倍、0.7582倍,即在"部分知情选择"时期,育龄妇女人工流产发生率之比低于"未知情选择"时期3.89%(0.0389 =1 -0.9611);在"全面知情选择"时期,育龄妇女的人工流产发生率之比低于"未知情选择"时期24.18%(0.2418 =1 -0.7582)。这表明,随着知情选择由"部分知情选择"到"全面知情选择"的逐步深入,已婚育龄妇女人工流产发生的可能性在逐步降低。

5. 模型小结

本案例先后进行了泊松截距模型、泊松模型、负二项模型与零膨胀负二项模型的实际比较,发现零膨胀负二项模型的拟合曲线比其他三种模型更为逼近于真实观测值的分布,呈基本重合态势。同时,Vuong检验与拟合优度的卡方检验等都表明零膨胀负二项模型优于一般模型,数据结果也比其他模型更切合实际,更具有理论解释性。由此可见零膨胀负二项模型在处理零过多的计数资料上具有较强的优越性。

6. 案例的进一步分析

通过零膨胀负二项模型的结果可以发现,虽然长期以来影响我国

已婚育龄妇女人工流产行为的因素是多方面的，但就中国近 30 年来整体的人工流产历程而言，个体特征、社会经济及计划生育政策等是较为重要的影响因素。即不同的个体特征（如民族、年龄、户口、现有子女数、子女性别等）、社会经济特征（如地域、受教育程度等）及计划生育政策强度（如是否开展知情选择、知情选择程度）等都对已婚育龄人群的人工流产行为有较为显著的影响。也就是说，个体的人工流产行为是政策、社会经济及个人特征综合作用的结果。

民族、地区、户口等变量对人工流产的影响无不凸显着宏观计划生育政策的渗透与调停的烙印。国家的宏观计划生育政策通过个体特征，并与个体特征形成合力，作用于个体的人工流产行为。本案例中汉族、农村、西部地区、受教育水平低的妇女，其人工流产发生的可能性分别高于非汉族、城市、中东部地区及受教育水平高的妇女，其原因可能是在这几个变量的作用中有一种共性的隐性机制链在发挥着干预效能，即计划生育政策。如我国的计划生育政策对非汉族人群相对宽松，其政策生育率为 $0\rightarrow\infty$，而汉族则为 $1\rightarrow1.5$，汉族妇女一旦怀孕，其政策性人工流产的可能性相对较高；受教育水平低的妇女多为农村户口，而计划生育政策在农村及西部地区的执行力度普遍较强，“一孩上环、二孩结扎、三孩流产”的单调的、强硬的避孕模式在广大农村及西部地区依然存在，会导致受教育水平低、农村及西部地区的妇女人工流产的可能性相对较高。但往往具有这些特征的人群存在着较强的、未满足的生育数量意愿和较为普遍的男孩偏好（Poston，2002；陈卫，2005），这就容易引发个体意愿与国家政策之间的冲突与博弈。

更进一步来说，即使育龄个体的生育意愿与国家生育政策有所对抗与冲突，处于弱势的育龄个体在其生育行为中也并不能达到完全的社会理性选择，进行选择时遵循的也不再是“利益最大化”原则，而是“满意”原则（Simmon，1957）。尤其当个体的自由行动权受限时，个体就只能寻求一种与国家利益共享的均衡路径来达致自我的“满意”，这种路径往往要借助于一些个体妥协或折衷的“弱武器”来实现（Scott，1985），如通过增加意外妊娠或性别选择性人工流产等手段来实现“多生”或“男孩偏好”。这一点也在数据分析中得到了部分佐证。

其一，“现有子女数”越多，育龄妇女人工流产发生的可能性就越低。可能的原因是随着存活子女的增多，育龄妇女家庭结构相对稳固，生育愿望基本满足。同时，在计划生育政策的直接干预下，子女数量越

多的妇女越有可能采用绝育等长效医控型避孕措施,这就意味着避孕失败的可能性会减少,其人工流产发生的可能性也随之降低。

其二,“最小子女是女孩”的育龄妇女,其人工流产的可能性低于“最小子女是男孩”的妇女。这似乎与以往有关避孕研究的结论相左,如避孕研究曾表明,“最小子女是女孩”的妇女采用短效避孕措施的可能性大大高于“最小子女是男孩”的妇女(王存同,2009a)。这意味着,基于短效避孕措施容易导致高意外妊娠率的事实,又囿于相对严格的生育控制政策,“最小子女为女孩”的妇女,其人工流产的可能性理应高一些,但本案例却发现“最小子女为女孩”的妇女选择人工流产的可能性较低,个中原因可能是这部分妇女选择了另一种路径,即计划外生育。这种推测在相关的定性调查中得到了验证。在甘肃定西及河南南阳、山东鲁南部分地区进行的实地调查中发现,虽然国家长期号召并鼓励“一对夫妇只生一个孩子”、“生男生女都一样”,但传统的“多子多福”、“重男轻女”的观念还根深蒂固。纵使这种生育观念赖以存在的经济基础和制度保障已不复存在,但还是表现出了较强的滞后影响,尤其在农村,留存了传统生育的深刻记忆。当地“最小子女为女孩”的夫妇,尤其是“独女户”,多有生育男孩的愿望,在当下的避孕措施选择时,多倾向于采用避孕套或其他短效自控型避孕措施。这样不但可以自行控制避孕与否,还可以使当地计生干部有“脸面”(若避孕率不达标,计生干部要被扣奖金或降职),多能得以“避孕失败”而“意外妊娠”(当地计生部门计划生育工作年终总结语)。若地方计生干部“睁只眼、闭只眼”,育龄妇女则可以与地方计生干部完成“沉默的共谋”,逃脱政策性人工流产的惩罚,不但能成功地实现计划外生育,还可在多生育的基础上增加生男孩的机率。据说,她们“这一招是跟超生娃娃的大姐们那里学来的”,并已成为当地公开的“秘密”(王存同,2009b)。

除了对社会经济及个体特征的影响进行考察外,本案例还对计划生育政策(“知情选择”变量)的影响进行了直接度量。知情选择作为我国计划生育改革及贯彻“以人为本”理念的标志性产物,也是我国由单调的“一孩上环、二孩结扎、三孩流产”强制型避孕模式向“自主选择”多元化模式转变的分水岭(王存同,2009b),知情选择开展前后及其进程对人工流产的影响反映的恰恰是不同政策力度所引起的效果差异。本案例定量分析发现,计划生育政策的确对人工流产有着较为明显的干预作用与指导性影响,表现为伴随我国知情选择政策的全面展

开，育龄妇女人工流产发生的可能性逐步降低，并在各个孩次上都有明显的体现。这一结果与国家统计报表基本一致，即“全面知情选择”时期总人工流产率（27.3%）（潘贵玉，2003）低于“部分知情时期”（32.3%）（蒋正华，2000）。应该说，人工流产可能性的下降至少在某种程度上说明我国知情选择的开展产生了一定的正面效果，其中原因可能是随着知情选择的逐步展开，妇女有权并能自主选择适宜的避孕措施，相应提高了避孕效果，降低了意外妊娠的风险，从而减少了人工流产的发生，这与知情选择推广的初衷相吻合。

当然，这种统计结果囿于现有数据对知情选择考察的局限，只是在控制其他变量的情况下大致反映知情选择对人工流产的影响。人工流产的减少，也可能意味着计划外出生的增加，但没有确凿的国家层面上的有关数据，目前还不能对这一假设进行量化验证。

四、讨　论

本研究以中国已婚育龄妇女人工流产影响因素的探索为例，通过比较零膨胀模型与一般计数模型的差异，试图说明零膨胀模型是处理零值过多计数资料的适宜工具，并实例示范该模型在社会科学领域中的具体应用。通过案例分析，发现已婚育龄妇女个体的人工流产行为是计划生育政策、社会经济及个人特征综合作用的结果。其中，计划生育政策长期占据干预性与主导性地位，但随着时代的变迁，社会经济与个人特征的作用日渐凸显。同时，也发现多数健康个体避孕成功而没有人工流产的经历是导致数据中较多零值出现的原因，即在人工流产次数的分析中，那些身体健康或者即使避孕失败但由于种种原因而没有进行人工流产的次数报告为零次。这种可解释的现象及系列的回归分析、拟合比较、Vuong 值检验等都支持本研究选择零膨胀模型。相对一般计数模型，零膨胀模型不但提供了更符合实际情况的结果解释，而且通过其中的 logit 回归还有可能判断零事件的具体来源。

有关零膨胀模型在社会科学实证研究中的具体应用，本研究认为有必要再次强调及注意的问题是：

首先，应用该模型之前要辨清数据中是否存在零过多的情况，即判断数据中的零值是否超过了一般计数模型（泊松/负二项模型）的预测

能力？这可利用描述性统计分析观察零值的百分比或采用 alpha 检验。若数据出现较多的零计数，提示选用零膨胀模型。

第二，要判断数据的均值与方差是否等同。等离散时可选用零膨胀泊松模型，过离散时可选用零膨胀负二项模型。

第三，应用零膨胀模型时，要判断该模型是否优于一般泊松或负二项模型。这可利用本研究示范的拟合分布图形比较、拟合优度检验及 Vuong 值检验。其中，Vuong 检验在复合模型（ZIP、ZINB）和一般计数模型（泊松/负二项模型）的比较分析中有相当高的检验效能。

第四，零膨胀模型的实质是将零值分成两个不同质的亚群，即一组中的个体根本不会发生相应的事件，另一组中的个体可能发生事件并假定服从泊松/负二项分布，将数据中的零看成“过多的零”和“真实的零”，前者来自于第一个过程，后者来自第二个过程。但需要注意的是，对于一个观察值为零的个体，在实际中有时并不能明确地判断此个体来源于上述的哪个过程（曾平，2009）。最好能借助于定性研究，对零的来源有明确的认识，并寻找产生过多零事件的原因和机制，这对研究者建立模型与合理解释大有裨益。

第五，零膨胀模型属于二项分布和计数分布的混合分布，对于解决较多的零值计数资料有一定的普适性。零膨胀模型对数据分别拟合 logit 回归和一般计数模型（泊松/负二项模型），logit 回归主要回答协变量影响事件发生与否的问题，而计数模型主要回答协变量影响事件发生次数多少的问题。

第六，零膨胀模型的参数估计可采用极大似然法，作为混合指数族分布，其对数似然函数比单纯的广义线性模型对数似然函数要复杂得多。一般情况下，BHHH 估计量容易计算也容易达到迭代的收敛。

第七，当同样的自变量用于零膨胀模型中的 logit 与泊松/负二项模型时，有时会发现 logit 与泊松/负二项模型的相应系数呈相反方向（Long & Freese，2005）。这是因为 logit 模型预测了永远为零组的归属，所以正系数代表了更可能为零产出（本例中的产出即人工流产次数），而计数回归预测了产出数量，所以负系数代表了较低的产出。

第八，当样本量较小时，零膨胀模型假设检验的结果并不太可靠或迭代不能收敛，宜采用精确检验（exact test）。

最近 10 年社会科学领域中有关零膨胀现象的分析方法取得了长足的进步。如多层零膨胀模型分析框架及基于结构方程模型的零膨胀

潜变量分析框架(Heck，2001；Muthen，1997；Raudenbush & Bryk，2001)及 Bootstrap 抽样法(Fan，2003)的应用等。这方面的一些最新进展以及针对性分析软件的出现使得以前认为不可能建立的模型已成为研究者现在常用工具的一部分。

参考文献：

陈卫，2005，《中国的人工流产——趋势、模式与影响因素》，北京：科学技术文献出版社。

郭志刚，2008，《高级社会统计专题》(授课 PPT)。

国家计生委计划财务司，2003，《全国计划生育调查资料汇编》，国家计生委内部资料。

蒋正华，2000，《1997 年中国生育率抽样调查数据集》，北京：中国人口出版社。

刘爽，2005，《中国育龄夫妇的生育"性别偏好"》，《人口研究》第 3 期。

潘贵玉，2003，《2001 年全国计划生育生殖健康调查数据集》，北京：中国人口出版社。

王存同，2009a，《中国已婚育龄人群避孕措施选择的偏好与影响因素分析》，《人口与发展》第 1 期。

——，2009b，《中国计划生育下的避孕行为：1960－2004》，北京大学博士学位论文。

吴晓刚，2007，《定量社会分析》(Quantitative Analysis in Social Science)(授课讲义)。

曾平，2009，《零过多计数资料回归模型及其医学应用》，山西医科大学硕士学位论文。

曾平、刘桂芬、曹红艳，2008，《零膨胀模型在心肌缺血节段数影响因素研究中的应用》，《中国卫生统计》第 5 期。

Bilgic，Abdulbaki & Wojciech J. Florkowski 2007，"Application of A Hurdle Negative Binomial Count Data Model to Demand for Bass Fishing in the Southeastern United States." *Journal of Environmental Management* 83(4).

Cameron，A. C. & P. K. Trivedi 1998，*Regression Analysis of Count Data*. New York：Cambridge University Press.

Dalrymple，M. L.，I. L. Hudson & R. P. K. Ford 2003，"Finite Mixture，Zero-inflated Poisson and Hurdle Models with Application to SIDS." *Computational Statistics & Data Analysis* 41(3－4).

Fan，X. 2003，"Using Commonly Available Software for Bootstrapping in Both Substantive and Measurement Analyses." *Educational and Psychological Measurement* 63(1).

Greene，W. 1994，"Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models." Working Paper (EC－94－10)，Department of Economics，New York University.

Gigerenzer，Gerd & Reinhard Selten (eds.) 2001，*Bounded Rationality*：*The Adaptive Toolbox*. Cambridge，MA：MIT Press.

Hamilton，C. 2006，*Statistics with Stata* (*Updated for Version* 9). Belmont，CA：Thomson－Brooks/Cole，Thomson.

Heck，R. H. 2001，"Multilevel Modeling with SEM." In G. A. Marcoulides & R. E. Schumacker (eds.)，*New Developments and Techniques in Structural Equation Modeling*. Mahwan，N. J.：

Lawrence Erlbaum Associates.

Johnson, N. L. & S. Kotz 1969, *Distributions in Statistics: Discrete Distributions* (2nd ed.). Boston: Houghton Mifflin Company.

Lambert, D. 1992, "Zero-inflated Poisson Regression, with An Application to Defects in Manufacturing." *Technometrics* 34(1).

Liao, Tim Futing 2002, *Statistical Group Comparison*. Hoboken, N. J.: Wiley & Sons Publication.

Long, J. S. & J. Freese 2005, *Regression Models for Categorical Outcomes Using Stata* (2nd ed.). College Station, TX: Stata Press.

Mullahy, J. 1986, "Specification and Testing of Some Modified Count Data Models." *Journal of Econometrics* 33(3).

—— 1997, "Heterogeneity, Excess Zeros, and the Structure of Count Data Models." *Journal of Applied Econometrics* 12(3).

Muthen, B. 1997, "Latent Variable Modeling of Longitudinal and Multilevel Data." *Sociological Methodology* 27(1).

Poston, Dudley L. 2002, "Son Preference and Fertility in China." *Journal of Biosocial Science* 34(3).

Powers, Daniel A. & Yu Xie 2008, *Statistical Methods for Categorical Data Analysis* (2nd ed.). Boston: Emerald Group Publishing.

Powers, Daniel A. 2009, *Categorical Data Analysis*. Handouts in Peking University & Michigan University Summer Course.

Raudenbush, S. W. & A. S. Bryk 2001, *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). California: Sage Publication.

Scott, J. C. 1985, *Weapons of the Weak*. London: Yale University Press.

Shiferaw, Gurmu 1998, "Generalized Hurdle Count Data Regression Models." *Economics Letters* 58(3).

Simon, H. 1957, *Administrative Behavior*. New York: Free Press.

Singer, J. D. & J. B. Willett 2003, *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford University Press.

Van den Broek, J. 1995, "A Score Test for Zero Inflation in A Poisson Distribution." *Biometrics* 51(2).

Vuong, Q. H. 1989, "Likelihood Ratio Tests for Model Selection and Non – nested Hypotheses." *Econometrica* 57(2).

Welsh, A. H., R. B. Cunningham, C. F. Donnelly & D. B. Lindenmayer 1996, "Modeling the Abundance of Rare Species: Statistical Models for Counts with Extra Zeros." *Ecological Modeling* 88(1 – 3).

作者单位:中央财经大学社会发展学院社会学系

责任编辑:杨 可

Abstract: This study introduced Zero-inflated Poisson/Negative Binomial modeling (ZIP/ZINB) technique and its applications to the research of social sciences. Hoping to help readers to better understand and feel comfortable to use these analytical tools, the author took an example about induced abortion in China to demonstrate major concepts, analytical strategies, and advantages of ZIP/ZINB modeling technique by comparing the analytical results of ZIP/ZINB with the Poisson/NB results.

Abstract: The post-war Japanese intellectual community, which grew out of associations and publications, is a mechanism of social introspection. The way it affects the society is different from that of the establishment. The value orientation of post-war Japanese intellectual community bears an internal connection to Japan´s economic growth and social change, in both theory and methodology, it has gone from Modernism, Marxism, Anti-Modernism, to a mass-based perspective on histories and social histories. This thesis attempts to examine the transformation of post-war Japanese intellectual communities by tracing the historical use of terminologies like "the people", "the general public" and "the citizen", thereby argues for a specific approach to analyze post-war Japanese intellectual communities and their implications.

Abstract: This paper tries to explore the situation of the concept of space in the history of western sociology, and lays open the sources of thought, the basic characteristics and the theoretical meaning of the turn of space in contemporary western sociology. This turn understood the concept of space as a social ontological concept, absorbed and transcended the objective and subjective theories of space. It criticized objectivism, universalism and historicism. But its perspectives of space tangled with Cartesian dualism in different degrees, and have been limited by the dualism of time and space. Hereby the author forecasts a reconstruction of the concept of space.