

应用简报

Logistic回归模型在人口问题中的应用 *

赵 旭 陈立萍 程维虎

(北京工业大学应用数理学院, 北京, 100124)

摘要

处理和分析“全国第五次人口普查”数据, 主要解决以下两个问题: 一是分析影响妇女超生的因素和影响程度; 二是预测妇女超生概率, 样本回代正确率98%以上.

关键词: Logistic回归, 人口普查, 因子分析, 分析与预测.

学科分类号: O212.

§1. 引言

在我国, 实行计划生育, 控制人口数量, 提高人口素质是实现现代化建设宏伟目标的重大决策. 中国人口发展经历了20世纪50年代和60年代无计划的高速增长、70年代生育水平的大幅度下降和80年代的徘徊波动后, 人口出生率在很短的时间内下降了一半以上, 并基本上完成了由传统的高出生、高死亡、高增长的人口模型向低出生、低死亡、低增长的人口模型转变. 由于人口转变和低生育水平并非完全靠社会经济的发展而实现自然转变, 在低生育水平的背后仍然有强大的反弹势头, 任何外部环境的变化都可能引发生育水平的波动. 我国的人口形势不容乐观, 有必要分析影响妇女超生的因素和影响程度以及预测超生概率.

§2. 模型选择

2.1 数据来源及定义变量

数据来源: 2000年全国第五次人口普查0.95%的抽样数据.

因变量: 当妇女所生子女数大于等于3时, 认为超生, y 取1; 否则取0.

自变量: 年龄 X_1 ; 住房面积 X_2 ; 户口性质 X_3 : 非农业取1, 否则取0; 民族 X_4 : 少数民族取1, 否则取0; 学历 X_5 : 未上过学取0, 小学取1, 初中取2, 高中取3, 高中以上取4; 流动程

*北京市自然科学基金(1154005)和高等学校博士学科点专项科研基金(20131103120027)资助.

本文2013年12月30日收到, 2015年4月25日收到修改稿.

doi: 10.3969/j.issn.1001-4268.2015.06.005

度 X_6 : 未流动取0, 县内流动取1, 跨县流动取2, 跨省流动取3; 住宅外墙墙体材料 X_7 : 竹木取0, 砖石取1, 钢筋混凝土取2, 其他取3; 婚姻状况 X_8 : 未婚取0, 初婚取1, 离婚取2, 再婚取3; 工作状况 X_9 : 有工作取1, 否则取0.

2.2 模型选择

Logistic回归模型是对二分类因变量进行回归分析时应用最为普遍的多元化分析方法. 模型如下(王济川和郭志刚, 2001):

$$\ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \sum_{l=1}^k \beta_l x_{li},$$

其中 $p_i = P(y_i = 1|x_{1i}, x_{2i}, \dots, x_{ki})$, $y_i = 1$ 表示第*i*个反应变量发生, $y_i = 0$ 则表示未发生.

§3. 因子分析

在“五普”数据中, 变量之间存在一定的相关关系. 为实现降维和消除变量间相关性这两个目的, 首先选择对数据进行因子分析. 经过反复试验, 最后入选的变量为在原有9个自变量的基础上, 加入互动项: 年龄×民族、年龄²×民族、年龄³×民族、工作状况×户口性质、住宅外墙墙体材料×婚姻状况, 删去工作状况、年龄、婚姻状况后, 共11个变量.

3.1 确定公因子

可求出前4个公因子的累积贡献率为

$$\frac{1}{11} \sum_{j=1}^4 \sum_{i=1}^{11} l_{ij}^2 = 82.867\%,$$

超过了80%, 这说明前4个公因子便可解释原有变量信息. 因此, 可将公因子个数确定为4.

确定公因子个数后, 为增强公因子对变量的解释力, 可对因子矩阵进行旋转, 在此选择方差极大法进行旋转(Johnson和Wichern, 2001). 根据旋转后的因子成分矩阵可以得出:

- (1) 与第1个公因子 F_1 有关的变量有: 民族、年龄×民族、年龄²×民族、年龄³×民族. 因此, 我们大致把第1个公因子解释为个人背景因子.
- (2) 与第2个公因子 F_2 有关的变量有: 学历、户口性质、工作状况×户口性质. 因此, 我们大致把第2个公因子解释为个人经历状况因子.
- (3) 与第3个公因子 F_3 有关的变量有: 住宅外墙墙体材料、住宅外墙墙体材料×婚姻状况. 因此, 我们大致把第3个公因子解释为个人经济状况因子.
- (4) 与第4个公因子 F_4 有关的变量有: 住房面积、人员流动程度. 因此, 我们大致把第4个公因子解释为个人生活稳定状况因子.

3.2 计算因子得分

从因子得分矩阵可以求出因子得分函数:

$$\begin{aligned}F_1 = & -0.015X_2 + 0.145X_3 + 0.943X_4 - 0.195X_5 - 0.052X_6 + 0.017X_7 \\& + 0.990X_1X_4 + 0.986X_1^2X_4 + 0.953X_1^3X_4 + 0.147X_3X_9 + 0.031X_7X_8, \\F_2 = & 0.290X_2 + 0.870X_3 - 0.091X_4 - 0.677X_5 - 0.234X_6 + 0.318X_7 \\& - 0.099X_1X_4 - 0.099X_1^2X_4 - 0.096X_1^3X_4 + 0.843X_3X_9 - 0.337X_7X_8, \\F_3 = & 0.041X_2 + 0.240X_3 - 0.032X_4 - 0.235X_5 + 0.005X_6 + 0.919X_7 \\& - 0.040X_1X_4 - 0.045X_1^2X_4 - 0.047X_1^3X_4 + 0.244X_3X_9 + 0.912X_7X_8, \\F_4 = & 0.692X_2 - 0.165X_3 - 0.003X_4 + 0.201X_5 - 0.741X_6 + 0.045X_7 \\& + 0.008X_1X_4 + 0.017X_1^2X_4 + 0.023X_1^3X_4 - 0.072X_3X_9 + 0.044X_7X_8.\end{aligned}$$

§4. 拟合模型

通过多次试验发现, 将年龄分为15–29岁和30–50岁两段后, 拟合出的Logistic回归模型最为理想。这样划分满足其特有的时代背景, 在“第五次人口普查”时, 30–50岁的妇女于60–80年代适孕, 60–80年代是人口出生率从无计划的高速增长到大幅度下降的时期; 15–29岁的妇女于80年代以后适孕, 80年代以后是人口出生率徘徊波动的时期。因此, 以80年代为界限对年龄划分, 80年代以前, 生育水平动荡不安, 即对应“五普”时30–50岁妇女数据; 80年代以后, 生育水平徘徊波动, 即对应“五普”时15–29岁妇女数据。

用因子分析得到的4个公因子(个人背景因子、个人生活稳定状况因子、个人经历状况因子、个人经济状况因子)和一个互动项(个人经济因子×个人经历因子)拟合Logistic回归模型。首先将以上5个变量都作为候选自变量, 然后由截距模型开始, 变量一律根据比分检验的概率大小依次进入方程, 对于加入到方程的变量, 再根据似然比检验概率大小移出方程, 最终所剩变量即为模型自变量(王济川和郭志刚, 2001)。

4.1 拟合15–29岁数据的Logistic回归模型

$$p = \frac{\exp(-4.734 - 0.729F_1 + 1.118F_2 + 0.22F_3 + 0.067F_4 - 0.384F_2 \times F_3)}{1 + \exp(-4.734 - 0.729F_1 + 1.118F_2 + 0.22F_3 + 0.067F_4 - 0.384F_2 \times F_3)}. \quad (4.1)$$

由式(4.1)可以看出, 除常数项以外个人经历状况因子的回归系数最大, 说明个人经历状况因子对妇女超生概率的影响最为显著, 而与个人经历状况因子有关的变量为学历、户口性质和工作状况, 换句话说, 即学历、户口性质和工作状况对妇女超生概率的影响是最

为显著的。因此，降低妇女超生的概率可以主要从学历、户口性质和工作状况这几方面入手。

表1 无偏分类表

观测值		预测值					
		预测样本			确认样本		
		超生		%	超生		%
超生	否	否	是		否	是	
	是	29900	0	100.0	30389	0	100.0
总百分比 (%)		98.5			98.4		

从表1可以看出，由预测样本(随机抽取的50%个数据)建立回归模型并用该模型计算预测样本的正确率达98.5%，由预测样本建立回归模型并用该模型预测确认样本(另外的50%个数据)的正确率达98.4%，将样本分为两类建立无偏分类表的好处是避免了用一组数据建立模型后，又用该组数据产生预测的分类表。同时，从表1还可看出，该模型对于不同样本的预测正确率接近，也就是模型的预测准确性是稳定的。

4.2 拟合30–50岁数据的Logistic回归模型

通过同样的标准和方法，可以得出30–50岁数据的Logistic回归模型，并且该模型确认样本的预测准确率也达到了80%以上。

$$p = \frac{\exp(-2.072 + 0.291F_1 + 1.701F_2 + 0.053F_3 + 0.311F_4 - 0.085F_2 \times F_3)}{1 + \exp(-2.072 + 0.291F_1 + 1.701F_2 + 0.053F_3 + 0.311F_4 - 0.085F_2 \times F_3)}. \quad (4.2)$$

4.3 分段的Logistic回归模型

$$p = \begin{cases} \frac{\exp(-4.734 - 0.729F_1 + 1.118F_2 + 0.22F_3 + 0.067F_4 - 0.384F_2 \times F_3)}{1 + \exp(-4.734 - 0.729F_1 + 1.118F_2 + 0.22F_3 + 0.067F_4 - 0.384F_2 \times F_3)}, & 15 \leq X_1 \leq 29; \\ \frac{\exp(-2.072 + 0.291F_1 + 1.701F_2 + 0.053F_3 + 0.311F_4 - 0.085F_2 \times F_3)}{1 + \exp(-2.072 + 0.291F_1 + 1.701F_2 + 0.053F_3 + 0.311F_4 - 0.085F_2 \times F_3)}, & 30 \leq X_1 \leq 50. \end{cases}$$

4.4 回归系数显著性检验

表2 进入回归方程的变量

15—29岁	背景因子	经历因子	经济因子	生活稳定因子	经济因子×经历因子
模型系数	-0.729	1.118	0.220	0.067	-0.384
Wald统计量	62.473	255.572	22.768	4.003	47.042
30—50岁	背景因子	经历因子	经济因子	生活稳定因子	经济因子×经历因子
模型系数	0.291	1.701	0.053	0.311	-0.085
Wald统计量	1398.978	6479.913	11.042	937.370	19.391

从表2可以看出,无论年龄在15—29岁还是30—50岁的模型,所用自变量(个人背景因子、生活稳定因子、个人经历因子、个人经济因子、个人经济因子×个人经历因子)的Wald统计量均大于3.841,所以它们都在 $\alpha = 0.05$ 水平上统计显著.因而以上5个自变量对于妇女超生的概率都有显著影响的.

§5. 结 论

Logistic回归模型在研究人口统计中有重要的应用价值,通过建立Logistic回归模型,对我国的“第五次人口普查”数据进行分析,得到如下结论:

(1) 通过大量试验发现,对影响因素年龄进行分段得到的分段模型,对于15—29岁妇女超生与否有很强的预测能力,样本回代正确率98%以上.

(2) 个人经历状况因子对妇女超生影响最为显著,并且,个人经历状况因子对于妇女超生概率具有正作用,与个人经历状况因子有关的变量包括学历、户口性质和工作状况,换句话说,学历、户口性质和工作状况对妇女超生与否的影响是最为显著的.因此,降低妇女超生的概率可以主要从学历、户口性质和工作状况这几方面入手.

参 考 文 献

- [1] 王济川, 郭志刚, Logistic回归模型——方法与应用, 高等教育出版社, 北京, 2001.
- [2] Johnson, R.A., Wichern, D.W.著, 陆璇译, 实用多元统计分析, 清华大学出版社, 北京, 2001.
- [3] 蒲诗松, 王静龙, 数理统计, 华东师范大学出版社, 上海, 1990.
- [4] 王松桂, 陈敏, 陈立萍, 线性统计模型: 线性回归与方差分析, 高等教育出版社, 北京, 1999.
- [5] 余建英, 何旭宏, 数据统计分析与SPSS应用, 人民邮电出版社, 北京, 2003.
- [6] 张文彤, 世界优秀统计工具SPSS 11.0统计分析教程(高级篇), 北京希望电子出版社, 北京, 2002.
- [7] Kleinbaum, D.G., Kupper, L.L., Muller, K.E. and Nizam, A., 应用回归分析和其他多元方法(英文版·第3版), 机械工业出版社, 北京, 2003.
- [8] 阮桂海等, SAS统计分析实用大全, 清华大学出版社, 北京, 2003.