中国人口时间序列预测模型的探讨

陈爱平[1], 安和平[2]
(1. 贵州大学 成人教育学院, 贵州 贵阳 550003; 2. 贵州大学人口研究中心, 贵州 贵阳 550025)

摘 要: 以 1952~2002 年人口序列数据为基础, 构建了中国人口时间序列预测模型, 采用一般回归法、后退法和逐步回归法三种方法对所设计的模型进行估计, 得到 72 个模型。然后, 从中筛选出估计标准误较小, 自相关影响基本消除的预测模型。作为中国人口总量的预测模型, 我们提出的预测模型有估计精度高, 误差低的特点, 是目前用自回归方法估计得出的较好的中国人口预测模型。

关键词: 中国人口总量; 自回归模型; 人口预测
中图分类号: C924.2 文献标识码: A 文章编号: 1000-4149(2004)06-0063-05

A Study on Chinese Population Forecast by Time Series Model

CHEN Ai-ping[1] AN He-ping[2]
(1. Educational Colleges of Adults, Guizhou University, Guiyang City, Guizhou Province, 550003;
2. Population Research Center, Guizhou University, Guiyang City, Guizhou Province 550025)

Abstract: Based on population data from 1952 to 2002, the authors developed Chinese population time-series forecast model. By using ordinary regression, backward and stepwise regression, the authors examined the designed models, and got 72 models. Then the authors chose the best forecast model with relatively small standard error and collinear relationships as the final Chinese population forecast model. This model is relatively good Chinese population forecast model by using auto-regression method.

Keywords: Chinese population; auto-regression model; population forecast

一、引言

以时间序列分析方法为基础的自回归模型用于中国人口预测已有了一定进展[1,2]。时间序列分析方法建立模型不考虑以人口理论或经济理论为依据的解释变量的作用, 而是依据变量本身的变化规律, 利用外推机制描述和预测时间序列的变化。建立时间序列模型的前提是时间序列必须具有平稳性。文献[1]中的综合模型, 由于考虑了与被解释变量同期的 GDP 值、同期的人口出生率和人口死亡率等作为解释变量, 对数据的拟合有很高的精度, 在一定程度上, 揭示了人口总量与被解释变量的关系。但是, 由于解释变量是被解释变量的同期变量, 从而使模型丧失了预测功能; 而文献[2]中提出的模型, 虽有预测功能, 其预测精度不高, 误差较大, 且复相关系数作为评价指标处于失真状态。以上模型之所以有缺陷, 一

收稿日期: 2004-04-12
作者简介: 陈爱平(1956-), 女, 贵州省安顺市人, 贵州大学成人教育学院讲师, 主要研究方向为应用统计学、社会学。

是建立模型时遗漏了其他重要解释变量或是模型设计不合理，二是没有考虑人口总量时间序列是非平稳时间序列，使用自回归模型建立模型时没有将其转化为平稳时间序列或作其他变换处理。本文通过将人口总量序列（Y_t）变换为人口年增长量形式的人口差分序列（$\triangle Y_t$），以人口差分序列（$\triangle Y_t$）作为被解释变量，人口年增长量的滞后差分序列（$\triangle Y_{t-i}$）作为解释变量。以 1952~ 2002 年中国人口序列数据为基础，对设计的模型，采用一般回归法、后退法和逐步回归法三种方法进行估计，得 72 个预测模型。然后，根据估计标准误和 DW 值等评价指标，从中筛选出自相关影响已消除，估计标准误最小的模型作为中国人口预测模型。本文所提出的预测模型对 2003 年中国人口总量进行预测，其预测值为 129244.139 万人。与 2003 年实际人口总量 129227 万人相比仅偏高 17.139 万人[3]，其相对误差仅为 0.013%。

二、中国人口时间序列过程分析

年人口总量用 Y_t 表示，其人口总量序列为 $\{Y_t\}$；年人口增长量用 $\triangle Y_{t-1}$ 表示，即人口一阶差分序列为 $\{\triangle Y_t\}$。则，1952~ 2002 年中国人口时间序列数据变化见图 1、图 2。由图 1 可以看出中国人口总量除在 1960、1961 年出现回落外，其他年份基本上保持增长趋势。从 1952~ 2002 年中国人口年平均增长 1419.42 万人，年平均增长率为 15.2‰。而 1974 年前年平均增长率为 21.2‰，从 1974~ 2002 年人口年平均增长率为 12.2‰，说明计划生育政策对控制人口已取得明显效果。从人口总量的变化特征看出，这是一个非平稳时间序列。

图 2 展示了中国人口年增长量的变化特征。建国初期由于进入和平环境，国民经济得以恢复，人口年增长量从 1950 年的 1029 万人，到 1957 年猛增到 1825 万人。在三年经济困难时期的 1960 年和 1961 年，由于粮食短缺等原因，人口出现负增长。此后，随着经济形势的好转，1962 年人口年增长量恢复到 1436 万人，随后呈连续递增态势。在 1965~ 1973 年平均人口年增长量为 2079.111 万人，平均人口自然增长率为 25.7‰，是中国人口增长最快的时期。其中，1970 年是人口增长最快的年份，年人口增长量为 2321 万人，自然增长率高达 28.0‰。随着上世纪 70 年代计划生育政策执行力度的加强，从 1974 年开始，年人口增长量开始逐步下降，至 1980 年基本回落到建国初期水平。由于受上世纪 50 年代和 60 年代初高出生率的影响，从 1981 年开始年人口增长量呈逐年上升趋势，到 1987 年出现了中国人口年增长量的第二个高峰年（该年人口增长量为 1793 万人）。此后，年增人口开始逐步回落，到 2000 年降至年增人口 1000 万人以下，2003 年年增人口已降到 774 万人，中国人口增长过快的势头基本得到有效控制。从人

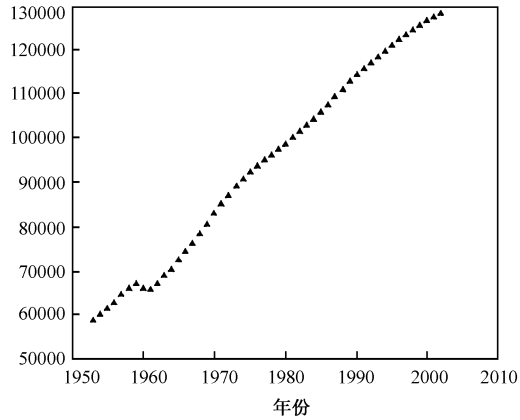

图 1　1952~ 2002 年中国人口总量序列

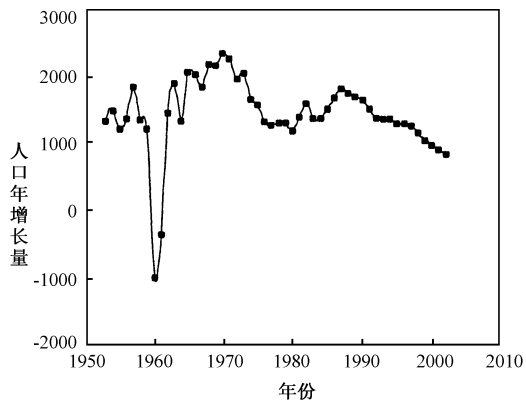

图 2　1952~ 2002 年中国人口年增长量序列

口总量序列 $\{Y_t\}$ 的变化特征看，1960、1961 年数据可看为两个异常值，其他年份则表现为平稳特征。但也不是白噪声序列，而是一个含有自相关和（或）移动平均成份的平稳时间序列。由于建立时间序列模型的前提是时间序列必须具有平稳性。如果时间序列是非平稳的，建立模型之前应先把它换成平稳的时间序列，同时仍保持原时间序列的随机性。因此，我们只考虑人口年增长量序列，即人口一阶差分序列 $\{\triangle Y_t\}$ 的建模。

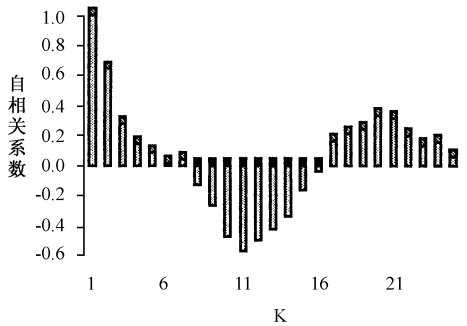

图 3　自相关图

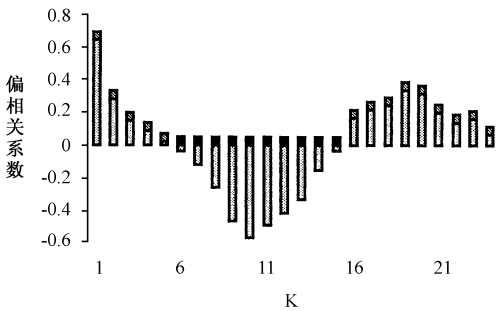

图 4　偏相关性图

从图 3、图 4 看出 $\triangle Y_t$ 的自相关系数、偏相关系数按正弦衰减，并有周期性，在第二周期的偏相关系数在最大值比第一周期的最大值小，有随周期的加大而逐步减小趋势，由于周期性的存在，很难判别 AR 模型阶数。为此，我们采取设计出所有可能的自回归过程，用修正后的复相关系数（Adjuste R^2）、DW（Durbirr Watson）、残差均方（Residual mean Square）、估计标准误（Std. Error of the Estimate）、F 检验的显著性等作为所建模型的评价指标，从中筛选出较为理想的模型作为预测模型。为使人口回归模型有预测功能，我们设计的模型不包括当期变量，即 i 必须满足 i ≥1。

自回归过程模型的一般形式[4]：

$$\triangle Y_t = \alpha_0 + \sum_{i=1}^{n} \alpha_i \triangle Y_{t-i} + \mu_t \qquad (2.1)$$

这样我们就用（2.1）式作为人口年增长量的自回归预测模型。通过下式：

$$Y_t = Y_{t-1} + \triangle Y_t \qquad (2.2)$$

得到人口总量序列，即计算出各年人口总量。在（2.1）和（2.2）式中 $\triangle Y_t$ 或 Y_t 仅依靠滞后信息来预测，这样它们就有了预测的功能。下面的关键就是如何设计具体的模型、选择恰当的估计方法和评价指标，筛选出相对较优的预测模型。

三、模型的建立与评价

根据建立回归模型时要求样本的个数（N）应大于解释变量的个数（n），且满足 N> n+ 1。用 1952~ 2002 年人口序列数据建立模型时，得（2.1）式中的最大滞后期为 i= 24。用一般回归法对（2.1）式进行估计使全部解释变量进入（2.1）式，得到的模型称为全模型；用逐步回归法和后退法[5]对（2.1）式进行估计，得到的模型称为选模型。在使用 SPSS 软件计算时，设置 Collinearity diagnostics，进行共线性诊断，剔出有共线性影响的解释变量；用修正后的复相关系数（Adjuste R^2）、DW 值（Durbirr Watson）、残差均方（Residual mean Square）、估计标准误（Std. Error of the Estimate）、F 检验的显著性等作为所建模型的评价指标，其结果见表 1。由于所有模型的 F 值随 i 的增大而增大，在显著性水平为 P= 0.000 时全部通过显著性检验，因而不再列入表 1 中。由表 1 看到，三种不同估计方法对（2.1）式估计得到的模型其参数有一定差异：从 DW 值看，逐步回归法估计的结果中只有 i = 9，18，22，23，24 受自相关的影响较大，在消除自相关影响的情况下，当 i= 16 时估计标准误最小，修正后的复相关系数相对最大；一般回归法估计的结果中，有 i= 7~ 12，16，17，23，24 估计的模型受自相关的影响较大，在消除自相关影响的情况下，当 i= 19 时所建

模型估计标准误最小，修正后的复相关系数相对最大；由后退法估计得到的模型中，在消除自相关影响的情况下，当 i= 23 时所建模型估计标准误最小，修正后的复相关系数相对最大。三种方法估计的模型经综合比较，以后退法所建模型中，i= 23 时估计的结果最好，不仅估计标准误最小，从 DW 值看，模型在 1% 的显著性水平下已不受自相关的影响[6~7]，并且该模型只含有 6 个解释变量。因此，我们将其作为本文提出的中国人口预测模型。由表 2 看到该预测模型的回归系数都通过显著性检验。用该模型对 2003 年中国人口总量进行预测，其残差仅为 17.139；从对 1976~ 2002 年人口年增长量拟合的结果看，有 39.3% 的年份相对误差< 1.0%，25% 的年份相对误差在 1.1% ~ 3.0%，28.6% 的年份相对误差在 3.1%~ 5.0%，仅有 7.1% 的年份相对误差大于 5.0%，说明预测精度是比较高的。从图 5 看出，该模型的残差基本成随机排列，表明自相关影响已消除。

表 1　不同方法的估计结果

模型号*	一般回归法			后退法				逐步回归法			
	DW 值	修正后的复相关系数	估计标准误	DW 值	修正后的复相关系数	估计标准误	剔除变量个数	DW 值	修正后的复相关系数	估计标准误	引入变量
1	1.713*	0.400	439.1	1.713*	0.400	439.1	0	1.713*	0.400	439.1	ΔY_{t-i}
2	1.916*	0.416	437.7	1.715*	0.400	443.4	1	1.715*	0.400	443.4	ΔY_{t-i}
3	1.968*	0.416	441.6	1.716*	0.402	446.8	2	1.716*	0.402	446.8	ΔY_{t-i}
4	1.994*	0.404	451.1	1.718*	0.402	451.6	3	1.718*	0.402	451.6	ΔY_{t-i}
5	1.940*	0.392	458.3	1.721*	0.409	451.6	4	1.721*	0.409	451.6	ΔY_{t-i}
6	1.998*	0.386	465.6	1.735*	0.416	454.0	5	1.735*	0.416	454.0	ΔY_{t-i}
7	1.293	0.374	475.1	1.734*	0.416	459.0	6	1.734*	0.416	459.0	ΔY_{t-i}
8	1.527	0.744	238.7	1.210	0.755	233.4	2	1.937*	0.714	252.3	ΔY_{t-1}, i= 1, 2, 3
9	2.326	0.663	219.0	2.248**	0.633	228.6	4	1.193	0.507	265.1	ΔY_{t-i}, i= 1.9
10	2.401	0.765	185.0	2.279**	0.791	174.7	4	1.915*	0.728	199.2	ΔY_{t-i}, i= 1.5
11	1.556	0.848	149.1	1.589	0.858	144.1	5	1.421**	0.830	157.7	ΔY_{t-i}, i= 1, 2, 3, 5
12	1.831	0.871	138.8	1.330	0.887	129.7	7	1.990*	0.861	144.2	ΔY_{t-i}, i= 14
13	2.032**	0.863	141.3	1.825**	0.879	132.7	7	1.617*	0.868	138.6	ΔY_{t-i}, i= 1.8
14	2.087**	0.878	132.3	2.371**	0.877	132.5	11	2.371*	0.877	132.5	ΔY_{t-i}, i= 1, 8, 14
15	2.035**	0.868	137.8	1.686**	0.896	122.5	10	1.529**	0.866	139.1	ΔY_{t-i}, i= 1.7
16	1.576	0.934	94.2	1.575	0.946	85.4	8	1.839*	0.941	88.6	ΔY_{t-i}, i= 1, 4, 8, 9, 14, 16
17	1.568	0.925	96.3	1.487	0.942	84.6	9	1.728**	0.924	96.9	ΔY_{t-i}, i= 1, 5, 9, 14, 16
18	2.383**	0.962	62.2	2.194**	0.960	64.1	9	1.323	0.912	94.9	ΔY_{t-i}, i= 1, 3, 14, 16
19	1.934**	0.953	62.7	2.152**	0.949	64.7	10	1.539**	0.847	112.5	ΔY_{t-i}, i= 1, 16
20	1.961**	0.936	69.3	1.602	0.957	56.5	6	1.756**	0.890	90.9	ΔY_{t-i}, i= 1, 4, 14, 16
21	1.956**	0.900	78.6	1.443	0.946	57.5	8	1.734**	0.862	92.3	ΔY_{t-i}, i= 1, 4, 14, 16
22	1.989**	0.891	81.1	2.167*	0.944	58.1	11	1.280	0.861	91.6	ΔY_{t-i}, i= 1, 16, 21
23	2.824	0.941	60.2	1.784**	0.964	46.8	17	0.766	0.892	80.7	ΔY_{t-i}, i= 1, 4, 16, 21
24	3.504	0.994	19.4	2.908	0.997	13.3	4	1.113	0.845	99.1	ΔY_{t-i}, i= 1, 21

* 模型号与滞后期 i 对应。* 在显著性水平为 0.05 下无自相关影响；** 在显著性水平为 0.01 下无自相关影响

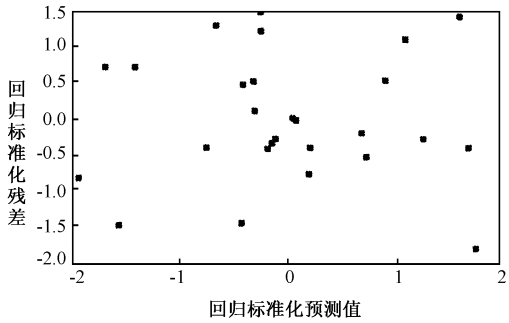

图 5　残差图

(被解释变量：人口年增数)

四、讨论

从三种估计方法的估计效果看，以后退法估计的效果较好，其次是一般回归法，而逐步回归法的估计效果相对较差。从表 1 看出，同一组数据在用同一公式，用不同的估计方法建立模型其结果有较大差异。从一般回归法的估计模型看，在 i> 10 时，仍有许多较优的模型，这些模型中虽包括部分不显著的解释变量，但我们不能盲目地断定它们是多余或有负面影响的变量。表 1 中一般回归法所建立的模型中，模型 13，14，15，18，19，20，正是由于这些不显著解释变量的存在，模型的估计精度才得到提高，其他评价指标才得到改善。后退法所建立的模型仅包含回归系数显著的变量，其中部分变量是不显著的变量，但在 i= 9，10，11，13，18，22，23 时，建立的模型是相对较优的估计模型。有人认为，所建模型包含的解释变量应当少而精，只选取有显著影响的解释变量建立模型才是最佳模型。我们认为，要根据研究的目的，应考虑变量的存在是否有利于提高估计精度和改善其他评价指标，并考虑解释变量数据获得的成本，难易程度等。因此，不能片面地认为模型中不显著的变量就是多余的变量或会增加估计误差的变量。

表 2　预测模型的回归系数

参数	非标准化回归系数	标准化回归系数	t 值	显著性水平
α_0	- 1193.072		- 8.725	0.000
α_1	0.799	0.740	17.427	0.000
α_8	0.296	0.401	6.842	0.000
α_{14}	0.247	0.348	6.690	0.000
α_{16}	0.127	0.369	7.554	0.000
α_{19}	0.131	0.384	6.866	0.000
α_{23}	0.112	0.327	6.759	0.000

本文所建模型是依靠滞后信息建立的人口预测模型，其数据采集成本已不再考虑，在消除自相关影响的情况下，表 1 中带* 或* * 的模型可分别作为是 i= 1，2，…23 时的中国人口预测模型。从各项指标综合考虑，以 i= 23 时后退法估计的模型相对较优。我们提出的预测模型对 2003 年中国人口总量的预测值仅偏高 17.139 万人，与文献 [1]、[2] 提出的模型相比，文献 [1]、文献 [2] 的预测结果分别偏高 46.3367 万人和 447.267 万人。说明本文提出的预测模型不仅有预测功能，而且有较高的预测精度，是目前较好的用自回归方法建立的中国人口预测模型。

参考文献：

[1] 赵进文. 中国人口总量与 GDP 总量关系模型研究. 中国人口科学，2003，(3).

[2] 范柏乃，刘超英. 中国人口总量预测模型新探——与赵进文教授商榷. 中国人口科学，2003，(6).

[3] 国家统计局. 中华人民共和国 2003 年国民经济和社会发展统计公报，2004.

[4] 张晓峒. 计量经济分析. 北京：经济科学出版社，2000.

[5] 何晓群，刘文卿. 应用回归分析. 北京：中国人民大学出版社，2001.

[6] 黄海波主编. 经济计量学精要习题集. 北京：机械工业出版社，2003.

[7] (美) 拉姆• 拉玛纳山 (Ramu Ramanathna). 薛菁睿 译. 应用经济学 (原书第五版). 北京：机械工业出版社，2003.

[8] 国家统计局. 中国统计年鉴. 北京：中国统计出版社，1991~ 2003.

[责任编辑　　齐明珠]