中国人口增长预测模型的建立与应用

罗 翔
(华中师范大学数学与统计学院,湖北 武汉 430079)

摘要:参考1978-2005年《中国国民经济与社会发展公报》中的数据,根据一般多元线性回归、二元二项式回归方法建立了中短期中国人口预测模型和长期预测模型,随后建立了费尔哈斯模型与之比较。结果表明,本文所建立的模型有预测结果精确度高、检验充分和模型稳定性好等优点,并针对模型中存在的不足提出了改进方案。

关键词:多元线性回归;二元二项式回归;费氏模型;插值拟合
中图分类号:O221.2 **文献标识码**:A **文章编号**:1673-6060(2008)03-0145-04

我国是一个人口大国,人口问题始终是制约我国发展的关键因素之一。根据已有数据,运用数学建模的方法,对中国人口做出分析和预测是一个重要的问题。

本文根据文献资料结合中国的实际情况和人口增长特点建立中国人口增长的数学模型,并由此对中国人口增长的中短期和长期趋势做出预测。

1 模型的建立和求解

1.1 中短期人口预测模型

对于中短期的人口预测,根据多元线性回归原理并运用MATLAB软件,实现二次多项式(曲面)拟合,并对模型的有效性进行检验。

1.1.1 多元线性回归模型

1.1.1.1 多元线性回归模型 根据多元线性回归的统计分析原理建立模型

$$\begin{cases} y_i=\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}+\varepsilon_i \\ \varepsilon_i\sim N(0,\sigma^2),i=1,2,\cdots,n \end{cases} \quad (1)$$

依据最小二乘准则

$$Q(\beta)=\sum_{i=1}^{n}\varepsilon_i^2=(Y-X\beta)^T(Y-X\beta)$$

估计模型中的参数β,得到β的线性无偏最小方差估计$\hat{\beta}$,其中$\hat{\beta}=(X^TX)^{-1}X^TY$。

(1)对误差方差σ^2进行无偏估计得到:

$$s^2=\hat{\sigma}^2=\frac{Q}{n-m-1}$$

且Q的自由度为$n-(m+1)$。

(2)根据概率统计中的区间估计得到回归系数β_0的置信区间:

$$[\hat{\beta}_0-t_{(n-2),1-\alpha/2}s\sqrt{\bar{x}^T(\bar{X}^T\bar{X})^{-1}\bar{x}+\frac{1}{n}},$$

$$\hat{\beta}_0+t_{(n-2),1-\alpha/2}s\sqrt{\bar{x}^T(\bar{X}^T\bar{X})^{-1}\bar{x}+\frac{1}{n}}$$

其中$\bar{x}^T=(\bar{x}_1,\cdots,\bar{x}_m)$。

(3)对模型进行有效性检验——决定系数和F统计量。

1.1.1.2 多元线性回归模型的预测原理 当模型(1)通过有效性检验后,可由自变量的任一给定值$x=(x_1,\cdots x_m)$预测因变量的理论值y,记作$\hat{y}$。显然

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_mx_m \quad (2)$$

与一元线形回归一样,$\hat{y}$是无偏的,并且均方误差$E(\hat{y}-y)^2$最小。

在给定显著性水平α下y的预测区间为

$[\hat{y}-\delta(x),\hat{y}+\delta(x)]$ $\delta(x)=t_{(n-2),1-\alpha/2}s\sqrt{(x-\bar{x})^T(\bar{X}^T\bar{X})^{-1}(x-\bar{x})+\frac{1}{n}+1}$

当n很大且x接近$\bar{x}$时,上述预测区间简化为

$[\hat{y}-u_{1-\alpha/2}s,\hat{y}+u_{1-\alpha/2}s]$

1.1.2 包含线性项和完全二次项的人口预测模型 我们通过对线性人口预测模型、包含线性项和交互项的人口预测模型以及包含线性项和纯二次项人口预测模型的研究,在包含线性项和纯二次项人口预测模型的基础上加入了交互项,建立模型

$$y_4=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2+\beta_4x_1^2+\beta_5x_2^2 \quad (3)$$

该模型充分考虑了完全二次项(x_1^2、x_2^2、x_1x_2)对模型的影响,利用MATLAB编程,计算其结果见表1。

收稿日期:2008-02-10
作者简介:罗翔(1986-),女,河南新乡人。

表 1　模型 3 的计算结果

模型 4：$y_4=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2+\beta_4x_1^2+\beta_5x_2^2+\varepsilon$		
回归系数	回归系数估计值	回归系数置信区间
β_0	-667.0660	[-828.3975 -505.7344]
β_1	-4.9940	[-9.1606 -1.2567]
β_2	220.1049	[166.9005 274.9094]
β_3	0.8131	[0.1233 1.5029]
β_4	-0.0100	[-0.2973 0.9668]
β_5	-17.8741	[-22.5967 -13.1514]
$R^2=0.9389$	$F=67.5778$	$P=0.1317*10^{-11}$　$S^2=0.0938$

通过分析和比较，可知模型(3)中 $R^2=0.9389$ 是几个模型中最大的，且 $S^2=0.0938$ 是最小的，其模型的精度是最高的，采取该模型对人口总数量(1997-2005)进性预测结果如表 2。

表 2　模型 3 的预测值

年份	人口总数(亿)	Fullquadratic 预测值	预测误差	5% 显著预测值下限	5% 显著预测值上限
1978	9.6259	9.7317	-0.1058	9.0670	10.3964
1979	9.7542	9.4344	0.3198	8.6296	10.2392
1980	9.9705	10.7124	-0.7419	10.2332	11.1916
1981	10.0072	9.7269	0.2803	8.8720	10.5818
1982	10.1654	10.852	-0.6866	10.2946	11.4094
1983	10.3000	10.2186	0.0814	9.3596	11.0776
1984	10.4357	9.2852	1.1505	8.4377	10.1327
1985	10.5851	10.9893	-0.4042	10.4767	11.5019
1986	10.7507	10.4336	0.3171	9.6902	11.1770
1987	10.9300	10.7527	0.1773	10.0766	11.4288
1988	11.1026	10.9236	0.1790	10.4007	11.4465
1989	11.2704	11.1213	0.1491	10.7005	11.5421
1990	11.4333	11.2484	0.1849	10.8762	11.6206
1991	11.5823	11.4537	0.1286	11.0933	11.8141
1992	11.7171	11.8190	-0.1019	11.4037	12.2343
1993	11.8517	11.8451	0.0066	11.4213	12.2689
1994	11.9850	11.8192	0.1658	11.3842	12.2542
1995	12.1121	12.0932	0.0189	11.6831	12.5033
1996	12.2389	12.1222	0.1167	11.7161	12.5283
1997	12.3626	12.1705	0.1921	11.7704	12.5706
1998	12.7610	12.3813	0.3797	12.0148	12.7478
1999	12.5786	12.5619	0.0167	12.2015	12.9223
2000	12.6743	12.6948	-0.0205	12.3288	13.0608
2001	12.7627	12.8187	-0.0560	12.3980	13.2394
2002	12.8453	12.9082	-0.0629	12.4039	13.4125
2003	12.9227	13.0004	-0.0777	12.4234	13.5774
2004	12.9968	13.0657	-0.0689	12.5023	13.6291
2005	13.0756	13.0286	0.0470	12.1409	13.9163

通过表 1 的数据可以知道其模型的预测值与实际的数据的拟合程度较高，对模型(3)的残差分析图(见图 1)可以发现其中的数据基本上趋于正常，可见该模型的精度较高。

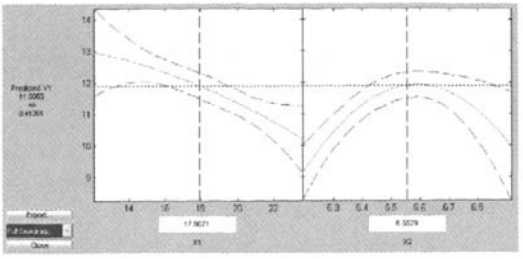

图 1　二元二项式回归交互界面

最终综合比较分析得到精度较高的回归方程为：

$y_4=-667.0660-4.9940x_1+220.1049x_2+0.8131x_1x_2-0.0100x_1^2-17.8741x_2^2$

对中短期(2007-2021)的人口进行预测，计算结果见表 3。

表 3　模型 3 的预测值

年份	2007	2008	2009	2010	2011	2012	2013	2014
预测值	13.1111	13.3954	14.0231	15.2001	15.6617	15.7543	15.5665	15.6324
年份	2015	2016	2017	2018	2019	2020	2021	
预测值	15.5607	15.4798	15.7228	15.7205	15.7560	15.7501	15.1171	

1.2　长期人口预测模型

由于人口性别比在长期内不是稳定的，且人口性别比会影响人口总数，为了使模型精度更高，结果更准确。充分考虑人口性别比对模型的影响，建立如下模型：

$$y_1=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3 \tag{4}$$

$$y_2=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_1x_2+\beta_5x_1x_3+\beta_6x_2x_3 \tag{5}$$

$$y_3=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_1^2+\beta_5x_2^2+\beta_6x_3^3+\beta_7x_1x_2+\beta_8x_2x_3+\beta_9x_2x_3 \tag{6}$$

$$y_4=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_1^2+\beta_5x_2^2+\beta_6x_3^3 \tag{7}$$

此处的 x_3 记为人口性别比，在建立长期人口预测模型中，关键是预测出人口性别比、出生率和死亡率，该模型采用多次曲线的插值拟合的方法预测，采用 MATLAB 编程求其预测结果见表 4、5。

表 4　人口性别比的预测

年份	2007	2008	2009	2010	2011	2012
性别比	118.1031	118.1482	118.1422	118.0827	117.9673	117.7938
年份	2013	2014	2015	2016	2017	2018
性别比	117.5599	117.2630	116.9010	116.4715	115.9721	115.4005
年份	2019	2020	2021	2022	2023	2024
性别比	114.7544	114.0314	113.2291	112.3454	111.3777	110.3237
年份	2025	2026	2027	2028	2029	2030
性别比	109.1812	107.9478	106.6211	105.1988	103.6786	102.0581
年份	2031	2032	2033	2034	2035	2036
性别比	100.3349	98.5068	96.5715	94.5264	92.3684	90.0981

表5　出生率和死亡率的预测

年份	2007	2008	2009	2010	2011	2012	2013	2014
出生率	17.1481	18.4664	19.4606	20.2944	20.9311	21.3842	21.6627	21.7934
死亡率	6.1836	6.2344	6.3086	6.4609	6.5742	6.7031	6.7813	6.7656
年份	2015	2016	2017	2018	2019	2020	2021	2022
出生率	21.7765	21.6299	21.3669	21.0011	20.5459	20.0149	19.421	18.7788
死亡率	6.7852	6.8008	6.7031	6.7148	6.6836	6.5938	6.5625	6.5547
年份	2023	2024	2025	2026	2027	2028	2029	2030
出生率	18.1008	17.4006	16.6919	15.988	15.3024	14.6486	14.0401	13.4902
死亡率	6.5469	6.543	6.5078	6.5313	6.5117	6.4688	6.4258	6.3672
年份	2031	2032	2033	2034	2035	2036		
出生率	13.0124	12.6203	12.3272	12.1467	12.3372	12.2467		
死亡率	6.3633	6.332	6.375	6.5508	6.4649	6.4883		

对长期人口预测模型(4)(5)(6)(7)的求解原理和思想都符合线性最小二乘原理,用 MATLAB 编程求解长期 linear 模型的计算结果、长期 interaction 模型的计算结果、长期 fullquadratic 模型的计算结果和长期 purequadratic 模型的计算结果,结果表明长期 purequadratic 模型中 $R^2 = 0.9989$ 接近 1 是模型中的最大的,$s^2 = 0.00404496$ 是四个模型中最小的(见表6),可见该模型的精度最高,最后可以得出最优预测的模型为

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 \quad (8)$$

表6　长期 purequadratic 模型的计算结果

包含 purequadratic 的二次项的线性模型:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2$

回归系数	回归系数估计值	置信区间上下限
β_0	-27.7135	[-143.5190 88.0919]
β_1	0.1228	[-0.1199 0.3655]
β_2	4.6578	[-13.6001 22.9156]
β_3	6.1429	[-1.5894 1.8825]
β_4	-0.0031	[-0.0086 0.0024]
β_5	-0.3622	[-1.7435 1.0192]
β_6	0.0006	[-0.0073 0.0084]
$R^2 = 0.9989$	$F = 1387.4$	$P = 0.1012 * 10^{-6}$　$s^2 = 0.00404496$

由于出生率、死亡率、人口性别比均已预测出,可通过该模型预测出长期(2007 - 2036)的人口总数,预测结果为见表7。

1.3　费尔哈斯人口预测模型的建立

费氏模型是在马尔萨斯模型基础上的修正,增加了一个二次项,生长和发展受制于环境约束:

$$\frac{dP(t)}{dt} = aP(t) - b[P(t)]^2$$

式中 $P(t)$ 为年的预测的人口数;a、b 分别为一次、二次的常系数;M 为区域人口上限,$M = a/b$

当 $P(t_0) = p_0$ 时的费氏模型的解析解为:

$$P(t) = \frac{M}{1 + \left(\frac{M}{p_0} - 1\right)e^{-\lambda M(t - t_0)}} \quad (9)$$

参数 a 和 b 的确定。要应用费氏模型进行人口预测,首先必须确定常系数,确定费氏模型的参数的方法,在这里我们用微分方程求解法。

微分方程求解就是利用至少3个的历史数据通过微分的方法计算式(9)左端的导数,并将两组导数和人口数代入式(9),从而得到二元一次方程组,最终获得参数 a、b,如果3个人口数据分别为 P_0、P_1 和 P_2,对应的时间分别为 T_0、T_1 和 T_2,则和的导数为

表7　长期模型的人口预测值(亿)

年份	2007	2008	2009	2010	2011	2012	2013	2014
预测值	13.0784	13.4954	13.7938	14.3106	14.9450	15.3027	15.2966	15.2064
年份	2015	2016	2017	2018	2019	2020	2021	2022
预测值	15.9239	15.6242	15.7900	15.4338	15.2258	15.3740	14.9300	14.9435
年份	2023	2024	2025	2026	2027	2028	2029	2030
预测值	14.9976	14.4644	14.5762	14.3673	14.4313	14.7972	14.1384	14.5723
年份	2031	2032	2033	2034	2035	2036		
预测值	14.4756	14.6752	14.7023	14.6750	14.9872	14.8976		

$$P_1' = \frac{P_1 - P_0}{T_1 - T_0}$$

$$P_2' = \frac{P_2 - P_1}{T_2 - T_1}$$

将 P_1 和 P_1' 以及 P_2 和 P_2' 分别代入式30,得到二元一次方程组,解该方程组即可获得参数 a 和 b 的计算结果为:

$$a=\frac{P_1'P_2^2-P_2'P_1^2}{P_1P_2(P_2-P_1)}$$

$$b=\frac{P_1'P_2-P_2'P_1}{P_1P_2(P_2-P_1)}$$

我们不妨取 2003、2004、2005 年的数据作为计算数据,运用上述公式在 Matlab 中编写函数求得参数 a、b 和 M

$a=0.0014, b=-3.3966\times10^{-4}, M=-4.2370$

那么在此数据基础上的费氏模型函数为:

$$P(t)=\frac{-4.237}{1-1.3279e^{-0.0014(t-2003)}}$$

我们可以运用此函数对 2007 - 2036 这 30 年的总人口数进行预测,结果见表 8。

表 8　费尔哈斯人口预测模型的计算结果(亿)

年份	2007	2008	2009	2010	2011	2012	2013	2014
预测值	13.2206	13.2973	13.3747	13.4529	13.5320	13.6118	13.6925	13.7714
年份	2015	2016	2017	2018	2019	2020	2021	2022
预测值	13.8565	13.9398	14.0239	14.1090	14.1950	14.2819	14.3698	14.4586
年份	2023	2024	2025	2026	2027	2028	2029	2030
预测值	14.5484	14.6392	14.7310	14.8329	14.9178	15.0127	15.1087	15.2059
年份	2031	2032	2033	2034	2025	2036		
预测值	15.3041	15.4035	15.5041	15.6058	15.7087	15.8129		

2　预测结果分析

2.1　由中短期的人口预测模型的结果可以看出,模型预测 2021 年时总人口数为 14.9300 亿已经很接近《国家人口发展战略研究报告》中提到的 2020 年的战略目标 14.5 亿。由于模型不仅可以预测人口总数,还可以预测当年的出生率,所以政府部门和计生部门可以通过控制出生率将总人口控制在目标附近。

2.2　对于长期的人口预测模型,在中短期的模型基础上加入男女性别比这一重要的因素再次运用 Matlab 统计工具箱二元二项式回归,通过对比后选择包含线性项和纯二次项的回归模型。利用该模型对 2007 - 2036 三十年内的人口总数,出生率,死亡率和性别比例进行了预测。三十年后也就是 2036 年的人口总数为 14.8976 亿,且回归曲线走势已趋于平缓,所以到本世纪中叶我国人口基本可以稳定在 15 亿左右,即可以达到《国家人口发展战略研究报告》的目标将人口控制在 15 亿左右。

2.3　费尔哈斯模型中短期的预测结果和用二元二项式回归模型进行预测所得结果在大体趋势上是一致的,这从一个侧面反映中短期预测模型的精确度较高。但随着时间的推移和其他相关因素的影响,其预测值和实际情况将会有较大偏差。

参考文献:

[1] 中国国民经济与社会发展公报[EB/OL]. http://cn.chinagate.com.cn/reports/2007-03/02/content_2368315.htm,2007-9-21.

[2] 姜启源,张立平,何青,等. 数学实验[M]. (第二版). 北京:高等教育出版社,2006.

[3] 王家文,王浩,刘海. Matlab7.0 编程基础[M]. 北京:机械工业出版社,2006.

[4] 裘书服,胡炳清,李秀珍. 费氏人口预测模型的求解及应用研究[J]. 森林工程,2006,3(22):54-56.

[5] 魏高峰,龙克柔. 中国人口演化模型与中国未来人口预测研究[J]. 科技咨询导报,2007(13):102-104.

[6] 韩中庚. 数学建模竞赛-获奖论文精选与点评[M]. 北京:科学出版社,2007.

(责任校对:刘明)