

# 目 录

<b>摘要</b>	I
<b>ABSTRACT</b>	II
<b>第一章 引言</b>	1
<b>第二章 初等函数模型</b>	1
2.1 一元线性函数与对数线性函数模型 . . . . .	1
2.2 指数函数模型 . . . . .	4
2.3 多项式模型 . . . . .	7
2.4 幂函数模型 . . . . .	9
<b>第三章 线性回归模型</b>	11
<b>第四章 微分方程模型</b>	17
4.1 马尔萨斯人口模型 . . . . .	17
4.2 阻滞增长模型 . . . . .	17
4.3 收割模型 . . . . .	21
<b>第五章 时间序列模型</b>	23
5.1 时间序列简单移动平均方法 . . . . .	23
5.2 加权一次移动平均预测 . . . . .	24
5.3 指数平滑预测 . . . . .	24
5.4 时间序列的 ARIMA 预测模型 . . . . .	25
<b>第六章 人口拟合与预测模型的对比分析</b>	29
<b>第七章 状态转移模型</b>	37
7.1 马尔可夫预测模型 . . . . .	37
7.2 人口增长的马尔可夫性 . . . . .	38
7.3 状态转移模型建模过程 . . . . .	39
7.4 状态转移模型的性质 . . . . .	41
7.5 模型评价与改进 . . . . .	42

7.6 应用实例 . . . . .	42
<b>第八章 其他模型与方法简述</b>	<b>44</b>
8.1 偏微分方程模型 - 宋健人口模型 . . . . .	44
8.2 差分方程模型 . . . . .	44
8.3 灰色系统 -G(1,1) 模型 . . . . .	45
8.4 神经网络模型 . . . . .	46
8.5 遗传算法模型 . . . . .	50
8.6 分形数据拟合方法 . . . . .	52
8.7 组合预测模型 . . . . .	55
8.8 非参数模型 . . . . .	55
<b>第九章 对基于数据的模型和基于因素的模型的一些思考</b>	<b>57</b>
<b>谢辞</b>	<b>60</b>
<b>参考文献</b>	<b>61</b>
<b>附录一 部分推导与证明</b>	<b>63</b>
A.1 线性自然增长模型的建立 . . . . .	63
A.2 指数自然增长模型的建立 . . . . .	63
A.3 偏微分方程模型的建立 . . . . .	64
A.4 差分方程模型的建立 . . . . .	65
<b>附录二 部分数据</b>	<b>67</b>
B.1 中国历年人口总数 . . . . .	67
<b>附录三 部分程序</b>	<b>68</b>
C.1 人口拟合预测的 BP 神经网络模型 Matlab 代码 . . . . .	68
C.2 人口拟合的分形程序 Matlab 代码 . . . . .	69

## 摘要

人口问题是中国面临的重大问题。准确的人口预测对制定人口政策有重要作用，从而直接影响社会、经济的发展。人口系统是一个复杂的系统，它受众多因素的影响。本文总结各种人口预测方法，目的在于得到一个适用于中国人口预测的较优的单模型。基于多模型的组合预测方法（如文献 [24]）可能会得到更好的结果，但不在本文的关注范围之内。通过对各个模型进行编程（Matlab/SAS）实验，讨论各模型对于中国人口增长的拟合预测效果，并对方法进行了分析比较，讨论了不同类型的模型的建模思想。

对初等函数模型，灰色系统，遗传算法，组合预测，非参数估计等方法作了简要介绍。在理论上探讨了马尔可夫预测模型在人口预测中应用的可行性，提出了具体实现方法和一些改进。

重点讨论了线性回归模型、微分方程模型和时间序列 ARIMA 模型。拟合、预测的对比结果表明时间序列 ARIMA 模型显著优于其他两种模型。此外，实验结果还发现神经网络模型对人口数据拟合预测十分有效。时间序列 ARIMA 模型和神经网络模型可以作为中国人口增长拟合、预测的两种较优选择。然而，在得出更加肯定的结论之前，还有许多问题值得研究。

### 关键词：

人口预测 线性回归 微分方程 时间序列 模型对比分析

# ABSTRACT

Population problem is a crucial issue in China. Forecasting population precisely is essential to make accordingly policy and hence will directly affect the growth of economy and the development of society. The population system is a complicate system influenced by many aspects. This paper is focus on finding a good single model for predicting population in China, combining forecasting based on several models (e.g., Liu and Xu [24]) is mentioned but not investigated, though it may give better results. In this paper, some common models for population prediction are discussed, experiments with Chinese population using Matlab/SAS are done for further analysis and comparison with these methods, basic modelling ideas of different models are also studied.

Elementary function model, gray system model, genetic algorithm, combining forecasting and nonparametric estimation are talked over briefly. Markov chains are studied to predict population in theory, specific method is proposed and some improvements are put forward.

Three models are discussed in details, they're linear regression model, differential equation model and ARIMA model. The comparison shows that ARIMA model is much better than the others. However, in another experiment, neural network model has been found to be much more efficient. Both ARIMA model and neural network are good chooses in predicting Chinese population, but lots of work remains to be done before I can draw any firm conclusions.

## Key Words:

population forecasting linear regression differential equation time series  
models comparison

# 第一章 引言

人口问题是一个世界性的问题，人口过度增长会引发许多问题，特别是一些经济不发达国家的人口过度增长，影响了整个国家的经济发展、社会安定和人民生活水平的提高。中国是目前世界人口最多的国家，人口问题也是中国面临的重大问题。控制人口的过度增长，必须要对未来人口作出准确的预测。本文探讨了多种人口拟合与预测的方法，通过编程计算，将各种方法应用于中国人口的拟合及预测中，并比较了各种模型和方法的特点和应用价值。

其中，对线性回归模型、微分方程模型和时间序列 ARIMA 模型三种模型进行了充分的讨论，并对三种模型实际结果做出了详细的对比分析。状态转移模型作为一种理论性的随机模型，对模型形式和性质进行了理论上的研究。神经网络模型等一些新模型也作了简要介绍，并对其中部分比较实用的模型进行了编程实验。

# 第二章 初等函数模型

初等函数模型使用简单的线性函数、对数线性函数、指数函数、多项式和幂函数等初等函数来表达人口增长规律。通过统计数据对几种初等函数中的有限参数进行拟合，便得到最终的模型形式。初等函数模型中的参数常被人为地赋予某些实际含意，如人口基数、增长率等，使初等函数模型具有直观的解释意义。

## §2.1 一元线性函数与对数线性函数模型

一元线性函数模型即线性自然增长模型，认为人口以最简单的线性方式自然增长。对数线性函数模型是一元线性函数模型的一种变形，下面主要讨论线性自然增长模型。

### 参数说明:

$t$  时间

$i$  人口净增长率

$P(t)$   $t$  时的人口总数

注:如无特殊说明, 以上参数含义在全文中一致.

### 基本假定:

$$P(t+s) - 1 = (P(t) - 1) + (P(s) - 1) \quad t \geq 0, s \geq 0. \quad (2.1.1)$$

其中  $P(t)$  一阶可微,  $P(0) = 1$ .

实际意义:1 单位人口经过  $t+s$  时期人口净增长量, 等于它经过  $t$  个时期净增长量加上经过  $s$  个时期的净增长量.

注:线性自然增长类似利息理论中的单利, 它在同样时期的增长的绝对量保持为常数, 即  $P(t+s) - P(t)$  不依赖于  $t$ .

**模型形式:** 由基本假定可以得到一元线性函数模型形式为:(详见附录 A.1)

$P(t) = a + b \cdot t, t \geq 0$ . 其中  $b = i$  为人口净增长率.

对数线性函数模型形式为: $\log P(t) = a + b \cdot t, t \geq 0$

**拟合与预测:** 使用 Matlab 对人口数据作一阶多项式拟合:

```
>> t=1978:1:2007;
>> y=[96259 97542 98705 100072 101654 103008 104357 105851 ...
107507 109300 111026 112704 114333 115823 117171 118517 ...
119850 121121 122389 123626 124761 125786 126743 127627 ...
128453 129227 129988 130756 131448 132129];
>> result=polyfit(t,y,1)
```

结果为:

```
result =
1.0e+006 *
0.0013    -2.4781
```

拟合模型:  $P(t) = -2478100 + 1300 \cdot t, i = 1300$

实际拟合值与相对误差如表:

年份 (1978-2007)	实际总人口 (单位: 万人)	拟合总人口 (单位: 万人)	相对误差 (%)
1978	96259.00	97377.92	-1.16
1979	97542.00	98679.98	-1.17
1980	98705.00	99982.04	-1.29
1981	100072.00	101284.09	-1.21
1982	101654.00	102586.15	-0.92
1983	103008.00	103888.21	-0.85
1984	104357.00	105190.27	-0.80
1985	105851.00	106492.33	-0.61
1986	107507.00	107794.39	-0.27
1987	109300.00	109096.44	0.19
1988	111026.00	110398.50	0.57
1989	112704.00	111700.56	0.89
1990	114333.00	113002.62	1.16
1991	115823.00	114304.68	1.31
1992	117171.00	115606.74	1.34
1993	118517.00	116908.80	1.36
1994	119850.00	118210.85	1.37
1995	121121.00	119512.91	1.33
1996	122389.00	120814.97	1.29
1997	123626.00	122117.03	1.22
1998	124761.00	123419.09	1.08
1999	125786.00	124721.15	0.85
2000	126743.00	126023.21	0.57
2001	127627.00	127325.26	0.24
2002	128453.00	128627.32	-0.14
2003	129227.00	129929.38	-0.54
2004	129988.00	131231.44	-0.96
2005	130756.00	132533.50	-1.36
2006	131448.00	133835.56	-1.82
2007	132129.00	135137.62	-2.28

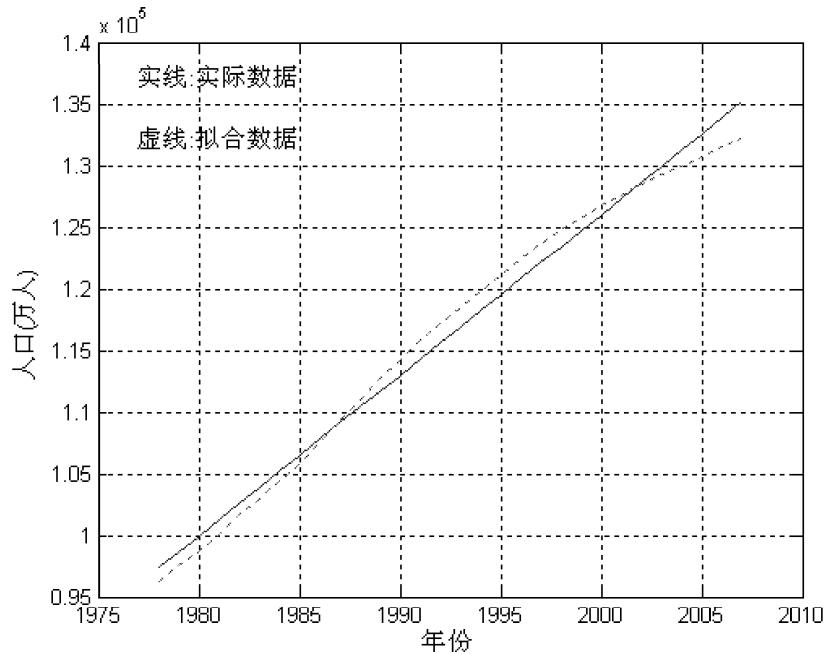


图 2.1: 线性增长

**评价:** 可见中国人口走势大致符合线性自然增长规律, 但线性自然增长由于假设条件忽略太多因素, 拟合误差较大, 多用于人口增长率基本保持不变的人口增长平稳期, 也可用于人口增长率的分析, 不适用于预测.

## §2.2 指数函数模型

指数函数模型用指数函数形式的简单函数来描述人口增长, 又称为指数自然增长模型.

**基本假定:**

$$P(t+s) = P(t) \cdot P(s) \quad t \geq 0, s \geq 0. \quad (2.2.1)$$

其中  $P(t)$  一阶可微,  $P(0) = 1$ .

实际意义: 1 单位人口经过  $t+s$  时期增长人口总量, 等于以经  $t$  个时期增长后的人口总量为起点, 再增长  $s$  个时期所得到的人口总量.

注: 指数自然增长类似利息理论中的复利, 它是增长的相对比率保持为常数, 即  $\frac{P(t+s)-P(t)}{P(t)}$  不依赖于  $t$ .

**模型形式:** 由基本假定可以得到指数函数模型形式为 (详见附录 A.2)

$$P(t) = ae^{bt}, t \geq 0 \text{ 或 } \log P(t) = \log a + b \cdot t, t \geq 0.$$

**拟合与预测:** 使用 Matlab 对人口数据作对数一阶多项式拟合:

```
>> t=1978:1:2007;
>> y=[96259 97542 98705 100072 101654 103008 104357 105851 ...
107507 109300 111026 112704 114333 115823 117171 118517 ...
119850 121121 122389 123626 124761 125786 126743 127627 ...
128453 129227 129988 130756 131448 132129];
>> ly=log(y);
>> result=polyfit(t,ly,1)
```

结果为:

```
result =
0.0114 -10.9869
```

拟合模型:  $\log P(t) = -10.9869 + 0.0114 \cdot t$ ,  $\log(1+i) = 0.0114$ ,

或  $P(t) = e^{-10.9869+0.0114 \cdot t}$ ,  $i = 0.0114$ .

实际拟合值与相对误差如表:

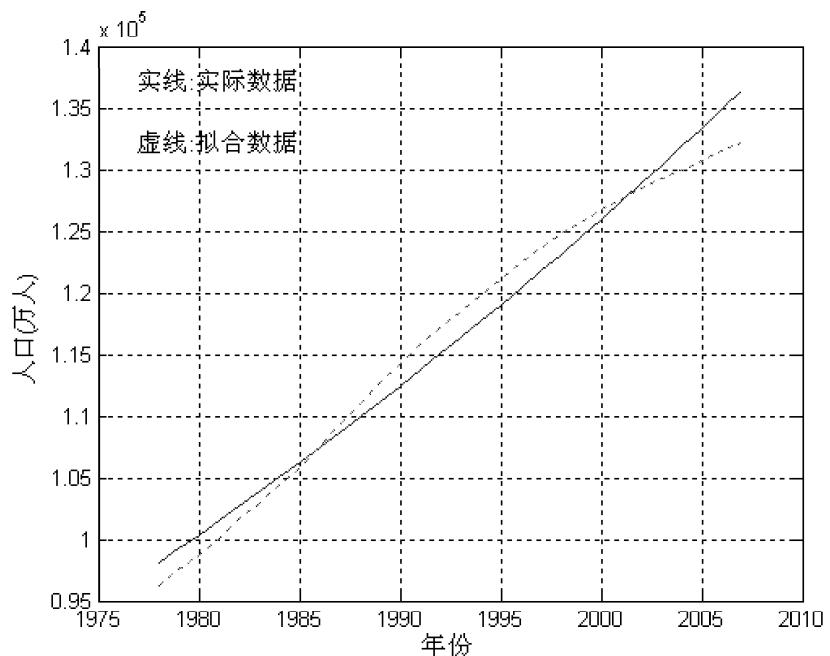


图 2.2: 指数增长

年份 (1978-2007)	实际总人口 (单位: 万人)	拟合总人口 (单位: 万人)	相对误差 (%)
1978	96259.00	98113.41	-1.93
1979	97542.00	99234.87	-1.74
1980	98705.00	100369.15	-1.69
1981	100072.00	101516.40	-1.44
1982	101654.00	102676.76	-1.01
1983	103008.00	103850.38	-0.82
1984	104357.00	105037.42	-0.65
1985	105851.00	106238.02	-0.37
1986	107507.00	107452.35	0.05
1987	109300.00	108680.56	0.57
1988	111026.00	109922.81	0.99
1989	112704.00	111179.25	1.35
1990	114333.00	112450.06	1.65
1991	115823.00	113735.40	1.80
1992	117171.00	115035.42	1.82
1993	118517.00	116350.31	1.83
1994	119850.00	117680.22	1.81
1995	121121.00	119025.34	1.73
1996	122389.00	120385.83	1.64
1997	123626.00	121761.87	1.51
1998	124761.00	123153.64	1.29
1999	125786.00	124561.32	0.97
2000	126743.00	125985.09	0.60
2001	127627.00	127425.13	0.16
2002	128453.00	128881.63	-0.33
2003	129227.00	130354.79	-0.87
2004	129988.00	131844.77	-1.43
2005	130756.00	133351.80	-1.99
2006	131448.00	134876.04	-2.61
2007	132129.00	136417.71	-3.25

**评价:** 指数自然增长规律其假设条件同样忽略太多因素, 误差比线性自然增长更大一些, 可见不适用于中国人口预测.

### §2.3 多项式模型

**模型形式:**  $n$  阶多项式模型  $P(t) = \sum_k^n a_k t^k, t \geq 0$  其中  $n = 1$  的情况已在自然增长模型中讨论过, 由于高阶多项式会产生振荡, 下面主要讨论  $n = 2, 3$  的情况.

**拟合与预测:** 使用 Matlab 对人口数据作对数二阶多项式拟合:

```
>> t=1978:1:2007;
>> t=t/1000;
>> y=[96259 97542 98705 100072 101654 103008 104357 105851 ...
107507 109300 111026 112704 114333 115823 117171 118517 ...
119850 121121 122389 123626 124761 125786 126743 127627 ...
128453 129227 129988 130756 131448 132129];
>> result=polyfit(t,y,2)
```

结果为:

```
result = 1.0e+007 *
-1.8183    7.3760   -7.4663
```

拟合模型:  $P(t) = 1.0e + 007 * (-1818.3t^2 + 7376.0t - 7.4663)$

同样可得三阶多项式拟合模型

$$P(t) = 1.0e + 010 * (-91.7t^3 + 546.0t^2 - 1084.2t - 0.7175)$$

二阶多项式实际拟合值与相对误差如表:

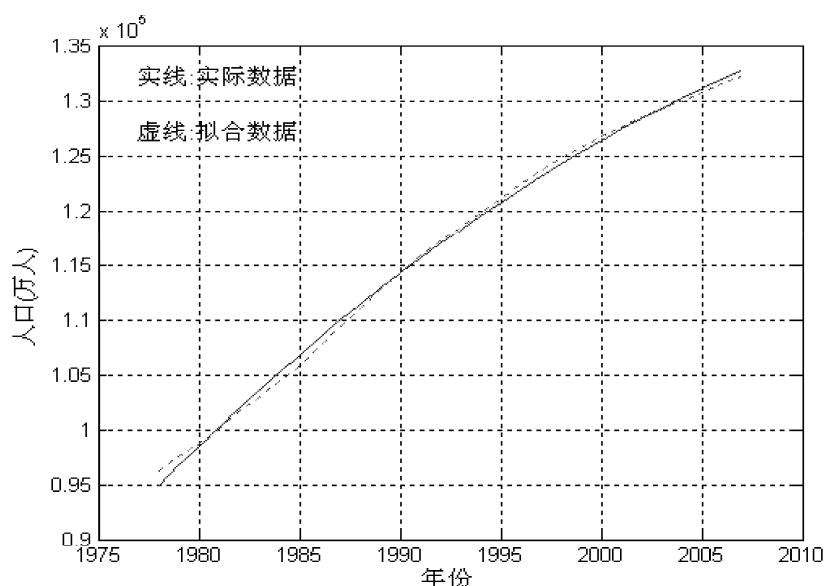


图 2.3: 二阶多项式拟合

年份 (1978-2007)	实际总人口 (单位: 万人)	拟合总人口 (单位: 万人)	相对误差 (%)
1978	96259.00	94917.21	1.39
1979	97542.00	96728.38	0.83
1980	98705.00	98503.18	0.20
1981	100072.00	100241.62	-0.17
1982	101654.00	101943.70	-0.28
1983	103008.00	103609.41	-0.58
1984	104357.00	105238.76	-0.84
1985	105851.00	106831.74	-0.93
1986	107507.00	108388.35	-0.82
1987	109300.00	109908.60	-0.56
1988	111026.00	111392.49	-0.33
1989	112704.00	112840.01	-0.12
1990	114333.00	114251.16	0.07
1991	115823.00	115625.95	0.17
1992	117171.00	116964.37	0.18
1993	118517.00	118266.43	0.21
1994	119850.00	119532.12	0.27
1995	121121.00	120761.45	0.30
1996	122389.00	121954.41	0.36
1997	123626.00	123111.01	0.42
1998	124761.00	124231.24	0.42
1999	125786.00	125315.11	0.37
2000	126743.00	126362.61	0.30
2001	127627.00	127373.75	0.20
2002	128453.00	128348.52	0.08
2003	129227.00	129286.93	-0.05
2004	129988.00	130188.97	-0.15
2005	130756.00	131054.65	-0.23
2006	131448.00	131883.96	-0.33
2007	132129.00	132676.90	-0.41

**评价:** 二阶多项式拟合误差较小, 优于自然增长模型, 但模型参数意义不明, 也不适用于预测.

## §2.4 幂函数模型

模型形式:  $P(t) = at^b, t \geq 0$

拟合与预测: 使用 Matlab 对人口数据作幂数模型拟合: 拟合参数即最优化问题

$$\min F(a, b) = \sum_{j=1}^{30} [P(t_j) - at_j^b]^2$$

的解. 编写 M- 文件 curvefun.m

```
function f=curvefun(x,tdata)
f=x(1)*(tdata.^x(2))%x(1)为a,x(2)为b
```

主程序:

```
>> t=1978:1:2007;
>> t=t/2000;
>> y=[96259 97542 98705 100072 101654 103008 104357 105851 ...
107507 109300 111026 112704 114333 115823 117171 118517 ...
119850 121121 122389 123626 124761 125786 126743 127627 ...
128453 129227 129988 130756 131448 132129];
>> y=y/10000;
>> x0=[0,0];%迭代初始值
>> result=lsqcurvefit('curvefun',x0,t,y)
```

结果为:

```
result =
12.5767    22.0781
```

拟合模型:  $P(t) = 10000 * 12.5767(t/2000)^{22.0781}$

实际拟合值与相对误差如表:

年份 (1978-1987)	实际总人口 (单位: 万人)	拟合总人口 (单位: 万人)	相对误差 (%)
1978	96259.00	98516.82	-2.35
1979	97542.00	99622.33	-2.13
1980	98705.00	100739.67	-2.06
1981	100072.00	101868.98	-1.80
1982	101654.00	103010.36	-1.33
1983	103008.00	104163.95	-1.12
1984	104357.00	105329.86	-0.93
1985	105851.00	106508.23	-0.62
1986	107507.00	107699.18	-0.18
1987	109300.00	108902.84	0.36

年份 (1988-2007)	实际总人口 (单位: 万人)	拟合总人口 (单位: 万人)	相对误差 (%)
1988	111026.00	110119.33	0.82
1989	112704.00	111348.78	1.20
1990	114333.00	112591.34	1.52
1991	115823.00	113847.13	1.71
1992	117171.00	115116.28	1.75
1993	118517.00	116398.93	1.79
1994	119850.00	117695.22	1.80
1995	121121.00	119005.29	1.75
1996	122389.00	120329.27	1.68
1997	123626.00	121667.31	1.58
1998	124761.00	123019.54	1.40
1999	125786.00	124386.12	1.11
2000	126743.00	125767.18	0.77
2001	127627.00	127162.87	0.36
2002	128453.00	128573.34	-0.09
2003	129227.00	129998.74	-0.60
2004	129988.00	131439.22	-1.12
2005	130756.00	132894.94	-1.64
2006	131448.00	134366.03	-2.22
2007	132129.00	135852.67	-2.82

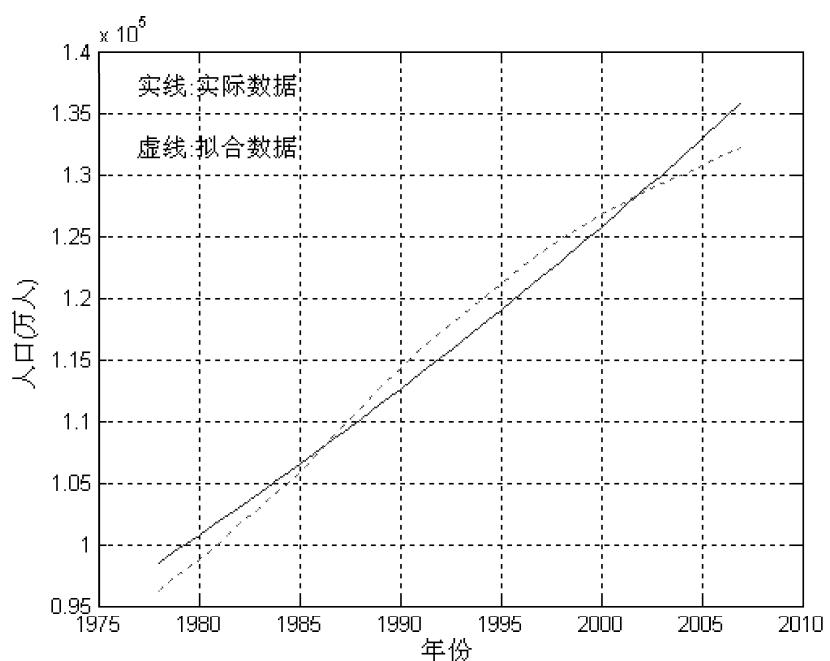


图 2.4: 幂函数增长

**评价:** 幂函数增长规律, 参数拟合难度大, 拟合误差也较大.

### 第三章 线性回归模型

线性回归模型是在实际中得到广泛应用的一种经典的线性统计模型. 线性回归模型形式简单, 参数意义明显, 模型可以解释, 具有显式解, 便于计算, 并且解具有良好的统计性质. 正是由于这些优点, 线性回归模型得到了广泛的研究与应用. 线性回归模型其实在前面的自然增长模型和初等函数模型中已有论述, 虽然不是用于人口拟合预测的最佳选择, 但鉴于该模型特点, 仍具有很好的对比和参考价值.

**模型形式:** 含有  $p - 1$  个自变量的线性回归模型的一般形式为

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + e,$$

其中  $Y$  为因变量 (观测变量),  $X_i$  为自变量 (设计变量),  $\beta_i$  为未知参数,  $e$  为随机误差.

对因变量  $Y$  和自变量  $X_1, \dots, X_{p-1}$  进行了  $n$  次观察, 得到  $n$  组数据

$(y_i, x_{i1}, \dots, x_{ip-1}), i = 1, \dots, n$ , 满足

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1} + e_i, i = 1, \dots, n$$

$$\text{记 } y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix},$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}, e = \begin{pmatrix} e_0 \\ e_1 \\ \vdots \\ e_n \end{pmatrix},$$

假设  $\text{rank}(X) = p$ , 即  $X$  列满秩;  $e_i (i = 1, \dots, n)$  互不相关, 均值皆为零, 且有公共方差  $\sigma^2$

线性回归模型的矩阵形式为  $y = X\beta + e, E(e) = 0, \text{Cov}(e) = \sigma^2 I$ .

称  $\beta_0$  为常数项,  $\beta' = (\beta_1, \dots, \beta_{p-1})$  为回归系数.

**模型的估计与性质 - 最小二乘估计:** 模型的最小二乘估计为  $\hat{\beta} = (X'X)^{-1}X'y$ ,

最小二乘估计  $\hat{\beta}$  的性质:

(1) 无偏性:  $E(\hat{\beta}) = \beta$

(2) 最小方差性 (Gauss-Markov 定理): 对任意  $p \times 1$  向量  $c'\hat{\beta}$  为  $c'\beta$  的唯一的最小方差无偏估计, 这里  $\hat{\beta}' = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_{p-1})$

(3)  $\hat{\sigma}^2 = \|y - X\hat{\beta}\|^2/(n-p)$  为  $\sigma^2$  的无偏估计.

**区间估计:**  $\beta_i$  的置信度为  $1-\alpha$  的置信区间为

$$[\hat{\beta}_i - t_{\alpha/2} \sqrt{(X'X)^{-1}_{ii}} \hat{\sigma}, \hat{\beta}_i + t_{\alpha/2} \sqrt{(X'X)^{-1}_{ii}} \hat{\sigma}]$$

其中  $t_{\alpha/2} = t_{\alpha/2}(n-p-1)$  为自由度  $n-p-1$  的 t 分布的左分位数.

### 点预测与区间预测

1. 点预测: 对于给定的一组自变量值  $X_0 = (X_{10}, \dots, X_{p0})$ , 代入回归方程可计算出因变量  $Y_0$  的点预测值为  $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_{10} + \dots + \hat{\beta}_{p0} X_{p0} = X_0 \hat{\beta}$

2. 区间预测: 对于给定的显著性水平  $\alpha (0 < \alpha < 1)$ ,  $Y_0$  的置信度  $1-\alpha$  的置信区间为  $[\hat{Y}_0 - t_{\alpha/2} \sqrt{\frac{S_e}{n-p-1}} \sqrt{1+M_0}, \hat{Y}_0 + t_{\alpha/2} \sqrt{\frac{S_e}{n-p-1}} \sqrt{1+M_0}]$  其中  $M_0 = X_0(X'X)^{-1}X'_0$ ,

$$S_e = \sum_{i=1}^n (Y_i - E(Y))^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2, t_{\alpha/2} = t_{\alpha/2}(n-p-1).$$

### 可线性化的模型

1. 双曲函数模型:  $\frac{1}{Y} = a + b \frac{1}{X} + e, Y^* = \frac{1}{Y}, X^* = \frac{1}{X}$  化为一元线性回归模型  $Y^* = a + bX^* + \mu$

2. 幂函数模型:  $\frac{1}{Y} = aX^b e^\mu, \ln Y = \ln a + b \ln X + \mu, Y^* = \ln Y, a^* = \ln a, X^* = \ln X$  化为一元线性回归模型  $Y^* = a^* + bX^* + \mu$

3. 指数函数模型:  $Y = ae^{bX+\mu}, \ln Y = \ln a + bX + \mu, Y^* = \ln Y, a^* = \ln a$  化为一元线性回归模型  $Y^* = a^* + bX + \mu$

4. 倒指数函数模型:  $Y = ae^{\frac{b}{X}+\mu}, \ln Y = \ln a + \frac{b}{X} + \mu, Y^* = \ln Y, a^* = \ln a, X^* = \frac{1}{X}$  化为一元线性回归模型  $Y^* = a^* + bX^* + \mu$

5. 对数函数模型:  $Y = a + b \ln X + \mu, X^* = \ln X$  化为一元线性回归模型  $Y = a + bX^* + \mu$

6.S型曲线模型:  $Y = \frac{1}{a+be^{-x}+\mu}, \frac{1}{Y} = a + be^{-x} + \mu, Y^* = \frac{1}{Y}, X^* = e^{-x}$  化为一元线性回归模型  $Y^* = a + bX^* + \mu$

**拟合与预测:** 使用 Matlab 对人口数量 (因变量) 与时间 (自变量) 作一元线性回归

```
>> t=[1978:1:2007]';
>> X=[ones(30,1) t];
>> Y=[96259 97542 98705 100072 101654 103008 104357 105851 ...
107507 109300 111026 112704 114333 115823 117171 118517 ...
119850 121121 122389 123626 124761 125786 126743 127627 ...
128453 129227 129988 130756 131448 132129]';
>> [b,bint,r,rint,stats]=regress(Y,X);
>> b,bint,stats
```

结果为:

```
b =
1.0e+006 *
-2.4781
0.0013
```

```

bint =
1.0e+006 *
-2.5961 -2.3601
0.0012 0.0014
stats =
1.0e+006 *
0.0000 0.0020      0      1.87791

```

$$\hat{\beta}_0 = 1.0e + 006 * -2.4781, \text{ 置信区间 } [1.0e + 006 * -2.5961, 1.0e + 006 * -2.3601];$$

$$\hat{\beta}_1 = 0.0013 * 1.0e + 006 \text{ 置信区间 } [1.0e + 006 * 0.0012, 1.0e + 006 * 0.0014];$$

$r^2 = 0, F = 1.0e + 006 * 0.0020, p = 0$ , 由误差方差  $1.0e + 006 * 1.87791 p < 0.05$  知线性模型成立.

综上得到拟合模型:  $P(t) = 1.0e + 006 * -2.4781 + 1.0e + 006 * 0.0013 \cdot t$

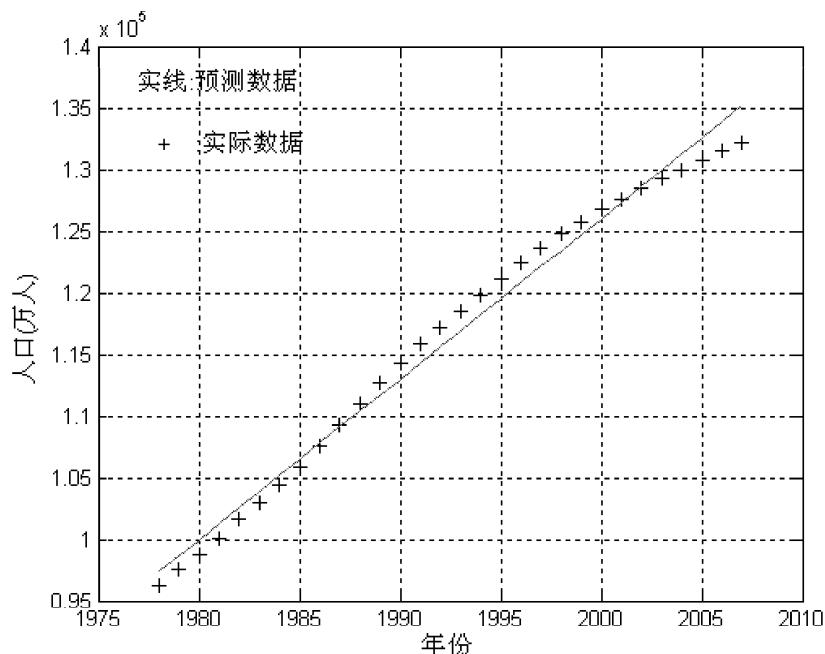


图 3.1: Matlab 一元线性回归

使用 Matlab 对人口数量 (因变量) 与时间 (自变量)、人均国内生产总值 (自变量)、居民消费价格指数 (自变量)、普通高等学校数 (自变量) 作多元线性回归:

```

>> t=[1978:1:2007]';
>> pt=[381 419 463 492 528 583 695 858 963 1112 1366 1519 1644 ...
1893 2311 2998 4044 5046 5846 6420 6796 7159 7858 8622 ...
9398 10542 12336 14053 16165 18934]’%人均国内生产总值
>> pr=[100.7 101.9 107.5 102.5 102.0 102.0 102.7 109.3 ...
106.5 107.3 118.8 118.0 103.1 103.4 106.4 114.7 124.1 ...
117.1 108.3 102.8 99.2 98.6 100.4 100.7 99.2 101.2 ...
103.9 101.8 101.5 104.8]’%居民消费价格指数
>> un=[598 636 675 743 811 879 947 1016 1028 1040 1052 ...
1064 1075 1075 1053 1065 1080 1054 1032 1020 1022 ...
1071 1041 1225 1396 1552 1731 1792 1867 1908]’%普通高等学校数
>> X=[ones(30,1) t pt pr un];
>> Y=[96259 97542 98705 100072 101654 103008 104357 105851 ...
107507 109300 111026 112704 114333 115823 117171 118517 ...
119850 121121 122389 123626 124761 125786 126743 127627 ...
128453 129227 129988 130756 131448 132129]’;
>> [b,bint,r,rint,stats]=regress(Y,X);
>> b,bint,stats

```

结果为:

```

b =
1.0e+006 *
-3.1397
0.0016
-0.0000
0.0000
-0.0000

bint =
1.0e+006 *
-3.2218 -3.0576
0.0016 0.0017
-0.0000 -0.0000
0.0000 0.0001
-0.0000 -0.0000

stats =
1.0e+005 *
0.0000 0.0723 0 1.3353

```

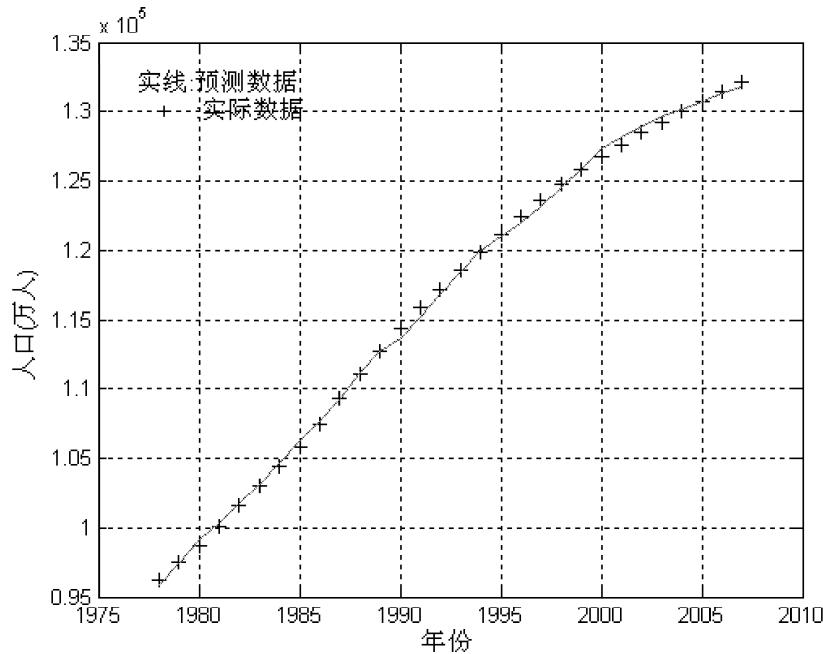


图 3.2: Matlab 多元线性回归

使用 SAS 对人口数量 (因变量) 与时间 (自变量) 作一元线性回归:

```
data population;
input time pop gdp price un;
cards;
1978 96259 381 100.7 598
.../*数据输入,数据中间部分省略*/
2007 132129 18934 104.8 1908 ;
proc reg data=population;
model pop=time;
plot pop*time;
run;
```

结果见图 3.3:

使用 SAS 对人口数量 使用逐步回归, 作多元线性回归, 结果见图 3.4:

```
proc reg data=population;
model pop=time gdp price un/selection=stepwise;
plot pop*time;
run;
```

**评价:** 线性回归模型可对中国人口作出较好的拟合, 考虑到人口受众多因素影响, 其长期变化规律不可能用较少的固定参数决定, 所以线性回归模型只适用于短期预测. 另外, 多元回归拟合精确度比一元回归更高一些, 但要考虑自变量选择、统计数据来源等问题.

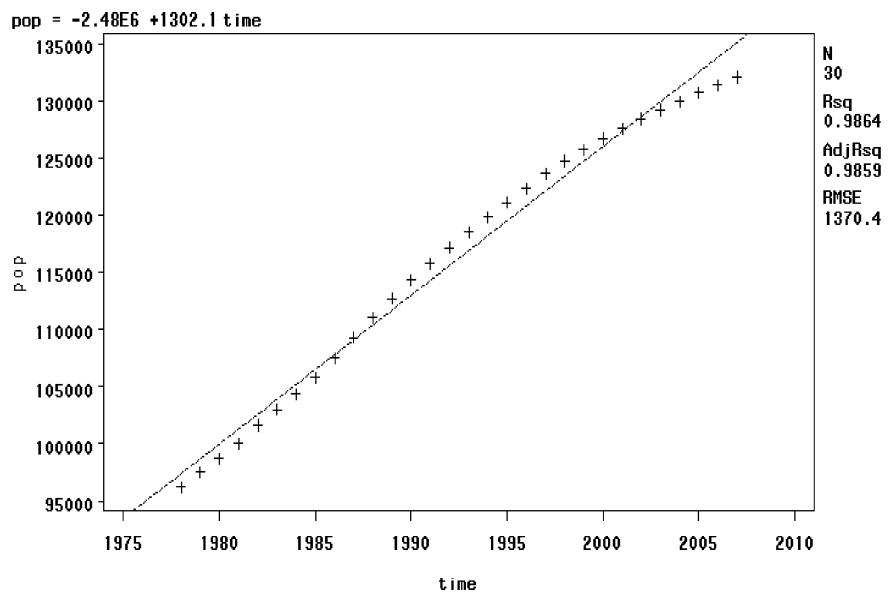


图 3.3: SAS 一元线性回归

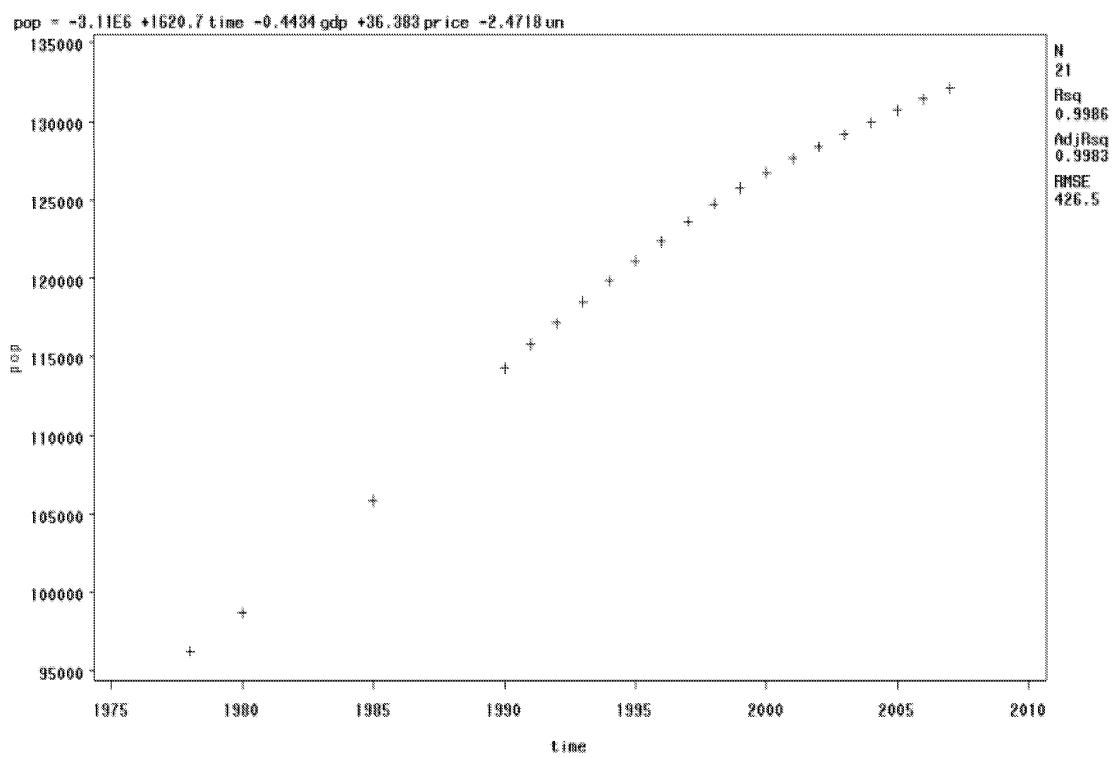


图 3.4: SAS 多元线性回归

## 第四章 微分方程模型

微分方程在现实生活中有着极其广泛的应用，许多复杂的系统都可以通过各种假设和简化使用微分方程来描述。微分方程本质上是先建立变量和变量变化率之间的方程关系，然后通过求解建立的微分方程，得到因变量对自变量的显式表达式，模型最终由此显式解确立。人口系统同样可以使用微分方程进行描述。利用微分方程建立起的人口增长模型在历史上已被充分的研究利用。最早的两种基于微分方程的经典人口模型是马尔萨斯（Malthus）人口模型和阻滞（Logistic）增长模型。

### §4.1 马尔萨斯人口模型

**模型建立：**假设人口增长与原人口数成正比，即

$$P(t+1) - P(t) = iP(t)$$

令  $\alpha = i + 1$

$$P(t+1) = \alpha P(t)$$

这是一个线性映射  $f(x) = \alpha x$  的迭代，易得

$$P(t) = \alpha P(t-1) = \alpha^2 P(t-2) = \cdots = \alpha^t P(0)$$

即人口增长呈几何级数。

**注：**马尔萨斯人口模型即前面数讨论的指数自然增长模型。

**评价：**马尔萨斯人口模型在人口具有充分的生存和增长条件时，比如人口增长初期是可取的，但实际上影响人口增长的因素非常多，必须对模型加以修正。

### §4.2 阻滞增长模型

**基本假定：**（1）总量增长较小时，总量增长率几乎与总量成正比。  
（2）总量增长较大时，总量增长率就变为负值。

**模型建立:** 微分方程

$$P' = aP\left(1 - \frac{P}{N}\right), a > 0, N > 0.$$

满足假设:  $a$  为较小时的总量增长率, 而  $N$  代表一种理想总量或承载量,  $P$  较小时  $1 - \frac{P}{N} \approx 1$  方程近似为  $P' = aP$ , 而  $P > N$  时  $P' < 0$ .

不失一般性, 设  $N = 1$ ,  $P(t)$  表示  $t$  时刻人口占理想总量的比例,

方程简化成  $P' = f_a(P) = aP(1 - P)$ , 这是一个一阶、自治、非线性方程.

**模型解释:** 将模型化为离散形式,

$$P(t+1) - P(t) = aP(t) - aP(t)^2$$

由于生存资源的有限性, 人口之间的竞争约束必定影响它们的数量增长, 其中  $-aP(t)^2$  为竞争约束项, 表明单位时间内由于竞争或约束而减少的群体个数与成员相遇次数的统计平均值 (从而与  $P^2$ ) 成正比.

**模型分析与求解:** 对原微分方程分离变量积分得  $\int \frac{dP}{P(1-P)} = \int adt$ ;  $\int \left(\frac{1}{P} + \frac{1}{1-P}\right) = \int adt$  解得  $P(t) = \frac{Ke^{at}}{1+Ke^{at}}$ ;  $K$  为微积分常数, 将  $t = 0$  代入上式可得  $K = \frac{P(0)}{1-P(0)}$  从而

$$P(t) = \frac{P(0)e^{at}}{1 - P(0) + P(0)e^{at}}$$

平衡解: 由  $P' = aP(1 - P) = 0$  得  $P = 0$  和  $P = 1$ , 代入解的形式分别有  $P = 0$  和  $P = 1$ , 故  $P(t) \equiv 1$  和  $P(t) \equiv 0$  为原方程的两个平衡解. 解图像:

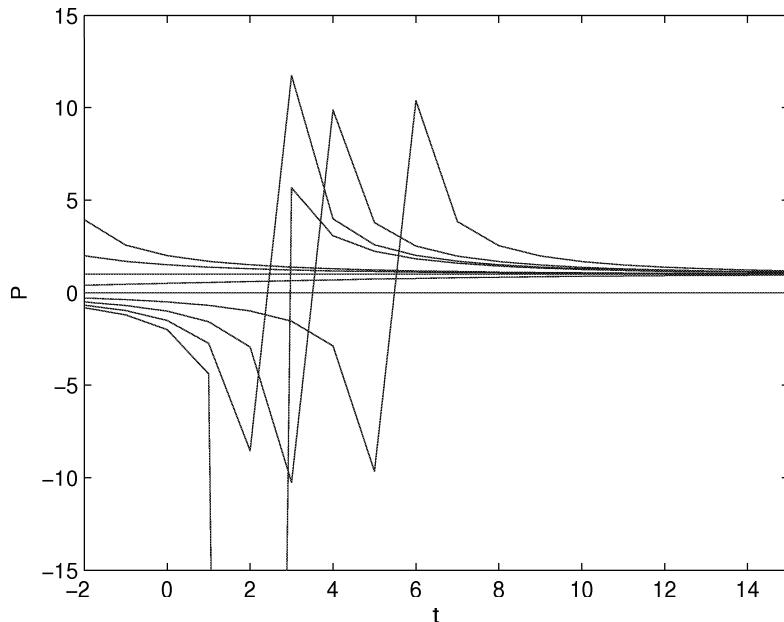


图 4.1: Logistic 增长模型的解图像

**注:** 由解图像可看出所有对应于  $P(0) > 0$  的解都趋向于理想总量  $P(0) = 1$ , 满足之前的假设.

阻滞增长模型建立在一个经典的非线性映射 -Logistic 映射之上, 通过研究该映射, 可以更好地理解阻滞增长模型所描述的规律以及人口系统的复杂性.

**Logistic 映射** 非线性映射  $f_a(P) = aP(1 - P)$ ,  $P \in [0, 1]$  和迭代  $P_{n+1} = aP_n(1 - P_n)$ ,  $n$  为正整数, 可以用于模拟某个生物种群的一代与下一代的关系: 其中种群第  $n$  代的总数为  $P_n$ ,  $P_n$  已除以理想总量 (种群数目可能达到的某个最大值) 作规范化, 故  $P_n \in [0, 1]$ . 模型表明下一代的总数  $P_{n+1}$  主要依赖于上一代的总数  $P_n$ , 还依赖于资源供应系数  $a$ , 由前面的分析, 该模型还考虑了种群之间的竞争.

映射  $f(x) = \mu x(1 - x)$ ,  $\mu \in [0, 4]$ ,  $x \in [0, 1]$  称为 Logistic 映射, Logistic 映射形成的迭代称为 Feigenbaum 迭代:  $x_{n+1} = \mu x_n(1 - x_n)$

**混沌与分岔** Feigenbaum 迭代的轨迹图:

$\mu = 2, x_0 = 0.2$  时, 前 4 个值为 0.3, 0.42, 0.4872, 0.4997, 第 5 个数及以后都为 0.5.

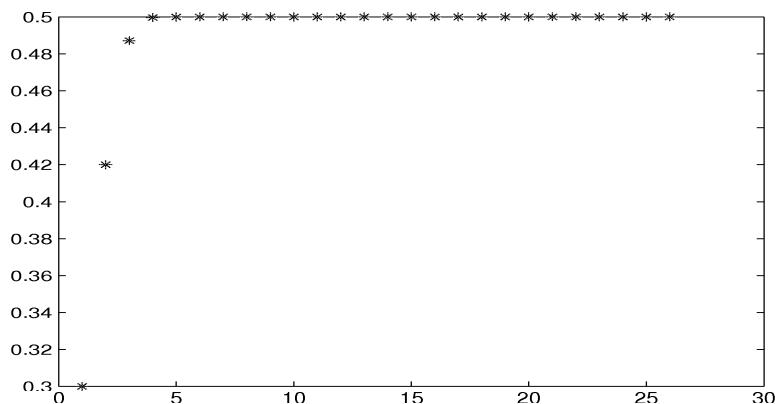


图 4.2: 稳定的 Feigenbaum 迭代

$\mu = 4, x_0 = 0.07$  时, 迭代轨迹不稳定, 没有趋近于常数的迹象.

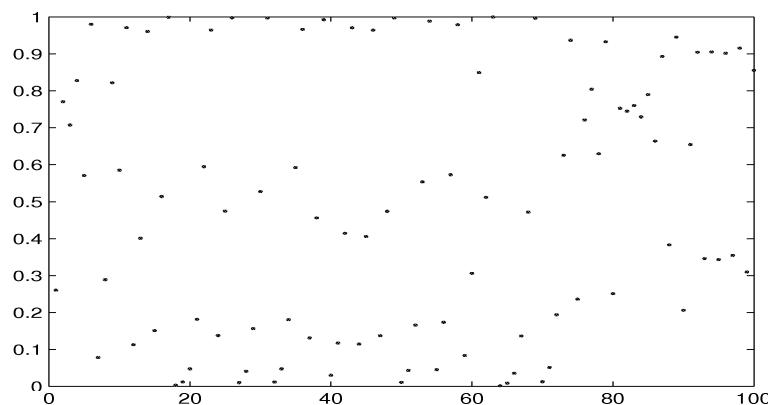


图 4.3: 混沌的 Feigenbaum 迭代

**不动点:**

令

$$f(x) = \mu x(1 - x) = x$$

解得

$$x_1 = 0, x_2 = \frac{\mu - 1}{\mu}$$

$f(x) = 3.2x(1 - x)$  的周期点:

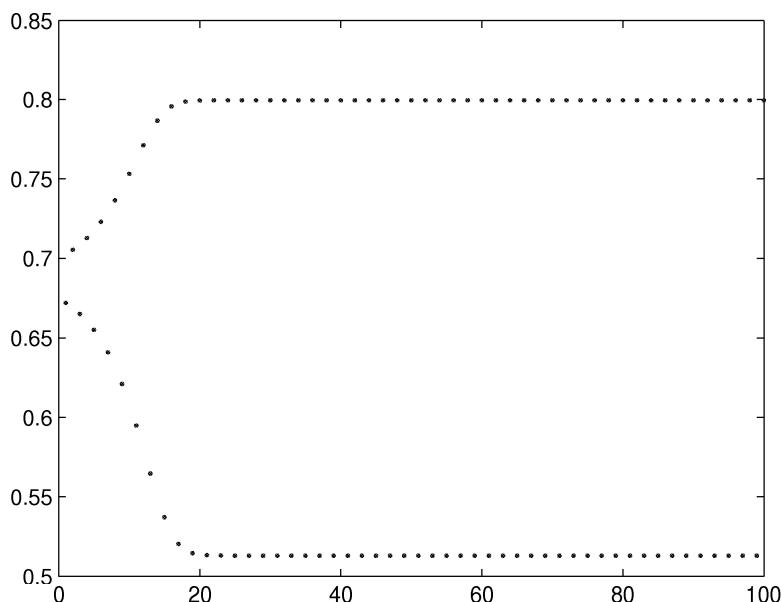


图 4.4: 迭代的周期点

轨迹在两个点之间跳动.

**迭代分岔:**

1. 当  $0 < \mu < 1$  时, 对  $\forall x_0 \in [0, 1]$  有  $x_n \rightarrow 0(n \rightarrow +\infty)$
  2. 当  $1 < \mu < 3$  时, 对  $\forall x_0 \in (0, 1)$  轨道将逐渐趋向于某个数值.
  3. 当  $3 < \mu < \sqrt{6} + 1$  时,  $x_n$  轨道将围绕两个数值振动,
- $x_{2k-1} \rightarrow x_3^*, x_{2k} \rightarrow x_4^*(k \rightarrow +\infty)$  原有的稳定点失稳并产生新的稳定的周期点, 出现分叉.
4. 当  $c_k < \mu < c_{k+1}, (k = 1, 2 \dots)$  时, 映射  $f$  有稳定的周期  $2^k$  点, 而周期  $2^{k-1}$  点不稳定, 且极限  $\lim_{k \rightarrow \infty} c_k = c_\infty = 3.569945557391 \dots$

**分岔图:**

Logistic 映射  $f(x) = \mu x(1 - x)$  当  $3 < \mu < 4$  时的分岔图:

```

>> clear;
>> x=0.2;
>> for u=2.6:0.0001:4
    for i=1:18
        x1=u*x*(1-x);
        x=x1;
        if i>9
            plot(u,x);
            hold on;
        end
    end
end

```

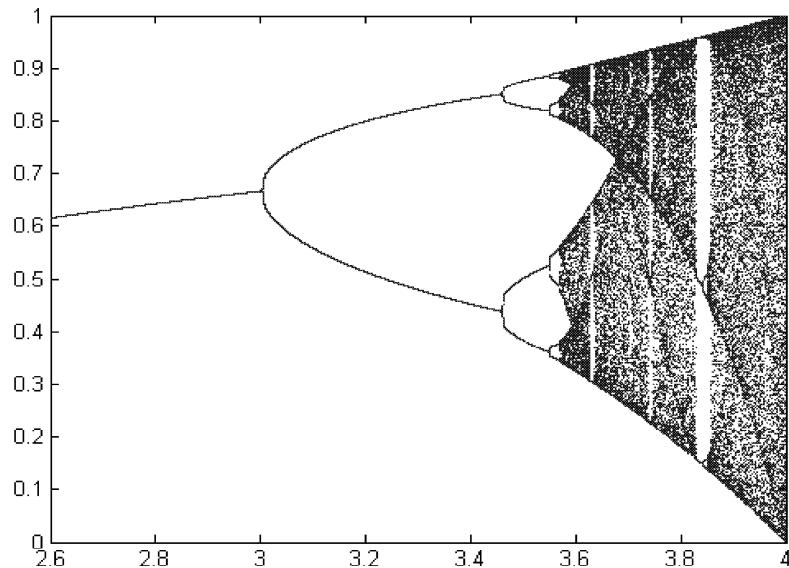


图 4.5: 分岔图

### §4.3 收割模型

收割模型是对阻滞增长模型的一种简单改进, 但方程却不再有显式解.

**常值收割模型** 考虑物种收割来修改阻滞 Logistic 增长模型, 假设增长遵循  $\alpha = 1$  的 Logistic 增长模型同时以常速率  $h$  被收割, 微分方程为

$$P' = P(1 - P) - h, h \geq 0$$

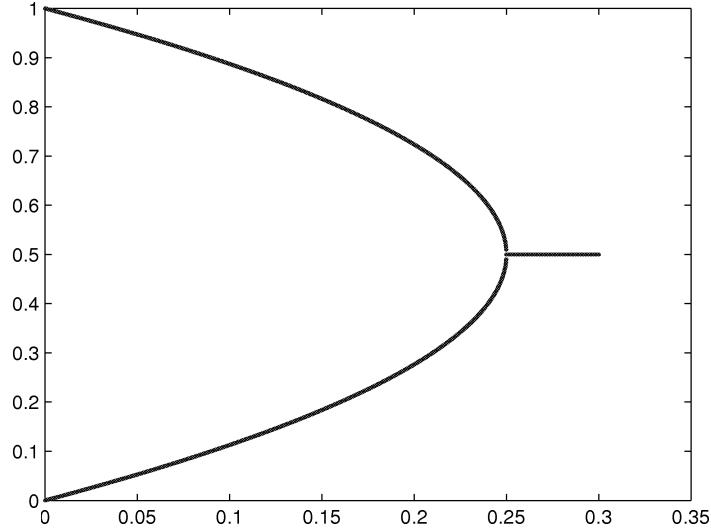


图 4.6: 周期分岔图

当收割率  $h \leq 1/4$  时只要初始总量足够大, 总量就能保持. 当收割率  $h = 1/4$  时收割率微小变化将导致物种数量的大变化, 收割率  $h > 1/4$  时物种将灭绝.

**周期收割模型** 假设物种收割率是周期变化的, 可得非自治的微分方程

$$P' = f(t, P) = aP(1 - P) - h(1 + \sin(2\pi t)),$$

此时方程不可分离变量, 无法求出解析解, 只能定性分析.

**拟合与预测:** 将 Logistic 模型一般化为  $P' = \alpha P - \beta P^2$  分离变量并积分得 Logistic 模型的显式解

$$P(t) = \frac{\alpha P_0 e^{\alpha(t-t_0)}}{\alpha - \beta P_0 + \beta e^{\alpha(t-t_0)}}$$

其中  $\alpha, \beta$  为待拟合参数. 理论上只需将任意两年的年份与人口数据  $t_1, P(t_1); t_2, P(t_2)$  代入显式解, 得二元方程组即可解得  $\alpha, \beta$ , 而所得模型便可用于预测, 但所得二元方程组为超越方程组, 求解困难. 文献 [8] 提出数值解法即确定适当  $\alpha, \beta$ , 使得

$$f = \sum_i \left( \frac{\alpha P_0 e^{\alpha(t-t_0)}}{\alpha - \beta P_0 + \beta e^{\alpha(t-t_0)}} - P_i \right)^2$$

取得最小值, 则有

$$\frac{\partial f}{\partial \alpha} = 0, \frac{\partial f}{\partial \beta} = 0.$$

可通过程序在一定范围内搜索  $\alpha, \beta$ .

## 第五章 时间序列模型

时间序列模型是专门针对时间序列数据而建立起的一种预测模型. 现实中感兴趣的系统往往是与时间有关的 - 系统随时间不断变化. 大部分的统计数据也都是时间序列数据. 人口系统显然是随时间变化的系统, 人口统计数据也多是时间序列数据. 对人口增长问题的研究可以归结为对人口时间序列数据的研究. 移动平均方法是最简单的时间序列预测方法, 但效果较差. 时间序列 ARIMA 模型是一种较实用的时间序列预测模型, 非常适合人口预测.

### §5.1 时间序列简单移动平均方法

**简单一次移动平均预测** 设人口数量时间序列为  $\{P_t\}$ , 取移动平均项数为  $n$ , 第  $t+1$  期预测值为

$$\hat{P}_{t+1} = M_t^{(1)} = \frac{P_t + P_{t-1} + \cdots + P_{t-n+1}}{n} = \frac{1}{n} \sum_{j=1}^n P_{t-n+j}.$$

$P_t$  表示第  $t$  期的实际值.  $\hat{P}_{t+1}$  表示第  $t+1$  期的预测值.  $M_t^{(1)}$  表示第  $t$  期一次移动平均数. 其预测的标准误差为  $S = \sqrt{\frac{\sum(P_{t+1} - \hat{P}_{t+1})^2}{N-n}}$ ,  $N$  为时间序列  $\{P_t\}$  所含原始数据的个数.

**注:**一般取  $n$  包含周期变动的时期, 这样可以减少周期影响. 若无周期变动, 应根据历史数据趋势, 选择  $n$ . 如果历史数据呈平稳趋势, 可选择较大的  $n$ , 反之, 数据若呈上升或下降趋势, 应选取较小的  $n$ .

简单移动平均方法拟合结果:

年份 (1978-1989)	实际总人口 (单位: 万人)	拟合总人口 (单位: 万人)	相对误差 (%)
1978	96259.00	0.00	100.00
1979	97542.00	0.00	100.00
1980	98705.00	0.00	100.00
1981	100072.00	0.00	100.00
1982	101654.00	98144.50	3.45
1983	103008.00	99493.25	3.41
1984	104357.00	100859.75	3.35
1985	105851.00	102272.75	3.38
1986	107507.00	103717.50	3.52
1987	109300.00	105180.75	3.77
1988	111026.00	106753.75	3.85
1989	112704.00	108421.00	3.80

年份 (1990-2007)	实际总人口 (单位: 万人)	拟合总人口 (单位: 万人)	相对误差 (%)
1990	114333.00	110134.25	3.67
1991	115823.00	111840.75	3.44
1992	117171.00	113471.50	3.16
1993	118517.00	115007.75	2.96
1994	119850.00	116461.00	2.83
1995	121121.00	117840.25	2.71
1996	122389.00	119164.75	2.63
1997	123626.00	120469.25	2.55
1998	124761.00	121746.50	2.42
1999	125786.00	122974.25	2.24
2000	126743.00	124140.50	2.05
2001	127627.00	125229.00	1.88
2002	128453.00	126229.25	1.73
2003	129227.00	127152.25	1.61
2004	129988.00	128012.50	1.52
2005	130756.00	128823.75	1.48
2006	131448.00	129606.00	1.40
2007	132129.00	130354.75	1.34

评价: 简单移动平均计算方便, 但误差比较大.

## §5.2 加权一次移动平均预测

加权一次移动平均预测公式如下:

$$\hat{P}_{t+1} = \frac{W_1 P_t + W_2 P_{t-1} + \cdots + W_n P_{t-n+1}}{W_1 + W_2 + \cdots + W_n} = \frac{\sum_{i=1}^n W_i P_{t-i+1}}{\sum_{i=1}^n W_i}$$

$P_t$  表示第  $t$  期的实际值.  $\hat{P}_{t+1}$  表示第  $t+1$  期的预测值.  $W_i$  表示权数.  $n$  表示移动平均的项数.

注: 时间序列简单移动平均方法是把参与平均的数据在预测中所起的作用同等看待, 加权一次移动平均预测参与平均的各期数据所起的作用是不同的, 权重代表了作用的大小. 取  $W_i = W_j, i \neq j$  即简单一次移动平均预测. 加权一次移动平均预测的缺点之一是权重选择的主观性.

## §5.3 指数平滑预测

**一次指数平滑预测** 一次指数平滑预测是以  $\alpha(1 - \alpha)^i; 0 < \alpha < 1, i = 0, 1, 2, \dots$  为权数, 对时间序列进行加权平均的一种方法. 其中加权方式为:  $P_i$  的权数为  $\alpha$ ,  $P_{t-1}$  的权数为  $\alpha(1 - \alpha)$ ,

$P_{t-2}$  的权数为  $\alpha(1 - \alpha)^2, \dots$ , 以此类推. 计算公式为:

$$\hat{P}_{t+1} = S_t^{(1)} = \alpha P_t + (1 - \alpha)S_{t-1}^{(1)}$$

其中  $\hat{P}_{t+1}$  表示第  $t + 1$  期的预测值;  $S_{t-1}^{(1)}, S_t^{(1)}$  分别表示第  $t - 1, t$  期一次指数平滑值;  $\alpha$  表示平滑系数,  $0 < \alpha < 1$ . 预测标准误差为  $S = \sqrt{\frac{\sum_{t=1}^{n-1} (P_{t+1} - \hat{P}_{t+1})^2}{n-1}}$

$n$  为时间序列所含原始数据的个数.

平滑系数  $\alpha$  的选择对预测值的影响较大, 但一般只能根据经验确定. 当时间序列数据是水平型的发展趋势类型,  $\alpha$  可取较小的值, 一般在  $0 \sim 0.3$  之间. 时间序列数据是上升或下降型的发展趋势类型,  $\alpha$  一般取较大的值, 一般在  $0.6 \sim 1$  之间. 实际应用中可选择不同的  $\alpha$  值比较选取最优者. 指数平滑预测的初值  $S_0^{(1)}$  可取时间序列的第一项或前几项的算术平均值为初值.

一次指数平滑预测适用于变化比较平稳、增长或下降趋势不明显的时间序列的下一期预测.

**二次指数平滑预测** 二次指数平滑预测是对一次指数平滑预测在作一次指数平滑预测来进  
行预测的一种方法, 但第  $t + 1$  期的预测值并非第  $t$  期的二次指数平滑值, 而采用下列公式进行  
预测:

$$\begin{cases} S_t^{(1)} = \alpha P_t + (1 - \alpha)S_{t-1}^{(1)}, \\ S_t^{(2)} = \alpha S_t^{(1)} + (1 - \alpha)S_{t-1}^{(2)}, \\ \hat{P}_{t+T} = a_t + b_t T \end{cases}$$

其中  $a_t = 2S_t^{(1)} - S_t^{(2)}$ ,  $b_t = \frac{\alpha}{1-\alpha}(S_t^{(1)} - S_t^{(2)})$   $P_t$  表示第  $t$  期的实际值;  $\hat{P}_{t+T}$  表示第  $t + T$  期的预测值;  $S_t^{(1)}$  表示第  $t$  期一次指数平滑值;  $S_t^{(2)}$  表示第  $t$  期二次指数平滑值;  $\alpha$  表示平滑系数.  $S_0^{(2)}$  的取值方法与  $S_0^{(1)}$  相同. 预测标准误差为  $S = \sqrt{\frac{\sum_{t=1}^{n-1} (P_{t+1} - \hat{P}_{t+1})^2}{n-2}}$  二次指数平滑预测适用于时间序列呈线性增长趋势下的短期预测.

## §5.4 时间序列的 ARIMA 预测模型

时间序列的 ARIMA 模型又称自回归滑动平均模型, 由自回归模型部分和滑动平均模型两部分组成.

**自回归模型:** 自回归模型 -AR(p) 的形式为  $X_t = c + \sum_{i=1}^p \phi_i X_{t-i} + \varepsilon_t$  其中  $\phi_i$  为模型参数,  $c$  为一常数,  $\varepsilon_t$  为白噪声

**滑动平均模型:** 滑动平均模型 -MA(q) 的形式为  $X_t = \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i}$  其中  $\theta$  为模型参数,  $\varepsilon_t$  为白噪声.

**自回归滑动平均模型:** 自回归滑动平均模型 -ARMA(p,q) 的形式为  $X_t = c + \varepsilon_t + \sum_{i=1}^p \phi_i X_{t-i} + \sum_{i=1}^q \theta_i \varepsilon_{t-i}$  自回归滑动平均模型包含自回归模型和滑动平均模型. 中  $\phi_i, \theta_i$  为模型参数,  $c$  为一常数, 一般讨论的模型为零均值, 即  $c = 0, \varepsilon_t$  为白噪声, 一般假设为独立正态同分布序列, 即  $\varepsilon_t \sim N(0, \sigma^2)$ .

ARMA 模型对数据性质有一定要求, 建模前需要对数据进行预处理, 常见的两种处理是平稳化和零均值化.

**平稳化:** 大多数的时间序列都是非平稳的, 建模之前需要对序列进行平稳化处理, 一般采用差分的方法. 看序列有什么趋势性, 如果存在线形趋势就进行一阶差分, 如果存在二次趋势就进行二阶差分, 以此类推. 如果存在季节趋势, 就进行季节差分.

**零均值化:** 为了处理方便, 一般假设所讨论的序列是零均值的, 如果时间序列不是零均值的, 可用样本均值作为序列的估计值, 建模前用样本数据减去均值, 然后对所得的零均值过程进行建模.

**ARIMA 模型的自相关函数:**

$$\begin{cases} \rho_0 = 1 \\ \rho_1 = \frac{(1-\phi_1\theta_1)(\phi_1-\theta_1)}{1+\theta_1^2-2\phi_1\theta_1} \\ \rho_k = \phi_1\rho_{k-1}, k \geq 2. \end{cases}$$

**ARIMA 模型的偏自相关函数:** 可由 *Yule – Wolker* 方程解得

$$\begin{cases} \theta_{k1}\rho_0 + \theta_{k2}\rho_1 + \cdots + \theta_{kk}\rho_{k-1} = \rho_1, \\ \theta_{k1}\rho_1 + \theta_{k2}\rho_0 + \cdots + \theta_{kk}\rho_{k-2} = \rho_2, \\ \quad \vdots \\ \theta_{k1}\rho_{k-1} + \theta_{k2}\rho_{k-2} + \cdots + \theta_{kk}\rho_0 = \rho_k, \end{cases}$$

**ARIMA 模型建模步骤:** ARIMA 模型建模具体步骤如图 5.1:

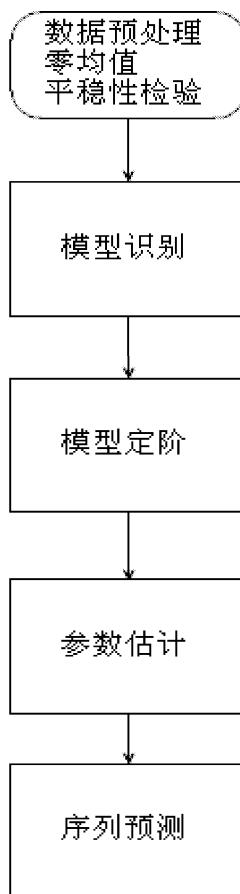


图 5.1: ARIMA 模型建模步骤

**ARIMA 模型用于人口预测:** 使用 SAS 时间序列工具包对人口进行拟合预测: 由于人口数据具有增长趋势性, 故取一阶差分, 使用  $ARIMA(p, 1, q)$  模型.

```

/*数据输入*/
data pop; input p@@; time=_n_;
cards; 96259 97542
98705 100072 101654 103008 104357 105851 107507 109300 111026
112704 114333 115823 117171 118517 119850 121121 122389 123626
;
run;
/*模型识别,相对最优定阶*/
proc arima data=pop;
identify Var=p(1) nlag=8 minic p=(0:8) q=(0:8);
run;

```

得到  $MinimumTableValue : BIC(1, 3) = 3.729405$  即最优模型为  $ARIMA(1, 1, 3)$

```
/*参数估计*/  
estimate p=4 q=2; run;
```

结果为:

```
Autoregressive Factors  
Factor 1: 1 - 0.72333 B**(1)  
Moving Average Factors  
Factor 1: 1 + 0.70678 B***(1) - 0.23813 B***(2) + 0.05508 B***(3)
```

```
/*序列预测*/  
forecast lead=10 id =time out=result;  
run;  
proc gplot data=result;  
plot p*time=1 forecast*time=2 195*time=3 u95*time=3/overlay;  
symbol1 c=black i=none v=star;  
symbol2 c=red i=join v=none;  
symbol3 c=green i=join v=none l=32;  
run;
```

结果为:

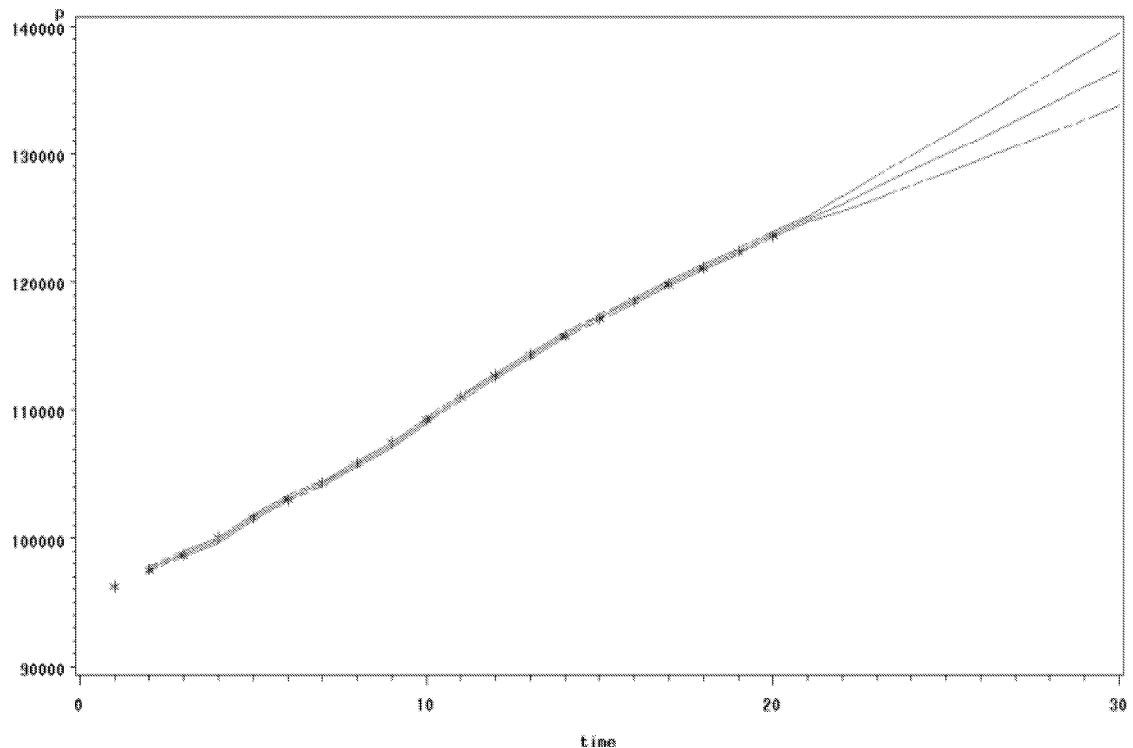


图 5.2: ARIMA 模型人口拟合预测结果

星号为实际值, 中间实线为拟合及预测值, 两端实线为 95% 置信区间上下界.

## 第六章 人口拟合与预测模型的对比分析

这里主要比较一元线性回归模型, 微分方程模型和时间序列 ARMA 模型三种最具代表性的模型的拟合效果和预测效果. 拟合效果 (1978 年到 1998 年 (共 20 年)) 的对比指标:

$$\text{拟合误差的绝对加和平均 } asef = \sum_{i=0}^{20} |P'_i - P_i| / 20$$

$$\text{拟合标准误差 } Sf = \sqrt{\frac{\sum_{i=0}^{20} (P'_i - P_i)^2}{20}}$$

$$\text{拟合相对误差 } re_i = \frac{P_i - P'_i}{P_i} * 100$$

预测效果 (1999 年到 2007 年 (共 10 年)) 的对比指标:

$$\text{预测误差的绝对加和平均 } asep = \sum_{i=21}^{30} |P'_i - P_i| / 10$$

$$\text{预测标准误差 } Sp = \sqrt{\frac{\sum_{i=21}^{30} (P'_i - P_i)^2}{10}}$$

$$\text{预测相对误差 } re = \frac{P_i - P'_i}{P_i} * 100$$

## 一元线性回归模型拟合预测结果

年份 (1978-1997)	实际总人口 (单位: 万人)	拟合总人口 (单位: 万人)	相对误差 (%)
1978	96259.00	95835.93	0.44
1979	97542.00	97331.17	0.22
1980	98705.00	98826.42	-0.12
1981	100072.00	100321.66	-0.25
1982	101654.00	101816.91	-0.16
1983	103008.00	103312.15	-0.30
1984	104357.00	104807.39	-0.43
1985	105851.00	106302.64	-0.43
1986	107507.00	107797.88	-0.27
1987	109300.00	109293.13	0.01
1988	111026.00	110788.37	0.21
1989	112704.00	112283.62	0.37
1990	114333.00	113778.86	0.48
1991	115823.00	115274.11	0.47
1992	117171.00	116769.35	0.34
1993	118517.00	118264.59	0.21
1994	119850.00	119759.84	0.08
1995	121121.00	121255.08	-0.11
1996	122389.00	122750.33	-0.30
1997	123626.00	124245.57	-0.50

拟合误差的绝对加和平均  $asef = 314.6034$ , 拟合标准误差  $Sf = 354.8352$ ,

年份 (1998-2007)	实际总人口 (单位: 万人)	预测总人口 (单位: 万人)	相对误差 (%)
1998	124761.00	125740.82	-0.79
1999	125786.00	127236.06	-1.15
2000	126743.00	128731.30	-1.57
2001	127627.00	130226.55	-2.04
2002	128453.00	131721.79	-2.54
2003	129227.00	133217.04	-3.09
2004	129988.00	134712.28	-3.63
2005	130756.00	136207.53	-4.17
2006	131448.00	137702.77	-4.76
2007	132129.00	139198.02	-5.35

预测误差的绝对加和平均  $asep = 3.7776e + 003$ , 预测标准误差  $Sp = 4.2604e + 003$ ,

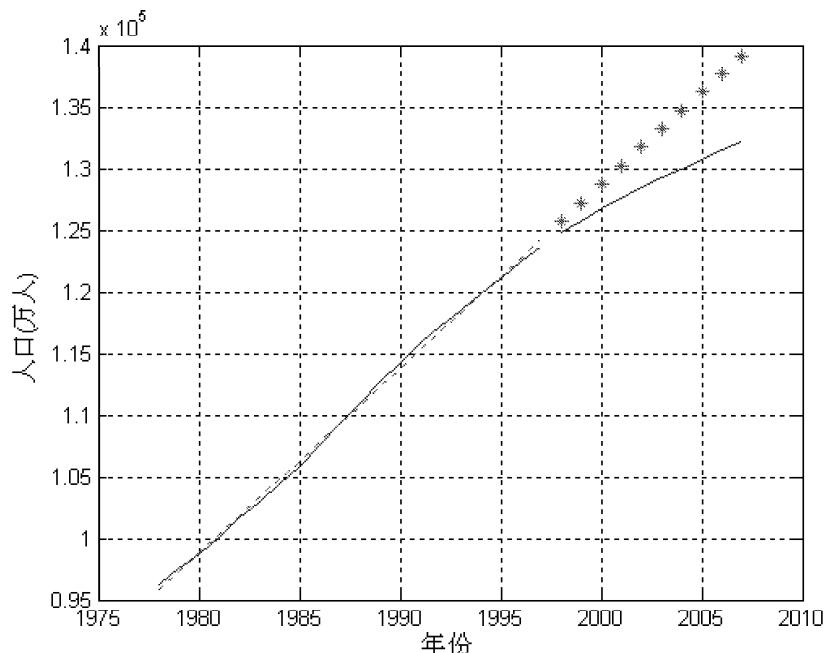


图 6.1: 一元线性回归模型拟合预测结果

## 微分方程模型拟合预测结果

年份 (1978-1997)	实际总人口 (单位: 万人)	拟合总人口 (单位: 万人)	相对误差 (%)
1978	96259.00	97507.37	-1.30
1979	97542.00	98770.15	-1.26
1980	98705.00	100047.45	-1.36
1981	100072.00	101339.40	-1.27
1982	101654.00	102646.10	-0.98
1983	103008.00	103967.70	-0.93
1984	104357.00	105304.29	-0.91
1985	105851.00	106656.00	-0.76
1986	107507.00	108022.96	-0.48
1987	109300.00	109405.26	-0.10
1988	111026.00	110803.04	0.20
1989	112704.00	112216.41	0.43
1990	114333.00	113645.48	0.60
1991	115823.00	115090.36	0.63
1992	117171.00	116551.19	0.53
1993	118517.00	118028.06	0.41
1994	119850.00	119521.08	0.27
1995	121121.00	121030.38	0.07
1996	122389.00	122556.07	-0.14
1997	123626.00	124098.24	-0.38

拟合误差的绝对加和平均  $asef = 685.4995$ , 拟合标准误差  $Sf = 790.2744$ ,

年份 (1998-2007)	实际总人口 (单位: 万人)	预测总人口 (单位: 万人)	相对误差 (%)
1998	124761.00	125657.02	-0.72
1999	125786.00	127232.51	-1.15
2000	126743.00	128824.81	-1.64
2001	127627.00	130434.03	-2.20
2002	128453.00	132060.29	-2.81
2003	129227.00	133703.67	-3.46
2004	129988.00	135364.29	-4.14
2005	130756.00	137042.23	-4.81
2006	131448.00	138737.61	-5.55
2007	132129.00	140450.52	-6.30

预测误差的绝对加和平均  $asep = 4.2589e + 003$ , 预测标准误差  $Sp = 4.8884e + 003$ ,

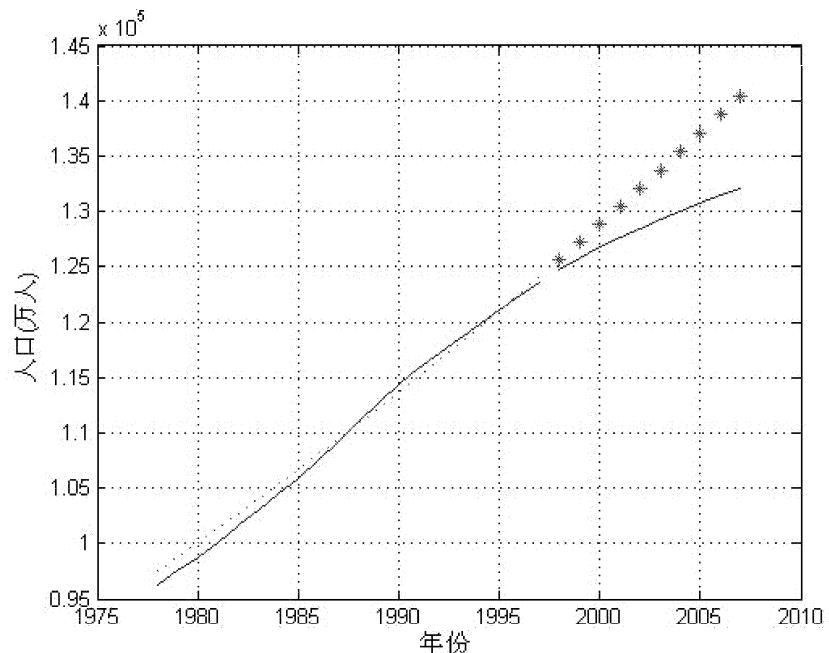


图 6.2: 微分方程模型拟合预测结果

### 时间序列 ARIMA 模型拟合预测结果

年份 (1978-1997)	实际总人口 (单位: 万人)	拟合总人口 (单位: 万人)	相对误差 (%)
1978	96259.00	105829.94	-9.94
1979	97542.00	95918.56	1.66
1980	98705.00	97482.60	1.24
1981	100072.00	98746.78	1.32
1982	101654.00	100770.36	0.87
1983	103008.00	102359.83	0.63
1984	104357.00	104124.91	0.22
1985	105851.00	105376.72	0.45
1986	107507.00	107080.42	0.40
1987	109300.00	108570.89	0.67
1988	111026.00	110425.17	0.54
1989	112704.00	111906.76	0.71
1990	114333.00	113497.72	0.73
1991	115823.00	114800.74	0.88
1992	117171.00	116083.73	0.93
1993	118517.00	117044.15	1.24
1994	119850.00	118102.50	1.46
1995	121121.00	119021.19	1.73
1996	122389.00	119958.13	1.99
1997	123626.00	120797.33	2.29

拟合误差的绝对加和平均  $asef = 71.1457$ , 拟合标准误差  $Sf = 91.6410$ ,

年份 (1998-2007)	实际总人口 (单位: 万人)	预测总人口 (单位: 万人)	相对误差 (%)
1998	124761.00	121667.92	2.48
1999	125786.00	119161.34	5.27
2000	126743.00	116213.74	8.31
2001	127627.00	112802.74	11.62
2002	128453.00	109181.16	15.00
2003	129227.00	105383.31	18.45
2004	129988.00	101643.40	21.81
2005	130756.00	98029.13	25.03
2006	131448.00	94743.09	27.92
2007	132129.00	91856.50	30.48

预测误差的绝对加和平均  $asep = 1.9936e + 003$ , 预测标准误差  $Sp = 2.4602e + 003$ ,

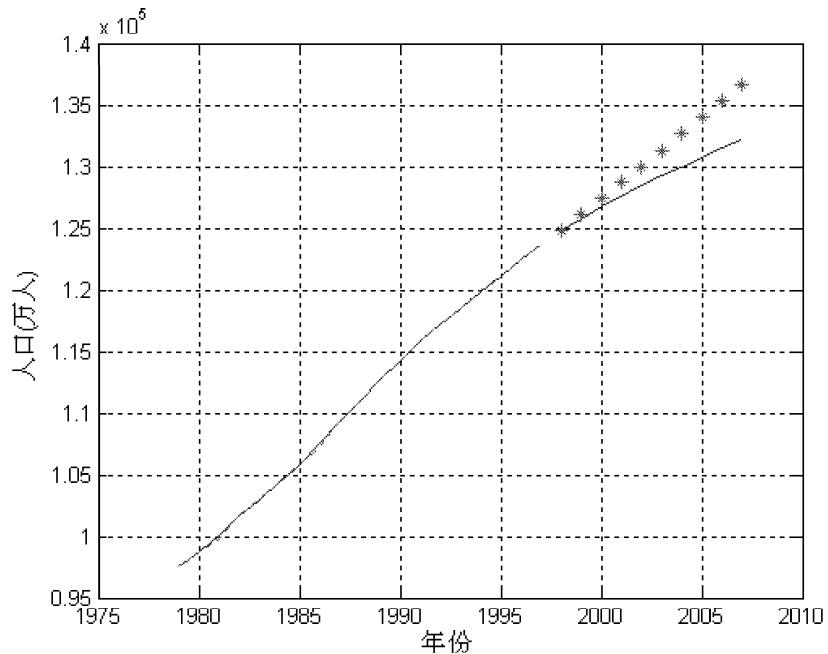


图 6.3: 时间序列模型拟合预测结果

#### 模型对比分析:

模型	asef	Sf	asep	Sp
线性回归	315	355	3778	4260
微分方程	685	790	4259	4888
时间序列	71	92	1994	2460

通过拟合结果可以看出, 这里选用的一元线性回归模型拟合与预测的误差比较小, 拟合曲线部分基本与人口数据统计曲线重合, 预测曲线部分偏高, 拟合预测误差的绝对加和平均和标准误差是三者中较小的. 一元线性回归模型效果好是 1979-1997 年间人口增长的线性趋势的反映,

可知人口在短期内仍会呈线性增长，但显然不可能持续的线性增长，所以一元线性回归模型只适用于短期预测，而不适用于长期预测。

微分方程模型是对线性增长的改进，拟合初期稍微偏高，中后期基本符合人口统计数据的走势。预测数据有些偏高，这与微分方程模型中的理想总量选择偏误有一定关系，可考虑进行改进。由于一元线性回归模型具有最小二乘的良好性质以及 1979-1997 年间中国人口增长的特征，微分方程各项指标高于一元线性回归模型，但这不足以说明一元线性回归模型优于微分方程模型。

时间序列模型拟合预测效果比较理想。时间序列模型的拟合难点在于模型的选择，如果选用的是时间序列 ARMA 模型，则由于数据具有趋势性、不平稳，拟合初期会出现显著的下降趋势，预测部分也会出现显著的下降趋势（如图 20），这违背了人口增长实际情况。拟合中间部分误差也会较大。时间序列模型出现较大误差，与模型参数选择有很大关系。而选用时间序列 ARIMA 模型，即对人口数据作一次差分，则趋势性消失，得到较为理想的拟合结果，但预测结果同样偏高。

三种模型有着各自的不同特点。一元线性回归模型最为简单，计算方便，但拟合与预测效果一般。微分方程模型的缺点在于要进行非线性参数拟合，拟合难度较大，虽然可以通过数值方法求解，但无法保证得到最优解。微分方程模型在实验中未达到预期的理想效果，但有较大的改进价值。时间序列 ARIMA 模型拟合和预测效果出众，明显优于前两种模型，具有较好的实用价值。三种模型的预测结果都有偏高的趋势，其中 ARIMA 模型偏离趋势最小，再次说明了 ARIMA 模型的优势。另外，预测偏高的趋势可能与实验选取的预测截断点（1997 年）有关，可以选取不同的截断点进行多次实验以更全面地研究三种模型的实际特点，克服一次实验的片面性。

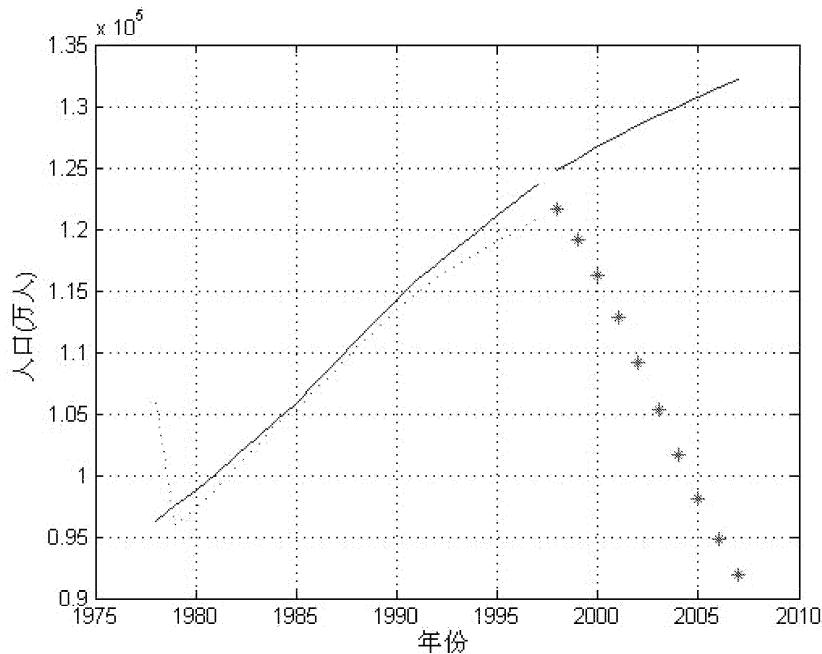


图 6.4: 未作一次差分的时间序列模型拟合预测结果

## 第七章 状态转移模型

状态转移模型是一种马尔可夫链过程模型, 其基本思想是将人口增长过程视为许多不同状态组成的链. 认为人口增长规律状态变化具有马尔可夫性 - 下一状态是什么只与当前状态有关, 而与上一状态无关. 人口序列是一种变化的随机波动性较大的数据序列, 马尔可夫链理论的转移概率可以反映随机因素的影响程度, 适用于预测随机波动较大的动态过程.

### §7.1 马尔可夫预测模型

设  $\{X_n, n = 0, 1, 2, \dots\}$  是一个离散随机变量序列, 而每个  $X_n$  所可能取的值属于一个有限实数集合  $E$ , 通常  $E = \{0, 1, 2, \dots\}$ . 如果随机变量序列  $\{X_n, n = 0, 1, 2, \dots\}$  中的  $X_{n+1}$  的条件概率分布只依赖于  $X_n$  的值, 而与前面的值无关, 即

$$P(X_{n+1} = i_{n+1}|X_n = i_n, X_{n-1} = i_{n-1}, \dots, X_0 = i_0) = P(X_{n+1} = i_{n+1}|X_n = i_n)$$

则称该随机过程是一个有限状态的马尔可夫链,  $E$  称为状态空间,  $\{0, 1, 2, \dots\}$  称为时间参数集.

**转移概率** 描述马尔可夫链的多维分布时条件概率  $P(X_{k+1} = i_{k+1}|X_k = i_k)$  称为时刻  $k$  时的一步转移概率  $p_{ij} = P(X_{k+1} = j|X_k = i)$ , 表示在时刻  $k$  时  $X_k$  取  $i$  值的条件下, 在下一时刻  $k+1$  时  $X_{k+1}$  取  $j$  值的概率.

$n$  步转移概率 - 在  $t = m$  时处于状态  $i$ , 在经过  $n$  步即时刻  $t = m + n$  时处于状态  $j$  的概率  $p_{ij}^{(n)} = P(X(m+n) = j|X(m) = i)$

性质:

$$(1) p_{ij}^{(n)}(m) \geq 0$$

$$(2) \sum_{j \in E} p_{ij}^{(n)}(m) = 1, \forall i, j \in E$$

特别的, 规定 0 步转移概率  $p_{ij}^{(0)}(m) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$ .

$n = 1$  时  $p_{ij}^{(0)}(m)$  为一步转移概率.

齐次马尔可夫链的  $n$  步转移概率与时间  $t$  无关, 记为  $p_{ij}^{(n)}$ .

设  $P$  代表一步概率  $p_{ij}$  所组成的矩阵, 对于有限状态的齐次马尔可夫链

$$P = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0m} \\ p_{10} & p_{11} & \cdots & p_{1m} \\ \cdots & \cdots & \cdots & \cdots \\ p_{m0} & p_{m1} & \cdots & p_{mm} \end{bmatrix}$$

称为一步转移概率矩阵, 显然对  $i = 0, 1, \dots, m$  有

$$(1) p_{ij}(k) \geq 0,$$

$$(2) \sum_{j=0}^m p_{ij} = 1$$

**切普曼 - 柯尔莫哥洛夫方程 (C-K 方程):** 设  $\{X_n, n = 0, 1, 2, \dots\}$  为齐次马尔可夫链则有  $p_{ij}^{(n)} = \sum_{k \in E} p_{ik}^{(m)} p_{kj}^{(n-m)}$ ,  $i, j \in E, 1 \leq m \leq n$  相应的  $n$  步转移概率表示为  $P^{(n)} = P^{(m)} P^{(n-m)}$

**初始分布:** 设马尔可夫链的状态集为  $E = \{0, \pm 1, \pm 2, \dots\}$  或其有限子集, 其初始时刻  $n = 0$  的概率记为  $p_i = p_i(0) = P\{X(0) = i\}$ , 称集合  $\{p_i(0)\}$  为马尔可夫链的初始分布.  $\{p_i(0)\}$  满足性质:

$$(1) p_i(0) \geq 0, i \in E;$$

$$(2) \sum_{i \in E} p_i(0) = 1$$

**绝对分布:** 设马尔可夫链的状态集为  $E = \{0, \pm 1, \pm 2, \dots\}$  或其有限子集, 其绝对时刻  $n$  的概率记为  $p_i(n) = P\{X(n) = i\}, i \in E$  满足性质:

$$(1) p_i(n) \geq 0, i \in E;$$

$$(2) \sum_{i \in E} p_i(n) = 1$$

绝对分布由初始分布和相应的转移概率唯一确定, 即  $p_j(n) = \sum_{i \in E} p_i(0) p_{ij}^{(n)}(0)$ ,

若  $\{X_n, n = 0, 1, 2, \dots\}$  为齐次马尔可夫链, 则  $p_j(n) = \sum_{i \in E} p_i(0) p_{ij}^{(n)}$

## §7.2 人口增长的马尔可夫性

**定理 7.2.1** 设  $X_1, X_2, \dots, X_n, \dots$  是独立随机变量序列, 概率密度函数为

$$f_{X_n}(x) = f_n(x) \text{ 现令 } Y_1 = X_1, Y_2 = X_1 + X_2, \dots, Y_n = X_1 + X_2 + \dots + X_n$$

则  $Y_1, Y_2, \dots, Y_n, \dots$  是马尔可夫序列. 即独立增量序列是马尔可夫序列.

证明:  $f(y_1, y_2) = f_{Y_2}(y_2 | Y_1 = y_1) f_{Y_1}(y_1)$ ,

$$\text{由已知条件 } f_{Y_1}(y_1) = f_{X_1}(y_1) = f_1(y_1) \quad f_{Y_2}(y_2 | Y_1 = y_1) = f_{X_1+X_2}(y_2 | X_1 = y_1)$$

$$= f_{X_2}(X_2 = y_2 - y_1 | X_1 = y_1) = f_{X_2}(y_2 - y_1) = f_2(y_2 - y_1)$$

故有  $f(y_1, y_2) = f_1(y_1) f_2(y_2 - y_1)$  推广到  $n$  个随机变量  $f(y_1, y_2, \dots, y_n)$

$$= f_1(y_1) f_2(y_2 - y_1) \cdots f_n(y_n - y_{n-1})$$

而  $f(y_n|y_{n-1}, \dots, y_1) = \frac{f(y_1, y_2, \dots, y_n)}{f(y_1, y_2, \dots, y_{n-1})} = f_n(y_n - y_{n-1})$

由上式知  $f(y_n|y_{n-1}, \dots, y_1)$  与  $y_{n-1}, \dots, y_1$  无关, 因此序列  $Y_n$  是一马尔可夫序列.

**人口生灭过程** 生灭过程是一种描述随机生灭的齐次马尔可夫链, 物理、化学、生物、医学等的许多实际模型都可以用生灭过程来描述.

模型假设若  $X(t) = n$ , 对人口在  $t$  到  $t + \Delta t$  的出生和死亡作如下假设:

1. 出生一人的概率与  $\Delta t$  成正比, 记作  $b_n \Delta t$ ; 出生二人及二人以上的概率为  $o(\Delta t)$

2. 死亡一人的概率与  $\Delta t$  成正比, 记作  $d_n \Delta t$ ; 死亡二人及二人以上的概率为  $o(\Delta t)$

3. 出生与死亡是相互独立的随机事件.

4.  $b_n$  和  $d_n$  均与  $n$  成正比, 记  $b_n = \lambda n, d_n = \mu n$   $\lambda$  和  $\mu$  分别是单位时间内  $n = 1$  时一个人出生和死亡的概率.

模型建立考察随机事件  $X(t + \Delta t) = n$  由假设 1、2、3 与出生或死亡一人的概率相比, 出生或死亡二人及二人以上的概率, 出生一人且死亡一人的概率均可忽略. 这样  $X(t + \Delta t) = n$  可以分解为三个互不相容事件之和:  $X(t) = n - 1$  且  $\Delta t$  内出生一人,  $X(t) = n + 1$  且  $\Delta t$  内死亡一人,  $X(t) = n$  且  $\Delta t$  内没有人出生或死亡.

由全概率公式  $P_n(t + \Delta t) = P_{n-1}(t)b_{n-1}\Delta t + P_{n+1}(t)d_{n+1}\Delta t + P_n(t)(1 - b_n\Delta t - d_n\Delta t)$

由此可得  $P_n(t)$  的微分方程  $\frac{dP_n}{dt} = b_{n-1}P_{n-1}(t) + d_{n+1}P_{n+1}(t) - (b_n + d_n)P_n(t)$

特别地, 在假设 4 下方程为  $\frac{dP_n}{dt} = \lambda(n - 1)P_{n-1}(t) + \mu(n + 1)P_{n+1}(t) - (\lambda + \mu)nP_n(t)$

若初始时刻 ( $t = 0$ ) 人口为确定数量  $n_0$ , 则  $P_n(t)$  的初始条件为

$$P_n(0) = \begin{cases} 1, & n = n_0 \\ 0, & n \neq n_0 \end{cases}$$

$X(t)$  的求解过程非常复杂, 而  $X(t)$  的期望  $E(X(t)) = \sum_{n=1}^{\infty} nP_n(t)$

容易求得:  $\frac{dE}{dt} = \lambda \sum_{n=1}^{\infty} n(n - 1)P_{n-1}(t) + \mu \sum_{n=1}^{\infty} n(n + 1)P_{n+1}(t) - (\lambda + \mu) \sum_{n=1}^{\infty} n^2 P_n(t)$

注意到  $\sum_{n=1}^{\infty} n(n - 1)P_{n-1}(t) = \sum_{k=1}^{\infty} k(k + 1)P_k(t), \sum_{n=1}^{\infty} n(n + 1)P_{n+1}(t) = \sum_{k=1}^{\infty} k(k - 1)P_k(t)$

有  $\frac{dE}{dt} = (\lambda - \mu) \sum_{n=1}^{\infty} nP_n(t) = (\lambda - \mu)E(t)$  初始条件  $E(0) = n_0$

故  $E(t) = n_0 e^{rt}, r = \lambda - \mu$  这个结果与指数模型  $x(t) = x_0 e^{rt}$  形式上完全一致. 从含义上看随机模型中的出生概率  $\lambda$  与死亡概率  $\mu$  之差  $r$  可称为净增长率, 人口的期望值  $E(t)$  呈指数增长. 在人口数量很多的情况下将  $r$  视为平均意义上的净增长率, 那么  $E(t)$  就可看成确定模型中的人口总数  $x(t)$ .

### §7.3 状态转移模型建模过程

将人口增长过程分段离散化, 假设人口的增长过程是一个独立增量过程, 即前一段时段人口净增长总量  $\Delta P_n = P_n - P_{n-1}$  与下一段时段人口净增长总量  $\Delta P_{n+1} = P_{n+1} - P_n$  相互独立, 则由定理 7.2.1 人口增长过程满足马尔可夫性. 只需得到人口状态集, 初始分布, 求出人口增长过程的状态转移矩阵, 然后由切普曼 - 柯尔莫哥洛夫方程即可得到未来人口情况. 人口状

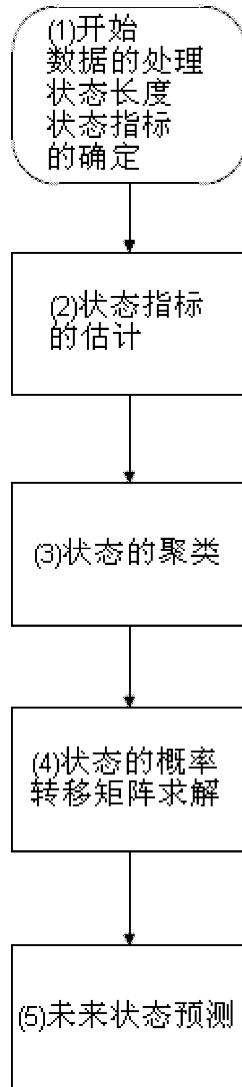


图 7.1: 状态转移模型建模过程

态集可以通过对已知的人口统计数据进行聚类分析得到, 但首先要确立或构造一个合适的状态指标. 由相邻两年的人口统计数据又可以求出人口增长过程的状态转移矩阵.

具体步骤 (如图 7.1): (1) 确定状态长度, 首先将连续人口模型离散化, 确定分段间隔, 一般以一年确定一个状态, 即状态长度为一年. 此外, 还要确定状态指标, 即构造或确定某一可计算的数量指标来标示状态、区分不同状态. 状态指标可以有许多不同的选择, 一般可选取此分段间隔的斜率, 因为斜率是一个相对量, 表达了增长速度这一信息.

(2) 状态指标的计算. 以选取斜率为状态指标为例, 统计数据较少时可以直接用斜率公式计算, 统计数据较多时可以通过一元线形回归估计出斜率项.

(3) 由于已知统计数据的状态指标是连续的, 必须进行聚类分析以确定有限状态集和每一年人口所处的状态.

(4) 状态概率的求解. 根据所确定的状态集和统计数据的状态, 研究每相邻两年的状态, 统计出不同状态之间相互转移的次数, 除以总次数即得到任意两个状态之间的转移概率.

(5) 状态预测. 首先计算出初始分布. 初始分布计算, 可以根据已知统计的数据得到每个状态的样本量, 除以总样本量, 即得初始分布概率. 最后由切普曼 - 柯尔莫哥洛夫方程即得未来状态的预测.

## §7.4 状态转移模型的性质

### 绝对分布对初始分布的稳定性

**定理 7.4.1** 设  $\xi = (\xi_n, \Pi, P)$  和  $\tilde{\xi} = (\tilde{\xi}_n, \tilde{\Pi}, P)$  是两个马尔可夫链, 具有不同的初始分布, 相应的为  $\Pi = (p_1, \dots, p_r)$ ,  $\tilde{\Pi} = (\tilde{p}_1, \dots, \tilde{p}_r)$ . 设  $\Pi^{(n)} = (p_1^{(n)}, \dots, p_r^{(n)})$  和  $\tilde{\Pi}^{(n)} = (\tilde{p}_1^{(n)}, \dots, \tilde{p}_r^{(n)})$ , 则  $\sum_{i=1}^r |\tilde{p}_i^{(n)} - p_i^{(n)}| \leq 2(1 - r\varepsilon)^n$

可见当  $\varepsilon$  足够小时 ( $0 < \varepsilon \leq \frac{1}{r}$ ), 转移概率阵相同, 绝对分布对初始分布的扰动具有稳定性.

**绝对分布对转移概率阵的稳定性** 初始分布相同时, 转移概率阵的微小变化对绝对分布的影响较为复杂, 下面只讨论一些特殊情况. 设初始分布向量为  $p_{(1^r)}^{(0)}$ , 绝对分布向量为  $p_{(1^r)}^{(n)}$ , 转移概率阵为  $P_{(r^r)}$ , 转移概率阵的扰动矩阵为  $dP_{(r^r)}$ , 其中扰动矩阵各元素绝对值  $|dP_{ij}| < \varepsilon$ ,  $\varepsilon$  为一较小的正数.

若  $P * dP = dP * P$  即两矩阵  $P$  和  $dP$  可交换 (比如  $dP = kE$ ,  $E$  为单位矩阵,  $k$  为某较小的数), 则  $(P + dP)^n - P^n = P^n + C_n^1 P^{n-1} dP + C_n^2 P^{n-2} dP^2 + \dots + dP^n - P^n = C_n^1 P^{n-1} dP + C_n^2 P^{n-2} dP^2 + \dots + dP^n$  绝对分布误差  $dp_n = p_n - p'_n = p_0((P + dP)^n - P^n)$ , 通过进一步假设可估计出  $dp_n$  的一个上界. 特别的, 如果  $dP$  是一收敛矩阵, 即  $\lim_{n \rightarrow \infty} dP^n = 0$  则  $\lim_{n \rightarrow \infty} dp_n = 0$

另外, 如果  $P$  与  $P + dP$  可以同时被对角化, 通过进一步的假设, 也可得到一些  $dp_n$  上界的估计.

### 状态链的遍历性与平稳分布

**遍历性:** 设对齐次马尔可夫链  $\{X_n, n = 0, 1, 2, \dots\}$  的所有状态  $i, j \in E$ , 存在不依赖于  $i$  的常数  $\pi_i$ , 有  $\lim_{n \rightarrow \infty} p_{ij}^{(n)} = \pi_j$ ,  $i, j \in E$  则称此齐次马尔可夫链具有遍历性, 并称  $\pi_i$  为状态  $j$  的稳态概率, 此时记  $\lim_{n \rightarrow \infty} P^{(n)} = \lim_{n \rightarrow \infty} P^n = \Pi$ . 若人口马尔可夫链具有遍历性则表明人口增长达到了某个有限的平稳状态.

**平稳分布:** 对齐次马尔可夫链  $\{X_n, n = 0, 1, 2, \dots\}$ , 存在实数集合  $\{r_j, j \in E\}$  满足  $r_j \geq 0, j \in E; \sum_{j \in E} r_j = 1; r_j = \sum_{i \in E} r_i p_{ij}, j \in E$  则称  $\{X_n, n \geq 0\}$  是一平稳齐次马尔可夫链,  $\{r_j, j \in E\}$  是该过程的一个平稳分布.

显然, 若平稳齐次马尔可夫链的初始分布为平稳分布, 则绝对概率等于初始概率.

**遍历性与平稳性分布定理:** 设齐次马尔可夫链  $\{X_n, n = 0, 1, 2, \dots\}$  的状态空间为  $E = \{1, 2, \dots, N\}$ , 若存在正整数  $m$ , 使对任意的  $i, j \in E$ , 其  $m$  步转移概率均大于零, 即  $p_{ij}^{(m)} > 0, i, j \in E$  则此链具有遍历性; 且各状态的稳态概率  $\pi_j, j \in E$  为方程组  $\pi = \pi P$  的唯一解.

## §7.5 模型评价与改进

- (1) 状态长度的确定: 一种改进是变长度的状态长度, 对样本点充分, 数据变化复杂的地方使用较短的状态长度, 对样本点少, 数据变化平稳的地方使用较长的状态长度. 对预测部分, 如果只需了解未来总体变化情况可使用较长的状态长度, 反之用较短的状态长度.
- (2) 状态指标的选取与构造: 状态指标不仅可以选取斜率, 还可以选取方差等统计量, 从而状态转移模型不仅可以预测未来人口的数量、增长率, 还可以预测未来人口的一些统计性质, 如变化幅度.

## §7.6 应用实例

使用 Matlab 对人口数据建立状态转移模型: 首先对斜率进行聚类

```
y=[96259 97542 98705 100072 101654 103008 104357 105851 ...
107507 109300 111026 112704 114333 115823 117171 118517 ...
119850 121121 122389 123626 124761 125786 126743 127627 ...
128453 129227 129988 130756 131448 132129];
for i=1:29 k(i)=y(i+1)-y(i);
end
T=clusterdata(k', 0.9);
T'
```

聚类结果:

```
T =
Columns 1 through 15
5 3 1 4 1 1 4 4 4 4 4 4 4 1 1
Columns 16 through 29
1 5 5 5 3 2 2 6 6 6 6 6 6 6 6
```

求解初始分布:

```
p0=zeros(1,max(T));
for i=1:max(T)
for j=1:length(T)
if T(j)==i p0(i)=p0(i)+1;
end end end
p0=p0/length(T);
p0
```

结果:

```
p0 =  
0.2069    0.0690    0.0690    0.2759    0.1379    0.2414
```

求解概率转移矩阵:

```
P=zeros(max(T),max(T));  
for j=1:length(T)-1  
P(T(j),T(j+1))=P(T(j),T(j+1))+1;  
end  
for j=1:max(T)  
st(j)=sum(P(j,:));  
end  
for i=1:max(T)  
for j=1:max(T)  
P(i,j)=P(i,j)/st(i);  
end  
end  
P
```

结果:

```
P =  
0.5000      0      0      0.3333    0.1667      0  
0      0.5000      0      0      0      0.5000  
0.5000    0.5000      0      0      0      0  
0.2500      0      0      0.7500      0      0  
0      0      0.5000      0      0.5000      0  
0      0      0      0      0      1.0000
```

易知该人口齐次马尔可夫链是遍历的, 且  $\lim_{n \rightarrow \infty} P^n =$

$$\begin{bmatrix} 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \\ 0 & 0.0000 & 0 & 0 & 0 & 1.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 1.0000 \\ 0 & 0 & 0 & 0 & 0 & 1.0000 \end{bmatrix}$$

注: 该概率转移矩阵含有较多的零元素, 零元素表明相应的两个状态不会发生转移. 然而另一种可能是统计数据的信息缺失 (未包含相应状态的相关信息), 一种较简单的改进是将零元素替换为较小概率值.

## 第八章 其他模型与方法简述

本节主要简要介绍、讨论其他一些人口预测和拟合模型. 其中许多模型在建模思想上与传统的经典模型有较大区别. 这些模型不仅在思想上是一种创新, 在实际中也体现出了较强的应用价值. 建模简便, 拟合、预测精度高等优势引起广泛关注与研究.

### §8.1 偏微分方程模型 - 宋健人口模型

偏微分方程模型类似于微分方程模型, 但引入了更多的影响人口系统的因素, 模型结构也更加复杂 - 由使用微分方程变为使用偏微分方程描述.

- 基本假设:**
1. 将所研究的社会人口当作一个整体、系统考虑.
  2. 所有表征和影响人口变化因素都是在整个社会人口平均意义下确定的.
  3. 把时间的流逝、婴儿的出生、人口死亡率、和居民的迁移看成人口状态变化的全部因素.

**模型形式:** 宋健偏微分方程人口模型具体形式为 (推导过程详见附录 A.3)

$$P(r, t) = \begin{cases} P_0(r-t)e^{-\int_{r-t}^r \mu(\rho)d\rho} & 0 < t < r \\ \varphi(t-r)e^{-\int_0^r \mu(\rho)d\rho} & r < t \end{cases} \quad (8.1.1)$$

其中  $r$  表示年龄,  $t$  表示时间,  $\mu(r, t)$  为人口相对死亡率,

初始条件:  $P(r, 0) = P_0(r)$ ;  $P_0(r)$  为初始时刻人口密度,

边界条件:  $P(0, t) = \varphi(t) = \mu(t)N(t)$ ;  $\mu(t)$  为相对出生率,  $N(t)$  为  $t$  时刻该地区人口总数.

### §8.2 差分方程模型

使用一阶偏微分方程研究人口模型在实际用中很不方便, 需要建立相应的离散模型. 作为输入的统计数据都是离散的, 人们想得到的预测值输出往往也是离散的, 连续模型解的表达式中包含了未知函数, 解析方法迭代求解非常困难. 与其用数值方法解连续模型, 不如直接建立离散模型.

**模型形式:** 向量形式的一解差分方程

$$P(t+1) = A(t)P(t) + \beta(t)B(t)P(t) \quad (8.2.1)$$

就是人口发展方程, 其中  $A(t)$  为死亡率矩阵,  $B(t)$  为生育率矩阵,(具体形式与模型建立详见附录 A.4),  $\beta(t)$  为总和生育率. 当初始人口分布  $P(0)$  已知, 并由统计资料确定参数矩阵  $A(t)$ ,  $B(t)$  并给定了控制变量总和生育率  $\beta(t)$  便可预测人口发展方程的状态变量  $P(t)$ . 在稳定的社会环境下可认为死亡率、生育模式和女性比不随时间变化, 即  $A(t)$ ,  $B(t)$  为常数矩阵, 有

$$P(t+1) = AP(t) + \beta(t)BP(t)$$

### §8.3 灰色系统 -G(1,1) 模型

灰色系统是一种部分信息已知、部分信息未知的系统, 它介于一无所知的黑色系统与全部可知的白色系统之间, 大量运用于对经济、社会、农业等系统进行预测. 灰色模型预测是在数据不存在一定的规律下而采取的一种建模与预测方法, 其预测得到的数据与原始数据存在一定的规律相似性. 影响人口增长的变化因素很多, 其中不乏一些不确定因素. 可以将人口系统看作一个灰色系统, 使用灰色系统预测法对人口变动进行分析.

**模型建立:** (1) 对原始数据的加工:

设  $x^{(0)}(k)$  表示以往连续  $n$  年间人口总数序列  $X^{(0)}$  的通项 (单位: 万人)( $k = 1, 2, \dots, n$ )

对该数列进行一次累加生成  $x^{(1)}(k) = \sum_{i=1}^k x^{(0)}(i)$ ,  $k = 1, 2, \dots, n$

用  $X^{(1)}$  表示累加生成数列,

用  $x^{(2)}(k) = -\frac{1}{2}[x^{(1)}(k) + x^{(1)}(k-1)]$ ,  $k = 2, 3, \dots, n$  表示对生成数列进行两项平均, 即均值生成序列  $X^{(2)}$

(2) 建立矩阵  $B$  和  $Y_N$ : 令  $B = \begin{bmatrix} x^{(2)}(2) & 1 \\ x^{(2)}(3) & 1 \\ \vdots & \vdots \\ x^{(2)}(n) & 1 \end{bmatrix}$ ,  $Y_N = [x^{(0)}(2), x^{(0)}(3), \dots, x^{(0)}(n)]^T$

(3) 计算参数向量  $\bar{a} = [a, u]^T$ : 灰色微分方程  $X^{(0)}(k) + aX^{(2)} = u$  的最小二乘估计参数序列满足  $\bar{a} = (B^T B)^{-1} B^T Y_N$

(4) 建立  $G(1, 1)$  模型

对生成数列  $X^{(1)} = (x^{(1)}(1), x^{(1)}(2), \dots, x^{(1)}(n))$  建立微分方程

$$\frac{dX^{(1)}}{dt} + aX^{(1)} = u$$

其时间响应函数为

$$\bar{x}^{(1)}(t+1) = (x^{(1)}(0) - \frac{u}{a})e^{-at} + \frac{u}{a}$$

其中  $a, u$  为 (3) 中计算出的常数, 且  $x^{(1)}(0) = x^{(0)}(1)$

这就是  $G(1, 1)$  模型, 可以据此对未来任意年间的人口数据作出预测.

**模型的检验:** (1) 残差检验: 令残差为  $\varepsilon(k) = \frac{x^{(1)}(k) - \bar{x}^{(1)}(k)}{x^{(1)}(k)}$ , ( $k = 1, 2, 3, \dots, n$ )

如果  $\varepsilon(k) < 0.2$  则达到一般要求;

如果  $\varepsilon(k) < 0.1$  则认为达到较高要求.

(2) 级比偏差值检验:

计算考察数据级比  $\lambda_0(k) = \frac{x^{(0)}(k-1)}{x^{(0)}(k)}$ ,  $k = 2, 3, \dots, n$ , 再用发展系数  $a$  求出相应的级比偏差  $\rho(k) = 1 - \frac{1-0.5a}{1+0.5a} \lambda_0(k)$ ,

如果  $\rho(k) < 0.2$  则可认为达到了一般要求;

如果  $\rho(k) < 0.1$  则认为达到了较高的要求.

**拟合与预测:** 使用灰色系统拟合预测结果:

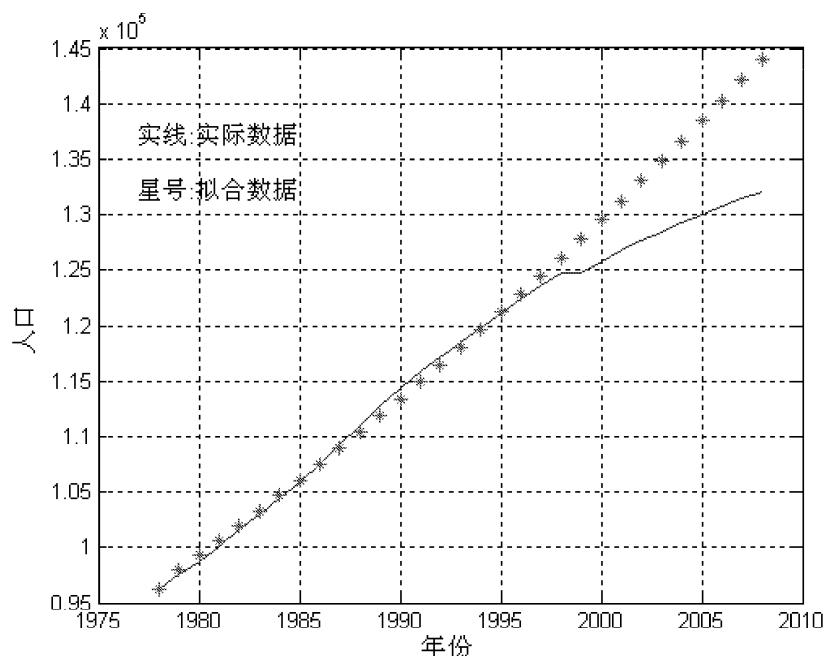


图 8.1: 灰色拟合预测结果

**评价:** 灰色模型适用于预测时间短, 数据资料少, 波动不大的系统对象, 其预测趋势都是一条较为平滑的曲线, 对于随机波动性较大的数据序列拟合性较差, 预测精度较低.

## §8.4 神经网络模型

人口系统是一个复杂的非线性系统, 线性模型与简单的非线性模型都很难做出较好的预测. 神经网络对复杂的非线性系统有较好的拟合能力, 其高度的非线性映射能力可以以任意精度逼近任意非线性函数. 神经网络具有较好的自学习、自适应、联想记忆、并行处理和非线性转换能力, 避免了复杂的数学推导, 在样本缺损和参数漂移的情况下仍能保证稳定的输出.

- 基本假设:**
1. 人口是时间的一维函数.
  2. 过去人口变化的规律对未来认识是适用的.

**BP 神经网络模型** 学习过程有信号的正向传播与误差的逆向传播两个过程组成. 正向传播时, 模式作用于输出层, 经隐层处理后, 传向输出层. 若输出层未能得到期望输出, 则转入逆向传播阶段, 将输出误差通过隐层向输入层逐层返回, 并“分摊”给各层所有单元, 从而获得各层单元的误差信号, 以作为修改各个单元权值的依据. 权值的不断修改过程, 也就是网络的学习训练过程. 学习训练过程一直进行到网络输出的误差逐渐减少到可接受的程度或达到设定的学习次数为止. 具体程序流程图:

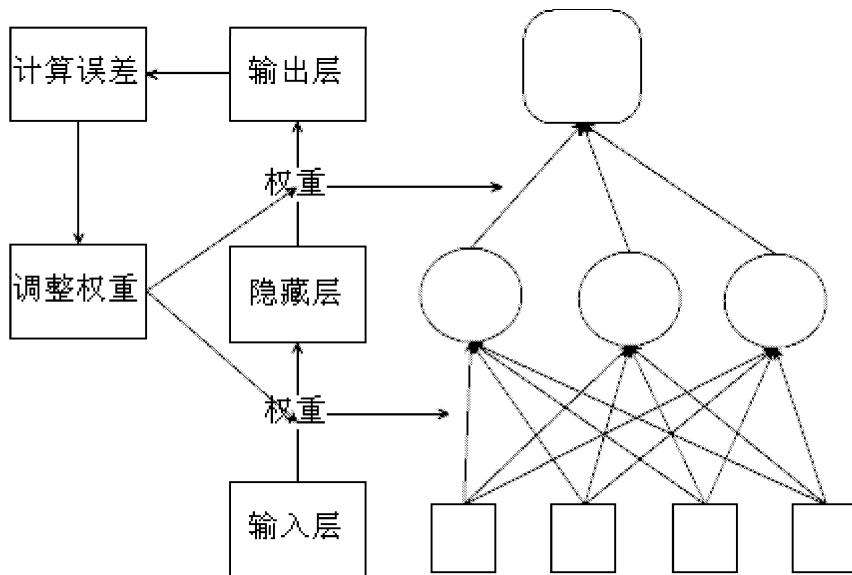


图 8.2: BP 神经网络模型

基于 BP 神经网络模型的人口预测时间序列方法只需以历史数据作为输入, 通过抑制或激活神经网络调节点, 自动决定影响性能及其影响程度, 自动形成模型, 无需模型假设. 程序流程图:

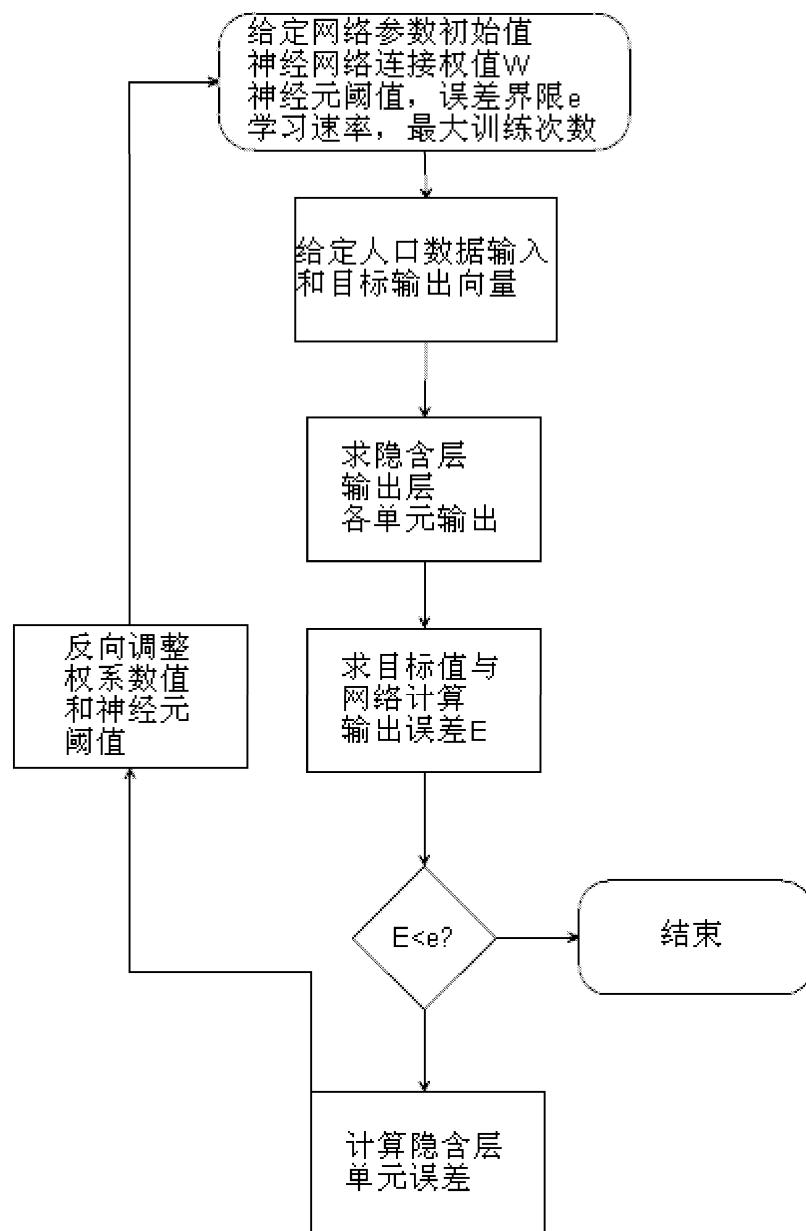


图 8.3: BP 神经网络模型用于人口预测

**拟合与预测:** 通过 Matlab 程序 (见附录 C.1), 使用 1978-1998 中国人口作为训练样本, 对 1999-2008 年度人口作出预测, 得到了很好的结果.

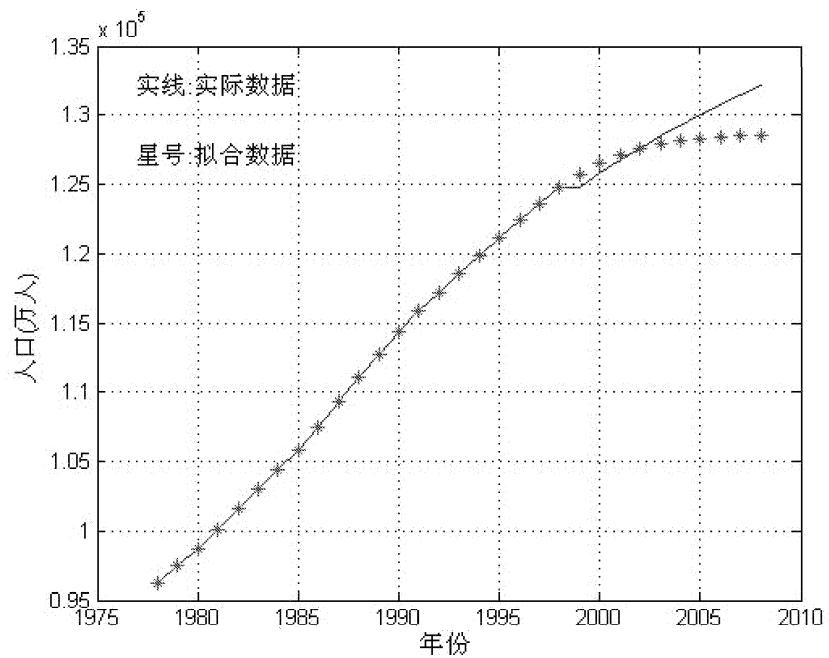


图 8.4: BP 神经网络模型人口预测结果

年份 (1978-1997)	实际总人口 (单位: 万人)	拟合总人口 (单位: 万人)	相对误差 (%)
1978	96259.00	96259.00	0.00
1979	97542.00	97542.00	0.00
1980	98705.00	98705.00	0.00
1981	100072.00	100072.00	0.00
1982	101654.00	101654.00	0.00
1983	103008.00	103008.00	0.00
1984	104357.00	104357.00	0.00
1985	105851.00	105851.00	0.00
1986	107507.00	107507.00	0.00
1987	109300.00	109300.00	0.00
1988	111026.00	111026.00	0.00
1989	112704.00	112704.00	0.00
1990	114333.00	114333.00	0.00
1991	115823.00	115823.00	0.00
1992	117171.00	117171.00	0.00
1993	118517.00	118517.00	0.00
1994	119850.00	119850.00	0.00
1995	121121.00	121121.00	0.00
1996	122389.00	122389.00	0.00
1997	123626.00	123626.00	0.00

年份 (1998-2007)	实际总人口 (单位: 万人)	拟合总人口 (单位: 万人)	相对误差 (%)
1998	124761.00	124763.01	-0.00
1999	124761.00	125741.67	-0.79
2000	125786.00	126534.77	-0.60
2001	126743.00	127145.93	-0.32
2002	127627.00	127598.65	0.02
2003	128453.00	127924.22	0.41
2004	129227.00	128153.38	0.83
2005	129988.00	128312.25	1.29
2006	130756.00	128421.24	1.79
2007	131448.00	128495.46	2.25

训练过程很快收敛, 达到预定误差界  $1^{-12}$ . 其最终结果可以看出, 拟合误差全部约为 0, 近期预测误差比较小, 远期预测误差也在可以接受范围之内. 神经网络模型结果明显优于前面的大部分模型, 是人口数据拟合预测模型的一个很好的选择.

## §8.5 遗传算法模型

之前讨论的各种模型的形式都是确定的, 限制了模型的求解空间, 难以提高预测精度, 难以反映人口系统的变化规律. 为扩大搜索空间, 可以在一个形式不确定的空间中寻找人口模型. 利用演化计算思想, 通过遗传程序设计可以是人口模型选择的一个新思路.

- 基本假设:**
1. 人口是时间的一维函数.
  2. 过去人口变化的规律对未来认识是适用的.

**模型建立:** 已知过去  $n$  年人口数, 求函数  $y = f(t)$  使

拟合误差  $\sum_{i=1}^n |f(t_i) - y_i|$  最小.

**程序设计:** 遗传算法是从代表问题可能潜在的解集的一个种群 (population) 开始的, 而一个种群则由经过基因 (gene) 编码的一定数目的个体 (individual) 组成. 每个个体实际上是染色体 (chromosome) 带有特征的实体. 染色体作为遗传物质的主要载体, 即多个基因的集合, 其内部表现 (即基因型) 是某种基因组合, 它决定了个体的形状的外部表现, 如黑头发的特征是由染色体中控制这一特征的某种基因组合决定的. 因此, 在一开始需要实现从表现型到基因型的映射即编码工作. 由于仿照基因编码的工作很复杂, 往往进行简化, 如二进制编码, 初代种群产生之后, 按照适者生存和优胜劣汰的原理, 逐代 (generation) 演化产生出越来越好的近似解, 在每一代, 根据问题域中个体的适应度 (fitness) 大小选择 (selection) 个体, 并借助于自然遗传学的遗传算子 (genetic operators) 进行组合交叉 (crossover) 和变异 (mutation), 产生出代表新的解集的种群. 这个过程将导致种群像自然进化一样的后代种群比前代更加适应于环境, 末代种群中的最优个体经过解码 (decoding), 可以作为问题近似最优解. 具体程序流程图:

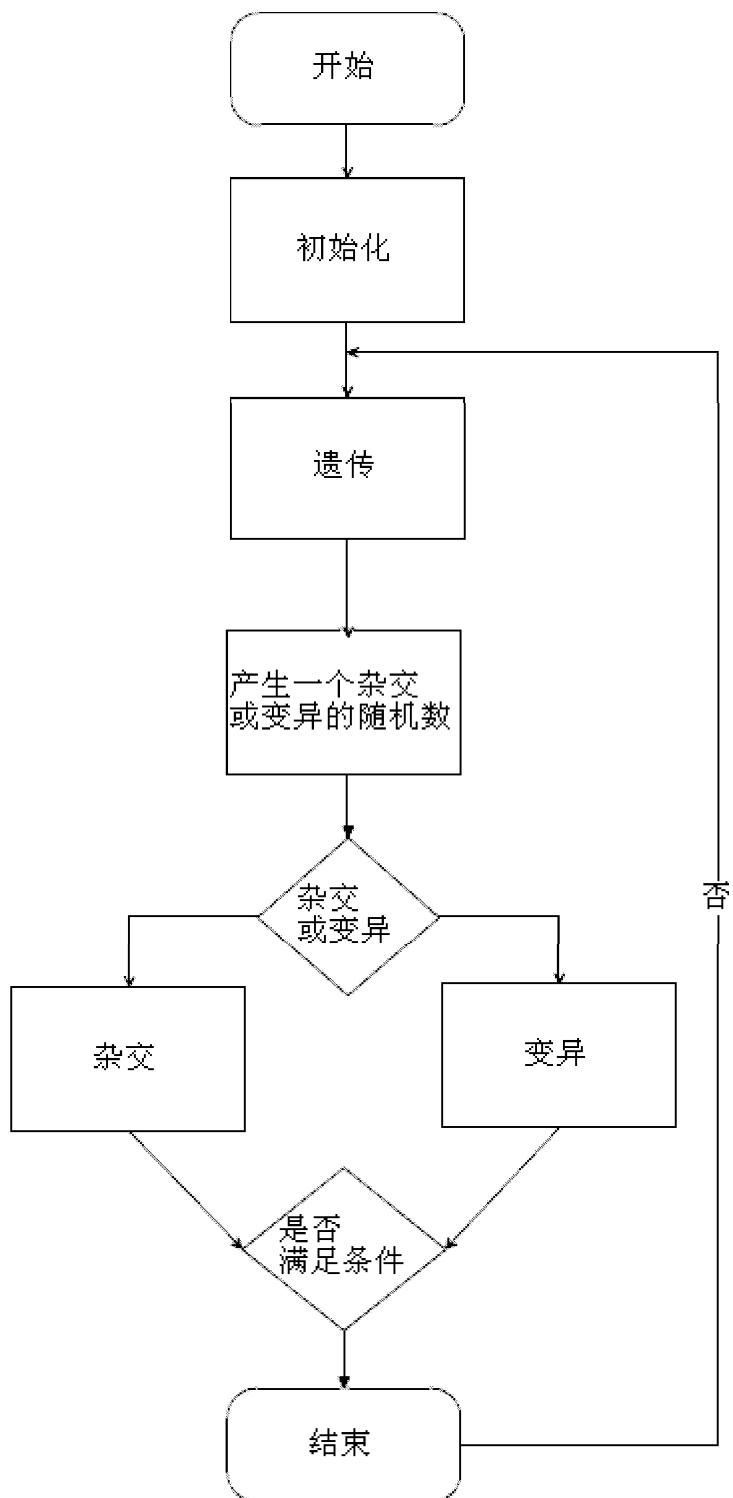


图 8.5: 遗传算法程序设计

通过遗传算法, 可以对函数集进行智能选择确定出拟合效果较好的人口模型.

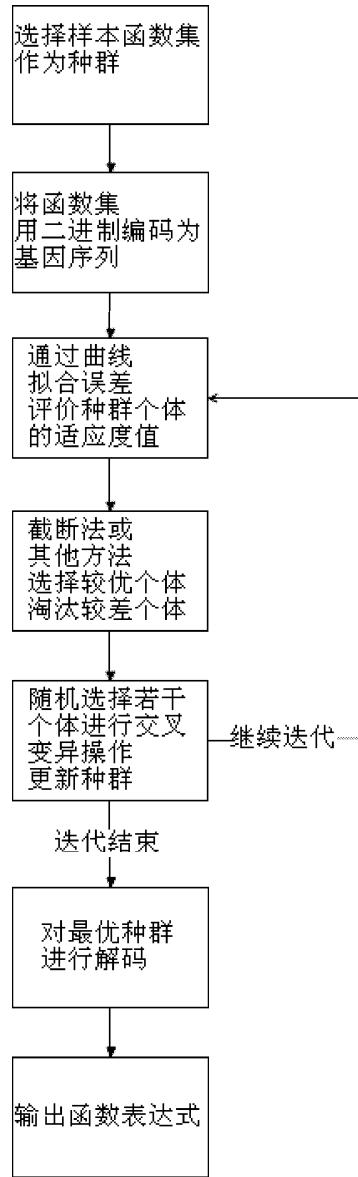


图 8.6: 遗传算法用于人口模型选择

**应用实例:** 文献 [21] 选择函数集  $F = \{+, *, /, \wedge, \sin, \cos, \exp, \ln\}$  作为种群, 通过遗传算法, 对湖北省 1990 – 1997 年人口数据, 找到较优拟合函数  $y(t_i) = 1000[5.587769 + \sin(56.3476 + \frac{t_i - 1989}{15.951538})]$ , 模型拟合误差为 0.004% ~ 0.334%. 可见通过遗传算法智能确定人口模型函数形式确实可以取得较好效果.

## §8.6 分形数据拟合方法

前面所讨论的拟合与预测模型大都是经典函数模型 - 最终的数据模型可用一个显式的函数来表示. 这样函数模型一般都具有光滑性等良好的分析性质. 然而, 真实的数据模型可能极其复杂, 数据函数很可能不具有良好的分析性质. 分形理论是自然界不规则现象的一种理论, 它

承认世界的局部可能在一定条件下, 过程中, 在某一方面表现出与整体的相似性, 它认为空间维数的变化不仅可以是离散的还可以是连续的. 分形理论用于数据拟合, 其拟合结果不再有良好的分析性质, 而是具有真实世界中更为普遍的一种性质 - 自相似性. 时间序列数据往往都具有一定的自相似性, 人口数据可能也存在一定的自相似性. 基于分形理论的数据拟合方法为人口模型的拟合预测从一个相反的方向提供了思路.

**分形插值原理:** 分形插值函数是通过迭代函数系统来实现的.

令数据集  $(x_i, y_i) : i = 0, 1, \dots, N$  给定, 考虑  $R^2$  上的一个 IFS,

它的吸引子  $G$  是内插数据的连续函数  $f : [x_0, x_N] \rightarrow R$  的图像.

考虑  $IFS\{R^2 : W_n, n = 1, 2, \dots, N\}$ , 其中  $W_n$  是具有如下形式的仿射变换:

$$W_n \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_n & 0 \\ c_n & d_n \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e_n \\ f_n \end{bmatrix}$$

且

$$W_n \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} = \begin{bmatrix} x_{n-1} \\ y_{n-1} \end{bmatrix}$$

$$W_n \begin{bmatrix} x_N \\ y_N \end{bmatrix} = \begin{bmatrix} x_n \\ y_n \end{bmatrix}$$

上式具体写为:

$$\begin{cases} a_n x_0 + e_n = x_{n-1} \\ a_n x_N + e_n = x_n \\ c_n x_0 + d_n y_0 + f_n = y_{n-1} \\ c_n x_N + d_n y_N + f_n = y_n \end{cases}$$

上式有四个方程五个未知数, 选择  $d_n$  为自由变量. 令  $|d_n| < 1$  (否则, 该 IFS 不收敛), 解方程组, 令  $L = x_N - x_0$ , 则

$$\begin{cases} a_n = L^{-1}(x_n - x_{n-1}), \\ e_n = L^{-1}(x_N x_{n-1} - x_0 x_n), \\ c_n = L^{-1}[y_n - y_{n-1} - d_n(y_N - y_0)], \\ f_n = L^{-1}[x_N y_n - 1 - x_0 y_n - d_n(x_N y_0 - x_0 y_N)]. \end{cases}$$

**注:**上面定义的矩阵变换是一伸长变换, 它把平行于  $y$  轴的线段影射到平行于  $y$  轴的另一线段且两线段长度之比为  $|d_n|$ , 故  $d_n$  又称为变换  $W_n$  的垂直比例因子.

**拟合与预测:** 通过 Matlab 程序 (见附录), 对 1978-2007 中国人口数据进行拟合, 结果如下:

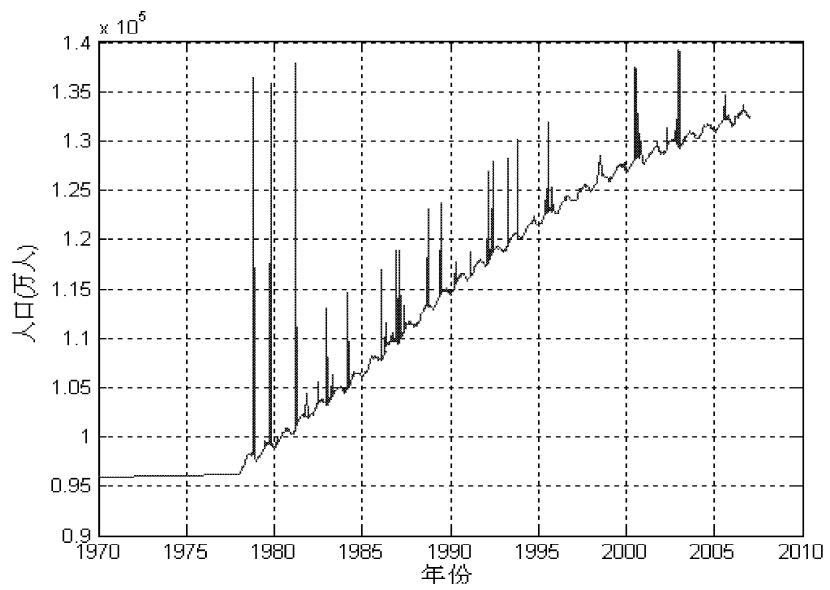


图 8.7: 分形数据拟合结果

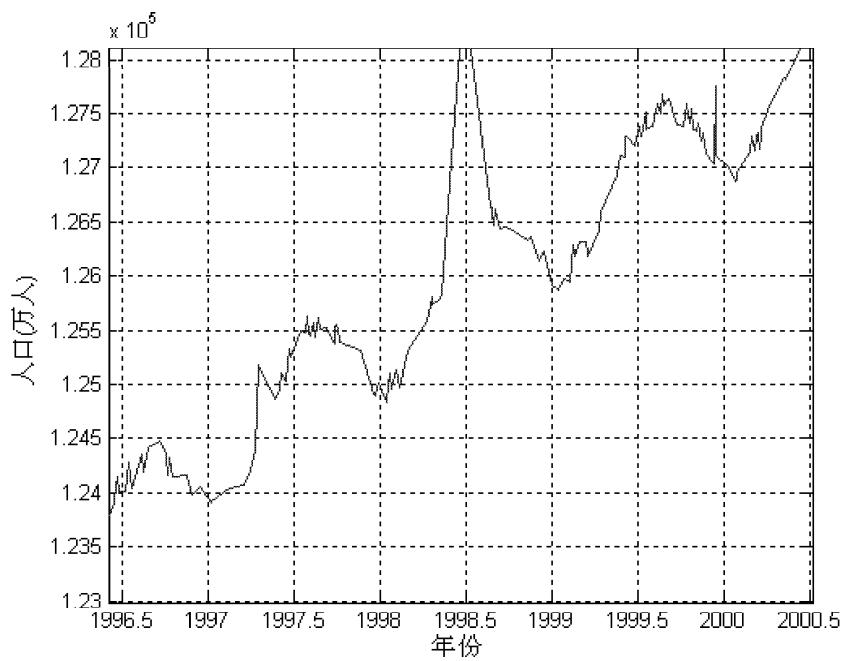


图 8.8: 分形数据拟结果局部放大图

## §8.7 组合预测模型

自 J.M. Bates 和 C.W. Granger (1969) 首次提出组合预测以来, 组合预测方法得到广泛研究和应用.

**模型思想:** 人口分析预测方法很多, 对此问题, 可以采取多种不同的方法进行分析, 充分利用原始数据中所包含的不同信息, 进行组合预测分析, 获得更高的预测精度和可靠度, 弥补单一方法的不足.

**组合模型理论:** 对一个预测问题, 用  $y_i$  表示实际观测值 ( $t = 1, 2, \dots, n$ )  
n 为样本规模. 有 m 种预测方法,  $f_i$  表示第  $i$  种方法的预测值 ( $i = 1, 2, \dots, n$ ),  $w_i$  表示第  $i$  种方法在组合预测中的权重, 则组合预测模型可表示为:

目标函数

$$\min z = \sum (y_i - \sum w_i f_i)^2, \text{ s.t. } \sum w_i = 1$$

如果预测值是实际值的无偏估计, 即  $E(f_i) = y_i$  则  $\sum w_i f_i$  也是  $y_i$  的无偏估计, 此模型表示组合预测与原始数据偏差平方最小为目标建立组合预测模型. 此模型可用 Lagrange 乘子法求解.

**组合模型求解过程:** (1) 对于单个模型计算, 获得相应预测数据;  
(2) 采用二次规划模型对组合模型进行建模;  
(3) 对组合模型进行误差分析, 修正模型;  
(3) 使用修正模型和已知数据对未来发展情况进行预测;

注:一般目标函数中还要增加条件  $w_i > 0$ .

## §8.8 非参数模型

之前讨论的人口模型大都是参数模型, 基本思想都是先通过假设, 确定模型形式和有限的未知模型参数, 然后用已知数据估计出未知参数, 从而得到模型的确定形式, 并用于预测. 非参数模型并不对人口模型的发展规律做出任何假设, 而是直接从数据出发, 通过研究数据的统计性质, 拟合相适应的数据模型. 非参数模型的拟合难点在于参数空间是不定的, 参数数量为无限多个. 实际的应用中, 常对参数空间做出一些假设以缩小范围, 从而产生了不同的非参数预测方法与非参数模型. 为了求出显示解, 参数数量也限制为有限多个.

**基本假设:** 人口是时间的函数, 可用  $P = r(t) + \epsilon$  来描述人口数量随时间的变化规律, 其中  $P$  表示人口数量,  $t$  表示时间,  $r$  为描述人口数量随时间的变化规律的未知函数.

**模型建立:** 非参数拟合的目的就是寻找未知函数  $r(\cdot)$ , 使其能较好描述人口数量随时间的变化规律 (使拟合误差较小), 同时还要求函数  $r(\cdot)$  要有较为理想的函数性质.

## 常见的非参数估计方法

**局部平均:** 局部平均的思想很简单, 选取固定或可变步长  $h(t)$ , 对每个时间点  $t$ , 得到区间  $[t - h(t), t + h(t)]$ , 而函数  $r(t)$  在区间  $[t - h(t), t + h(t)]$  的值就取为落入区间  $[t - h(t), t + h(t)]$  的所有统计数据的平均值. 局部平均方法简单, 确点也是明显的, 拟合函数  $r(t)$  不具连续、可导等性质, 也无法用于人口预测.

**核估计:** 核估计其实是一种加权的局部平均, 常见形式有  $r_h = \frac{\sum_{j=1}^n P_j K((t-t_j)/h)}{\sum_{j=1}^n K((t-t_j)/h)}$ ,

$0 \leq x \leq 1$ , 加权函数  $K$  称为核函数, 可采用不同的形式.

**傅立叶估计:** 对  $r(t)$  作傅立叶级展开有  $r(x; m) = \phi_0 + 2 \sum_{j=1}^m \phi_j \cos(\pi j x)$ ,

$x \leq 1$  其中级数项数  $m$ , 级数系数  $\phi_j$ , 为要估计的参数.

**多项式估计:** 多项式估计就是假设  $r(t)$  为一多项式形式, 所要估计的便是多项式系数. 此外还有分段多项式估计 - 对不同时间段采用不同多项式, 样条估计 - 要求分段多项式估计的分段点处有较好的光滑性等.

**有理函数估计:** 假设  $r(t)$  为有理函数, 即  $r_{p,q}(x) = \frac{\beta_0 + 2 \sum_{j=1}^q \beta_j \cos(\pi j x)}{|1 + \alpha_1 \exp(\pi i x) + \dots + \alpha_p \exp(\pi i x)|^2}$ ,  
 $x \leq 1$ .

**小波展开估计:** 对函数进行小波展开  $r(x) = c_0 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j-1} c_{j,k} 2^{j/2} \psi_H(2^j x - k)$ ,  
 $x \leq 1$ , 其中  $\psi(\cdot)$  为小波函数, 可有不同形式,  $c_{j,k}$  为要估计参数.

## 第九章 对基于数据的模型和基于因素的模型的一些思考

**数据模型与因素模型** 本文所探讨的大部分人口拟合与预测的模型，基本上分为两类 - 基于数据的模型和基于因素的模型。基于数据的模型指的是模型本身是从（统计）数据出发，通过研究已知的历史数据本身的各种（统计）性质、特点，对未知的数据作出估计、预测。文中基于数据的模型有初等函数模型、线性回归模型、时间序列模型、状态转移模型、灰色系统模型、分形拟合方法以及非参数模型。

基于因素的模型指的是模型本身从决定、影响系统的各种因素出发，通过理论分析与推导，将较为复杂的系统用各种因素显式地表示，并引入表征各种因素的影响作用大小的参数（系数），最后通过统计数据估计出各个参数的大小或变化规律，从而得到具体描述系统本身结构的模型，并用其对未来的数据作出预测。文中基于因素的模型有：初等函数模型中的自然增长模型、微分方程模型、偏微分方程模型和差分方程模型。

**基于因素的模型** 基于因素的模型需要对未知的系统进行假设，确立出影响系统的主要因素。其优点是模型是可解释的，模型的参数具有明显的实际意义，模型一旦确立，就意味着系统是在假设下可知的。基于因素的模型常用于对影响模型的因素作出分析，研究因素与系统的关系，从而达到通过控制因素来影响、控制整体系统的目的。

基于因素的模型的缺点也是显然的，影响系统各个因素往往本身就是复杂的，想要通过这些因素去研究原系统，往往容易把问题复杂化。比如微分方程模型要用到人口的自然增长率，偏微分方程模型要用到性别比例和生育率，这些因素明显的是比人口数量更加复杂、抽象的概念，所以实际应用中往往要对这些因素作出种种简单的假设，比如因素不变，但这些假设的合理性有时是很难验证的。所以，基于因素的模型似乎都默认了一个原则 - 变量的变化率的变化规律比变量本身的变化规律简单。比如，如果认为人口呈线性增长，前提假设就是人口自然增长率是常数。然而，并非所有的函数的导函数都要比原函数形式简单，类似人口系统这样的复杂系统，其内部因素的变化规律也可能比我们感兴趣的系统最终输出还要复杂。退一步，假设这一原则对大部分系统，比如人口系统，是成立的，那么利用导函数变化的规律来估计原函数，基本方法就是积分运算。这样一来，一个重要的问题是误差的积累与放大，函数的积分运算涉及维数的增加，比如，一条只有长度的平面曲线积分后得到的是面积。导函数的微小误差在积分后可能会变得不可忽略。对因素的假设是与实际存在一定的误差的，因素一多，误差便会积累，如果模型不是很稳健，结果就会与实际产生很大的出入。可见基于因素的模型的稳健性是十分重要的。另一个问题是认识的局限性。我们所能想到的因素必然不会包含实际影响系统的全部因素，许多因素，甚至是十分重要的因素都可能被归结入随机误差中。所假设的因素还可能包含了一些无关或影响很小的因素。此外，对各个因素如何影响系统，即模型的形式的估计与认知也可能

是错误的。另一个问题是统计数据的限制。我们不一定能得到系统的主要因素的历史数据，这对模型的最终的估计带来阻碍。种种的问题影响了基于因素的模型的实用价值。

**基于数据的模型** 基于数据的模型的基本假设都是关于数据的。基于数据的模型不需要对具体问题的系统作出过多的假设。基于数据的模型的基本思想是真实的系统是未知的，而统计数据中包含了系统的信息，通过提取历史统计数据的某些信息，以及历史统计数据本身变化的信息，我们可以得到其中包含的关于未来的信息，从而对未来的工作可以作出一定的估计。基于数据的模型假设系统的历史与未来的信息是相关的，不是独立的。在实际中，我们所感兴趣的系统正是这样的系统，比如人口系统，否则，如果历史与未来的信息毫无关系，这样的系统是完全不稳定、不可知的，系统没有任何利用价值，对其研究也是无意义的。在基于数据的模型所关注的系统中，未来的数据仿佛受到了历史的信息的一种牵制，距离现在越近的数据受到的牵制力越大，预测就越准确，较远的数据受到的牵制力越小，甚至可能完全不受影响。通过研究这种“牵制力”，我们就可以绕开复杂的系统本身，从系统外部对未来作出预测。对不同的系统，历史统计数据信息的研究可以采取相同的方法，这就使基于数据的模型具有普遍可用性和简单性。

然而，基于数据的模型也有一些固有的缺陷。系统的输出所包含的信息究竟有多少，如何有效提取所需的信息，显然对不同的系统是不一样的。对此，出现了多种多样的不同形式的基于数据的模型，但针对不同的系统，如何选择出最有效的模型，又是一个困难的问题。此外，历史输出的信息能够提供对系统内部结构哪些认知，又该如何描述这些信息，是需要继续探讨的问题。

**基于模拟的智能模型** 本文还简要介绍了两种不同于以上两种模型的模型 - 神经网络模型和遗传算法模型。这两种模型都可以归结为智能模型。所谓智能模型，指的是一种动态的、具有自适应能力和学习能力的模型。这种模型的基本思想是对未知系统的结构进行统一形式的模拟。比如神经网络模型，其用内部的统一的复杂结构 - 隐藏层去模拟未知系统内部的复杂结构，通过学习、训练的动态思想去建立自身结构，来模拟未知结构。这种对未知系统结构的模拟思想从根本上解决了基于因素的模型和基于数据的模型一些固有缺陷。同时，智能模型融合了基于因素的模型和基于数据的模型两者的优点，不需要太多的模型假设，既具有基于数据的模型的普遍可用性和简单性又具有基于因素的模型的特点 - 结构。实际结果也验证了智能模型的优势与价值。

我们也可以认为基于数据的模型本身的模型形式代表了一种“结构”，比如不同的  $p, q$  值和自回归、滑动平均系数的 ARMA 模型具有不同的结构，但这种结构的复杂程度远不如神经网络模型，因而不具备较强的“结构模拟”功能。未知结构的系统本身像是一种“黑箱”系统，输入和输出是已知的，但我们对其内部结构一无所知。智能模型的结构是对现实模型的各种结构的一个“抽象”，而且是具有学习、模拟能力的一种抽象，是可以通过输入和输出进行内部结构的自我建立的一种抽象。两个“黑箱”系统的输入和输出相同，其内部结构却可以有不同的实现方式。智能模型正是利用这一“黑箱”原理，实现了对现实的模拟。

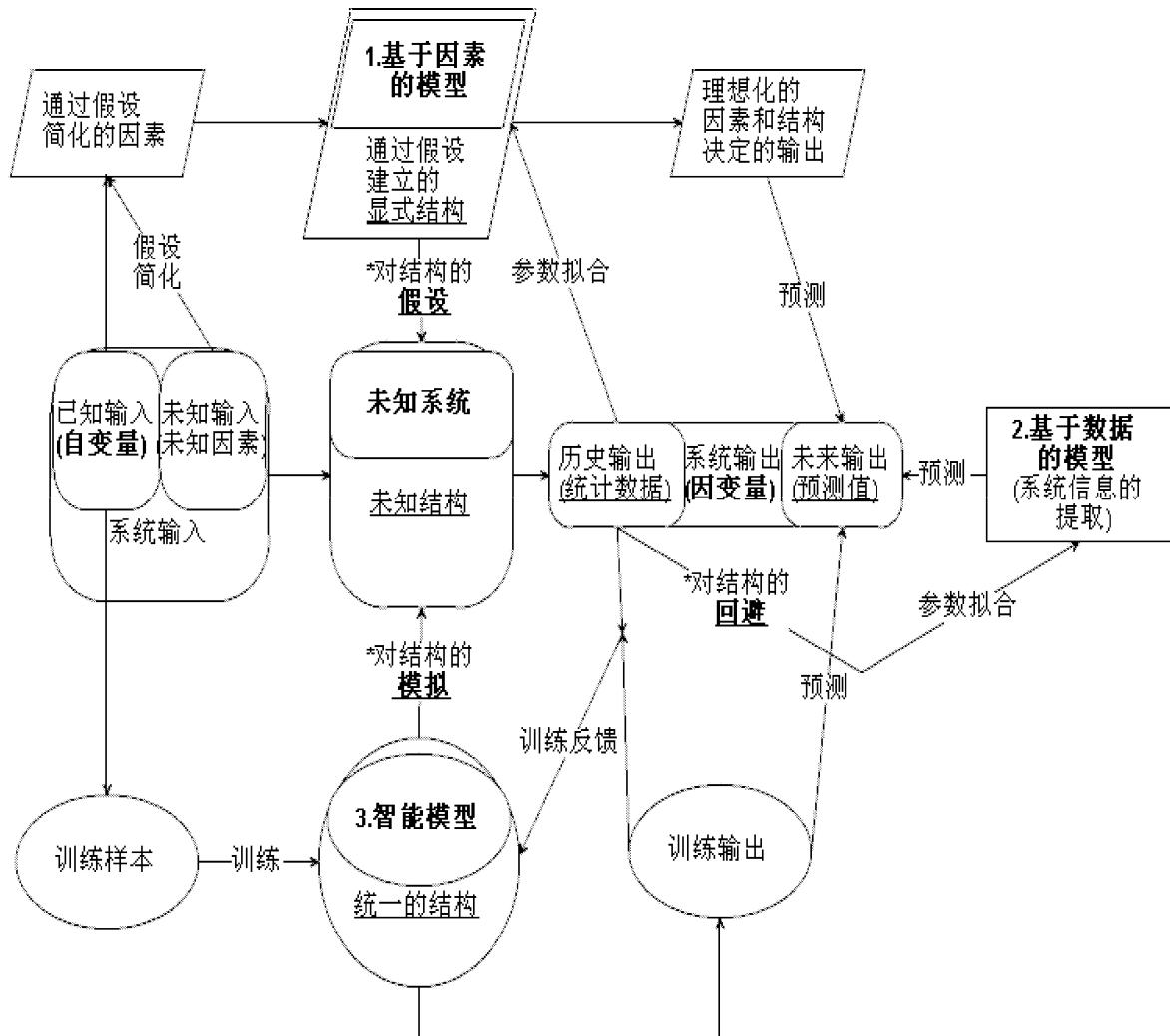


图 9.1: 模型的比较

**结论** 基于数据的线性回归模型过于简单, 不适用于人口预测. 基于因素的微分方程模型需要太多的假设和统计数据, 实用性不强. 基于数据的时间序列的 ARIMA 模型和基于模拟的神经网络模型可以作为中国人口增长拟合、预测的单模型的两个较优选择. 可能的改进有使用基于数据的模型处理微分方程模型中各种因素的混合模型, 以及基于多模型的组合预测模型 (如文献 [24]).

## 谢    辞

论文在选题、写作、修改中多次得到林路教授的指导. 林路教授在论文写作各环节中给予我及时地检查、多次询问并提出了许多建议. 首先向林路教授表示感谢.

感谢数学学院各位老师对我的教育培养, 他们一丝不苟的作风, 严谨求实的态度, 踏实的精神将使我终生受益.

感谢我的室友、同学四年米对我学习、生活的关心和帮助.

最后, 向我的父亲、母亲致谢, 感谢他们对我的支持和鼓励.

## 参考文献

- [1] (美)Sheldon M. Ross, 陈典发译, 数理金融初步, 机械工业出版社,2005.
- [2] 王松桂, 史建红, 尹素菊, 吴密霞, 线性模型引论, 科学出版社,2005.
- [3] 赵静, 但琦, 数学建模与数学实验, 高等教育出版社,2002.
- [4] 姜启源, 谢金星, 叶俊, 数学模型, 高等教育出版社,2003.
- [5] 唐焕文, 贺明峰, 数学模型引论, 高等教育出版社,2005.
- [6] (美) Morris W. Hirsch, Stephen Smale, Robert L. Devaney, 甘少波译, 微分方程、动力系统与混沌导论, 人民邮电出版社,2008.
- [7] 于万波, 混沌的计算实验与分析, 科学出版社,2008.
- [8] 王周喜, 胡斌, 王洪萍, 人口预测模型的非线性动力学研究, 数量经济技术经济研究,2002 年第 8 期.
- [9] Precision, bias, and uncertainty for state population forecasts: an exploratory analysis of time series models Jeff Tayman, Stanley K. Smith, Jeffrey Lin Springer, Science+Business Media B.V.,2007.
- [10] 熊建平, 吴建华, 万国金, AR 模型在人口增长预测中的应用, 计算机与现代化,2005 年第 10 期.
- [11] 王燕, 应用时间序列分析, 中国人民大学出版社,2008.
- [12] 李裕奇, 刘垸, 随机过程习题解答, 国防工业出版社,2008.
- [13] 周荫清, 李春升, 陈杰, 随机过程习题集, 清华大学出版社,2004.
- [14] (俄罗斯) A. H. 施利亚耶夫, 周概容译, 概率, 高等教育出版社,2007.
- [15] 高惠璇, 应用多元统计分析, 北京大学出版社,2005.
- [16] 张文, 温荣生, 邱淑芳, 基于背景值重构的灰色一马尔柯夫模型及其应用, 东华理工学院学报, 第 30 卷第 1 期,2007 年 3 月.
- [17] World Population Projections Using Metabolic GM (1,1)Model, Caimei Lu, Yonghong Hao, Xuemeng Wang Proceedings of 2007 IEEE International Conference on Grey Systems and Intelligent Services, November 18-20, 2007.
- [18] An Artificial Neural Net approach to forecast the population of India, Goutami Bandyopadhyay and Surajit Chattopadhyay.
- [19] 毕小龙, 袁勇, 基于 BP 神经网络的人口预测方法研究, 武汉理工大学学报(交通科学、与工程版), 第 31 卷第 3 期 2007 年 6 月.
- [20] 尹春华, 陈雷, 基于 BP 神经网络人口预测模型的研究与应用, 人口学刊,2005 年第 2 期.
- [21] 何朗, 成浩, 王宗跃, 朱慧颖, 黄樟灿, 基于遗传程序设计的单因素人口预测模型, 武汉理工大学学报(信息与管理工程版), 第 25 卷第 5 期,2003 年 10 月.
- [22] 李水根, 分形, 高等教育出版社,2004.
- [23] (英) Kenneth Falconer, 分形几何, 人民邮电出版社,2007.
- [24] 刘鑫, 徐世英, 基于组合预测模型的西藏人口分析, 中央民族大学学报(自然科学版)第 17 卷第 3 期,2008 年 8 月.
- [25] Jeffery D.Hart, Nonparametric Smoothing and Lack-of-Fit Tests, Springer 出版社,1997.
- [26] 肖燕彩, 邱成, MATLAB 语言及实践教程, 清华大学出版社, 北京交通大学出版社,2005.

- [27] 李东风, 统计软件教程:SAS 系统与 S 语言, 人民邮电出版社,2006.
- [28] 国家统计局人口和就业统计司, 中国人口和就业统计年鉴, 中国统计出版社,2008.

## 附录一 部分推导与证明

### §A.1 线性自然增长模型的建立

由基本假设 (2.1.1),  $P(t)$  可一阶可微,

$$\begin{aligned} P'(t) &= \lim_{s \rightarrow 0} \frac{P(t+s) - P(t)}{s} \\ &= \lim_{s \rightarrow 0} \frac{[P(t) + P(s) - 1] - P(t)}{s} \\ &= \lim_{s \rightarrow 0} \frac{P(s) - 1}{s} \\ &= \lim_{s \rightarrow 0} \frac{P(s) - P(0)}{s} \\ &= P'(0) \end{aligned}$$

可见  $P'(0)$  为一常数.

上式用  $r$  代替  $t$ , 两端从 0 到  $t$  积分,

$$\begin{aligned} \int_0^t P'(r) dr &= \int_0^t P'(0) dr \\ P(t) - P(0) &= t \cdot P'(0) \\ P(t) &= 1 + t \cdot P'(0) \end{aligned}$$

令  $P'(t) = P'(0) = i$  有  $P(t) = 1 + t \cdot i$  对  $t \geq 0$ .

### §A.2 指数自然增长模型的建立

由基本假设 (2.2.1),  $P(t)$  可一阶可微,

$$\begin{aligned} P'(t) &= \lim_{s \rightarrow 0} \frac{P(t+s) - P(t)}{s} \\ &= \lim_{s \rightarrow 0} \frac{P(t) \cdot P(s) - P(t)}{s} \\ &= P(t) \cdot \lim_{s \rightarrow 0} \frac{P(s) - 1}{s} \\ &= P(t) \cdot P'(0). \end{aligned}$$

$$\frac{P'(t)}{P(t)} = \frac{d}{dt} \log_e P(t) = P'(0)$$

上式用  $r$  代替  $t$ , 两端从 0 到  $t$  积分,

$$\begin{aligned} \int_0^t \frac{d}{dr} \log_e P'(r) dr &= \int_0^t P'(0) dr \\ \log_e P(t) - \log_e P(0) &= t \cdot P'(0) \\ \log_e P(t) &= t \cdot P'(0) \quad (\text{其中 } \log_e P(0) = \log_e 1 = 0) \end{aligned}$$

令  $P(1) = 1 + i$  则  $\log_e P(1) = \log_e(1 + i) = P'(0)$  即  $P(t) = (1 + i)^t \quad t \geq 0$  或  $\log_e P(t) = t \cdot \log_e(1 + i) = t \cdot k \quad t \geq 0, k = \log_e(1 + i), k \geq 0$ .

### §A.3 偏微分方程模型的建立

#### 参数说明:

$N(t)$  t 时刻该地区人口总数

$r_m$  人口的最高寿命

$F(r, t)$  人口函数;  $r$  表示年龄,  $t$  表示时间, 该函数表示  $t$  时刻该地区一切年龄小于  $r$  的人  
数;  $r_2 > r_1$  时  $F(r_2, t) \geq F(r_1, t)$

$P(r, t)$  人口年龄分布密度函数;  $P(r, t) = \frac{\partial F}{\partial r}$  该函数表示在  $t$  时刻, 年龄为  $r$  的人数; 显然有

$$P(r, t) \geq 0, P(r_m, t) = 0, F(r, t) = \int_0^r P(\xi, t) d\xi$$

$$F(r_m, t) = \int_0^{r_m} P(\xi, t) d\xi = \int_0^\infty P(\xi, t) d\xi = N(t)$$

$$t$$
 时刻年龄在  $r_1$  到  $r_2$  之间的人口为  $F(r_2, t) - F(r_1, t) = \int_{r_1}^{r_2} P(\xi, t) d\xi$

$M(r, t)$  人口死亡分布函数, 表示  $t$  时刻该地区年龄为  $r$  的人的死亡数

$\mu(r, t)$  相对死亡率  $\mu(r, t) = \frac{M(r, t)}{P(r, t)}$

$k(r, t)$  女性在人口中占的比例

$b(r, t)$   $t$  时刻平均每个  $r$  岁女性的生育数

$r_1, r_2$  育龄区  $t$  时刻, 出生的婴儿总数为  $f(t) = \int_{r_1}^{r_2} b(r, t) k(r, t) P(r, t) dr$  令  $b(r, t) = \beta(t) h(r, t)$ ,

$$\text{其中 } \int_{r_1}^{r_2} h(r, t) dr = 1; \int_{r_1}^{r_2} b(r, t) dr = \int_{r_1}^{r_2} \beta(t) h(r, t) dr = \beta(t) \int_{r_1}^{r_2} h(r, t) dr = \beta(t)$$

$\beta(t)$  一个女性一生总生育数

$h(r, t)$  一个女性在  $r$  岁时的生育概率

$$\text{如果 } \beta(t), h(r, t) \text{ 都与 } t \text{ 无关, 则 } f(t) = \beta(t) \int_{r_1}^{r_2} h(r, t) k(r, t) P(r, t) dr = \beta \int_{r_1}^{r_2} h(r) k(r, t) P(r, t) dr$$

**模型建立:**  $t$  时刻年龄在  $[r, r + \Delta r]$  的人数为  $P(r, t) \Delta r$  过了  $\Delta t$  时间后, 死亡人数为  $\mu(r, t) P(r, t) \Delta r \Delta t$

另一部分活到了  $t + \Delta t$  时刻, 其年龄位于区间  $[r + \Delta r', r + \Delta r + \Delta r']$

显然有  $\Delta r' = \Delta t$  即在  $t + \Delta t$  时刻, 年龄在  $[r + \Delta r', r + \Delta r + \Delta r']$  中的人口数为  $P(r + \Delta r, t + \Delta t) \Delta r$

下式显然成立

$$P(r, t) \Delta r - P(r + \Delta r', t + \Delta t) \Delta r = \mu(r, t) P(r, t) \Delta r \Delta t$$

即

$$P(r + \Delta r', t + \Delta t) \Delta r - P(r, t + \Delta t) \Delta r + P(r, t + \Delta) \Delta r - P(r, t) \Delta r = -\mu(r, t) P(r, t) \Delta r \Delta t$$

两边同除以  $\Delta r \Delta t$  有:

$$\frac{P(r + \Delta r', t + \Delta t) - P(r, t + \Delta t)}{\Delta r'} + \frac{P(r, t + \Delta) - P(r, t)}{\Delta t} = -\mu(r, t) P(r, t)$$

取极限:

$$\frac{\partial P(r, t)}{\partial r} + \frac{\partial P(r, t)}{\partial t} = -\mu(r, t)P(r, t)$$

初始条件:  $P(r, 0) = P_0(r)$ ;  $P_0(r)$  为初始时刻人口密度

边界条件:  $P(0, t) = \varphi(t) = \mu(t)N(t)$ ;  $\mu(t)$  为相对出生率

综上得人口模型的微分方程, 当  $\mu(r, t)$  不依赖于  $t$ , 仅依赖于  $r$  时, 可解得:

$$P(r, t) = \begin{cases} P_0(r - t)e^{-\int_{r-t}^r \mu(\rho)d\rho} & 0 < t < r \\ \varphi(t - r)e^{-\int_0^r \mu(\rho)d\rho} & r < t \end{cases}$$

## §A.4 差分方程模型的建立

**模型建立:** 时间以年为单位, 年龄按周岁计算, 设最大年龄为  $m$  岁,

记  $x_i(t)$  为第  $t$  年  $i$  岁 (满  $i$  周岁而不到  $i+1$  周岁) 的人数,

$t = 0, 1, 2, \dots, i = 0, 1, 2, \dots$ .

只考虑由于生育率、老化和死亡引起的人口演变、不计迁移等社会因素影响.

记  $d_i(t)$  为第  $t$  年  $i$  岁人口的死亡率. 即  $d_i(t) = \frac{x_i(t) - x_{i+1}(t+1)}{x_i(t)}$

于是  $x_{i+1}(t+1) = (1 - d_i(t))x_i(t), t = 0, 1, 2, \dots, i = 0, 1, 2, \dots, m-1$

记  $b_i(r)$  为第  $t$  年  $i$  岁女性生育率, 即每位女性平均生育婴儿数,  $[i_1, i_2]$  为育龄区间,  $k_i(t)$  为第  $t$  年  $i$  岁人口的女性比,

则第  $t$  年出生的人数为

$$f(t) = \sum_{i=i_1}^{i_2} b_i(t)k_i(r)x_i(t)$$

记  $d_{00}(t)$  为第  $t$  年婴儿的死亡率, 即第  $t$  年出生但未活到人口统计时刻的婴儿比例,

$$d_{00}(t) = \frac{f(t) - x_0(t)}{f(t)}$$

综上  $x_1(t+1) = (1 - d_{00}(t))(1 - d_0(t)) \sum_{i=i_1}^{i_2} b_i(t)k_i(r)x_i(t)$

将  $b_i(t)$  分解为  $b_i(t) = \beta(t)_j h_i(t)$  其中  $h_i(t)$  是生育模式, 用以调整育龄妇女在不同年龄时生育率的高低,

满足  $\sum_{i=i_1}^{i_2} h_i(t) = 1, \beta(t) = \sum_{i=i_1}^{i_2} b_i(t)$

可知  $\beta(t)$  表示第  $t$  年每个育龄妇女平均生育婴儿数, 若设在  $t$  年后的一个育龄时期内各个年龄女性生育率  $b_i(t)$  都不变,

则  $\beta(t)$  又可表示为  $\beta(t) = b_{i_1}(t) + b_{i_1+1}(t+1) + \dots + b_{i_2}(t+i_2-i_1)$

即  $\beta(t)$  是第  $t$  年  $i_1$  岁的每位妇女一生平均生育婴儿数, 称总和生育率, 或生育胎次.

记  $b'_i(t) = (1 - d_{00}(t))(1 - d_0(t))h_i(t)k_i(t)$  则  $x_1(t+1) = \beta(t) \sum_{i=i_1}^{i_2} b'_i(t)x_i(t)$

记  $x(t) = [x_1(t), x_2(t), \dots, x_m(t)]^T$

$$A(t) = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 - d_1(t) & 0 & \cdots & 0 \\ 0 & 1 - d_2(t) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 - d_{m-1}(t) & 0 \end{bmatrix}_{m*m}$$

$$B(t) = \begin{bmatrix} 0 & \cdots & 0 & b'_{i_1}(t) & \cdots & b'_{i_2}(t) & 0 & \cdots & 0 \\ 0 & \cdots & 0 \\ \vdots & & & & & & & & \vdots \\ 0 & \cdots & 0 \end{bmatrix}_{m*m}$$

综上  $x(t+1) = A(t)x(t) + \beta(t)B(t)x(t)$ .

## 附录二 部分数据

### §B.1 中国历年人口总数

年份 (1978-1994)	总人口 (单位: 万人)	年份 (1993-2007)	总人口 (单位: 万人)
1978	96259	1993	118517
1979	97542	1994	119850
1980	98705	1995	121121
1981	100072	1996	122389
1982	101654	1997	123626
1983	103008	1998	124761
1984	104357	1999	125786
1985	105851	2000	126743
1986	107507	2001	127627
1987	109300	2002	128453
1988	111026	2003	129227
1989	112704	2004	129988
1990	114333	2005	130756
1991	115823	2006	131448
1992	117171	2007	132129

数据来源: 中国人口和就业统计年鉴 2008.

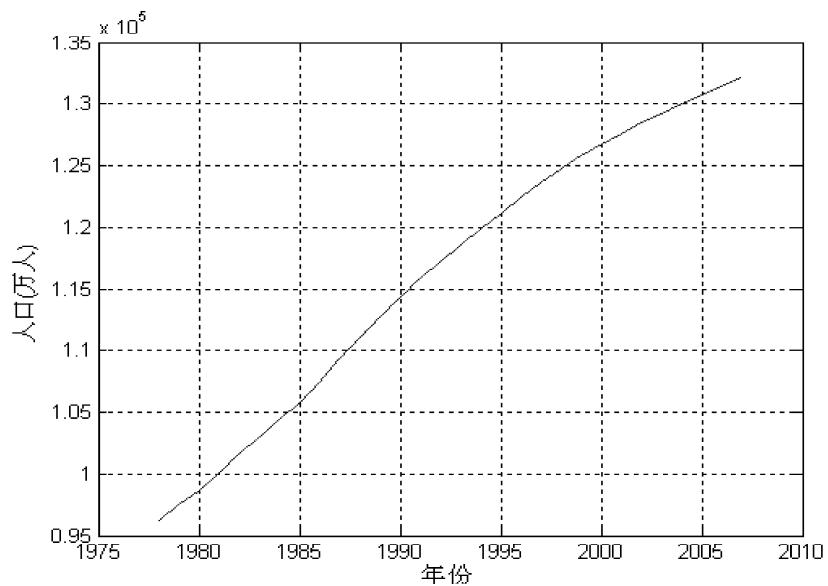


图 B.1: 中国历年人口总数

### 附录三 部分程序

#### §C.1 人口拟合预测的 BP 神经网络模型 Matlab 代码

```
%数据输入
P=1:20;
y=[96259 97542 98705 100072 101654 103008 104357 105851...
107507 109300 111026 112704 114333 115823 117171 118517 ...
119850 121121 122389 123626 124761 124761 125786 126743 ...
127627 128453 129227 129988 130756 131448 132129];
%数据标准化
[YY,minp,maxp]=premnmx(y);
T=YY(1:20);
PR=minmax(P);
r=size(T,1);
```

```
%建立神经网络
netw=newff(PR,[8 r],{'tansig','purelin'});
net=netw;
err=1e-12;
net.trainParam.goal=err;
net.trainParam.lr=0.001;
net.trainParam.epochs=2000;
net.trainParam.show=50;
%训练神经网络
netw=train(net,P,T);
```

```
%预测
P1=[11];y1=[YY(21)];
T1=sim(netw,P1);
P2=22:30;
T2=sim(netw,P2);
P=[P1,P2];
X=P+1978;
T=[T1,T2]
Y=T;
OY=postmnmx(Y,minp,maxp);
```

## §C.2 人口拟合的分形程序 Matlab 代码

```
%主程序
t=[1978:1:2007];
P=[96259 97542 98705 100072 101654 103008 104357 105851 ...
107507 109300 111026 112704 114333 115823 117171 118517 ...
119850 121121 122389 123626 124761 125786 126743 127627 ...
128453 129227 129988 130756 131448 132129];
d=ones(31,1)/4; [a,e,c,f]=GetAECF(t,P,d);
IFSpt(a,e,d,c,f,t(1),P(1),2000,30)
```

```
%程序文件GetAECF.m
function [a,e,c,f]=GetAECF(x,y,d)
L=x(length(x))-x(1);
for i=1:length(x)-1
    a(i+1)=( x(i+1) - x(i) )/L;
    e(i+1)=( x(length(x))*x(i) - x(1)*x(i+1) )/L;
    c(i+1)=( y(i+1)-y(i) - d(i+1)*( y(length(y))-y(1) ) )/L;
    f(i+1)=( x(length(x))*y(i) - x(1)*y(i+1) - d(i+1)*...
        ( x(length(x))*y(1) - x(1)*y(length(y)) ) )/L;
end
```

```
%程序文件IFSpt.m
function IFSpt(a,e,d,c,f,x0,y0,ll,lp)
x(1)=x0; y(1)=y0;
for i=2:ll
    r=ceil(rand(1)*lp);
    x(i)=a(r)*x(i-1)+e(r);
    y(i)=c(r)*x(i-1)+d(r)*y(i-1)+f(r);
end
in=1;
for i=1:ll
if y(i)<140000
    xx(in)=x(i);
    yy(in)=y(i);
    in=in+1;
end
end
[B,ind]=sort(xx);
for i=1:length(yy)
    yyy(i)=yy(ind(i));
end
plot(B,yyy)
axis([1970,2010,90000,140000])
grid on, xlabel('年份'), ylabel('人口(万人)')
```