分类号_____ 密级_____

UDC _____ 编号_____

河南科技大學

硕 士 学 位 论 文

中国人口预测的半参数模型

学 位 申 请 人： 韩玉涛

指 导 教 师： 杨万才

武新乾

学 科 专 业： 应用数学

学 位 类 别： 理学

2011 年 5 月

论文题目：中国人口预测的半参数模型
专　　业：应用数学
研 究 生：韩玉涛
指导教师：杨万才教授　　武新乾副教授

摘　要

有关中国人口预测的模型有很多种，传统的线性模型在实际应用中常常存在设定误差，非参数回归模型则是假定变量间的关系未知，对整个回归函数进行估计，能更好的拟合样本数据，并对数据做出较为精确的预测。半参数模型由于融合了非参数模型和线性模型的优点，具有更强的解释能力，而且还较好的避免了“维数祸根”这一问题，受到诸多学者的广泛关注。近年来，半参数模型在人口建模中也有所应用，但是半参数方法的应用主要集中于核类方法，对于全局光滑的样条方法和基于重抽样思想的 Bootstrap 方法在人口预测中的应用还有待进一步的研究。

本文主要从三个不同的方面建立中国人口预测的半参数回归模型。首先，基于时间序列分析建立线性自回归模型，同时基于多项式样条估计理论建立了半参数自回归模型，将这两种模型分别对中国人口进行拟合与预测对比，结果显示半参数自回归模型优于线性模型；其次，考虑到中国人口与 GDP 存在着高度的相关性，建立以 GDP 为外生变量的半参数回归模型，并与以 GDP 为非参数函数的非参数模型和以 GDP 为线性主部的半参数模型进行对比，结果显示本文建立的以 GDP 为非参数函数的的半参数回归模型更优一些；最后，尝试采用 Bootstrap 方法和多项式样条估计对建立的半参数模型中的参数和非参数函数进行点估计，得到半参数回归方程对中国人口进行拟合及预测，结果表明基于 Bootstrap 方法的半参数回归模型对中国人口拟合和预测精度均较好。

关 键 词：中国人口，线性回归，半参数回归，多项式样条，Bootstrap 方法

论文类型：应用研究

Subject: **Semi-parametric Regressive Prediction Model for Population of China**

Specialty: **Applied Mathematics**

Name: **Han Yu-tao**

Supervisor: **Yang Wan-cai Wu Xin-qian**

ABSTRACT

There are many different prediction models about the total population of China. Setting errors are existed in the practical applications in the traditional linear models, however the whole regression function is estimated and the sample data can be fitted better and predicted accurately in the nonparametric regression models because of assuming the relationships among the variables of the models are unknown. Semi-parametric model has more explanatory power and can avoid the "dimension curse" better with merging the advantages of non-parametric model and linear model, therefore the semi-parametric models are populated by many scholars. In recent years, semi-parametric regressive prediction model has also been applied in the model for population, but the applications of estimation of the semi-parametric model are focuse of kernel methods, spline smoothing method of global methods and Bootstrap method based on the resampling are not found and applied in the population in the literatures.

Semi-parametric regressive prediction model for pupulation of China is done under three main cases. Firsrt, the linear regressive model is done based on theory of time series analysis, and semi-parametric regressive model is estimated based on the theory of spline estimation, then the observations of Chinese population are fitted and predicted using the two models, simulated results show the established semiparametric regression model is superior to some traditional models. Second, a semi-parametric autoregression model with an exogenous variable for population of China is presented considering high correlation between the GDP and chinese population, comparing the prediction results with the non-parametric model with GDP is the non-parmetric funtion and semi-parametric regressive prediction model with GDP is the linear main part of model, the result show that semi-parametric regressive model done in this paper is another better model. Finally, semi-parametric regressive prediction model for pupulation of China is done based on the theory of spline and bootstrap method,

semiparametric autoregression model is set up, and the Chinese population are fitted and predicted,the results show that accuracy of fit and prediction of the total population of China are higher.

KEY WORDS: Population of China, Linear autoregression, Semi-parametric auto-regression, Polynomial spline estimation, Bootstrap method

Dissertation Type: Applied research

目　　录

第1章 绪论

人口问题依然是目前世界普遍关心的重大问题之一，尤其是发展中国家人口的快速增长，引起了国际社会的广泛关注。人口的过快增长已对各国的经济发展产生了重大影响。改革开放至今的 30 多年来，我国经济、社会等各方面的发展都取得了举世瞩目的成就，人民生活得到了极大改善，社会结构和人口素质也发生了深刻变化。然而，人口众多、人均资源少依然是我国的基本国情，人口与经济、社会、资源和环境等方面的不够协调仍是我国发展面临的重大问题之一。因而，对我国人口的发展趋势进行研究有利于政府职能部门的科学决策、有利于人与自然的和谐发展，富有重要的现实意义和应用价值。

1.1 人口模型国内外研究动向及进展

早在 300 多年前, Graunt 和 Erler 等人就开始从人口进化的数量和年龄分布的角度出发对人口的数量进行研究。至到 20 世纪后，Lotka、Sharpe 和 Leslie 等人开始用决定性模型来研究人口动力学的解析理论[1]。随着人口的过快增长，人口对社会影响也与日俱增，人口问题成为世人关注的重大问题之一，于是有关人口建模问题也成为人们研究的又一热点问题，关于人口预测的模型也出现很多种，常见的模型有：

（1） 微分方程模型

上世纪六十年代，中国人口已超过了 8 亿。宋健敏锐的意识到人口过快增长的问题，把控制论的方法引用到人口学中，在 1980 年提出了“人口控制论”，建立人口发展方程预测模型。直到现在，这一理论仍在世界上被广泛应用于人口发展的研究，相关方面可参看文献[2-5]。人口发展方程[1]，假定社会中人口的增长只取决于人口的出生率、死亡率和迁移率，并从这三个方面对人口进行预测。

然而，人口自然增长率[6]是衡量人口变化的又一重要指标，国家统计局对人口自然增长率的计算方法规定为，一定时期人口自然增长数与年平均总人数之比。针对人口自然增长率，出现了对人口预测的自然增长率模型。但模型的假定条件都比较强，不太符合现实。

人口年龄移算法[7]，根据特定时点和特定年龄段的人口按照特定的生存比例推算出同样时点的未来人口的一种建模方法。年龄移算法，其形式相对简单，因此对人口作短期预测时常常被采用。谢建文等(2008)[8]在宋健人口模型上，综合

应用了机理分析、参数辨识以及统计学的一般原理用年龄移算法对未来中国老年化趋势进行了预测。但此方法也有不足之处，该模型是依据某一年的特定数据对未来年份的人口做出预测，使得残差表现为一种有规律的变化趋势，即随着预测时间的逐年增加，残差不断增大，误差也不断增加[7]，不适宜作长期预测。

18 世纪末，英国人 Malthus 假定人口净相对增长率是常数，建立 Malthus 人口指数模型。其形式相对简单，因此在对人口作短期预测时也常被采用。但该模型也有不足之处：把净相对增长率看作常数[9]，不能描述也不能预测较长期的人口发展过程。

1938 年由荷兰数学生物学家 Verhulst 修正了 Malthus 人口模型的基本假设，提出 Logistic 阻滞增长模型。该模型优点在于它不仅考虑了自然资源，而且还考虑了环境条件等因素对人口增长的阻滞作用，较好地描述了人口的增长规律，且预测精度较高，因此得到广泛应用。李百岁(2007)[10]、冯守平(2008)[11]等分别用 Logistic 模型对内蒙古城市化人口及中国人口进行预测，结果均显示，该模型可以弥补指数增长模型的不足，能够对相对较长时期的人口做出预测，而且预测精度较高。但由于仅考虑人口的总增长率，而不涉及人口的年龄结构，又如果阻滞作用并不严重时，预测效果就不是很佳。

以上人口模型都是基于微分方程建立的模型，虽然预测精度较高，但是，如果考虑的变量增多的话，方程就变得比较复杂，求解起来就会增加难度。

（2） Leslie 模型

自 E. G. L ewis(1942)和 P. H. Leslie (1945)建立 Leslie 矩阵以来，Leslie 矩阵(离散模型)在预测人口和分析生物种群等各方面都起着非常重要的作用，随着时间的推移，Leslie 模型也得到了不断地改进和发展,现已成为人口模型中最常用的模型之一。陈文权等(2008)[12]充分利用了各个年龄阶段的指标数据，用修正的 Leslie 模型对我国人口及结构进行预测。Leslie 矩阵较适合于进行时间跨度大及范围大的预测。由于各年龄段人口的出生率与死亡率各不相同，因而不适宜对自然条件下的人口增长进行分析，这就在一定程度上限制了 Leslie 模型的应用。

（3） 随机预测模型

Lee、Tuljapurkar(1991)及 Carter(1990)提出了随机人口预测的 LTC 方法，其中，随机死亡与生育预测克服了 Pollard(1975)经典随机人口模型无法处理的若生育与死亡率发生随机性变化时如何得到相应概率的困难[13-15]。李南、Tuljapurkar S. (1995)[16]将模型扩展为时间-区域序列模型，同时解决了 LTC 方法对发展中国家面临的死亡率进行分析时数据不足的困难，也成功地对中国的随机死亡率进行了预测；李南、申卯兴(1995)[17]用 LTC 方法对中国随机生育率进行了预测；

李南、胡华清(1998)[18]用 LTC 方法实现了中国随机人口的预测。

（4）数理统计方法

回归分析由于其运算简单，易于使用且预测精度较高，因此在人口预测中有着广泛的应用。付莹(2000)[19]用一元线性回归分析了中国 1978-1987 年的人口变化趋势，并对 2000 年的人口进行了预测；由于影响人口发展的因素很多，又提出多元回归分析法，李旭东(2007)[20]应用多元回归模型探讨了喀斯特地质环境下，人口分布的自然地带规律。考虑到人口的发展是个动态的过程，安和平和陈爱平(2004)[21]用自回归模型对我国人口作了预测，其预测精度有所提高。虽然人口的发展是动态的发展过程，然而用线性回归模型对中国人口进行预测时，却忽视了其中存在的非线性关系。

（5）灰色模型

灰色系统由于它含有已知信息，同时又含有未知信息，也是近年来研究的热点之一。用灰色系统模型进行预测，由于所需信息量少、方法简单，且预测精度高，因此在对人口进行预测时，常常被采用。薛臻(2008)[22]、周诗国(2005)[23]用灰色模型对我国人口进行预测，分析了我国人口的增长状况；蒿建华(2008)[24]对西安未来的人口规模进行了预测，均取得了理想的结果。但是，灰色预测模型也有其局限性，虽然反映了数据的规律性，却不能完全反映各种非规律性的社会因素对预测指标的影响[23]。

（6）BP(Back Propagation)神经网络模型

近些年来，随着计算机的不断发展以及误差反向传播算法的提出，使得人工神经网络理论及其在各方面应用都取得了很大的发展。此方法简单且易于使用，同时预测精度较好，因此成为中国人口预测应用最广泛的模型之一。张静和王兴华(2001)[25]利用神经网络对襄樊人口数量进行预测，结果显示较其它的方法精度高；赵方等(2006)[26]对中长时期内人口自然增长率作了预测，与传统的线性回归模型相比预测精度较好。

BP 神经网络模型不仅具有强的非线性映射能力，而且具有柔性网络结构，适合于对非线性方面的问题进行分析，但也容易陷入局部极小的缺陷。

（7）线性时间序列模型

Box-Jenkins 建模法是由博克斯(Box)和詹金斯(Jenkins)建立的一种时间序列分析预测方法。它依据自相关函数和偏自相关函数的统计特性，先确定所要建立的模型类别，再对模型定阶、对模型中的参数估计，之后对模型进行适应性检验，最后采用建立的模型对人口做出预测。由于预测精度较高，因而也受到广泛的应用。胡秋灵和姚文辉(2007)[27]、张静(2009)[28]等用 Box-Jenkins 建模法分别

对我国人口 2004 年及 2005 年的自然增长率作了预测，结论显示此方法预测效果好且简单易行。

以上是人口预测中常用的模型，每一种模型都有其使用范围、优点和缺陷。为了使人口预测的精度更高，也为了扩大模型的适用范围，又有学者将两种或多种方法结合，组成新的模型进行预测，同时也将最新的研究成果用于模型当中。

（8）其它模型

叶小青(2008)[29]将灰色动态 GM(1，1)与微分方程结合，对我国总人口作出预测，估计值与真值比较，误差较小。李国成等(2009)[30]将神经网络和灰色系统结合，用灰色人工神经网络对人口总量进行预测，兼具两个模型的优点，结果显示比两种模型单一使用精度要高；刘金月和许少华(2008)[31]将小波变换和神经网络结合，对人口作出预测，不仅精度更高，且具有较快的收敛速度。

王丽敏和莫君慧(2009)[32]将“小波去噪”与灰色动态 GM(1，1)和 BP 神经网络相结合，对中国人口结构、分布、出生率等七个指标进行预测，结果不仅对人口发展做出了预测，也反映了我国人口的结构、分布等其它的一些特点。

巩永丽、张德生和武新乾(2007)[33]用非参数自回归预测模型对人口增长率作了分析，结果显示相对于线性回归模型，非参数自回归模型能够较好地解决人口增长预测这一非线性问题,且预测精度较高。姜爱平等(2007)[34]用具有外生变量的半参数自回归模型，用核估计法对人口进行预测，并与自回归模型进行对比，结果显示用半参数方法，预测效果更好。

1.2 半参数模型国内外研究动向及进展

Engle 等（1986）[35]在研究气候条件对电力需求影响时提出半参数回归模型

$$Y_i = X_i'\beta + g(T_i) + e_i \quad , \quad i = 1, 2, \cdots, n ,$$

其中 X_i 和 T_i 为解释性变量，Y_i 为被解释性变量，β 为线性主部参数，$g(\cdot)$ 为未知非参数函数，e_i 为随机误差。

由于它结合了线性回归模型和非参数回归模型的特点，吸收了各自的优点，不论是在理论研究上还是实际应用中都具有重要意义，因此受到了诸多统计学者的关注，成为近年来统计研究的热点之一。

（1） 基本模型上的研究

Heckman(1988)[36]、Rice(1986)[37]分别研究了在 $\{(X_i, T_i), 1 \le i \le n\}$ 为 $i \cdot i \cdot d \cdot$ 随机设计、固定设计下，$g(\cdot)$ 选取一类样条估计时，β 的加权最小二乘估计 $\hat{\beta}_n$ 的渐近正态性，及 β 的估计的协方差函数的渐近性质。

我国学者洪圣岩(1992)、高集体（1992）等分别研究了模型中当 $g(\cdot)$ 取 Parzen-

Rosenblatt 核估计及取一类核估计序列时，β 的加权最小二乘估计 $\hat{\beta}_n$ 的渐近正态性及 $\hat{\beta}_n$ 和 $\hat{g}_n$ 的强弱收敛速度[38]，还进一步给出了 $g(\cdot)$ 取一类近邻估计时 β 的估计性质。

柴根象等(1995)假定 $\{X_i\}$ 为 p 维固定设计向量，$\{T_i\}$、$\{e_i\}$ 为 $i\cdot i\cdot d\cdot$ 点列且相互独立，用两阶段估计方法，在比较弱的条件下证明了 $\hat{\beta}_n$ 的强收敛速度及 $\hat{\beta}_n$ 的渐近正态性。钱伟民等(1998)在此基础上，又讨论了 $\hat{\sigma}_n^2$ 的 t 统计量的渐近正态性及 $\hat{\sigma}_n^2$ 的 Bootstrap 逼近问题，证明了适当条件下 Bootstrap 逼近成立[38]。

施沛德(1992)研究了当 $g(\cdot)$ 取逐点多项式逼近及 B −样条逼近时参数分量和非参数分量的 M −估计的若干性质。薛留根(2002)在一般条件下证明了用随机加权统计量的分布逼近原估计量误差的分布的强有效性，给出了 M 估计的最优强收敛速度[38]。

陈明华(1999)、胡舒合(1994)、吴本忠(1998)等研究了误差 e_i 为独立序列、ρ −混合和 φ −混合序列时，β 和 $g(\cdot)$ 的加权最小二乘估计 $\hat{\beta}_n$ 和 $\hat{g}_n$ 的强相合性和 $\hat{g}_n$ 的一致强相合性；柴根象(2003)进一步研究了 e_i 为 α −混合序列时，估计的渐近正态性和收敛速度；潘光明和胡舒合(2003)则将误差 e_i 逐步推广 L^P −Mixin−gale 情形，研究了估计量的 q −阶 $(q>1)$ 平均相合性[38]。

（2）数据存在截断或删失时的研究及推广

潘建敏(1997)[39]、薛留根(1999)[40]等分别对污染数据、随机删失及不同的截断数据的情况下，对半参数回归模型进行了详细的研究，证明了估计量的渐近正态性和强相合性。叶阿忠、吴相波和黄志刚(2009)[41]提出半参数计量经济联立模型的局部线性工具变量变窗宽估计，并证明了参数分量和非参数分量估计(在内点处)的渐近正态性和一致性，得到它们的收敛速度快于非参数模型估计的收敛速度，克服和弥补了非参数计量经济联立模型估计收敛速度慢的缺陷，也使得联立模型的估计计理论更具有实用价值。

（3） 半参数模型的应用

黄四民、梁华(1994)[42]把半参数模型用在居民消费结构方面，分析我国居民消费结构变动趋势，通过与线性回归模型得到的结果进行比较，证明了半参数方法的预测精度具有更高的优越性；王一兵(2005)[43]将模型用于商品房价格指数，由于半参数模型融合了线数模型和非参数模型的优点，事先不需要对未知函数的分布做任何假设，并且可以充分利用前期的信息来构建模型的线性部分，因此能够准确的描述和预测价格的剧烈变化或者异常的发生，与最小二乘法进行对比，结果表明用半参数回归分析，其拟合效果明显优于即使很复杂的最小二乘法；冯春山、蒋馥和吴家春(2005)[44]等用半参数方法，对石油市场收益实证分析，结果

也表明在 99%置信水平下半参数方法的风险计量效果高于通常基于正态分布假设的参数法。

半参数方法正日益受到越来越多的学者的重视，然而，半参数方法的应用主要集中于核类方法，这种方法属于局部方法，对于全局光滑的样条方法和基于重抽样思想的 Bootstrap 方法在中国人口预测中的应用还尚未见有文献研究。

1.3 本文的研究目的和研究内容

人口问题依然是目前世界普遍关心的重大问题之一。人口众多、人均资源少依然是我国的基本国情，人口与经济、社会、资源和环境等各方面的不够协调仍是我国目前发展面临的重要问题之一。 因而，对我国人口的发展趋势进行研究有利于政府职能部门的科学决策、有利于人与自然的和谐发展，富有重要的现实意义和应用价值。

有关经济发展和人口增长的关系问题，国内外学者已进行了大量的实证研究[52-56]。他们已证明了中国人口总量与 GDP 之间存在着高度的相关性，但二者之间并非是简单的线性回归关系。非参数模型假定变量之间的关系未知，回归函数的形式可以是任意的，因此能够更好的拟合样本数据，并能对数据做出比较精确的预测。

半参数模型由于融合了非参数模型和线性模型的优点，与传统的线性模型和非参数模型相比，半参数模型具有更强的解释能力，而且还能很好的避免“维数祸根”这一问题，因而受到了诸多学者的广泛关注，成为又一个热点研究方向。近年来，半参数模型在人口建模中也有所应用。但是对于半参数回归方法的应用主要集中于核类方法，这种方法属于局部方法，它不能够给出所要拟合的模型简单的显式表达式，并且计算量大、运行时间较长，而多项式样条估计是全局光滑方法，能较好地克服上述核估计的弊端[50]，对于全局光滑的样条方法在人口预测中还尚未见有文献研究。

同时考虑到人口数据主要受到前期人口基数的影响，GDP 总量对人口数据趋势作局部调整，本文将人口数据的滞后变量作为半参数模型的线性主部变量，以人口数据的另一显著滞后变量和 GDP 总量分别作为非参数部分变量, 基于多项式样条估计理论，建立中国人口预测的半参数自回归模型及具有外生变量的中国人口预测的半参数回归模型。

Bootstrap方法基于重抽样思想，在具体应用中只根据给定的观测信息，不需要事先假设总体的分布，也不用增加新的样本信息，只是对样本重构并不断计算相关的估计值，进而对总体的分布特性做出推断，一定程度上解决了误差项分布

未知而导致的推断失误。Bootstrap方法做为一种新兴的非参数估计方法，还未见Bootstrap方法应用在半参数回归模型中对中国人口进行预测。因此，本文尝试基于多项式样条方法和Bootstrap方法建立中国人口的半参数自回归模型。

1.4 本文的结构安排

本文结构安排如下：

第一章 绪论。较为详细地介绍了人口预测模型及半参数模型的发展状况，并对本文的研究目的和研究内容做了简单介绍。

第二章 中国人口预测的半参数自回归模型。首先介绍半参数模型及多项式样条估计，其次建立中国人口预测的线性自回归模型，再次基于线性回归选择显著滞后变量，利用多项式样条方法估计建立半参数自回归方程。

第三章 中国人口预测的具有外生变量的半参数回归模型。考虑到中国人口总量和 GDP 总量之间的关系，建立中国人口预测的具有外生变量的半参数回归模型，用多项式样条估计对所建立模型的参数进行估计，建立半参数回归方程，对中国人口进行拟合及预测。

第四章 基于 Bootstrap 方法的中国人口预测的半参数模型。首先对 Bootstrap 方法做了简单概述，其次建立中国人口预测的半参数模型，对模型中线性主部的参数和非参数函数做 Bootstrap 估计，建立 Bootstrap 半参数回归方程，最后对中国人口进行拟合和预测。

第五章 结论。总结本论文主要研究的工作，同时为尚待研究的问题指明研究方向。

1.5 本章小结

本章首先对人口预测模型国内外的发展状况做了概述，接着介绍了半参数模型的研究状况，然后介绍了本文的研究背景、思路及内容，最后给出了结构安排。

第2章 中国人口预测的半参数自回归模型

本章首先基于时间序列分析理论对中国人口建立线性自回归模型，用最小二乘估计建立线性自回归方程；其次，基于线性回归选择显著滞后变量和多项式样条方法估计理论，建立中国人口预测的半参数自回归方程；再次，将建立的两种模型对中国人口分别进行拟合及预测分析，同时将本章建立的半参数自回归模型与Logistic模型、Leslie模型、灰色神经网络模型对中国人口预测的结果进行对比分析；最后基于本章建立的半参数自回归模型对2010-2013年中国总人口数量进行预报。

2.1 中国人口预测的线性自回归模型的建立

2.1.1 数据的平稳化处理

本章中用到的原始人口数据来源于中华人民共和国国家统计局官方网站“http://www.stats.gov.cn/”。为了便于同其它模型对比，本章中用到的样本数据为中国 1949-2008 年总人口。

对中国1949-2008年60个原始人口数据作时序图，得到图2-1。

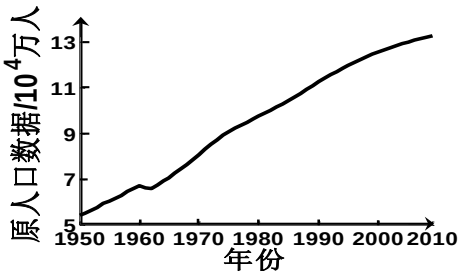


图2-1 1949-2008年原人口数据时序图

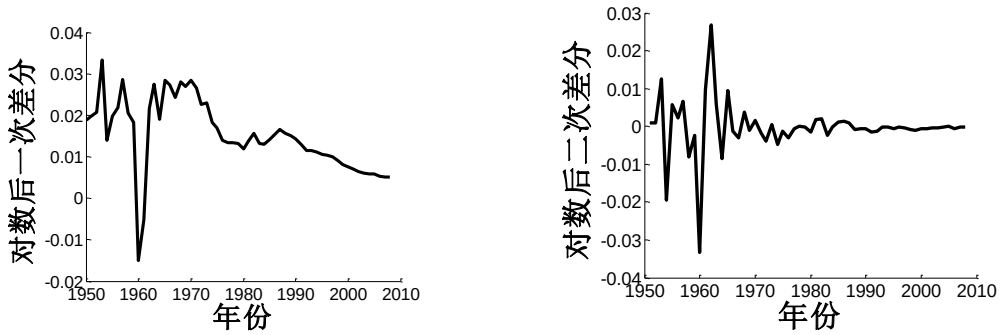
Fig.2-1 Timing diagram of the original population of 1949-2008

从图2-1可以看到数据是不平稳的。根据线性自回归模型的要求，先对原始人口数据序列做对数处理，对对数变换后的序列做一次差分，见图2-2(a)。

从图2-2(a)中可以看到一次差分后的序列依然是不平稳的。对一次差分后的序列，再进行一次差分，即对对数变换后的序列进行二次差分，若记$\{Y_t\}$为中国总人口序列，$\{\nabla^2 \ln(Y_t)\}$为对数变换后二次差分序列，令

$$W_t = \nabla^2 \ln(Y_t) - \overline{\nabla^2 \ln(Y_t)},$$

其中 $\overline{\nabla^2 \ln(Y_t)}$ 为 $\{\nabla^2 \ln(Y_t)\}$ 的均值，∇ 为差分符号，见图2-2(b)。



(a)一次差分时序图　　(b) 二次差分时序图

(a) Timing diagram of one difference　　(b) Timing diagram of two differences

图2-2 对数变换后差分时序图

Fig.2-2 Differential timing diagram of one difference after log transformation

2.1.2 数据的平稳性检验

从图2-2(b)可直观的判断序列 $\{W_t\}$ 是平稳的。为进一步说明序列的平稳性，再进行游程检验。

游程检验法[50]，是在保持序列原有顺序的情况下，设序列 $\{X_t\}$ 的均值为 $\bar{X}$，序列中把大于或小于 $\bar{X}$ 的观察值分别用符号“+”和“－”表示。这样得到一个符号序列，在记号序列中每一段连续相同的记号序列叫做一个游程。

设序列长度为 N，$N = N_1 + N_2$，其中 N_1 和 N_2 分别是序列中“+”、“－”出现的次数，游程总数为 r。当 N_1 和 N_2 不超过 15 时，游程总数服从 r 分布：

$$E(r) = \frac{2N_1N_2}{N} + 1,\quad D(r) = \frac{2N_1N_2(2N_1N_2 - N)}{N^2(N-1)}。$$

当 N_1 和 N_2 大于 15 时，统计量 $Z = \dfrac{r - E(r)}{\sqrt{D(r)}}$ 渐近服从于 $N(0,1)$ 分布。

在显著水平 $\alpha = 0.05$ 下，查表若统计量 $|Z| \le 1.96$，则认为此序列是平稳的，否则认为是非平稳的。

由上可知，游程总数 $r = 24$，序列长度 $N = 53$，“+”、“−”出现的次数分别为 $N_1 = 25$，$N_2 = 28$，这里 N_1 和 N_2 均大于15，则统计量 $Z = \dfrac{r - E(r)}{\sqrt{D(r)}}$ 渐近服从于 $N(0,1)$ 分布。

$$E(r)=2N_1N_2/N+1=27.415,$$
$$D(r)=2N_1N_2(2N_1N_2-N)/\left[N^2(N-1)\right]=12.910,$$
$$Z=(r-E(r))/\sqrt{D(r)}=(24-27.415)/\sqrt{12.910}=-0.9504。$$

在显著性水平 $\alpha=0.05$ 下，$|Z|<1.96$，因此判定序列 $\{W_t\}$ 是平稳的。

2.1.3 模型阶数的确定

本文选取AIC、BIC准则及残差方差图来初步确定 *AR* 模型的阶数。由MATLAB运行结果(见图2-3和2-4)，可确定滞后7阶是较为理想的。

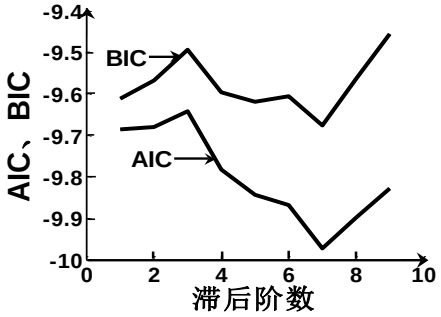


图2-3 AIC、BIC图

Fig.2-3 Figures of AIC& BIC

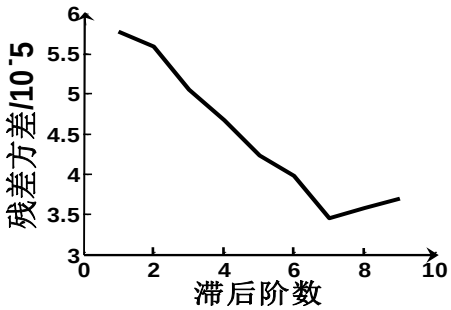


图2-4 残差方差图

Fig.2-4 Figure of residual variance

再用F检验法[50]检验线性自回归模型 $AR(p)$ 的阶数。首先对 $\{W_t\}$ 分别拟合 $AR(6)$ 和 $AR(7)$ 模型，两种模型的残差平方和 Q_1 和 Q_0 分别为0.0016和0.0013，则

$$F=\frac{(Q_1-Q_0)/S}{Q_0/(N-r)}=\frac{(0.0016-0.0013)/1}{0.0013/(53-7)}=10.62,$$

其中 S 为舍弃因子的个数，N 为样本容量，r 为回归因子个数。给定显著性水平 $\alpha=0.05$，查 F 分布表得 $F_\alpha(1,46)=4.05$，$F>>F_\alpha$，说明 $AR(6)$ 和 $AR(7)$ 有显著的差异，模型阶数有上升的可能。再拟合 $AR(8)$ 模型，其残差平方和为0. 0013，与 $AR(7)$ 比较有

$$F=\frac{(0.0013-0.0013)/1}{0.0013/(53-8)}=0。$$

同理查表得 $F_\alpha(1,45)=4.05$，$F<<F_\alpha$，故 $AR(7)$ 与 $AR(8)$ 没有显著差异，即选择 $AR(7)$ 是合适的。

2.1.4 参数的估计

根据2.1.1中平稳化处理后的1951-2003年中国人口数据，对平稳序列 $\{W_t\}$ 建

立$AR(7)$模型，用最小二乘估计确定其中的参数，建立线性自回归方程：

$$W_t = 0.15W_{t-1} - 0.53W_{t-2} - 0.12W_{t-3} - 0.11W_{t-4} - 0.35W_{t-5} + 0.24W_{t-6} - 0.35W_{t-7}。 \quad (2\text{-}1)$$

对建立的线性自回归方程(2-1)中各变量进行显著性检验，在显著性水平$\alpha = 0.05$下，只有W_{t-2}，W_{t-5}，W_{t-7}的系数是显著的(见表2-1)。

表2-1 方程(2-1)各变量系数的显著性检验

Tab.2-1 Significance test of the variable coefficients of equation (2-1)

参数	估计值	标准差	t-统计量	显著性概率
C	-0.0004	0.0004	-0.8815	0.3836
C(1)	0.1512	0.1472	1.0271	0.3109
C(2)	-0.5331	0.1393	-3.8272	0.0005
C(3)	-0.1233	0.1423	-0.8664	0.3917
C(4)	-0.1070	0.1387	-0.7717	0.4451
C(5)	-0.3480	0.1301	-2.6747	0.0110
C(6)	0.2438	0.1254	1.9437	0.0594
C(7)	-0.3517	0.1314	-2.6759	0.0109

注：C为常数项，$C(i)(i=1,\cdots,7)$为方程(2-1)中$\{W_t\}$相应变量的系数。

选取显著性变量W_{t-2}，W_{t-5}和W_{t-7}，重新估计相应系数，得到线性自回归方程

$$W_t = -0.44W_{t-2} - 0.29W_{t-5} - 0.30W_{t-7}。 \quad (2\text{-}2)$$

方程(2-2)的残差平方和为0.0017，将其与方程(2-1)比较，同上做F检验，得$F = 2.83 < F_\alpha = 4.05$，说明两个线性自回归方程没有显著差异。

2.1.5 模型的适应性检验

由文献[50]知，若拟合的模型合适，统计量$Q = N\sum_{k=1}^{L(N)}\rho_k^2(N,\varepsilon_t)$，近似服从$x^2(L(N)-p-q)$分布，其中$N$为样本容量，$\rho_k^2(N,\varepsilon_t)$为残差序列$\{\varepsilon_t\}$的自相关函数，$L(N)=\sqrt{N}$为自相关系数的个数，$p$和$q$为模型参数个数。

通过计算得Q=7.66，在显著性水平$\alpha = 0.05$下，查表得$x_{0.95}^2(4) = 9.49$，$Q < x_{0.95}^2(7-3)$，说明ε_t是独立的，即模型是合适的，可选取方程(2-2)对中国

2004-2009年人口进行预测。

2.2 半参数自回归模型及半参数模型的建立

2.2.1 半参数自回归模型和多项式样条估计

本文考虑的半参数回归模型的一般形式为：

$$Y_t = \alpha^T X_t + g(Z_t) + \varepsilon_t, \quad t = 1,2,\cdots, n, \tag{2-3}$$

其中Y_t为被解释变量，α是未知参数向量，$X_t = (X_{t1},\cdots,X_{tp})^T = (Y_{t-1},\cdots,Y_{t-p})^T$为解释性变量，线性主部$\alpha^T X_t$把握被解释变量的大势走向；$g(\cdot)$为未知非参数光滑函数，对被解释变量作局部调整；随机误差序列$\{\varepsilon_t\}$独立同分布，满足：

$$E(\varepsilon_t) = 0, \quad \mathrm{Var}(\varepsilon_t) = \sigma^2 < \infty,$$

且ε_t与$Y_s(s<t)$相互独立。

对模型(2-3)中参数向量α和非参数函数$g(\cdot)$的估计，本文采用武新乾等(2007)[49]中的多项式样条方法。不妨在紧区间$[a,b]$上建立k次多项式样条空间，其结点序列为：

$$a = z_0 < z_1 < \cdots < z_{N_n} < z_{N_n+1} = b,$$

基函数为$B_s(\cdot)$ $(s=1,\cdots,K)$，则存在常数向量$\beta = (\beta_1,\cdots,\beta_K)^T$，使得

$$g(z) \approx \sum_{s=1}^{K} \beta_s B_s(z),$$

这里$K = N_n + k + 1$，k为多项式样条的次数。

记$Y = (Y_1,\cdots,Y_n)^T$，I为$n\times n$单位矩阵，

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}, \quad B = \begin{pmatrix} B_1(Z_1) & B_2(Z_1) & \cdots & B_K(Z_1) \\ B_1(Z_2) & B_2(Z_2) & \cdots & B_K(Z_2) \\ \vdots & \vdots & & \vdots \\ B_1(Z_n) & B_2(Z_n) & \cdots & B_K(Z_n) \end{pmatrix},$$

$$A = B(B^T B)^{-1} B^T, \; H = I - X\left[X^T (I-A) X\right]^{-1} X^T (I-A),$$

最小化

$$m(\alpha,\beta) = \sum_{t=1}^{n} \left\{ Y_t - \alpha^T X_t - \sum_{s=1}^{K} \beta_s B_s(Z_t) \right\}^2,$$

可得α和β的估计，即

$$\hat{\alpha} = \left[X^T (I-A) X\right]^{-1} X^T (I-A) Y, \; \hat{\beta} = (B^T B)^{-1} B^T H Y,$$

从而得到 g 的样条估计为 $\hat{g}(z)=\sum_{s=1}^{K}\hat{\beta}_s B_s(z)$。

在实际应用中，取内结点 $N_n=\left[n^{2/(4k+3)}\right]$，即 N_n 为 $n^{2/(4k+3)}$ 的整数部分，基函数 $B_s(\cdot)$ 为 $1,z,\cdots,z^k,(z-z_1)_+^k,\cdots,(z-z_{N_n})_+^k$ ，其中

$$(z-z_i)_+=\max\{0,z-z_i\}(i=1,2,\cdots,N_n)。$$

2.2.2 中国人口半参数模型的建立

基于线性回归选取的显著性变量，分别选取滞后2阶、5阶和7阶做为非参数部分，其余二变量做为线性部分，由MATLAB7.0 运行结果，得到相应的半参数自回归方程：

$$W_t=-0.45W_{t-2}-0.29W_{t-5}+\hat{g}(W_{t-7}),\quad (2\text{-}4)$$

$$W_t=-0.45W_{t-2}-0.30W_{t-7}+\hat{g}(W_{t-5}),\quad (2\text{-}5)$$

$$W_t=-0.23W_{t-5}-0.26W_{t-7}+\hat{g}(W_{t-2})。\quad (2\text{-}6)$$

将建立的三个半参数自回归方程分别对中国 1958-2003 年人口进行拟合，并得到拟合的均方误差，结果详见表 2-2。

表 2-2 三个半参数自回归方程对中国人口拟合及预测

Tab.2-2 Fitting and prediction results of Chinese population are done using three semi-parametric autoregressive equations

年份	实际人口（万人）	半参数方程(2-4)		半参数方程(2-5)		半参数方程(2-6)	
		拟合人口（万人）	相对误差(%)	拟合人口（万人）	相对误差(%)	拟合人口（万人）	相对误差(%)
1958	65994	66153.55	0.24	66168.05	0.26	66355.28	0.55
1959	67207	67475.00	0.40	67487.30	0.42	67565.13	0.53
1960	66207	68201.68	3.01	68267.91	3.11	67391.71	1.79
1961	65859	65498.60	-0.55	65587.71	-0.41	65424.96	-0.66
1962	67295	66182.70	-1.65	66203.79	-1.62	67294.27	0.00
1963	69172	68522.19	-0.94	68525.78	-0.93	68611.41	-0.81
1964	70499	70065.29	-0.62	70102.58	-0.56	70518.91	0.03
1965	72538	72469.68	-0.09	72493.89	-0.06	72496.07	-0.06

续表 2-2

Continued 2-2

1966	74542	74720.70	0.24	74720.21	0.24	73650.35	-1.20
1967	76368	76434.12	0.09	76376.02	0.01	76546.35	0.23
1968	78534	77803.42	-0.93	77867.61	-0.85	77888.81	-0.82
1969	80671	80460.43	-0.26	80350.65	-0.40	80172.36	-0.62
1970	82992	82255.50	-0.89	82304.25	-0.83	82577.93	-0.50
1971	85229	85596.45	0.43	85607.42	0.44	85550.65	0.38
1972	87177	87165.23	-0.01	87221.61	0.05	87410.51	0.27
1973	89211	89123.66	-0.10	89114.88	-0.11	89034.83	-0.20
1974	90859	91518.91	0.73	91498.15	0.70	91070.85	0.23
1975	92420	92286.70	-0.14	92303.16	-0.13	92438.37	0.02
1976	93717	94233.96	0.55	94213.31	0.53	93639.07	-0.08
1977	94974	95083.60	0.12	95075.78	0.11	95023.27	0.05
1978	96259	96364.77	0.11	96346.38	0.09	96059.78	-0.21
1979	97542	97783.65	0.25	97762.85	0.23	97749.30	0.21
1980	98705	98793.24	0.09	98781.49	0.08	98872.08	0.17
1981	100072	100062.41	-0.01	100044.73	-0.03	100076.86	0.00
1982	101654	101518.12	-0.13	101498.87	-0.15	101435.57	-0.21
1983	103008	103205.86	0.19	103187.78	0.17	103391.77	0.37
1984	104357	104249.85	-0.10	104234.87	-0.12	104452.03	0.09
1985	105851	105810.92	-0.04	105794.78	-0.05	105583.75	-0.25
1986	107507	107255.74	-0.23	107245.44	-0.24	107325.18	-0.17
1987	109300	109049.28	-0.23	109035.13	-0.24	109226.83	-0.07
1988	111026	110987.58	-0.03	110983.94	-0.04	111178.50	0.14
1989	112704	112582.31	-0.11	112581.98	-0.11	112772.32	0.06
1990	114333	114431.24	0.09	114411.63	0.07	114409.47	0.07
1991	115823	115910.31	0.08	115897.30	0.06	115931.20	0.09
1992	117171	117205.33	0.03	117200.75	0.03	117242.80	0.06
1993	118517	118507.13	-0.01	118498.87	-0.02	118434.66	-0.07
1994	119850	119854.94	0.00	119844.71	0.00	119783.09	-0.06
1995	121121	121187.26	0.05	121166.47	0.04	121246.86	0.10

续表 2-2

Continued 2-2

1996	122389	122417.81	0.02	122396.33	0.01	122464.37	0.06
1997	123626	123703.32	0.06	123681.47	0.04	123703.77	0.06
1998	124761	124867.45	0.09	124845.13	0.07	124929.69	0.14
1999	125786	125913.18	0.10	125890.42	0.08	125949.81	0.13
2000	126743	126820.09	0.06	126801.54	0.05	126795.09	0.04
2001	127627	127695.54	0.05	127677.38	0.04	127668.08	0.03
2002	128453	128510.82	0.05	128490.18	0.03	128524.03	0.06
2003	129227	129278.13	0.04	129258.95	0.02	129288.96	0.05
拟合的均方误差		428.71		427.80		318.13	
2004	129988	130000.98	0.01	129983.08	0.00	130001.98	0.01
2005	130756	130784.31	0.02	130727.18	-0.02	130788.50	0.02
2006	131448	131555.70	0.08	131445.72	0.00	131600.51	0.12
2007	132129	132288.98	0.12	132116.51	-0.01	132422.47	0.22
2008	132802	132992.34	0.14	132743.83	-0.04	133251.16	0.34
预测的均方误差		40.21		9.78		82.41	

注：1. 资料来源：中华人民共和国国家统计局官方网站(以下各表同)；

2. 表 2-2 中相对误差$=\left(\hat{Y}_i - Y_i\right)\Big/Y_i \times 100$，$i=1,\cdots,n$。

从表 2-2 可以看到，三个半参数自回归方程对中国 1958-2003 年人口拟合的相对误差均较小，半参数回归方程(2-6)对人口拟合的均方误差最小，但是对中国 2004-2008 年的总人口进行预测时，相对误差的绝对值整体上大于其余两个半参数回归方程。半参数回归方程(2-5)对中国人口拟合的相对误差略小于方程(2-6)，在对中国 2004-2008 年人口预测时，其预测的相对误差的绝对值最小，并且对中国人口进行预测的均方误差也明显小于其余两个半参数回归方程。因此选取半参数回归方程(2-5)与线性回归方程(2-2)做对比。

2.3 不同模型对人口预测的对比分析

2.3.1 线性时间序列模型与半参数回归模型对中国人口的拟合

利用本章建立的线性自回归方程(2-2)与半参数回归方程(2-5)分别对 1958-

2003 年中国人口进行拟合，结果详见表 2-3。

图2-2(b)可以直观的看到在1958-1970年间的序列波动比较大，1970年之后序列的波动较小。图形中波动越大，说明此序列在此区间越不平稳。从表2-3中可以看到，总的来说线性自回归模型和半参数回归模型对中国1958-2003年人口拟合的较好，但较之于半参数回归模型对中国人口拟合的均方误差，线性回归模型的均方误差略大一些。

从表2-3中还可以看到，对于序列不太平稳(1958-1970年)的时候，线性回归模型与半参数回归模型对人口拟合的相对误差均较大，但是随着时间的推移，序列渐趋于平稳化，此时，半参数回归模型对人口拟合的相对误差的绝对值从整体上看要略小于线性回归模型。

表 2-3 两种模型对 1958-2003 年中国人口拟合的结果对比

Tab.2-3 Fitting results of Chinese population from 1958 to 2003 are done using the two models

年份	实际人口(万人)	线性自回归模型			半参数自回归模型		
		总人口拟合（万人）	拟合误差（万人）	相对误差(%)	总人口拟合（万人）	拟合误差（万人）	相对误差(%)
1958	65994	66167.30	173.30	0.26	66168.05	174.05	0.26
1959	67207	67491.69	284.69	0.42	67487.30	280.30	0.42
1960	66207	68280.87	2073.87	3.13	68267.91	2060.91	3.11
1961	65859	65600.22	-258.78	-0.39	65587.71	-271.29	-0.41
1962	67295	66217.92	-1077.08	-1.60	66203.79	-1091.21	-1.62
1963	69172	68545.27	-626.73	-0.91	68525.78	-646.22	-0.93
1964	70499	70120.92	-378.08	-0.54	70102.58	-396.42	-0.56
1965	72538	72500.01	-37.99	-0.05	72493.89	-44.11	-0.06
1966	74542	74724.59	182.59	0.24	74720.21	178.21	0.24
1967	76368	76403.29	35.29	0.05	76376.02	8.02	0.01
1968	78534	77880.25	-653.75	-0.83	77867.61	-666.39	-0.85
1969	80671	80380.37	-290.63	-0.36	80350.65	-320.35	-0.40
1970	82992	82308.53	-683.47	-0.82	82304.25	-687.75	-0.83
1971	85229	85631.60	402.60	0.47	85607.42	378.42	0.44
1972	87177	87251.11	74.11	0.09	87221.61	44.61	0.05
1973	89211	89132.22	-78.78	-0.09	89114.88	-96.12	-0.11
1974	90859	91525.65	666.65	0.73	91498.15	639.15	0.70

续表 2-3

Continued 2-3

1975	92420	92326.75	-93.25	-0.10	92303.16	-116.84	-0.13
1976	93717	94243.07	526.07	0.56	94213.31	496.31	0.53
1977	94974	95107.23	133.23	0.14	95075.78	101.78	0.11
1978	96259	96372.74	113.74	0.12	96346.38	87.38	0.09
1979	97542	97793.82	251.82	0.26	97762.85	220.85	0.23
1980	98705	98811.19	106.19	0.11	98781.49	76.49	0.08
1981	100072	100075.62	3.62	0.00	100044.73	-27.27	-0.03
1982	101654	101528.32	-125.68	-0.12	101498.87	-155.13	-0.15
1983	103008	103215.08	207.08	0.20	103187.78	179.78	0.17
1984	104357	104263.66	-93.34	-0.09	104234.87	-122.13	-0.12
1985	105851	105827.36	-23.64	-0.02	105794.78	-56.22	-0.05
1986	107507	107271.07	-235.93	-0.22	107245.44	-261.56	-0.24
1987	109300	109060.42	-239.58	-0.22	109035.13	-264.87	-0.24
1988	111026	111018.94	-7.06	-0.01	110983.94	-42.06	-0.04
1989	112704	112614.06	-89.94	-0.08	112581.98	-122.02	-0.11
1990	114333	114440.33	107.33	0.09	114411.63	78.63	0.07
1991	115823	115926.67	103.67	0.09	115897.30	74.30	0.06
1992	117171	117231.58	60.58	0.05	117200.75	29.75	0.03
1993	118517	118534.47	17.47	0.01	118498.87	-18.13	-0.02
1994	119850	119880.23	30.23	0.03	119844.71	-5.29	0.00
1995	121121	121201.42	80.42	0.07	121166.47	45.47	0.04
1996	122389	122433.13	44.13	0.04	122396.33	7.33	0.01
1997	123626	123718.74	92.74	0.08	123681.47	55.47	0.04
1998	124761	124879.90	118.90	0.10	124845.13	84.13	0.07
1999	125786	125925.75	139.75	0.11	125890.42	104.42	0.08
2000	126743	126838.54	95.54	0.08	126801.54	58.54	0.05
2001	127627	127713.51	86.51	0.07	127677.38	50.38	0.04
2002	128453	128526.82	73.82	0.06	128490.18	37.18	0.03
2003	129227	129297.17	70.17	0.05	129258.95	31.95	0.02
均方误差：			429.68			427.80	

注：资料来源：中华人民共和国国家统计局官方网站。

2.3.2 线性时间序列模型与半参数回归模型对中国人口的预测

选取线性自回归方程(2-2)与半参数自回归方程(2-5)分别对2004-2009年人口进行预测(见表2-4)。

从表2-4可以看到，线性模型的短期(2年)预测效果还是比较好的，但是随着年数的增加，预测误差递增的速度比较快：从第一年误差的33万人很快的增长到第六年的489万人。相对于线性自回归模型，半参数自回归模型对人口中预测的相对误差的绝对值明显均较小，且均方误差也显著小于线性回归模型的均方误差。虽然半参数回归模型对中国人口预测的误差也在逐年增大，但是预测6年的误差约为线性自回归模型的1/5.6、1/2.6、1/24.4、1/13.8、1/5.6、1/3.7。

表2-4 线性自回归模型和半参数回归模型对2004-2009年人口预测结果

Tab.2-4 Predictions of Chinese population from 2004 to 2009 are done using linear regression model and semi-parametric regression model

年份	实际人口（万人）	线性自回归模型			半参数自回归模型		
		人口拟合（万人）	预测残差（万人）	相对误差(%)	人口拟合（万人）	预测残差（万人）	相对误差(%)
2004	129988	130021.26	33.26	0.03	129982.05	-5.95	0.00
2005	130756	130841.36	85.36	0.07	130724.08	-31.92	-0.02
2006	131448	131656.50	208.50	0.16	131439.47	-8.53	-0.01
2007	132129	132444.99	315.99	0.24	132106.04	-22.96	-0.02
2008	132802	133219.34	417.34	0.31	132728.05	-73.95	-0.06
2009	133474	133963.55	489.55	0.37	133336.95	-137.05	-0.10
均方误差		16615.05			65.71		

注：1. 资料来源：中华人民共和国国家统计局官方网站；

2. 表2-4中相对误差＝$\left(\hat{Y}_i - Y_i\right)\Big/Y_i \times 100$，预测残差＝$\hat{Y}_i - Y_i$，$i = 1, \cdots, n$。

2.3.3 半参数回归模型与其它模型的对比

将本章建立的半参数自回归模型和其它模型对中国2005-2008年人口预测的结果进行对比(见表2-5)。

从表2-5中可以看到Logistic模型和灰色神经网络模型对中国人口进行预测时

的预测残差较小，但是从整体上看半参数自回归模型对中国人口进行预测时的残差的绝对值最小。

表2-5 几种模型对中国2005-2008年人口的预测(单位:万人)

Tab.2-5 Predictions of Chinese population from 2005 to 2008 are done using other models

年份	原人口	半参数自回归模型		Logisti 模型[45]		Leslie 模型[12]		灰色神经网络模型[30]	
		总人口拟合	预测残差	总人口拟合	预测残差	总人口拟合	预测残差	总人口拟合	预测残差
2005	130756	130724.08	-31.92	130777	21	125680	-5076	130801	45
2006	131448	131439.47	-8.53	131461	13	126120	-5328	131547	99
2007	132129	132106.04	-22.96	132104	-25	126530	-5599	132222	93
2008	132802	132728.05	-73.95	132709	-93	126910	-5892	132663	-139

注：1. 资料来源：中华人民共和国国家统计局官方网站；

2. 表2-5中其它模型的数据分别来自参考文献[45、12、30]。

最后，利用半参数自回归模型对中国2010-2013年人口进行预测(见表2-6)。

表2-6 半参数回归模型对2010-2013年中国人口进行预测

Tab.2-6 Predictions of Chinese population from 2010 to 2013 are done using the semi-parametric regression model

年份	2010	2011	2012	2013
预测人口(万人)	134110.53	134524.1	134971.32	135372.10

2.4 结论

本章基于时间序列分析、半参数回归和非参数的多项式样条估计理论，建立中国人口预测的线性自回归模型和半参数自回归模型，采用两种模型分别对中国人口进行拟合和预测，并且将半参数模型与 Logistic、Leslie 和灰色神经网络模型对中国人口预测的结果做了相比，结果显示半参数自回归模型能够给出所拟合数据的显式表达式，计算量小, 运行时间较快, 并且预测精度也有所提高。

第3章 中国人口预测的具有外生变量的半参数回归模型

考虑到人口数据主要受到前期人口基数的影响，GDP 总量对人口数据趋势作局部调整，本章将人口数据的滞后变量作为半参数模型的线性主部变量，GDP 总量作为非参数部分变量, 建立中国人口预测的具有外生变量的半参数回归模型。基于线性回归理论和多项式样条估计理论得到半参数回归方程，对 1972-2000 年中国人口进行拟合，同时对 2001-2009 年中国人口分三种情况进行预测对比，最后对中国 2010-2015 年的人口进行预测。

3.1 具有外生变量的半参数回归模型的建立

本章中用到的原始人口数据和GDP数据来源于中华人民共和国国家统计局官方网站“http://www.stats.gov.cn/”，用到的样本数据为1952-2009年的中国总人口及GDP数据。

本章中考虑的半参数回归模型的一般形式为：

$$Y_t = \alpha^T X_t + g(Z_t) + \varepsilon_t, \quad t = 1,2,\cdots,n, \tag{3-1}$$

其中Y_t为被解释变量，α是未知参数向量，$X_t = (X_{t1},\cdots,X_{tp})^T = (Y_{t-1},\cdots,Y_{t-p})^T$为解释性变量，线性主部$\alpha^T X_t$把握被解释变量的大势走向；$g(\cdot)$为未知非参数光滑函数，对被解释变量作局部调整；随机误差序列$\{\varepsilon_t\}$独立同分布且满足：

$$E(\varepsilon_t) = 0, \quad \mathrm{Var}(\varepsilon_t) = \sigma^2 < \infty,$$

且ε_t与$Y_s(s<t)$相互独立。

对半参数回归模型(3−1)的线性主部参数向量α和非参数函数$g(\cdot)$的估计，采用第二章中讲到的多项式样条估计方法，这里不再作介绍。

3.1.1 数据的平稳化处理

记$\{Y_t\}$为中国总人口序列，$\{Z_t\}$为中国GDP序列，根据半参数回归模型的要求，两个序列必须是平稳的。由软件Matlab7.0对中国原始人口及GDP数据作时序图，如图3-1所示。

从图3-1可以看到两个序列均是不平稳的。根据数据特点，对1952-2009年58个原始人口及GDP数据先取其自然对数，再进行二次差分。令

$$P_t = \nabla^2 \ln(Y_t) - \overline{\nabla^2 \ln(Y_t)},\quad G_t = \nabla^2 \ln(Z_t) - \overline{\nabla^2 \ln(Z_t)}$$

其中$\overline{\nabla^2 \ln(Y_t)}$、$\overline{\nabla^2 \ln(Z_t)}$分别为$\{\nabla^2 \ln(Y_t)\}$和$\{\nabla^2 \ln(Z_t)\}$的均值，$\nabla$为差分符号，则$\{P_t\}$、$\{G_t\}$为零均值序列，如图3-2所示。

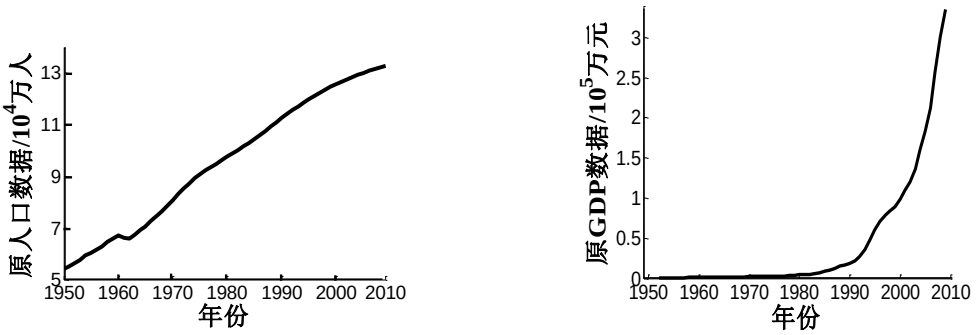


图3-1 1952-2009年原人口及原GDP数据时序图

Fig.3-1 Timing diagram of the original population and GDP of China from 1952 to 2009

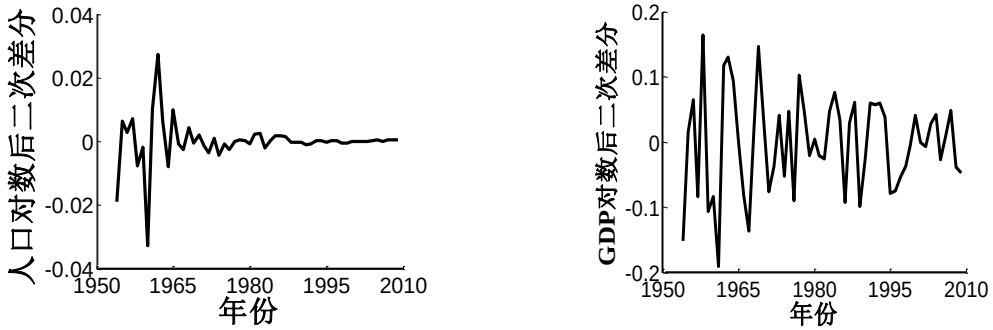


图3-2 人口与GDP变换数据时序图

Fig.3-2 Differential timing diagram after log transformation of population and GDP

3.1.2 数据的平稳性检验

从图3-2可直观地判断序列$\{P_t\}$和$\{G_t\}$都是平稳的。为了定量说明序列的平稳性，对序列$\{P_t\}$进行游程检验。游程总数$r=26$，序列长度$N=56$，“＋” 和“—” 出现的次数分别为$N_1=29$，$N_2=27$，

$$E(r)=2N_1N_2/N+1=28.9643,$$

$$D(r)=2N_1N_2(2N_1N_2-N)/\left[N^2(N-1)\right]=13.7098,$$

$$Z=(r-E(r))/\sqrt{D(r)}=(26-28.9643)/\sqrt{13.7098}=-0.2162,$$

在显著性水平 $\alpha=0.05$ 下，$|Z|<1.96$，因此判定序列 $\{P_t\}$ 是平稳的。

同样对 $\{G_t\}$ 做游程检验，在显著性水平 $\alpha=0.05$ 下，$|Z|=0.0026<1.96$，所以序列 $\{G_t\}$ 也是平稳的。

3.1.3 模型中线性部分的定阶

做 $\{P_t\}$ 和 $\{G_t\}$ 的关系图，见图3-3。

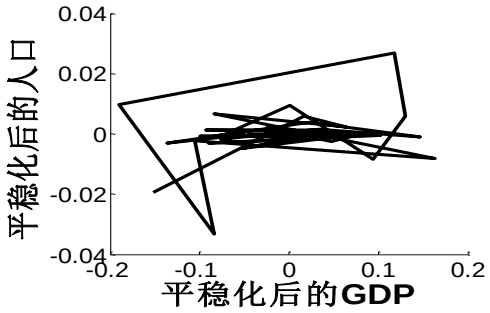


图3-3 $\{P_t\}$ 和 $\{G_t\}$ 的关系图

Fig.3-3 Diagram of $\{P_t\}$ and $\{G_t\}$

从图3-3可以看到，$\{P_t\}$ 和 $\{G_t\}$ 之间并没有明显的线性关系，因此我们将GDP做为半参数回归模型的非参数函数部分，而人口序列做为线性主部变量。

本章依然基于线性自回归 $\mathrm{AR}(p)$ 模型滞后阶数的选取方法来确定半参数模型(3-1)中线性主部的阶数。

先用AIC，BIC准则[50]初步确定线性主部的阶数，其中

$$\mathrm{AIC}(p)=\ln(\hat{\sigma}_a^2)+2p/n,\quad \mathrm{BIC}(p)=\ln(\hat{\sigma}_a^2)+p\cdot\ln(p)/n,$$

这里 p 为滞后阶数，n 为样本容量，$\hat{\sigma}_a^2$ 为拟合残差方差。由计算结果知当滞后阶数为18阶时，AIC和BIC值均达到最小，见图3-3和图3-4。

再用F检验法[50]确定线性部分的阶数。首先对 $\{P_t\}$ 分别拟合 $\mathrm{AR}(17)$ 和 $\mathrm{AR}(18)$ 模型，两种模型的残差平方和 Q_1 和 Q_0 分别为 6.79×10^{-6} 和 5.01×10^{-6}，则

$$F=\frac{(Q_1-Q_0)/S}{Q_0/(n-p)}=\frac{(6.79\times10^{-6}-5.01\times10^{-6})/1}{5.01\times10^{-6}/(52-18)}=12.10,$$

其中 S 为舍弃因子的个数，n 为样本容量，p 为回归因子个数。给定显著性水平 $\alpha=0.05$，查 F 分布表得 $F_\alpha(1,34)=4.13$，$F>F_\alpha$，说明 $\mathrm{AR}(17)$ 和 $\mathrm{AR}(18)$ 有显著的差异，模型阶数有上升的可能。再拟合 $\mathrm{AR}(19)$ 模型，其残差平方和为 4.85×10^{-6}，与 $\mathrm{AR}(19)$ 比较有

$$F=\frac{\left(5.01\times10^{-6}-4.85\times10^{-6}\right)/1}{4.85\times10^{-6}/(52-19)}=1.06\text{。}$$

同理查表得 $F_\alpha(1,33)=4.14$， $F<F_\alpha$，故 $\mathrm{AR}(18)$ 与 $\mathrm{AR}(19)$ 没有显著差异，因此在显著水平 $\alpha=0.05$ 上，选择滞后18阶是合适的。

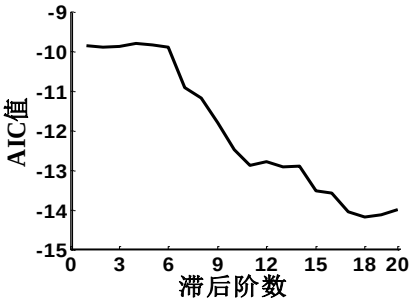


图3-3 AIC

Fig.3-3 Figure of AIC

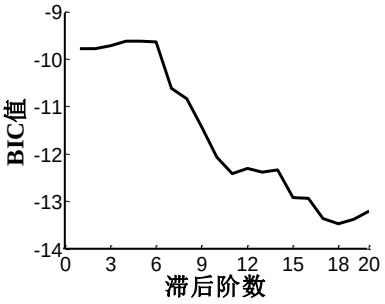


图3-4 BIC

Fig.3-4 Figure of BIC

3.1.4 显著性变量的选取及方程的建立

根据3.2.1节中平稳化处理后的1952-2005年人口数据，建立线性自回归 $\mathrm{AR}(18)$ 模型，其参数估计及其检验结果见表3-1。

表3-1 $\{P_t\}$ 建模各变量系数的显著性检验

Tab.3-1 Significance test of the variable coefficients of $\{P_t\}$

参数	估计值	标准差	t-统计量	显著性概率
$C(1)$	0.7985	0.1878	4.2518	0.0006
$C(2)$	-0.7538	0.2148	-3.5097	0.0029
$C(3)$	0.5973	0.1333	4.4793	0.0004
$C(4)$	-0.0949	0.1218	-0.7793	0.4472
$C(5)$	0.1195	0.1114	1.0733	0.2991
$C(6)$	-0.0575	0.1036	-0.5550	0.5866
$C(7)$	0.1484	0.1050	1.4133	0.1767
$C(8)$	-0.0586	0.0916	-0.6401	0.5312
$C(9)$	0.1647	0.0888	1.8551	0.0821
$C(10)$	-0.0783	0.0586	-1.3348	0.2006
$C(11)$	0.0211	0.0515	0.4097	0.6874

续表 3-1

Continued 3-1

$C(12)$	-0.0787	0.0411	-1.9163	0.0734
$C(13)$	-0.0292	0.0301	-0.9693	0.3468
$C(14)$	0.1252	0.0232	5.4030	0.0001
$C(15)$	-0.1165	0.0321	-3.6253	0.0023
$C(16)$	0.2245	0.0337	6.6688	0.0000
$C(17)$	-0.1131	0.0360	-3.1444	0.0063
$C(18)$	0.1070	0.0328	3.2645	0.0049

注： $C(i)$（$i=1,\cdots,7$）为$\{P_t\}$中相应变量的系数。

从表3-1易知，在$\alpha=0.05$的显著性水平下，只有P_{t-1}，P_{t-2}，P_{t-3}，P_{t-14}，P_{t-15}，P_{t-16}，P_{t-17}，P_{t-18}的系数是显著的。

基于线性回归模型选取的显著性变量做为半参数回归模型(3-1)的线性主部变量，G_t做为非参数部分变量，对模型中线性主部变量的参数和非参数函数进行三次样条估计，得到半参数回归方程:

$$\begin{aligned} P_t &= 0.63P_{t-1} - 0.36P_{t-2} + 0.51P_{t-3} + 0.11P_{t-14} - 0.08P_{t-15} \\ &\quad +0.17P_{t-16} - 0.07P_{t-17} + 0.05P_{t-18} + \hat{g}(G_t) \end{aligned} \quad 。 \tag{3-2}$$

3.2 不同模型对中国人口进行拟合及预测

3.2.1 对中国 1972-2000 年人口进行拟合

为了与文献[49]中建立的非参数模型和文献[45]中利用核方法建立的半参数模型做对比，基于第3.1节建模方法，以1952-2000年中国人口及GDP数据对中国人口建立半参数回归模型，得到新的回归方程：

$$\begin{aligned} P_t &= 0.63P_{t-1} - 0.37P_{t-2} + 0.51P_{t-3} + 0.11P_{t-14} - 0.08P_{t-15} \\ &\quad +0.17P_{t-16} - 0.07P_{t-17} + 0.04P_{t-18} + \hat{g}(G_t) \end{aligned} 。 \tag{3-3}$$

利用半参数回归方程(3-2)和(3-3)分别对中国1972-2000年人口进行拟合，并与以外生变量GDP为非参数函数的回归模型和基于核估计以GDP为线性主部的半参数模型进行对比，结果详见表3-2。

从表3-2可以看到，半参数回归模型对1972-2000年中国人口拟合的相对误差的绝对值明显小于非参数模型。同时与文献[34]中的以GDP为线性主部的半参数

回归模型的拟合情况相比，本章中基于多项式样条估计的半参数回归模型对人口拟合的相对误差的绝对值也较小。

半参数回归方程(3-2)和(3-3)对中国人口拟合的相对误差相差无几，但是比较二者的均方误差，可以看到半参数回归方程(3-2)的均方误差要略小于回归方程(3-3)，这是由于半参数回归方程(3-2)建模包含的样本信息要比方程(3-3)丰富一些的缘故。

表 3-2　三种模型对 1972-2000 年中国人口拟合结果对比分析

Tab.3-2 Fitting results of Chinese population from 1972 to 2000 are done using three different models

年份	人口(万人)	半参数回归方程（3-2）		半参数回归方程（3-3）		非参数模型[49]		半参数模型/核估计[34]	
		总人口拟合(万人)	相对误差(%)	总人口拟合(万人)	相对误差(%)	总人口拟合(万人)	相对误差(%)	总人口拟合(万人)	相对误差(%)
1972	87177	87131.84	-0.05	87135.84	-0.05	85991.00	-1.36	86711.37	-0.53
1973	89211	89200.61	-0.01	89198.88	-0.01	88145.80	-1.19	88248.60	-1.08
1974	90859	90900.08	0.05	90897.72	0.04	88836.80	-2.23	89737.70	-1.23
1975	92420	92427.86	0.01	92428.43	0.01	90791.80	-1.76	91269.10	-1.25
1976	93717	93758.16	0.04	93758.34	0.04	90303.60	-3.64	92712.83	-1.07
1977	94974	94972.13	0.00	94977.38	0.00	92550.30	-2.55	94251.31	-0.76
1978	96259	96319.78	0.06	96320.76	0.06	95694.00	-0.59	95826.52	-0.45
1979	97542	97581.45	0.04	97581.15	0.04	98215.80	0.69	97383.40	-0.16
1980	98705	98666.49	-0.04	98667.50	-0.04	100545.80	1.86	98940.43	0.24
1981	100072	100058.24	-0.01	100066.02	-0.01	101886.70	1.81	100452.28	0.38
1982	101654	101575.97	-0.08	101582.66	-0.07	103237.00	1.56	101969.39	0.31
1983	103008	103172.72	0.16	103165.31	0.15	104683.00	1.63	103509.83	0.49
1984	104357	104339.74	-0.02	104330.70	-0.03	106177.90	1.74	105124.40	0.74
1985	105851	105796.00	-0.05	105797.25	-0.05	106837.10	0.93	106770.17	0.87
1986	107507	107346.96	-0.15	107345.86	-0.15	107042.80	-0.43	108304.11	0.74
1987	109300	109203.05	-0.09	109206.49	-0.09	107620.70	-1.54	109861.59	0.51
1988	111026	111059.71	0.03	111050.40	0.02	109866.10	-1.04	111480.52	0.41
1989	112704	112820.44	0.10	112826.55	0.11	112006.30	-0.62	112976.24	0.24

续表 3-2

Continued 3-2

1990	114333	114330.71	0.00	114338.52	0.00	113861.90	-0.41	114420.08	0.08
1991	115823	115917.76	0.08	115909.88	0.08	116849.00	0.89	115920.99	0.08
1992	117171	117109.88	-0.05	117102.93	-0.06	118959.40	1.53	117474.66	0.26
1993	118517	118419.36	-0.08	118411.04	-0.09	117382.30	-0.96	119080.22	0.48
1994	119850	119794.67	-0.05	119794.75	-0.05	119249.60	-0.50	120719.16	0.73
1995	121121	121116.66	0.00	121110.72	-0.01	122547.80	1.18	122247.38	0.93
1996	122389	122321.21	-0.06	122314.27	-0.06	121819.60	-0.47	123662.75	1.04
1997	123626	123666.98	0.03	123663.57	0.03	122871.40	-0.61	124984.01	1.10
1998	124761	124863.21	0.08	124866.12	0.08	124503.00	-0.21	126227.47	1.18
1999	125786	125834.98	0.04	125846.75	0.05	126261.50	0.38	127439.75	1.31
2000	126743	126770.64	0.02	126768.48	0.02	127901.60	0.91	128678.51	1.53
均方误差：		4598.96		4987.75					

注：1. 资料来源：中华人民共和国国家统计局官方网站；

2. 表3-2中相对误差＝$\left(\hat{Y}_i - Y_i\right)\Big/Y_i \times 100$，$i = 1, \cdots, n$。

3.2.2 三种模型对中国人口预测的结果比较

利用半参数回归方程(3-3)对中国2001-2005年人口进行预测，并与文献[34]和[49]的结果进行对比，详见表3-3。

表3-3 三种模型对中国人口预测的对比结果

Tab.3-3 Predictions of Chinese population are done using three different models

年份	人口（万人）	半参数回归方程(3-3)		半参数模型/核估计[34]		非参数模型[49]		变系数模型[57]	
		总人口预测（万人）	相对误差（%）	总人口预测（万人）	相对误差（%）	总人口预测（万人）	相对误差（%）	总人口预测（万人）	相对误差（%）
2001	127627	127653.85	0.02	127667.25	0.03	126434	-0.93	126613.1	-0.79
2002	128453	128491.76	0.03	128469.37	0.01	128982	0.41	127281.5	-0.91
2003	129227	129277.43	0.04	129286.62	0.05	129176	-0.04	128313.8	-0.71

续表 3-3

Continued 3-3

2004	129988	130005.36	0.01	130023.05	0.03		129183.5	-0.62
2005	130756	130674.69	-0.06	130896.25	0.11		129870.3	-0.68
均方误差：		48.33		72.58		754.03	966.06	

注：1. 资料来源：中华人民共和国国家统计局官方网站；

2. 表3-3中相对误差＝$\left(\hat{Y}_i - Y_i\right)\Big/Y_i \times 100$，$i=1,\cdots,n$。

从表3-3可以看到，对中国2001-2005年人口的预测，半参数回归模型的预测的相对误差绝对值均明显小于非参数模型。与文献[57]建立的变系数模型节点取2时对中国人口进行预测的结果相比，本文建立的半参数回归模型对中国人口预测的相对误差，均明显较小。另外从对人口预测的总体来看，与文献[34]建立的半参数模型相比，基于样条估计的半参数回归模型的预测残差和相对误差的绝对值也较小。同时与其它几个模型的均方误差相比，本章建立的半参数回归模型的均方误差也最小。

3.2.3 最优模型的选取

将半参数回归方程(3-2)与(3-3)分别对 2006-2009 年中国人口进行预测对比分析，结果详见表 3-4。

表3-4说明，两个半参数回归方程对中国人口预测的残差和相对误差的绝对值均较小，然而回归方程(3-2)预测的均方误差要小于方程(3-3)预测的均方误差。此外，由表3-2易知，半参数回归方程(3-2)对人口进行拟合的均方误差也小于回归方程(3-3)。综上可知，半参数回归方程(3-2)对中国人口拟合和预测的精度均较高，因此本文把它做为较为理想的中国人口预测模型。

最后，基于半参数回归方程(3-2)和一步预测法，对2010-2015年中国人口进行预测，具体结果见表3-5。

3.3 结论

考虑到 GDP 对中国人口的影响，但是二者并非是简单的线性关系，本文基于半参数回归和多项式样条估计理论，将人口数据的滞后变量作为半参数模型的线性主部变量，GDP 总量作为非参数部分变量, 建立中国人口的具有外生变量的

半参数回归模型。与以 GDP 做为外生变量的非参数模型、以 GDP 做为线性部分的半参数模型和变系数模型相比，本文建立的以 GDP 做为非参数部分的半参数回归模型，对中国人口的拟合及预测精度较高，并且采用多项式样条估计，能够给出所拟合数据的显式表达式，计算量小、运行时间较快。此外，本文基于建立的半参数模型还预测了 2010-2015 年中国人口总量的变化趋势，这对于中国第六次人口普查分析和“十二五”期间社会经济发展目标的规划、调整和决策具有一定的借鉴意义。

表3-4 基于两个半参数回归方程对2006-2009年中国人口预测的对比结果

Tab.3-4 Predictions of Chinese population from 2006 to 2009 are obtained based on two different semi-parametric autoregression models

年份	人口(万人)	半参数回归方程(3-2)			半参数回归方程(3-3)		
		总人口预测(万人)	预测残差(万人)	相对误差(%)	总人口预测(万人）	预测残差(万人)	相对误差(%)
2006	131448	131485.5	37.5	0.03	131485.6	37.6	0.03
2007	132129	132161.31	32.31	0.02	132161.58	32.58	0.02
2008	132802	132793.73	-8.27	-0.01	132794.3	-7.7	-0.01
2009	133474	133417.82	-56.18	-0.04	133390.59	-83.41	-0.06
均方误差：		1418.75			2372.67		

注：1. 资料来源：中华人民共和国国家统计局官方网站；

2. 表3-4中预测残差= $\hat{Y}_i - Y_i$，相对误差= $\left(\hat{Y}_i - Y_i\right)\Big/Y_i \times 100$，$i = 1, \cdots, n$。

表3-5 基于半参数回归方程(3-2)对2010-2015年中国人口的预测结果

Tab.3-5 Predictions of Chinese population from 2010 to 2015 are done using the semi-parametric regression equation (3-2)

年份	2010	2011	2012	2013	2014	2015
预测人口(万人)	134131.88	134756.81	135354.18	135920.56	136439.98	136944.60

第4章 基于 Bootstrap 方法的中国人口预测的半参数模型

本章提出建立基于 Bootstrap 方法的中国人口预测的的半参数回归模型，根据线性回归理论选取半参数回归模型中线性主部的显著性变量，基于多项式样条理论，采用 Bootstrap 方法对半参数回归模型中线性主部参数和非参数函数进行点估计，得到中国人口预测的半参数回归方程，对 1972-2005 年和 2006-2009 年中国人口分别进行拟合和预测，最后对 2010-2013 年中国人口进行预测。

4.1 Bootstrap 方法概述

Bootstrap方法基于重抽样思想，在具体应用中只根据给定的观测信息，不需要事先假设总体的分布，也不用增加新的样本信息，只是对样本重构并不断计算相关的估计值，进而对总体的分布特性做出推断。在实际应用中为了减小推断误差，同时为了保证 $\alpha*(B+1)$ 为整数，进行Bootstrap重抽样的次数 B 一般选为99、199、499、999等，其中 α 是显著性水平。Bootstrap方法也为无法获得大量样本而可能导致推断失误这一问题提供了有效的思路和解决途径。

Bootstrap方法是由美国斯坦福大学Efron教授(1979)在总结前人的基础上首次提出，并且对Bootstrap方法做了系统地介绍。Bootstrap方法做为一种新兴的统计方法受到了广大统计学家的关注，自提出后的近30多年间，Bootstrap方法不论是在理论上，还是在实际应用中比如在金融、经济学、医学等等领域都得到了广泛的发展(参看文献[58-67])。目前常用的Bootstrap方法有：残差Bootstrap方法，Wild Bootstrap方法，Sieve Bootstrap方法等。

本文中用到的 Bootstrap方法是残差Bootstrap方法。

残差Bootstrap方法[67]，是对模型的残差进行有放回地随机抽样，其中要求误差项与回归元之间是相互独立的，同时误差项满足独立同分布。

4.2 中国人口的半参数模型的建立

本章中用到的1949-2009年原中国人口数据来源于中华人民共和国国家统计局官方网站“http://www.stats.gov.cn/”。

本章中考虑的半参数回归模型的一般形式仍为：

$$Y_t = \alpha^T X_t + g(Z_t) + \varepsilon_t \ , \quad t = 1,2,\cdots,n \ , \tag{4-1}$$

其中 Y_t 为被解释变量，α 是未知参数向量，$X_t = (X_{t1},\cdots,X_{tp})^T = (Y_{t-1},\cdots,Y_{t-p})^T$ 为

解释性变量，线性主部 $\alpha^T X_t$ 把握被解释变量的大势走向；$Z_t=(Z_{t1},\cdots,Z_{tq})^T=(Y_{t-p-1},\cdots,Y_{t-p-q})^T$，$g(\cdot)$为未知非参数光滑函数，对被解释变量作局部调整；随机误差序列$\{\varepsilon_t\}$独立同分布且满足：

$$E(\varepsilon_t)=0,\ \mathrm{Var}(\varepsilon_t)=\sigma^2<\infty,$$

且ε_t与$Y_s(s<t)$相互独立。

4.2.1 数据的平稳化处理

基于第二章中提到方法对中国1949-2009年数据进行平稳化处理，基于线性回归选取的显著性变量，分别选取滞后2阶、5阶和7阶做为非参数部分，其余二变量做为线性部分。此时，半参数回归模型分别为：

$$W_t=\alpha_1 W_{t-5}+\alpha_2 W_{t-7}+g(W_{t-2})+\varepsilon_t, \quad (4\text{-}2)$$

$$W_t=\alpha_1 W_{t-2}+\alpha_2 W_{t-7}+g(W_{t-5})+\varepsilon_t, \quad (4\text{-}3)$$

$$W_t=\alpha_1 W_{t-2}+\alpha_2 W_{t-5}+g(W_{t-7})+\varepsilon_t。 \quad (4\text{-}3)$$

4.2.2 模型参数的确定

本文中用多项式样条估计和残差Bootstrap方法对半参数回归模型(4-2)、(4-3)及(4-4)中线性主部的参数和非参数函数进行估计。具体步骤如下(以模型(4-2)为例说明)：

① 用多项式样条估计方法对模型(4-2)的线性主部的参数和非参数函数进行样条估计，得到参数和非参数函数的样条估计量$\hat{\alpha}$、$\hat{g}(z)$，并得到误差项的估计量$\hat{\varepsilon}_t$；

② 对①中的误差项估计量$\hat{\varepsilon}_t$进行有放回地抽取，得到第一组残差Bootstrap样本$\hat{\varepsilon}_t^{(1)}$；

③ 将①中得到的$\hat{\alpha}$、$\hat{g}(z)$和②中得到的$\hat{\varepsilon}_t^{(1)}$带回半参数回归模型(4-2)，得到$W_t$相应的Bootstrap估计值$\hat{W}_t^{(1)}$；

④ 由③中得到的$\hat{W}_t^{(1)}$带到模型(4-2)中，由步骤①得到第一组参数和非参数函数的Bootstrap估计量$\hat{\alpha}^{(1)}$、$\hat{g}^{(1)}(z)$；

⑤ 重复步骤②$B=999$次，得B组残差Bootstrap样本$\hat{\varepsilon}_t^{(i)}$，进而可得到$B$组参数估计量$\hat{W}_t^{(i)}$，又可得到$B$组$\hat{\alpha}^{(i)}$、$\hat{g}^{(i)}(z)$（$i=2,\ 3,\cdots,\ \mathrm{B}+1$）；

⑥ 由步骤④和⑤总共得到1000组参数和非参数函数的估计量Bootstrap估计值$\hat{\alpha}^*$、$\hat{g}^*(z)$，对得到的1000组$\hat{\alpha}^*$、$\hat{g}^*(z)$进行从小到大排序，在5%的置信区间内，可得到$\hat{\alpha}^*$、$\hat{g}^*(z)$的区间估计；

⑦ 再取$\hat{\alpha}^*$、$\hat{g}^*(z)$区间估计值的平均值，得到α、g的Bootstrap方法的点估

计 $\hat{\alpha}_B^*$、$\hat{g}_B^*$。

按照以上确定参数的方法，得到半参数回归模型(4-2)线性主部的参数和非参数函数的 Bootstrap 点估计，此时方程的显式表达式为：

$$W_t = -0.28W_{t-5} - 0.26W_{t-7} + \hat{g}(W_{t-2}), \quad (4\text{-}5)$$

同样的方法可得到半参数回归模型(4-3)和(4-4)的显式表达式，分别为：

$$W_t = -0.45W_{t-2} - 0.30W_{t-7} + \hat{g}(W_{t-5}), \quad (4\text{-}6)$$

$$W_t = -0.45W_{t-2} - 0.29W_{t-5} + \hat{g}(W_{t-7}). \quad (4\text{-}7)$$

4.3 对中国人口进行拟合及预测

4.3.1 三个半参数回归方程对中国人口进行拟合

由得到的三个半参数回归方程(4-5)、(4-6)和(4-7)分别对中国 1958-2004 年人口进行拟合，并得到拟合值的均方误差，详见表 4-1。

表4-1 三个半参数回归方程分别对中国1958-2000年人口拟合

Tab.4-1 Fitting results of Chinese population from 1958 to 2000 are done using three semi-parametric equations

年份	人口（万人）	半参数回归方程（4-5）		半参数回归方程（4-6）		半参数回归方程（4-7）	
		总人口拟合（万人）	误差 (%)	总人口拟合（万人）	误差 (%)	总人口拟合（万人）	误差 (%)
1958	65994	66355.95	0.55	66180.70	0.28	66155.73	0.25
1959	67207	67566.35	0.53	67491.84	0.42	67477.15	0.40
1960	66207	67393.16	1.79	68273.81	3.12	68204.09	3.02
1961	65859	65425.53	-0.66	65591.52	-0.41	65500.93	-0.54
1962	67295	67295.49	0.00	66210.29	-1.61	66184.92	-1.65
1963	69172	68613.02	-0.81	68528.21	-0.93	68524.41	-0.94
1964	70499	70520.14	0.03	70105.04	-0.56	70067.63	-0.61
1965	72538	72497.28	-0.06	72497.76	-0.06	72471.97	-0.09
1966	74542	73651.96	-1.19	74730.61	0.25	74723.15	0.24
1967	76368	76547.90	0.24	76427.49	0.08	76436.88	0.09

续表 4-1

Continued 4-1

1968	78534	77889.54	-0.82	77874.42	-0.84	77806.10	-0.93
1969	80671	80173.31	-0.62	80353.45	-0.39	80463.64	-0.26
1970	82992	82578.95	-0.50	82315.32	-0.82	82258.27	-0.88
1971	85229	85551.40	0.38	85610.90	0.45	85599.27	0.43
1972	87177	87411.43	0.27	87224.66	0.05	87168.20	-0.01
1973	89211	89035.64	-0.20	89120.92	-0.10	89126.56	-0.09
1974	90859	91071.89	0.23	91501.88	0.71	91521.87	0.73
1975	92420	92439.24	0.02	92307.94	-0.12	92289.74	-0.14
1976	93717	93640.27	-0.08	94216.97	0.53	94237.00	0.55
1977	94974	95024.16	0.05	95079.06	0.11	95086.68	0.12
1978	96259	96060.76	-0.21	96350.90	0.10	96367.89	0.11
1979	97542	97750.19	0.21	97766.21	0.23	97786.81	0.25
1980	98705	98873.00	0.17	98785.42	0.08	98796.44	0.09
1981	100072	100077.76	0.01	100048.33	-0.02	100065.65	-0.01
1982	101654	101436.49	-0.21	101503.16	-0.15	101521.40	-0.13
1983	103008	103392.82	0.37	103192.44	0.18	103209.21	0.20
1984	104357	104453.10	0.09	104239.42	-0.11	104253.23	-0.10
1985	105851	105584.78	-0.25	105798.98	-0.05	105814.35	-0.03
1986	107507	107326.15	-0.17	107251.27	-0.24	107259.22	-0.23
1987	109300	109227.87	-0.07	109041.08	-0.24	109052.82	-0.23
1988	111026	111179.61	0.14	110988.00	-0.03	110991.19	-0.03
1989	112704	112773.41	0.06	112586.87	-0.10	112585.98	-0.10
1990	114333	114410.49	0.07	114417.41	0.07	114434.95	0.09
1991	115823	115932.24	0.09	115903.21	0.07	115914.07	0.08
1992	117171	117243.86	0.06	117206.53	0.03	117209.15	0.03
1993	118517	118435.76	-0.07	118503.75	-0.01	118510.99	-0.01
1994	119850	119784.19	-0.05	119849.74	0.00	119858.84	0.01
1995	121121	121247.95	0.10	121171.57	0.04	121191.18	0.06
1996	122389	122465.49	0.06	122401.17	0.01	122421.78	0.03
1997	123626	123704.89	0.06	123686.38	0.05	123707.32	0.07

续表 4-1

Continued 4-1

1998	124761	124930.82	0.14	124850.63	0.07	124871.49	0.09
1999	125786	125950.95	0.13	125895.92	0.09	125917.26	0.10
2000	126743	126796.24	0.04	126806.87	0.05	126824.20	0.06
2001	127627	127669.23	0.03	127683.00	0.04	127699.68	0.06
2002	128453	128525.19	0.06	128495.72	0.03	128514.99	0.05
2003	129227	129290.13	0.05	129264.26	0.03	129282.32	0.04
2004	129988	130029.17	0.03	129988.18	0.00	130007.65	0.02
均方误差：		314.82		423.30		424.26	

注：1. 资料来源：中华人民共和国国家统计局官方网站；

2. 表 4-1 中相对误差＝$\left(\hat{Y}_i - Y_i\right)\Big/Y_i \times 100$， $i=1,\cdots,n$。

从表 4-1 可以看到三个半参数回归方程对人口拟合的相对误差的绝对值均较小，相差也不是很大。但是半参数回归方程(4-5)的均方误差最小，而半参数回归方程(4-6)的均方误差略小于方程(4-7)的均方误差。

4.3.2 三个半参数回归方程对中国人口的预测

将本章中建立的三个半参数回归方程(4-5)、(4-6)和(4-7)分别对中国 2005-2009 年人口进行预测，同时得到预测的均方误差，详见表 4-2。

表 4-2 三个半参数回归方程对中国 2005-2009 年人口的预测

Tab.4-2 Predictions of Chinese population from 2005 to 2009 using three semi-parametric regression equations

年份	人口（万人）	半参数回归方程（4-5）		半参数回归方程（4-6）		半参数回归方程（4-7）	
		总人口拟合（万人）	相对误差(%)	总人口拟合（万人）	相对误差(%)	总人口拟合（万人）	相对误差(%)
2005	130756	130761.60	0.00	130742.60	-0.01	130762.43	0.00
2006	131448	131558.58	0.08	131475.08	0.02	131535.01	0.07
2007	132129	132366.62	0.18	132163.08	0.03	132271.86	0.11

续表 4-2

Continued 4-2

2008	132802	133184.39	0.29	132811.95	0.01	132978.48	0.13
2009	133474	133996.31	0.39	133414.77	-0.04	133651.56	0.13
均方误差：		101.87		10.99		43.93	

注：1. 资料来源：中华人民共和国国家统计局官方网站；

2. 表 4-2 中相对误差＝$\left(\hat{Y}_i - Y_i\right)\Big/Y_i \times 100$，$i=1,\cdots,n$。

从表 4-2 可知，在对 2005-2009 年中国人口进行的预测中，半参数回归方程(4-6)和的相对误差的绝对从值从整体上看小于半参数回归方程(4-5)和(4-7)，同时预测的均方误差也明显小于其余两个半参数回归方程。从表 4-1 可以看到三个半参数回归方程对中国人口拟合的效果都较好，但是综合考虑，认为半参数回归方程(4-6)为较理想的中国人口预测模型。

4.3.3 基于选取的半参数回归模型对中国人口进行拟合

为了得到更加精确的预测结果，对半参数回归模型(4-3)的线性部分的参数向量和非参数函数重新进行 Bootstrap 估计，同 4.2.2 中确定参数的方法，这里让 B=4999，即可得到 5000 组α、g的 Bootstrap 估计值$\hat{\alpha}^*$、$\hat{g}^*(z)$。

通过 Matlab 编程可实现，在α、g分别取置信区间内的第 250，1000，2000，…，4750 时的估计值$\hat{\alpha}^*$、$\hat{g}^*(z)$时，半参数回归方程对 1958-2004 年中国人口拟合及对 2005-2009 年中国人口进行预测的均方误差，详细结果见表 4-3。从均方误差表中选取均方误差较小时对应的参数估计值，建立具体的半参数回归方程。

表 4-3　取区间内的不同值时对中国人口拟合及预测的均方误差

Tab.4-3 MSE of fitting and prediction of Chinese population using different variables in range

参数区间	对 1958-2004 年中国人口拟合的均方误差	对 2005-2009 年中国人口预测的均方误差
取值 250	424.31	16.38
取值 1000	423.69	15.52
取值 2000	423.24	13.65

续表 4-3

Continued 4-3

取值 2500	423.3	11.91
取值 3000	423.35	11.42
取值 3900	423.61	10.44
取值 4000	423.64	10.40
取值 4100	423.76	10.41
取值 4500	424.48	10.78
取值 4750	425.88	11.95

从表 4-3 中可以看到，当半参数回归模型参数值(4-3)线性主部的参数和非参数函数取区间内 4000 点时的参数值时，对中国人口拟合及预测的均方误差相对较小，因此可得到一个新的半参数回归方程：

$$W_t = -0.4455W_{t-2} - 0.3028W_{t-7} + \hat{g}\left(W_{t-5}\right)。 \tag{4-8}$$

将建立的半参数回归方程(4-8)和方程(4-6)分别对中国 2005-2009 年人口进行预测，见表 4-4。

表 4-4 对 2005-2009 年中国人口进行预测

Tab.4-4 Predictions of Chinese population from 2005 to 2009

年份	人口（万人）	半参数回归方程(4-6)			半参数回归方程(4-8)		
		总人口拟合（万人）	预测残差（万人）	误差(%)	总人口拟合（万人）	预测残差（万人）	误差(%)
2005	130756	130744.21	-11.79	-0.01	130744.02	-11.98	-0.01
2006	131448	131479.95	31.95	0.02	131478.94	30.94	0.02
2007	132129	132172.13	43.13	0.03	132170.35	41.35	0.03
2008	132802	132826.11	24.11	0.02	132823.41	21.41	0.02
2009	133474	133435.24	-38.76	-0.03	133431.38	-42.62	-0.03
均方误差：		10.42			10.40		

注：1. 资料来源：中华人民共和国国家统计局官方网站；

2. 表 4-4 中预测残差= $\hat{Y}_i - Y_i$，相对误差= $\left(\hat{Y}_i - Y_i\right)/Y_i \times 100$，$i = 1, \cdots, n$。

从表 4-4 可以看到半参数回归模型(4-3)对中国人口预测效果还是比较好的，预测误差的绝对值没有超过 50 万人，相对误差的绝对值也在 0.03%内，半参数回归方程(4-8)的预测的均方误差只略小于方程(4-6)。

4.3.4 与第二章建立的半参数自回归方程的对比

将本文建立的半参数回归方程(4-8)与第二章中建立的半参数回归方程(2-5)分别对 1958-2004 年中国人口进行拟合，并对 2005-2009 年中国人口进行预测，详见表 4-5。

表 4-5 半参数回归方程(4-8)与（2-5）对中国人口拟合与预测对比

Tab.4-5 Predictions of Chinese population from 2010 to 2015 are done using the semi-parametric regression equation (4-8) and (2-5)

年份	人口（万人）	半参数回归方程(2-5)			半参数回归方程(4-8)		
		总人口拟合(万人)	拟合误差(万人)	误差(%)	总人口拟合(万人)	拟合误差(万人)	误差(%)
1958	65994	66168.00	174.00	0.26	66198.78	204.78	0.31
1959	67207	67487.25	280.25	0.42	67490.66	283.66	0.42
1960	66207	68267.93	2060.93	3.11	68279.26	2072.26	3.13
1961	65859	65587.78	-271.22	-0.41	65591.52	-267.48	-0.41
1962	67295	66203.82	-1091.18	-1.62	66214.45	-1080.55	-1.61
1963	69172	68525.84	-646.16	-0.93	68529.82	-642.18	-0.93
1964	70499	70102.64	-396.36	-0.56	70108.54	-390.46	-0.55
1965	72538	72493.90	-44.10	-0.06	72492.83	-45.17	-0.06
1966	74542	74720.20	178.20	0.24	74742.06	200.06	0.27
1967	76368	76376.03	8.03	0.01	76528.68	160.68	0.21
1968	78534	77867.62	-666.38	-0.85	77881.05	-652.95	-0.83
1969	80671	80350.70	-320.30	-0.40	80356.79	-314.21	-0.39
1970	82992	82304.22	-687.78	-0.83	82329.28	-662.72	-0.80
1971	85229	85607.53	378.53	0.44	85611.12	382.12	0.45
1972	87177	87221.71	44.71	0.05	87227.16	50.16	0.06
1973	89211	89114.94	-96.06	-0.11	89124.37	-86.63	-0.10
1974	90859	91498.26	639.26	0.70	91502.50	643.50	0.71
1975	92420	92303.24	-116.76	-0.13	92310.51	-109.49	-0.12

续表 4-5

Continued 4-5

1976	93717	94213.43	496.43	0.53	94217.75	500.75	0.53
1977	94974	95075.90	101.90	0.11	95080.57	106.57	0.11
1978	96259	96346.48	87.48	0.09	96352.08	93.08	0.10
1979	97542	97762.97	220.97	0.23	97767.11	225.11	0.23
1980	98705	98781.60	76.60	0.08	98786.93	81.93	0.08
1981	100072	100044.85	-27.15	-0.03	100049.21	-22.79	-0.02
1982	101654	101498.98	-155.02	-0.15	101504.38	-149.62	-0.15
1983	103008	103187.89	179.89	0.17	103193.93	185.93	0.18
1984	104357	104234.98	-122.02	-0.12	104241.18	-115.82	-0.11
1985	105851	105794.91	-56.09	-0.05	105800.27	-50.73	-0.05
1986	107507	107245.53	-261.47	-0.24	107253.90	-253.10	-0.24
1987	109300	109035.23	-264.77	-0.24	109043.77	-256.23	-0.23
1988	111026	110984.07	-41.93	-0.04	110990.05	-35.95	-0.03
1989	112704	112582.10	-121.90	-0.11	112589.03	-114.97	-0.10
1990	114333	114411.74	78.74	0.07	114419.34	86.34	0.08
1991	115823	115897.41	74.41	0.06	115905.57	82.57	0.07
1992	117171	117200.87	29.87	0.03	117208.96	37.96	0.03
1993	118517	118499.00	-18.00	-0.02	118505.53	-11.47	-0.01
1994	119850	119844.85	-5.15	0.00	119851.52	1.52	0.00
1995	121121	121166.61	45.61	0.04	121173.23	52.23	0.04
1996	122389	122396.48	7.48	0.01	122402.82	13.82	0.01
1997	123626	123681.62	55.62	0.04	123687.99	61.99	0.05
1998	124761	124845.27	84.27	0.07	124852.37	91.37	0.07
1999	125786	125890.56	104.56	0.08	125897.62	111.62	0.09
2000	126743	126801.68	58.68	0.05	126808.64	65.64	0.05
2001	127627	127677.51	50.51	0.04	127684.87	57.87	0.05
2002	128453	128490.32	37.32	0.03	128497.52	44.52	0.03
2003	129227	129259.09	32.09	0.02	129266.06	39.06	0.03
2004	129988	129983.01	-4.99	0.00	129989.97	1.97	0.00
均方误差：			423.25			423.64	

续表 4-5

Continued 4-5

2005	130756	130737.22	-18.78	-0.01	130744.02	-11.98	-0.01
2006	131448	131458.86	10.86	0.01	131479.38	31.38	0.02
2007	132129	132132.86	3.86	0.00	132171.17	42.17	0.03
2008	132802	132764.48	-37.52	-0.03	132824.69	22.69	0.02
2009	133474	133345.56	-128.44	-0.10	133433.21	-40.79	-0.03
均方误差：			19.78			10.40	

注：1. 资料来源：中华人民共和国国家统计局官方网站；

2. 表 4-5 中拟合误差= $\hat{Y}_i - Y_i$，相对误差= $\left(\hat{Y}_i - Y_i\right)/Y_i \times 100$，$i = 1, \cdots, n$。

从表 4-5 可以看到，半参数回归方程(4-8)与(2-5)对 1958-2004 年中国人口拟合及对 2005-2009 年人口进行预测的相对误差均较小。但是结合多项式样条估计和 Bootstrap 方法建立的半参数回归方程(4-8)，对人口预测的相对误差的绝对值在 0.03%以内，因此本文尝试基于 Bootstrap 方法建立的半参数回归模型也是较理想的人口模型。

最后，基于半参数回归方程(4-8)对 2010-2014 年中国人口进行预测，结果详见表 4-6。

表4-6　基于半参数回归方程(4-8)对2010-2014年中国人口的预测

Tab.4-6 Predictions of Chinese population from 2010 to 2014 are done using the semi-parametric regression equation (4-8)

年份	2010	2011	2012	2013	2014
预测人口（万人）	134076.93	134616.88	135099.1	135544.67	135928.11

4.4 结论

本章基于时间序列分析、半参数回归和多项式样条估计理论及 Bootstrap 方法，建立了中国人口预测的半参数回归模型。利用多项式样条估计与 Bootstrap 方法对半参数模型中线性主部的参数和非参数函数进行估计，根据均方误差选取最优的半参数回归模型，建立了半参数回归方程，并对中国人口进行了拟合及预测，并与第二章建立的半参数自回归模型对中国人口进行拟合及预测的结果做对比，结果显示基于 Bootstrap 方法建立的半参数回归模型也是比较理想的模型。

第5章 结论

5.1 主要研究成果

本文主要做了如下的工作：

1．中国人口预测的模型有很多种，但在现实生活中，各变量间的关系未必是线性或可线性化的非线性关系，非参数模型则是假定变量之间的关系未知，对回归函数进行估计，因而能更好的拟合样本数据。而半参数回归模型整合了非参数模型和线性模型的优点，具有更强的解释能力，但是半参数方法的应用主要集中于核类方法，对全局光滑的样条方法在人口预测中还有待做进一步研究，因此本文基于时间序列分析、半参数线性回归和非参数的多项式样条估计理论，建立中国人口预测的线性自回归模型和半参数自回归模型，对中国人口进行拟合和预测。结果显示，半参数回归模型优于传统的线性模型。

2．考虑到 GDP 对中国人口的影响，但是二者并非是简单的线性关系，本文基于半参数回归和多项式样条估计理论，将人口数据的滞后变量作为半参数模型的线性主部变量，GDP 总量作为非参数部分变量，建立中国人口预测的具有外生变量的半参数回归模型。与以 GDP 做为外生变量的非参数模型、以 GDP 做为线性部分的半参数模型和变系数模型相比，本文建立的以 GDP 做为非参数部分的半参数回归模型，对中国人口的拟合及预测精度较高，并且采用多项式样条估计，能够给出所拟合数据的显式表达式，计算量小、运行时间较快。最后，基于建立的半参数模型还预测了 2010-2015 年中国人口总量的变化趋势，这对于中国第六次人口普查分析和“十二五”期间社会经济发展目标的规划、调整和决策具有一定的借鉴意义。

3．现实应用中的传统模型往往存在设定误差，而 Bootstrap 方法基于重抽样思想，一定程度上解决了误差项分布未知而导致的推断失误，因此本文尝试基于 Bootstrap 方法建立中国人口预测的的半参数回归模型，根据线性模型选取半参数回归模型中线性主部的显著性变量，基于多项式样条方法和 Bootstrap 方法给出线性主部参数和非参数函数的点估计，得到中国人口预测的半参数回归方程对中国人口进行拟合和预测。最后，与只采用多项式样条估计建立的半参数自回归方程对中国人口拟合及预测结果做了对比，结果显示基于多项式样条估计与 Bootstrap 方法建立的中国人口预测的的半参数回归模型也是较为理想的人口预

测模型。

5.2 尚待研究的问题

在建立的中国人口预测的半参数回归模型中，

1．模型中线性主部滞后阶数的选取是基于线性回归选取显著性变量的方法而确定的，方法虽然简单而且因此建立的半参数回归方程对中国人口预测的相对误差较小，下一步的工作就是尝试其它方法选取线性主部的显著性变量，以提高对中国人口预测的精度；

2．对具有外生变量的半参数模型，模型中线性主部的参数向量和非参数函数进行的多项式样条估计，在具体应用中，结点个数的选取和结点位置的放置还有待做进一步的研究；

3．对于中国人口预测的半参数回归模型，利用 Bootstrap 方法对模型的参数和非参数函数进行点估计，得到了半参数回归方程，为进一步为相关部门提供更精确的信息，下一步的工作就是，对中国人口进行区间预测；

4．近年来对非参数函数的估计，出现了新的理论方法，如小波分析法，下一步的工作就是尝试将小波估计法应用到半参数模型中对中国人口进行预测。

参考文献

[1] 宋健. 人口控制[J]. 自动化学报，1989，15(5)：385-391.

[2] 宋健，于景元. 人口控制论[J]. 软科学研究，1989，(1)：1-7.

[3] 龚跃，党宏. 人口预测的模型与方法[J]. 长春光学精密机械学院学报，1991，14(3)：83-92.

[4] 王伟华，宋宇婷，王福胜等. 一类人口发展方程解的存在性和收敛性的研究[J]. 哈尔滨师范大学自然科学学报，2006，22(2)：15-17.

[5] 何朗，赵韫，管坤等. 人口发展的参数预测模型[J]. 武汉理工大学学报(信息与管理工程版), 2008, 30(3)：494-497.

[6] 邹春松. “人口自然增长率”计算方法亟需改进[J]. 统计与决策，1991，(4)：44-44.

[7] 茆长宝，程琳. 两种人口预测模型的精确度比较—以人口年龄移算法和灰色预测模型为例[J]. 南京人口管理干部学院学报，2009，(1)：29-32.

[8] 谢建文，张元标，王志伟等. 基于宋健人口模型的老年化预测[J]. 2008，36(12)：5227-5229.

[9] 王勇. Logistic 人口模型的求解问题[J]. 哈尔滨商业大学学报(自然科学版)，2006，22(5)：58-59.

[10] 李百岁，同力嘎. 内蒙古人口城市化 Logistic 模型及其应用[J]. 干旱区资源与环境，2007，21(2)：32-36.

[11] 冯守平. 中国人口增长预测模型[J]. 安徽科技学院学报，2008，22(6)：73-76.

[12] 陈文权，赵兹，李得胜. Leslie 修正模型在人口预测中的应用[J]. 世界科技研究与发展，2008，30(2)：219-224.

[13] Lee R. D., Tuljapurkar, S. Stochastic Popultation Forecasts for the US Meetings of the Population[J]. Association of Amercia in Washington D.C., 1991.

[14] Lee. R.D. ,Carter L. Modeling and Forecasting US Moratlity[C]. Annual Meetings of the Population Association of America (Toronto)，1990.

[15] Pollard.H.. Mathematical Models for the Grouth of Human Population[M].University Press, Cambridge，1975:112-134.

[16] 李南，Tuljapurkar S. 基于时间一区域序列的随机死亡率预测及对中国数据的应用[J]. 人口研究，1995，(4)：58-63.

[17] 李南，申卯兴. 基于随机模型的中国生育率预测[J]. 预测，1996，15(6)：33-36.

[18] 李南，胡华清．基于随机方法的中国人口预测与规划[J]．系统工程理论方法应用，1998，7(1)：37-41.

[19] 付莹．回归分析在人口预测中的应用[J]．辽宁高职学报，2000，2(1)：56-58.

[20] 李旭东．贵州喀斯特高原人口分布的自然环境因素Ⅱ．多元回归分析与地带性研究[J]．干旱区研究，2007，24(2)：280-286.

[21] 陈爱平，安和平．中国人口时间序列预测模型的探讨[J]．人口与经济，2004，(6)：63-67.

[22] 薛臻．我国人口增长预测数学模型[J]．河南科技学院学报(自然科学版)，2008，36(1)：123-127.

[23] 周诗国．我国人口的灰色预测模型研究及其应用[J]．数理医药学杂志，2005，18(4)：307-309.

[24] 蒿建华．灰色模型在人口预测中的应用[J]．西安文理学院学报（社会科学版），2008，11(3)：42-44.

[25] 张静，王兴华．利用神经网络预测人口数量[J]．襄樊学院学报，2001，22(5)：73- 76.

[26] 赵方，王晓雷，蔡森等．神经网络在人口自然率预测中的应用[J]．中原工学院学报，2006，10，17(5)：13-15.

[27] 胡秋灵，姚文辉，李红霞．基于 Box-Jenkins 建模法的人口自然增长率预测模型[J]．统计与决策，2007，(3)：4-6.

[28] 张静．基于 Box-Jenkins 建模法对中国人口增长率预测及结果分析[J]．黑龙江科技信息，2009，(14)：24-24.

[29] 叶小青．人口问题的随机微分方程模型[J]．统计与决策，2008，(6)：161-162.

[30] 李国成，吴涛，徐沈．灰色人工神经网络人口总量预测模型及应用[J]．计算机工程与应用，2009，45(16)：215-218.

[31] 王丽敏，莫君慧．基于灰色 BP 神经网络的中国人口预测模型[J]．新乡学院学报(自然科学版)，2009，26(2)：3-5.

[32] 刘金月，许少华．基于自适应小波过程神经元网络的人口预测研究[J]．长江大学学报(理工卷)，2008，5(4)：236-237,240.

[33] 巩永丽，张德生，武新乾．人口增长率的非参数自回归预测模型[J]．数理统计与管理，2007，26(5)：759-764.

[34] 姜爱平，张德生, 武新乾等．预测我国人口总量的具有外生变量的半参数自回归模[J]．河南科技大学学报(自然科学版), 2007，28(1)：97-100.

[35] Engk -lerF. RiceJ. Semiparametric estimates of the relation between weather and electricity sales [J]. Journal of the American Statistical Association ,1986,81:310-320.

[36] Paul，Speckman. Kemel smoothing in partial linear models[J]. Journal of the Royal Statistic Society:Series B, 1988,50:413-436.

[37] RiceJ. Convergence rates for partially splined models[J]. Statistics and Probability Letters,1986,4:203-208.

[38] 王成勇. 国内半参数回归模型研究进展[J]. 襄樊学院学报，2008，29(2)：8-13.

[39] 潘建敏. 污染数据半参数回归模型的估计方法[J]. 工程数学学报，1997，14(3).

[40] 薛留根. 随机删失下半参数回归模型的估计理论[J]. 数学年刊（A 辑），1999，20A(6)：745-754.

[41] 叶阿忠，吴相波，黄志刚. 半参数计量经济联立模型的变窗宽估计理论[J]. 管理科学学报，2009，12(2)：60-66.

[42] 黄四民，梁华. 用半参数部分线性模型分析居民消费结构[J]. 数量经济技术经济研究，1994，(10)：33-38.

[43] 王一兵. 半参数回归模型在商品房价格指数中的应用研究[J]. 统计研究，2005，(4)：25-29.

[44] 冯春山，蒋馥，吴家春. 应用半参数方法计算市场风险的受险价值[J]. 系统工程理论方法应用，2005，14(4)：379-381.

[45] 王学保，蔡果兰. Logistic 模型的参数估计及人口预测[J]. 北京工商大学学报(自然科学版), 2009，27 (6)：75-78.

[46] 彭志搠. AR(p)模型在中国总人口预测中的应用[J]. 河北工程大学学报(自然科学版), 2007, 24(4)：109-112.

[47] 叶阿忠. 非参数计量经济学[M]. 南开大学出版社，2003.

[48] 张慧芳, 张德生, 武新乾等. 我国人口总量的非参数预测模型[J]. 延边大学学报(自然科学版), 2007，33(2)：90-93.

[49]武新乾, 田铮, 韩四儿. 具有外生变量部分线性自回归模型的样条估计[J]. 数学年刊, 2007, 28A(3)： 377-386.

[50] 王振龙. 时间序列分析[M]. 中国统计出版社，2000.

[51] 李建民，王金营. 中国生育率下降经济后果的计量分析[J]. 中国人口科学，2000，(1)：8-16.

[52] 王谦, 郭震威. 人口增长对经济增长的影响分析—与胡鞍钢博士商榷[J]. 人口研究, 2001, 25(1)：20-23.

[53] 冯利华. 人口增长的综合预测分析[J]. 系统工程, 2000，18(1)：71-75.

[54] 赵进文. 中国人口总量与 GDP 总量关系模型的研究[J]. 中国人口科学，2003，(3)：25-31.

[55] 陈理洪. 基于 Copula 函数的中国人口总量与 GDP 总量相关性研究[J]. 南阳师范学院学报，2008，7(9)：27-28.

[56] 姜爱平. 我国人口时间序列的变系数预测模型[J]. 数理统计与管理，2008，27(5)：755-759.

[57] Efron,B. Bootstrap methods:Another look at the jackknife[J]. Annals of Statistics,1979, 7(1):1-26.

[58] Efron,B. An Introduction to the Bootstrap[M]. New York, Chapman and Hall, 1993.

[59] Yong P. Jackknife and bootstrap resampling methods in statistical analysis to correct for bias[J]. Statistical Science,1996,11:189-228.

[60] Bühlmann. P. Sieve bootstrap for time series[J]. Bernoulli 3,1997(1):123-148.

[61] Efron,B. The bootstrap and modern statistics[J]. Journal of the American Statistical Association, 2000,(1):1293-1296.

[62] Bühlmann. P. Bootstraps for time series[J]. Statistical Science,2002,17(1):52-72.

[63] Chang Y. J.Y.Park. A sieve bootstrap for the test of a unit root[J]. Journal of Time Series Analysis,2003,24(4):379-400.

[64] Politis,D.N. The impact of bootstrap methods on time series analysis[J]. Statistical Science,2003,18(2):219-230.

[65] James G. MacKinnon. Bootstrap Methods in Econometrics[J]. Queen's Economics Department Working Paper No.1028,2006,(2):1-28.

[66] Jinhong You,Gemai Chen. Wild bootstrap estimation in partially linear models with beteroscedasticity[J]. Statistics & Probability Letters, 2006,76(4):340-348.

[67] 龙志和，欧变玲. Bootstrap 方法在经济计量领域的应用[J]. 工业技术经济，2008，27(7)：132-135.

致　　谢

本论文完成之际，我深深地感谢我的导师杨万才教授和武新乾副教授。在我攻读硕士学位期间，杨老师和武老师的言传身教不仅使我在学业上受益匪浅，而且在树立正确的人生观、价值观上也给了我很大启发。他们渊博的知识、敏锐的思维、和蔼的态度、严谨的学风、一丝不苟的工作作风以及对科学的敬业精神都将使我受益终身。在此，对他们再次表达我衷心的感谢！

在完成论文的过程中，河南科技大学理学院文洪江、郭文博以及我的师弟、朋友们也都给予了大力支持，在此对他们表示衷心的感谢！

在我攻读硕士学位期间，数学与统计学院的领导和许多老师，为我的学习和研究工作提供了许多帮助和便利条件。研究生处的出色管理，使我按学校要求完成了硕士生的学习和研究工作。在此也向他们表示我衷心的感谢！

最后，诚心地感谢我的家人对我的全力支持和照顾，使我能全身心地投入到学习和研究工作当中，按时完成学业！

攻读硕士学位期间的研究成果

本人在攻读硕士学位期间发表的学术论文

[1] 韩玉涛，杨万才，武新乾. 我国人口预测的半参数自回归模型[J]. 河南科技大学学报(自然科学版)2011，32(1)：100-104.

[2] Wu Xinqian, Yang Wancai, Han Yutao. Spline estimation for nonparametric regression with linear process errors. Data Processing and Quantitative Economy Modeling-Conference Proceedings of the 3rd International Institute of Statistics & Management Engineering Symposium, Weihai, China 2010, 156-160.

[3] 贾长生，武新乾，韩玉涛. 洛阳市人口发展趋势的 ARIMA 模型[J]. 洛阳师范学院学报，2011，30(5)：15-18.

[4] 韩玉涛，杨万才，武新乾. 中国人口预测的具有外生变量的半参数回归模型[J]. 数理统计与管理(已录用).