中南财经政法大学
青年学术文库

半参数回归模型及其应用

Semiparametric Regression Model and Its Application

金　林◎著

中国出版集团
世界图书出版公司

中南财经政法大学
青年学术文库

半参数回归模型及其应用

Semiparametric Regression Model and Its Application

金　林◎著

中国出版集团
世界图书出版公司
广州·上海·西安·北京

图书在版编目（CIP）数据

半参数回归模型及其应用 / 金林著 .— 广州：世界图书出版广东有限公司 , 2016.10

ISBN 978-7-5192-1936-9

Ⅰ . ①半… Ⅱ . ①金… Ⅲ . ①回归分析－半参数模型－研究 Ⅳ . ① O212.1

中国版本图书馆 CIP 数据核字（2016）第 255581 号

半参数回归模型及其应用

责任编辑　孔令钢

出版发行　世界图书出版广东有限公司

地　　址　广州市新港西路大江冲 25 号

http:// www.gdst.com.cn

印　　刷　北京振兴源印务有限公司

规　　格　710mm×1000mm　1/16

印　　张　16.75

字　　数　287 千

版　　次　2016 年 10 月第 1 版　2016 年 10 月第 1 次印刷

ISBN　978-7-5192-1936-9/O・0052

定　　价　52.00 元

《中南财经政法大学青年学术文库》

编辑委员会

摘　要

在统计建模过程中经常碰到非线性关系，但线性函数形式似乎成了约定俗成的形式，通常很少验证线性函数假设是否成立。在统计建模过程中忽略变量之间的非线性会导致严重的问题。常见的处理非线性关系的方法有数据转换方法和神经网络，支持向量机，投影寻踪和基于树的方法等高计算强度方法。然而，这些方法要么在实际使用过程中有很大的局限性，要么得到的结果无法解释。使用非参数和半参数回归方法来处理非线性关系则在一定程度可以避免这些问题。非参数回归模型和半参数回归模型不对回归模型的函数形式进行假设，而是从数据本身来估计合适的函数形式。半参数回归模型及其推广模型对经典线性模型进行了多方面的推广，使其发展成为一个理论框架。该框架能够涵括许多重要的统计模型，因此具有重要的理论意义。同时，半参数回归模型及其推广模型在统计建模中也发挥着重要作用，具有很大的应用前景。

本书主要讨论半参数回归模型及其推广模型，并利用它们研究了一个重要的官方统计问题：中国全国及地区 GDP 数据准确性评估。导论部分交代本书的研究背景和意义，给出了半参数回归模型及其推广模型的文献综述，并简述了本书的研究思路、主要内容、可能的创新和未来研究方向。本书接下来的部分具体分为六章：第 2 章讨论非参数光滑方法；第 3 章详细讨论了半参数回归模型和广义可加模型；第 4 章讨论半参数混合模型的相关问题；第 5 章利用半参数回归模型及其统计诊断理论对中国全国 GDP 数据准确性进行了检测；第 6 章利用半参数混合模型对中国地区 GDP 数据准确性进行了检测；第 7 章是结语。各章具体内容为：

第 2 章主要讨论了非参数回归模型。首先通过介绍两种简单的光滑方法，揭示了非参数回归背后的基本思想。接下来主要介绍了两种重要的光滑方法：局部多项式回归和光滑样条。在局部多项式回归中，给出了局部多项式光滑器，并讨论了局部多项

式回归光滑程度的选择、置信带构建和假设检验等问题。在光滑样条中，首先给出了几种常见的样条：简单样条，二次和三次样条，B 样条等。选择合适的节点个数和位置，利用这些样条就可以拟合非线性。但这些样条容易发生过拟合现象，而光滑样条在一定程度可以避免过拟合。为此继续讨论了光滑样条。此外还对光滑样条的置信带构建和假设检验等统计推断问题进行了讨论。最后，讨论了非参数回归中的自动光滑方法: 广义交叉验证和利用似然方法估计光滑参数。

第 3 章把第 2 章讨论的非参数回归模型在可加性条件下推广到多元情况，得到了可加模型和半参数模型。对于可加模型和半参数回归模型，详细讨论了其模型设定、估计方法和统计推断。为把半参数回归模型推广到广义可加模型，比较详细地讨论了广义线性模型，包括其三个组成部分：系统部分、随机部分和联接函数。最后，讨论了广义可加模型的模型设定、估计方法、置信带构建和假设检验。

第 4 章把半参数回归模型推广到混合模型，也即半参数混合模型。首先讨论线性混合模型，给出线性混合模型的模型设定和参数估计。接下来讨论如何把半参数方法加入到混合模型中去，得到半参数混合模型的模型设定。最后讨论广义线性模型、半参数方法和混合模型三者结合得到的广义可加混合模型 (GAMM)，包括其模型设定和估计方法。

第 5 章首先综述了中国 GDP 数据准确性评估方法，并确定采取统计诊断方法作为评估方法，接下来简要介绍了 GDP 数据的经济背景——经济增长理论，并给出了生产函数的一般形式。在收集并整理了 1953—2010 年中国 GDP、资本存量、就业人员和教育支出经费数据的基础上，同时估计了线性回归模型和半参数回归模型，通过模型比较发现，半参数回归模型优于线性回归模型。因此选定半参数回归模型为最优模型。为对 GDP 数据准确性进行评估，讨论了半参数回归模型统计诊断理论后，对估计的半参数回归模型计算了相关诊断统计量的值，并通过图示和严格的假设检验得到了模型的异常点。最后，对这些异常点所在的经济背景进行了分析，得到了经济意义上的中国 GDP 数据异常点。

第 6 章首先给出了中国地区 GDP 总和与全国 GDP 差距明显的事实，并简单分析了该现象发生的原因。解决地区 GDP 虚高现象关键性措施之一就是对各地区的 GDP 准确性进行评估。因此，这一章在综述中国地区 GDP 相关研究的基础上，以新经济增长理论为基础，收集并整理了 1990—2010 年中国地区一级的 GDP、资本存

量、就业人员和平均受教育年限面板数据。对于这个面板数据集，我们分别建立了线性回归模型，带有方差结构的线性回归模型，带有哑变量的线性回归模型，线性混合模型和半参数混合模型，通过统计意义上的模型比较和选择，并结合实际经济意义，得到最优模型为线性混合模型。在讨论了线性混合模型的统计诊断理论后，对估计的线性混合模型进行了统计诊断，得到了统计模型的异常点。

第 7 章首先总结了中国 GDP 数据准确性检测的结论，并对统计诊断方法用于 GDP 准确性检测进行了评价。最后在实证基础上总结了半参数回归模型及其推广模型在实际应用过程中应注意的问题。

本书完成的主要工作和得到的主要结论为：

1. 本书对包括半参数回归模型在内的众多主流和前沿的回归方法从实际应用角度进行了比较全面的讨论，包括模型的基本思想、发展历程和模型之间的内在逻辑联系。此外，还讨论了半参数回归模型和线性混合模型的统计诊断问题。本书还通过实证分析完整体现了复杂情况下统计模型的建模思路和流程，总结了半参数回归模型应用条件和注意事项。

2. 从全国层面 GDP 数据准确性实证分析来看，1958、1959、1961、1991 和 1994 年的 GDP 是建立的半参数模型的异常点。5 个异常点分布在两个时间段：1958—1961 年和 1991—1994 年。考虑到这两个时间段的具体情况，1958—1961 年中国一方面由于各种政治因素的影响，另一方面“三年自然灾害”的发生，导致正常的经济活动遭到破坏，有可能导致真实 GDP 与其它年份相比相差较大，因此异常点 1958、1959、1961 年的 GDP 数据的准确性可能没有问题，较好反映了这几年真实 GDP 情况。1991—1994 年是中国通货膨胀非常严重的时期，但通货膨胀不一定对国家实际总产出造成非常严重的影响。因此 1991、1994 年的 GDP 数据的异常可能是由于统计准确性问题所造成的。

3. 从地区层面 GDP 数据准确性实证分析来看，一共检测出 27 个异常点。从异常点地区分布来看，具有异常点的地区有山西、内蒙古、辽宁、吉林、上海、安徽、山东、河南、广西、海南、四川和云南，其中河南和辽宁的情况最为严重，分别有 7 年和 5 年 GDP 数据为异常值，吉林有 3 年数据 GDP 为异常值，山东、山西和四川各有 2 年 GDP 数据为异常值，其它省份各有 1 年 GDP 数据为异常值。从异常值的分布时间来看，出现异常值的年份主要集中在三个时间段：1990—1991 年，2004—2006

年和 2009—2010 年。在 27 个异常点中，只有 5 个异常点不是在这三个时间段之内。

基于上述研究，本书可能的创新主要体现在对统计模型的梳理归纳，实证研究框架拓展，实证研究时间范围拓展，实证研究精度提高和全文研究架构较新等方面。

Abstract

In the process of statistical modeling we often encounter nonlinear relationships. However, linear models are often the conventional choice to approach such analyses. we seldom verify if linearity assumptions are satisfied. Ignoring the nonlinear relationships among variables in statistical modeling may lead to serious problems. There are two types of methods addressing nonlinearity in statistical modeling: data transformation and the computationally intensive methods such as neural networks, support vector machines, projection pursuit, and tree-based methods. As useful as these methods can be, they either face major limitations in application or produce results that are difficult to interpret.

The aforementioned limitations can be overcome by employing nonparametric and semiparametric regression methods。 Nonparametric and semiparametric regression models do not make assumptions regarding specific functional forms. Instead, they estimate functional forms based on the actual characteristics of data. Semiparametric regression models and its generalization have improved classical linear models in several ways and developed into a new theoretical framework. The framework consists of a number of important statistical models and therefore contributes significantly to the development of the relevant theories. In addition, the semiparametric regression models and its generalization play an important role in statistical modeling and have great prospects in application.

This book focused on the semiparametric regression models and its generalized models, and employed them to investigate a problem in the official statistics: the degree of accuracy in China' s national and regional GDP data. In the introduction section , the research background and significance were discussed. Literature review of the semiparametric regression models and its generalized models were also presented. Research design and possible break-through of this book were outlined. This book can be divided into six chapters: The first chapter discusses nonparamet-

ric smoothing methods; Semiparametric regression models and generalized additive models are discussed in detail in the second chapter; The third chapter discusses the semiparametric mixed model; The fourth chapter assesses the accuracy of China's national GDP data with semiparametric regression model and its statistical diagnostic theory. The fifth chapter assesses the accuracy of China's regional GDP data with semiparametric mixed model and its statistical diagnostic theory. The sixth chapter is the conclusion. The main contents of each chapter are:

Chapter 2 discussed the non-parametric regression models. First, the basic ideas underlying nonparametric regression were illustrated. Second, two important smoothing methods are introduced: local polynomial regression and smoothing spline. Local polynomial smoothers are introduced in local polynomial regression. The choice of smoothness parameter, confidence belt construction and hypothesis testing are also discussed. Several common splines are introduced, including simple splines, quadratic and cubic splines and B-splines. By selecting the appropriate node number and location, these splines can fit nonlinearity. However, these splines are prone to overfit data, which can be prevented by smoothing spline. So smoothing spline is discussed, the confidence belt construction and hypothesis testing of smoothing spline are also discussed. Finally, two automated smoothing techniques are discussed: generalized cross-validation and likelihood method to estimate the smoothing parameter.

Chapter 3 discusses additive models and semiparametric regression models. Nonparametric regression models can be extended to multiple situations with the additive assumption: the extended models are additive models and semiparametric regression models. First, model specification, estimation methods and statistical inference of additive models and semiparametric regression models are discussed. Then, semiparametric regression models can be extended into generalized additive models. Generalized linear models (GLM) are the basis of generalized additive models. Therefore, generalized linear models are outlined, including three components of GLM: systematic component, stochastic component and link function. Finally, the model specification, estimation methods, confidence belt construction and hypothesis testing of generalized additive models are discussed.

Chapter 4 extends semiparametric models to mixed model framework, Semiparametric mixed models are discussed. First, model specification, estimation methods and computation of linear mixed models are outlined. Then, the relationship of nonparametric methods and mixed models are reviewed, model specification of semiparametric mixed models are presented. Finally, generalized linear models, semiparametric models and mixed models are combined to one: generalized additive mixed models (GAMM). Model specification and estimation methods of GAMM are also

discussed.

Chapter 5 assesses the accuracy of China's national GDP data. Firstly, We review assessment methods of accuracy of China's GDP data and take the statistical diagnostic method as our assessment method. The economic background of the GDP data - economic growth theories are briefly introduced, the general form of the production function is also proposed. Secondly, we collect and organize 1953-2010 China's GDP, capital stock, employment and education spending funds data. Linear regression model and semiparametric model are estimated with the data. Through the model comparison, the semiparametric regression model appears to fit GDP data better than the linear regression model, therefore was chosen as the optimal model. Thirdly, to evaluate the accuracy of the GDP data, statistical diagnostic theory of semiparametric regression models are discussed. we calculate the value of diagnostic statistics of the estimated semiparametric regression model. Based on these diagnostic statistics, figure observation and rigorous hypothesis testing are used to detect outliers. Finally, we analyze the economic background of these outliers and decide China's GDP data real outliers.

Chapter 6 first presents the fact that the gap between the sum of the regional GDP and national GDP is rather big and analyze the causes of this discrepancy. One of the key methods in curbing the artificial inflation of regional GDP is effective assessment of the accuracy of the region's GDP data. Then, we review the research on China region's GDP data, collect and organize 1990-2010 China region's GDP , capital stock, employment and education spending funds panel data. Thirdly, with this panel data, we establish linear regression model, linear regression model with variance structure, linear regression model with dummy variables, linear mixed model and semiparametric mixed model. With model comparison and selection, and combined with the economic meaning, the optimal model is linear mixed model. Finally, statistical diagnostic theory of linear mixed models is discussed. We calculate the value of diagnostic statistics of the estimated linear mixed model. Based on these diagnostic statistics, statistical outliers are identified with figure observation and rigorous hypothesis testing.

Chapter 7 first summarizes the conclusions of the China's GDP data accuracy and reviews statistical diagnostic method for the assessment of the accuracy of GDP data. Finally, based on empirical study we summarize the conditions and potential concerns when using semiparametric regression models and its generalized models in practical application.

The main work and major findings of this book are:

1. In this book, the semiparametric regression and many other mainstream and

frontier regression models are comprehensively discussed from a practical point of view. The discussions include the basic ideas and the course of development of these models, and the internal logical connections between these models. In addition, statistical diagnostics of semiparametric regression models and linear mixed model are also discussed. Through the process of empirical analysis, this book also demonstrates the protocol and process of statistical modeling and summarizes the conditions and potential concerns of the application of semiparametric regression models.

2. Empirical analysis from China's national level shows that GDP data in 1958, 1959, 1961, 1991 and 1994 are the outliers of the estimated semiparametric model. These five outliers are distributed in two time periods: 1958-1961 and 1991-1994. We take into account China's specific circumstances of the two time periods to decide real GDP outliers. During 1958-1961, political impact and Three Years of Natural Disasters resulted in the destruction of normal economic activity, which may lead to real GDP data of 1958-1961 being quite different compared to the other years. Therefore, GDP data of 1958, 1959 and 1961 may have no accuracy problem and reflect the real economic activities in these years. During the period of 1991-1994, China suffered serious inflation, but inflation does not necessarily cause a very serious impact on the country's total output. The outliers of 1991 and 1994 may be due to the accuracy or lack thereof in official statistical work.

3. The empirical analysis of China's district GDP data identified a total of 27 outliers. Geographically, these 27 outliers are distributed among the provinces of Shanxi, Inner Mongolia, Liaoning, Jilin, Shanghai, Anhui, Shandong, Henan, Guangxi, Hainan, Sichuan and Yunnan. Henan and Liaoning had the most serious accuracy problems of GDP data, Henan had 7 outliers and Liaoning had 5 outliers. Jilin province had 3 outliers, Shandong, Shanxi and Sicuan had 2 outliers, the other six provinces had one outlier each. Most of these 27 outliers are occurred within three time periods :1990-1991, 2004-2006 and 2009-2010. Only 5 out of the 27 occurred outside of these three time periods. Based on the above studies, the possible break-through of this book may be reflected by: synthesis of statistical models, expansion of empirical research framework, expansion of empirical research time period, improvement of empirical research accuracy and reproducible research of this book.

符号列表

AIC	赤池信息量准则 (Akaike Information Criterion)
BIC	贝叶斯信息准则 (Bayesian Information Criterion)
BLUP	最佳线性无偏预测 (Best Linear Unbiased Prediction)
CV	交叉验证 (Cross Validation)
DPQL	双重罚拟似然 (估计)(Double Penalized Quasi Likelihood)
GAM	广义可加模型 (Generalized Additive Model)
GAMM	广义可加混合模型 (Generalized Addictive Mixed Model)
GCV	广义交叉验证 (Generalized Cross Validation)
GDP	国内生产总值 (Gross Domestic Product)
GLM	广义线性模型 (Generalized Linear Model)
GLMM	广义线性混合模型 (Generalized Linear Mixed Model)
GLS	广义最小二乘 (估计)(Generalized Least Squares)
IRLS	迭代再加权最小二乘 (估计)(Iteratively Reweighted Least Squares)
loess	局部回归 (local regression)
lowess	局部加权回归 (local weighted regression)
MCMC	马尔科夫链蒙特卡罗 (Markov Chain Monte Carlo)
ML	极大似然 (估计)(Maximum Likelihood)
MPS	物质产品平衡表体系 (Material Product System)
OLS	普通最小二乘 (估计)(Ordinary Least Squares)
REML	限制极大似然 (估计)(Restricted Maximum likelihood)
SNA	国民帐户体系 (System of National Accounts)
WLS	加权最小二乘 (估计)(Weighted Least Squares)

目　录