青年科技创新人才学术文库

纵向数据半参数回归模型的估计理论

ZONGXIANGSHUJU BANCANSHU HUIGUIMOXING DE GUJI LILUN

田 萍 编著

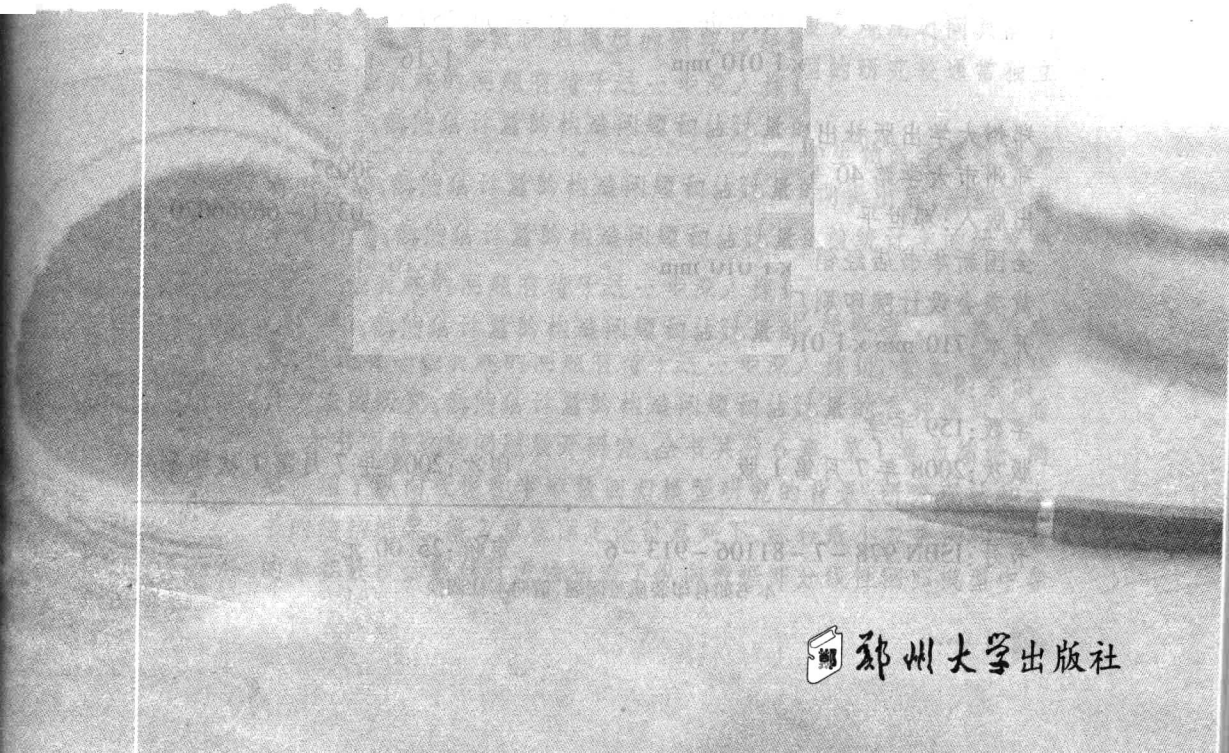
郑州大学出版社

青年科技创新人才学术文库

纵向数据半参数回归模型的估计理论

ZONGXIANGSHUJU BANCANSHU HUIGUIMOXING DE GUJI LILUN

田 萍 编著



郑州大学出版社

内容简介

本书在纵向数据下研究了两类半参数回归模型:纵向数据部分线性回归模型和纵向数据部分线性单指标回归模型中参数分量和非参数分量的估计问题.纵向数据半参数回归模型在计量经济学、生物医学和统计学等领域有着广泛的应用.

图书在版编目(CIP)数据

纵向数据半参数回归模型的估计理论/田萍编著.—郑州:郑州大学出版社,2008.7
ISBN 978-7-81106-913-6

Ⅰ.纵… Ⅱ.田… Ⅲ.回归分析-统计模型-研究 Ⅳ.0212.1

中国版本图书馆 CIP 数据核字(2008)第 092257 号

郑州大学出版社出版发行
郑州市大学路 40 号　　邮政编码:450052
出版人:邓世平　　发行部电话:0371-66966070
全国新华书店经销
黄委会设计院印刷厂印制
开本:710 mm×1 010 mm　　1/16
印张:8
字数:159 千字
版次:2008 年 7 月第 1 版　　印次:2008 年 7 月第 1 次印刷

书号:ISBN 978-7-81106-913-6　　定价:25.00 元

前 言

本书在纵向数据下研究了两类半参数回归模型:纵向数据部分线性回归模型和纵向数据部分线性单指标回归模型中参数分量和非参数分量的估计问题.

纵向数据是指对同一组受试个体在不同时间上的重复观测数据.这类数据与截面数据不同.截面数据指仅仅在某一时间点对同一组个体作一次观测.假设有 n 个观测个体,记 t_{ij} 为第 i 个个体第 j 次观测的时间,x_{ij} 和 y_{ij} 分别是第 i 个个体在时间 t_{ij} 的协变量和响应变量,m_i 为第 i 个个体重复观测的数目,则纵向数据集可记为 $\{(t_{ij},y_{ij},x_{ij}^T),1\leqslant i\leqslant n,1\leqslant j\leqslant m_i\}$.对纵向数据研究的兴趣通常集中在评价时间和协变量对响应变量的效应.在纵向数据中,尽管在不同个体之间的观测是独立的,但由于对同一个体进行重复观测,因此同一个体内的不同观测之间可能是相关的.对纵向数据进行分析必须考虑到这种数据的非独立性.由于重复观测之间具有内相关性,这就使得对纵向数据半参数回归模型的研究较通常独立数据情形下半参数回归模型的研究更为复杂.

纵向数据半参数回归模型在计量经济学和生物医学等领域都有广泛的应用.这类模型自20世纪90年代初期提出后,一些学者采用不同的方法进行了研究,并逐步形成了目前统计学的研究热点之一.

虽然纵向数据半参数回归模型的研究已经取得了可喜的成果,但还有一些有趣的问题有待于进一步深入探讨.譬如,新的估计方法的研究,新的估计量的构造问题和估计量的各种渐近性质等.本书围绕这些问题展开研究,全书共分6章.第1章为绪论,简要介绍了纵向数据和半参数回归模型研究的背景、研究现状和本书的结构框架.第2章在固定设计点列下,综合最小二乘法和一般的非参数权函数估计方法构造了纵向数据部分线性回归模型中参

数分量和非参数分量的半参数最小二乘估计，并研究了估计量的渐近性质. 第3章在固定设计点列下，基于参数分量的广义估计方程和非参数分量的一般权函数方法构造了纵向数据部分线性回归模型的广义半参数最小二乘估计，并研究了估计量的渐近性质. 第4章在固定设计点列下，构造了纵向数据部分线性回归模型中未知参数的广义经验对数似然比函数. 在适当条件下，证明了所提出的比依分布收敛于χ^2分布. 第5章在固定设计点列下，对于纵向数据部分线性单指标回归模型提出未知参数的广义惩罚样条最小二乘估计方法，并在一定条件下讨论了估计的渐近性质. 第6章在随机设计点列下，基于纵向数据组间独立的特点提出了纠偏的经验似然方法. 该方法有优良的性质，它可以避免欠光滑(undersmoothing)来保证参数估计的$\sqrt{n}$-相合性，即在估计非参数$g(\cdot)$和其导数$g'(\cdot)$时对窗宽的选取范围放宽，使得只需使用同一个窗宽利用局部线性光滑方法就可以同时获得$g(\cdot)$和$g'(\cdot)$的估计量.

借本书出版之际，向我的老师北京工业大学应用数理学院的博士生导师薛留根教授表示衷心的感谢，本书的许多成果都是在他的悉心指导下完成的. 感谢张忠占教授、王松桂教授、杨振海教授、李寿梅教授和程维虎教授的启发和指导. 感谢李高荣同学的鼓励与帮助，书中不少成果是大家合作完成的.

本书的出版得到了郑州大学出版社和崔青峰先生的大力支持和帮助，还得到了许昌学院出版基金的资助，对此作者向他们表示诚挚的谢意.

由于作者水平有限，书中错误或不当之处在所难免，恳请专家及广大读者批评指正.

田　萍

2008年1月

目 录

第1章 绪 论

回归模型是数理统计中发展较早，理论丰富而且应用性较强的统计模型. 自 F. Galton 于 1886 年首次提出回归模型以来，回归模型一直受到人们的重视. 在过去的几十年里，回归模型被广泛地应用于工业，农业，气象，地质，经济管理以及医药卫生等各个领域，取得丰富的理论和应用成果. 同时，由于实际应用的需要，为了更加接近现实，更好地解释数据，回归模型也一直处于不断发展进步之中，其模型由初期的参数回归模型发展到非参数回归模型，20 世纪 80 年代以来又兴起了半参数回归模型，其理论研究和应用领域都在不断深入和扩大.

1.1 参数回归模型

参数回归模型假定回归方程 $E(y|x)=m(x;\beta)$ 的函数形式已知，回归模型可表示为

$$y_i=m(x_i;\beta)+e_i,\quad i=1,2,\cdots,n, \tag{1.1}$$

其中 β 为未知的有限维参数，e_i 为均值为 0 的随机误差. 例如大家熟知的线性回归模型和非线性回归模型. 参数回归模型有一个共同的特点：函数形式假定是已知的，未知的仅仅是其中的有限个参数. 由于参数回归模型容易处理，且对其研究已有相当长的历史，因而已形成一套成熟的理论和方法.

参数回归模型通常由经验和历史资料提供了大量的额外信息，因而当模型假定成立时，其推断有较高的精度. 例如熟知的线性回归模型的最小二乘估计，其方差有 $O(n^{-1})$ 的阶. 但响应变量与解释变量之间的函数关系形式不总是已知的，在许多实际应用中，没有证据表明已知函数关系的存在. 因而，参数回归模型在实际应用中往往存在模型的设定误差，当模型假定不成立时，基于模型假定所作的统计推断其表现可能很差，甚至导致错误的结论.

1.2 非参数回归模型

为了减少参数回归的模型偏差,统计学家提出了一个假设更宽松更自由的模型 - 非参数回归模型. 非参数回归模型可表示为

$$y_i = m(x_i) + e_i, \quad i = 1,2,\cdots,n, \tag{1.2}$$

其中 $m(\cdot)$ 为一未知函数,一般假定 $m(\cdot)$ 具有一定的光滑性,e_i 为均值为 0 的随机误差. 非参数回归模型的特点是:回归函数的形式可以任意,对解释变量和响应变量的分布也很少限制,因而有较大的适应性和稳健性. 非参数回归模型是基于数据的模型,它假定响应变量和解释变量之间的关系形式未知,由观测数据本身对整个回归函数进行估计,是较参数回归模型更符合现实的模型.

回归函数 $m(\cdot)$ 的估计可以通过核、局部多项式或样条光滑等非参数回归方法来获得. 尽管非参数回归模型比参数回归模型需要更多的计算,但由于计算机的快速发展和计算能力的增强,近些年来非参数回归方法受到人们的普遍关注.

1.3 维数灾祸

在上述非参数回归模型中,回归函数既可以是一元的,也可以是多元的. 理论上讲,一元非参数回归的估计方法可以直接推广到多元非参数回归,但是有一个很困难的问题即维数问题. 由于非参数回归估计方法本质上讲都是局部估计或局部光滑,要想使 $m(\cdot)$ 在 x 点得到比较充分的估计,就必须使得 x 的领域包含有足够多的数据. 当 x 的维数增大时,局部光滑所需要的数据点个数成指数倍增加. 例如,如果一个局部领域沿着每一个坐标轴包含 10 个数据点,那么在相应的 d 维领域就需要 10^d 个数据点. 而实际上,高维数据具有内部稀疏性,即随着维数 d 的增加,一个局部领域所包含的数据点个数在整个样本中所占的比例越来越小. 为了获取非参数光滑所必需的足够多的数据,我们可以有两种选择,或者增大窗宽,或者增加样本数目. 增大窗宽必然导致估计的偏差增加,而要获取数目非常多的样本在许多情况下是不实际的. 由此可见,当解释变量 x 的维数增加时,多元非参数回归估计的精度下降很快,人们称这种现象为"维数灾祸"(curse of dimensionality).

关于"维数灾祸"方面的详细讨论可参见文献[1~5].

1.4 降维模型

为了克服“维数灾祸”问题，统计学家在寻找既能达到数据降维，同时又能保留非参数光滑优点的方法方面已经作了许多研究，提出了多种降维模型，各种降维模型都对响应变量 y 与解释变量 x 之间的关系作了某些假定. 一般地，降维模型可以分为两大类：基于可加的模型和基于投影的模型. 可加类模型假定 p 个解释变量对响应变量的效应是可加的（也可以存在交互效应）. 这些模型包括交替条件期望模型（alternating conditional expectation）（参见文献[6]），可加回归模型（additive regression model）（参见文献[4,7～13]），可变系数回归模型（varying coefficient regression model）（参见文献[14～17]）和部分线性回归模型（partially linear regression model）（参见文献[18～54]）等；投影类模型假定 p 个解释变量对响应变量的效应可以归结为解释变量的 k 个线性组合 $(x^T\beta_1,\cdots,x^T\beta_k)$ 的效应. 这些模型包括投影追踪回归模型（projection pursuit regression model）（参见文献[55～58]），切片逆回归模型（sliced inverse regression model）（参见文献[59～62]），单指标回归模型（single－index regression model）（参见文献[63～79]），部分线性单指标回归模型（partially linear single－index regression model）（参见文献[80～83]），PHD 分析（principal Hessian direction）（参见文献[84]），SAVE 法（sliced average variance esimation）（参见文献[85]）等.

下面仅对本文将要涉及的部分线性回归模型和（部分线性）单指标回归模型予以简要介绍.

1.4.1 部分线性回归模型

部分线性回归模型假设响应变量 y 依赖于 p 维协变量 $x=(x_1,x_2,\cdots,x_p)^T$ 和一维协变量 t，且 y 与 x 之间呈线性关系，y 与 t 之间呈非线性关系，其模型形式为

$$y_i=x_i^T\beta+g(t_i)+e_i,\quad i=1,2,\cdots,n, \tag{1.3}$$

其中 $(x_i^T,t_i)_{i=1}^n$ 可以是随机设计也可以是固定设计. $\beta=(\beta_1,\cdots,\beta_p)^T$ 是未知参数向量，$g(\cdot)$ 是一元未知函数. e_i 是期望为 0 的随机误差.

自 Engle 等[18]在研究气象条件对电力需求影响这一实际问题时首次提出部分线性回归模型以来，该模型已出现了一系列丰富的研究成果. 由于部分线性回归模型结合了参数回归模型和非参数回归模型，因此，文献中对这种模型的处理方法一般都是融合了参数回归中常用的方法和近些年来发展起来的非参数方法.

Robinson[19]在非参数分量 $g(\cdot)$ 取 Nadaraga－Waston 核估计时，构造了参数

分量β的加权最小二乘估计$\hat{\beta}$和非参数分量$g(\cdot)$的估计$\hat{g}(\cdot)$，在一些正则条件下，研究了估计$\hat{\beta}$的渐近正态性及$\hat{\beta}$和$\hat{g}(\cdot)$的弱收敛速度.

Speckman[20]采用参数化形式Wr逼近非参数分量$g(\cdot)$，其中W为某个给定的$n\times q$的满秩矩阵，r是附加的$q\times 1$的未知参数向量，部分线性模型(1.3)可以用矩阵形式表示为：

$$y = x\beta + Wr + e. \tag{1.4}$$

考虑同时极小

$$\| y - x\beta - Wr \|^2 = \min!, \quad \beta \in R^p, r \in R^q,$$

可得β的估计

$$\hat{\beta} = (x^T(I - P_W)x)^{-1}x^T(I - P_W)y, \tag{1.5}$$

其中$P_W = W(W^TW)^{-1}W^T$为投影阵. 在适当的条件下，Speckman研究了该估计量的渐近行为. 由于在β估计的构造上已经消除了t对x和y的影响，因此β的估计是渐近无偏的.

Green等[21]提出可以用任一光滑矩阵S(不必为投影阵)替代(1.4)式中的W，由此可得β的估计

$$\tilde{\beta} = (x^T(I - S)x)^{-1}x^T(I - S)y, \tag{1.6}$$

由于这种估计是由Green等提出的，因此也称之为G. J. S. 估计.

Engle等[18]，Heckman[22]，Rice[23]，Whaba[24]，Green等[25]和Eubank等[26]使用光滑样条方法定义了β和$g(\cdot)$的惩罚估计量为极小化

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - x_i^T\beta - g(t_i)\}^2 + \lambda\int\{g''(u)\}^2 du$$

的解，其中λ是一个惩罚参数，它起到在拟合程度与光滑程度之间的平衡作用. 这种方法既考虑到估计量同数据的拟合，又顾及到非参数分量估计的光滑性. Heckman[22]在x_i和t_i独立的情况下研究了β的惩罚最小二乘估计的相合性和渐近正态性. 其他作者也分别在不同的条件下对估计的大样本性质进行了研究.

Hamilton和Truong[27]采用局部线性回归构造了参数和非参数分量的估计，并证明了估计量的渐近正态性. Mammen和Van de Geer[28]应用经验过程理论构造了惩罚拟似然估计，并推导了该估计的渐近性质.

Severini等[29]和Härdle等[30]研究了模型(1.3)的推广形式：广义部分线性回归模型

$$E(y_i|x_i,t_i) = H\{x_i^T\beta + g(t_i)\}, \quad i = 1,2,\cdots,n, \tag{1.7}$$

其中H(称为联系函数)是一个已知的函数. 为了估计β和$g(\cdot)$，Severini等引进了拟似然估计方法，该方法有类似于似然函数的性质，但仅需指定y的二阶矩而不是完全分布. 基于Severini等人的方法，Härdle等[30]考虑了$g(\cdot)$的线性检验问

题. 他们对检验问题的研究补充了 Severini 等人的工作.

我国学者在部分线性回归模型的研究上也作了大量相当深刻的工作. Liang[31] 系统地研究了多种场合下 β 的渐近有效估计的构造. Shi[32] 利用分块多项式逼近方法得到了 β 和 $g(\cdot)$ 的稳健 M 估计 $\hat{\beta}$ 和 $\hat{g}(\cdot)$,在一定条件下证明了 $\hat{\beta}$ 具有渐近正态性,并得到了 $\hat{\beta}$ 和 $\hat{g}(\cdot)$ 的弱收敛速度. 柴根象等[33~35] 将小波方法引入部分线性回归模型,建立了回归参数 β 和未知函数 $g(\cdot)$ 的小波估计,证明了它们具有优良的大样本性质,并讨论了误差方差的小波估计及其渐近性质.

以上研究大都是在 (x_i^T, t_i) 是随机设计点列的情况下进行讨论的. 当 (x_i^T, t_i) 是固定设计时的研究成果相对较少. 而非随机设计情形并不是随机设计情形的特例,随机设计情况下的结果往往并不能简单的推广到非随机设计情形,因此二者的处理方法和假设条件也有区别. 胡舒合[36],高集体等[37~39] 分别研究了当 $g(\cdot)$ 的估计取一类非参数权函数估计时,β 的最小二乘估计和加权最小二乘估计的强相合性,渐近正态性,收敛速度,Berry - Esseen 界限以及重对数律等方面的大样本性质. 王启华[40,41] 在截断样本下研究了 β 和 $g(\cdot)$ 的估计的强相合性,$p(p \geq 2)$ 阶平均相合性和渐近正态性. 陈明华[42,43] 讨论了 β 和 $g(\cdot)$ 的估计的强相合性,$p(p \geq 2)$ 阶平均相合性和收敛速度.

柴根象等[44] 基于部分线性回归模型的可加性,提出了新的二阶段估计方法,得到了 β 和 $g(\cdot)$ 的核权函数形式的估计量 $\hat{\beta}$ 和 $\hat{g}(\cdot)$. 在 x_i 为固定设计点列,t_i 为随机设计点列的情形下,证明了 $\hat{g}(\cdot)$ 的强相合性和一致强相合性,其一致强收敛速度可达到非参数回归函数估计的最优一致强收敛速度 $(n^{-1}\log n)^{\frac{1}{3}}$,同时得到了 $\hat{\beta}$ 的渐近正态性.

薛留根[45~47] 将随机加权方法应用于部分线性回归模型的研究,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,并给出了估计量的最优收敛速度和随机加权逼近速度. 薛留根[48,49] 研究了部分线性回归模型中误差方差估计之分布的 Berry - Esseen 界限和非一致收敛速度.

近些年来,对部分线性回归模型的研究一直是统计学界的一大热点,其研究不断向各方面有所发展. 例如对删失数据,异方差,相依混合序列等情形下的研究都取得了可喜的研究成果. 具体可参见文献[50~52].

关于部分线性回归模型方法的详细讨论可进一步参考著作[53,54].

1.4.2 单指标回归模型

单指标回归模型首先考虑 p 维解释变量 x 的线性组合,把所有的解释变量投影到一个线性空间上,然后在这个一维线性空间上拟合一个一元函数. 单指标回归

模型形式如下

$$y_i = g(x_i^T\beta) + e_i, \quad i = 1,2,\cdots,n, \tag{1.8}$$

其中 $g(\cdot)$ 为一元未知函数，β 为 p 维未知投影参数向量，e_i 为均值为 0 的随机误差. $x^T\beta$ 称为指标. 由于指标 $x^T\beta$ 合并了 x 的维数，把 p 维解释变量降维到一元指标，从而使得单指标回归模型避免了多元非参数回归中出现的“维数灾祸”问题. 研究单指标回归模型的一大任务就是对未知函数和未知参数进行估计.

单指标回归模型作为一种广义的回归模型，是 20 世纪 80 年代中后期发展起来的一种重要的统计模型，该模型在金融经济，生物医学等领域具有广泛的应用背景. 国外自 20 世纪 80 年代末以来，一些统计文献从不同角度，根据不同的假设条件，对该模型做了一定的研究，并提出了一系列方法，而在国内，有关该模型的相关文献还很少.

下面简要回顾一下该模型的研究现状.

Ichimura[63] 研究了单指标回归模型的可识别性. 由于指标系数向量 β 的任意非 0 倍数可以被函数 $g(\cdot)$ 吸收，因此指标系数 β 可以是任意的. 考虑到模型的可识别性，必须要求 $\|\beta\| = 1$，并且 β 的第一个非 0 元素大于 0. Manski[64] 讨论了二元响应模型的可识别性，Horowitz[65] 讨论了单指标模型的其他可识别性要求.

假定 β 已知，则未知函数 $g(\cdot)$ 可以通过经典的 y 对 $u = x^T\beta$ 的一元非参数回归进行估计. 尽管 NW 核估计在各种方法中不是最优的，但由于计算简便，较容易实施，NW 核估计在单指标回归模型的讨论中使用最为广泛.

设 $\{(x_i^T, y_i), i = 1,\cdots,n\}$ 为 i. i. d. 样本，定义 $u_i = x_i^T\beta$. 记 K 为核函数，通常 K 为有界对称概率密度函数. h 表示窗宽. 回归函数 $g(\cdot)$ 的 NW 核估计为

$$\hat{g}^{\beta}(u) = \frac{\sum_{i=1}^{n} K(\frac{u-u_i}{h}) y_i}{\sum_{i=1}^{n} K(\frac{u-u_i}{h})}. \tag{1.9}$$

由于 β 是未知的，因此估计量 $\hat{g}^{\beta}(u)$ 在实际中不可行. 如果能够获得 β 的估计 $\hat{\beta}$，则 $g(\cdot)$ 的估计为

$$\hat{g}^{\hat{\beta}}(u) = \frac{\sum_{i=1}^{n} K(\frac{u-\hat{u}_i}{h}) y_i}{\sum_{i=1}^{n} K(\frac{u-\hat{u}_i}{h})}, \tag{1.10}$$

其中 $\hat{u}_i = x_i^T\hat{\beta}$.

Horowitz[65] 的研究结果表明，β 的 $\sqrt{n}$ 相合估计是存在的. 也就是说，存在 $\hat{\beta}$，使得

$$\hat{\beta} - \beta = O_p(n^{-\frac{1}{2}}).$$

这是参数估计所能达到的最好的收敛速度，而回归函数的非参数估计通常不能达到这一速度. 由于 $\hat{\beta}$ 的收敛速度比 $g(\cdot)$ 的任何非参数估计的收敛速度都来得快，因此，估计 $\hat{g}^{\beta}(u)$ 和 $\hat{g}^{\hat{\beta}}(u)$ 的区别是渐近可忽略的. 特别地，有下述结论：

$$(nh)^{\frac{1}{2}}[\hat{g}^{\hat{\beta}}(u)-g(u)]=(nh)^{\frac{1}{2}}[\hat{g}^{\beta}(u)-g(u)]+o_p(1).$$

这说明 β 的 $\sqrt{n}$ 相合估计对 $g(\cdot)$ 的 NW 估计的渐近分布不会产生任何影响，一旦 β 的估计 $\hat{\beta}$ 被得到，就可以用(1.10)式来获得 $g(\cdot)$ 的 NW 估计. 详细讨论可参见文献[65].

关于 β 的估计，文献中已经提出了许多方法. 概括而言，根据是否需要求解非线性最优化问题，可以把 β 的估计方法分为两大类：M 估计(M-estimators)和直接估计(direct estimators).

如果 $g(\cdot)$ 已知，β 的 M 估计具有如下形式

$$\hat{\beta}=\arg\max_{\beta}\frac{1}{n}\sum_{i=1}^{n}\Psi(y_i,g(x_i^T\beta)),$$

其中 Ψ 为满足一定正则条件的 $R^2\to R$ 上的函数. 在单指标回归模型中，为了消除估计的偏差，用分离一点 NW 估计 $\hat{g}_{(-i)}(x_i^T\beta)$ 替代未知函数 $g(\cdot)$. 由此，极大化

$$\frac{1}{n}\sum_{i=1}^{n}\Psi(y_i,\hat{g}^{\beta}_{(-i)}(x_i^T\beta))$$

可得 β 的估计. $g(\cdot)$ 的分离一点估计 $\hat{g}_{(-i)}(x_i^T\beta)$ 等于除去第 i 个观测后用剩余的 $n-1$ 个观测得到的 NW 估计.

β 的 M 估计包括半参数最小二乘估计(semiparametric least squares)和半参数极大似然估计(semiparametric maximum likelihood).

Ichimura[63] 和 Härdle 等[66] 详细地研究了半参数最小二乘估计方法. 该方法采用了参数回归模型中最小二乘估计的思想. 在未知参数给定的情况下，用分离一点 NW 估计方法估计未知函数 $g(\cdot)$，所得的估计是未知参数 β 的一个函数，然后用非线性最小二乘的思想极小化残差，可得未知参数的估计，估计量为

$$\hat{\beta}=\arg\min_{\beta\in R^p}\sum_{i=1}^{n}W(x_i)[y_i-\hat{g}^{\beta}_{(-i)}(x_i^T\beta)]^2,$$

其中 $W(\cdot)$ 为一个非负权函数.

由于该方法需要解决复杂的非线性最优化问题，因此计算较为复杂. 关于该方法的进一步推广可参看文献[67~69].

半参数极大似然估计沿用了参数极大似然估计的思想. 在单指标模型中，x 和 y 的联合分布以及给定 x 的情况下 y 的条件密度都既依赖于参数 β 又依赖于回归函数 $g(\cdot)$. 假定 y 的条件密度通过指标 $x^T\beta$ 依赖于 x，并把联合分布和条件密度分别记为 $l_{g,\beta}(\cdot,\cdot)$ 和 $l_{g,\beta}(\cdot\mid\cdot)$，则似然函数可表示为

$$L_g(\beta) = \prod_{i=1}^{n} l_{g,\beta}(x_i, y_i) = \prod_{i=1}^{n} l_{g,\beta}(y_i \mid x^T\beta = x_i^T\beta) f(x_i),$$

其中 f 是 x 的边缘密度. 因此对数似然函数为

$$LL_g(\beta) = \sum_{i=1}^{n} \log l_{g,\beta}(y_i \mid x^T\beta = x_i^T\beta) + \sum_{i=1}^{n} \log f(x_i).$$

易见项 $\sum_{i=1}^{n} \log f(x_i)$ 不依赖于 β 和 $g(\cdot)$，极大化 $LL_g(\beta)$ 等于极大化 $\sum_{i=1}^{n} \log l_{g,\beta}(y_i \mid x^T\beta = x_i^T\beta)$. 用回归函数 $g(\cdot)$ 的分离一点 NW 估计代替 $g(\cdot)$，则 β 的半参数极大似然估计是下面的极大化问题的解

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \log l_{\hat{g}^{\beta}_{(-i)},\beta}(y_i \mid x^T\beta = x_i^T\beta).$$

Delecroix 等[70]证明了 β 的半参数极大似然估计是渐近有效的，它保持了参数极大似然估计的最重要的性质.

M 估计具有有效性、渐近正态性和自动窗宽选择等许多优点，但该方法的最大缺点是它需要解复杂的最优化问题. 尽管直接估计方法的理论性质不如 M 估计，但由于其能够提供估计量的解析形式，因此，在单指标模型的估计中直接估计方法具有一定的吸引力.

Stoker[71]和 Härdle 等[72]提出平均导数方法拟合单指标回归模型，其基本思想是考察平均导数向量

$$\gamma = \mathrm{E}(\nabla m(x)) = \mathrm{E}(g^{(1)}(x^T\beta))\beta,$$

其中 $\nabla m(x) = \{\frac{\partial m}{\partial x_j}, j = 1, 2, \cdots, p\}$ 是 p 元函数 $m(\cdot)$ 的偏导数向量. 然后关于 x 的边际分布求期望. 由此，从 γ 的估计就可以获得单指标系数 β 的直接估计. 当估计出未知参数 β 之后，可以用一元非参数回归方法来估计未知函数 $g(\cdot)$. Härdle 等人证明了 $\hat{\beta}$ 的相合性，并讨论了 $g(\cdot)$ 的估计的大样本性质. 但是，其方法需首先获得 x 的边际密度的核估计，然后获得 γ 的估计，这一过程在实际应用中是比较复杂的.

关于平均导数方法进一步的讨论可参看文献[73,74]等.

Powell 等[75]描述了密度加权平均导数估计方法. 这种方法不需要迭代并且容易计算. 密度加权平均导数估计适用于所有 x 的分量都是连续随机变量的情形，它基于如下关系式

$$\beta \propto \mathrm{E}[p(x)\partial g(x^T\beta)/\partial x] = -2\mathrm{E}[y\partial p(x)/\partial x],$$

其中 $p(\cdot)$ 是 x 的概率密度函数. 第二个等式是通过对第一个式子进行分部积分得到的. 通过用 $p(\cdot)$ 的非参数估计取代 $p(\cdot)$ 和用样本均值取代总体均值 E 就可以估计等式右边的表达式，从而可以对 β 进行直接估计. Horowitz 等[76]把这种方

法推广到 x 含有离散变量的情形，并且举例说明了单指标回归模型的应用，Hristache 等[77]对这种方法提出了进一步的推广和修正.

Samarov[78]，Ichimura 等[79]进一步讨论了单指标回归模型(1.8)的推广形式多指标回归模型，关于多指标回归模型这里不再作过多的介绍.

1.4.3 部分线性单指标回归模型

在一些情况下，响应变量 y 可能与一些解释变量 z 之间是线性关系，同时与另一些解释变量 x 之间是非线性关系，在这种情况下，我们可以考虑把单指标回归模型推广到部分线性单指标回归模型

$$y_i = z_i^T\alpha + g(x_i^T\beta) + e_i, \quad i = 1,2,\cdots,n, \tag{1.11}$$

其中 α 为 q 维未知参数向量，刻画了 y 与 z 之间的线性关系. β 为 p 维未知参数向量，刻画了 x 的线性组合，$g(\cdot)$ 为一元未知函数，函数 $g(\cdot)$ 和线性组合 $x^T\beta$ 刻画了 y 与 x 之间的非线性关系. e_i 为均值为 0 的随机误差.

部分线性单指标回归模型(1.11)是一个比较广泛的模型.

(1)若 $z=0$，则模型(1.11)就简化为单指标回归模型(1.8).

(2)若 β 为一维变量，则模型(1.11)就简化为部分线性回归模型(1.3).

(3)若 $z=0$ 且 β 为一维变量，则模型(1.11)就简化为一元非参数回归模型(1.2).

(4)若 $z=0$，$g(\cdot)$ 为正态分布函数或 logistic 分布函数，则模型(1.11)即为 probit 回归模型或 logistic 回归模型.

(5)若 $g(\cdot)=0$，则模型(1-11)就简化为古典的线性回归模型.

部分线性单指标回归模型对解释变量 x 使用降维方法，避免了“维数灾祸”问题，同时该模型又把参数回归模型和非参数回归模型有机地结合起来，因此，部分线性单指标回归模型较单纯的参数回归模型或非参数回归模型有更大的适应性. 部分线性单指标回归模型不仅有实际的应用背景，而且有广泛的应用前景，在金融经济和生物医学等领域具有很大的应用价值. 部分线性单指标回归模型为数据处理提供了十分有用的统计推断工具，它们的理论和方法在最近几年越来越受到人们的重视. 一些学者对该模型进行了研究并逐步形成了目前统计界的热门课题之一.

Carroll 等[80]讨论了广义部分线性单指标回归模型

$$G(\mu(x_i,z_i)) = z_i^T\alpha + g(x_i^T\beta), \quad i = 1,2,\cdots,n, \tag{1.12}$$

其中 $G(\cdot)$ 是已知的联系函数，当 $G(\cdot)$ 是单位函数时模型(1.12)即简化为部分线性单指标回归模型(1.11). 对于未知参数向量 α,β 和未知函数 $g(\cdot)$ 的估计，Carroll 等利用局部拟似然方法提出了如下算法：

(1)利用参数模型,例如多元线性回归模型,获得未知参数 α,β 的初始估计 $\hat{\alpha}$, $\hat{\beta}_1$. 对单指标参数 β 的估计量 $\hat{\beta}_1$ 进行标准化:$\hat{\beta}=\hat{\beta}_1/\|\hat{\beta}_1\|$.

(2)使用局部线性方法,关于 a,b 极大化局部拟似然函数

$$\sum_{i=1}^{n} Q[G^{-1}\{a+b(x_i^T\hat{\beta}-u)+z_i^T\hat{\alpha}\},y_i]K_h(x_i^T\hat{\beta}-u),$$

得到 $g(\cdot)$ 的估计 $\hat{g}(u;h,\hat{\alpha},\hat{\beta})=\hat{\alpha}$.

(3)利用估计 $\hat{g}(\cdot)$,关于 α,β 极大化全局拟似然函数

$$\sum_{i=1}^{n} Q[G^{-1}\{\hat{g}(u;h,\hat{\alpha},\hat{\beta})+z_i^T\alpha\},y_i],$$

获得 α,β 的新估计 $\hat{\alpha},\hat{\beta}$.

(4)重复第(2)步和第(3)步直到收敛.

(5)用 α,β 的最终估计 $\hat{\alpha},\hat{\beta}$ 获得 $g(\cdot)$ 的最终估计 $\hat{g}(\cdot)$.

Carroll 等详细讨论了估计 $\hat{\alpha},\hat{\beta}$ 和 $\hat{g}(\cdot)$ 的渐近分布,并且给出了模拟和实际例子进一步解释了模型和提出的估计方法. 该方法的缺点是:它要求 α,β 的初始估计接近于最终结果,且当 x 的维数较高时,这种算法计算比较困难.

Yu 等[81]提出了惩罚样条估计方法. 该方法首先通过样条函数

$$g(u)=\delta_0+\delta_1 u+\cdots+\delta_p+\sum_{k=1}^{K}\delta_{p+k}(u-s_k)_+^p$$

估计一元未知函数 $g(\cdot)$(其中 $s_k,k=1,2,\cdots,K$ 为结点),然后对惩罚的误差平方和求极小化得到未知参数 α,β 的估计 $\hat{\alpha},\hat{\beta}$. 在参数空间为紧空间的情况下,讨论了部分线性单指标回归模型估计量的存在性,相合性和渐近正态性. 进一步地,Yu 等[82]在一般的 Euclidean 空间下也给出了估计量的相合性结果,文中对该方法进行了模拟研究,给出了实际应用例子.

部分线性单指标回归模型的惩罚样条方法提供了一个直接拟合方法,它计算简便且富有弹性. 惩罚函数不仅适用于连续数据也适用于非连续数据.

Xue 等[83]对部分线性单指标回归模型的未知参数 α,β 提出了三种对数经验似然比统计量. 在适当的条件下,证明了所提出的统计量依分布收敛于 χ^2 分布. 所得结果可以用来构造未知参数的置信域. 所提出的方法也适合于纯粹的单指标模型. 并通过模拟研究说明了经验似然方法在置信域的精度及其覆盖概率大小方面优于最小二乘估计方法.

在部分线性单指标回归模型(1.11)中,由于函数 $g(\cdot)$ 是未知的,并且函数 $g(\cdot)$ 的自变量不是简单的实数,而是包含指标参数向量 β 的乘积形式 $x^T\beta$,这样就使得处理此类回归模型较一般的部分线性回归模型更具有复杂性和挑战性. 纵观以上单指标回归模型或部分线性单指标回归模型的研究,我们可以发现,对于未