中央财经大学学术著作基金资助出版

FUZA SHUJUXIA BANCANSHU
DANDIAO HUIGUI MOXING DE TONGJI TUIDUAN

复杂数据下半参数单调回归模型的统计推断

孙志猛 著

本书重点探讨了测量误差数据、缺失数据、随机删失数据等几种复杂数据类型的半参数单调回归模型建模方法。

经济科学出版社
Economic Science Press

中央财经大学学术著作基金资助出版

二零一二

FUZA SHUJUXIA BANCAN
DANDIAO HUIGUI MOXIN

复杂数据下半参数单调回归模型的统计推断

孙志猛 著

经济科学出版社

图书在版编目（CIP）数据

复杂数据下半参数单调回归模型的统计推断 / 孙志猛著. — 北京 ： 经济科学出版社，2015.7

ISBN 978-7-5141-5878-6

Ⅰ. ①复… Ⅱ. ①孙… Ⅲ. ①半参数模型—统计推断 Ⅳ. ①O211.3

中国版本图书馆 CIP 数据核字（2015）第 150140 号

责任编辑：袁 溦
责任校对：靳玉环
责任印制：邱 天

复杂数据下半参数单调回归模型的统计推断

孙志猛 著

经济科学出版社出版、发行 新华书店经销

社址：北京市海淀区阜成路甲 28 号 邮编：100142

总编部电话：010-88191217 发行部电话：010-88191537

网址：www.esp.com.cn

电子邮件：esp@esp.com.cn

北京京鲁数码快印有限责任公司印装

710×1000 16 开 9.5 印张 100000 字

2015 年 8 月第 1 版 2015 年 8 月第 1 次印刷

ISBN 978-7-5141-5878-6 定价：32.00 元

（图书出现印装问题，本社负责调换。电话：010-88191502）

前　言

半参数回归模型是参数回归模型和非参数回归模型的结合，其既含有参数分量，又含有非参数分量，在描述实际问题时更具灵活性和解释能力. 在用半参数回归模型解决实际问题时，经常碰到模型的非参数分量与其解释变量的关系具有明显的单调性的情形. 受此启发，统计学者提出了半参数单调回归模型. 在实际问题中，我们经常会遇到各种复杂类型数据，如测量误差数据、缺失数据、删失数据等. 由于这些数据具有复杂的结构，忽略其结构的统计推断方法往往会降低统计推断的效率. 目前，在复杂数据下，对于半参数单调回归模型的研究还不多见. 因此，研究复杂数据下半参数单调回归模型的统计推断方法具有一定的理论意义和实用价值.

本书主要研究复杂数据下半参数单调回归模型的估计问题，考虑了测量误差数据、缺失数据和随机右删失数据等数据类型.

首先，研究了半参数单调回归测量误差模型的参数部分和非

参数部分的估计问题. 利用核估计方法和纠偏的思想得到了参数的相合估计, 并利用参数部分的估计进一步构造了非参数部分的单调估计. 证明了参数部分估计的渐近分布为正态分布, 非参数部分估计的渐近分布为双边的标准布朗运动与时间参数平方的和的最大凸弱函数在 0 点的左斜率. 通过随机模拟比较了考虑测量误差和忽略测量误差两种方法得到的估计的有限样本性质.

其次, 研究了响应变量随机右删失、解释变量带有测量误差情况下半参数单调回归模型的估计问题. 采用了对随机右删失数据进行无偏变换的方法来处理删失问题、对最小二乘目标函数纠偏的思想来处理测量误差问题. 首先把半参数模型转化为线性模型构造了参数部分的 $\sqrt{n}$ 相合估计, 然后构造了非参数部分的单调约束最小二乘估计, 保证了非参数部分估计的单调性. 证明了参数部分估计的渐近正态性, 同样得到了非参数部分估计的渐近分布. 通过随机模拟研究了不同删失概率下考虑测量误差和忽略测量误差两种方法得到的估计的有限样本性质.

再次, 研究了响应变量随机缺失、解释变量带有测量误差情况下半参数单调回归模型的估计方法. 采用对缺失数据进行借补的思想处理缺失问题, 提出了参数部分和非参数部分的两步估计. 证明了参数部分估计的渐近分布为正态分布, 得到了非参数部分的收敛速度. 通过随机模拟研究了不同缺失概率下考虑测量误差和忽略测量误差两种方法得到的估计的有限样本性质.

最后, 研究了解释变量带有测量误差情况下半参数可加单调

回归模型的估计. 证明了参数部分估计的渐近分布为正态分布, 同时得到了非参数部分估计的渐近分布, 证明了每一个非参数部分估计的 Oracle 性质, 也即从渐近分布上说, 每一个非参数部分的估计都和其他分量已知的情况下得到的估计有一样的精度. 通过随机模拟研究了估计的有限样本性质.

作 者

2015 年 6 月

目　录

第 1 章

绪　论

回归分析是一种重要的统计分析方法. 借助回归分析, 人们可以通过分析实验数据或观测数据研究变量之间相互依赖的定量关系. 例如, 人们要研究一个国家或地区的香烟的销售量是否与抽烟者的年龄、受教育程度、收入以及香烟的销售价格等社会经济因素有关系. 这种关系可以用包含响应变量和解释变量的等式来描述. 在上述香烟的消费例子中, 香烟的消费量是响应变量, 抽烟者的年龄、受教育程度、收入和香烟的销售价格等因素是解释变量. 记响应变量为 Y, 各解释变量分别为 $X_1, X_2, \cdots, X_p$, 其中 p 为解释变量的个数. Y 和 $X_1, X_2, \cdots, X_p$ 之间的关系可以由如下的回归模型来描述:

$$Y = f(X_1, X_2, \cdots, X_p) + \varepsilon, \tag{1-1}$$

其中, ε 为随机误差, 代表除 $X_1, X_2, \cdots, X_p$ 之外的 Y 的影响因素,

函数 $f(X_1, X_2, \cdots, X_p)$ 用来描述 Y 与 $X_1, X_2, \cdots, X_p$ 之间的关系, 通常被称为回归函数. 假设在试验中对 $(Y, X_1, X_2, \cdots, X_p)$ 作 n 次观测得到随机样本 $\{(Y_i, X_{1i}, X_{2i}, \cdots, X_{pi}), i = 1, \cdots, n\}$. 在回归分析中, 统计推断的基本问题是基于 $\{(Y_i, X_{1i}, X_{2i}, \cdots, X_{pi}), i = 1, \cdots, n\}$ 估计回归函数 $f(\cdot)$, 从而进行统计预测和决策. 自然, 统计推断方法与回归函数的基本假设密切相关. 因此, 可以根据回归函数的基本假设, 把回归模型分为参数回归模型、非参数回归模型和半参数回归模型三种不同的类型.

参数回归模型在传统的统计分析方法中占有重要的地位, 研究成果也较为成熟. 非参数回归模型近年来也得到了快速的发展. 半参数回归模型是参数回归模型和非参数回归模型的结合. 鉴于其在实际应用中的灵活性, 半参数回归模型越来越受到统计学者的关注, 涌现出了大量的研究成果.

1.1 模型

1.1.1 参数回归模型

一个统计模型是一类定义在同一个样本空间上的概率测度的集合. 这类统计模型也可表示为 $\mathscr{P} = \{p_v : v \in \mathscr{V}\}$, $\mathscr{V}$ 为参数空间. 若 $\mathscr{V}$ 为有限维空间或有限维空间的子集, 分布族 $\mathscr{P}$ 的分布函数形式是已知的, 未知的仅是有限个参数, 则模型 $\mathscr{P}$ 称为参数模型. 具体到回归问题中, 参数回归模型是指除回归函数中的有限个未知参数外其他因素都是已知的回归模型. 相对于非参数回归模型

和半参数回归模型来说, 参数回归模型的估计和统计推断问题相对容易, 其发展历史也较为久远, 已经形成了一套比较成熟的理论和方法. 当回归函数是未知参数的线性函数时, 回归模型也称为线性回归模型. 线性回归模型是一类最简单的回归模型, 查特吉和哈迪 (Chatterjee and Hadi)[1] 结合大量的实际数据研究了线性回归模型的参数估计、假设检验和回归诊断等问题, 系统介绍了参数估计的渐近正态性、误差异方差时的加权最小二乘方法、误差自相关的检测和迭代估计、解释变量多重共线性的诊断和主成分方法以及模型的变量选择等问题. 文献 [2]~[6] 也对线性模型进行了系统的研究. 正态线性模型是一类应用最广泛的参数线性回归模型, 其形式如下:

$$Y = X^{\top}\beta + \varepsilon,$$

其中, $X = (X_1, \cdots, X_p)^{\top}$ 为 p 维解释变量, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^{\top}$ 为 p 维未知回归参数, $\top$ 表示转置, ε 为随机误差, 服从于均值为 0 的正态分布, 且独立于 X.

对于正态线性回归模型, 若得到 β 的估计, 则自然也得到了回归函数的估计, 从而可以进行统计预测和决策. 在正态线性回归模型下, 估计 β 的常用方法是最小二乘法, 其思想是得到的参数估计应该使响应变量的真值和预测值在平方距离的意义下最接近, 即使预测误差的平方和达到最小. 在观测样本 $\{(Y_i, X_{1i}, X_{2i}, \cdots, X_{pi}), i = 1, \cdots, n\}$ 下, β 最小二乘估计可以表达为

$$\hat{\beta}_n = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} [Y_i - X_i^{\top}\beta]^2 \right\}. \tag{1-2}$$

求解最优化问题 (1-2) 式, 关于 β 对目标函数对求偏导, 令偏导数等于 0, 得正则方程

$$\sum_{i=1}^{n} X_i(Y_i - X_i^{\top}\beta) = 0.$$

显然, 当 $[\sum_{i=1}^{n} X_i X_i^{\top}]^{-1}$ 存在时, 正则方程的解为

$$\hat{\beta}_n = \left[\sum_{i=1}^{n} X_i X_i^{\top}\right]^{-1} \cdot \sum_{i=1}^{n} X_i Y_i.$$

参数线性回归模型的回归函数形式较为简单, 估计方便, 且由于该模型仅依赖于有限个回归参数, 因此当实际问题与假设模型较为接近时, 其统计推断往往具有较高的精度. 然而, 由于参数线性回归模型的模型假定较强, 当实际问题与模型假定相差较远甚至相背离时, 用该模型拟合实际问题的效果就较差, 甚至得出错误的结论. 这促使统计学者寻找一种适用性更强的统计方法. 非参数回归模型正是朝着这个方向努力的一种统计方法.

1.1.2 非参数回归模型

在参数模型中, 总体的分布是已知的, 未知的仅是有限个参数. 然而在许多实际问题中, 数据并不是来自假设已知的总体, 或者数据出现被污染的情况. 这时, 参数模型就不能有效地解决实际问题. 于是, 人们希望不对总体的分布形式作特殊假定, 而是尽量利用样本本身的信息进行统计推断. 这正是非参数统计的思想. 与参数统计模型不同的是, 非参数统计模型往往不是依赖于有限个未知参数, 其参数空间往往是某个无穷维参数空间, 即一个非参数统

计模型可以表示为 $\mathscr{P}=\{p_v : v \in \mathscr{V}\}$, $\mathscr{V}$ 为某无穷维参数空间. 具体到回归问题中, 一个常见的非参数回归模型为:

$$Y = h(X) + \varepsilon, \tag{1-3}$$

其中, Y 为一维响应变量, X 为 p 维协变量, $h \in \mathcal{H}$, $\mathcal{H}$ 为 R^p 上某个函数空间, ε 为均值为 0 的随机误差, 且与 X 相互独立.

非参数回归模型的基本问题是估计未知回归函数 $h(\cdot)$. 近年来, 随着统计科学的发展, 涌现出了许多非参数回归函数的估计方法, 如核回归、局部多项式回归、样条逼近方法和小波方法等. 范 (Fan)[7] 提出了回归函数的局部线性光滑估计, 分别计算了估计的均方误差和积分均方误差, 得到了估计的收敛速度; 范和吉贝尔斯 (Fan and Gijbels)[8] 进一步研究了回归函数的局部多项式估计, 讨论了估计的渐近性质; 多诺霍 (Donoho)[9] 研究了基于小波的估计方法, 证明了小波方法具有空间适应性, 即该方法对于空间的非齐次曲线具有很好的适应性; 尤班克 (Eubank)[10] 对样条方法进行了全面的介绍; 何和史 (He and Shi)[11] 研究了非参数分位数回归模型, 得到了回归函数 B 样条估计的收敛速度. 在本书的研究方法中, 主要用到了核回归, 下面仅就一维解释变量的情况重点介绍之. 核回归是一种局部平均的方法, 由纳达拉亚 (Nadaraya)[12] 和沃森 (Watson)[13] 分别提出. 通常通过核函数来定义权函数. 核函数 $K(\cdot)$ 一般是光滑函数, 满足

$$K(x) \geqslant 0, \int K(x)\mathrm{d}x = 1, \int xK(x)\mathrm{d}x = 0, \int x^2K(x)\mathrm{d}x > 0.$$

核函数通常取某个均值为 0 的概率密度函数. 下面是一些常用的核函数:

$$\begin{aligned}
\text{Boxcar}\ \ \text{核}: K(x) &= \frac{1}{2} I[x] \\
\text{Gaussian}\ \ \text{核}: K(x) &= \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}} I[x] \\
\text{Epanechnikov}\ \ \text{核}: K(x) &= \frac{3}{4}(1-x^2) I[x] \\
\text{Tribube}\ \ \text{核}: K(x) &= \frac{70}{81}(1-|x|^3)^3 I[x]
\end{aligned}$$

其中,

$$I[x] = \begin{cases} 1, & |x| \leqslant 1, \\ 0, & |x| > 1. \end{cases}$$

假设来自模型 (1-3) 的样本为 $\{(Y_i, X_i), i = 1, \cdots, n\}$. 定义权函数

$$\varpi_{ni}(x) = \frac{K(\dfrac{x - X_i}{h_n})}{\displaystyle\sum_{j=1}^{n} K(\frac{x - X_j}{h_n})}, i = 1, \cdots, n.$$

其中, 光滑参数 h_n 称为窗宽. 则模型 (1-3) 回归函数的核估计定义为

$$\hat{h}(x) = \sum_{i=1}^{n} \varpi_{ni}(x) Y_i.$$

核估计有一个有意思的解释: 它是由局部加权最小二乘得到的局部常数估计. 为了说明这一点, 定义 $l_{ni} = K(\dfrac{x - X_i}{h_n})$, 选择 $a \equiv \hat{h}(x)$ 使得加权平方和

$$\sum_{i=1}^{n} l_{ni}(x)(Y_i - a)^2$$

达到最小. 由初等微积分不难得到解为

$$\hat{h}(x)=\frac{\sum_{i=1}^{n} l_{ni}(x)Y_i}{\sum_{i=1}^{n} l_{ni}(x)},$$

这正是核估计.

在核估计中, 窗宽 h_n 的作用是用来控制光滑的程度. 由核估计的定义不难看出, 核估计既同样本有关, 也同核函数和窗宽的选取有关. 在给定样本之后, 核估计性能的好快取决于核函数和窗宽的选取是否恰当. 若窗宽选取较大, 会使分布的某些局部特征被掩盖起来, 从而造成了过度平均, 有增大偏差的趋势; 若窗宽选取较小, 整个估计又会出现较大的波动, 有增大方差的趋势. 因此, 在核回归中, 窗宽的选取是一项重要的工作. 而对于核函数, 理论和实践均表明, 核函数的选取对核估计的性质并没有本质的影响. 德罗伊和瓦格纳 (Devroye and Wagner)[14] 证明了回归函数核估计的均方收敛性.

一方面, 由于非参数统计方法不依赖于总体分布, 因此, 即使对总体的分布所知甚少甚至完全未知, 非参数统计方法仍然可以得到可靠的结论. 这使得非参数统计方法具有更广泛的适用性. 另外, 非参数统计方法仅利用样本的一般信息, 对模型的限制较少, 因此若真实模型与假定的理论模型有不大的偏离, 该方法得到的估计仍然具有良好的性质, 至少不至于变得太坏, 因此非参数统计方法具有天生的稳健性. 另一方面, 当借助历史经验或专业知识对

总体的分布有较多的了解甚至总体分布已知时, 不利用任何先验知识就成为非参数统计方法的缺点. 在这样的情况下, 由于它没有充分利用总体分布的已知的信息, 所以得出的统计推断结论就不如参数方法精确.

1.1.3 非参数单调回归模型

在模型 (1-3) 中, 若参数空间 $\mathcal{H}$ 为单调函数的全体构成的集合, 则该回归模型称为非参数单调回归模型. 非参数单调回归模型是一类重要的非参数回归模型, 这是因为在许多实际应用中, 研究者强烈的相信真实的回归函数具有单调性. 因此, 很自然地, 我们只能在单调函数的集合中选择回归函数. 下面是单调回归方面的实际例子.

例 1.1.3.1 表 1-1 列出了波特霍夫和罗伊 (Potthoff and Roy) 讨论过的来自牙科研究的部分数据 (Robertson et al. [15]). 该研究在北卡罗来纳大学牙科学院 (University of North Carolina Dental School) 进行. 研究测量了一定数量研究个体的脑垂体裂缝的尺寸 (单位是毫米). 表 1-1 的数据包含了三个 8 岁女生、三个 10 岁女生、三个 12 岁女生和两个 14 岁女生的测量数据. 在该研究中, 假设个体的脑垂体裂缝尺寸随着年龄的增长而增长是合理的. 因此脑垂体裂缝尺寸和年龄关系可以用非参数单调回归模型来拟合.