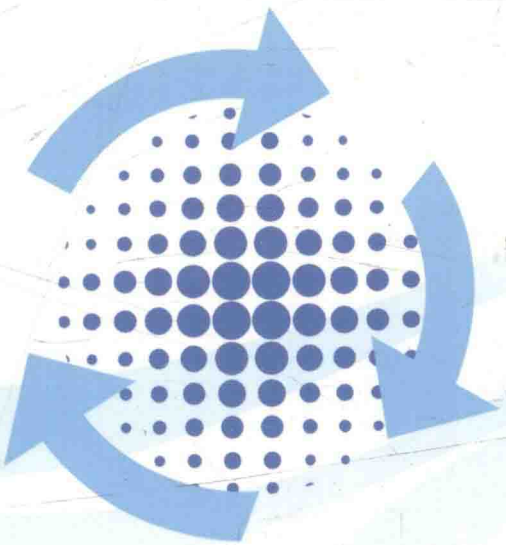
半参数模型的约束统计推断及应用

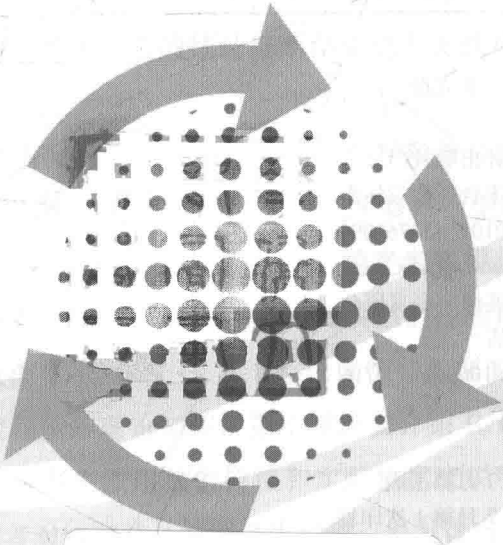
丁建华 著



中国统计出版社
China Statistics Press

半参数模型的约束统计推断及应用

丁建华 著



中国统计出版社
China Statistics Press

图书在版编目（CIP）数据

半参数模型的约束统计推断及应用 / 丁建华著. -- 北京 : 中国统计出版社, 2017.7
ISBN 978-7-5037-8146-9

Ⅰ. ①半… Ⅱ. ①丁… Ⅲ. ①非参数统计—研究 Ⅳ. ①O212.7

中国版本图书馆 CIP 数据核字(2017)第 108396 号

半参数模型的约束统计推断及应用

作　　者/丁建华
责任编辑/李潇潇　王立群
封面设计/李雪燕
出版发行/中国统计出版社
通信地址/北京市丰台区西三环南路甲 6 号　邮政编码/100073
电　　话/邮购（010）63376909　书店（010）68783171
网　　址/http://www.zgtjcbs.com/
印　　刷/河北鑫兆源印刷有限公司
经　　销/新华书店
开　　本/710mm×1000mm　1/16
字　　数/210 千字
印　　张/16.5
版　　别/2017 年 7 月第 1 版
版　　次/2017 年 7 月第 1 次印刷
定　　价/30.00 元

前　言

约束统计推断是数理统计学的一个重要分支。众所周知，在经济学、生物学和社会学等领域，变量与变量之间联系的先验信息往往是可以利用的，用好这些信息会使统计推断更有效，更合理。例如儿童生长曲线的研究，身高随着年龄而增高，因而生长曲线满足单调非降性。在生存分析中危险率函数可能满足递增、递减、常数、凸凹性等各种形状，如由于老化或损耗危险率递增，而大多数总体的危险率则服从凸约束。鉴于此，我们在做统计推断时，这些约束条件必须放在统计模型中，统计推断问题中有了约束条件，需要新的方法来解这些问题，无约束统计问题中所使用的数学方法常常不能再使用，它的统计推断结果也与无约束统计中相应的问题不同。

本书基于 R 语言案例分析系统地介绍了半参数部分线性模型在约束条件下的统计推断问题，共十章。前四章介绍半参数模型的约束最小二乘估计及稳健估计，第五章在 Bayes 框架下讨论半参数模型在约束条件下的估计问题，第六章介绍约束条件下半参数模型的检验问题，第七章介绍部分线性变系数模型的估计，第八章给出函数型数据的约束估计方法，第九章介绍经验过程基础理论知识，第十章简要介绍 R 语言并给出部分章节的主要程序。本书力图用数值模拟和实际数据来说明基本思想和基本方法，尽量使读者对约束统计推断产生兴趣，引发读者建立约束统计模型去认识和解决实际问题。

本书从问题背景、方法引进、理论证明，计算机 R 语言以及应用实例等方面介绍半参数模型的约束统计推断方法。每章配有实际例子说明

方法的实用性，附录中给出部分章节的程序，方便读者利用 R 语言进行编程。另本书的证明要用到经验过程的基本知识，侧重方法论的读者可略去证明。

书稿虽经多次检查、修改、但疏漏和不足之处恐怕仍在所难免，真诚欢迎读者批评指正。

丁建华

于山西大同大学统计系

2017 年 3 月

内容简介

本书介绍在约束条件下半参数模型的统计推断。内容主要包括参数估计、假设检验的方法和理论及Bayes估计问题。本书侧重内容的科学性，体现学术思想，注重方法论、模拟计算和实例分析。

本书对高等院校数学与统计专业的研究生、教师及相关科研机构研究人员具有参考价值。

目 录

第 1 章
引言

1.1　研究的问题

近年来, 非参数统计和半参数统计已经成为统计研究的重要分支。在非参数回归模型

$$Y = \psi(Z) + \varepsilon \tag{1-1}$$

中, 当回归函数在没有任何形状限制的条件下, 已经有比较完善的理论。估计未知回归函数方法的有核估计 (Nadaraya 1964[1]; Watson 1964[2]), 样条估计 (Eubank 1988[3]), 局部多项式估计 (Fan 1993[4]) 等方法。但在实际应用中, 由经验或专业知识可知 $\psi(\cdot)$ 满足条件 $\psi^{(r)}(\cdot) \geqslant 0$ (或 $\psi^{(r)}(\cdot) \leqslant 0$), 这包括非负性，单调性和凸凹性等形状约束, 这类型形状约束问题经常出现在各种领域中。例如, 在空间流行病学研究中, 研究工人某种疾病的发生风险与其距不同辐射源的距离之间的关系时, 人们相信距离各种辐射源的距离越近, 疾病发生的风险越大。又如由经济学理论知生产函数或恩格尔曲线具有凸性。当回归函数满足上述形状约束时, 由随机误差, 通常所使用的无约束估计不一定满足所要求的形状约束。自从 Brunk(1955)[5] 给出单调参数的最大似然估计, 具有单调约束的统计模型的推断受到了广泛关注, 统计学者一直致力于寻找满足所需形状限制的计算上行之有效的非参数光滑估计。

非参数回归模型的自然推广是半参数回归模型:

$$Y = \boldsymbol{X}^T\boldsymbol{\beta} + \psi(Z) + \varepsilon \qquad (1-2)$$

其中 $(\boldsymbol{X}, Z)$ 是协变量或解释变量, Y 是响应变量, $\boldsymbol{\beta}$ 是未知回归系数向量, ε 为均值为0的随机误差且与 $(\boldsymbol{X}, Z)$ 相互独立。半参数模型介于参数回归模型和非参数回归模型之间, 在应用上, 这种模型可描述许多实际问题, 比单纯的参数模型和非参数模型有更大的适应性。该模型自从Engle等[6]提出以来就受到了许多统计学者的关注。例如, Heckman(1986)[7]利用惩罚似然估计方法研究了 $\boldsymbol{\beta}$ 的估计的渐近性质。Chen(1988)[8]利用分段多项式近似 $\psi(\cdot)$, 并证明了 $\boldsymbol{\beta}$ 的估计可达到 $n^{-1/2}$ 收敛速度和最小的渐近方差。Speckman(1988)[9]利用核光滑方法研究了部分线性模型的理论性质。Mammen和van der Geer(1997)[10]应用经验过程理论研究了参数 $\boldsymbol{\beta}$ 的惩罚拟似然估计的渐近性质。Hamilton和Truong(1997)[11]利用局部线性光滑方法导出了 $\boldsymbol{\beta}$ 和 ψ 估计的渐近分布。更多关于半参数回归模型的文献有 [12-15]。与模型 (1-1) 类似, 实际中也往往需要模型 (1-2) 中的 $\psi(\cdot)$ 满足某种形状约束, 我们看下面的例子。

例 1.1 CD4细胞是人体免疫系统的重要组成部分, 人类免疫缺陷病毒HIV对CD4细胞具有破坏作用。研究者想了解平均CD4百分数随着时间的衰减趋势, 并评价抽烟, 感染HIV时的年龄, 感染前CD4百分数(PreCD4)对感染后平均CD4百分数(PerCD4)的影响。取感染HIV后个体在不同时间点上CD4细胞的百分数为响应变量, 并考虑三个协变量: 抽烟状态, 年龄和感染前CD4百分数。统计分析表明只有变量PreCD4与PerCD4线性相关, 抽烟和年龄对PerCD4没有显著影响。由医学常识可知, 个体感染HIV病毒后, HIV病毒对CD4细胞有破坏性的作用, 因此我们有理由认为, 响应变量PerCD4随着时间的推移是单调递减的。综上

所述, 对 HIV 数据集我们可建立如下单调约束半参数部分线性模型:

$$\mathrm{CD4}_{ij} = \psi(\mathrm{time}_{ij}) + \beta \mathrm{PreCD4}_{ij} + \varepsilon_{ij},$$
$$i = 1, \cdots, n_i;\ j = 1, \cdots, m$$

其中 time_{ij} 是第 i 个艾滋病人第 j 次测量的时间, 系数 β 是 PreCD4 的固定效应, $\psi(\cdot)$ 是时间的单调递减函数。显然, 在这个例子中, $\psi(\cdot)$ 的拟合应满足单调性假设。然而, 在无约束情形下, 拟合上述模型并不能保证这种单调性。我们在第二章 2.4.2 节详细分析了这个例子, 图 2-2 给出的无约束拟合曲线在两端边界部分出现上升的趋势, 违反了 $\psi(\cdot)$ 是时间的单调递减函数的假设。为此, 本书讨论在半参数部分线性模型下, 当非参数函数满足几种形状约束条件时, 对感兴趣的参数分量和约束非参数分量进行估计等统计推断。

1.2 形状约束条件下回归模型的估计方法

对非参数模型 (1-1), 当回归函数是单调函数时, Ayer 等 (1955)[16] 提出 PAVA 算法 (Pool adjacent violators algorithm)。Friedmant 和 Tibshirani(1984)[17] 和 Mukerjee(1988)[18] 分别给出单调函数的非参数光滑估计。Ramsay(1988)[19] 和 Wang 和 Li(2008)[20] 使用光滑样条方法给出单调函数的估计。Mammen 等 (2001)[21] 给出单调函数估计的投影方法。Geng 和 Shi(1990)[22] 给出在伞型序约束条件下的保序回归。Shi(1988)[23] 提出伞型序约束下正态均值齐性的似然比检验。Sun 和 Zhang(2013)[24] 讨论了基于核估计的具有测量误差的半参数可加模型的保序回归方法。Ding 和 Zhang(2014)[25] 讨论了半参数部分线性模型的基于单调多项式的估计方法。Du 等 (2013)[26] 考虑了半参数模型在单调约束条件下的 M- 估计。梁宝生等 (2013)[27] 研究了半参数单调变系数部分线性 EV (Error-in-

Variable) 模型的估计问题。Ding(2017)[28] 研究了函数型数据部分线性模型的单调估计。

当回归函数为凹 (或凸) 函数时, Hildreth(1954)[29] 首次使用约束最小二乘估计凹函数。Fraser 和 Massam(1989)[30] 和 Wu(1982)[31] 给出估计凹函数的有效算法。Ait-Sahalia 和 Duarte(2003)[32] 在形状约束下给出期权定价模型的递减凸函数估计。对于形状约束条件下的大样本性质, Hanson 和 Pledger(1976)[33] 给出凹函数估计的相合性。Mammen(1991)[34] 和 Groeneboom 等 (2001)[35] 给出凸函数估计的收敛速度。然而, Birke 和 Dette(2007)[36] 指出, 基于约束最小二乘和投影方法的估计可能损失光滑度, 即使已知回归函数是光滑函数。特别, PAVA 算法得到的估计经常是阶梯函数。另外, 其中一些估计方法计算困难, 尤其当样本量较大时。例如, Dykstra(1983)[37] 和 Han(1988)[38] 使用迭代循环计算最小二乘凸估计。而且, 当样本量非常大时 ($n > 10000$) 时, 一些算法非常费时。另一个主要的缺点是有些估计仅在观测值点满足给定的形状约束, 在未观测的支撑点不一定满足约束条件。因此当预测变量远离观测数据时, 想要的形状约束可能不满足, 如 Hildreth(1954)[29] 提出的约束最小二乘凹估计, Villalobos 和 Wahba(1987)[39] 提出的仅在格点上强加约束的二元样条光滑方法。近年来, 统计工作者提出同时结合形状约束与光滑过程的方法, 如基于核估计的方法有 Mamman(1991)[40] 和 Hall 和 Huang(2001)[41] 和 Dette 等 (2006)[42]。基于样条的估计方法有 He 和 Shi(1998)[43], Pal 等 (2007)[44] 和 Meyer(2008)[45]。基于多项式基的估计方法有 Curtis 和 Ghosh(2011)[46], Chang 等 (2007)[47] 和 Ding 和 Zhang(2015)[48]。基于重排算子的两步估计方法, 如 Birke 和 Dette (2007)[36]。

- **核估计**: 设 $(Z_1, Y_1), \cdots, (Z_n, Y_n)$ 是取自模型 (1-1) 的独立同分布样本,

要估计条件均值 $\psi(z)=E(Y|Z=z)$, 无约束核估计 [1−2] 表示为:

$$\hat{\psi}(z)=n^{-1}\sum_{i=1}^{n}A_i(z)Y_i \qquad (1-3)$$

其中权函数 A_i 与 Z_i 有关, $A_i(z)=h^{-1}K\{(z-Z_i)/h\}$, K 为核函数, 通常取有界对称且具有紧支撑的概率密度函数, h 是窗宽。假设函数 ψ 在区间 $[0,1]$ 上单调递增, 由于随机误差, 无约束估计 $\hat{\psi}(z)$ 在 $[0,1]$ 上不一定满足单调约束条件 $\hat{\psi}'(z)\geqslant 0$, 因此, Hall 和 Tuang(2002)[41] 推广 (1-3) 式为:

$$\hat{\psi}(z|\boldsymbol{p})=\sum_{i=1}^{n}p_iA_i(z)Y_i \qquad (1-4)$$

其中 $\boldsymbol{p}=(p_1,\cdots,p_n)$ 是定义在集合 $\{X_1,\cdots,X_n\}$ 上的概率分布。在条件 $\hat{\psi}'(z|\boldsymbol{p})\geqslant 0$ 下, 选择 $\boldsymbol{p}=\hat{p}$ 使得 $\boldsymbol{p}$ 到均匀分布: $\boldsymbol{p}_{\text{unif}}=(1/n,\cdots,1/n)$ 的距离 $D(\boldsymbol{p})$ 最小。更一般地, 可要求 $\hat{\psi}'(z|\boldsymbol{p})\geqslant\varepsilon$, ε 为给定的正数。由于 $\boldsymbol{p}$ 是概率测度, 需要满足条件 $\sum_i p_i=1$, 且 $\lim p_i\geqslant 0$。距离 $D(\boldsymbol{p})$ 的选择可参见 Cressie 和 Read(1984)[49]。核估计简单易于理解, 但计算方法比较困难, 可利用 Matlab 二次规划程序如 NAG library 中的函数 E04UCF 求解。

重排方法: 设 $\mathscr{Z}$ 是紧区间, 不失一般性, 假设 $\mathscr{Z}=[0,1]$, 可测函数 $\psi(z)$ 的定义域为 $\mathscr{Z}$, 值域为 $\mathscr{H}$, $\mathscr{H}$ 为 $\mathbb{R}$ 的有界子集。当 Z 服从区间 [0,1] 上的均匀分布时, $F_\psi(y)=\int_{\mathscr{Z}}I\{\psi(u)\leqslant y\}du$ 表示 $\psi(Z)$ 的分布函数。令

$$\psi^*(z)=Q_\psi(z)=\inf\{y\in\mathbb{R}:F_\psi(y)\geqslant z\}$$

为 $F_\psi(y)$ 的分位数函数。因此,

$$\psi^*(z)=\inf\{y\in\mathbb{R}:[\int_{\mathscr{Z}}I\{\psi(u)\leqslant y\}du]\geqslant z\} \qquad (1-5)$$

函数 ψ^* 称为函数 ψ 的递增重排 (rearrangement)(Chernozhukov 2007[50])。重排算子把函数 ψ 变换到它的分位数函数 ψ^*, 即当 $Z\sim U(0,1)$ 时,

$z \mapsto \psi^*(z)$ 是随机变量 $\psi(Z)$ 的分位数函数。设 ψ_0 为模型 (1-1) 的真实函数, 且单调递增, $\hat{\psi}$ 是 ψ_0 的初始估计, 可以是核估计、局部多项式估计或样条估计。由于随机误差, 初始估计不一定满足单调性, 对估计 $\hat{\psi}$ 进行重排运算, 可得到单调约束估计 $\hat{\psi}^*$, 当 $\hat{\psi}$ 不是单调估计时, 重排得到的估计 $\hat{\psi}^*$ 在 L_p 范数下具有更小的估计误差 [50]。重排方法虽然估计误差较小, 但在已知函数是光滑函数时可能损失光滑度。

Seive 估计方法: (i) 样条: 设 $(Z_1, Y_1), \cdots, (Z_n, Y_n)$ 是来自模型 (1-1) 的样本观测值, 为简单起见, 不妨设 $Z_1 \leqslant \cdots \leqslant Z_n$。对于 k 阶回归样条, 选择 l 个格点 $t_{k+1}, \cdots, t_{k+l}$, 定义节点为 $Z_1 = t_1 = \cdots = t_k < \cdots < t_{l+k+1} = \cdots = t_{l+2k} = Z_n$, 则次数为 $k-1$ 的分段多项式样条空间的 $m = l + k$ 个基函数可定义为 $\delta_1(z), \cdots, \delta_m(z)$, 基函数可选择为 B- 样条基或加号基, 其具体构造方法可参见 De Boor(2001)[51]。如果 $\boldsymbol{B}$ 是 $n \times m$ 设计矩阵, 其第 j 列是向量 $(\delta_j(Z_1), \cdots, \delta_j(Z_n))^T$, 令 $\tilde{b} = (\boldsymbol{B}^T\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{Y}$, 则

$$\hat{\psi}(z) = \sum_{j=1}^{m} \tilde{b}_j \delta_j(z) \tag{1-6}$$

是无约束回归样条估计。无约束估计往往不满足想要的形状约束。He 和 Shi(1998)[43] 对 B- 样条基的系数向量施加不等式约束: $b_1 \leqslant \cdots \leqslant b_m$ 从而得到单调 B- 样条。Lu(2011)[52] 在半参数模型中利用此单调 B- 样条来逼近非参数函数。Ramsay(1988)[19] 给出单调样条基的递归构造方法。Meyer(2012)[53] 通过对二次 B- 样条在节点处的导数限制为非负来获得单调样条, 对三次 B- 样条在节点处的二阶导数限制为非负来得到凸样条。然而, 样条对节点的个数和位置较为敏感, 另外, 对三次或更高次的单调样条, 约束不再是线性约束 [53], 这使得计算或推断更加困难。

(ii) Bernstein 多项式: 设 ψ 为定义于 [0,1] 上的函数, 与之相关的 Bern-

stein 多项式定义为

$$B_N(\psi;z)=\sum_{k=0}^{N}\psi(\frac{k}{N})\binom{N}{k}z^k(1-z)^{N-k},\ N=1,\cdots \qquad (1-7)$$

这样对任一函数 ψ, 可得到一个 Bernstein 多项式序列。如果函数 ψ 在 [0,1] 上连续, 则由 Weierstrass 定理 (Lorentz 1986[54]), 当 $N\longrightarrow\infty$ 时, $B_N(\psi;z)$ 在 [0,1] 上一致收敛于 ψ, 即 $B_N(\psi;z)\to\psi(z),\forall z$。

设 $B_N(z)=\sum_{k=0}^{N}\alpha_i\binom{N}{k}z^k(1-z)^{N-k}$ 是 N 阶 Bernstein 多项式, 如果 $\alpha_0\leqslant\alpha_1\leqslant\cdots\leqslant\alpha_N$, 则对任一 $z\in[0,1]$, $B'_N(z)\geqslant 0$, 因此 Bernstein 多项式是单调的。如果对每一 $k=1,\cdots,N-1$, $\alpha_{k+1}+\alpha_{k-1}\leqslant 2\alpha_k$, 则对任一 $z\in[0,1]$, $B''_N(z)\leqslant 0$, Bernstein 多项式是凹函数。

Bernstein 多项式由于在所有多项式中具有最优的形状约束性质, 例如, 若函数 $\psi(\cdot)$ 是凸函数, 则 $(1-7)$ 式的 Bernstein 多项式也为凸函数 (Car- nicer 和 Pena 1993[55]), 并且它的导数有同样的收敛性质 (Lorentz 1986[54])。Bernstein 多项式估计适用于各种不同的形状约束, 例如非负性, 单调性, 凸性, 增凹性, 以及它们的各种变体及组合。Bernstein 多项式估计不仅只在观测点满足所要求的形状约束, 而在所有点上都满足所要求的形状约束。因此, 使用 Bernstein 多项式, 我们能够把相当广泛的形状约束问题转换为仅具有线性约束的最小二乘问题。Bernstein 多项式估计已被应用于非参数单调曲线估计, 例如, Chak 等 (2005)[56] 给出半参数模型的 Bernstein 多项式估计方法; Chang 等 (2007)[47] 给出单调和凸回归函数的 Bayes 估计方法; Curtis 和 Ghosh(2011)[46] 提出保序回归模型的变量选择方法。Ding 和 Zhang(2016)[57] 给出非参数回归模型的 Bayes 估计方法。Ding 和 Zhang(2016)[58] 给出非参数混合效应模型的 Bayes 估计方法。Petrone (1999)[59] 给出随机 Bernstein 多项式和 Stadtmuller(1986)[60] 等给出非参数估计的渐近性质。Ding 和 Zhang(2017)[61] 基于 Bernstein 多项式