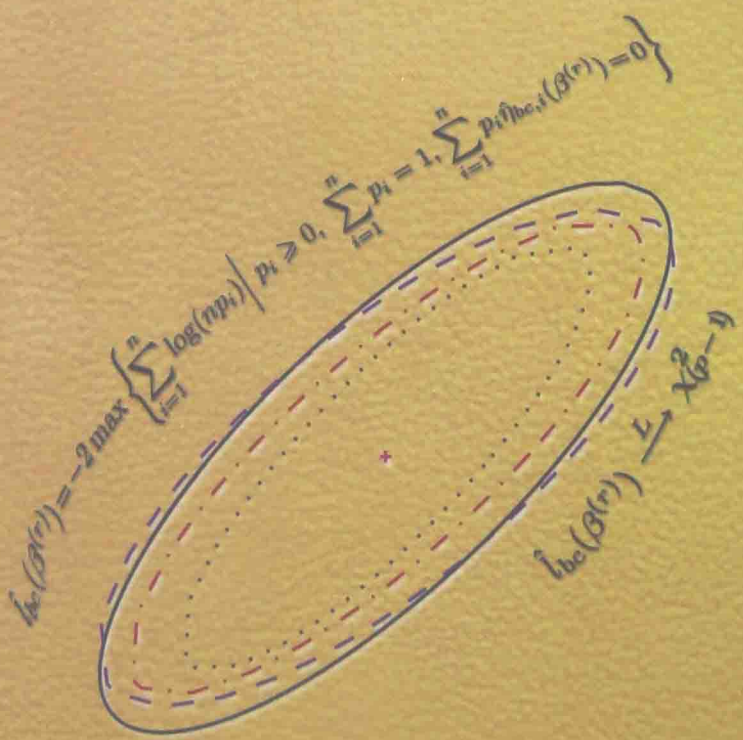
纵向数据半参数模型

李高荣　杨宜平 著



科学出版社

纵向数据半参数模型

李高荣　杨宜平　著

科学出版社

北京

内 容 简 介

纵向数据半参数模型目前是统计学和计量经济学研究的热门研究课题之一，并在生物学、医学、传染病学、经济学、金融学和遥感等领域有着广泛的应用. 本书共分 8 章，主要针对几种纵向数据半参数回归模型，重点阐述这些模型的估计方法、统计推断及渐近结果. 本书除介绍这些模型的发展动态，还特别详细介绍了一些最新研究成果，使读者对纵向数据统计模型的方法和统计思想有一个较为全面的了解，并起到抛砖引玉的作用.

本书适合理工院校数理统计专业、数学专业和计量经济学专业的高年级大学生、研究生、教师和一般科学技术人员使用. 另外，本书还可供各行各业应用统计科学工作者和医学、生物学、经济学、金融学、社会学、心理学和工业工程等专业人士参阅.

图书在版编目(CIP)数据

纵向数据半参数模型/纵高荣，杨宜平著. —北京：科学出版社，2015.2

ISBN 978-7-03-043419-7

Ⅰ.①纵… Ⅱ.①李…②杨… Ⅲ.①数据模型-半参数模型 Ⅳ.①O211.3

中国版本图书馆 CIP 数据核字(2015) 第 033295 号

责任编辑：王丽平 / 责任校对：邹慧卿

责任印制：张 倩 / 封面设计：陈 敬

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

文林印务有限公司 印刷

科学出版社发行 各地新华书店经销

*

2015 年 2 月第 一 版 开本：720 × 1000 1/16

2015 年 2 月第一次印刷 印张：18 3/4

字数：430 000

定价：108.00 元

(如有印装质量问题，我社负责调换)

前　言

纵向数据广泛应用于医学、生物学、社会学、金融学和经济学等诸多领域. 在社会学和经济学中, 纵向数据又称为面板数据. 纵向数据之所以得到如此广泛的应用, 是由于纵向数据是同一个个体按时间顺序测量得到的数据, 它将截面数据与时间序列数据结合到一起, 能很好地分析出个体随时间变化的趋势, 反映个体间的差异和个体内部的变化, 起到截面数据或者时间序列数据模型不可替代的作用. 但是纵向数据这种 "个体间独立、个体内相关" 的特点给统计分析提出一定的挑战. 统计学家都在寻找和发展纵向数据统计模型估计和统计推断的有效方法.

半参数回归模型是 20 世纪 80 年代发展起来的一种重要统计模型, 包含部分线性模型、单指标模型、部分线性单指标模型、变系数模型和可加模型等. 该类模型既含有参数分量, 又含有非参数分量, 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. 在理论上, 处理这类模型的方法融合了参数回归模型中常用的方法和近年发展起来的非参数方法, 但也并非这两类方法的简单叠加. 总之, 可以认为其复杂性和难度都超过了单一性质的回归模型. 而纵向数据的半参数建模和统计分析已经成为统计学和计量经济学研究的热门研究课题之一, 并在生物学、医学、传染病学、经济学、金融学和遥感等领域有着广泛的应用.

全书分为 8 章. 第 1 章绪论, 介绍纵向数据的定义、例子和纵向数据分析遇到的一些困难之处, 并对半参数回归模型进行简要的介绍. 第 2 章从指数族出发, 介绍广义线性模型的建立, 进一步推广到纵向数据广义线性模型的边际建模, 以及分析纵向数据广义线性模型的广义估计方程 (GEE) 方法、二次推断函数 (QIF) 方法、广义经验似然方法和惩罚变量选择方法, 重点介绍一种新的变量选择方法 —— 光滑门限广义估计方程变量选择方法. 第 3 章针对纵向数据部分线性模型, 首先介绍一些代表性研究成果, 然后对作者和其合作者近几年的研究成果进行详细的介绍, 如广义经验似然方法、测量误差模型修正的 QIF 方法和惩罚变量选择方法. 第 4 章介绍纵向数据单指标模型的四种经验似然方法, 并构造了指标参数向量的经验似然置信域. 第 5 章继续讨论单指标模型, 分别在独立数据和纵向数据下给出单指标模型中联系函数同时置信带的构造问题. 第 6 章对纵向数据部分线性单指标模型, 发展纠偏的经验似然方法、纠偏的 GEE 方法、纠偏的 QIF 方法, 以及纠偏的 GEE 和 QIF 变量选择方法. 第 7 章针对纵向数据变系数模型, 除了介绍已有的研究成果, 重点介绍纵向数据测量误差变系数模型的经验似然统计推断. 第 8 章对面板数据非参数和部分线性固定效应模型提出消除固定效应的方法, 并讨论模型中非参数

函数同时置信带的构造问题.

本书第 2, 4, 5, 6, 8 章内容由李高荣负责撰写和整理, 第 1, 3, 7 章内容由重庆工商大学的杨宜平博士负责整理, 全书最后由李高荣定稿. 由于本书的时间比较仓促, 特别感谢杨宜平博士在本书撰写过程中既要照顾小孩, 又要完成本书的撰写和整理工作, 使得本书得以按计划完成并交稿.

在本书出版和修改过程中, 得到了许多老师、同事和研究生的帮助, 借本书出版之际, 对他们表示衷心的感谢. 首先特别感谢我的导师薛留根教授及香港浸会大学朱力行教授多年以来对我的关心和支持, 感谢他们在我博士研究生和博士后期间引导我进入统计学这个充满魅力的领域, 让我学到了许多统计思想和统计学的理论和方法. 也特别感谢合作者山东大学林路教授、香港浸会大学彭衡副教授和童铁军副教授、新加坡南洋理工大学练恒博士、南京信息工程大学来鹏博士、深圳大学张君博士、天津商业大学杨随根博士和师弟冯三营博士, 在与他们合作过程中, 我受益匪浅. 感谢北京工业大学的杨振海教授、王松桂教授、张忠占教授、程维虎教授、李寿梅教授、陈立萍副教授和谢琍副教授等各位老师多年来给予我的大力支持和帮助. 同时感谢硕士生杨秀娟和李海斌对本书的帮助, 并提出一些有益的建议.

本书的完成得到了国家自然科学基金 (11101014, 11471029, 11301569)、北京市自然科学基金 (1142002) 和北京市教育委员会科技计划面上项目 (KM201410005010) 的支持, 还特别得到了北京工业大学京华人才支持计划项目对本书的资助, 作者谨在此表示感谢.

最后感谢我的爱人和女儿李睿菡多年来对我科研工作的支持和理解, 特别是在本书撰写过程中她们在背后默默的支持和付出让我有足够的时间来完成本书. 在本书出版过程中, 得到了科学出版社领导和王丽平等编辑的支持和帮助, 在此一并表示感谢.

由于作者知识有限, 书中内容定有许多不足之处, 还恳请各位专家和同行及实际应用者多提宝贵意见, 敬请有关专家与广大读者批评指正.

李高荣

2014 年 8 月于北京工业大学

符 号 表

$\mathbb{R}$	实数集合
$\mathbb{R}^p$	p 维 Euclidean 空间
$\|\cdot\|$	Euclidean 模
i.i.d.	独立同分布
$=:$	“定义为” 或 “记为”
T	向量或矩阵的转置
$\mathrm{tr}(A)$	矩阵 A 的迹
$A \otimes B$	矩阵 A 和 B 的 Kronecker 乘积
$A^{\otimes 2}$	AA^{T}
$\mathrm{mineig}(A)$	矩阵 A 的最小特征值
$\mathrm{diag}(a_1, \cdots, a_n)$	由元素 $a_1, \cdots, a_n$ 组成的对角矩阵
$\xrightarrow{L}$	依分布收敛
$\xrightarrow{P}$	依概率收敛
a.s.	几乎处处收敛, 或者依概率 1 强收敛
$y = O(1)$	y 是有界变量
$y = o(1)$	y 是无穷小量
$\xi_n = o_P(\eta_n)$	对任一 $\varepsilon > 0$, 有 $P\{\|\xi_n\| \geqslant \varepsilon \|\eta_n\|\} \to 0$
$\xi_n = o_P(1)$	ξ_n 依概率收敛到 0
$\xi_n = O_P(1)$	ξ_n 依概率有界
$f'(x)$	函数 $f(x)$ 的一阶导数
$K(\cdot)$	核函数
h和h_n	窗宽
$N(\mu, \Sigma)$	均值为 μ, 协方差阵为 Σ 的正态分布
χ_p^2	自由度为 p 的 χ^2 分布
$\|\cdot\|_p$	p 范数
$\mathrm{sgn}(t) = I(t > 0) - I(t < 0)$	符号函数
x_+	变量 x 的正部
$\|g\|_\infty$	对于任意的函数 $g(u)$, 定义 $\|g\|_\infty = \sup_{u \in [0,1]} \|g(u)\|$
$\|A\|_\infty$	对任意的矩阵 $A(u) = (a_{ij}(u))_p$, 定义 $\|A\|_\infty = \left(\sum_{i=1}^{p} \sum_{j=1}^{p} \|a_{ij}\|_\infty^2 \right)^{1/2}$

目　录

第1章 绪 论

1.1 纵向数据

1.1.1 纵向数据介绍及例子

纵向数据 (longitudinal data) 是指对同一组受试个体或者受试单元在不同时间点上重复观测若干次, 得到的由截面和时间序列融合在一起的数据 (Diggle et al., 2002).

纵向数据在实际中的例子很多, 广泛应用于医学、生物学、社会学、经济学和金融学等诸多领域, 反映了个体间的差异和个体内部的变化. 纵向数据综合了截面数据和时间序列数据的特点和优点, 同时随着计算机性能的飞速发展, 使得纵向数据的统计分析研究越来越受到人们的重视. 例如, 如果要研究儿童阅读能力随时间变化趋势的问题, 可以随机抽取一些儿童, 在不同年龄段对其阅读能力进行测试, 这样得到的数据就是纵向数据. 这些儿童的阅读能力, 随着年龄的增长均有提高, 但是每个儿童在进行观测时的初始阅读能力却不一样, 有些儿童在年龄较小时的阅读能力反而比有些年龄较大的儿童阅读能力要强. 也就是说, 纵向数据模型既考虑了个体间的差异 (初始的阅读能力不同), 也考虑了个体内部的变化 (阅读能力随着年龄的增长而提高). 这个例子也反映了纵向数据最大的特点: 对不同个体观测所得到的数据是独立的, 但是对同一个体观测所得到的数据往往具有相关性. 如果对此研究采用截面数据的方法进行分析, 就忽略了儿童的初始阅读能力, 从而使得分析出的结果违背了实际情况. 所以, 纵向数据是同一个体按时间顺序观测得到的, 它将截面数据和时间序列数据结合在一起, 能很好地分析出个体随时间变化的趋势, 反映了个体间的差异和个体内部的变化. 对比仅利用截面数据或者时间序列数据模型, 纵向数据模型有不可替代的作用, 有很高的应用价值. 同时随着计算机性能的飞速发展, 纵向数据的统计分析研究也越来越受到人们的重视.

首先介绍如下四个纵向数据的例子, 在本书中将会对这几个例子进行分析.

例 1.1.1(多中心艾滋病群组研究) Kaslow 等 (1987) 公布了一组来自于多中心艾滋病群组研究的数据. 该研究是计划在 1984~1991 年, 对 283 位 HIV(human immunodeficiency virus) 呈阳性的同性恋患者每半年进行一次定期检查, 记录他们看病的医院地址和感染的情况. 但是由于部分患者没有定期来检查或者因病情发作而不到半年就需要检查一次, 每位患者重复测量的次数不同. 每位患者在这 8 年

内至少检查过 1 次, 最多检查过 14 次. 对于这组数据, 响应变量是 HIV 感染后患者血液内所含 CD4 细胞的比例, 协变量是患者的年龄、吸烟状况、HIV 感染前 CD4 细胞的比例及其交互作用. 大家感兴趣的问题是, 如何识别出真正对 HIV 感染后血液内 CD4 细胞比例的变化有影响的协变量, 以及进一步了解它们分别产生了怎样不同的影响.

例 1.1.2 (多发性硬化症临床试验)　多发性硬化症临床试验的数据集最初被 Petkau 等 (2004), Petkau 和 White (2003) 分析过, 并且在 Song (2007) 的专著中也被多次分析. 该实际数据集涉及一个纵向的临床试验, 用来评价复发缓解多发性硬化症 (MS) 中的干扰素 β-1b(IFNB) 的中和抗体的影响, 它是一种可破坏包围神经的髓鞘的疾病. 该数据集是来自英国哥伦比亚大学承担的 Betaseron 临床试验的磁共振成像 (MRI) 研究的子课题, 涉及 50 个复发缓解多发性硬化症患者, 每个患者每隔 6 周来大学进行一次治疗. 对于 17 个预定的治疗访问周期, 该数据集对每个患者包含 3 个响应变量, 分别是: ① 主动扫描 (active scan), 是一个二元响应变量, 如果上次进行基线扫描后本次治疗进行了扫描, 记录为 1, 否则为 0; ② 病情恶化情况 (exacerbation), 也是一个二元响应变量, 即指进行 MRI 扫描检查是否出现病情加重的情况, 病情加重用 1 表示, 否则用 0 来表示; ③ 疾病负担 (burden of disease), 一个正的连续型响应变量, 表示每次扫描后所有切片上 MS 病变的总面积 (单位: mm^2). 本数据记录了 7 个协变量或解释变量: 治疗 (Trt)、时间 (T, 单位: 周)、时间的平方、年龄 (Age)、性别 (Gender)、患病的持续年限 (Dur, 单位: 年) 和一个额外的基线协变量扩大残疾状态等级 (EDSS) 评分. 50 个患者被随机分成 3 个治疗组, 具体分配为 17 个患者服用安慰剂 (placebo) 进行治疗、17 个患者服用低剂量 (low dosage) 药剂治疗, 还有 16 个患者服用高剂量 (high dosage) 药剂治疗. 该数据集中不仅存在缺失数据, 而且为非平衡纵向数据. MS 临床试验的主要目的是研究药物治疗对减轻疾病症状的影响.

例 1.1.3 (癫痫病发作数据)　这是一个临床随机对照试验, 通过将一种新研发的抗癫痫的药物与能降低癫痫病发作频率的安慰剂进行比较, 来考察该新研发药物的疗效, 见参考文献 Thall 和 Vail (1990), Wang 等 (2005b). 研究者将新药和安慰剂随机的分给 59 位患者服用, 其中 28 个患者服用安慰剂, 31 个患者服用新研发的抗癫痫药物. 在接下来的 8 周内, 每两周对患者进行一次定期检查, 记录在这两周内癫痫发作的次数 (表 1.1.1 中 Y_1, Y_2, Y_3, Y_4). 同时, 在进入试验之初, 研究者会记录每位患者的基本情况, 包括年龄 (Age)、进入试验初期未服药前癫痫的发作次数 (表 1.1.1 中 Base)、试验中服用的药物 (表 1.1.1 中 Trt, 其中 0 表示服用安慰剂, 1 表示服用新药) 等. 对于这组数据, 响应变量是患者每两周的发病次数, 协变量是基于患者的基本情况得到的各种指标, 包括年龄的对数和基准癫痫病数 (除以 4 后取对数). 对于该数据的研究, 大家非常关心的一个科学问题是药物是否有助于减

少癫痫发作率. 对该问题的研究可参考文献 Thall 和 Vail (1990), Wang 等 (2005b), Bai 等 (2009), Pang 和 Xue (2012), Yang 等 (2014c).

表 1.1.1 癫痫病发作数据

ID	Y_1	Y_2	Y_3	Y_4	Trt	Base	Age	ID	Y_1	Y_2	Y_3	Y_4	Trt	Base	Age
104	5	3	3	3	0	11	31	103	0	4	3	0	1	19	20
106	3	5	3	3	0	11	30	108	3	6	1	3	1	10	30
107	2	4	0	5	0	6	25	110	2	6	7	4	1	19	18
114	4	4	1	4	0	8	36	111	4	3	1	3	1	24	24
116	7	18	9	21	0	66	22	112	22	17	19	16	1	31	30
118	5	2	8	7	0	27	29	113	5	4	7	4	1	14	35
123	6	4	0	2	0	12	31	117	2	4	0	4	1	11	27
126	40	20	23	12	0	52	42	121	3	7	7	7	1	67	20
130	5	6	6	5	0	23	37	122	4	18	2	5	1	41	22
135	14	13	6	0	0	10	28	124	2	1	1	0	1	7	28
141	26	12	6	22	0	52	36	128	0	2	4	0	1	22	23
145	12	6	8	4	0	33	24	129	5	4	0	3	1	13	40
201	4	4	6	2	0	18	23	137	11	14	25	15	1	46	33
202	7	9	12	14	0	42	36	139	10	5	3	8	1	36	21
205	16	24	10	9	0	87	26	143	19	7	6	7	1	38	35
206	11	0	0	5	0	50	26	147	1	1	2	3	1	7	25
210	0	0	3	3	0	18	28	203	6	10	8	8	1	36	26
213	37	29	28	29	0	111	31	204	2	1	0	0	1	11	25
215	3	5	2	5	0	18	32	207	102	65	72	63	1	151	22
217	3	0	6	7	0	20	21	208	4	3	2	4	1	22	32
219	3	4	3	4	0	12	29	209	8	6	5	7	1	41	25
220	3	4	3	4	0	9	21	211	1	3	1	5	1	32	35
222	2	3	3	5	0	17	32	214	18	11	28	13	1	56	21
226	8	12	2	8	0	28	25	218	6	3	4	0	1	24	41
227	18	24	76	25	0	55	30	221	3	5	4	3	1	16	32
230	2	1	2	1	0	9	40	225	1	25	19	8	1	22	26
234	3	1	4	2	0	10	19	228	2	3	0	1	1	25	21
238	13	15	13	12	0	47	22	232	0	0	0	0	1	13	36
101	11	14	9	8	1	76	18	236	1	4	3	2	1	12	37
102	8	7	9	4	1	38	32								

例 1.1.4(荷尔蒙纵向数据) 纵向荷尔蒙数据是收集了 34 个健康妇女在一个月经周期的尿样, 每隔一天试验尿的孕激素. 在 34 个参与者中, 每个妇女按时提供 11~28 次观测, 共得到 492 个观测值, 平均每个妇女进行 14.5 次观测. He 等 (2002) 与薛留根和朱力行 (2007) 对该荷尔蒙纵向数据利用部分线性模型进行拟合, 他们

考虑响应变量为孕激素值的对数, 两个协变量分别为年龄 (Age) 和体重指数 (BMI).

从上面 4 个例子中, 可以看出纵向数据是同一个体在不同时刻的多次重复观察而得到的数据集, 对于每个个体, 都得到一个变量集. 但是, 它又不同于一般意义上的多元统计数据. 在多元统计分析中, 每一个个体也得到一个变量, 但是这个变量是同一个体多个指标的一次观察得到的向量, 并无重复的含义. 因此纵向数据一个显著的特点是 "个体间独立、个体内相关", 有的文献中也称为 "组间独立、组内相关". 对于这些纵向数据分析最大的挑战就是需要考虑同一观测个体的不同次观测之间的相关性.

对比截面数据的研究, Song (2007) 指出纵向数据的研究具有以下 3 个方面的挑战:

(1) 由于纵向数据的概率机制非常复杂, 并很难表示出来, 所以纵向数据分析是一个非常具有挑战的问题. 在大部分情况下, 纵向数据的极大似然推断要么不存在, 要么太复杂而使得数值计算很难实施. 为了解决这个困难, Liang 和 Zeger (1986) 提出了分析纵向数据非常流行的广义估计方程 (generalized estimating equations, GEE) 方法, GEE 方法不要求指定数据的概率模型, 是拟似然方法的一种推广 (详见第 2 章的讨论), 且 GEE 方法仅仅要求指定数据的一阶矩和二阶矩, 并把纵向数据中的组内相关参数作为讨厌参数;

(2) 纵向数据中常存在缺失数据, 这也使得纵向数据分析变得非常困难. 主要原因是纵向数据中的缺失模式比截面数据中的更加复杂. 例如, 在截面数据中, 每个个体只有一个样本点, 如果这个数据点缺失, 在数据分析时把这个个体删掉就可以了. 但对于纵向数据, 在一个时间点上的数据缺失并不意味着整个个体就完全没有信息, 因为在其他时间点上仍然有测量数据被记录. 进一步, 对于纵向数据中缺失情况时遇到的缺失机制的表示和组内相关结构等问题, 给统计分析也提出了许多新的机遇和挑战;

(3) 当纵向数据时间序列的长度很大时, 纵向数据的建模模式或回归分析等成为统计分析的一个主要任务. 在目前文献中, 大部分纵向数据的文献都是集中在重复测量的次数有限的情形, 而当重复次数趋于无穷大时, 在这种情况下, 如果纵向数据的组内相关结构不再是讨厌参数时, 发展相应的统计推断方法也成为纵向数据分析的一个具有挑战的任务.

1.1.2　纵向数据的表示

考虑来自 n 个个体的数据, 其中第 i $(i = 1, \cdots, n)$ 个个体有 m_i 次观测, 总的观测次数为 $N = \sum_{i=1}^{n} m_i$. 设 Y_{ij} 和 (X_{ij}, t_{ij}) 分别表示对第 i 个个体进行第 j 次观测 $(j = 1, \cdots, m_i)$ 所得到的响应变量和协变量的观测值, 这里 $X_{ij} =$

$(x_{ij,1},\cdots,x_{ij,p})^{\mathrm{T}}\in\mathbb{R}^p$, t_{ij} 表示观测时间. 在更一般的集合中, t_{ij} 不一定表示时间, 但一定是模型中非参数部分依赖于时间的协变量. 所有的观测数据构成一个纵向数据集, 表示为

$$\{(X_{ij},Y_{ij},t_{ij}),i=1,\cdots,n,j=1,\cdots,m_i\}.$$

该纵向数据集合如表 1.1.2 所示.

表 1.1.2 纵向数据表示表格

个体	重复测量	响应变量	协变量			
1	1	Y_{11}	$x_{11,1}$	$\cdots$	$x_{11,p}$	t_{11}
1	2	Y_{12}	$x_{12,1}$	$\cdots$	$x_{12,p}$	t_{12}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
1	m_1	Y_{1m_1}	$x_{1m_1,1}$	$\cdots$	$x_{1m_1,p}$	t_{1m_1}
2	1	Y_{21}	$x_{21,1}$	$\cdots$	$x_{21,p}$	t_{21}
2	2	Y_{22}	$x_{22,1}$	$\cdots$	$x_{22,p}$	t_{22}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
2	m_2	Y_{2m_2}	$x_{2m_2,1}$	$\cdots$	$x_{2m_2,p}$	t_{2m_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
n	1	Y_{n1}	$x_{n1,1}$	$\cdots$	$x_{n1,p}$	t_{n1}
n	2	Y_{n2}	$x_{n2,1}$	$\cdots$	$x_{n2,p}$	t_{n2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
n	m_n	Y_{nm_n}	$x_{nm_n,1}$	$\cdots$	$x_{nm_n,p}$	t_{nm_n}

1.2 半参数模型

半参数回归模型是 20 世纪 80 年代发展起来的一种重要统计模型, 此模型介于参数回归模型和非参数回归模型之间. 在不少实际问题中, 要考察对象 Y(响应变量) 同影响 Y 的因素 X(解释变量或协变量) 之间的关系. 传统的线性模型当假设模型成立时, 其推断有较高的精度, 但当参数假定与实际背离时, 其拟合情况就很差. 若用非参数模型去处理, 则有可能会丢失有经验或历史资料得到的信息, 因而采用两者的混合, 即采用半参数回归模型. 这种模型既有参数分量, 又含有非参数分量. 在理论上, 处理这种模型的方法融合了参数回归模型中常用的方法和较近发展起来的非参数方法, 但并非这两类方法的简单叠加. 总之, 可以认为其复杂性和难度都超过了单一性质的回归模型. 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. 例如, 在生物学、医学、传染病学、经济学、金融学和遥感等领域有着广泛的应用.

半参数回归模型发展至今, 在解决实际问题中, 实际工作者和学者们提出了许多类型的半参数回归模型, 下面就涉及的几种半参数模型进行简要介绍.

1.2.1　非参数模型

假设 Y 为响应变量, X 为影响 Y 的协变量, 则非参数回归模型的形式为

$$Y = g(X) + \varepsilon, \tag{1.2.1}$$

其中 $g(x) = E(Y|X = x)$ 为未知的回归函数, ε 为模型误差, 且满足 $E(\varepsilon|X) = 0$.

非参数回归模型的优点是回归函数 $g(\cdot)$ 的任意形式, 而且模型的假设少, 可以很好地拟合实际数据. 但非参数回归模型的缺点是当 $X \in \mathbb{R}^p$, 且 X 的维数 p 较高时, 对非参数模型进行估计和统计推断会遇到所谓的 "维数灾祸" 问题. 在第 8 章讨论了面板数据非参数固定效应模型的同时置信带的构造问题. 非参数回归模型经常考虑 $p = 1$ 或 $p = 2$ 的情形, 即一元或者二元回归模型. 对于协变量更高维的情形, 即 $p \geqslant 3$ 时, 且协变量为 $X = (X_1, \cdots, X_p)^{\mathrm{T}}$, 考虑如下的线性模型

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon.$$

这时回归函数变为 $g(x) = E(Y|X = x) = \beta_0 + x_1\beta_1 + \cdots + x_p\beta_p$, 即模型退化为经典的线性回归模型. 如果响应变量 Y 为非高斯分布, 如泊松 (Poisson) 分布、伽马 (Gamma) 分布、二项 (binomial) 分布、指数 (exponential) 分布等, 可以考虑广义线性模型, 关于广义线性模型, 第 2 章给了较为详细的介绍, 并给出了纵向数据广义线性模型的一些估计方法的介绍.

在实际应用中, 为了保留参数模型的优点及非参数回归模型数据适应性的优点, 同时避免 "维数灾祸" 问题, 统计学者提出并发展了很多半参数回归模型, 如部分线性模型、单指标模型、部分线性单指标模型和变系数模型等, 这些模型已经广泛应用到了生物医学和计量经济学等领域中.

1.2.2　部分线性模型

部分线性模型假设响应变量 Y 依赖于 p 维协变量 X 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

$$Y = X^{\mathrm{T}}\beta + g(T) + \varepsilon, \tag{1.2.2}$$

其中 $\{X^{\mathrm{T}}, T\}$ 可以是随机设计也可以是固定设计, $\beta = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ 是未知参数向量, $g(\cdot)$ 是一元未知函数, ε 是期望为 0 的模型随机误差.

自 Engle 等 (1986) 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. 由于部分线性模型结