首都经济贸易大学出版基金资助
SHOUDU JINGJI MAOYI DAXUE CHUBAN JIJIN ZIZHU

复杂数据下测量误差模型的估计理论与方法

FUZA SHUJU XIA CELIANG WUCHA MOXING DE GUJI LILUN YU FANGFA

刘 强 ◎ 著

首都经济贸易大学出版社
Capital University of Economics and Business Press

首都经济贸易大学出版基金资助

国家社科基金项目

复杂数据下测量误差模型的估计理论与方法

FUZA SHUJU XIA CELIANG WUCHA MOXING DE GUJI LILUN YU FANGFA

刘 强 ◎ 著

首都经济贸易大学出版社

Capital University of Economics and Business Press

·北 京·

图书在版编目(CIP)数据

复杂数据下测量误差模型的估计理论与方法/刘强著. —北京:首都经济贸易大学出版社,2012.5

ISBN 978-7-5638-1996-6

Ⅰ.①复… Ⅱ.①刘… Ⅲ.①测量误差—模型—估计理论—研究 Ⅳ.①P207

中国版本图书馆 CIP 数据核字(2012)第 070996 号

复杂数据下测量误差模型的估计理论与方法

刘 强 著

出版发行	首都经济贸易大学出版社
地　　址	北京市朝阳区红庙(邮编 100026)
电　　话	(010)65976483　65065761　65071505(传真)
网　　址	http://www.sjmcb.com
E-mail	publish@cueb.edu.cn
经　　销	全国新华书店
照　　排	首都经济贸易大学出版社激光照排服务部
印　　刷	北京泰锐印刷有限责任公司
开　　本	880 毫米×1230 毫米　1/32
字　　数	169 千字
印　　张	6.625
版　　次	2012 年 5 月第 1 版第 1 次印刷
书　　号	ISBN 978-7-5638-1996-6/P·2
定　　价	19.00 元

图书印装若有质量问题,本社负责调换

前　言

在经济、社会、人口、医学等众多研究领域中，人们通过各种方式收集数据，然后对数据进行统计分析，利用分析结果指导社会实践. 但是在收集数据的过程中，经济变量能够精确观测的假定在实际问题中往往不能成立. 一方面，对经济变量进行测量时，经常会受到多种客观因素的影响，导致一些偏差，如抽样误差、记录误差等. 例如，为了研究不同地区城镇居民生活中的储蓄和消费状况，往往需要通过各种途径进行数据调查，然而在调查中，由于种种因素的影响，我们得到的一般不是真实的储蓄和消费数据；研究身高和体重的关系时，测量身高体重时所得的测量值往往不是很精确，收集到的数据经常带有误差；研究某种药物对某种病症的影响疗效时，在将病症量化时也不可能做到精确；调查研究影响工资收入的因素时，某些被调查者可能因为种种原因，不愿意透露真实的工资，此时调查得到的数据也带有测量误差；等等. 另一方面，人们考察变量之间的关系时，常常只关心主要因素的影响，其他影响不大的因素之效应将体现于偏差中. 通常文献中称这种观测数据带有误差的问题为“测量误差问题”，分析这些数据的统计模型通常称为“测量误差模型”或“EV(Errors - in - variables)模型”. 大多数情形下，人们在进行数

据分析时，往往忽略这种误差，直接用带有测量误差的数据代替真实数据进行分析，当测量误差较小时，产生的结果可能与真实结果相差不大，在可以接受的范围内. 但是如果误差太大，分析结果的真实性则会受到很大的影响. 因此关于测量误差模型（EV 模型）的研究无论是在理论还是在实践上都具有极其重要的意义.

EV 模型最早是在 19 世纪 70 年代出现的，之后百余年里发展缓慢，直到 20 世纪 80 年代，人们在经济、金融、医学、生物等领域的研究中发现很多数据用普通的回归模型处理时得到的结果不太理想，究其原因是数据误差过大，无法利用传统的回归模型进行分析，因此测量误差模型得到了广泛的应用，随之关于测量误差模型（EV 模型）的理论研究开始受到重视并迅速发展. 在过去的 20 多年里，线性、非线性测量误差模型的参数估计问题已有大量的研究，未知参数置信域的构造以及模型检验问题也有了长足的发展. 随着经济、金融、医学、生物、人口、环境等学科研究中的实验技术、检验方法以及数据分析手段的日益提高，所获得的数据在结构上越来越复杂精细，数据所提供的信息也越来越繁杂，获得的变量个数越来越多，关于复杂数据的统计推断应运而生.

复杂数据主要包括删失数据、纵向数据、缺失数据、Panel 数据、污染数据、高维数据等统计数据. 复杂数据的分析与建模已经成为当今国内外统计学界和计量经济学界的一个研究热点，也是一个研究难点，已经引起国内外众多统计学家和计量经济学家的高度关注. 由于复杂数据

形式多样,存在数据的缺失、数据的删失、纵向数据等各种情况,这给模型的建立、数据的分析造成了很大的困难.相对于独立同分布数据而言,复杂数据在现实生活中更为普遍,关于复杂数据下 EV 模型的估计与检验问题更富有现实意义,但处理过程却更为棘手.也正因为如此,在该领域中还有许多的空白亟待研究.另外,误差类型以及模型的种类也在不断增加.常见的误差类型有:可加误差、乘积误差、Berkson 误差等;模型类型主要有:线性模型、非线性模型、广义线性模型、部分线性模型、混合效应模型、变系数模型、单指标模型等.基于上述原因,对于不同的数据类型、误差类型以及模型类型,需要用不同的统计方法来进行分析,因此复杂数据下各种 EV 模型(如线性 EV 模型、广义线性 EV 模型、混合效应 EV 模型以及变系数 EV 模型等)的研究还有许多具有挑战性的问题有待去研究解决.如何在纵向数据、删失数据等复杂数据情形下讨论各种 EV 模型的大样本性质是当今统计学界和计量经济学界亟待解决的热点课题之一.

关于测量误差模型的一些基本结果被总结在 Fuller 的专著《 Measurement error model》(1987)和 Carroll 等的专著《 Measurement Error in Nonlinear models》(2006)中.关于复杂数据的一些结果可以参见 Diggle 等的专著《Analysis of Longitudinal Data》(1994), Little 和 Rubin 的专著《 Statistical Analysis with Missing Data》(2002), Desu 和 Raghavarao 的专著《 Nonparametric Statistical Methods for Complete and Censored Data》(2003), Tsiatis 的专著《Semi-

parametric Theory and Missing Data》(2006), Wu 和 Zhang 的专著《Nonparametric Regression Methods for Longitudinal Data Analysis》(2006),金勇进和邵军的专著《缺失数据的统计处理》(2009),等等.

本书研究的模型主要包括:线性 EV 模型、非线性 EV 模型以及半参数 EV 模型等统计模型. 研究的主要目的是:在纵向数据、删失数据、缺失数据等复杂数据下研究几类 EV 模型中兴趣参数及兴趣函数估计的大样本问题,如估计量的渐近正态性、相合性及其收敛速度等统计性质.

本书的结构安排如下. 第 1 章是绪论,主要介绍本书涉猎的一些基本模型、基本方法和常见数据. 第 2 ~ 7 章主要介绍作者近几年对复杂数据下测量误差模型估计理论与估计方法的一些研究成果. 具体而言,第 2 章主要讨论了删失数据下线性 EV 模型的估计问题,给出了未知参数估计的渐近正态性,构造了未知参数的两种经验对数似然比统计量,证明了所构造的经验似然比统计量渐近于 χ^2 分布;第 3 章考虑了缺失数据下两类 EV 模型的估计问题,讨论了估计量的大样本性质;第 4 章考虑了纵向数据下两类 EV 模型的经验似然估计问题;第 5 章讨论了非线性 EV 模型的降维估计问题;第 6 章研究了协变量带有 Berkson 测量误差的非线性半参数模型,通过利用核估计和最小距离方法给出了未知参数 γ 和未知函数 $g(\cdot)$ 的估计,证明了未知参数估计的相合性和渐近正态性,同时也给出了未知函数估计的收敛速度;第 7 章研究了纵向数据下变系数 EV 模型的经验似然推断问题,给出了未

知函数估计的渐近正态性，构造了未知参数的两种经验对数似然比统计量，证明了所构造的经验似然比统计量的分布渐近于χ^2分布，所得结果可以用来构造未知非参数函数的渐近置信域.模拟研究给出了所提出的统计量在有限样本情形下的实际表现，最后将上述方法运用到了AIDS临床试验数据分析.

本书是国家社科基金项目（编号:10CTJ001），北京市教委社科计划项目（编号:SM201110038014），北京市属高等学校人才强教计划资助项目（编号:PHR201008214）的阶段性研究成果之一，同时也受到北京市教委统计学特色专业建设项目和北京市教委科研水平提高经费的资助.

该课题的研究得到了北京工业大学薛留根教授和西安交通大学吴可法教授的悉心指导.在此书出版之际，谨向薛留根教授和吴可法教授表示诚挚的感谢！

在本书的撰写过程中，首都经济贸易大学统计学院纪宏教授、马立平教授、刘黎明教授、刘娟教授、沈大庆教授给予了极大的支持和热心的帮助，在此表示诚挚的感谢！另外，还需特别感谢首都经济贸易大学统计学院的张娟博士和首都经济贸易大学出版社的赵杰老师、薛晓红编辑，没有他们的帮助和关心，本书不可能很快出版。

由于作者水平有限，不妥之处在所难免，恳请各位同行和读者批评指正.

刘　强

2012年3月于首都经济贸易大学

E－mail：cuebliuqiang@163.com

符号表

$\triangleq$	“定义为”或“记为”
T	向量或矩阵的转置
$\xrightarrow{L}$	依分布收敛
$\xrightarrow{P}$	依概率收敛
$a.s.$	强收敛（依概率 1 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
$O_p(1)$	随机有界
$K(\cdot)$	核函数
h 或 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 的卡方分布
i. i. d.	独立同分布
$\|\cdot\|$	Euclidean 范数
$A^{\otimes 2}$	AA^T
$\mathrm{tr}(A)$	方阵 A 的迹
$\mathrm{mineig}(A)$	矩阵 A 的最小特征值
$\mathrm{diag}(a_1, \cdots, a_n)$	由元素 $a_1, \cdots, a_n$ 组成的对角阵
c	正的常数，在不同地方可以表示不同的值

目 录

1 绪论 …… 1

1.1 统计模型 …… 3

1.2 复杂数据 …… 15

1.3 非参数估计方法 …… 17

1.4 降维估计 …… 23

1.5 本书的内容及结构 …… 24

1.6 参考文献 …… 26

2 删失数据下线性 EV 模型的估计 …… 38

2.1 方法与主要结果 …… 39

2.2 模拟研究 …… 46

2.3 实际数据分析 …… 48

2.4 定理的证明 …… 50

2.5 参考文献 …… 55

3 缺失数据下 EV 模型的估计 …… 58

3.1 非线性 EV 模型响应变量均值的估计 …… 60

3.2 非线性 EV 模型中未知参数的估计 …… 83

3.3 非线性半参数 EV 模型的估计 …… 95

3.4 参考文献 …… 121

4 纵向数据下 EV 模型的估计 …… 125
4.1 半参数 EV 模型的估计 …… 127
4.2 混合效应 EV 模型的估计 …… 143
4.3 参考文献 …… 154
5 非线性 EV 模型的降维估计 …… 157
5.1 方法与主要结果 …… 158
5.2 定理的证明 …… 164
5.3 参考文献 …… 166
6 Berkson 测量误差模型的估计 …… 167
6.1 方法与主要结果 …… 168
6.2 定理的证明 …… 171
6.3 参考文献 …… 178
7 变系数 EV 模型的经验似然推断 …… 180
7.1 引言 …… 180
7.2 方法与主要结果 …… 182
7.3 数值研究 …… 187
7.4 定理的证明 …… 191
7.5 参考文献 …… 197

1

绪 论

在经济、社会、人口、医学等众多研究领域中，人们通过各种方式收集数据，然后对数据进行统计分析，利用分析结果指导社会实践. 但是在收集数据的过程中，经济变量能够精确观测的假定在实际问题中往往不能成立. 一方面，对兴趣变量进行测量时，经常会受到多种客观因素的影响，导致一些偏差，如抽样误差、记录误差等. 例如，为了研究不同地区城镇居民生活中的储蓄和消费状况，往往需要通过各种途径进行数据调查，然而在调查中，由于种种因素的影响，我们得到的一般不是真实的储蓄和消费数据；在研究身高和体重的关系时，测量身高和体重时测量值往往不是很精确，得到的数据经常带有误差；研究某种药物对某种病症的影响，在将病症量化时也不可能做到精确；调查研究影响工资收入的因素时，某些被调查的人可能因为种种原因，不愿意透露真实的工资，此时调查得到的数据也带有测量误差；等等 . 另一方面，人们考察变量之间的关系时，常常只关心主要因素的影响，其他影响不大的因素的效应将体现于偏差中 . 文献中通常称这种观测数据带有误差的问题为“测量误差问题”，称分析这些数据的统计模型通常为“测量误差模型”或“EV(Errors - in - variables)模型”. 在大多数情形下，人们在进行数据分析时，往往忽略这种误差，直接用带有测量误差的数据代替真实数据进行分析，当测量误差较小时，产生的结

果可能与真实值相差不大,在可以接受的范围内. 但是如果误差太大,分析所得结果的真实性则会受到很大的影响. 因此,关于测量误差模型(EV 模型)的研究无论是在理论还是在实践上都具有极其重要的意义.

EV 模型最早是在 19 世纪 70 年代出现的,之后百余年里发展缓慢,直到 20 世纪 80 年代,人们在经济、金融、医学、生物等领域的研究中发现很多数据用普通的回归模型处理时得到的结果不太理想,究其原因是数据误差过大,无法利用传统的回归模型进行分析,因此测量误差模型得到了大量应用,随之该问题的理论研究开始受到重视并迅速发展. 在过去的 20 多年里,线性、非线性测量误差模型的参数估计问题已有大量的研究,未知参数置信域的构造以及模型检验问题也有了长足的发展. 随着生物、医学、人口、环境、经济、金融等学科研究中的实验技术、检验方法以及数据分析手段的日益提高,所获得的数据在结构上越来越复杂精细,数据所提供的信息也越来越繁杂,获得的变量个数也越来越多,关于复杂数据 (Complex Data)的统计推断应运而生.

复杂数据主要包括删失数据 (Censored Data)、纵向数据 (Longitudinal Data)、缺失数据 (Missing Data)、面板数据 (Panel Data)、高维数据 (High - dimensional Data)等统计数据. 复杂数据的分析与建模已经成为当今国内外统计学界和计量经济学界中的一个研究热点,也是一个研究难点,已经引起国内外众多统计学家和计量经济学家的高度关注. 由于复杂数据形式多样,存在数据的缺失、数据的删失、纵向数据等各种情况,这给模型的建立、数据的分析造成了很大的困难. 相对于独立同分布数据而言,复杂数据在现实生活中更为普遍,关于复杂数据下 EV 模型的估计与检验问题更富有现实意义,但处理过程却更为棘手,也正因为如此,在该领域中还有许多的空白亟待研究. 另外,误差类型以及模型的种类

也都在不断增加. 常见的误差类型有:可加误差、乘积误差、Berkson 误差等;模型的类型主要有:线性模型、非线性模型、广义线性模型、部分线性模型、混合效应模型、变系数模型以及单指标模型等. 基于以上这些因素,对于不同的数据类型、误差类型以及模型类型,需要用不同的统计方法来进行分析,因此复杂数据下各种 EV 模型(如线性 EV 模型、广义线性 EV 模型、混合效应 EV 模型、变系数 EV 模型等)的研究还有许多具有挑战性的问题有待我们去研究解决. 如何在纵向数据、删失数据等复杂数据情形下讨论各种 EV 模型的大样本性质,是当今统计学界和计量经济学界亟待解决的热点课题之一.

本书研究的模型主要包括:线性 EV 模型、非线性 EV 模型以及半参数 EV 模型等统计模型. 研究的主要目的是:在纵向数据、删失数据、缺失数据等复杂数据下研究几类 EV 模型中兴趣参数及兴趣函数估计的大样本问题,如估计量的渐近正态性、相合性及其收敛速度等统计性质.

下面分别对半参数回归模型、测量误差模型(EV 模型)、混合效应模型、变系数模型、复杂数据、降维估计、经验似然方法以及它们的研究现状进行简单的回顾.

1.1 统计模型

1.1.1 半参数回归模型

根据回归函数的形式,回归模型可以分为参数型、非参数型、半参数型三种形式. 参数回归模型(Parametric Regression Model)主要分为两类:线性回归模型(Linear Regression Model)和非线性回归模型(Nonlinear Regression Model). 线性回归模型的一般形式为

$$Y = X^T\beta + \varepsilon \tag{1.1.1}$$

其中，Y 为响应变量，X 为 p 维协变量，β 为 p 维未知参数，ε 为模型误差.

非线性回归模型的一般形式为

$$Y = f(X;\beta) + \varepsilon \tag{1.1.2}$$

其中，Y 为响应变量，X 为 p 维协变量，β 为 q 维未知参数，$f(\cdot)$ 为已知的函数，ε 为模型误差.

参数回归模型对回归函数提供了大量的额外信息（通常由经验和历史资料提供），因而当假设模型成立时，其推断具有很高的精度. 但当参数假定与实际（真实模型）相背离时，基于假定模型所做的统计推断可能表现很差.

由于参数模型适应性较差，Stone[1] 于 1977 年提出了如下非参数回归模型（Nonparametric Regression Model）：

$$Y = m(X) + \varepsilon \tag{1.1.3}$$

其中，Y 为响应变量，X 为 p 维协变量，$m(\cdot)$ 为未知的 Borel 函数，ε 为模型误差，一般假定 $\mathrm{E}(\varepsilon|X) = 0$.

非参数回归模型的特点是：回归函数的形式是任意的，变量的分布也很少限制，因而具有较大的适应性. 由于回归函数未知，在实际问题未提供任何信息时，非参数模型会明显降低其解释能力.

为弥补非参数回归模型的不足，一个努力的方向就是 Engle 等[2] 在研究气象条件对电力需求的影响这一实际问题时提出了部分线性回归模型（Partially Linear Regression Model）. 部分线性回归模型由两个部分组成，线性部分称为参数分量，非线性部分称为非参数分量，其中参数分量中的回归参数 β 是有限维，而非参数分量涉及无限维的未知参数（详见文献[3]）. 该模型自提出以来一直是统计学界和计量经济学界研究的一个热点课题. 作为部分线性模型的一般推广，半参数回归模型（Semiparametric Regression Model）可以表示如下：

$$Y=f(X,\beta)+g(T)+\varepsilon \tag{1.1.4}$$

其中,$f(\cdot)$为已知的联系函数,$g(\cdot)$为未知的非参数函数,$\{(X^T,T)\}$为协变量,β为p维未知参数,ε为随机误差.

作为半参数回归模型的特例,文献中所谓的"部分线性回归模型"的一般形式为

$$Y=X^T\beta+g(T)+\varepsilon \tag{1.1.5}$$

半参数回归模型(包括部分线性回归模型)的优点在于:它既含有参数分量,又含有非参数分量,可以概括和描述众多实际问题,因而自模型提出以来就引起统计学界和计量经济学界的广泛重视. 半参数回归模型的核心在于它既集中了主要部分(参数分量)的信息,又兼顾了干扰因素(非参数部分)的信息,因而具有较强的解释能力. 随着半参数回归模型在理论和方法上的日益成熟,它必将具有广阔的应用前景.

关于非参数、半参数回归模型的具体问题可以参见柴根象与洪圣岩的专著《半参数回归模型》[3], Härdle, Liang 和 Gao 的专著《Partially Linear Models》[4], Ruppert 等的专著《Semiparametric Regression》[5],叶阿忠的专著《非参数计量经济学》[6], Desu 和 Raghavarao 的专著《Nonparametric Statistical Methods for Complete and Censored Data》[7] 以及 Tsiatis 的专著《Semiparametric Theory and Missing Data》[8]. 文献[3]着重研究大样本领域的一些传统问题,包括估计的相合性、渐近分布、收敛速度和渐近效率等. 文献[4]涉及若干较新题材,包括 Bootstrap、截尾数据、非线性和非参数时间序列等. 文献[5]对 2003 年前的参数、半参数回归模型统计推断的研究成果进行了总结,讨论的模型囊括了参数回归模型、广义参数回归模型、半参数回归模型、可加模型、半参数混合效应模型、广义可加模型、测量误差模型、贝叶斯半参数回归模型以及交互模型等统计模型,研究的内容主要有密度估计、参数估计、假设

检验、变量选择等等. 文献[6]主要讨论非参数、半参数模型在计量经济学中的应用及计算机模拟程序实现. 文献[7]侧重于删失数据下非参数统计方法的探讨. 文献[8]主要侧重于探讨缺失数据下半参数模型的统计推断问题.

1.1.2 EV 模型

在科学实验、工农业生产、经济分析以及社会调研等领域中,从客观的角度来看,人们在对兴趣变量进行观测时,往往会受到多种因素的影响,导致一些观测偏差;从主观的角度来看,人们考察变量之间的关系时,往往将诸多影响不大的因素的效应归结在兴趣变量取值的偏差中. 因此,观测变量带有测量误差的数据在实际问题中是非常普遍的.

自 1877 年 Adcock[9] 最先开始讨论两个变量的观测均含有误差的直线拟合以来,EV 模型一直受到人们的广泛重视. 通常的可加 EV 模型可写为

$$\begin{cases} Y = f(X) + \varepsilon \\ \xi = X + u \end{cases} \tag{1.1.6}$$

其中,X 为真值,ξ 为观测值,ε 和 u 分别为模型误差和测量误差.

相应于回归模型的分类形式,根据 f 的不同形式,EV 模型可分为参数 EV 模型、非参数 EV 模型以及半参数 EV 模型. 若函数 f 具有形式 $g(\cdot,\beta)$,且 g 的形式已知,则模型(1.1.6)称为参数 EV 模型;若 g 的形式完全未知,相应模型称为非参数 EV 模型;若函数 f 具有形式 $g(\cdot,\beta)+h(\cdot)$,其中 g 形式已知而 h 未知,则相应模型称为半参数 EV 模型,半参数 EV 模型是介于参数 EV 模型与非参数 EV 模型之间的一类模型.

为了处理协变量带有测量误差问题,Fuller[10] 提出如下线性 EV 模型: