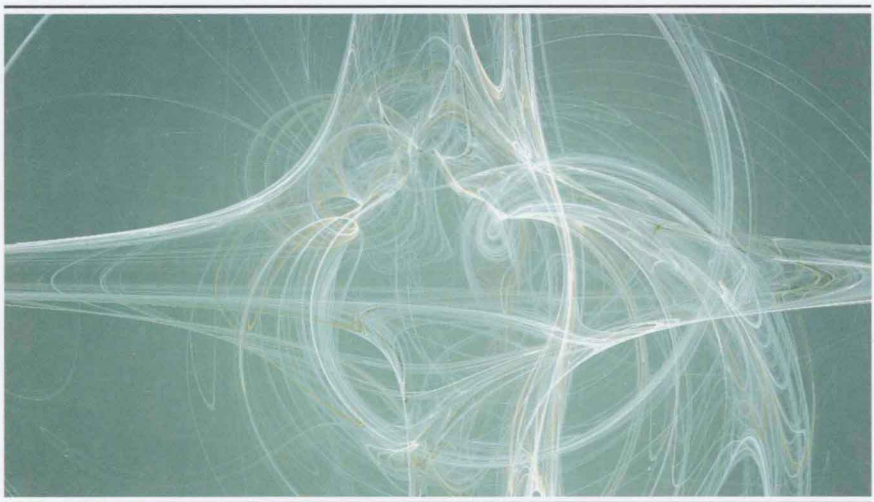
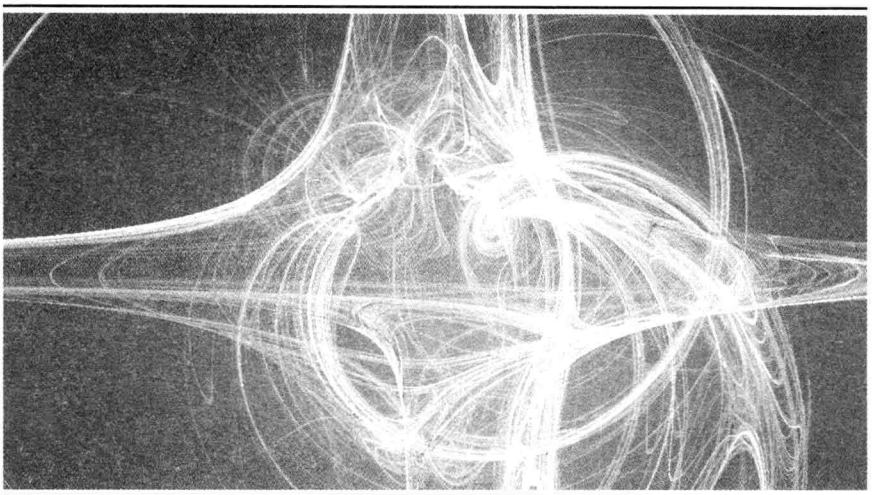
基于复杂数据的统计推断及其应用



宇世航　张良勇　魏蕴波◆著

黑龙江大学出版社
HEILONGJIANG UNIVERSITY PRESS

基于复杂数据的统计推断及其应用



宇世航　张良勇　魏蕴波◆著

黑龙江大学出版社
HEILONGJIANG UNIVERSITY PRESS

图书在版编目(CIP)数据

基于复杂数据的统计推断及其应用 / 宇世航，张良勇，魏蕴波著. -- 哈尔滨 : 黑龙江大学出版社，2013.7
ISBN 978-7-81129-644-0

Ⅰ. ①基… Ⅱ. ①宇… ②张… ③魏… Ⅲ. ①统计分析-研究 Ⅳ. ①O212.1

中国版本图书馆 CIP 数据核字(2013)第 166700 号

基于复杂数据的统计推断及其应用

JIYU FUZA SHUJU DE TONGJI TUIDUAN JI QI YINGYONG

宇世航　张良勇　魏蕴波　著

责任编辑　张永生　于　丹
出版发行　黑龙江大学出版社
地　　址　哈尔滨市南岗区学府路 74 号
印　　刷　哈尔滨市石桥印务有限公司
开　　本　720×1000　1/16
印　　张　12.25
字　　数　233 千
版　　次　2013 年 7 月第 1 版
印　　次　2013 年 7 月第 1 次印刷
书　　号　ISBN 978-7-81129-644-0
定　　价　28.00 元

前　言

统计学能得到迅速发展，主要原因就是它可与其他学科融合，根据实际问题的需要，不断探索新的数据分析方法，逐渐形成新的理论。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。随着经济、人口、医学、生物、金融、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进步，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。复杂数据包括时间序列数据、缺失数据、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统计学界和计量经济学界的研究热点。

本书主要研究的内容有：时间序列数据下测量误差模型统计推断的方法，包括均值估计、递归型核密度估计、解释变量有误差的自回归模型、全部变量有误差的自回归模型；基于整值时间序列数据的风险模型；负相依数据下部分和之和的大数定律、中心极限定理、核回归估计；缺失数据非参数回归函数加权核估计；微阵列数据下的多重假设检验；等等。

这些内容仅就某类复杂数据的某个方面做了研究，列出研究成果的同时，也对相应问题的研究现状做了简要的介绍，以便读者对此类问题的发展状况有所了解。

本书得到了黑龙江省教育厅科学技术研究项目（编号：11551543）、黑龙江省应用技术研究与开发计划项目（编号：2013R0150）、齐齐哈尔大学青年教师科学技术类科研启动支持计划项目（编号：2010K－M28，2012K－M30）的支持。本书第1章、第2章、第3章、第4章由宇世航执笔，第6章、第7章由魏蕴波执笔，第5章、第8章由张良勇执笔。

由于作者水平有限，疏漏不足在所难免，恳请同行及广大读者批评指正。

作者于2013.7

目　录

第1章 绪 论

1.1 测量误差

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，对兴趣变量进行观测时，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：在研究不同地区城镇居民的储蓄和消费状况时，往往需要通过各种途径进行数据调查，但在调查中，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，身高和体重的测量值往往不是很精确，得到的数据经常带有误差；在研究某种药物对某种病症的影响时，在将病症量化时也不可能精确；在调查影响工资收入的因素时，某些被调查的人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。通常称这种观测数据带有误差的问题为“测量误差问题”或“EV(Errors - in - Variables)问题”。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的数据进行分析。当测量误差较小时，产生的结果与真实值相差不大，在可以接受的范围内，但是如果误差太大，分析结果会受到很大的影响。因此，关于测量误差模型(EV 模型)的研究无论是在理论上还是在实践上都是具有极其重要意义的。

最原始的测量误差模型在 19 世纪 70 年代就已经出现，但并未受到特别的重视。其后百余年，回归模型无论在理论上还是在应用上均有长足的进展，而且在统计分析中起到了极为重要的作用。反观测量误差模型却进展缓慢，真正的原因很难厘清，但模型的复杂程度可能是其主要因素。直观来说，能收集到的数据是以 (X,Y) 的形式出现的，只知道 Y 和 X 之间有关系，但 Y 和 X 之间的关系却不明确，在此情形下，任何统计推断或多或少都会碰到困难。从技术上来说，对模型的参数估计(点估计和区间估计)均有一定的困难和障碍，这也是测量误差模型在一般应用上不受重视的原因。到 20 世纪 80 年代，测量误差模型才开始受到重视，其原因

是人们在经济、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，究其原因是数据因误差过大无法利用传统的回归模型进行分析。

1.1.1 简单的测量误差模型

称如下形式的测量误差模型

$$\tilde{X} = X + u \tag{1.1}$$

为简单的测量误差模型，其中，X 与 u 独立，X 的密度 $f(x)$ 未知，u 是均值为 0、分布已知的测量误差。模型(1.1)中所描述的测量有不可忽略误差的现象广泛存在于显微荧光光度术(Microfluorophotometry)、电泳疗法(Electrophoresis)、生物统计、抽样调查和其他领域中。这一方面的工作大体分为两个阶段。

第一阶段是在独立同分布(i. i. d.)数据下的测量误差模型研究，比较有代表性的成果有：Fuller(1987)对当时线性测量误差模型的研究成果进行了概括和总结；Cui 和 Chen(2003)讨论了线性测量误差模型的经验似然推断问题，并提高了置信域和覆盖概率的精度；Hsiao 在随机变量 X 的密度函数已知的条件下证明了未知参数的最小二乘估计是相合的，且是渐近正态的。Carroll 等对 1995 年以前的关于非线性测量误差模型的研究做了详尽的归纳和总结，描述了矩估计法、似然函数法等，同时也包括参数模型中的参数估计、区间估计、假设检验，非参数模型中的密度估计以及连接函数的估计等统计推断问题。自 1995 年以来，非线性测量误差模型的研究又有了新的发展，如薛留根(2005)讨论了非线性测量误差模型中参数的估计问题，Cui 和 Kong(2006)及刘强等人(2005)分别利用经验似然和小波估计法研究了半参数模型参数及未知函数的估计。

第二阶段是在复杂数据下考虑测量误差模型。实践中，我们经常会遇到一些不完全数据，如测量误差数据、删失数据和缺失数据；还会遇到一些复杂数据，如组间独立而组内相关的纵向数据、Panel 数据以及时间序列数据。从几个统计学顶级杂志上可以看到，近年来对复杂数据测量误差模型的研究在不断深入，如：Qin 等人(2009)和刘强、薛留根(2012)讨论了缺失数据下线性和非线性测量误差模型参数的经验似然置信域，Heuchenne、Keilegom (2007)和陈放等人(2010)研究了删失数据下的部分线性测量误差模型和非线性测量误差模型，Xue、Zhu(2007)和 Liu(2011)考虑了纵向数据下半参数测量误差模型和部分线性测量误差模型。随着数据类型的不同，处理方法也各不相同。

1.1.2 一般的测量误差模型

称如下形式的测量误差模型

$$\tilde{X} = \varphi(X,u) \tag{1.2}$$

为一般的测量误差模型.其中 $\varphi(\cdot)$ 为任意函数。对于此类问题,文献中一般利用核实数据进行处理。

关于一般的测量误差模型(1.2),已经有了很多研究成果。例如,Wang(2000)和 Wang、Rao(2002)研究了线性测量误差模型的估计问题;Sepanski、Lee(1995)和 Stute 等人(2007)研究了非线性测量误差模型的统计推断问题;Wang、Yu(2007)和薛留根(2006)等人分别研究了半参数测量误差模型、部分线性测量误差模型、非线性半参数测量误差模型的统计推断问题;刘强、薛留根(2010)讨论了单指标测量误差模型的估计问题,利用核实数据构造了未知参数的两种经验对数似然比统计量,即估计的经验对数似然比统计量和调整的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置信域。

1.1.3 测量误差模型的常用处理方法

1.1.3.1 变量的平均变换与分解卷积方法

在实际应用中,估计真实变量变换的均值十分必要,而真实变量的观测往往含有测量误差。例如,在抽样调查中,希望估计一批产品中 n 个球的平均体积,但直径 X 的测量带有误差 u,观测到 $\tilde{X}_i = X_i + u_i$ ($1 \leqslant i \leqslant u_i$),平均体积 $\pi E(X^3)/6$ 需要根据样本 $\{\tilde{X}_1,\cdots,\tilde{X}_n\}$ 进行估计。在实际抽样中,样本量一般很大,如果误差分布足够,X 的分解卷积非参数密度估计应该是可行的。一般来说,X 的分布关于光滑可积函数 $h(\cdot)$ 的平均变换定义为

$$\theta = E[h(X)] = \int h(x)f(x)\mathrm{d}x$$

其中,$f(\cdot)$ 是 X 的密度函数。

Qin 和 Feng(2003)根据来自模型(1.1)的 $\{\tilde{X}_1,\cdots,\tilde{X}_n\}$ 构造了 θ 的卷积核估计

$$\hat{\theta}_{nd} = \int h(x)\hat{f}_n(x)\mathrm{d}x = \frac{1}{na_n}\sum_{j=1}^{n}\int K_n\left(\frac{x-\tilde{X}_j}{a_n}\right)h(x)\mathrm{d}x$$

其中,$\hat{f}_n(x) = (na_n)^{-1}\sum_{j=1}^{n}K_n\left(\frac{x-\tilde{X}_j}{a_n}\right)$ 是密度 f 的分解卷积核密度估计,

$$K_n(x) = \frac{1}{2\pi}\int\frac{\varphi_K(t)}{\varphi_u(t/a_n)}\exp(-itx)\mathrm{d}t$$

是核函数 $K(\cdot)$ 的 Fourier 变换 $\varphi_K(t)$ 的分解卷积核,a_n 是窗宽序列,$i = \sqrt{-1}$ 且 $\varphi_u(t)$ 是 u 的特征函数。

由于分解卷积核密度估计的极限行为极其依赖 φ_u 的尾部,在普通光滑的情形

下，Fan（1991a，1991b）证明了 $f_n^{(l)}(x) - f^{(l)}(x)$ 具有最优平均收敛速度 $O(n^{-(m+\alpha-l)/[2(m+\alpha+\beta)+1]})$（关于 f 一致，其中，β 为 φ_u 的普通光滑指数，m 和 α 为 f 可导指数）以及标准化 $f_n(x)$ 后具有渐近正态性。

Cui（2005）在普通光滑的情形下获得了 $E(\hat{\theta}_{nd}) - \theta$ 和 $\hat{\theta}_{nd} - \theta$ 的表示定理，从而证明了 $\hat{\theta}_{nd}$ 的渐近正态性。当 $h(x)$ 为多项式时，分解卷积 σ_0^2，$K(x) = \exp(-x^2/2)/\sqrt{2\pi}$ 可以得到矩估计 $\hat{\theta}_{nm} = (1/n)\sum_{j=1}^{n}\tilde{X}_j^2 - \sigma_0^2$，并且 $\hat{\theta}_{nd} = \hat{\theta}_{nm} + a_n^2$。

1.1.3.2 SIMEX 方法与 EXPEX 方法

众所周知，正态分布是超光滑分布的重要代表，研究一般超光滑情形下的测量误差模型较为困难。Fan（1991a，1991b）证明了 $f_n^{(l)}(x) - f^{(l)}(x)$ 具有最优平均收敛速度 $O((\log n)^{-(m+\alpha-l)/\beta})$（关于 f 一致，其中，β 为 φ_u 的超光滑指数，m 和 α 为 f 可导指数）以及标准化 $f_n(x)$ 后具有渐近正态性，表明在测量误差模型中超光滑情形下，用分解卷积法进行与 X 密度有关的非参数估计的收敛速度极慢，并且不可能改进。Cook、Stefanski（1994），Stefanski、Bay（1996），Staudenmayer、Ruppert（2004）等给出了 SIMEX 方法，即当测量误差分布已知正态时，用原数据加上模拟误差数据（方差 $\delta > 0$，可变化）产生新数据，并在新数据下进行估计。随着 δ 变化，可找出估计的变化规律，进而拟合出变化曲线，再外推插值至 $\delta = -1$ 时估计的值，即得所求的估计。这一方法对误差分布已知正态且方差较小时的估计比较有效。

Cui（2005）在超光滑情形下，研究了简单测量误差模型的 X 的平均变换估计，提出了 EXPEX 方法，这是相对于 SIMEX 方法的一种估计方法，它能有效地避免由误差超光滑所带来的困难，从本质上提高估计的收敛速度。

令 $Z \sim N(0,\sigma_0^2)$，$h^*(y) = \mathrm{Re}\{E_Z[h(y+iZ)]\} = \lim_{\lambda\to -1}\mathrm{Re}\{E_Z[h(y+\sqrt{\lambda}Z)]\}$，其中，$i = \sqrt{-1}$，Re 表示复数的实部。称 $h^*(y)$ 是 $h(y)$ 的 EXPEX 函数，θ 的 EXPEX 估计构造如下：

$$\hat{\theta}_{ne} = \frac{1}{n}\sum_{j=1}^{n} h^*(\tilde{X}_j)$$

在一定的条件下，Cui（2005）建立了 $\hat{\theta}_{ne}$ 的 $\sqrt{n}$ 渐近正态性。

1.1.3.3 线性测量误差模型的正交回归与 M 估计法

称如下形式的测量误差模型

$$\begin{cases} Y = X^{\tau}\beta_0 + \varepsilon \\ \tilde{X} = X + u \end{cases} \tag{1.3}$$

为线性测量误差模型，其中 $\tilde{X}$ 为取值于 $\mathbb{R}^p$ 上的可观测随机向量，X 为 p 维不可观测

的随机向量，β_0 为 $p\times 1$ 未知参数向量，$(\varepsilon,u^\tau)^\tau$ 为 $p+1$ 维对称向量，即 $(\varepsilon,u^\tau)^\tau \overset{d}{=} RU_{p+1}$（其中，$R$ 为非负随机向量，U_{p+1} 为 $\Omega_p=\{a\mid a\in\mathbb{R}^p,\ \|a\|=1\}$ 上的均匀随机向量，并且 R 与 U_{p+1} 独立），$\sigma^2=ER^2/(p+1)>0$ 未知，$(\varepsilon,u^\tau)^\tau$ 与 X 独立（球对称误差的要求是为了使模型可识别）。模型(1.3)为线性测量误差模型，有着广泛的应用背景（如经济、林业、建筑、生物、遥感等领域）。对模型(1.3)的研究主要是利用极大似然法、广义最小二乘法分别给出 β_0、σ^2 的估计 $\hat{\beta}_n$、$\hat{\sigma}_n^2$，并获得它们的相合性与渐近正态性，这一方面的重要工作可参见文献(Glesser,1990)。但随着稳健统计方法的发展，人们已不满足于广义最小二乘估计。1989 年，Zamar 给出了测量误差模型中 β_0 的估计 $\hat{\beta}_n$，并在一些不易验证的条件下获得了 $\hat{\beta}_n$ 的强相合性。在正态误差假设下，Cheng 和 van Ness(1999)应用正交回归和极大似然法研究了结构线性误差的稳健估计问题。

假设 $\{\tilde{X}_i=(X_{i1},X_{i2},\cdots,X_{ip})^\tau,Y_i\mid 1\leqslant i\leqslant n\}$ 为来自模型(1.3)的一组独立同分布随机样本，即

$$\begin{cases}Y_i=X_i^\tau\beta_0+\varepsilon_i\\ \tilde{X}_i=X_i+u_i\end{cases}\qquad i=1,2,\cdots,n$$

$(\varepsilon_i,u_i{}^\tau)^\tau$ $(1\leqslant i\leqslant n)$ 为独立同分布球对称随机误差向量，有 $E(\varepsilon_1,u_1{}^\tau)^\tau=0$，$Cov(\varepsilon_1,u_1{}^\tau)^\tau=\sigma^2I_{p+1}$。为了获得 β_0 的 M 估计，选取一适当的 $\rho(\cdot)$ 函数，则 β_0 的正交回归 M 估计定义为下述极值问题的解：

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{Y_i-\tilde{X}_i^\tau\hat{\beta}_n}{\sqrt{1+\|\hat{\beta}_n\|^2}}\right)=\min\left[\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{Y_i-\tilde{X}_i^\tau\beta}{1+\|\beta\|^2}\right)\right]$$

称 $\hat{\beta}_n$ 为 β_0 的 M 估计，并由此定义 σ^2 的估计 $\hat{\sigma}_n^2$

$$\hat{\sigma}_n^2=\frac{1}{n}\sum\frac{(Y_i-\tilde{X}_i^\tau\hat{\beta}_n)^2}{1+\|\hat{\beta}_n\|^2}$$

Cui(1997a)研究了测量误差模型中 β_0 的 M 估计问题，在一般的 $\rho(\cdot)$ 函数下，获得了 β_0 的估计 $\hat{\beta}_n$，在一些基本的假设下，得到了 $\hat{\beta}_n$ 的强相合性与渐近正态性，并同时得到了 σ^2 的估计 $\hat{\sigma}_n^2$ 及其渐近性质。需要指出的是 Cui(1997b)在球对称误差向量假设下，研究了线性测量误差模型中广义最小二乘估计的渐近性质，并说明了对不可观测的点列或随机向量所施加的条件及对误差向量所施加的矩条件本质上是不可改进的。

1.1.3.4 核实数据法

在许多实际问题中，准确测得我们感兴趣的变量 X 是很困难的，或者所需成本

很高,或者需要花费大量时间。因此通常用可观测的替代变量 $\tilde{X}$ 来代替,这些替代变量可由一些相对简单的测量方法获得。一般地,替代变量和真正的变量之间的关系相对于经典的可加的误差结构(模型(1.1))来说要复杂。事实上,在许多实际问题中,很难明确指出真正的变量和替代变量之间的关系,如模型(1.2)。也就是说,最现实的情况是不对真正的变量与替代变量间关系做任何模型假设。然而,这会给获得准确的统计分析带来很大的困难。如果没有额外的观察和信息,由测量误差引起的偏差将很难评估。一种解决方法就是使用核实数据捕捉真正的变量与替代变量间的关系。获得核实数据的例子很多,一些例子可在 Amemiya、Fuller(1984),Chen、van Ness(1999),Cook、Stefanski(1994)的文献中找到。如劳动力市场调查报告中的工资和工作时间是有测量误差的,但是,劳动力的工资和工作时间的准确数据可由工资表记录获得。又如对校园吸烟行为的调查通常由问卷调查的方式进行,但采集的数据有误差,若对学生的唾液进行化学分析,所需的成本又高,于是仅对一小部分研究对象的唾液进行化学分析。这里带有误差的变量(例如劳动力市场调查报告中的工资和工作时间,校园吸烟行为自我报告)被用作替代变量,相应的对研究对象的一小部分子集的精测数据被用作核实数据。

在核实数据的帮助下,一些统计学家已经基于替代变量发展了统计推断技巧。Carroll 和 Wand(1991)对于 Logistic 测量误差模型发展了使用核回归技巧的半参数方法。Pepe 和 Fleming(1991)也考虑了一个代替离散随机变量的类似问题。Sepanski(1996)考虑非线性模型

$$Y = g_1(X,\beta_0) + e$$

这里 $g_1(\cdot)$ 是一已知函数,X 有测量误差,或者 Y 有测量误差,或者 X 和 Y 都有测量误差。他在每种情况下利用核实数据对此模型发展了一种所谓的半参数回归方法,也就是在给定替代变量的条件下没有指定真正的变量的分布假设,他定义了 β_0 的估计。Wang(1999)考虑部分线性模型 $Y = X'\beta + g(T) + e$。这里 X 为有测量误差的 p 维协变量向量,T 和反映变量 Y 精确测量。值得注意的是 Wang 考虑的模型没有包括在 Sepanski 所考虑的模型中,因为这里 $g(\cdot)$ 是一个未知函数。他应用半参数的方法获得 β 和 $g(\cdot)$ 的估计,并证得 β 的强相合性、渐近正态性以及获得 $g(\cdot)$ 的估计的弱相合收敛率。Wang(2003)仍考虑上面提到的部分线性模型,但这里反映变量 Y 有测量误差,T 和协变量向量 X 精确测量,应用半参数降维的技巧获得 β 和 $g(\cdot)$ 的估计,并证得 β 的渐近正态性。研究者(宇世航,2010;宇世航,赵世舜,2012)分别构造了核实数据下均值估计和递推核密度估计的方法。

1.2 时间序列数据

按时间次序排列的随机变量序列

$$X_1, X_2, \cdots \tag{1.4}$$

称为时间序列。如果用

$$x_1, x_2, \cdots, x_N \tag{1.5}$$

分别表示随机变量 $X_1, X_2, \cdots, X_N$ 的观测值，就称序列(1.5)是时间序列(1.4)的 N 个观测样本。这里 N 是观测样本的个数。如果用

$$x_1, x_2, \cdots \tag{1.6}$$

表示 $X_1, X_2, \cdots$ 的依次观测值，就称序列(1.6)是时间序列(1.4)的一次实现或一条轨道。

时间序列数据是一种复杂的数据对象，在社会生活的各个领域中有大量的时间序列数据有待进一步的分析和处理。时间序列分析的主要任务就是根据观测数据的特点为数据建立尽可能合理的统计模型，然后利用模型的统计特性去解释数据的统计规律，以达到预报或控制的目的。

1.2.1 传统的时间序列模型

传统的时间序列模型处理的都是连续数据，例如某地区的月平均气温、股票日交易价格等。ARMA 模型是时间序列分析中应用最为广泛的一类模型。在近十几年里 ARCH 模型也取得了极为迅速的发展。ARCH 模型作为一种度量金融时间序列数据波动性的有效工具，应用于与波动性有关的研究领域，包括政策研究、理论命题检验、季节性分析等方面。采用 ARCH 模型来模拟波动性，将会给期货交易制度设计、风险控制制度设计和投资组合风险管理策略研究提供一个更为广阔的研究空间。几类模型的数学描述如下：

AR(p) 模型(Auto - Regressive Model)：

如果 $\{\varepsilon_t\}$ 是白噪声 WN(0, σ^2)，实数 $a_1, a_2, \cdots, a_p$ ($a_p \neq 0$) 使得多项式 $A(z)$ 的零点都在单位圆外，

$$A(z) = 1 - \sum_{j=1}^{p} a_j z^j \neq 0, \ |z| \leqslant 1$$

就称 p 阶差分方程，

$$X_t = \sum_{j=1}^{p} a_j X_{t-j} + \varepsilon_t, \quad t \in \mathbb{Z}$$

是一个 p 阶自回归模型。

MA(q) 模型(Moving Average Model)：

如果 $\{\varepsilon_t\}$ 是白噪声 WN(0, σ^2)，实数 $b_1, b_2, \cdots, b_q$ ($b_q \neq 0$) 使得

$$B(z) = 1 + \sum_{j=1}^{q} b_j z^j \neq 0, \ |z| < 1$$

就称

$$X_t = \varepsilon_t + \sum_{j=1}^{q} b_j \varepsilon_{t-j}, \quad t \in Z$$

是一个 q 阶滑动平均模型。

ARCH **模型**(Autoregressive Conditional Heteroscedasticity Model):

$$X_t = \beta_0 + \beta_1 X_{t-1} + \cdots + \beta_p X_{t-p} + u_t,$$

$$\sigma_t^2 = Eu_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_q u_{t-q}^2$$

1.2.2 整数值时间序列数据

离散取非负整数值的时间序列数据(即计数数据)在实际生活中普遍存在,比如每月的失业人数、一个医院每天的住院病人数、一个车站的日客流量、未支付的信用分期付款(在信用评分中用)、事故或者事故索赔次数(在确定保费时用)、住房抵押贷款中的预付费人数(在定价住房抵押贷款证券时用),因此统计学家对整数值时间序列的研究越来越感兴趣,它已成为统计学者的研究热点问题。

整数值时间序列数据广泛地、大量地存在,而且彼此之间各有差异。20 世纪 80 年代,由 Al - Osh 和 Alzaid (1987)提出了 INAR 模型。Weiβ(2008)综述了 2008 年之前整数值时间序列数据的研究成果。Kim、Park(2008)和 Zhang 等人 (2010)提出了符号稀疏算子(Thinning Operator),基于此算子的 INAR 模型可以处理负整数值和负相关的时间序列数据,特别是在处理非平稳的整数值时间序列数据时发挥了重要的作用。近年来,整数值自回归模型再次引起人们的兴趣,Drost 等人(2009a,2009b)分别在 JRSS B 和 Bernoulli 等统计学高档次杂志发表了最新研究成果。上述模型属于稀疏算子模型,即利用稀疏算子来构建模型。这类模型是结构的状态空间模型,它常用 Poisson (泊松)分布来构造模型。一个具有重大影响的模型是 Ferland 等人 (2006)提出的整数值广义自回归条件异方差(Generalized Autoregressive Conditional Heteroskedasticity, GARCH)模型,它可以处理整数值时间序列数据中的异方差性,即假设整数值时间序列数据的条件分布是 Poisson 分布,条件均值(强度过程)是历史观察值和历史强度的线性函数。由于这个模型是处理金融数据非常成功的 GARCH 模型的整数值推广,因而一出现就引起了广泛关注,并引发了一些相关研究。Fokianos 和 Rahbek(2009)考虑了模型的遍历性和基于似然的参数估计量,Fokianos 和 Fried(2010)研究了干预(Intervention)效应问题,Weiβ(2009)建立了自协方差函数的递推关系,Weiβ(2010a, 2010b) 给出了边际分布和高阶矩,Neumann(2011)给出了模型的混合性质和遍历性。

INAR **模型**:

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t, \quad 0 < \alpha < 1, t = 1, 2, \cdots$$

其中 $\alpha \circ X_{t-1} = \sum_{i=1}^{X_{t-1}} B_i$, X_0 为非负随机变量, $\{\varepsilon_t\}$ 是 i. i. d. 取非负整数值的随机变

量序列，$\{B_i\}$ 是 i. i. d. 伯努利随机变量序列，$P(B=1)=\alpha$。

INMA 模型：

$$X_t=\alpha\circ\varepsilon_{t-1}+\varepsilon_t,\quad 0<\alpha<1,t=1,2,\cdots$$

其中 $\alpha\circ\varepsilon_{t-1}=\sum_{i=1}^{\varepsilon_{t-1}}B_i$，$X_0$ 为非负随机变量，$\{\varepsilon_t\}$ 是 i. i. d. 取非负整数值的随机变量序列，$\{B_i\}$ 是 i. i. d. 伯努利随机变量序列，$P(B=1)=\alpha$。

1.3 负相依数据

近几年来，人们在研究独立随机变量的同时，相继提出了各种相依随机变量的概念。如自 20 世纪 50 年代引入的各种混合随机变量，60 年代引入的正相依（Positively Associated，PA）随机变量，80 年代引入的负相依（Negatively Associated，NA）随机变量，等等。这些相依随机变量的引入不仅是理论研究的需要，如在马氏链、随机场理论、多元统计分析等分支就早已提出了一些随机变量的相依性概念，而且也是解决实际问题的需要，如在一些实际问题中统计样本的观测值存在着非独立的情形。其中，NA 随机变量是 Joag - Dev 和 Proschan（1983）提出的如下一类包含独立随机变量在内的相依随机变量的概念：

定义 1.1 称随机变量 $X_1,X_2,\cdots,X_n(n\geqslant 2)$ 是 NA 的，若对 $\{1,2,\cdots,n\}$ 的任何两个不相交的非空子集 A，B 均有

$$Cov(f_1(X_i,i\in A),f_2(X_j,j\in B))\leqslant 0$$

其中 f_1 和 f_2 是任意两个使上述协方差存在的且对每个变元均非降（或均非升）的函数。

称随机变量序列 $\{X_j,j\in n\}$ 是 NA 序列，对任何 $n\geqslant 2$，$X_1,X_2,\cdots,X_n$ 都是NA 的。

有着统计学特色的 NA 随机变量不仅在可靠性理论、渗透理论、多元统计分析理论中应用，而且在通信、气象等工程领域及风险分析中应用，因此对 NA 序列在极限理论、统计及应用等方面的研究有着十分重要的意义。人们希望具有统计学特色的 NA 序列的极限理论同独立序列的极限理论一样完善。

NA 随机变量的一些基本且重要的性质：

性质 1：假设 $\{X_i,1\leqslant i\leqslant n\}$ 为 NA 随机变量，$A_1,A_2,\cdots,A_m$ 为 $\{1,2,\cdots,n\}$ 的两两不交空子集，$\alpha_i=\#(A_i)$ 表示 A_i 中元素的个数，如果 $f_i:R^{di}\to R,i=1,2,\cdots,m$ 为按分量增（或降）的函数，则 $\{f_i(x_i,j\in A_i),1\leqslant i\leqslant m\}$ 仍为 NA 的。进一步，如果 $f_i\geqslant 0,i=1,2,\cdots,m$，则还有

$$E\Big[\prod_{i=1}^{m}f_i(x_j,j\in A_i)\Big]\leqslant\prod_{i=1}^{m}=f_i(x_j,j\in A_i)$$

性质 2：独立随机变量序列必为 NA 随机变量序列，NA 随机变量序列的子列仍是 NA 的。

性质 3：若 $\{X_n, n\geq 1\}$ 与 $\{Y_n, n\geq 1\}$ 为相互独立的 NA 随机变量序列，则它们的并 $\{X_n, Y_n, n\geq 1\}$ 仍为 NA 的。

1.4 缺失数据

经典的统计方法与理论大都建立在完全数据分析的基础上，然而在实践中，常常因为各种原因一些数据不能获得，如一些被抽样的个体不愿提供所需要的信息、一些不可控的因素产生信息损失及一些调研者因本身的原因不能收集正确的调查信息，市场调研、邮寄问卷调查、社会经济研究、医学研究、观察研究及其他学科实验中常常产生缺失数据（不完全数据）。在这种情况下，标准的统计方法不能直接应用到这些缺失数据的统计分析中，一个简单直接的方法是排除那些有缺失数据的个体，只对有完全数据的个体进行分析，这是所谓的完全情形（CC）分析。然而，这一方法在大多数情况下都有严重的偏差，并且由于一些有缺失数据的个体被删除产生不必要的信息损失，常常导致无法分析。实际上，缺失数据统计分析方法的有效性很大程度上取决于数据是否依赖于数据集中的变量及与哪些变量有关，即是否依赖于缺失机制。

1.4.1 缺失机制

设 Z 是一个完全观测向量，当数据缺失时，设 Z_{obs} 是 Z 中总能被观察到的分量组成的向量；而设 Z_{min} 是 Z 中可能缺失的分量组成的向量；δ 是示性函数，若 Z 被完全观测，其取值为 1，否则取值为 0。下面介绍三种主要的缺失机制。

1.4.1.1 完全随机缺失（MCAR）机制

如果数据缺失不依赖于任何其他变量，即 $P(\delta = 1 \mid Z) = P(\delta = 1)$，则称数据缺失是 MCAR。

1.4.1.2 随机缺失（MAR）机制

如果数据缺失仅依赖于被观察到的变量 Z_{obs}，但不依赖于可能缺失的变量 Z_{min}，即 $P(\delta = 1 \mid Z) = P(\delta = 1 \mid Z_{obs})$，则称数据缺失是 MAR。

1.4.1.3 不可忽略缺失机制

如果数据缺失仅依赖于 Z 的缺失部分，这样的数据缺失称为不可忽略缺失或非随机缺失。

MCAR 机制意味着观察数据是所有数据的随机抽样，在 MCAR 机制假设下，上

面所述的CC分析可能损失效率,但并不引起偏差。MAR机制是比MCAR机制更加现实的假设,MCAR机制是MAR机制的特殊情形,CC分析在MAR机制假设下通常既可能产生无效推断,也可能产生偏差。容易看到不可忽略缺失机制是比另两种缺失机制更强的假设,一般地,在MAR机制下有效的方法在不可忽略缺失机制下并不有效。

1.4.2 缺失数据常用处理方法

设 X 是 p 维协变量,Y 是反映变量,实践中 Y 或 X 的某分量缺失。简单的CC分析通常不应用到这种缺失数据分析,因此,人们致力于寻求缺失数据的统计分析方法,以使不完全情形的信息得到使用,从而获得更加有效的推断。

1.4.2.1 似然方法

假设给定协变量 X,Y 的条件概率密度或(X,Y)的联合概率密度有参数形式,在 Y 缺失的情况下,无须对缺失机制做任何假设即可用CC分析做似然推断,并定义相合的极大似然估计。其渐近方差估计可用对数似然二阶微分获得。更进一步,Qin(2000)通过联合经验似然与参数似然,基于所有观察数据发展了半参数似然方法,这一方法利用辅助信息改进推断。但应该指出这种方法对模型假设是敏感的,即若模型假设错误,将定义有严重偏差的估计。而当协变量缺失时,获得极大似然估计的方法和技术有因子分解法、Newton - Raphson算法、拟Newton算法及EM算法等。

1.4.2.2 插补方法

插补方法就是使用某种规则或方法对缺失项填充数值,使有缺失的数据集变成完整的数据集。插补有单一插补与多重插补,是常用、简单、方便的方法。一般地,有均值插补、回归插补、随机回归插补、热平台插补、冷平台插补、替代插补等方法。关于这些插补方法的详细介绍可参见文献(Little,Rubin,2002),这里仅简介如下:

(1)均值插补就是以响应单元均值填补缺失值;

(2)回归插补就是用单元缺失项对观测项回归,用预测值填补相应的缺失值;

(3)随机回归插补就是用回归插补值再加上一个随机项填补相应的缺失值;

(4)热平台插补是由“类似”响应单元中抽取的值填补相应的缺失值;

(5)冷平台插补是用其他来源获得的数据代替某一项目中的缺失数据;

(6)替代插补就是用总体中未选到的备择单元代替不响应单元,如一个户主无法取得联系,那么用同一住宅区内一个先前没有被选中的户主代替。

1.4.2.3 HT逆概率加权法

CC分析通常会定义不相合的估计或给出有严重偏差的统计分析结果,然而对