复杂数据下两类回归模型的统计推断

闫 莉 著

科学出版社

复杂数据下两类回归模型的统计推断

闫 莉 著

科 学 出 版 社

北 京

内 容 简 介

本书介绍基于复杂数据的部分线性模型和广义线性模型的统计推断, 主要内容包括稳健 M 估计、经验似然推断和变量选择等问题的研究. 全书共 8 章. 第 1 章概括本书所讨论的模型、数据类型和主要研究方法. 第 2 章研究随机适应误差下部分线性模型的 M 估计. 第 3 章给出鞅差序列下回归函数的估计及其渐近性质. 第 4 章研究鞅差序列部分线性模型估计的渐近性质. 第 5 章讨论鞅差序列部分线性模型的经验似然推断. 第 6 章讨论数据有测量误差时, 部分线性模型中的经验似然推断. 第 7 章研究缺失数据广义线性模型的统计推断. 第 8 章讨论高维数据广义线性模型的变量选择问题.

本书可供高等院校数学、概率统计、生物统计和计量经济等相关专业的师生使用, 也可供数学、生物、医学、经济、金融等领域的科技人员参考.

图书在版编目(CIP)数据

复杂数据下两类回归模型的统计推断/闫莉著. —北京: 科学出版社, 2016.10

ISBN 978-7-03-050123-3

I.①复… II.①闫… III.①回归分析－统计模型－统计推断 IV. O212.1

中国版本图书馆 CIP 数据核字(2016) 第 238437 号

责任编辑: 李 萍 杨 丹 孙翠勤 / 责任校对: 赵桂芬

责任印制: 徐晓晨 / 封面设计: 红叶图文

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

北京中石油彩色印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销

*

2016 年 10 月第 一 版 开本: 720×1000 B5

2016 年 10 月第一次印刷 印张: 15 1/4

字数: 307 000

定价: 85.00 元

(如有印装质量问题, 我社负责调换)

前　言

部分线性模型是一类具有较强实际背景的半参数模型, 由 Engle 等在 1986 年首次提出, 之后有大量研究与众多应用. 广义线性模型理论是对线性模型经典理论的重要推广, 可以用来分析实际问题中遇到的各种不同类型的数据. 它对应用问题的研究, 特别是在社会、经济以及生物和医学数据的统计分析中有着重要的理论和实际意义. 另外, 在实际应用中常常遇到诸如缺失数据、测量误差数据、相依数据和高维数据等类型的复杂数据, 而关于复杂数据的统计分析已成为当前统计研究的前沿与热点问题之一. 因此, 研究基于复杂数据的部分线性模型和广义线性模型中的统计推断具有一定的理论意义和实用价值.

经验似然方法是由美国斯坦福大学教授 Owen 于 1988 年提出的一种非参数统计推断方法. 这一方法与经典的正态逼近或现代的 Bootstrap 等方法比较, 在构造置信域方面有很多突出的优点. 例如, 无须估计渐近方差、置信域形状由数据自行决定、Bartlett 纠偏性以及域保持性等. 经验似然方法的出现开辟了统计推断方法的新篇章, 引起了许多研究者的关注, 他们将该方法应用到处理各种数据下的统计建模及各种应用领域. 本书将经验似然方法应用到复杂数据下部分线性模型和广义线性模型中的估计问题中, 拓宽了经验似然方法的应用领域.

本书主要在相依数据、测量误差数据、缺失数据和高维数据等复杂数据下, 研究部分线性模型和广义线性模型中的统计推断问题. 首先, 讨论了相依序列下的部分线性回归模型, 得到了参数和非参数 M 估计的渐近性质, 推广和改进了已有结果; 在此基础上, 还研究了鞅差误差序列下回归函数的估计及其渐近性质, 以及鞅差误差序列部分线性模型的统计推断问题. 其次, 研究了协变量含有测量误差时的部分线性回归模型, 考虑了未知参数的经验似然推断问题. 另外, 将经验似然方法应用于广义线性模型, 讨论了数据有缺失时, 未知参数拟似然估计和经验似然问题. 最后, 讨论了高维数据广义线性模型的变量选择问题, 在参数维数趋于无穷时, 得到了广义线性模型的自适应 Lasso 估计和桥估计, 并证明了它们的渐近统计性质.

本书的出版得到科学出版社李萍编辑的支持和关心, 在此表示诚挚的谢意. 另

外, 要感谢武汉大学概率统计系刘禄勤教授和高付清教授在作者攻读硕士学位期间的关心和照顾. 感谢陕西师范大学数学与信息科学学院李建林教授在作者攻读博士学位期间的支持和爱护. 同时要感谢国家自然科学基金项目 (11201276) 和中央高校基本科研业务费专项资金项目 (GK201503012, GK201503015) 提供的支持. 还要感谢陕西师范大学数学与信息科学学院的领导和同事, 感谢他们一直以来的支持和关心. 最后特别感谢家人的理解和支持.

由于作者水平有限, 书中疏漏或不妥之处在所难免, 恳请广大读者不吝赐教.

闫　莉

2016 年 3 月

目　　录

第1章 绪 论

在实际应用中, 需要经常分析来自不同领域的实际数据. 为了更好地对它们进行分析和讨论, 首先需要在对其实际背景研究的基础上建立统计模型. 这样的统计模型只是对这些数据的近似表达, 而好的统计模型的建立能够较好地解释数据并预测, 所以统计学家的目标就是寻求更加合理的统计模型, 其中统计回归模型越来越受到人们的重视.

本书研究的第一个回归模型是部分线性模型 (partly linear model, PLM), 是半参数回归模型 (semi-parametric regression model) 的一种. 它是由 Engle 等于 1986 年在研究用电需求量与电价、收入以及气温等变量之间的关系时引入的 [1], 之后有大量研究和众多应用. 部分线性模型将回归函数分为参数和非参数结构, 其中参数结构为关于参数的线性形式, 它的出现受到了统计学界的极大关注.

对于部分线性模型和半参数回归的系统介绍, 可参见 2000 年 Härdle, Liang 和 Gao 的关于部分线性模型的专著 [2], 以及 2003 年 Ruppert, Wand 和 Carroll 的关于半参数回归的专著 [3].

本书研究的第二个回归模型是广义线性模型 (generalized linear model, GLM), 它是常见的线性回归模型的重要推广. 广义线性模型可用于分析连续数据, 更重要的是分析如计数数据和属性数据等离散型数据. 这在实际问题的研究中, 对社会管理、经济金融、生命科学等数据的统计分析有重要的意义.

广义线性模型的个别特例起源很早, 1919 年 Fisher 曾使用广义线性模型的个别特例来分析社会学中的一些现象. 二十世纪四五十年代 Berkson 等也曾使用过广义线性模型中最重要的特例 Logistic 模型来研究社会和经济中的一些问题. 自 1972 年统计学家 Nelder 和 Wedderburn 引入广义线性模型一词以来 [4] , 其相关研究工作逐渐增加. 1983 年统计学家 McCullagh 和 Nelder 出版了系统论述广义线性模型的专著 [5], 并于 1989 年再版. 该书对广义线性模型的基本知识和理论有详细的论述, 但应用举例较少. 1994 年, Fahrmeir 出版了关于广义线性模型中的多元统计建模的专著 [6], 该书的应用举例较多, 内容广泛, 但缺乏深度. 两本书都没有从数

学上严格证明所列的理论结果. 1985 年, Fahrmeir 与 Kaufmann 首次系统讨论了广义线性模型参数极大似然估计的强弱相合性和渐近分布问题 [7], 并从理论上给予了严格证明.

在当今信息快速发展的时代, 人类面临着更加复杂的人口、环境、经济、医药卫生等科学和社会问题. 获得的数据结构十分复杂, 并且具有高维 (high dimensional)、相依 (dependent) 以及数据不完全 (incomplete) 等特点. 这给统计学提出了更多更复杂的问题, 同时统计学家也面临着巨大的挑战和机遇, 其中基于高维、相依和不完全数据 (incomplete data) 等复杂数据的统计建模和统计分析是当今统计学界的一个前沿和热点研究问题 [8].

本书主要研究在相依数据 (dependent data)、缺失数据 (missing data)、测量误差数据 (measurement error data) 以及高维数据 (high-dimensional data) 等复杂数据下, 部分线性模型以及广义线性模型中的统计推断问题. 本书的研究成果将为部分线性模型和广义线性模型在生物医学以及计量经济学等领域应用过程中, 对相依数据、测量误差数据、缺失数据以及高维数据的统计分析提供一定的理论依据和方法支撑, 因此有着一定的理论意义和应用价值. 下面将对部分线性模型和广义线性模型的研究背景和现状, 以及关于相依数据、缺失数据、测量误差数据和等复杂数据的背景和处理方法进行介绍.

1.1 模 型 介 绍

1.1.1 部分线性模型

为了解决具体的实际问题, 统计学者提出了不同类型的半参数回归模型. 例如, 部分线性模型、单指标模型、部分线性单指标模型、变系数模型以及半参数变系数部分线性模型等. 由于部分线性模型在半参数回归模型的重要地位, 本书考虑部分线性模型. 它于 1986 年由 Engle 等在处理用电需求与气温变化之间的关系时引入 [1]. 这个模型自引入以来, 在工农业、经济管理、医疗卫生、金融等领域得到了广泛的应用, 受到了理论和应用研究者的广泛关注和重视.

考虑如下的部分线性模型

$$Y_i = X_i^{\mathrm{T}}\beta + g(T_i) + e_i, \quad 1 \leqslant i \leqslant n, \tag{1.1.1}$$

其中 Y_i 是响应变量, $X_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}} \in \mathbb{R}^p$, $T_i \in [0,1]$ 是协变量或解释变量, (X_i, T_i) 是独立同分布的随机设计点列或固定的设计点列, $\beta = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ 是 p 维未知参数向量, $g(\cdot)$ 是定义在 $[0,1]$ 上的未知光滑函数, e_i 是随机误差, 一般假定 $E(e_i) = 0$ 和 $\mathrm{Var}(e_i) = \sigma^2 > 0$ (未知).

Engle 等用部分线性模型分析了气温和用电量的关系, 他们基于如下的数据进行了研究.

响应变量 y 表示 4 个城市的月销售电量, x_1 和 x_2 分别表示月电价和收入, t 表示日平均气温, $x_3, \cdots, x_{13}$ 表示 11 个月的虚变量(dummy variable), Engle 等首先考虑了 t 对用电需求 y 的影响是一非参数光滑函数 $g(t)$, 然后考虑了 $x_1, \cdots, x_{13}$ 对 y 的影响是线性函数 $\sum_{j=1}^{13} \beta_j x_j$, 这两部分通过可加的方式共同影响 y, 模型为

$$y = \sum_{j=1}^{13} \beta_j x_j + g(t) = X^{\mathrm{T}}\beta + g(t),$$

其中 $X = (x_1, \cdots, x_{13})^{\mathrm{T}}$.

对部分线性模型 (1.1.1) 的研究主要是未知参数 β、未知的误差方差 σ^2 和未知函数 $g(\cdot)$ 的估计以及它们的大样本性质, 主要问题包括: 第一, 参数 β 估计的相合性问题, 包括估计的强弱相合性, 这是因为参数估计的相合性是估计量良好性质的最基本要求. 第二, 参数 β 和误差方差 σ^2 估计的收敛速度、渐近分布以及 Berry-Esseen 界限等问题. 收敛速度主要看参数估计能否达到重对数律的要求; 渐近分布主要考虑其极限分布为正态分布的情况; Berry-Esseen 界限是指若其估计具有渐近正态分布, 考虑其向正态逼近的速度界限的问题. 第三, 参数分量和非参数分量的稳健 M 估计. 所谓估计的稳健性, 是指当实际数据与所研究模型相符时该估计具有良好性质, 而有较小偏差时, 对该估计的性质也影响不大. 当有较大偏离时, 其性质可能会变差, 但也基本可以接受. 第四, 非参数函数 g 的估计及其收敛速度问题, 主要考察在估计非参数函数 g 时选取不同的方法, 如核估计法、近邻估计法、样条估计法以及局部多项式估计法等, 并讨论所构造估计量的收敛速度, 特别是最优收敛速度问题 [11].

自 Engle 等提出部分线性模型之后, 众多研究者对上述问题进行了研究. 例如, 在 (x_i, t_i) 为独立同分布随机点列的情形, Heckman, Rice, Speckman, Chen, Gao 以及 Hamilton 等先后讨论了用不同的估计方法对非参数函数 g 进行估计时, 参数 β 的最小二乘估计的渐近性质和非参数函数估计的收敛速度等问题 [12–17]. 钱伟民

和柴根象将小波方法应用于部分线性模型, 得到了回归参数 β、误差方差 σ^2 和非参数函数 g 的小波估计及其渐近统计性质 [18, 19]. 由于最小二乘估计不稳健, 施沛德和李国英将稳健 M 估计方法应用于部分线性模型, 得到了参数和非参数部分的 M 估计, 其中非参数函数 g 由逐段多项式逼近, 他们得到了参数 β 的 M 估计的渐近分布为正态分布, 同时得到了参数 β 和非参数函数 g 的 M 估计的收敛速度, 其中参数 M 估计达到了 $n^{-1/2}$ 的收敛速度, 非参数函数 g 的 M 估计的收敛速度达到了非参数回归的最优收敛速度 [20, 21]. 张日权和王静龙讨论了非参数函数 g 由局部线性方法逼近时, 参数 β 和非参数函数 g 的 M 估计及其渐近统计性质 [22]. 童行伟等考虑了部分线性模型中的非参数函数 g 由光滑 B 样条函数逼近, 参数 β 的 Huber-Dutter 估计的渐近性质 [23].

在 (x_i, t_i) 为固定非随机点列的情形下, 许多研究者基于非参数函数取一类非参数估计, 包括常见的核估计和近邻估计, 得到了参数 β 的最小二乘估计和加权最小二乘估计的渐近正态性以及非参数函数 g 的估计的强弱收敛速度等问题 [16–26]. 在 M 估计研究方面, 施沛德和滕新东研究了固定设计下的部分线性模型, 得到了参数 β 的 M 估计的渐近性质 [27].

关于部分线性模型大样本性质的详细研究, 可参见高集体的博士论文 [28]. 关于 M 估计的详细讨论, 可参见施沛德的博士论文 [29]. 关于估计的渐近有效性, 参见梁华的博士论文 [30]. 关于不完全数据部分线性模型的统计推断, 参见王启华的博士论文 [31](该论文曾获 1999 年首届全国优秀博士学位论文). 2000 年 Härdle 等详细讨论了部分线性模型中关于参数和非参数若干估计的统计推断问题 [2].

1.1.2　广义线性模型

广义线性模型从形式上直接推广了正态线性模型, 它不仅可以对连续数据进行统计分析, 还可对如属性数据和计数数据等离散数据类型进行分析. 这对于广义线性模型在实际问题中的应用, 如对社会、经济、生物和医学数据的统计分析有重要的意义. 广义线性模型的提出开辟了一个新的领域, 并在应用中发挥了重要的作用.

早期的广义线性模型出现在 19 世纪初, 用来分析人口增长等社会现象. 其中, 研究较多的是 Logistic 模型和 Probit 模型, 在相当长的时期内, 这两类模型只作为研究个别社会和经济问题的一种工具. 直到 1972 年, Nelder 和 Wedderburn 正式引入广义线性模型的概念, 并利用极大似然方法处理广义线性模型中的估计问题 [4].

1974 年, Wedderburn 首次考虑了广义线性模型中响应变量的分布未知, 但具有期望和方差的情形, 提出了广义线性模型中的拟似然方法 [32]. 自此研究工作逐渐增加, 广义线性模型也成为实际数据建模的重要工具.

下面介绍广义线性模型的引入. 经典的线性模型为

$$y_i = z_i^{\mathrm{T}}\beta + e_i, \quad i = 1, \cdots, n,$$

其中 $y_i \in \mathbb{R}$ 为响应变量, $z_i = z(x_i)$ 是依赖于协变量 x_i 的设计向量, β 是未知的参数向量, 独立误差序列 $e_i \sim N(0, \sigma^2)$.

为了更好地理解引入广义线性模型的意义, 把上述线性模型用另一种方式来描述. 响应变量 y_i 相互独立且服从正态分布, 即

$$y_i \sim N(\mu_i, \sigma^2), \quad i = 1, \cdots, n,$$

且 $\mu_i = E(y_i)$. 均值 μ_i 与设计向量 z_i 的关系如下:

$$\mu_i = z_i^{\mathrm{T}}\beta, \quad i = 1, \cdots, n.$$

如果协变量 x_i 是独立随机变量序列, 上述正态分布应该理解为在给定协变量 x_i 的条件下, 响应变量 y_i 的条件分布, 并且 y_i 是条件独立的.

考虑响应变量 y 是多维的情形, 总结通常的线性回归的特征如下:

(1) $E(y) = \mu = Z^{\mathrm{T}}(x)\beta$ (关于 β 为线性), 这里 β 为 p 维未知参数, y 为 q 维响应变量, x 为协变量或解释变量. $Z(x)$ 为 x 的 $p \times q$ 设计矩阵 (例如, 若 x 为 1 维, $z^{\mathrm{T}}(x)$ 可以是 $(1, x), (1, x, x^2)$ 等. 若 x 是多维, $z^{\mathrm{T}}(x)$ 可取 $(1, x^{\mathrm{T}})$ 等), 简记为 Z.

(2) 变量 $x, Z(x), y$ 多取连续值的情形, 如工农业生产中的产品产量、人的身高、体重等特征测量.

(3) 响应变量 y 的分布为正态或近似于正态分布.

线性回归模型应该说是到目前为止研究的最为广泛的统计模型之一, 理论上已经达到了一个相当完善的程度. 可以说, 上述特点在线性模型的理论和应用发展中起着很重要的作用. 当研究数据有正态分布数据的特点且呈现某种线性形式时, 研究者一般采用上述线性模型进行讨论. 但在很多实际问题中, 如对于计数数据以及属性数据等离散型数据的分析, 可能不具有以上线性模型的特点, 若用该模型进行统计建模和分析, 则可能会导致较大的偏差, 因此很难把它们纳入经典线性模型的理论框架.

基于此, 广义线性回归从线性回归的三个特征分别进行推广:

(1) $E(y)=\mu=h(Z^{\mathrm{T}}\beta)$, h 一般假定为已知的非线性函数, 且是一个充分光滑的单调函数. $g=h^{-1}$(h 的反函数) 称为联系函数 (link function), 满足

$$g(\mu)=Z^{\mathrm{T}}\beta,$$

注意此时 $E(y)$ 已不等于 $Z^{\mathrm{T}}\beta$, 而是关于 $Z^{\mathrm{T}}\beta$ 的某一函数.

(2) 变量 $x, Z(x), y$ 可取连续或离散值的情形, 具体在实际问题的研究中更多的是针对离散值的情形, 如 $\{0,1\}$ 以及 $\{0,1,2,\cdots\}$ 等.

(3) 响应变量 y 的分布为指数型分布, 其具体形式可表示为

$$C(y)\exp(\theta^{\mathrm{T}}y-b(\theta))\mathrm{d}\nu(y),\quad \theta\in\Theta\ (\text{参数空间}),\tag{1.1.2}$$

其中 θ 称为自然参数, $b(\theta)$ 为 θ 的已知函数. ν 为一测度. 常见的情形有:

① 当 y 为连续变量时, $\mathrm{d}\nu(y)$ 为 Lebesgue 测度: $\mathrm{d}\nu(y)=\mathrm{d}y$.

② 当 y 为离散变量时, y 取有限个值 $a_1,\cdots,a_m$ 或可列个值 $a_1,a_2,\cdots$, 这时

$$\nu(\{a_i\})=1,\ i=1,\cdots,m \text{ 或 } \nu(\{a_i\})=1,\ i=1,2,\cdots.$$

可见正态分布是指数型分布的一个特例.

在广义线性模型的研究中, 大家所关注的主要问题是未知参数 β 的估计及其渐近统计性质. 在标准的广义线性模型中 (即响应变量的分布属于指数型分布), 一般利用极大似然方法, 得到参数 β 的极大似然估计 (maximum likelihood estimate, MLE). MLE 性质的好坏, 如相合性和渐近正态性, 直接影响到广义线性模型统计方法的研究及其在实际中的应用. Fahrmeir 和 Kaufmann 在 1985 年建立了广义线性模型极大似然估计的大样本理论, 讨论了广义线性模型参数的极大似然估计的存在性、强弱相合性和渐近正态性成立的一般条件 [7]. Qian 和 Wu 研究了 Logistic 模型中参数极大似然估计的重对数律和模型选择问题 [33]. 丁洁丽和陈希孺改进和推广了 Fahrmeir 和 Kaufmann [7] 的工作, 讨论了广义线性模型极大似然估计的强相合性和渐近正态性 [34, 35].

上述讨论都是在响应变量 y 服从指数型分布的假定下进行的. 这个假定是由于经常要对离散数据进行统计分析, 而常见的离散型数据的分布多属于指数型分布, 如二项分布、Poisson 分布等. 但是, 在一些实际问题的分析中, 所获得的数

据分布可能并不属于指数型分布. 1974 年 Wedderburn 发现若有关于 y 的期望和方差函数的正确设定 (甚至方差函数也可以未知), 则仍可利用类似极大似然的方法来处理相关统计推断问题, 于是在不要求指数型分布的假定下提出了拟似然方法 [32]. 自此, 关于广义线性模型的理论和应用得到了长足的发展, 出现了一大批研究成果. 关于拟似然方法及其研究进展将在 1.3.2 小节中详细介绍.

1.2 复杂数据类型

1.2.1 相依数据

在当今科技和信息快速发展的时代, 一些数据具有很复杂的结构, 并且这些数据之间往往不是独立的, 而是具有某种相依关系. 于是研究者把独立数据下的相关理论成果推广到了相依数据的情形 [36], 扩大了其应用范围.

在实际应用中, 众多研究者对各种相依数据的统计模型进行了详细的研究 [37], 特别是在相依数据部分线性模型的研究方面取得了重要进展. 例如, 施沛德和郑忠国考虑了基于严平稳 β 混合观测的部分线性模型, 讨论了 M 估计的收敛速度问题, 并在一定的条件下, 证明了参数分量 β 的 M 估计具有渐近正态性, 非参数分量 g 的 B 样条 M 估计达到了非参数回归的最优收敛速度 [38]. Gao 和 Anh 考虑了误差为长相依序列下, 半参数回归模型的估计问题 [39]. 任哲和陈明华与 Baek 和 Liang 在误差序列为负相关 (NA) 的情形下, 得到了部分线性模型中参数 LS 估计的渐近性质 [40–42]. 凌能祥考虑了线性过程误差下半参数回归模型的相关问题和鞅差误差序列下部分线性模型参数估计的相合性 [43, 44]. 在误差为鞅差序列的情形下, 李国亮讨论了几类误差为鞅差的回归模型的大样本理论 [45]. 李国亮和刘禄勤利用核估计方法估计非参数函数 g, 得到了部分线性模型参数 β 的 LS 估计的强相合性 [46]. Chen 和 Cui 考虑了具有鞅差误差下部分线性模型的经验似然推断 [47]. 于卓熙系统讨论了若干相依误差下部分线性模型的经验似然推断问题 [48].

1.2.2 缺失数据

在实际应用的研究中, 常出现因为各种主观和客观原因而导致的数据不完整. 例如, 在一些涉及个人隐私 (如年龄、工资、信用) 等敏感问题的调查中, 某些被访问者可能不愿提供所需要的信息; 在医学疾病追踪的试验中由于被追踪者意外死亡或失踪等产生数据丢失; 还有一些是调研人员的主观原因 (如责任心不强) 导致收

集到的信息不完整等. 总之, 数据缺失的现象在医疗研究、市场调研、金融经济研究、民意调查以及生物、化学、物理等实验科学中经常见到.

Little 与 Rubin 详细探讨了关于数据缺失的类型、数据缺失的机制以及处理缺失数据的方法 [9], 其中缺失机制的研究对于处理缺失数据中的相关问题显得尤其重要. 通常数据缺失的机制分为三类: 一是完全随机缺失 (missing completely at random, MCAR); 二是随机缺失 (missing at random, MAR), MCAR 和 MAR 中数据缺失的出现都是随机的, 故可称之为可忽略的缺失机制; 三是非随机缺失 (not missing at random, NMAR), 也可称之为不可忽略缺失 (non-ignorably missin, NI). 下面详细介绍这三种缺失机制.

设 Y 为研究的数据, Y_{mis} 表示该数据缺失的部分, Y_{obs} 表示该数据可观测到的部分, 则 $Y = (Y_{\text{obs}}, Y_{\text{mis}})$. 记 M 为缺失数据指示变量矩阵, 它是由 0 和 1 组成的矩阵. 其中若 Y 的元素能观测到, 相应 M 的元素取为 1; 若 Y 的元素缺失, 相应 M 的元素取为 0. 记 $f(M|Y,\theta)$ 表示给定 Y 时 M 的条件分布, 可以用 $f(M|Y,\theta)$ 来刻画数据缺失的机制类型, 其中 θ 是未知参数. 上述三种不同的缺失机制类型可描述如下 [9]:

(1) 如果数据的缺失不依赖于 Y 的值 (不管是缺失的还是观测到的), 即

$$f(M|Y,\phi) = f(M|\phi), \quad \forall\, Y, \phi,$$

则缺失机制为完全随机缺失, 即这种缺失不依赖于数据本身. 这是缺失数据问题中最简单的一种, 表示缺失现象是完全随机发生的, 指该变量的缺失与否没有系统差异性. 这种缺失可以忽略, 对统计分析没有大的影响. 由于在实际问题中, 没有原因的缺失 (如 MCAR) 不太常见, 所以这个缺失机制的应用很少.

(2) 如果缺失仅依赖于 Y_{obs}(观测到的), 不依赖于 Y_{mis}(缺失的), 即满足

$$f(M|Y,\phi) = f(M|Y_{\text{obs}},\phi) \quad \forall\, Y_{\text{mis}}, \phi,$$

则称其为随机缺失. 即这种缺失只依赖于数据观测到的部分 Y_{obs}. 这是在实际应用中较为常见的缺失数据机制, 也是在众多研究中经常假设的数据缺失机制类型.

(3) 如果缺失依赖于 Y_{mis}(缺失的), 也有可能与 Y_{obs}(观测到的) 有关, 则称其为非随机缺失或不可忽略机制. 这种情形在缺失数据的处理问题中最为复杂, 研究者甚少.