# 缺失数据的模型检验及其应用

许王莉 著

科学出版社

# 缺失数据的模型检验及其应用

许王莉　著

科学出版社

北京

## 内 容 简 介

本书主要研究缺失数据模型的检验问题. 全书共分为 8 章. 第 1 章主要介绍数据的不同缺失机制, 包括协变量缺失和因变量缺失, 以及在不同缺失机制下常见的统计分析方法. 第 2 章介绍一些常见的检验方法, 主要包括蒙特卡罗检验和得分类型的检验. 在蒙特卡罗检验这部分, 着重介绍参数和非参数蒙特卡罗检验方法. 第 3 章介绍在数据不存在缺失的情况下, 几种常见模型的检验方法及其性质. 第 4 章是关于在因变量缺失时, 部分线性模型中非线性部分是否符合某类参数结构的拟合优度检验问题. 第 5 章讨论协变量随机缺失时, 广义线性模型本身的拟合优度检验问题. 第 6 章对于变系数模型, 在响应变量缺失的情况下, 研究变系数部分是否具有一定参数结构的检验. 第 7 章研究的是协变量缺失时候的统计推断问题. 第 8 章的主要内容是因变量随机缺失的情况下, 变系数模型本身的拟合优度检验问题. 第 4 章到第 8 章的检验统计量主要采用蒙特卡罗检验和得分类型的检验.

本书可作为概率统计、应用数学等专业高年级本科生及研究生教材, 也可供应用数据分析等相关专业人士参考.



图书在版编目(CIP)数据

缺失数据的模型检验及其应用/连玉莉著. —北京: 科学出版社, 2014.1

ISBN 978-7-03-039233-8

Ⅰ. ①缺… Ⅱ. ①许… Ⅲ. ①数据模型 Ⅳ. ①TP311.13

中国版本图书馆 CIP 数据核字 (2013) 第 290712 号

责任编辑: 李 欣 / 责任校对: 宣 慧
责任印制: 赵德静 / 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

文林印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2014 年 1 月第 一 版 开本: 720×1000 1/16
2014 年 1 月第一次印刷 印张: 8 3/4
字数:176 000

**定价: 45.00 元**

(如有印装质量问题, 我社负责调换)

# 前　言

在应用研究领域，缺失数据是一类常见的数据. 引起数据缺失的原因很多，比如，获取某些数据花费的代价很大；研究个体由于药物的副作用而停止试验等. 数据的缺失机制主要有完全随机缺失、随机缺失和非随机缺失三种. 用缺失数据拟合模型的统计推断已经有很多研究，但是大部分的研究还是在模型的估计方面. 如果用错误的模型拟合数据，得到的结果可能是不合理的. 所以关于模型的检验具有非常重要的意义，本书主要研究缺失数据模型的假设检验问题，其主要结论大部分是作者和合作者郭旭、朱力行已有的研究成果.

书中的各章是关于不同的模型在数据缺失下的检验，主要包括广义线性模型、半参数模型、变系数模型. 既包括非参数模型的检验，也包括参数模型的检验. 既包括协变量缺失的检验，也包括响应变量缺失的检验. 本书所研究的缺失机制都是随机缺失. 对于完全随机机制的情况，可以采用完全数据下的统计量. 对于非随机缺失机制下数据的假设检验问题，相关的文献较少，具有进一步研究的必要性.

本书统计量的构造主要采用经验过程和得分类型的方法. 这两类重要的方法在完全数据的情况下已经有很多应用. 由于经验过程的方法构造统计量的临界值很难计算得到，书中采用蒙特卡罗 (Monte Carlo) 逼近统计量的分布，此方法已经成为统计学中非常重要的方法. 这个方法的本质就是通过产生参考数据得到条件统计量，使得条件统计量在原假设和具有一定形式的备择假设下可以逼近原假设下统计量的分布. 对具有半参数结构的模型和分布，产生参考数据是一个具有挑战性的难题.

科学出版社 2008 年出版的朱力行和许王莉的《非参数蒙特卡罗检验及其应用》对此方法有非常详细的讨论. 书中提出一种新的方法，即非参数蒙特卡罗检验 (NMCT)，用于处理半参数和非参数结构的模型. 非参数蒙特卡罗方法是基于参数蒙特卡罗方法和其他蒙特卡罗逼近方法，比如，自助法提出来的一种逼近检验统计量分布的方法. 现在已经广泛应用于各种不同的统计推断方法中.

本书所设定的读者范围较广，适合学习缺失数据分析的初学者，也适合应用蒙特卡罗来分析实际问题的研究者，还可以作为研究生的参考教材. 本书中有定理的详细证明，也有实际数据的分析，可作为理论或者应用工作者的参考书籍.

本书得以出版, 感谢国家自然科学基金 (编号：11071253) 和北京市科技新星计划 (编号：2010B066) 的资助. 感谢香港浸会大学朱力行教授在本书的撰写过程中给予的宝贵建议. 感谢余味、牛翠珍、郭旭等同学为本书搜集、整理资料，并且对

本书进行校正和排版.

由于作者水平所限, 书中难免有不妥之处, 欢迎读者批评指正. 来函请发至 wxu.stat@gmail.com.

许王莉

2013 年 5 月

# 符 号 表

| | |
|---|---|
| $\mathbf{R}$ | 实数集合 |
| $x \in \mathbf{R}^n$ | 输入和 $n$ 维欧氏空间 |
| $y \in \mathcal{Y}$ | 输出和输出集合 |
| $(x_i, y_i)$ | 第 $i$ 个训练点 |
| $T = \{(x_1, y_1), \cdots, (x_l, y_l)\}$ | 训练集 |
| $l$ | 训练点个数 |
| $[x]_i,\ [x_i]_j$ | 向量 $x$ 的第 $i$ 个分量, 向量 $x_i$ 的第 $j$ 个分量 |
| $x = \Phi(x)$ | Hilbert 空间中的向量和输入空间到 Hilbert 空间的映射 |
| $[x]_i,\ [x_i]_j$ | 向量 $x$ 的第 $i$ 个分量, 向量 $x_i$ 的第 $j$ 个分量 |
| $(x \cdot x'), (\mathrm{x} \cdot \mathrm{x}')$ | $x$ 与 $x'$ 的内积, x 与 x′ 的内积 |
| $\mathcal{H}$ | Hilbert 空间 |
| $w$ | $\mathbf{R}^n$ 空间中的权向量 |
| $w_i$ | 权向量 $w$ 的第 $i$ 个分量 |
| $w$ | Hilbert 空间中的权向量 |
| $w_i$ | 权向量 $w$ 的第 $i$ 个分量 |
| $b$ | 阈值 |
| $K(x, x')$ | 核函数 $(\Phi(x) \cdot \Phi(x'))$ |
| $K$ | 核矩阵 (Gram 矩阵) |
| $\|\cdot\|_p$ | $p$ 范数 |
| $\|\cdot\|$ | 2 范数 |
| $\|\cdot\|_1$ | 1 范数 |
| $h$ | VC 维 |
| $C$ | 惩罚参数 |
| $\xi$ | 松弛变量 |
| $\xi_i$ | 松弛变量的第 $i$ 个分量 |
| $\alpha$ | 对偶变量, Lagrange 乘子 |
| $\alpha_i$ | 对偶变量的第 $i$ 个分量 |
| $\beta$ | 对偶变量, Lagrange 乘子 |
| $\beta_i$ | 对偶变量的第 $i$ 个分量 |
| $P(\cdot)$ | 通常表示概率分布或概率 |

# 目　录

# 第 1 章　缺 失 数 据

## 1.1　协变量缺失机制

在医学和流行病学等应用领域，协变量缺失处处存在. 数据缺失机制对于数据的统计推断是非常重要的，不同的缺失机制会导致不同的似然函数，进而得出不同的统计推断结果. 缺失机制的概念是由 Rubin (1976) 提出的，主要分为三大类：随机缺失 MAR (missing at random)、完全随机缺失 MCAR(missing completely at random) 和非随机缺失 NMAR(not missing at random)，其中非随机缺失也称为不可忽略缺失 (nonignorable missingness).

用 $Y$ 表示响应变量, $(X, Z)$ 表示协变量，$\delta$ 表示协变量 $X$ 是否缺失，等于 1 表示观测到，等于 0 表示缺失. 以下给出协变量 $X$ 三种不同缺失的定义.

(1) 完全随机缺失，也就是协变量 $X$ 是否缺失与协变量 $Z$ 和响应变量 $Y$ 没有任何关系. 用公式表示为 $P(\delta = 1|Y, X, Z) = P(\delta = 1)$.

(2) 随机缺失，也就是协变量 $X$ 缺失只和协变量 $Z$ 和响应变量 $Y$ 有关，与 $X$ 本身没有关系. 用公式表示为 $P(\delta = 1|Y, X, Z) = P(\delta = 1|Y, Z)$.

(3) 非随机缺失，在这种缺失机制下，协变量 $X$ 缺失可能与 $Z$ 和 $Y$ 有关，也可能与 $X$ 本身有关.

下面给出一个模拟说明上述所提到的三种不同的协变量缺失机制. 假定数据来自如下模型

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1.1.1}$$

设定 $(\beta_0, \beta_1) = (1, 1)$，$X$ 和 $\varepsilon$ 独立且都来自标准正态分布. 如下三种不同缺失函数分别表示三种不同的缺失机制.

(1) $P(\delta = 1) = 0.6$;

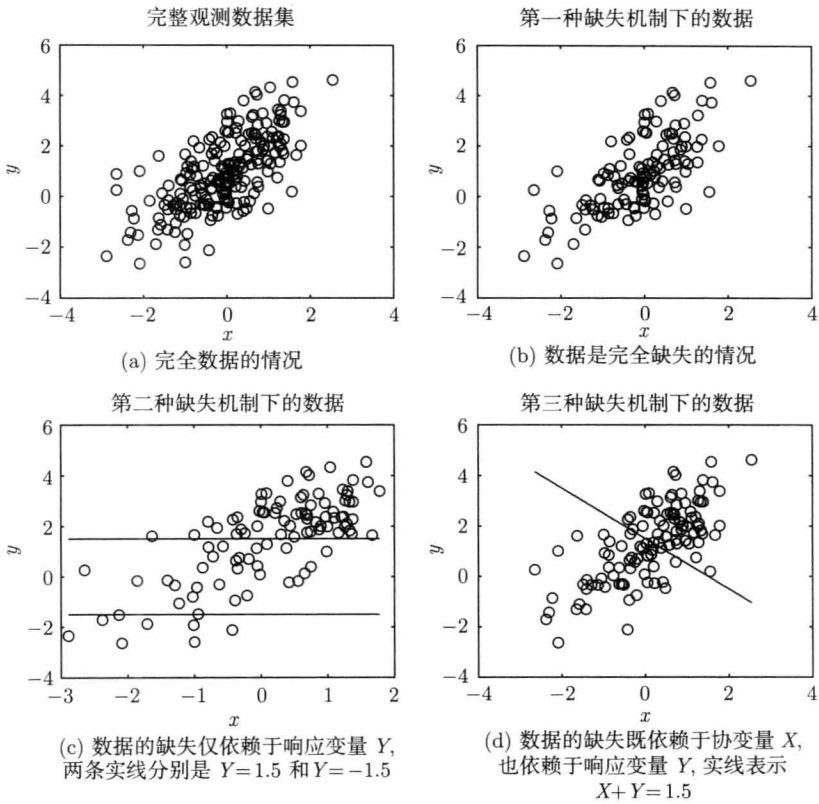
(2) $P(\delta|Y) = 0.30$，如果 $|Y| \leqslant 1.5$, 否则 $= 0.95$;

(3) $P(\delta|Y, X) = 0.40$，如果 $X + Y \leqslant 1.5$, 否则 $= 0.90$.

这三种缺失机制分别是完全随机缺失、只依赖响应变量 $Y$ 的缺失，以及既依赖于 $X$ 也依赖 $Y$ 的缺失. 在这三种不同的缺失机制下，数据缺失的概率都等于或者约等于 0.6.

我们随机产生 200 组数据，图 1.1.1 (a), (b), (c) 和 (d) 分别表示数据完全观测到的情况, 第一、第二以及第三种缺失机制下得到的数据. 从图 1.1.1 中可以看出，

图 1.1.1(b) 是图 1.1.1(a) 中的数据随机缺失 40% 的数据；图 1.1.1(c) 可以明显看出在 $|Y|>1.5$ 时，缺失的概率明显小于 $|Y|\leqslant 1.5$ 的情况；图 1.1.1(c) 也可以看到在 $X+Y\leqslant 1.5$ 的缺失概率明显小于其他情况.




(a) 完全数据的情况

(b) 数据是完全缺失的情况

(c) 数据的缺失仅依赖于响应变量 $Y$, 两条实线分别是 $Y=1.5$ 和 $Y=-1.5$

(d) 数据的缺失既依赖于协变量 $X$, 也依赖于响应变量 $Y$, 实线表示 $X+Y=1.5$

图 1.1.1　实际数据集的散点图

下面说明不同的缺失机制对极大似然估计的影响. 假定得到的数据为 $(Y_i, X_i, Z_i, \delta_i), i=1,\cdots,n$. 假设 $(Y_i, X_i, Z_i, \delta_i)$ 独立同分布，则基于此数据的似然函数为

$$\prod_{i=1}^{n} f(Y_i, X_i, Z_i, \delta_i)=\prod_{i=1}^{n} f(Y_i, X_i, Z_i)\prod_{i=1}^{n} f(\delta_i|Y_i, X_i, Z_i), \tag{1.1.2}$$

这里 $f(Y_i, X_i, Z_i)$ 是 $(Y_i, X_i, Z_i)$ 的联合密度函数，$f(\delta_i|Y_i, X_i, Z_i)$ 是二值指示 $\delta_i$ 的条件二项分布的密度函数. 在缺失机制为 MCAR 时，$f(\delta_i|Y_i, X_i, Z_i)=f(\delta_i)$, 此时式 (1.1.2) 可以化简为

$$\prod_{i=1}^{n} f(Y_i, X_i, Z_i, \delta_i)=\prod_{i=1}^{n} f(Y_i, X_i, Z_i)\prod_{i=1}^{n} f(\delta_i). \tag{1.1.3}$$

在缺失机制为 MAR 时，$f(\delta_i|Y_i,X_i,Z_i)=f(\delta_i|Y_i,Z_i)$，此时式 (1.1.2) 可以化简为

$$\prod_{i=1}^{n}f(Y_i,X_i,Z_i,\delta_i)=\prod_{i=1}^{n}f(Y_i,X_i,Z_i)\prod_{i=1}^{n}f(\delta_i|Y_i,Z_i). \tag{1.1.4}$$

在缺失机制是 NMAR 的情况下，条件密度 $f(\delta_i|Y_i,X_i,Z_i)$ 不能进一步简化，该概率依赖于缺失的 $X_i$ 和没有缺失的 $(Y_i,Z_i)$. 假定有 $m$ 个个体可以观测到协变量 $X$，对于数据的统计推断，一种简单的方法就是仅仅利用这 $m$ 个观测到的数据进行统计分析，这种分析方法称为 CC (completed cases) 方法. 如果缺失机制是 MCAR，用这一方法得到的统计推断结果是合理的，因为这 $m$ 个数据可以看成从 $(Y,X,Z)$ 的分布中独立得到. 相对于全部 $n$ 个数据的统计结论，由于数据量的减少，估计的有效性会降低. 如果缺失机制是 MAR 或者 NMAR，用 CC 的估计方法直接作统计推断可能会有误差. 下面给出一个模拟说明这一结论. 由于 NMAR 这种缺失机制比较复杂，相关的研究较少，本书主要研究考虑缺失机制是 MAR 的情况.

我们仍然采用模型 (1.1.1)，参数和缺失机制的设置都是一致的. 在三种不同的缺失机制下，我们采用 CC 的方法估计，所得的结果见表 1.1.1. 表 1.1.1 中 $\hat{\beta}_{\mathrm{C}}$，$\hat{\beta}_{\mathrm{MCAR}}$，$\hat{\beta}_{\mathrm{MAR}}$ 和 $\hat{\beta}_{\mathrm{NMAR}}$ 分别表示基于完整数据、第一、第二，以及第三种缺失机制下非缺失数据用 CC 方法得到的模拟结果. 表 1.1.1 中分别研究了 $n=100$ 和 $n=200$ 的情况. 从表 1.1.1 中可以看出，参数 $(\beta_0,\beta_1)$ 的估计 $\hat{\beta}_{\mathrm{C}}$ 和 $\hat{\beta}_{\mathrm{MCAR}}$ 都是无偏估计，只是 $\hat{\beta}_{\mathrm{MCAR}}$ 的标准方差要大一些. 然后 $\hat{\beta}_{\mathrm{MAR}}$ 和 $\hat{\beta}_{\mathrm{NMAR}}$ 得到的估计并不是无偏的. 进一步验证了用 CC 的方法对 MAR 和 NMAR 数据作统计推断是不合适的.

**表 1.1.1 不同缺失机制下参数的估计**

| | $n=100$ | | $n=200$ | |
|---|---|---|---|---|
| | 估计值 | 标准误 | 估计值 | 标准误 |
| $\hat{\beta}_{\mathrm{C}}$ | 1.000 | 0.010 | 1.001 | 0.005 |
| | 1.004 | 0.010 | 1.000 | 0.005 |
| $\hat{\beta}_{\mathrm{MCAR}}$ | 1.003 | 0.017 | 1.002 | 0.008 |
| | 1.006 | 0.017 | 1.000 | 0.008 |
| $\hat{\beta}_{\mathrm{MAR}}$ | 1.218 | 0.019 | 1.217 | 0.009 |
| | 1.099 | 0.017 | 1.094 | 0.008 |
| $\hat{\beta}_{\mathrm{NMAR}}$ | 1.150 | 0.018 | 1.149 | 0.009 |
| | 0.990 | 0.018 | 0.987 | 0.008 |

## 1.2　协变量缺失的处理方法

### 1.2.1　完整个体分析

1. 不加权的完整个体分析

不加权的完整个体分析是指直接将有缺失值的个体数据剔除，仅利用那些全部变量均有观测的数据. 这一方法优点是简单，但是有信息的损失. 首先是样本量变小造成的精度损失；其次，如果数据的缺失机制不是 MCAR 时，此种估计方法还会产生偏差.

假设 $\theta$ 为待估计的参数，$\hat{\theta}_{\mathrm{NM}}$ 为没有缺失值时的估计量，$\hat{\theta}_{\mathrm{CC}}$ 为使用完整个体分析得到的估计量，那么 $\hat{\theta}_{\mathrm{CC}}$ 估计量的方差可以表示为

$$\mathrm{Var}(\hat{\theta}_{\mathrm{CC}}) = \mathrm{Var}(\hat{\theta}_{\mathrm{NM}})(1 + \Delta_{\mathrm{CC}}), \tag{1.2.1}$$

其中 $\Delta_{\mathrm{CC}}$ 是由于信息损失带来的方差增加的百分数.

2. 加权的完整个体分析

用 $X_i = (x_{i1}, x_{i2}, \cdots, x_{ip})$ 表示第 $i$ 个个体的协变量观测值，$y_i$ 表示第 $i$ 个个体的响应变量观测值，记 $\pi_i$ 为第 $i$ 个个体的协变量全部观测到的概率，$C$ 表示协变量完整观测的个体的指标集. 若我们建立的模型为

$$y = f(x, \beta) + \varepsilon, \tag{1.2.2}$$

那么用加权的完整个体估计方法，参数 $\beta$ 的估计为

$$\hat{\beta} = \arg\min_{\beta} \sum_{i \in C} \pi_i D(y_i, f(x_i, \beta)), \tag{1.2.3}$$

其中 $D(\cdot)$ 表示距离函数，如平方距离函数等. 在实际中，$\pi_i$ 通常是未知的，所以用其估计值 $\hat{\pi}_i$ 来代替，这里的估计可以用核估计等估计方法. 加权的完整数据分析方法在缺失机制为 MAR 时估计是无偏的.

### 1.2.2　基于插补数据的方法

当有某个完全观测到的变量 $X_j$ 与有缺失值的变量 $X_k$ 具有很强的相关关系时，插补法是一个很好的方法. 也就是说，我们用 $X_k$ 对 $X_j$ 作回归，并用得到的回归模型来预测缺失的 $X_k$ 的值. 这个方法可以对每一个缺失项插补一个值 (单一插补)，或者在某些情形下插补多个值 (多重插补). 在本小节中先介绍单一插补的方法.

1. 单一插补

在这种方法中插补值是缺失值的预测分布的一个平均值或抽样值，也就是说，要求我们以观测到的数值为基础，建立一个模型以预测缺失值的分布. 建立模型的方法主要包括明确建模和模糊建模两类. 以下关于明确建模和模糊建模的论述主要参考 Little 和 Rubin (2002) 中的论述.

明确建模的方法有：①均值插补，即以变量观测到的部分的均值来插补缺失值；②回归插补，以有缺失值的变量对完整观测的变量作回归，用缺失变量的完整部分的数据估计回归模型的参数，再用此模型来预测缺失部分的值；③随机回归插补，用回归插补值加上一个随机项来预测缺失值. 例如，缺失变量是具有 0 和 1 两个属性的分类变量，用 Logistic 回归对确实变量和另外一些预测变量作回归，缺失项的回归预测值就是一个 0 到 1 之间的概率，而随机回归预测值就是一个以此概率抽出的 0 或 1.

模糊建模的方法通常有：①热平台插补，将缺失项的值用 “类似” 的样本点中的对应值代替，这里的 “类似” 样本点就是指用一些距离函数衡量的与缺失的样本点距离最小的样本点；②冷平台插补，用一个从其他来源的，比如以往调查中的一个完整个体值代替缺失值；③替代法，主要在抽样阶段使用此方法，是指当某个个体的属性有缺失时，对另外一个替代的个体进行调查. 以上三种模糊建模方法在一个研究中通常是综合使用的.

在选取插补方法时应该遵循以下三点原则：①要以观测到的数据为基础，并且尽量减少预测的偏差，保持观测到的变量和缺失变量之间的关系；②多变量缺失的时候应该保持缺失变量之间的联系；③如果操作方便，尽量从缺失项的预测分布来抽取，这比直接用均值估计往往能够得到更高的估计精度.

2. 多重插补

多重插补是指对每一缺失项都用一个插补向量来代替，其维数 $M \geqslant 2$，对于插补向量中的每一个插补值可以用单一插补中的方法来得到，比如随机回归插补. 这样我们可以构造 $M$ 个完整的数据集，当每一个缺失项都用其插补向量的第一个分量来代替时构成第一个数据集，$\cdots$，当每一个缺失项都用其插补向量的第 $M$ 个分量来代替时构成第 $M$ 个数据集. 最终的参数估计值就是由 $M$ 个完整数据集得到的估计值的某种综合. 多重插补是由 Rubin (1978) 首先提出的，并且得到了广泛的应用.

假设估计的目标参数为 $\theta$，我们对以上提到的每一个完整数据集用相同的方法进行估计，得到 $M$ 个估计值 $\hat{\theta}_m, m = 1, 2, \cdots, M$，其方差分别用 $V_m, m = 1, 2, \cdots, M$ 来表示，那么多重插补的估计值为

$$\hat{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m, \tag{1.2.4}$$

其方差为

$$T = V_W + \frac{M+1}{M}V_B, \tag{1.2.5}$$

其中 $V_W$ 表示插补内方差，$V_B$ 表示插补间方差，它们的表达式分别为

$$V_W = \frac{1}{M}\sum_{m=1}^{M}V_m, \tag{1.2.6}$$

$$V_B = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta} - \hat{\theta}_m)^2, \tag{1.2.7}$$

当 $\theta$ 为标量时，在大样本下 $(\theta - \hat{\theta})T^{-1/2}$ 服从自由度为 $d$ 的 $t$ 分布，其中

$$d = (M-1)\left(1 + \frac{1}{M+1}\frac{V_W}{V_B}\right)^2. \tag{1.2.8}$$

### 1.2.3　基于似然的方法

1. 有缺失值的似然函数

Rubin(1976a) 给出了用极大似然理论处理缺失数据的方法. 考虑有三个变量 $x, z, y$ 的模型，其中 $x, z$ 是协变量，$y$ 是响应变量，假设变量 $x$ 有缺失. 我们得到的数据为 $(X, Z, Y, \delta)$，其中 $X, Z, Y, \delta$ 都是维数为 $n \times 1$ 的向量，$\delta$ 是指示对应的 $X$ 的分量有无缺失的二元变量组成的 $n \times 1$ 向量，取值为 1 表示对应的 $X$ 观测到，取值为 0 表示对应的 $X$ 缺失. 另外，$X = (X_O, X_M)$，$X_O$ 表示观测到的值组成的向量，$X_M$ 表示缺失的值. 用 $\theta$ 表示待估参数，$\phi$ 表示缺失机制分布，那么

$$f(X, Z, Y, \delta|\theta, \phi) = f(X, Z, Y|\theta)f(\delta, Z, Y|X, \phi). \tag{1.2.9}$$

由于 $X$ 有缺失，所以实际观测到的数据只有 $(X_O, Z, Y, \delta)$，观测数据的分布由 $(X, Z, Y, \delta)$ 的分布积去 $X_M$ 得到，也就是

$$f(X_O, Z, Y, \delta|\theta, \phi) = \int f(X_O, X_M, Z, Y|\theta)f(\delta, Z, Y|X_O, X_M, \phi)\mathrm{d}X_M. \tag{1.2.10}$$

$(\theta, \phi)$ 的整个似然函数是正比于函数 (1.2.10) 的一个函数，也就是

$$L_{\text{full}}(\theta, \phi|X_O, Z, Y, \delta) \propto f(X_O, Z, Y, \delta|\theta, \phi). \tag{1.2.11}$$

当数据的缺失机制为 MAR 时，也就是说 $\delta$ 的分布不依赖于 $X_M$，有

$$f(\delta, Z, Y|X_O, X_M, \phi) = f(\delta, Z, Y|X_O, \phi), \tag{1.2.12}$$

那么根据函数 (1.2.10) 可以得到

$$f(X_O, Z, Y, \delta|\theta, \phi) = f(\delta, Z, Y|X_O, \phi) \times \int f(X_O, X_M, Z, Y|\theta)\mathrm{d}X_M \tag{1.2.13}$$

$$= f(\delta, Z, Y|X_O, \phi) f(X_O, Z, Y|\theta), \tag{1.2.14}$$

在这种情形下，可以得到一个较简单的似然函数

$$L_{\mathrm{ign}}(\theta|X_O, Z, Y) \propto f(\delta, Z, Y|X_O, \phi) f(X_O, Z, Y|\theta), \tag{1.2.15}$$

在似然推断中，当缺失机制为 MAR 并且参数 $\theta$ 和 $\phi$ 是独立的时候，可以忽略数据的缺失机制.

2. EM 算法

在 MAR 缺失机制下，可以通过极大化函数 (1.2.15) 而得到 $\theta$ 的 ML 估计. 但是在大多数情况下，此估计不能通过解析法从函数 (1.2.15) 中解出，因此我们考虑使用迭代算法，Newton-Raphson 算法和 Berndi 等 (1974) 所提出的算法都可以用于解决这个问题. 在数据有缺失的情况下，我们经常使用所谓的期望 —— 极大化算法，也称为 EM 算法，这是一种将依据 $L_{\mathrm{ign}}(\theta|X_O, Z, Y)$ 的 $\theta$ 的 ML 估计与基于完全数据似然 $L(\theta|X, Z, Y)$ 的 $\theta$ 的 ML 估计结合起来的算法.

EM 算法的每一步迭代都有一个 E 步 (期望步) 和 M 步 (极大化步). E 步是在给定观测数据和现有参数下，求缺失数据的条件期望，然后将缺失数据用这些条件期望代替. 假设迭代的初始值为 $\hat{\theta}_0$，那么第一个 E 步就是在 $\hat{\theta}_0$ 下，求完全数据似然的期望

$$L^{(1)}(\theta|\hat{\theta}^{(0)}) = \int L(\theta|X, Z, Y) f(X_M, Z, Y|X_O, \theta = \hat{\theta}^{(0)})\mathrm{d}X_M. \tag{1.2.16}$$

接下来进行 M 步，也就是极大化这个完全数据似然的期望来得到 $\hat{\theta}^{(1)}$，也就是

$$\hat{\theta}^{(1)} = \arg\max_{\theta} L^{(1)}(\theta|\hat{\theta}^{(0)}), \tag{1.2.17}$$

重复以上两步直到收敛，便得到了 $\theta$ 的 ML 估计.

EM 算法主要有两个缺点：①当数据缺失比例较大时，收敛可能很慢；② 对于有些问题，M 步的极大化计算仍然很困难. 因此产生了一些推广的 EM 算法，如 ECM，ECME，AECM 等，在这里就不一一介绍了.

## 1.3　响应变量缺失机制

类似于协变量的缺失机制，响应变量也有三种不同的缺失机制. 沿用前面的记号，用 $Y$ 表示响应变量, $(X, Z)$ 表示协变量，$\delta$ 表示响应变量 $Y$ 是否缺失，等于 1 表示观测到，等于 0 表示缺失. 以下给出响应变量 $Y$ 三种不同缺失的定义.

(1) 完全随机缺失，也就是响应变量 $Y$ 是否缺失与协变量 $(X, Z)$ 没有任何关系. 用公式表示为 $P(\delta = 1|Y, X, Z) = P(\delta = 1)$.

(2) 随机缺失，也就是响应变量 $Y$ 缺失只和协变量 $(X, Z)$ 有关，与 $Y$ 本身没有关系. 用公式表示为 $P(\delta = 1|Y, X, Z) = P(\delta = 1|X, Z)$.

(3) 非随机缺失，在这种缺失机制下，响应变量 $Y$ 缺失可能与 $X$ 和 $Z$ 有关，也可能与 $Y$ 本身有关.

下面给出一个模拟说明上述所提到的三种不同的响应变量缺失机制. 假定数据来仍然来自模型 (1.1.1)，即为

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

上述模型中参数的设置和随机变量的分布和模型 (1.1.1) 一致. 如下三种不同缺失函数分别表示三种不同的缺失机制.

(1) $P(\delta = 1) = 0.6$;

(2) $P(\delta|X) = 0.30$，如果 $X \leqslant 0.1$, 否则 $= 0.95$;

(3) $P(\delta|Y, X) = 0.40$，如果 $X + Y \leqslant 1.5$, 否则 $= 0.90$.

这三种缺失机制分别是完全缺失、只依赖响应变量 $Y$ 的缺失，以及既依赖于 $X$ 也依赖于 $Y$ 的缺失. 在这三种不同的缺失机制下，数据缺失的概率都等于或者约等于 0.6.

我们随机产生 200 组数据，如图 1.3.1 所示. 从图 1.3.1 中可以看到不同的缺失机制对缺失数据产生的影响. 图 1.3.1 和图 1.1.1(a)，(b) 和 (d) 展示的图一致. 为了方便比较，把这三种不同的数据在图 1.3.1 重新展示. 从图 1.3.1 (c) 中可以看出，不同于协变量缺失下随机缺失机制的结果，如果因变量缺失，在随机缺失机制下，缺失的数据受到协变量的影响.

类似于协变量缺失对似然函数的影响，不难得出响应变量的缺失机制对于似然函数的影响，这里不再赘述.