线性混合效应模型影响分析

费 宇　潘建新 著

科学出版社
www.sciencep.com

线性混合效应模型影响分析

费 宇 潘建新 著

科学出版社
北京

内 容 简 介

本书研究了线性混合效应模型的影响分析问题，将近两年刚刚发展起来的 Q 函数方法全面系统地应用于该模型的统计诊断，对 6 种协方差结构的模型给出了 Cook 型诊断统计量，并提出基于 Q 函数的二阶导数期望的 Cook 型诊断统计量，发展和推广了原有的 Q 函数方法；还讨论了方差结构对统计诊断的影响，指出方差结构的误定可能引起影响点的误判，最后讨论了个体水平和观测值水平影响分析的关系.

本书可供大专院校的学生、教师、科研人员及统计工作者参考.

图书在版编目(CIP)数据

线性混合效应模型影响分析/费 宇，潘建新著 —北京·科学出版社，2005
ISBN 7-03-015490-8

Ⅰ.线… Ⅱ ①费… ②潘… Ⅲ 统计模型 Ⅳ C8

中国版本图书馆 CIP 数据核字(2005)第 046741 号

责任编辑：吕 虹 祖翠娥/责任校对 李奕萱
责任印制 钱玉芬/封面设计 王 浩

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
双青印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2005 年 7 月第 一 版 开本：B5(720×1000)
2005 年 7 月第一次印刷 印张：12 3/4
印数：1—3000 字数：239 000

定价:30.00 元

(如有印装质量问题，我社负责调换〈环伟〉)

前　言

线性混合效应模型是最重要的一种常见回归模型，它通常用于纵向数据和空间数据分析，本书讨论该模型的影响分析问题.

关于线性混合效应模型影响分析问题，现有的文献大都是从传统的似然函数出发来讨论这个问题，这种方法可以处理独立方差结构下线性混合效应模型的影响分析，但对于稍具复杂而常用的非独立方差结构，如一阶自回归结构，从似然函数出发则难以获得相应的影响诊断统计量，如 Cook 统计量，一个重要的原因是有关的 Hessian 阵和 Fisher 信息阵太复杂.

Zhu 等人 2001 年在文献 [62] 中针对缺失数据模型提出了一种基于 Q 函数的广义 Cook 距离，这里的 Q 函数是 EM 算法中对数似然函数的条件期望. 在本书中，采用这种 Q 函数方法来研究线性混合效应模型的影响分析问题，使用的工具主要有两种 Cook 统计量：QD_i 和 QD_i^*. 前者是由 Zhu 等人的方法推导而出的诊断统计量，而后者是作者提出的，它比前者的数学形式更为简洁. 这两种统计量都可以方便地用于线性混合效应模型的影响分析，而且由于方法简单实用，预见它们也可以应用于非线性混合效应模型的诊断分析.

本书第 1 章将简要介绍统计诊断的背景，给出线性混合效应模型的简要介绍和本书的结构框架. 另外，为了便于以后的应用，还给出了一些矩阵运算和矩阵求导公式，以及线性混合效应模型的 Q 函数的具体形式等.

在第 2 章里，基于似然函数框架，对于独立方差结构的线性混合效应模型，导出了四种广义 Cook 距离，并将它们和真实的 Cook 距离通过数据分析加以比较，从中能够看出这四种距离都是真实距离的良好近似.

第 3 章致力于 Q 函数方法在线性混合效应模型诊断中的应用，在六种不同的方差结构假设下，导出了广义 Cook 统计量 QD_i 和 QD_i^*，这里 QD_i 是基于 Q 的二阶导数的统计量，而 QD_i^* 是基于 Q 的二阶导数的期望的统计量. 作为真实 Cook 距离 CD_i 的近似，这两种 Q 统计量方便于实际应用，特别是 QD_i^*，由于可以分解为三个相互正交的部分，分别对应于固定效应、个体间方差分量和个体内方差分量，具有简洁的解析形式和清晰的统计解释. 广义 Cook 统计量 QD_i^* 的这一特点对于研究子集参数的影响问题起到关键作用.

在第 4 章，将用第 3 章提出的 Q 统计量讨论方差结构对于统计诊断的影响. 实例分析表明方差结构的选择确实影响全参数和固定效应的影响评价，所以方差结构的误定可能引起强影响个体的误判，特别是当感兴趣的参数是固定效应的时候，方差结构的误定往往导致强影响个体的误判.

第 5 章讨论独立方差结构的线性混合效应模型个体水平和观测值水平之间的

统计诊断量的关系. 基于 Q 方法, 书中给出了在这两个不同水平下的广义 Cook 统计量之间的关系. 实例分析表明, 包含有强影响观测值的个体比那些不包含强影响观测值的个体更有可能成为强影响个体, 而强影响观测值也较有可能包含于强影响个体中. 全书的结论及待研究的一些课题也将在这一章给出.

最后, 本书中用到的所有数据列于附录里, 以便读者使用.

本书适用对象为统计专业的研究生、教师、科研人员及统计工作者.

本书研究的课题获得云南省“跨世纪学术带头人”基金和云南大学自然科学基金的鼎力支持, 同时在写作过程中得到了云南大学王学仁教授、马骏教授和施本植教授的大力支持, 还得到了东南大学韦博成教授和上海师范大学岳荣先教授许多有益的建议. 此外, 本书的出版得到了云南大学经济学院和会计系出版基金的资助, 得到了科学出版社科学分社吕虹编审的大力支持和帮助, 在此我们一并表示衷心的感谢.

作　者

2004 年 12 月

目　录

第1章 引 论

统计诊断是 20 世纪 70 年代中期发展起来的一门统计科学新分支，它是对从实际问题中建立的统计模型的合理性进行分析，通过一系列诊断统计量来检查数据、模型和推断方法中可能存在的“问题”，从而提出“治疗”方案. 影响分析是统计诊断的一种重要方法，它的目的是探测数据中对既定模型的统计推断影响特别大的点，即所谓“强影响点” (influential point). 点删除法是一种经典的影响分析方法，自从 1977 年 Cook 在文献 [11] 中提出这一方法以来，它已经成功地应用于许多模型的影响分析，如普通线性模型的影响分析 (参阅文献 [14]), 增长曲线模型的影响分析 (参阅文献 [40]). 本书将点删除法思想应用于线性混合效应模型的影响分析，采用广义 Cook 统计量作为影响度量，讨论了似然函数和 Q 函数框架下的影响分析问题.

第 1 章的 §1.1 简要介绍统计诊断的基本概念; §1.2 介绍线性混合效应模型; §1.3 介绍本书结构和主要结论; 最后，一些预备知识，如相关的矩阵运算、矩阵微商和线性混合效应模型的 Q 函数等在 §1.4 中给出，以便于后面章节的应用.

§1.1 基 本 概 念

1.1.1 统计诊断的概念

统计科学中，统计模型在数据分析、统计推断和预测等方面扮演重要角色，我们从实际现象中获得数据后，通常需要建立一个“好模型”来拟合这些数据. 根据统计模型，我们可以获得一些描述实际现象的重要信息，这些重要信息通常通过少量的几个参数来描述. 统计科学的发展表明，许多重要的统计模型如普通回归模型、广义线性模型、增长曲线模型和线性混合效应模型等，已经被成功地用来解决实际问题. 然而，选择模型的时候，难免面临一些问题，如:

- 什么是“好模型”?
- 基于选定的模型作出的统计推断对数据是否敏感?
- 数据中是否存在“强影响点”? 当它们被删除后，统计推断是否会有较大的改变? 如何探测这些所谓的“强影响点” (参阅文献 [40]) ?

在过去 20 多年里已有许多方法被用于解决这些问题，可供参阅的文献包括 Cook 和 Weisberg 1982 年出版的著作 [14], Chatterjee 和 Hadi 1988 出版的著作 [8], 以及 Pan 和 Fang 2002 出版的著作 [40], Cook 于 1977 年和 1986 年发表的论文 [11,12].

一般而言，有两种研究统计模型影响评价的方法，一种是所谓的“点删除”影响分析方法，它是一种常用的方法，可以评价单独的一个点对统计模型的推断(如模型参数的估计)的影响. 这一方法是统计学家 Cook 1977 年在文献 [11] 里提出的方法，这一方法也被称为整体影响分析，即评价一个个体或观测值被删除后对模型产生的影响. “点删除”影响分析方法是最常见的统计诊断方法，它已被广泛应用于各种统计模型 (参阅文献 [1] ~ [3], [8],[14], [15], [40], [58]).

第二种影响评价方法是 1986 年 Cook 在文献 [12] 中提出的所谓局部影响分析方法，这一方法采用几何曲率的模来度量扰动对模型拟合的影响，如影响度量可以是 Cook[12] 提出的似然距离或 Box 和 Tiao[7] 提出的贝叶斯 K-L 距离. 局部影响分析方法已经广泛地应用于统计模型的统计诊断，如应用于线性混合效应模型 (参阅文献 [5], [33]), 广义线性回归模型 (参阅文献 [56]), 非线性回归模型 (参阅文献 [55]) 及增长曲线模型 (参阅文献 [40]).

一般来说，局部影响分析比整体影响分析所揭示的信息要多，因为整体影响分析可能会出现所谓“掩盖”效应，即由于某种原因，诊断统计量没能探测出数据中所有的异常点或强影响点，这一现象称为“掩盖”效应. 整体影响分析还可能会出现所谓“淹没”效应，即由于某种原因，诊断统计量将某些不是异常点或强影响点的值诊断为异常点或强影响点，这一现象称为“淹没”效应. “掩盖”效应和“淹没”效应会使得探测到的强影响点有时过多或过少 (参阅文献 [52]).

尽管如此，作为一种最常用的诊断方法，“点删除”影响分析方法在影响分析中扮演着非常重要的角色，它已成功地应用于普通线性回归模型 (参阅文献 [8], [14]) 和增长曲线模型 (参阅文献 [40]) 等许多常用统计模型. 对于线性混合效应模型，从似然函数出发，1992 年 Christensen 等人在文献 [10] 中用“点删除”影响分析方法讨论了强影响观测值的探测问题，Banerjee 和 Frees 于 1997 年在文献 [3] 中用同样的方法讨论了强影响个体的探测问题，他们的讨论都基于这样一个假定：协方差阵是已知的，如果协方差阵未知，就用基于全模型的协方差阵的估计代替协方差阵，然后再作统计诊断. 实际中，颇具争议的问题是：协方差阵已知这样的假定是否合理？对于协方差阵未知的情况，用全模型的协方差阵的估计代替协方差阵作统计诊断分析又是否合理？

1.1.2　强影响观测值和强影响个体

我们知道，统计模型是对复杂的实际过程的近似描述，数据中的几个甚至仅一个不寻常的观测值可能对由统计模型获得的统计结论会有强烈的影响. 换言之，当这个观测值从数据中被删除之后，统计推论会有重大改变，这个观测值就是我们所说的强影响观测值 (参阅文献 [13],[40]).

强影响观测值的定义可以推广到纵向数据分析中，而纵向数据分析通常可以在线性混合效应模型的框架下讨论. 在纵向数据分析中，一个个体或实验单位往往

包含几个观测值，统计推断如回归参数的估计可能受来自于某个个体的几个或一个观测值的强烈影响，也就是说，当这个个体或观测值从数据中删除之后，统计推断会有实质性改变，这个个体或观测值就称为强影响个体或强影响观测值. 所以，纵向数据分析中影响分析可以分为两个水平：个体水平的影响分析和观测值水平的影响分析.

一个有趣的问题是：这两个水平的影响分析的关系怎样？本书第 5 章将讨论这个问题.

1.1.3 Cook 距离

记 θ 为回归模型的参数向量，全模型 (完全数据模型) 的参数向量 θ 的最大似然估计记为 $\hat{\theta}$, 数据删除模型 (假定第 i 个点被删除) 的 θ 的最大似然估计记为 $\hat{\theta}_{[i]}$, 那么 Cook 距离 CD_i 由下面的式子给出：

$$CD_i(M) = (\hat{\theta}_{[i]} - \hat{\theta})' M (\hat{\theta}_{[i]} - \hat{\theta}) \tag{1.1.1}$$

其中 M 是权矩阵， Cook 和 Weisberg 在 1982 出版的著作 [14] 里提出了 M 的几种选择. 一般而言，权矩阵 M 可选为“外尺度”或“内尺度” (参阅文献 [14]). 如果选“外尺度”，权矩阵 M 通常取为 $\{-\ddot{l}_{[i]}(\hat{\theta})\}$, 这里 $l_{[i]}$ 是数据删除模型 (删除第 i 号数据点后的模型) 的似然函数， $\ddot{l}_{[i]}$ 上面的点表示对 θ 求二阶导数，因此 $\{-\ddot{l}_{[i]}(\hat{\theta})\} = \left\{ -\dfrac{\partial^2 l_{[i]}(\theta)}{\partial\theta\partial\theta'} \right\}_{\theta=\hat{\theta}}$ 表示 $-l_{[i]}$ 的二阶导数在 $\theta = \hat{\theta}$ 处的取值，相应的渐进置信域为

$$\{\theta : (\theta - \hat{\theta})' \{-\ddot{l}_{[i]}(\hat{\theta})\} (\theta - \hat{\theta}) \leqslant \chi^2(\alpha; r)\} \tag{1.1.2}$$

其中 $\chi^2(\alpha; r)$ 是自由度为 r 的 χ^2 分布的上 α 分位点 (参阅文献 [62]).

如果 Cook 统计量 $CD_i(M)$ 的值较大，则认为第 i 号个体或观测值是强影响点，这里所谓“较大”是与 χ^2 分布的临界值比较而言的 (参阅文献 [14]).

§1.2 线性混合效应模型

线性混合效应模型 (参阅文献 [27]) 是最重要的回归模型之一，通常用于纵向数据分析，因为纵向数据分析的许多问题都可以纳入线性混合效应模型的框架下讨论 (参阅文献 [17]).

线性混合效应模型的一般定义为

$$\begin{cases} Y_i = X_i\beta + Z_i u_i + \epsilon_i \\ u_i \sim N(0, G) \\ \epsilon_i \sim N(0, R_i) \end{cases} \tag{1.2.1}$$

其中对于第 i 个个体，Y_i 是 $n_i\times 1$ 维的响应向量，X_i 是 $n_i\times p$ 的设计阵，β 是 $p\times 1$ 维的未知的固定效应参数，Z_i 是 $n_i\times q$ 的对应于 $q\times 1$ 维个体间随机效应 u_i 的设计阵，ϵ_i 是 $n_i\times 1$ 维随机误差向量，$i=1,\cdots,m$ 对应 m 个个体. 向量 u_i 与 ϵ_i 相互独立，矩阵 G 是个体间协方差阵，R_i 是个体内协方差阵. 我们用 Σ_i 记 Y_i 的协方差阵，即 $\Sigma_i=\mathrm{var}(Y_i)=Z_iGZ_i'+R_i$.

引入矩阵记号，线性混合效应模型 (1.2.1) 可以写为

$$\begin{cases} Y=X\beta+Zu+\epsilon \\ u\sim N(0,\mathcal{G}) \\ \epsilon\sim N(0,\mathcal{R}) \end{cases} \tag{1.2.2}$$

其中 $Y=(Y_1',\cdots,Y_m')'$ 是 $n\times 1\left(n=\sum_{i=1}^m n_i\right)$ 维响应向量，由 $Y_1,\cdots,Y_m$ 排列而成，$n\times p$ 的设计阵 X 由 $X_1,\cdots,X_m$ 排列而成，即 $X=(X_1',\cdots,X_m')'$，矩阵 $Z=\mathrm{diag}(Z_1,\cdots,Z_m)$ 是 $n\times mq$ 的对应随机效应 $u=(u_1',\cdots,u_m')'$ 的设计阵，$\epsilon=(\epsilon_1',\cdots,\epsilon_m')'$ 是 $n\times 1$ 维随机误差向量，这里 $\mathcal{G}=\mathrm{diag}(G,\cdots,G)$ 和 $\mathcal{R}=\mathrm{diag}(R_1,\cdots,R_m)$ 是块对角矩阵，$G=G(\alpha)$ 中的 α 是 $r\times 1$ 的 G 的参数向量，而 $R_i=R_i(\gamma)$ 中的 γ 是 $s\times 1$ 的 R_i 的参数向量. 显然，Y 的协方差阵 $\Sigma=Z\mathcal{G}Z'+\mathcal{R}$，我们用 $\theta=(\beta',\alpha',\gamma')'$ 记全模型 (1.2.2) 的全参数.

如果第 i 个个体被删除，我们得到模型 (1.2.2) 对应的数据删除模型：

$$\begin{cases} Y_{[i]}=X_{[i]}\beta+Z_{[i]}u_{[i]}+\epsilon_{[i]} \\ u_{[i]}\sim N(0,\mathcal{G}_{[i]}) \\ \epsilon_{[i]}\sim N(0,\mathcal{R}_{[i]}) \end{cases} \tag{1.2.3}$$

其中 $Y_{[i]}=(Y_1',\cdots,Y_{i-1}',Y_{i+1}',\cdots,Y_m')'$ 是 $(n-n_i)\times 1$ 响应向量 (从 Y 中删除 Y_i 而得)，$X_{[i]}=(X_1',\cdots,X_{i-1}',X_{i+1}',\cdots,X_m')'$ 是 $(n-n_i)\times p$ 设计阵 (从 X 中删除 X_i 而得)，$Z_{[i]}=\mathrm{diag}(Z_1,\cdots,Z_{i-1},Z_{i+1},\cdots,Z_m)$ 是 $(n-n_i)\times(m-1)q$ 块对角矩阵 (从 Z 中删除第 i 对角块 Z_i 而得)，$\mathcal{G}_{[i]}=\mathrm{diag}(G,\cdots,G)$ 是 $(m-1)q\times(m-1)q$ 块对角矩阵 (从 $\mathcal{G}$ 中删除第 i 对角块 G 而得)，$\mathcal{R}_{[i]}=\mathrm{diag}(R_1,\cdots,R_{i-1},R_{i+1},\cdots,R_m)$ 是 $(n-n_i)\times(n-n_i)$ 块对角矩阵 (从 $\mathcal{R}$ 中删除第 i 对角块 R_i 而得). 相应地，$Y_{[i]}$ 的协方差阵 $\Sigma_{[i]}=Z_{[i]}\mathcal{G}_{[i]}Z_{[i]}'+\mathcal{R}_{[i]}$.

线性混合效应模型常常用于纵向数据分析，但该模型统计诊断问题的研究却不多，Christensen 等 1992 年在文献 [10] 中用“点删除”法研究了该模型的统计诊断问题，他们在协方差阵 V 已知的条件下讨论了线性混合效应模型的 Cook 距离，当协方差阵 V 未知时，建议用 V 的估计代替 V 进行诊断. 这是一种方便的方法，但如果简单地忽略协方差的影响可能引起强影响观测值的误定 (我们将在第 4 章详细讨论这个问题). 另外，Christensen 等 1992 年的文章 [10] 只研究了固定效应参数的诊断问题，而没有讨论协方差参数和全参数 (固定效应和协方差参数) 的诊断问题，但协方差参数和全参数的诊断问题有重要的理论和实际意义.

与 Christensen 等人 [10] 的研究类似， Banerjee 和 Frees 于 1997 年在文献 [3] 中研究了个体的影响分析问题，他们的研究是从个体角度出发. 假定模型方差结构是在正确的前提下进行的，然而，这一假定在很多情况下是不成立的 (因为实际中方差结构往往是未知的)，这意味着在统计诊断中我们不能忽视协方差结构对诊断的影响.

为了解决上面提到的这些问题，受 Zhu 等人 [62] 提出的 Q 函数方法的启示，本书基于 Q 函数提出两种广义 Cook 距离 QD_i 和 QD_i^*，前者是由 Zhu 等人的方法推导而出的诊断统计量，而后者是我们提出的，它比前者的数学形式更为简洁，而且全参数的广义 Cook 距离 QD_i^* 可以分解为相互独立的三个广义 Cook 距离，分别对应固定效应、个体间协方差分量和个体内协方差分量，这有利于我们讨论子集参数的诊断问题，详细的讨论将放在第 2 章和第 3 章里. 我们获得的两种统计量都可以方便地用于线性混合效应模型的影响分析，而且由于方法简单实用，我们预见它们也可以应用于非线性混合效应模型的诊断分析.

§1.3 本书的结构

正如前面指出，本书的主要目的是介绍“点删除”诊断方法在线性混合效应模型诊断中的应用，我们的研究是在似然函数框架和 Q 函数框架下进行的. 由于非独立方差结构下基于似然函数的诊断统计量没有解析表达式，所以我们把讨论重点放在 Q 函数方法，诊断统计量 Cook 距离 (参阅文献 [14]) 将在全书的影响诊断中扮演十分重要的角色. 图 1.1 给出了本书的主要结构，详细的概要在以下子节中给出.

1.3.1 似然函数框架下的统计诊断

似然函数框架下的统计诊断这部分内容将在第 2 章里介绍，我们采用“点删除”影响分析法研究独立方差结构下线性混合效应模型的统计诊断，我们给出以下四种基于似然函数的 Cook 距离：

(1) C_i 基于 $\ddot{l}_{[i]}$，即数据删除模型 (1.2.3) 的似然函数的二阶导数；

(2) C_i^* 基于 $\ddot{l}$，即全模型 (1.2.2) 的似然函数的二阶导数；

(3) D_i 基于 $E\ddot{l}_{[i]}$，即 $\ddot{l}_{[i]}$ 的期望；

(4) D_i^* 基于 $E\ddot{l}$，即 $\ddot{l}$ 的期望.

最后三种距离 C_i^*,D_i 和 D_i^* 是第一种距离 C_i 的近似，与第一种距离相比，后三种距离的主要优势在于降低了计算量，特别是第四种基于 $\ddot{l}$ 的期望 $E\ddot{l}$ 的距离 D_i^*，它具有最简洁的数学表达式，在适当条件下，我们将看到在实际应用中 D_i^* 是一种合理的 Cook 距离.

为了说明以上四种 Cook 距离的关系，我们分析了两批实际数据 (称作牙齿数据和气雾剂数据).

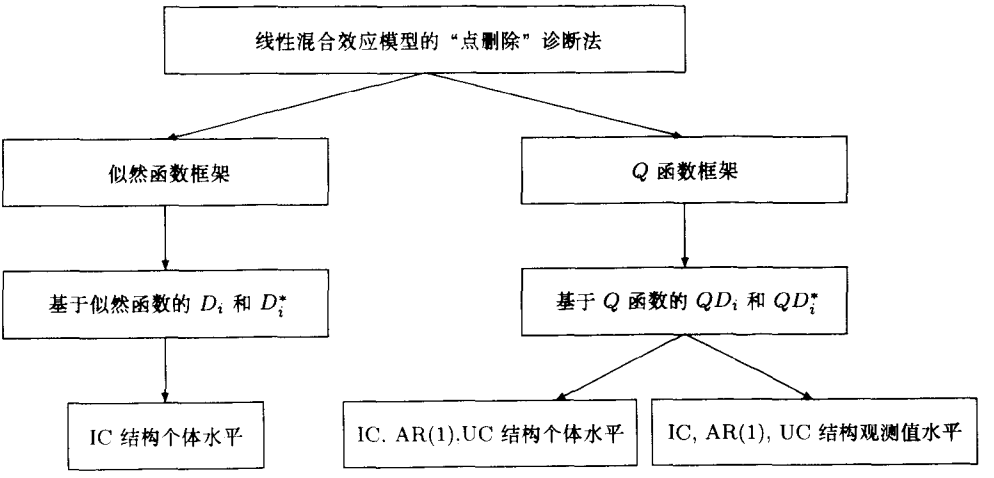

图 1.1　本书的结构

注：IC 结构即独立协方差结构，AR(1) 结构即一阶自回归结构，UC 结构即一般协方差结构

1.3.2　Q 函数框架下的统计诊断

在第 3 章，我们将研究 Q 函数框架下线性混合效应模型的“点删除”统计诊断，这里 Q 函数是对数联合似然函数的条件期望，它在众所周知的 EM 算法 (参阅文献 [16]) 中扮演重要的角色，在 §3.1 我们将给出 EM 算法的简要介绍.

基于 Q 函数，Zhu 等人 2001 年在文献 [62] 中给出了缺失数据模型的广义 Cook 距离，在他们的思想启发下，针对线性混合效应模型，我们基于 Q 函数导出了两种广义 Cook 距离用于评价删除一个个体对参数估计的影响. 我们给出的第一种广义 Cook 距离 QD_i 是基于 $\ddot{Q}$, 即 Q 函数的二阶导数；而第二种广义 Cook 距离 QD_i^* 是基于 $E\ddot{Q}$, 即 $\ddot{Q}$ 的期望. 作为前者的近似，后者在数学表达式上更简洁，而且统计意义也更清楚.

我们将在以下六种常见方差结构下讨论线性混合效应模型的统计诊断问题：

(1) 独立结构 (IC 结构)：G 和 R_i 都是独立结构；

(2) 自回归结构 I(AR(1)I 结构)：G 是独立结构而 R_i 是 AR(1) 结构；

(3) 自回归结构 II(AR(1)II 结构)：G 是 AR(1) 结构而 R_i 是独立结构；

(4) 自回归结构 III(AR(1)III 结构)：G 和 R_i 都是 AR(1) 结构；

(5) 一般结构 I(UC I 结构)：G 是一般结构而 R_i 是独立结构；

(6) 一般结构 II(UC II 结构)：G 是一般结构而 R_i 是 AR(1) 结构.

我们选择以上六种方差结构作讨论的主要原因是：对于纵向数据分析而言，以上六种方差结构是最常见和最重要的方差结构；很多纵向数据分析的实际问题都可以归入以上六种方差结构中的某一种结构下来讨论，如对于牙齿测量数据 (文献 [48]) 来说，采用 AR(1)I 结构就能很好地反映该数据的方差结构情况；虽然独立结构 (IC 结构) 在实际中是罕见的，但这一结构在理论上是最简单的结构，而且在独立方差结构下，我们可以找到似然函数观点下的诊断统计量的解析表达式，这便于比较 Q 函数观点下的诊断统计量与似然函数观点下的诊断统计量的诊断结果.

我们在以上六种方差结构下导出了两种广义 Cook 距离 QD_i 和 QD_i^*，并用这两种广义 Cook 距离分析了六个实例，结果显示 QD_i^* 在探测强影响个体时和 QD_i 一样有效，而统计度量 D_i,D_i^*,QD_i 及 QD_i^* 之间的比较则说明基于 Q 函数的度量与基于似然函数的度量的效果同样好. 第 2 章和第 3 章的讨论是在个体水平下进行的，而第 5 章将给出个体水平和观测值水平影响分析之间的关系.

1.3.3 方差结构对统计诊断的影响

正如 §1.2 指出，方差结构在影响分析中的作用很关键，第 4 章将讨论方差结构在线性混合效应模型的统计诊断中的作用. 换言之，方差结构是如何影响线性混合效应模型的统计诊断？如果由于某种原因引起方差结构的误定，那么对全参数的估计和固定效应的估计有何影响？

葡萄糖数据 (参阅文献 [60]) 的分析说明方差结构的选择确实影响全参数的 Cook 距离，也影响固定效应参数的 Cook 距离. 对后者的影响看起来比对前者的影响要大，这提示我们注意：当方差结构误定时， Christensen 等 [10] 和 Banerjee 与 Frees[3] 忽略方差结构正确判定的做法是值得商榷的. 我们的分析显示研究回归参数的诊断，特别是固定效应参数 β 的诊断时，考虑方差结构的影响是很有意义的.

1.3.4 两水平的影响分析

实际中，强影响观测值的探测和强影响个体的探测同样重要，在独立方差结构，即 IC 结构下，第 5 章给出了观测值水平线性混合效应模型的影响分析度量，还研究了个体水平和观测值水平两水平影响分析的关系.

两水平下的两种 Cook 距离 (这里以 QD 为例说明) 的关系如下：

(1) 当某些 (至少一个)QD_{ij}(第 i 个个体的第 j 个观测值的 Cook 距离) 较大时， QD_i(第 i 个个体的 Cook 距离) 可能较大. 换言之，含有至少一个强影响点的个体比不含任何强影响点的个体更有可能是强影响个体.

(2) 当 QD_i 较大时，一个 (或多个)QD_{ij} 可能较大. 换言之，强影响个体 (如第 i 个个体) 中至少会有一个或多个强影响观测值.

第 5 章的理论研究可以推广到其他常见方差结构，但对较为复杂的方差结构，技术处理上可能会较困难.

§1.4　预 备 知 识

为了以后章节的使用方便，一些预备知识，如矩阵的运算、矩阵微商及线性混合效应模型 (1.2.2) 的 Q 函数将在这一子节给出，许多矩阵的运算和矩阵微商法则在本书中将经常用到，我们在下面仅给出最基本法则，要了解更多的有关内容，可以参阅文献 [36] 及 [61].

引理 1.1　设 A 和 B 是 $p\times p$ 和 $q\times q$ 非奇异矩阵，C 是 $p\times q$ 矩阵，而 D 是 $q\times p$ 矩阵，那么

$$(A+CBD)^{-1}=A^{-1}-A^{-1}CB(B+BDA^{-1}CB)BDA^{-1} \tag{1.4.1}$$

(参见文献 [40].)

引理 1.2　设 A 是 $p\times p$ 非奇异矩阵有如下分解：

$$A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

其中 A_{11},A_{12},A_{21} 和 A_{22} 分别是 $k\times k, k\times(p-k), (p-k)\times k$ 和 $(p-k)\times(p-k)$ 子阵. 假定 A_{11} 和 A_{22} 非奇异，记 $A_{11.2}=A_{11}-A_{12}A_{22}^{-1}A_{21}$ 和 $A_{22.1}=A_{22}-A_{21}A_{11}^{-1}A_{12}$，则

$$\begin{aligned} A^{-1} &= \begin{pmatrix} A_{11.2}^{-1} & -A_{11}^{-1}A_{12}A_{22.1}^{-1} \\ -A_{22}^{-1}A_{21}A_{11.2}^{-1} & A_{22.1}^{-1} \end{pmatrix} \\ &= \begin{pmatrix} A_{11}^{-1}+A_{11}^{-1}A_{12}A_{22.1}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22.1}^{-1} \\ -A_{22}^{-1}A_{21}A_{11.2}^{-1} & A_{22.1}^{-1} \end{pmatrix} \end{aligned} \tag{1.4.2}$$

(参见文献 [40].)

引理 1.3　设 A 是 $q\times q$ 非奇异矩阵有形式

$$A=(\rho^{|j-k|})=\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{q-1} \\ \rho & 1 & \rho & \cdots & \rho^{q-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho^{q-1} & \rho^{q-2} & \rho^{q-3} & \cdots & 1 \end{bmatrix} \tag{1.4.3}$$

那么我们有

$$|A|=(1-\rho^2)^{q-1} \tag{1.4.4}$$

(参阅文献 [23].)

引理 1.4 设 $g(x)$ 是 $p\times 1$ 向量 x 的函数，则对于常数对称矩阵 A, 有

$$\frac{\partial\{y-g(x)\}'A\{y-g(x)\}}{\partial x}=-2D(x)'A\{y-g(x)\} \tag{1.4.5}$$

其中 $D(x)=\partial g(x)/\partial x'$.

(参阅文献 [57].)

引理 1.5 设 A 是 x 的矩阵函数，记 $|A|$ 为矩阵 A 的行列式，则

$$\frac{\partial\log|A|}{\partial x}=\mathrm{tr}\left(A^{-1}\frac{\partial A}{\partial x}\right),\qquad \frac{\partial A^{-1}}{\partial x}=-A^{-1}\frac{\partial A}{\partial x}A^{-1} \tag{1.4.6}$$

其中 $\mathrm{tr}(\cdot)$ 表示矩阵的迹.

(参阅文献 [53].)

引理 1.6 对于线性混合效应模型 (1.2.2), 响应向量和随机效应的联合对数似然函数有以下形式:

$$\begin{aligned}\log f(Y,u)=&-\frac{1}{2}\log|\mathcal{R}|-\frac{1}{2}(Y-X\beta-Zu)'\mathcal{R}^{-1}(Y-X\beta-Zu)\\&-\frac{m}{2}\log|G|-\frac{1}{2}u'\mathcal{G}u\end{aligned} \tag{1.4.7}$$

而 Q 函数，即给定 Y 和 $\tilde\theta=(\tilde\beta',\tilde\alpha',\tilde\gamma')'$ 的条件下，对数似然函数的条件期望，可以表示为

$$\begin{aligned}Q(\theta|\tilde\theta)\equiv&E\{\log f(Y,u)|Y,\tilde\theta\}\\=&-\frac{1}{2}\log|\mathcal{R}|-\frac{1}{2}\tilde e'\mathcal{R}^{-1}\tilde e-\frac{1}{2}\mathrm{tr}\{\mathcal{R}^{-1}Z\tilde\Phi^{-1}Z'\}\\&-\frac{m}{2}\log|G|-\frac{1}{2}\mathrm{tr}\{\mathcal{G}^{-1}(\tilde\Phi^{-1}+\tilde u\tilde u')\}\end{aligned} \tag{1.4.8}$$

其中 $\tilde\theta$ 是 EM 算法前一次迭代所得的参数的估计值，而

$$\begin{aligned}\tilde e&=Y-X\beta-Z\tilde u,&\tilde u&=\tilde{\mathcal{G}}Z'\tilde\Sigma^{-1}(Y-X\tilde\beta)\\\tilde\Phi&=\tilde{\mathcal{G}}^{-1}+Z'\tilde{\mathcal{R}}^{-1}Z,&\tilde\Sigma&=Z\tilde{\mathcal{G}}Z'+\tilde{\mathcal{R}}\end{aligned} \tag{1.4.9}$$

和 $\tilde\beta$, $\tilde{\mathcal{G}}=\mathcal{G}(\tilde\alpha)$, $\tilde{\mathcal{R}}=\mathcal{R}(\tilde\gamma)$ 是 EM 算法前一次迭代所得的参数的估计.

(参阅文献 [45].)

引理 1.7 对于具有一般协方差结构线性混合效应模型 (1.2.2), 我们有

$$E(\tilde e)=0,\qquad E(\tilde u\tilde u')=\tilde{\mathcal{G}}Z'\tilde\Sigma^{-1}Z\tilde{\mathcal{G}},\qquad E(\tilde e\tilde e')=Z\tilde\Phi^{-1}Z'+\mathcal{R} \tag{1.4.10}$$

其中 $\tilde e,\tilde u,\tilde\Phi,\tilde\Sigma$ 由 (1.4.9) 给出.

证明　因为 $\tilde{e} = Y - X\beta - Z\tilde{u} = Z(u - \tilde{u}) + \epsilon$, 所以 $E(\tilde{e}) = ZE(u - \tilde{u}) + E(\epsilon)$, 注意到

$$u|(Y, \tilde{\theta}) \sim N(\tilde{u}, \tilde{\Phi}^{-1}), \qquad \epsilon \sim N(0, \mathcal{R}) \tag{1.4.11}$$

(参阅文献 [45]) 立即可得 $E(\tilde{e}) = 0$.

由 (1.4.9) 并注意到 (1.4.11), 可得

$$E(\tilde{u}\tilde{u}') = \tilde{\mathcal{G}}Z'\tilde{\Sigma}^{-1}[E(Y - X\tilde{\beta})(Y - X\tilde{\beta})']\tilde{\Sigma}^{-1}Z\tilde{\mathcal{G}} = \tilde{\mathcal{G}}Z'\tilde{\Sigma}^{-1}Z\tilde{\mathcal{G}}$$

此外, 由 (1.4.9) 和 (1.4.11), 我们有

$$\begin{aligned} E(\tilde{e}\tilde{e}') &= E[Z(u - \tilde{u}) + \epsilon][Z(u - \tilde{u}) + \epsilon]' \\ &= Z[E(u - \tilde{u})(u - \tilde{u})']Z' + E(\epsilon\epsilon') \\ &= Z\tilde{\Phi}^{-1}Z' + \mathcal{R} \end{aligned}$$

证明完毕.