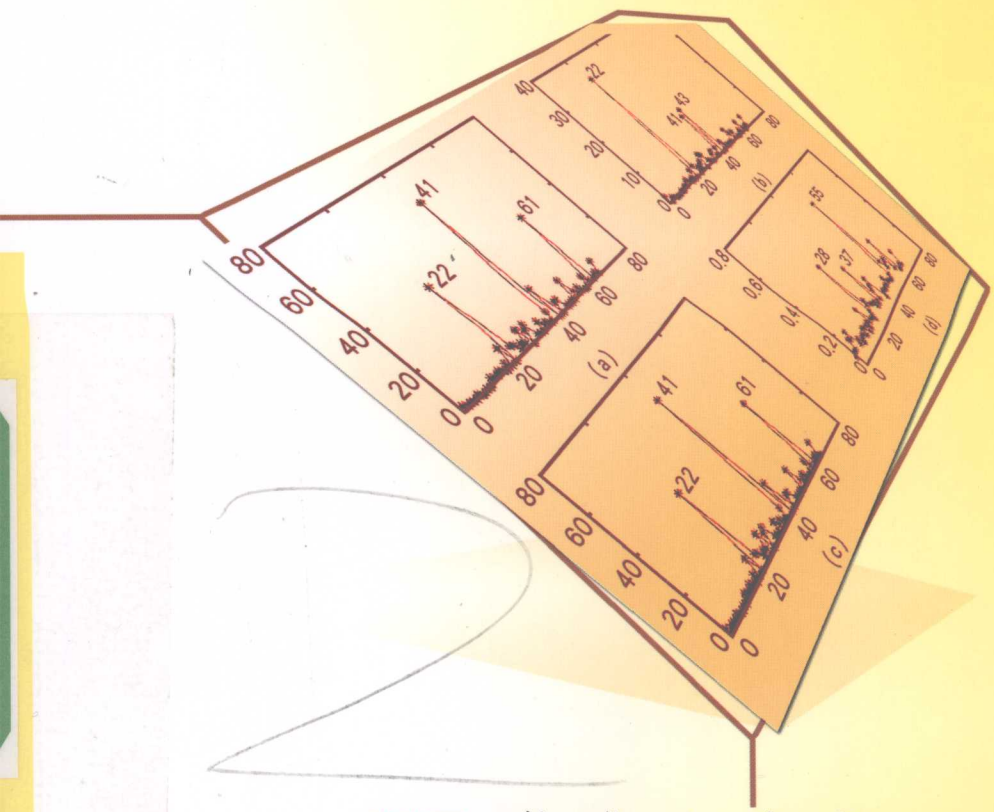
多水平模型及其统计诊断

石磊◎著



科学出版社
www.sciencep.com

多水平模型及其统计诊断

石 磊 著

科 学 出 版 社
北 京

内 容 简 介

本书系统介绍了多水平模型中的统计模型诊断问题. 主要内容是：介绍了异常值和影响点的定义和历史发展, 数据删除法和局部影响分析的一些主要方法和结果, 以及多水平模型的定义、参数估计及统计推断；总结了已知协方差结构的广义线性模型的统计诊断结果, 并采用删除多个数据点的方法研究了多水平模型中固定效应及随机效应参数的诊断统计量；研究了多水平模型下基于均值漂移模型的异常点探测问题、高水平单元的局部影响分析和单个观察值的局部影响分析. 本书还介绍了目前关于多水平模型的统计软件, 同时给出了利用Matlab语言编写的计算程序.

本书可作为数理统计专业本科生、硕士生和博士生的教材或参考书, 也可作为数学、生物、医学、工程、经济领域的教师及相关科技工作者的参考书.

图书在版编目(CIP)数据

多水平模型及其统计诊断/石磊著 —北京：科学出版社, 2008

ISBN 978-7-03-021278-8

I. 多… II. 石… III. 统计模型-研究 IV. C8

中国版本图书馆 CIP 数据核字(2008) 第 030792 号

责任编辑：王丽平 房 阳 / 责任校对：陈丽珠
责任印制：赵德静 / 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

源海印刷有限责任公司 印刷

科学出版社发行 各地新华书店经销
*
2008 年 5 月第 一 版 开本：B5(720 × 1000)
2008 年 5 月第一次印刷 印张：11 3/4
印数：1—3 000 字数：217 000

定价：40.00 元

(如有印装质量问题, 我社负责调换〈环伟〉)

前　言

在评价模型假设的恰当性、识别可能严重影响统计推断或需要特别关注的数据结构等方面, 识别异常值和影响点是一个非常重要的研究工作. 异常值问题一直受到较为广泛的关注, 最早可以追溯到基于数据得出结论的探索性研究 (Bernoulli, 1777), 最早尝试使用统计方法解决异常值问题的研究大约在 1850 年. 现在关于异常值的理论成果已经非常丰富, 主要的结果综述见 Barnett 和 Lewis, 1994; Bechman 和 Cook, 1983; Hawkins, 1980. Bechman 和 Cook 在他们 1983 年一篇综述报告中指出异常值可以被视为不协调点或者杂质点. 所谓不协调点是指那些与数据集的主体明显不协调, 使研究者感到惊讶的数据点; 杂质点是指与数据集的主体不是来自同一分布的数据值. 对于线性回归模型, 通常使用均值漂移模型研究异常值 (详见 Srikantan, 1961; Ferguson, 1961; Bechman 和 Cook, 1983; Hawkins, 1980). 在识别单个异常值时, Srikantan(1961), Cook 和 Weisberg(1982) 提出采用最大的学生化残差的平方做为检验统计量.

影响点是指那些对参数估计和统计推断有潜在影响的数据点 (Cook, 1977; Cook 和 Weisberg, 1982), 影响点的识别就称为影响分析. 在影响分析中, 应明确哪些统计量是我们感兴趣的 (或称之为影响的目标), 比如参数估计、假设检验、置信区间的估计和预测都是一些在影响分析中经常关注的方面. 影响分析中经常使用的方法是数据删除法 (case deletion), 这一方法的核心在于比较删除数据点前后我们感兴趣的统计量的差异. 使用删除数据点的方法识别影响点已有了大量的文献, 如 Belsley et al., 1980; Cook 和 Weisberg, 1982; Chatterjee 和 Hadi 1988; Cook 和 Weisberg, 1999; Atkinson 和 Riani, 2000 总结了主要的成果. 对线性回归模型, 如果回归系数是我们感兴趣的参数, 则 Cook 距离是这种情况下研究影响点的著名诊断统计量, 该统计量具有显式表达式并且可以被分解为杠杆值和学生化残差的函数, 许多统计软件都可以计算这一统计量并输出结果. 然而对一些较复杂的模型, 如非线性回归模型 (nonlinear model)、广义线性回归模型等 (generalized linear model, 注意它不同于 general linear model), 其参数估计只能通过迭代解出, 因此在这些模型下我们不可能得到度量数据删除影响的精确解析式 (或递推公式). 对于这些模型, 通常采用一种称为一步近似的方法获得近似的影响测度 (Cook 和 Weisberg, 1982; Pregibon, 1981). 局部影响分析是影响分析中一种较新的方法, 该方法由 Cook(1986) 首次提出, 研究同时干扰模型的某些部分而不是删除个别数据点时数据点的联合影响. 这一方法采用似然距离影响图的法曲率来度量影响, 人们最关心的诊断统计量是使得法曲率绝对值达到最大的方向. 局部影响分析方法允许我们对模型的不同部分进行

干扰并在这些扰动下展开影响分析评价, 其中较具有代表性的应用见 Beckman et al., 1987; Lawrance, 1988; Thomas 和 Cook, 1990; St. Laurent 和 Cook, 1993; Tsai 和 Wu, 1992; Shi, 1997; Wu 和 Luo, 1993; Poon 和 Poon, 1999; Zhu 和 Lee, 2001. Shi (1997) 在一般的扰动模式下定义了广义影响函数和广义 Cook 统计量, 该方法的一大优点是便于研究非极大似然类型统计量的局部影响. Shi (1997) 同时也证明该方法在似然结构下与 Cook 的方法等价.

近年来人们已经开始关注具有复杂协方差结构的线性模型. Christensen et al. (1992) 研究了混合模型下的数据删除诊断. 他们在假定协方差阵已知的情况下对回归参数估计进行了影响评价, 并且利用一步近似讨论了方差分量的影响诊断统计量. Martin (1992) 在具有已知协方差结构的广义线性模型下研究了删除单个或者多个数据点对回归参数估计产生的影响. Haslett 和 Hayes (1998) 在一般协方差结构的线性模型中引入了两种残差: 边缘残差和条件残差. Haslett(1999) 给出了数据删除诊断统计量的一种更简洁形式. Hodges (1998) 研究了分层模型下数据点的影响评价, 通过将分层随机效应模型写为一个线性模型的形式给出了一个近似的数据删除公式. Haslett 和 Dillane (2004) 导出线性混合模型下方差分量估计的数据删除诊断统计量.

在运用局部影响分析方法方面, De Gruttola et al. (1987) 基于不同个体单元之间方差相等的假设得到了三步广义最小二乘估计的一些影响测量和杠杆值. Beckman et al. (1987) 通过干扰单个观测值的常量方差得出了一种在线性混合 ANOVA 模型中的诊断方法. Schall 和 Dunne (1991) 提出四种干扰模式来研究回归 ARMA 模型的局部影响分析. Lesaffre 和 Verbeke (1998) 提出一些探测线性混和模型中影响点的诊断统计量, 这些统计量来自单个数据点的单项加权扰动得到的 Cook 法曲率的分解简化. Ouwens, Tan 和 Berger (2001) 研究了广义混合线性模型的局部影响分析, 并比较了基于个体和基于观测值的局部影响分析的区别.

对于具有未知协方差结构的模型, 采用迭代法同时估计回归系数和协方差矩阵中的未知参数是非常必要的. 然而上面提到的所有文章里, 研究者都是将回归系数估计的影响分析和协方差矩阵的未知参数估计的影响分析分开来讨论. 对于这种处理方式, 研究者往往忽略了一点, 他们使用的是数据删除法, 但为了处理方便, 对协方差矩阵中的未知参数是在整个数据集之上进行估计的. 正如 Hodges (1998, p507), Atkinson (1998, p521), 及 Haslett 和 Dillane (2004, p142) 指出的, 这种做法的不足之处在于删除数据点不仅会影响回归系数估计值也会影响到协方差矩阵未知参数的估计值. 然而到目前为止, 还没有对这方面的研究成果.

研究线性模型影响诊断的大量文献中, 关于多水平模型 (multilevel model) 的相关文献还比较少. Langford 和 Lewis (1998) 讨论了该模型下的异常点探测问题并建议使用残差、杠杆值、水平分析法来识别异常点. Shi 和 Ojeda (2004) 对协方差矩阵进行扰动, 研究了在这种扰动模式下多水平模型的局部影响结果. 模型其他

方面的影响分析还没有涉及, 例如数据删除诊断、基于均值漂移模型的异常点探测问题以及基于观察值的局部影响分析等.

本书是作者近年来研究多水平线性模型及其统计诊断的主要结果, 其中部分结果已发表或即将发表. 对多水平模型的统计诊断, 本书的结果同时考虑了协方差矩阵中未知参数的估计对回归系数估计的影响 (目前的所有文献中, 在研究固定效应参数的影响度量时, 都假设协方差矩阵中的未知参数是固定的, 并且用完全数据得到的估计代替), 使获得的诊断统计量更为精确和有效. 这在现有的统计模型诊断中是一个新的思想和方法.

为了使读者对多水平模型有比较好的了解, 本书的第 1 章介绍了异常值和影响点的定义和历史发展以及数据删除法和局部影响分析的一些主要方法和结果. 同时为了方便读者阅读, 作者给出了本书中可能涉及的一些预备知识, 包括矩阵代数、矩阵微商、分布、估计及检验的统计理论.

第 2 章介绍多水平模型的定义, 参数估计方法以及一些相关概念, 包括迭代广义最小二乘 (IGLS) 估计方法、限制迭代广义最小二乘 (RIGLS) 估计方法、参数的置信区间及统计检验、不同水平下的预测残差等. 通过两个例子来说明模型的结构, 并通过详细的分析介绍了多水平模型的建模过程. 这一内容对于对多水平模型感兴趣的读者是有帮助的.

第 3 章总结了目前已知协方差结构的广义线性模型 (general linear model, GLM) 已有的统计诊断结果以及作者研究的一些新结果. 多水平模型的参数可分为固定效应参数和随机效应参数. 如果我们预先假定随机效应参数是已知的, 那么多水平模型就可以看作是一个协方差结构已知的广义线性模型, 迭代广义最小二乘 (IGLS) 估计相应的就变为固定效应参数的广义最小二乘 (GLS) 估计. 因此, 对于已知协方差结构的广义线性模型的影响分析的回顾和研究是很有帮助的. 协方差结构已知的广义线性模型的数据删除诊断可参见 Christensen, Pearson 和 Johnson, 1992; Martin, 1992; Haslett, 1999. 其中杠杆值的定义是本书一个新的结果, 我们发现在标准线性回归模型 (standard linear model) 中经常使用的许多诊断统计量可以写成刻度化预测残差和广义杠杆测度的函数. 为了方便后几章内容的讲述, 我们在这一章中介绍了统计诊断中常用的一些方法, 如均值滑动模型、数据删除法、局部影响分析等. 此外, 我们还给出了已知协方差结构的广义线性模型的局部影响分析的结果, 包括分别扰动协方差矩阵、响应变量、解释性变量得到局部影响分析的一些结果.

第 4 章采用删除多个数据点的方法讨论了数据删除诊断, 得到固定效应和随机效应参数 IGLS 估计的一些数据删除测度, 给出两类新的近似公式, 且与基于 Cook 距离的影响测度进行了比较. 所得的结果可以用于研究任何水平下的个体对固定效应和随机效应参数估计带来的影响. 第 I 种近似就是我们经常在复杂模型中使用的一步近似. 这种近似准则是把固定效应和随机效应参数估计的诊断分析分开进

行, 得到简洁的数据删除测度的表达式, 但是这种近似会忽略或模糊删除多个数据点带来的真实影响. 第 II 种近似能克服这一缺点, 从而得到更有效的诊断统计量.

第 5 章研究了多水平模型下基于均值漂移模型的异常点探测问题. Langford 和 Lewis (1998) 讨论了多水平数据的异常点探测问题, 他们定义了基于残差的水平, 提出使用残差、杠杆值、水平分析法来识别异常点. 然而他们使用的基于残差的统计量是固定效应参数和随机效应参数一步估计的最小二乘回归, 忽略了随机效应参数估计的影响. 在均值漂移模型下, 我们可以同时研究固定效应和随机效应参数 IGLS 估计的异常值识别. 其中我们详细讨论了两水平模型下水平 1 和水平 2 产生均值漂移的结果, 而且每种情况都给出了多个异常点的检验统计量, 这些结果适用于研究任何水平下的异常点探测问题. 与一步近似诊断相比较, 我们的方法更为精确和可靠, 实例分析也证实了这点.

第 6 章研究了多水平模型下高水平单元的局部影响分析, 这是作者的研究结果 (Shi 和 Ojeda, 2004) 的主要内容. 通过同时扰动高水平单元的协方差矩阵, 给出了识别多水平模型固定效应参数和随机效应参数的局部影响诊断统计量.

第 7 章研究了多水平模型下单个观察值的局部影响分析. 这一内容是第 5 章研究内容的继续, 主要用于识别低水平观测值的强影响点. 本章讨论了 (同时) 扰动协方差矩阵的对角元、单个观察值的响应变量和解释性变量, 得到三种干扰模式下固定效应和随机效应参数估计的广义 Cook 统计量及评价观察值联合影响的局部影响诊断统计量, 并提出一种简化计算的方法.

第 8 章介绍目前关于多水平模型的统计软件, 同时给出了利用 Matlab 语言编制的计算程序, 以及在不同的多水平模型及方差结构下分析程序的编制方法. 这一思想还可以推广到更为复杂的多水平模型. 这些程序对热心于实际应用的读者是很有帮助的.

本书的读者要求具有较好的数理统计及线性模型的基础知识. 读者对象主要是数理统计及相关领域的研究生和研究人员. 其中第 2 章和第 8 章对从事多水平模型应用的统计学专业人员很有帮助. 由于作者水平有限, 不妥之处在所难免, 欢迎广大读者批评指正.

本书大部分内容是在我攻读博士学位期间完成的, 在此我要感谢 —— 加拿大 Calgary 大学陈歌迈教授的热心支持和鼓励. 要感谢我的妻子干文和女儿石安琪, 没有她们的关心和支持, 本书是无法完成的. 我的博士生鲁筠为书稿的录入做了许多工作, 国家自然科学基金 (项目编号 10761010 及 10261009) 为我的研究提供了积极的支持, 在此一并致谢!

石　磊

2007 年 12 月于云南财经大学

目　录

第1章 引 论

统计模型诊断是 20 世纪 70 年代中期发展起来的统计学领域一个新的研究方向, 其主要目的是评价统计模型的适当性以及识别数据中可能存在的异常值和强影响点. 在模型适当性的评价方面, 线性模型中目前主要采用残差分析来判断模型拟合的好坏. 异常值的识别主要在一定的异常模型假设下进行统计检验. 而影响点的识别主要是研究数据点 (或数据集) 对我们关注的某个内容的影响程度并识别数据中的特殊结构, 这项工作也称之为影响分析. 统计模型诊断可以为统计模型的改进提供重要的参考信息. 在某些领域中, 异常值及影响点还可以为我们提供某些特殊信息: 如在地质找矿中, 异常值及影响点可能对应着矿产资源富集信息; 而在经济领域, 异常的出现还可能是某种预警信息的表现. 本书主要讨论多水平模型中异常值和影响点的识别, 但主要集中于多水平线性模型.

本章主要给出统计模型诊断中涉及的基本概念和方法. 1.1 节主要回顾了异常值及影响点识别的发展及相关概念; 1.2 节给出了本书中经常使用的一些矩阵知识; 1.3 节以线性模型为例, 介绍了标准线性模型中数据删除法及相关结果; 1.4 节是关于局部影响分析的介绍.

1.1 基 本 概 念

1.1.1 异常值

异常值对我们现代人来说并不陌生, 我们甚至不自觉地会采用一些手段来处理现实生活中出现的异常现象. 最典型的就是在体育比赛中对裁判打分的平均算法 —— 去掉一个最高分和最低分, 再作平均. 人们对异常值的认识可以追朔到 16 世纪, Bernoulli 提到: "在 200 多年前, 人们丢掉异常值的处理方法已经是常见的现象." 处理和识别异常值的统计方法可追溯到 1850 年. 虽然异常值的识别和处理方法在现代已经发展很快, 但是对异常值的定义依然有不同的理解和争论.

例如 Edgeworth (1887) 认为: 不一致观测值 (discordant observation) 可以定义为那些与所在样本中其他数据点遵从的频率规则 (law of frequency) 不一致的观测值.

82 年后, Grubbs(1969) 又这样表述:

一个异常的观测值, 即异常值, 是严重偏离所在样本其他数据点的观测值.

这些表述实质上认为异常值是有目的的、后验的. 这种有目的性的识别异常值的方法, 一般只能在数据中的异常值可以预先通过视觉观察时才能使用 (在一元小样本中较多). 事实上, 对样本量较大或较为复杂的数据集, 比如回归、多元数据、试验设计等, 预先观察到异常值是很困难的. 因此, 就有在观察到异常值之前制定一种客观的准则, 这种准则大多依赖于异常值模型 (outlier model). 由于近几年来强调统计建模的重要性, 许多研究者认为异常值是那些来自于非目标总体 (某种统计模型) 的观察值. Hawkins (1980) 给出了一种比较明确的定义:

异常值是指污染的观测值或不一致观测值的总称. 不一致观测值是指那些让调查者感到吃惊或有较大偏差的数据点. 而污染的观测值是指来自非目标总体的观测值.

异常值模型对异常值的识别和处理是一个非常重要的工具. 从 20 世纪开始, 人们提出了大量的异常值模型来描述异常值的出现. 其中 Dixon(1950) 在正态样本情形建议使用两种分布的混合 (mixture). 第一种是均值滑动模型 (mean-shift model), 即认为数据的分布来自于 $N(\mu,\sigma^2)$ 及 $N(\mu+\lambda,\sigma^2)$ 的混合. 另一种是方差扰动模型, 其分布是 $N(\mu,\sigma^2)$ 与 $N(\mu,a^2\sigma^2)$ 的混合, 这里 λ, $a^2>1$ 是扰动参数. 这种异常值模型的引入导致我们可以通过传统的统计检验问题来拒绝异常值. 有些研究者还利用这种混合模型来改进参数估计, 处理一些具有重尾 (heavy-tail) 结构的数据.

在正态线性模型中, 异常值识别问题使用最为广泛的就是 mean-shift model:

$$\boldsymbol{Y}=\boldsymbol{X}\beta+\boldsymbol{D}\gamma+\boldsymbol{e},$$

其中 $\boldsymbol{X}$ 为 $n\times p$ 设计矩阵, $\boldsymbol{Y}$ 为观测向量, $\boldsymbol{e}$ 为随机误差, β 为未知的回归系数, $\boldsymbol{D}$ 为 $n\times m$ 矩阵, γ 为 $m\times 1$ 未知向量. 令 $\boldsymbol{D}=(d_1,\cdots d_j\cdots,d_m)$, 其中 d_j 为第 j 个元素为 1 其余元素为零的 $n\times 1$ 向量. 则上述模型即表示在 $I=(i_1,\cdots,i_m)$ 中的观察点 $(y_i,\boldsymbol{x}'_i)$ 出现均值滑动. 因此 $I=(i_1,\cdots,i_m)$ 中的观察值是否为异常值, 就对应于 $H_0:\gamma=0$ 的检验.

当 $m=1$ 时, 检验统计量等价于如下的刻度化残差

$$t_i^*=\frac{\hat{e}_i}{\hat{\sigma}_{[i]}\sqrt{(1-h_{ii})}},\quad i=1,\cdots,n$$

其中 $\hat{e}_i=y_i-\boldsymbol{x}_i'\hat{\beta}$, h_{ii} 是 $\boldsymbol{H}=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 的第 i 个对角元, $\hat{\sigma}_{[i]}$ 是删除第 i 个数据点后 σ^2 的估计. 在零假设下 $t_i^*\sim t_{n-p}$.

有了检验统计量的分布形式, 我们就很容易确定临界值, 进而给出检验过程. Barnett 和 Lewis(1994) 将上述均值滑动异常值模型区分为两种, 一种称为有标识的异常值模型 (labeled). 此时 $I=(i_1,\cdots,i_m)$ 被预先认为是异常值, 其检验过程可以基于 $t_I/(n-p)$ 确定其临近值. 另一种异常值模型假定我们并不知道 $I=(i_1,\cdots,i_m)$,

此时检验统计量应该是 $\max_I |t_I|$. 在 $k=1$ 时, 使用的检验统计量为

$$|t_k| = \max_{1\leqslant i\leqslant n} |t_i|.$$

上述检验过程首先由 Anscombe(1960), Daniel(1960) 提出的, 但此时由于 $|t_k|$ 的精确分布无法求出, 只能由近似的方法来确定临界值, 如 Bonferroni 不等式的应用是一种常用的近似方法, 具体内容可参考 Cook 和 Weisberg(1982).

异常值检验的主要内容及结果可参阅 Barnett 和 Lewis (1994), Bechman 和 Cook (1983), 及 Hawkins (1980). 此外在时间序列模型中, Fox (1972) 根据时间序列模型的特点, 提出了可加异常值 (additive outlier) 及革新异常值 (innovation outlier) 的概念, 并基于似然比检验给出了近似的检验过程. 之后, 时间序列中异常值的研究获得了许多丰富的成果, 代表性的文献有 Abraham 和 Box (1979), Box 和 Tiao (1975), Tsay (1986, 1988), Chan, Tiao 和 Chen (1988). 在具有随机效应的单向分类及双向分类模型中, 基于似然比检验, 石磊等 (1997, 1999, 2006, 2007) 在一系列论文中也给出了许多研究成果 (包括单个及多个异常值检验).

1.1.2 影响点

当数据中的某些点被识别为异常值时, 统计学家关心的一个问题是这些点对模型参数的估计和统计推断是否产生较大的影响, 这一问题的提出导致了回归诊断中的另一个研究分支 —— 影响分析. 最早给出影响分析的概念是文献 (Cook, 1977). 影响分析的主要任务是识别数据中那些对模型参数的估计和统计推断产生较大影响的点, 称之为影响点. 在影响分析的研究中, 首先要指明的是对什么的影响, 比如是对参数估计, 还是检验统计量, 或是置信区间. 例如, 假设 θ 是给定模型中的未知参数, $\hat{\theta}$ 是 θ 的某种估计. 这种估计可能是极大似然估计, Bayes 估计或是线性模型中的最小二乘估计. 如果我们想要了解数据对参数估计 $\hat{\theta}$ 的影响, 一种比较直观的方法是通过比较删除某个或某几个数据点后的参数估计与完全数据下的参数估计, 来判定这些点对参数估计的影响:

$$D_i = \frac{(\hat{\theta} - \hat{\theta}_{[i]})' M (\hat{\theta} - \hat{\theta}_{[i]})}{c}, \tag{1.1}$$

其中 $\hat{\theta}, \hat{\theta}_{[i]}$ 表示在原模型及第 i 个数据点删除后 θ 的估计. 这种方法称为数据删除法 (case deletion). 从理论上讲, 式 (1.1) 中的 $\hat{\theta}_{[i]}$ 可以通过删除第 i 个数据点后重新计算 θ 的估计获得 (这称为真实的影响), 但这就不是统计诊断了. 在现代计算机速度大幅提高的时代, 这样做似乎也是可行的, 但当数据量很大, 每次估计都需要迭代, 同时要对多个数据点 (的所有组合) 进行计算时, 计算工作量是很大的. 因此, 统计诊断内容是要通过某种技巧, 获得 $\hat{\theta}_{[i]}$ 的一个递推公式 (update formula),

即 $\hat{\theta}_{[i]}$ 的计算只能通过完全数据下的估计及相关的统计量获得, 即

$$\hat{\theta}_{[i]} = \hat{\theta} + \Delta(\hat{\theta}, \boldsymbol{Y}), \tag{1.2}$$

其中 $\boldsymbol{Y}$ 是观测数据. $\Delta(\hat{\theta}, \boldsymbol{Y})$ 是删除数据前后估计的差, 它仅仅通过在完全数据下的估计计算 (只进行一次估计计算) 获得. 异常值识别用到的统计量也是如此.

影响分析的研究发展很快, Cook(1986) 创新性地提出了一种新的方法, 称为局部影响分析. 该方法使影响分析得到了快速的发展, 这一内容我们将在 1.3 节中讨论.

1.1.3　异常值和影响点的处理

异常值和影响点是不同的概念, 但有时也很难区分. 异常值可能是影响点, 也可能不是影响点, 反之亦然. 这主要是因为它们的定义不同. 这方面的讨论可参考韦博成等 (1992) 的著作. 然而异常值和影响点都是相对于给定的模型而言的, 不同模型对应的异常值和影响点可能是不同的. 当模型变化后, 原来异常的点或有影响的点可能不再是异常值和影响点.

对异常值和影响点的处理也许是实际工作者最关心的问题, Barnett(1978), Barnett 和 Lewis(1994) 将其分为两类:

(1) 一类称为识别 (identification), 或识别以进一步研究. 这类方法可提供如下几种处理方式:

(a) 拒绝, 丢弃;

(b) 提供一种通常情况下无法注意的信息;

(c) 提供一种对原有模型式参数估计方法改进的信息;

(d) 识别出数据中有缺陷的数据, 以便进一步进行试验.

异常值的识别可以帮助对模型进行不断修正, 直至获得一个理想的模型. 这在统计建模中是一个非常重要的工作, 具体讨论可参阅 Box(1979, 1980) 及 Cook 和 Weisberg (1982).

(2) 第二类称之为调整 (accommodation). 在这类方法中, 我们对数据中存在异常值时处理的方法是对模型或分析方法进行适当修正. 例如, 利用混合模型来适应污染值, 当混合模型对称时, M 估计常常可提供一种对异常值比较稳健的估计. 当然, 直接用稳健估计也是调整的一种处理方式.

1.2　预 备 知 识

本节给出书中涉及的一些部分矩阵代数, 随机变量的分布、估计及检验理论. 除了给出证明的结果之外, 其余结论是大家熟知的, 可参阅 (王松桂等, 2004; Rao,

Toutenburg(1995), 张贤达 (2004), 孙文爽等 (1994).

1.2.1 一些矩阵代数

引理 1.1 设矩阵 A 有如下分块

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

若 $|A_{11}|\neq 0$, 则

$$A^{-1}=\begin{bmatrix} A_{11}^{-1}+A_{11}^{-1}A_{12}A_{22.1}^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22.1}^{-1} \\ -A_{22.1}^{-1}A_{21}A_{11}^{-1} & A_{22.1}^{-1} \end{bmatrix}.$$

若 $|A_{22}|\neq 0$, 则

$$A^{-1}=\begin{bmatrix} A_{11.2}^{-1} & -A_{11.2}^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}A_{11.2}^{-1} & A_{22}^{-1}+A_{22}^{-1}A_{21}A_{11.2}^{-1}A_{12}A_{22}^{-1} \end{bmatrix}.$$

其中 $A_{22.1}=A_{22}-A_{21}A_{11}^{-1}A_{12}$, $A_{11.2}=A_{11}-A_{12}A_{22}^{-1}A_{21}$.

引理 1.2 设 A 是一个 $n\times n$ 矩阵, B 和 C 分别是 $n\times m$ 和 $m\times n$ 矩阵, 且秩均为 m, 则

$$(A+BC)^{-1}=A^{-1}-A^{-1}B(I_m+CA^{-1}B)^{-1}CA^{-1}.$$

当 $m=1$, $B=b$, $C=c'$, 若 $1+c'A^{-1}b\neq 0$, 则上式变为

$$(A+bc')^{-1}=A^{-1}-\frac{A^{-1}bc'A^{-1}}{1+c'A^{-1}b}.$$

定义 1.1(Kronecker 乘积) 设 $A=(a_{ij})$ 和 $B=(b_{ij})$ 分别是 $n\times m$ 和 $p\times q$ 矩阵, 令 $a_{ij}B$ 表示 a_{ij} 乘以 B 的所有元素得到的 $p\times q$ 矩阵, A 与 B 的 Kronecker 乘积定义为

$$A\otimes B=\begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1m}B \\ a_{21}B & a_{22}B & \cdots & a_{2m}B \\ \vdots & \vdots & & \vdots \\ a_{n1}B & a_{n2}B & \cdots & a_{nm}B \end{bmatrix}.$$

定义 1.2(vec 运算) 设 $A=(a_1,a_2,\cdots,a_m)$ 是一 $n\times m$ 矩阵. 称

$$\mathrm{vec}(A)=(a_1',a_2',\cdots,a_m')'$$

为矩阵 A 的按列拉直运算.

引理 1.3　Kronecker 乘积及向量化运算有如下结果：

(1) $(A_1+A_2)\otimes B = A_1\otimes B + A_2\otimes B$;

(2) $(\alpha A)\otimes(\beta B) = \alpha\beta(A\otimes B)$;

(3) $(A_1\otimes B_1)(A_1\otimes B_1) = (A_1A_2)\otimes(B_1B_2)$;

(4) $(A\otimes B)' = A'\otimes B'$;

(5) $(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$;

(6) $\mathrm{vec}(A+B) = \mathrm{vec}(A) + \mathrm{vec}(B)$;

(7) $\mathrm{vec}(\alpha A) = \alpha\mathrm{vec}(A)$;

(8) $\mathrm{tr}(AB) = \mathrm{vec}'(A')\mathrm{vec}(B)$;

(9) $\mathrm{tr}(ABC) = \mathrm{vec}'(A)(I\otimes B)\mathrm{vec}(C)$;

(10) $\mathrm{tr}(A) = \mathrm{tr}(I_nA) = \mathrm{vec}'(I_n)\mathrm{vec}(A)$;

(11)$\mathrm{vec}(ABC) = (C'\otimes A)\mathrm{vec}(B)$;

(12) 设 a 及 b 分别为 $n\times 1$ 及 $p\times 1$ 向量, 则 $\mathrm{vec}(ab') = b\otimes a$.

引理 1.4　设 A 为 $n\times n$ 幂等方阵, 即 $A^2 = A$. 若 $\mathrm{tr}(A) = r \leqslant n$, 则

(1) A 的特征值为 1 或 0;

(2) $\mathrm{tr}(A) = \mathrm{rank}(A) = r$;

(3) 若 $r = n$, 则 $A = I_n$;

(4) 若 A 和 B 都是 $n\times n$ 幂等方阵, 且 $AB = BA$, 则 AB 也是幂等方阵;

(5) $I_n - A$ 也是 $n\times n$ 幂等方阵, 且 $A(I_n - A) = (I_n - A)A = 0$.

引理 1.5　X 为 $n\times p$ 矩阵, $\mathrm{rank}(A) = p$, 则有 $H = X(X'X)^{-1}X'$, $I_n - H$ 是幂等矩阵, 且

$$HX = X,\quad X'H = X',\quad (I_n - H)X = 0,\quad X'(I_n - H) = 0.$$

引理 1.6　设 A 为 $n\times n$ 对称阵, $\lambda_1 \leqslant \cdots \leqslant \lambda_n$ 为 A 的特征值, $\alpha_1,\cdots,\alpha_n$ 为其对应的特征向量, 则

(1) $\sup_{x\neq 0}\dfrac{x'Ax}{x'x} = \alpha_1'A\alpha_1 = \lambda_1$,　$\inf_{x\neq 0}\dfrac{x'Ax}{x'x} = \alpha_n'A\alpha_n = \lambda_n$;

(2) $\sup_{x'x=1} x'Ax = \alpha_1'A\alpha_1 = \lambda_1$,　$\inf_{x'x=1}\dfrac{x'Ax}{x'x} = \alpha_n'A\alpha_n = \lambda_n$.

引理 1.7(Cholesky 分解)　设 A 为 $n\times n$ 对称正定矩阵, 则存在唯一的一个下三角矩阵 G, 使得

$$A = GG'.$$

引理 1.8　假设 A 和 B 是两个阶数相同的对称阵, 且 A 可逆. 则对于任意小的 ϵ, 我们有

$$(A+B\epsilon)^{-1} = A^{-1} - A^{-1}BA^{-1}\epsilon + o(\epsilon^2), \tag{1.3}$$

$$(A+B\epsilon+o(\epsilon^2))^{-1}=A^{-1}-A^{-1}BA^{-1}\epsilon+o(\epsilon^2). \tag{1.4}$$

证明 首先证明对任意的对称矩阵 B 和任意小的 ϵ,

$$(I+B\epsilon)^{-1}=I-B\epsilon+o(\epsilon^2). \tag{1.5}$$

假设 B 可以分解为

$$B=\sum_{i=1}^{m}\lambda_i\alpha_i\alpha_i',$$

其中 λ_i 和 α_i, $i=1,\cdots,m$ 分别为 B 的非 0 特征值及其相应的特征向量, m 为 B 的秩, 则

$$\begin{aligned}(I+B\epsilon)^{-1}&=\left(I+\sum_{i=1}^{m}\lambda_i\alpha_i\alpha_i'\epsilon\right)^{-1}=\left(\sum_{i=1}^{m}(1+\lambda_i\epsilon)\alpha_i\alpha_i'\right)^{-1}\\&=\sum_{i=1}^{m}(1+\lambda_i\epsilon)^{-1}\alpha_i\alpha_i'=\sum_{i=1}^{m}(1-\lambda_i\epsilon+o(\epsilon^2))\alpha_i\alpha_i'\\&=I-B\epsilon+o(\epsilon^2).\end{aligned}$$

式 (1.5) 得证. 注意到

$$\begin{aligned}(A+B\epsilon)^{-1}&=A^{-\frac{1}{2}}(I+A^{-\frac{1}{2}}BA^{-\frac{1}{2}}\epsilon)^{-1}A^{-\frac{1}{2}}\\&=A^{-1}-A^{-1}BA^{-1}\epsilon+o(\epsilon^2),\end{aligned}$$

其中 $A^{\frac{1}{2}}$ 表示矩阵 A 的平方根分解. 类似的我们有

$$(A+B\epsilon+o(\epsilon^2))^{-1}=A^{-1}-A^{-1}BA^{-1}\epsilon+o(\epsilon^2).$$

结论得证.

1.2.2 矩阵微商

定义 1.3 设 $y=f(X)$ 是 $n\times m$ 矩阵 $X=(x_{ij})$ 的一元实值函数, 其对 x_{ij} 是可微的, 则 $y=f(X)$ 对 X 的微商定义为

$$\frac{\partial y}{X}=\begin{bmatrix}\dfrac{\partial f}{x_{11}} & \cdots & \dfrac{\partial f}{x_{1m}}\\ \vdots & & \vdots\\ \dfrac{\partial f}{x_{n1}} & \cdots & \dfrac{\partial f}{x_{nm}}\end{bmatrix}.$$

引理 1.9 设 x,y 为 n 维向量, A 为 $n\times n$ 对称阵, 则