线性和广义线性混合模型及其统计诊断

费　宇　陈　飞
喻达磊　韩俊林　著

科学出版社

线性和广义线性混合模型及其统计诊断

费 宇 陈 飞 喻达磊 韩俊林 著

科 学 出 版 社

北 京

内 容 简 介

本书系统介绍线性混合模型和广义线性混合模型的基本理论和方法. 主要包括两类模型的参数估计、假设检验、置信区域和统计诊断问题. 重点是两类模型的统计诊断分析, 采用数据删除方法研究两类模型影响点的探测问题, 基于EM算法中的Q函数, 来构建影响度量——广义Cook统计量, 解决了一般方差结构的两类混合模型统计诊断的困难. 而且, 获得的影响度量有很好的统计意义, 能够方便地用于全参数(均值参数与方差参数)和部分参数(均值参数或方差参数)的诊断分析.

本书可以作为统计专业高年级本科生及研究生的教材和参考书, 也可以作为数学、生物、医学和经济等领域教师和研究人员的参考书.

图书在版编目(CIP)数据

线性和广义线性混合模型及其统计诊断/费宇等著. —北京: 科学出版社, 2013.3

ISBN 978-7-03-036479-1

Ⅰ. ①线… Ⅱ. ①费… Ⅲ. ①线性模型–研究 ②线性模型–统计分析(数学) Ⅳ. ①O212

中国版本图书馆CIP数据核字 (2013) 第012694号

责任编辑: 李 欣 赵彦超/责任校对: 李 影
责任印制: 钱玉芬/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

北京佳艺恒彩印刷有限公司 印刷
科学出版社发行 各地新华书店经销
*
2013年3月第 一 版 开本: B5(720×1000)
2013年3月第一次印刷 印张: 11 1/4
字数: 208 000
定价: 48.00元
(如有印装质量问题, 我社负责调换)

前　言

线性混合模型和广义线性混合模型是两类重要的应用广泛的模型, 近年来关于两类模型的参数估计、假设检验、置信区域和统计诊断问题方面的研究有了很多新的进展. 本书试图向读者系统介绍近几年来关于两类模型的研究结果, 特别是两类模型和统计诊断分析是本书的重要内容.

第 1 章简要介绍普通线性模型、广义线性模型、线性混合模型和广义线性混合模型的定义, 统计诊断的含义和基本方法, 还给出了一些矩阵代数和矩阵微商法则, 方便以后应用.

第 2 章介绍线性混合模型的定义、常见类型. 在正态假定下讨论了线性混合模型参数的最大似然估计和限制最大似然估计, 在非正态假定下讨论了方差分量参数的估计方法, 接着研究了该模型的假设检验和置信区域理论, 最后介绍线性混合模型随机效应的预测问题和模型选择方法.

第 3 章研究线性混合模型的统计诊断. 先从一般似然函数出发讨论独立方差结构下线性混合模型的影响点探测问题, 然后从 EM 算法中的 Q 函数出发来构建影响度量, 来研究非独立方差结构下线性混合模型的统计诊断, 获得了两个重要的基于 Q 函数的广义 Cook 统计量, 它们有很好的统计意义, 能够方便地用于全参数 (均值参数与方差参数) 和部分参数 (均值参数或方差参数) 的统计诊断分析.

第 4 章介绍广义线性混合模型的定义、模型参数估计 (包括基于 EM 算法的最大似然估计和基于条件似然的参数估计), 讨论了估计量的大样本性质, 然后研究了参数的区间估计和假设检验问题, 最后介绍了广义线性混合模型的选择准则.

第 5 章研究广义线性混合模型的统计诊断. 首先在一般似然函数框架下研究模型的影响点探测问题, 然后从 EM 算法中的 Q 函数出发构造影响度量, 来讨论模型的统计诊断分析, 最后介绍了模型扰动的选择问题.

本书内容是国家自然科学基金项目 “线性混合模型和广义线性混合模型均值和方差协方差结构的同时拟合及其统计诊断”(项目编号：11061036) 的部分研究成果, 同时, 本书的出版也获得国家自然科学基金数学天元基金项目 “充分降维理论

中基于分布加权思想的压缩估计”(项目编号：11126297)、云南财经大学统计学博士点建设基金的支持, 在此表示诚挚的感谢.

作　者

2012 年 10 月于昆明

目　录

插图目录

表 格 目 录

第1章 引 论

1.1 线性模型简介

1.1.1 普通线性模型

定义 1.1 如下模型通常用来描述 y 与 x 之间的随机线性关系

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon, \tag{1.1}$$

其中 $x_1, \cdots, x_k$ 是非随机的自变量, y 是因变量, β_0 是常数项, $\beta_1, \cdots, \beta_k$ 是回归系数, ε 是随机误差项.

假设对 $y, x_1, \cdots, x_k$ 进行了 n 次观测, 得到 n 组观测值 $y_i, x_{1i}, \cdots, x_{ki} (i = 1, \cdots, n)$, 它们满足关系式

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i, \tag{1.2}$$

引入矩阵记号, 记

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

则模型 (1.2) 可以写成如下形式

$$y = X\beta + \varepsilon, \tag{1.3}$$

其中 y 是 $n \times 1$ 观测向量, X 是 $n \times (k+1)$ 已知设计阵, ε 是随机误差向量, β 是未知参数向量.

如果模型 (1.3) 满足条件: (1) $E(\varepsilon) = 0$, (2) $\mathrm{var}(\varepsilon) = \sigma^2 I$, (3) $x_1, \cdots, x_k$ 不相关, 则称模型 (1.3) 为普通线性回归模型 (ordinary linear regression model).

进一步, 如果模型的随机误差项服从正态分布, 即 $\varepsilon \sim N(0, \sigma^2 I)$, 则称模型

(1.3) 为普通正态线性回归模型.

下面给一个例子说明.

例 1.1(普通线性回归模型)　考虑变量 y 随变量 $x_1,\cdots,x_k$ 变化的情况, 随机观测了 n 组观测值 $(y_i,x_{1i},\cdots,x_{ki})(i=1,\cdots,n)$, 这里 y_i 是 y 的第 i 个观测值, x_{ji} 是 $x_j(j=1,\cdots,k)$ 的第 i 个观测值, 影响 y_i 取值的主要因素是 $x_{1i},\cdots,x_{ki}$. 此外, 随机误差也要考虑, 于是可以将 y 随 x 变化的情况写出如下形式

$$y_i=f(x_{1i},\cdots,x_{ki})+\varepsilon_i\quad(i=1,\cdots,n). \tag{1.4}$$

如果函数 f 是线性函数, 即 $f(x_{1i},\cdots,x_{ki})=\beta_0+\beta_1x_{1i}+\cdots+\beta_kx_{ki}$, 则模型 (1.4) 就是一个线性回归模型. 如果变量 $x_1,\cdots,x_k$ 相互独立, 随机误差项 ε_i 相互独立, 均值为 0, 方差为 σ^2, 则该模型是一个普通线性回归模型.

1.1.2　广义线性模型

模型 (1.3) 是一般的线性模型, 可以处理连续型变量, 即 y 变量是连续型变量的情况, 但在实际中, 很多变量不是连续的, 例如调查顾客是否购买了某种商品, 这里的响应变量 y 是二值变量, 比如 $y=1$ 表示购买了该商品, $y=0$ 表示没有购买该商品. 对于 y 是二值变量或其他非连续变量的情况, 不能直接采用一般线性模型 (1.3) 进行分析, 可以采用本节定义的广义线性模型 (generalized linear model, GLM).

定义 1.2　广义线性模型 (McCulloch et al., 2002) 是一般线性模型的推广, 其定义由以下三个部分组成:

(1) 随机成分 (random components): 设 $y_1,\cdots,y_n$ 是来自于指数分布族的随机样本, 即 y_i 的密度函数为

$$f(y_i,\alpha_i,\phi)=\exp\left\{\frac{y_i\alpha_i-b(\alpha_i)}{a_i(\phi)}+c_i(y_i,\phi)\right\}, \tag{1.5}$$

其中 ϕ 是散度参数 (dispersion parameter), $a_i(\cdot),b(\cdot)$ 和 $c_i(\cdot)$ 是已知函数.

(2) 系统成分 (system components): 对于第 i 个响应 y_i, 以下系统成分称为线性预测项 (linear predictor)

$$\eta_i=x_i^{\mathrm{T}}\beta=\sum_{j=1}^{k}x_{ij}\beta_j\quad(i=1,\cdots,n), \tag{1.6}$$

它是协变量 x_j 的线性组合.

(3) 关联函数 (link function)：设 $\mu_i = E(y_i)$ 是 y_i 的期望, 而 $g(\cdot)$ 是单调可微函数, 将随机成分的期望 μ_i 与系统成分联结起来, 即

$$g(\mu_i) = \eta_i = x_i^{\mathrm{T}}\beta \quad (i = 1, \cdots, n). \tag{1.7}$$

引入矩阵符号, 记

$$\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}_{n\times 1}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}_{n\times 1}, \quad X = \begin{bmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{bmatrix}_{n\times k},$$

则关联函数可以写成以下矩阵形式

$$g(\mu) = \eta = X\beta, \tag{1.8}$$

其中 β 是未知参数.

以上定义的广义线性模型是一类广泛的线性模型, 实际中应用广泛的概率单位模型 (probit model) 和逻辑斯谛模型 (logistic model) 就属于广义线性模型.

例 1.2 (概率单位模型) 设 y_i 服从参数为 p_i 的伯努利分布 (Bernoulli distribution), 即 $y_i \sim B(p_i)(i = 1, \cdots, n)$, 则 $\mu_i = E(y_i) = p_i > 0$, 采用正态关联函数, 即 $g(\mu_i) = \Phi^{-1}(p_i) = x_i^{\mathrm{T}}\beta$, 其中 $\Phi(\cdot)$ 是正态分布的分布函数, 此模型称为概率单位模型.

例 1.3 (逻辑斯谛模型) 设 y_i 服从参数为 p_i 的伯努利分布, 即 $y_i \sim B(p_i)(i = 1, \cdots, n)$, 则 $\mu_i = E(y_i) = p_i > 0$, 采用逻辑关联函数, 即 $g(\mu_i) = \mathrm{logit}(p_i) = \log\dfrac{p_i}{1-p_i} = \eta_i = x_i^{\mathrm{T}}\beta$, 此模型称为逻辑斯谛模型.

1.1.3 线性混合模型

线性模型 $y = X\beta + \varepsilon$ 中, 回归系数 β 是未知常数, 但实际中, 在某些情况下, 系数视为随机的更合理. 比如在医学研究中, 随机观测了 m 个病人的某个指标的测量值, 每个病人都作了若干次重复测量, 用 y_{ij} 记第 i 个病人的第 j 次观测值 $(i = 1, \cdots, m; j = 1, \cdots, n_i)$, 显然, 第 i 个病人的 n_i 次观测值是相关的, 假定 u_i 是第 i 个病人的随机效应 (random effect), 则可以采用以下线性混合模型 (linear mixed model, LMM) 来拟合 y_{ij}, 即

$$y_{ij} = x_{ij}^{\mathrm{T}}\beta + u_i + \varepsilon_{ij} \quad (i = 1, \cdots, m; j = 1, \cdots, n_i),$$

其中 x_{ij} 是已知协变量, β 是未知回归系数, u_i 是未知随机效应, 一般假定 u_i 是相互独立同分布的随机变量, 均值为 0, 方差为 σ_u^2; 而 ε_{ij} 是随即误差项, 均值为 0, 方差为 σ_ε^2.

定义 1.3　一般的线性混合模型 (Laid et al., 1982) 定义为

$$y = X\beta + Zu + \varepsilon, \tag{1.9}$$

其中 y 是响应观测向量, X 是已知协变量矩阵, β 是未知回归系数, 通常称为固定效应, Z 是已知矩阵, u 是随机效应, ε 是随机误差, 一般假定 u 与 ε 相互独立.

进一步, 不失一般性, 可以假定 $E(u) = 0, E(\varepsilon) = 0$, 而 $\mathrm{var}(u) = G, \mathrm{var}(\varepsilon) = R$, 如果假定 u 和 ε 均服从正态分布, 即 $u \sim N(0, G)$, $\varepsilon \sim N(0, R)$, 称模型为正态线性混合模型.

线性混合模型是一类应用非常广泛的模型, 可以用来拟合多种复杂数据, 比如纵向数据和面版数据等, 线性混合模型 (1.9) 包含了许多常用模型, 比如方差分量模型, 含协变量的两向随机效应模型、纵向模型、增长曲线模型等 (具体参见 2.2 节), 下面仅给一个例子说明.

例 1.4 (单向随机效应模型)　医学实验中, 通常要比较 k 种药治疗某种疾病的效果, 药效度量指标为 Y, 通常采用双盲实验法, 假设随机抽取了 $n = mk$ 个病人, 分为 k 组, 每组有 m 个人, 分别服用 k 种药, 记 y_{ij} 是服用第 i 种药的第 j 个病人的药效测量值, 则 y_{ij} 可以表示为

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad (i = 1, \cdots, k; j = 1, \cdots, m), \tag{1.10}$$

其中 μ 称为总平均, α_i 表示第 i 种药的效应, ε_{ij} 是随机误差.

在这个问题中, 我们感兴趣的因子是药品, 它有 k 个不同的品种, 称为因子的水平或处理, 模型 (1.10) 称为单向分类模型 (或单因素方差分析模型). 引入矩阵记号, 记 $X = \mathbf{1}_n, \beta = \mu$, 且

$$y = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1m} \\ y_{21} \\ \vdots \\ y_{km} \end{bmatrix}_{n\times 1}, \quad Z = \begin{bmatrix} \mathbf{1}_m & & & \\ & \mathbf{1}_m & & \\ & & \ddots & \\ & & & \mathbf{1}_m \end{bmatrix}_{n\times k}, \quad u = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{bmatrix}_{k\times 1}, \quad \varepsilon = \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1m} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{km} \end{bmatrix}_{n\times 1},$$

则上述模型可以表示为 $y = X\beta + Zu + \varepsilon$, 即线性混合模型 (1.9) 的形式.

1.1.4 广义线性混合模型

广义线性混合模型 (generalized linear mixed model, GLMM) 是广义线性模型 (GLM) 和线性混合模型 (LMM) 的有机结合, 可以用来拟合离散型非独立的一类数据, 这类数据无法单独运用广义线性模型或线性混合模型来处理, 下面先给出广义线性混合模型的定义, 然后给一个例子说明.

定义 1.4 广义线性模型 (Jiang, 2007) 的定义由以下三个部分组成:

(1) 随机成分 (random components): 设给定随机向量 $u = (u_1, \cdots, u_m)^{\mathrm{T}}$, 响应变量 $y_1, \cdots, y_m$ 是来自于指数分布族的变量, 其密度函数为

$$f(y_i, \alpha_i, \phi|u) = \exp\left\{\frac{y_i\alpha_i - b(\alpha_i)}{a_i(\phi)} + c_i(y_i, \phi)\right\}, \tag{1.11}$$

其中 ϕ 是散度参数 (dispersion parameter), $a_i(\cdot), b(\cdot)$ 和 $c_i(\cdot)$ 是已知函数.

(2) 系统成分 (system components): 对于第 i 个响应 y_i, 以下系统成分称为线性预测项 (linear predictor)

$$\eta_i = x_i^{\mathrm{T}}\beta + z_i^{\mathrm{T}}u \quad (i = 1, \cdots, n), \tag{1.12}$$

其中 x_i 和 z_i 是已知向量, 而 β 和 u 是未知的固定效应和随机效应.

(3) 关联函数 (link function): 在给定 u 的条件下, y_i 的条件期望是 μ_i, 即 $\mu_i = E(y_i|u)$, 而 $g(\cdot)$ 是单调可微函数, 将随机成分的期望 μ_i 与系统成分 η_i 联结起来, 即

$$g(\mu_i) = \eta_i = x_i^{\mathrm{T}}\beta + z_i^{\mathrm{T}}u_i \quad (i = 1, \cdots, n). \tag{1.13}$$

引入矩阵符号, 记

$$\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}_{n\times 1}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}_{n\times 1}, \quad X = \begin{bmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{bmatrix}_{n\times k},$$

$$Z = \begin{bmatrix} z_1^{\mathrm{T}} \\ z_2^{\mathrm{T}} \\ \vdots \\ z_n^{\mathrm{T}} \end{bmatrix}_{n\times m}, \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_m \end{bmatrix}_{m\times 1},$$

则关联函数可以写成以下矩阵形式

$$g(\mu)=\eta=X\beta+Zu. \tag{1.14}$$

例 1.5 (广义逻辑斯谛模型)　医学研究经常涉及到配对二值数据, 比如想了解癌症病人在主要的癌症治疗中心是否比在一般的社区医院能接受更有效的治疗, 我们不能简单地比较两个地方癌症治疗的有效率, 因为两个地方病人的数目差别很大, 比如, 癌症治疗中心可能提供较差的治疗, 原因是他们要治疗最难的癌症. 一个可能的方法是采用配对设计: 根据治疗的时间、方法和病人的年龄, 我们从癌症治疗中心随机抽取一个病人与从社区医院抽取的一个病人配成一对, 假设响应变量是90天内肿瘤大小有没有缩小, 用 $y_{ij}=1$ 表示缩小了, 用 $y_{ij}=0$ 表示没有缩小, 这里 $i=1,\cdots,m$ 表示病人对别; 而 j 表示医院类型, $j=1$ 表示癌症治疗中心, $j=2$ 表示一般社区医院; 协变量 x_{ij} 表示病人来自于哪种医院, $x_{i1}=0$ 表示来自癌症治疗中心, $x_{i2}=1$ 表示来自一般社区医院; 记 $p_{ij}=P(y_{ij}=1|u_i)$, 其中 u_i 表示第 i 个病人的随机效应. 于是, 给定 u_i 条件下, y_{ij} 服从参数为 p_{ij} 的伯努利分布, 即 $y_{ij}|u_i\sim B(p_{ij})(i=1,\cdots,m;j=1,2)$, 显然 $\mu_{ij}=E(y_{ij}|u_i)=p_{ij}>0$, 采用逻辑关联函数, 即 $g(\mu_{ij})=\mathrm{logit}(p_{ij})=\log\dfrac{p_{ij}}{1-p_{ij}}=\eta_{ij}=\alpha+\beta x_{ij}+u_i$, 这样得到的广义线性混合模型称为广义逻辑斯谛模型.

1.2　统计诊断概述

1.2.1　统计诊断的含义

统计模型是统计分析的重要手段, 采用统计模型拟合数据, 进行统计预测和推断是数据分析的常用方法, 只有模型是合理的情况下相应的统计预测和推断才是有效的, 而模型是否合理需要进行统计检验和统计诊断分析, 因此, 统计诊断是数据分析的重要环节. 统计诊断的主要任务是通过诊断统计量检测用既定模型 (postulated model) 拟合观测数据的合理性, 具体说来, 统计诊断主要检测两类数据点: 一是异常点 (outlier), 即严重偏离既定模型的数据点; 二是强影响点 (influential case), 即对统计推断 (比如参数估计、假设检验等) 的结果有重要影响的点.

需要指出的是, 异常点很可能是强影响点, 强影响点也很可能是异常点; 但二者之间的联系不是必然的, 即异常点也可能不是强影响点, 反之亦然. 异常点与强影响

点之间的关系是一个比较复杂的有争议的问题, 它与数据的实际背景有很大关系, 关于二者关系的讨论可以参阅 Beckman 和 Cook(1983), Cook 等 (1982), Chatterjee 和 Hadi(1988), 韦博成等 (1991), Pan 和 Fang(2002).

统计诊断方法总是结合数据点进行分析, 研究它们对于统计推断的影响, 所以也把这个过程称为影响分析 (influence analysis).

1.2.2 统计诊断的主要方法

统计诊断有两种主要方法: 点删除 (case deletion) 方法和局部影响分析 (local influence analysis) 方法, 下面简要介绍这两种方法.

1. 点删除方法

点删除方法也称数据删除方法 (Cook, 1977; Cook et al., 1982), 它通过比较点删除模型与原模型相应统计量之间的差异, 进行统计诊断分析, 它是统计诊断的最基本的方法, 适用于线性回归模型和其他更复杂的统计模型.

这里以线性回归模型简单说明点删除方法的含义. 给定一组数据集 $X = \{x_1, \cdots, x_n\}$, 假设 X 对应参数模型 $M(\theta)$, 即 $X \sim M(\theta)$. 参数 θ 的一个估计为 $\hat{\theta}$(MLE 或 LS 估计等), 研究数据点 x_i 对 $\hat{\theta}$ 的影响, 考虑删除 x_i 前后估计量的变化, 设删除 x_i 后 θ 的估计为 $\hat{\theta}_{[i]}$, 构造某种合适的"距离" D_i, 用来度量 $\hat{\theta}_{[i]}$ 与 $\hat{\theta}$ 之间的"差异", D_i 通常称为诊断统计量. 最常用就是如下定义的 Cook 距离

$$D_i(M) = (\hat{\theta}_{[i]} - \hat{\theta})^{\mathrm{T}} M (\hat{\theta}_{[i]} - \hat{\theta}), \tag{1.15}$$

其中 M 是权矩阵 (Cook et al., 1982).

如果 x_i 是一个正常的点, 则 $\hat{\theta}_{[i]}$ 与 $\hat{\theta}$ 应该相差不大, 从而 D_i 会比较小; 如果 D_i 比很大, 则 $\hat{\theta}_{[i]}$ 与 $\hat{\theta}$ 相差很大, 说明 x_i 的存在与否对 θ 的估计值有重大影响, 即 x_i 对 $\hat{\theta}$ 有很大影响, 那么 x_i 可能是异常点或强影响点. 本书采用的统计诊断方法是点删除方法, 主要使用类似 (1.15) 的 Cook 距离作为诊断统计量.

2. 局部影响分析

局部影响分析是由 Cook(1986) 基于似然函数提出的一种统计诊断方法, 它用于评价既定模型假定有微小变动对统计推断产生的局部影响. Cook(1986) 引入了扰动 (perturbation) 概念, 把异常点和强影响点归结为"比其他点受到更大扰动的