联合均值与方差模型

吴刘仓　徐登可　张忠占　著

科学出版社

联合均值与方差模型

吴刘仓　徐登可　张忠占　著

科学出版社
北京

内 容 简 介

本书系统介绍联合均值与方差模型及其拓展模型的理论、方法和应用. 内容主要包括：联合均值与方差模型的参数极大似然估计、变量选择、经验似然推断方法、缺失数据分析、基于频率和 Bayes 下统计诊断研究. 偏态(SN, StN)数据下联合位置与尺度模型和联合位置、尺度与偏度模型的参数极大似然估计、变量选择、缺失数据分析、统计诊断研究. 双重广义线性模型的经验似然推断、缺失数据分析、变量选择. 此外, 还介绍了这些模型的理论和方法在生物医学、产品的质量管理与控制、经济和金融学、产品设备的性能改进等领域的若干具体实际应用.

本书可作为统计学、生物医学、质量管理与控制、经济和金融中的风险管理、测量仪器或加工设备的精度提高或性能改进等相关专业研究生的教学参考书, 也可供相关专业的教师、科技人员和统计工作者参考.

图书在版编目(CIP)数据

联合均值与方差模型/吴刘仓, 徐登可, 张忠占著. —北京：科学出版社, 2019.1

ISBN 978-7-03-059218-7

Ⅰ. ①联… Ⅱ. ①吴… ②徐… ③张… Ⅲ. ①方差-统计模型 Ⅳ. ① C815

中国版本图书馆 CIP 数据核字(2018) 第 244266 号

责任编辑：李 欣 李香叶 / 责任校对：杨聪敏
责任印制：张 伟 / 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

北京九州逸驰传媒文化有限公司 印刷
科学出版社发行 各地新华书店经销
*
2019 年 1 月第 一 版 开本：720 × 1000 1/16
2019 年 1 月第一次印刷 印张：12 1/4
字数：247 000

定价：88.00 元

(如有印装质量问题, 我社负责调换)

前　言

联合均值与方差模型是 20 世纪 80 年代发展起来的一类重要的统计模型, 该模型既可对均值参数建模, 同时又可对方差参数建模, 相比单纯的均值回归模型具有更大的适应性, 可以概括和描述众多的实际问题. 在特别关注方差或波动的领域, 如产品的质量改进试验、经济和金融中风险管理、测量仪器或加工设备的精度提高等领域具有广泛的应用. 一方面, 该模型的一个特点是对方差的重视, 它能更好地解释数据变化的原因和规律, 这是数据分析中的一个重要的发展趋势. 这种思想也体现在质量管理方面, 比如日本田口学派的一个重要贡献是控制产品性能指标的方差. 控制产品性能指标的期望只表明平均来说产品性能指标合乎要求. 但若方差比较大, 则相当一部分产品仍然不合格, 因而控制方差的大小就与产品的合格率发生了紧密的联系. 另一方面, 为了研究影响方差的因素, 从而有效地控制方差, 有必要建立关于方差参数的模型. 联合均值与方差模型的实际背景主要来源于产品的质量改进试验, 典型的例子就是试验设计中的田口方法, 它是日本田口玄一 (Genichi Taguchi) 所创立的一种以低廉的成本实现高性能产品的稳健设计方法. 其基本观点是产品的质量高不仅表现在出厂时能让顾客满意, 而且在使用过程中给顾客和社会带来的损失要小. 用统计的语言描述就是, 使期望达到要求, 同时方差尽量小. 这便引出了均值和方差的同时建模问题. 用所建立的模型来选择使波动达到最小而均值达到要求的设计变量的实施条件. 然而, 为了更全面准确、更及时有效地分析复杂异方差数据, 本书从复杂数据和复杂模型的角度, 针对联合均值与方差模型建立了一套系统处理复杂异方差数据的统计推断方法, 重点研究了缺失数据、偏态数据等复杂数据下复杂联合均值与方差模型的估计理论、统计诊断、变量选择和经验似然推断方法及结合金融、经济、社会科学、气候科学、环境科学、工程技术和生物医学等学科中的一些实际复杂异方差数据作相关统计分析, 解释和分析这些学科中的复杂现象, 为这些学科的研究和发展提供了新的统计分析方法, 拓展和丰富了联合均值与方差模型的理论与方法.

我们希望本书的出版能引起回归分析、产品的质量管理与控制、经济和金融中风险管理、测量仪器或加工设备的精度提高等关注方差或波动领域方面的学者和实际使用者的兴趣. 特别地, 第 2—6 章的部分内容还可以继续深入研究, 希望有兴趣的读者通过本书的介绍能在相关领域进行进一步的研究工作. 全书共六章, 第 1 章主要介绍了联合均值与方差模型及其推广, 同时介绍了本书用到的变量选择方法、经验似然推断方法、统计诊断方法和缺失数据分析. 第 2 章研究了基于正

态数据下联合均值与方差模型的变量选择、响应变量随机缺失下的参数估计方法、经验似然推断方法、统计诊断方法. 第 3 章研究了基于偏正态数据和偏 t 正态数据下联合位置与尺度模型的变量选择. 特别地, 对偏正态数据下联合位置与尺度模型深入研究了统计诊断和响应变量随机缺失的参数估计方法. 第 4 章研究了基于 Box-Cox 变换下联合均值与方差模型的变量选择方法和变换参数的截面极大似然估计. 第 5 章研究了双重广义线性模型的经验似然推断和响应变量随机缺失下的参数估计方法以及 t 型双重广义线性模型的变量选择. 第 6 章分别研究了基于偏正态数据和偏 t 正态数据下联合位置、尺度与偏度模型的变量选择方法.

本书的出版得到了国家自然科学基金项目 (11861041, 11261025, 11771032, 11301485) 和昆明理工大学应用统计学科团队建设项目 (14078358), 浙江省自然科学基金项目 (LY17A010026) 经费的支持, 特此表示衷心感谢! 本书写作过程中, 自始至终得到科学出版社的关心与帮助, 特别要感谢李欣编辑, 她对本书的写作与出版都给予了大力的支持与帮助, 特此表示衷心感谢!

由于作者水平有限, 书中难免有不妥之处, 敬请同行专家、学者和广大读者批评指正.

吴刘仓　徐登可　张忠占

2018 年 8 月

目　录

第1章 绪 论

1.1 模 型

经典的回归模型中, 观测值的方差齐性是一个基本的假定. 在此假定下, 方可进行常规的统计推断. 然而在大多数社会经济现象和质量改进试验中, 存在大量的异方差数据, 所以这种假定不一定成立. 若方差非齐, 我们称为异方差. 处理异方差的方法常见的有两类. 第一类, 数据变换法, 如方差稳定化变换和经典的 Box–Cox 变换. 经过变换后转化为同方差处理. 第二类, 方差建模法. 不仅对均值而且也对方差建立统计模型, 称为异方差回归模型, 我们称为联合均值与方差模型. 因为, 一方面, 在许多应用领域, 特别在经济领域和工业产品的质量改进试验中, 非常有必要对方差建模, 以便更好地了解方差的来源, 达到有效控制方差. 例如, 田口玄一的稳健试验设计. 另一方面, 方差建模本身具有科学意义, 而且对有效估计和正确推断均值参数起到非常关键的作用 (Carroll, 1987; Carroll and Ruppert, 1988). 所以, 方差建模与均值建模具有同等重要的地位. 相比均值建模, 方差建模研究处于起步阶段.

1.1.1 线性回归模型

线性回归模型, 又称为线性模型, 是现代统计学中理论最丰富、应用最广泛的重要分支. 随着高速计算机的日益普及, 在生物、医学、经济、管理、农业、工业、工程技术等领域的应用获得了长足的发展 (Chatterjee and Hadi, 2006; Searle, 1971; Rao and Toutenburg, 1995; Wang and Chow, 1994; Christensen, 1987; 王松桂等, 2004).

正态线性模型的形式如下:

$$\begin{cases} y_i = x_i^{\mathrm{T}}\beta + \varepsilon_i, \\ \varepsilon_i \sim N(0, \sigma^2), \\ i = 1, 2, \cdots, n. \end{cases} \tag{1.1.1}$$

其中 $x_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$ 是 p 维解释变量, y_i 是其相应的响应变量, $\beta = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ 是 p 维未知回归参数, T 是转置.

由于正态线性回归模型的回归函数部分仅回归参数 β 是未知的, 因此若得到 β 的估计, 则自然也得到了回归函数的估计, 从而可以进行统计预测和决策. 在正态线

性回归模型下, 估计 β 的常用方法是极大似然估计法, 在观测样本 $(y_i, x_{i1}, x_{i2}, \cdots, x_{ip})$, $i = 1, \cdots, n$, β 的极大似然估计可以表达为

$$\hat{\beta} = \left(\sum_{i=1}^{n} x_i x_i^{\mathrm{T}} \right)^{-1} \sum_{i=1}^{n} x_i y_i.$$

线性回归模型的回归函数形式较为简单, 估计方便, 且由于该模型仅依赖于有限个回归参数, 因此当实际问题与假设模型较为接近时, 其统计推断往往具有较高的精度. 然而, 为准确、及时地分析来自各个领域的复杂现象, 一方面发展了大量有效的复杂模型, 比如: 非参数回归模型、半参数回归模型、变系数回归模型和部分线性变系数回归模型等. 但这些模型本质上都是对响应变量的均值建模, 把方差看作讨厌参数. 另一方面, 因为正态线性回归模型观测值的方差齐性是一个基本的假定, 在此假定下, 方可进行常规的统计推断. 然而在大多数社会经济现象中, 存在大量的异方差数据, 所以观测值的方差齐性这种假定有时并不切合实际. 而在许多应用领域, 特别在经济领域和工业产品的质量改进试验中, 非常有必要对方差建模, 以便更好地了解方差的来源, 达到有效控制方差. 此外, 方差建模本身具有科学意义, 而且对有效估计和正确推断均值参数起到非常关键的作用 (Carroll, 1987; Carroll and Ruppert, 1988). 所以, 方差建模与均值建模具有同等重要的地位. 近些年来, 同时对均值和方差建模引起了许多统计学者的研究兴趣. 下面介绍本书研究的模型.

1.1.2　联合均值与方差模型

Park(1966) 首次研究提出了联合均值与方差模型:

$$\begin{cases} y_i \sim N(\mu_i, \sigma_i^2), \\ \mu_i = x_i^{\mathrm{T}} \beta, \\ \log \sigma_i^2 = z_i^{\mathrm{T}} \gamma, \\ i = 1, 2, \cdots, n. \end{cases} \tag{1.1.2}$$

其中 $x_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$ 和 $z_i = (z_{i1}, \cdots, z_{iq})^{\mathrm{T}}$ 是解释变量, y_i 是其相应的响应变量, $\beta = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ 是 $p \times 1$ 的均值模型的未知参数向量, $\gamma = (\gamma_1, \cdots, \gamma_q)^{\mathrm{T}}$ 是 $q \times 1$ 的方差模型的未知参数向量. z_i 包含一些或者所有 x_i 和其他不在 x_i 的变量, 即均值模型和方差模型可能包含不同的解释变量或者相同的一些解释变量, 包含相同的解释变量但存在不同的影响方式. 注意 $x = (x_1, \cdots, x_n)^{\mathrm{T}}$ 和 $z = (z_1, \cdots, z_n)^{\mathrm{T}}$ 是解释变量矩阵. 当 $\sigma_i^2 = \sigma^2 (i = 1, 2, \cdots, n)$ 时, 该模型为正态线性回归模型 (1.1.1).

基于正态分布下联合均值与方差模型的研究, 最近这些年已经引起了许多统计学家的研究兴趣. Park(1966) 提出刻度参数的对数线性模型, 并给出了参数的二阶

段估计; Harvey(1976) 在一般条件下讨论了均值与方差效应的极大似然估计和似然比检验; Aitkin(1987) 给出了联合均值与方差模型的极大似然估计; Verbyla(1993) 将 Park 的对数刻度参数模型推广到更一般的函数形式, 给出了参数的极大似然估计和限制极大似然估计, 并讨论了异常点的诊断问题. Engel 和 Huele(1996) 应用联合均值与方差模型得到田口玄一的稳健试验设计; Taylor 和 Verbyla(2004) 针对异常点数据研究提出了基于 t 分布下联合位置与尺度模型, 并研究了该模型参数的估计和检验问题.

第 2 章系统研究了这个模型的变量选择、经验似然推断、响应变量随机缺失下的参数估计、基于频率和 Bayes 下的统计诊断等问题, 而且本书进一步推广该模型, 考虑了缺失数据、偏态数据、尖锋厚尾数据和异常点数据的联合建模的情形, 并研究了其相应的统计推断.

1. 基于偏正态分布下联合位置与尺度模型

在实际问题中, 如金融、经济、社会科学、气候科学、环境科学、工程技术和生物医学等领域, 经常遇到研究的数量关系的响应变量具有非对称性的情形. 常伴有尖锋厚尾特征, 而且存在大量的异方差数据, 人们还非常关注方差的变化, 了解方差的来源, 所以非常有必要对方差建模, 以便更好地了解数据波动的统计变化规律.

针对偏正态 (SN) 数据, 我们研究提出如下感兴趣的基于偏正态分布下联合位置与尺度模型:

$$\begin{cases} y_i \sim \mathrm{SN}(\mu_i, \sigma_i^2, \lambda), \\ \mu_i = x_i^{\mathrm{T}}\beta, \\ \log \sigma_i^2 = z_i^{\mathrm{T}}\gamma, \\ i = 1, 2, \cdots, n. \end{cases} \tag{1.1.3}$$

其中 μ_i 是位置 (location) 参数, σ_i 是尺度 (scale) 参数, λ 是偏度 (skewness) 参数, $x_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$ 和 $z_i = (z_{i1}, \cdots, z_{iq})^{\mathrm{T}}$ 是解释变量, y_i 是其相应的响应变量, $\beta = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ 是 $p \times 1$ 的位置模型的未知参数向量, $\gamma = (\gamma_1, \cdots, \gamma_q)^{\mathrm{T}}$ 是 $q \times 1$ 的尺度模型的未知参数向量. z_i 包含一些或者所有 x_i 和其他不在 x_i 的变量, 即位置模型和尺度模型可能包含不同的解释变量或者相同的一些解释变量, 包含相同的解释变量但存在不同的影响方式.

(1) 当 $\lambda = 0$ 时, 该模型为基于正态分布下联合均值与方差模型 (1.1.2).

(2) 当 $\sigma_i^2 = \sigma^2 (i = 1, 2, \cdots, n)$ 时, 该模型为偏正态线性回归模型.

若 $y \sim \mathrm{SN}(\mu, \sigma^2, \lambda)$, 则有

$$E(y) = \mu + \sqrt{\frac{2}{\pi}} \frac{\lambda}{\sqrt{1+\lambda^2}} \sigma,$$

$$\mathrm{Var}(y)=\left(1-\frac{2\lambda^2}{\pi(1+\lambda^2)}\right)\sigma^2.$$

第 3 章分别研究了这个模型的变量选择、统计诊断和响应变量缺失下参数估计等问题, 详见 3.1—3.3 节.

2. 基于 t 分布下联合位置与尺度模型

近年来, 随机误差为 t 的回归 (简称 t 回归) 越来越受到理论与实际应用工作者的关注, 因为对不少具有重尾 (heavy-tail) 的分布数据 (诸如经济、金融中的数据), t 回归的估计比较稳健, 拟合效果比正态回归更好 (Lange et al., 1989; Lin et al., 2009; Barroso and Cordeiro, 2005). 类似于联合均值与方差模型, 在金融、经济等实际问题中, 存在大量的异方差数据, 人们还非常关注方差的变化, 了解方差的来源, 所以非常有必要对方差建模, 以便更好地了解数据波动的统计变化规律. Taylor 和 Verbyla (2004) 针对异常点数据研究提出了如下感兴趣的基于 t 分布下联合位置与尺度模型:

$$\begin{cases} y_i\sim t(\mu_i,\sigma_i^2,\nu), \\ \mu_i=x_i^{\mathrm{T}}\beta, \\ \log\sigma_i^2=z_i^{\mathrm{T}}\gamma, \\ i=1,2,\cdots,n. \end{cases} \tag{1.1.4}$$

其中 μ_i 是位置参数, σ_i 是尺度参数, ν 是自由度, $x_i=(x_{i1},\cdots,x_{ip})^{\mathrm{T}}$ 和 $z_i=(z_{i1},\cdots,z_{iq})^{\mathrm{T}}$ 是解释变量, y_i 是其相应的响应变量, $\beta=(\beta_1,\cdots,\beta_p)^{\mathrm{T}}$ 是 $p\times 1$ 的位置模型的未知参数向量, $\gamma=(\gamma_1,\cdots,\gamma_q)^{\mathrm{T}}$ 是 $q\times 1$ 的尺度模型的未知参数向量. z_i 包含一些或者所有 x_i 和其他不在 x_i 的变量, 即位置模型和尺度模型可能包含不同的解释变量或者相同的一些解释变量, 包含相同的解释变量但存在不同的影响方式.

(1) 当 $\nu\to\infty$ 时, 该模型就成为基于正态分布联合均值与方差模型 (1.1.2).

(2) 当 $\sigma_i^2=\sigma^2(i=1,2,\cdots,n)$ 时, 该模型为 t 回归模型.

若 $y\sim t(\mu,\sigma^2,\nu)$, 则有 $E(y)=\mu$, $\mathrm{Var}(y)=\dfrac{\nu}{\nu-2}\sigma^2$.

3. 基于偏 t 正态 (StN) 分布下联合位置与尺度模型

在实际问题中, 如金融、经济、社会科学、气候科学、环境科学、工程技术和生物医学等领域, 经常遇到研究的数量关系的响应变量具有非对称性的情形, 常伴有尖锋重尾特征, 而且存在大量的异方差数据, 人们还非常关注方差的变化, 了解方差的来源, 所以非常有必要对方差建模, 以便更好地了解数据波动的统计变化规律. 而就像正态分布和 t 分布的差异一样, 偏 t 正态分布比偏正态分布尾部要重.

针对稳健性和数据的偏斜性, 我们研究提出如下感兴趣的基于偏 t 正态分布下联合位置与尺度模型:

$$\begin{cases} y_i \sim \mathrm{StN}(\mu_i, \sigma_i^2, \lambda, \nu), \\ \mu_i = x_i^{\mathrm{T}}\beta, \\ \log \sigma_i^2 = z_i^{\mathrm{T}}\gamma, \\ i = 1, 2, \cdots, n. \end{cases} \tag{1.1.5}$$

其中 μ_i 是位置参数, σ_i 是尺度参数, λ 是偏度参数, ν 是自由度, $x_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$ 和 $z_i = (z_{i1}, \cdots, z_{iq})^{\mathrm{T}}$ 是解释变量, y_i 是其相应的响应变量, $\beta = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ 是 $p \times 1$ 的位置模型的未知参数向量, $\gamma = (\gamma_1, \cdots, \gamma_q)^{\mathrm{T}}$ 是 $q \times 1$ 的尺度模型的未知参数向量. z_i 包含一些或者所有 x_i 和其他不在 x_i 的变量, 即位置模型和尺度模型可能包含不同的解释变量或者相同的一些解释变量, 包含相同的解释变量但存在不同的影响方式.

(1) 当 $\nu \to \infty$ 时, 该模型就成为基于偏正态分布下联合位置与尺度模型 (1.1.3).

(2) 当 $\lambda = 0$ 时, 该模型为基于 t 分布下联合位置与尺度模型 (1.1.4).

(3) 当 $\sigma_i^2 = \sigma^2 (i = 1, 2, \cdots, n)$ 时, 该模型为偏 t 正态线性回归模型.

若 $y \sim \mathrm{StN}(\mu, \sigma^2, \lambda, \nu)$, 则有

$$E(y) = \mu + \sqrt{\frac{2}{\pi}}\sigma\lambda \left(\frac{\nu}{2}\right)^{\frac{1}{2}} \frac{\Gamma\left(\dfrac{\nu-1}{2}\right)}{\Gamma\left(\dfrac{\nu}{2}\right)} E_V (V + \lambda^2)^{-\frac{1}{2}},$$

$$\mathrm{Var}(y) = \sigma^2 \left\{ \frac{\nu}{\nu-2} - \frac{\lambda^2\nu}{\pi} \left[\frac{\Gamma\left(\dfrac{\nu-1}{2}\right)}{\Gamma\left(\dfrac{\nu}{2}\right)} \right]^2 [E_V (V + \lambda^2)^{-\frac{1}{2}}]^2 \right\},$$

其中 $V \sim \mathrm{Gamma}\left(\dfrac{\nu-1}{2}, \dfrac{\nu}{2}\right)$.

第 3 章研究了这个模型的变量选择问题, 详见 3.4 节.

4. 基于 Box-Cox 变换下联合均值与方差模型

对于有偏斜的数据, 当分析数量关系时, 为了应用正态线性回归模型, 最常见的方法之一是对响应变量观察值 y 进行一个数据变换 $f(y)$, 使其同时满足线性性、方差齐性和正态性三个条件. 下面的 Box–Cox 变换 (Box and Cox, 1964) 是其中最著名的数据变换之一.

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \ln y, & \lambda = 0. \end{cases} \tag{1.1.6}$$

Box–Cox 变换通过引进一个新的变换参数 λ, 通过数据本身来决定如何进行变换, 即变换参数 λ 的自适应 (adaptive) 估计 (Atkinson, 1982; Cook and Weisberg, 1982).

然而, 在一定的条件下, 这些假定条件是非常不适合的. 仅仅要求一个变换就使其同时满足线性性、方差齐性和正态性大多数情况是做不到的 (Nelder and Lee, 1991). 例如, $y_i \sim \text{Poisson}(\mu_i), i = 1, 2, \cdots, n$, 为满足线性性需要作变换 $\ln y$, 为满足方差齐性需要作变换 $y^{1/2}$, 为满足正态性需要作变换 $y^{2/3}$. 特别变换后存在大量异方差的情形, 如 Carroll 和 Ruppert(1988). 然而, 当 Box–Cox 变换的同方差假定不成立时, 统计推断将遇到诸多问题, 因为未知参数有 $n+p$ 个, 参数估计的问题变得较为复杂, 甚至是不能识别的 (韦博成等, 2003). 为了克服上面的缺陷, 我们研究提出如下感兴趣的基于 Box–Cox 变换下联合均值与方差模型

$$\begin{cases} y_i^{(\lambda)} \sim N(\mu_i, \sigma_i^2), \\ \mu_i = x_i^{\mathrm{T}}\beta, \\ \log \sigma_i^2 = z_i^{\mathrm{T}}\gamma, \\ i = 1, 2, \cdots, n, \end{cases} \tag{1.1.7}$$

其中 $x_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$ 和 $z_i = (z_{i1}, \cdots, z_{iq})^{\mathrm{T}}$ 是解释变量, y_i 是其相应的响应变量, $\beta = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ 是 $p \times 1$ 的均值模型的未知参数向量, $\gamma = (\gamma_1, \cdots, \gamma_q)^{\mathrm{T}}$ 是 $q \times 1$ 的方差模型的未知参数向量. z_i 包含一些或者所有 x_i 和其他不在 x_i 的变量, 即均值模型和方差模型可能包含不同的解释变量或者相同的一些解释变量, 包含相同的解释变量但存在不同的影响方式.

(1) 当 $\lambda = 1$ 时, 该模型就成为基于正态分布下联合均值与方差模型 (1.1.2).

(2) 当 $\sigma_i^2 = \sigma^2 (i = 1, 2, \cdots, n)$ 时, 该模型为数据变换模型.

第 4 章研究了这个模型的变量选择和变换参数 λ 的估计问题.

1.1.3　双重广义线性模型

广义线性回归模型是经典线性回归模型极其重要的发展与推广. 目前, 对均值的广义线性回归模型, 已有大量的文献提出了许多有效和灵活的方法. 然而, 在许多应用方面, 特别在经济领域和工业产品的质量改进试验中, 非常有必要对散度建模, 了解方差的来源, 以便有效控制方差. 典型的例子之一就是试验设计中的田口方法, 它是日本田口玄一所创立的一种以廉价的成本实现高性能产品的稳健设计方法, 其基本观点就是产品质量高不仅表现在出厂时能让顾客满意, 而且在使用过程中给顾客和社会带来的损失要小. 用统计语言描述就是, 使期望达到要求, 同时方差尽量小. 这便引出了均值和散度的同时建模问题. 用所建立的模型来选择使得波动达到最小而均值达到要求的设计变量的实施条件 (王大荣, 2009; Lee Nelder, 2006). 另一方面, 散度建模本身具有科学意义. 而且对有效估计和正确推断均值参

数起到非常关键的作用 (Carroll, 1987; Carroll and Ruppert, 1988). 所以, 散度建模与均值建模同等重要. 相比均值建模, 散度建模研究处于起步阶段.

1. 双重广义线性模型

Pregibon(1984) 在一篇综述文章中首次提出了对散度参数建模的方法, 即考虑下面的双重广义线性模型

$$\begin{cases} \mathrm{Var}(y_i) = \phi_i V(\mu_i), \\ g(\mu_i) = x_i^{\mathrm{T}}\beta, \\ h(\phi_i) = z_i^{\mathrm{T}}\gamma, \\ i = 1, 2, \cdots, n. \end{cases} \tag{1.1.8}$$

其中 ϕ_i 是散度参数, $V(\mu_i)$ 是方差函数, $x_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$ 和 $z_i = (z_{i1}, \cdots, z_{iq})^{\mathrm{T}}$ 是解释变量, y_i 是其相应的响应变量, $\beta = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ 是 $p \times 1$ 的均值模型的未知参数向量, $\gamma = (\gamma_1, \cdots, \gamma_q)^{\mathrm{T}}$ 是 $q \times 1$ 的散度模型的未知参数向量. z_i 包含一些或者所有 x_i 和其他不在 x_i 的变量, 即均值模型和散度模型可能包含不同的解释变量或者相同的一些解释变量, 包含相同的解释变量但存在不同的影响方式. $g(\cdot), h(\cdot)$ 分别是均值与散度的联系函数, 而且要求 $h \geqslant 0$, g^{-1}, h^{-1} 存在且 $h'(\cdot) \neq 0$. 当 $\phi_i = \phi (i = 1, 2, \cdots, n)$ 时, 该模型为广义线性模型.

可以看到, y 的方差 $\mathrm{Var}(y) = \phi V(\mu)$ 由两部分的乘积构成: 一部分是与均值 μ 无关的散度参数 ϕ, 另一部分是均值的函数 $V(\mu)$, 称为方差函数. 方差函数 $V(\mu)$ 的选择依赖于所使用的分布. 例如, 常见的 Tweedie 分布, $V(\mu) = \mu^p$; 特别地, 对于正态分布, $V(\mu) = 1$; 对于对数正态分布和 Gamma 分布, $V(\mu) = \mu^2$; 对于拟高斯 (IG) 分布, $V(\mu) = \mu^3$; 对于 Poisson 分布, $V(\mu) = \mu$.

Smyth(1989) 称上述模型 (1.1.8) 为双重广义线性模型 (double generalized linear model, DGLM), Lee 和 Nelder(2006) 称之为联合广义线性模型 (joint generalized linear model, JGLM). 该模型在工业产品的质量改进试验中得到了广泛的应用.

双重广义线性模型已引起了许多统计学者的研究兴趣, 研究方法大概分为两类. 第一类, 不假定分布的情况下, 只需假定前二阶矩的存在. 主要是利用基于扩展拟似然 (extended quasi-likelihood, EQL)(Nelder and Pregibon, 1987) 和伪似然 (pseudo-likelihood, PL)(Engel and Huele, 1996) 两类推广的似然方法. 第二类, 假定双指数分布族 (double exponential family, DEF)(Efron, 1986). 基于 DEF 下双重广义线性模型的统计推断 (Galfand and Dalal, 1990; Dey et al., 1997; Gijbels, 2010). 关于该模型的参数估计问题, Smyth(1989) 给出了参数的极大似然估计; Nelder 和 Lee(1991, 1998) 利用扩展拟似然函数, 在分布前二阶矩的条件下, 给出了最大扩展拟似然估计 (MEQL); Smyth 和 Verbyla(1999, 2009), Smyth 等 (2001), Smyth(2002) 系统地研究了参数 ϕ 的 REML 类型的估计.

另一方面, 双重广义线性模型的思想与随机效应的思想相结合, 从而引出了许多新的模型和统计分析方法. Lee 和 Nelder(1996) 通过扩大随机效应的分布类, 提出了分层广义线性模型 (hierarchical generalized linear model, HGLM), 并对似然函数进行了扩展, 提出了 h-似然函数, 避免了计算高维积分. Lee 和 Nelder(2006) 通过对 HGLM 中的散度参数建模, 并将 HGLM 的思想融入散度参数模型中, 构造了双重分层广义线性模型 (double hierarchical generalized linear model, DHGLM), 该模型类包括了经济领域中的 ARCH 模型、GARCH 模型和 SV 模型.

第 5 章研究了双重广义线性模型 (1.1.8) 的经验似然推断和响应变量随机缺失下的参数估计方法, 详见 5.1 节、5.2 节.

2. t 型双重广义线性模型

双重广义线性模型已引起了许多统计学者的研究兴趣. 然而, 双重广义线性模型基于 EQL、PL 和 DEF 的估计方法, 受异常点数据的影响非常大, 所以非常有必要发展一种稳健的估计方法. 本书基于稳健的角度, 推广双重广义线性模型 (1.1.8), 研究提出了一类新的双重广义线性模型, 我们称为 t 型双重广义线性模型, 模型如下:

$$\begin{cases} \mathrm{Var}(y_i) = \dfrac{\nu}{\nu-2}\phi_i V(\mu_i), \\ g(\mu_i) = x_i^{\mathrm{T}}\beta, \\ h(\phi_i) = z_i^{\mathrm{T}}\gamma, \\ i = 1, 2, \cdots, n. \end{cases} \tag{1.1.9}$$

其中 $\nu > 0$ 是自由度, ϕ_i 是散度参数, $V(\mu_i)$ 是方差函数, $x_i = (x_{i1}, \cdots, x_{ip})^{\mathrm{T}}$ 和 $z_i = (z_{i1}, \cdots, z_{iq})^{\mathrm{T}}$ 是解释变量, y_i 是其相应的响应变量, $\beta = (\beta_1, \cdots, \beta_p)^{\mathrm{T}}$ 是 $p \times 1$ 的均值模型的未知参数向量, $\gamma = (\gamma_1, \cdots, \gamma_q)^{\mathrm{T}}$ 是 $q \times 1$ 的散度模型的未知参数向量. z_i 包含一些或者所有 x_i 和其他不在 x_i 的变量, 即均值模型和散度模型可能包含不同的解释变量或者相同的一些解释变量, 包含相同的解释变量但存在不同的影响方式. $g(\cdot), h(\cdot)$ 分别是均值与散度的联系函数, 而且要求 $h \geqslant 0$, g^{-1}, h^{-1} 存在且 $h'(\cdot) \neq 0$.

(1) 当 $\nu \to \infty$ 时, 该模型就是上述的双重广义线性模型 (1.1.8).

(2) 当 $\nu \to \infty$ 且 $\phi_i = \phi (i = 1, 2, \cdots, n)$ 时, 该模型为广义线性模型.

第 5 章研究了这个模型的变量选择等问题, 详见 5.2 节.

1.1.4　联合位置、尺度与偏度模型

在诸如金融、经济、社会科学、气候科学、环境科学、工程技术和生物医学等领域, 研究的数据大多不严格服从正态分布或 t 分布等对称分布, 而是有一定的偏