• 国家自科基金项目“风险信息共享背景下的个体风险评估研究”（71303045）成果
• 对外经济贸易大学学术创新团队（CXTD5-04）成果

广义GAMMA分布簇广义线性混合模型理论与应用

GUANGYI GAMMA FENBUCU
GUANGYI XIANXING HUNHE MOXING
LILUN YU YINGYONG

谢远涛 杨娟 著

对外经济贸易大学出版社
University of International Business and Economics Press

国家自科基金项目“风险信息共享背景下的个体风险评估研究”(71303045)成果
对外经济贸易大学学术创新团队(CXTD5－04)成果

广义 GAMMA 分布簇广义线性混合模型理论与应用

谢远涛　杨　娟　著

对外经济贸易大学出版社
中国·北京

图书在版编目（CIP）数据

广义 GAMMA 分布簇广义线性混合模型理论与应用 / 谢远涛，杨娟著. —北京：对外经济贸易大学出版社，2014

ISBN 978-7-5663-1008-8

Ⅰ. ①广… Ⅱ. ①谢… ②杨… Ⅲ. ①线性模型－研究 Ⅳ. ①O212

中国版本图书馆 CIP 数据核字（2014）第 079255 号

广义 GAMMA 分布簇广义线性混合模型理论与应用

谢远涛　杨　娟　著

责任编辑：赵　昕　王　京

对外经济贸易大学出版社

北京市朝阳区惠新东街 10 号　　邮政编码：100029

邮购电话：010－64492338　　发行部电话：010－64492342

网址：http://www.uibep.com　　E-mail：uibep@126.com

北京市山华苑印刷有限责任公司印装　　新华书店北京发行所发行

成品尺寸：170mm×240mm　　10.75 印张　　170 千字

2014 年 6 月北京第 1 版　　2014 年 6 月第 1 次印刷

ISBN 978-7-5663-1008-8

定价：43.00 元

中 文 摘 要

通过拿掉经典线性回归模型的假定，可以得到越来越接近现实和越来越能刻画真实数据的模型。广义线性混合模型把随机效应模型与重复观测效应模型统一起来了，还可以在更广义的指数分布簇框架下对响应变量进行建模。

然而遗憾的是，仅仅指数分布簇还是远远不够的，现实数据需要有更多灵活性的分布模型。在生存分析和保险精算中，应用很广的往往是广义 Gamma 分布簇。作为广义 Gamma 分布特例，Weibull 分布和在刻画损失风险、生存时间等变量的应用上最广泛，同时该分布还能把对数正态模型作为其极限分布来研究，可以轻易变换得到逆幂变换 Gamma 分布簇，进而把逆 Gamma 分布和逆 Weibull 分布作为特例包含进来。因此，本书在传统广义线性混合模型的框架下分析满足广义 Gamma 分布簇的响应变量的建模问题。

广义 Gamma 分布簇与指数分布簇不同，但是有联系，通过参数重整可以构造出指数分布簇模型中常用的拟似然函数，利用连接函数可以基于广义 Gamma 分布簇构造特殊的广义线性混合模型。一方面，可以充分利用广义线性模型和广义线性混合模型的研究成果，使估计推断有良好的理论支持并且容易实现；另一方面，使广义 Gamma 分布簇广义线性混合模型能够把随机常数项模型、变系数模型、重复观测模型和空间相关模型作为特例包含起来。

本书的估计推断部分通过横向比较六种常用的估计方法，找出这些方法的内在联系，选取一个可以把所有方法作为其特例的方法，使这些方法在估计上具有一致的形式；根据模型的具体特点对新引入的参数设计算法并编程实现，以确保参数估计具有良好的性质，如渐进正态性，基于此可

以进一步讨论假设检验问题。

本书按照得分检验的基本思路构建广义 Gamma 分布簇线性混合模型的假设检验，通过模型误设检验来实现参数的收缩，将三参数广义 Gamma 分布收缩到两参数的 Gamma 分布、Weibull 分布或指数分布，再到单参数指数分布，可以把指数分布、Weibull 分布和 Gamma 分布情况作为特例来分析，有效解决复杂性与准确性的折中。

在其他进阶讨论部分，本书还讨论了负响应变量观测值的处理以及模型结构的调整；以两类比例危险模型为例讨论广义 Gamma 分布簇广义线性混合模型如何引入半参数模型；然后讨论了广义 Gamma 分布簇广义线性混合模型与神经网络模型的联系；最后讨论了删失、打结时偏似然函数、伪似然或者拟似然的调整问题，以及联合多变量混合模型的构建问题。

第 6 章给出一个实际数据分析案例，研究如何利用广义 Gamma 分布簇广义线性混合模型绘制恢复曲线。

第 7 章为该模型在精算科学中的扩展与应用。在第 7.1 节中，我们考虑把分类费率厘定技术（广义线性模型、广义线性混合模型，也可以推广到广义 Gamma 分布簇广义线性混合模型）与个体经验费率厘定技术（信度模型）进行整合。第 7.2 节我们讨论另一类分类费率厘定技术和个体经验费率厘定技术整合的方案。第 7.3 节我们讨论对准备金计提技术的改进，以实现动态监管。

关键词：广义 Gamma 分布簇；广义线性混合模型；多变量生存分析；参数重整；受限极大（伪）似然估计；多水平模型；神经网络

Abstract

This article try to give a system research on Generalized Linear Mixed Models based on Generalized Gamma Distribution.

Generalized Gamma Distribution Family is different from Exponential Distribution Family, as well as some connection. Parameters rearrangement will do well in deriving quasi-likelihood function. Generalized Gamma Distribution Family models can be embedded into Generalized Linear Mixed Models by way of special link function. Estimation can be easily achieved with the theory support of Generalized Linear Mixed Models; some other models such as random intercept models, variable-coefficient models, repeated models, spatial models and so on can easily be derived from this Generalized Linear Mixed Models based on Generalized Gamma Distribution.

More than six methods are compared when it comes to the estimation. This paper tries to get the uniform for all these six methods and then put it into SAS coding. The reason for classical estimation methods is to gain the good characters which are very important in hypothesis test.

This article also proposes the score tests for Generalized Linear Mixed Models based on Generalized Gamma Distribution for model misspecification test on Gamma distribution, Weibull distribution, and exponential distribution and so on. This is the tradeoff between complication and accuracy. In the 5th chapter, this article discusses some advanced issue such as negative observation, semi-parameter models, artificial neural networks models, censored data, tied data and the unit modeling for multi heterogeneity variables.

This article gives a real application with Generalized Linear Mixed Models

based on Generalized Gamma Distribution in the final part. This paper proposed a analysis model for Generalized Gamma distribution data based on Generalized Linear Mixed Models. This paper estimates the parameters and plots the recovery curves with the dynamic investigation data. The recovery curve can explore the recovery of burned children visually and comprehensively. With this model it is easy to give a comprehensive evaluation on recovery from different aspects such as physics, social psychology and so on. The method proposed in this paper is also available in comprehensive evaluation which are generally used in social research.

In chapter 7 we discuss some extend models with important application in actuarial science. In section 7. 1 we try to construct a credibility model based on generalized linear mixed models. Suppose dependent variable follows exponential distribution family, natural parameters follows natural conjugate prior. This article introduces credibility factor into generalized linear mixed models to build a model with both priority information and posterior information included according to Bayesian theory, and concludes that the estimate of random effect coincide with the credibility formula. This article finds out the relationship between prior parameters by way of the limit of natural conjugate prior. Estimation can be achieved by way of pseudo-likelihood method. Bühlmann-Straub credibility and Bühlmann credibility can be treated as a special case of this model by transforming the Tweedie models which is discussed in this paper. This paper may contribute to rate making and credibility. In section 7. 2 we discuss some other ways to combine the linear mixed models with credibility analysis. In section 7. 3 we use operational time to redesign the runoff triangle, and construct double generalized linear mixed models and Tweedie based generalized linear mixed models by way of improving Horel curve to describe heterogeneous claim developments factors. These models are available for dynamic risk measure and management.

Key Words: Generalized Linear Mixed Models; multivariate survival analysis; Generalized Gamma Distribution Family; partial maximum likelihood estimate; multi-level models; neural networks

目　录

第 1 章　引言 / 1

1.1　背景与模型综述 / 1

1.2　文献综述 / 5

1.3　本书意义 / 11

1.4　内容简介 / 13

第 2 章　广义 Gamma 分布簇广义线性混合模型的构建 / 19

2.1　广义 Gamma 分布及其与指数分布簇的关系 / 19

2.2　似然函数和连接函数的构建 / 21

2.3　广义 Gamma 分布簇广义线性混合模型的构建 / 25

第 3 章　参数估计 / 31

3.1　似然函数的构建 / 31

3.2　参数估计 / 36

3.3　编程实现 / 54

第 4 章　其他模型推断 / 59

4.1　预测函数及方差估计 / 59

4.2　偏差及拟合优度检验 / 60

4.3　Wald 统计量和 F 统计量 / 62

4.4　固定效应的检验 / 63

4.5　随机效应的检验 / 64

第5章　其他进阶讨论 / 67

5.1　负响应变量观测值问题 / 67
5.2　广义 Gamma 分布簇线性混合模型得分检验 / 69
5.3　Gamma 分布线性混合模型的得分检验 / 71
5.4　Weibull 分布线性混合模型的得分检验 / 72
5.5　指数分布线性混合模型的得分检验 / 73
5.6　半参数广义 Gamma 分布簇广义线性混合模型 / 75
5.7　与数据挖掘工具的联系 / 77
5.8　删失与打结问题 / 81
5.9　模型推广 / 82

第6章　实例分析 / 85

6.1　数据简要介绍 / 85
6.2　模型拟合结果 / 90
6.3　得分检验 / 90
6.4　恢复曲线 / 91
6.5　总结 / 98

第7章　在精算科学中的扩展与应用 / 99

7.1　广义线性混合模型框架下的信度模型分析 / 99
7.2　一般线性混合模型下的信度分析扩展 / 109
7.3　基于操作时间和广义线性混合模型的准备金评估技术研究 / 120

附录 / 133

参考文献 / 137

致谢 / 157

后记 / 158

图表索引

图 1.1　形状参数为 0.5 的 Gamma 分布图 / 3
图 1.2　形状参数为 1 的 Gamma 分布图 / 4
图 1.3　形状参数为 3 的 Gamma 分布图 / 4
图 3.1　Nelder and Mead 直接搜索算法 / 50
图 3.2　非对称二分搜索算法 / 52
图 3.3　非对称有记忆二分搜索算法 / 54
图 5.1　人工神经网络结构 / 78
图 5.2　人工神经网络 BP 算法迭代计算流程 / 81
图 6.1　Emotions、physport、upperfx、school、comply 尺度的信度 / 87
图 6.2　Concernp、family、pain、appearance、satisfy 尺度的信度 / 87
图 6.3　Appearance 尺度的恢复曲线 / 92
图 6.4　Compliance 尺度的恢复曲线 / 92
图 6.5　Emotional health 尺度的恢复曲线 / 93
图 6.6　Family disruption 尺度的恢复曲线 / 93
图 6.7　Itch 尺度的恢复曲线 / 94
图 6.8　Pain 尺度的恢复曲线 / 94
图 6.9　Parental concern 尺度的恢复曲线 / 95
图 6.10　Physical function and sport 尺度的恢复曲线 / 95
图 6.11　Satisfaction 尺度的恢复曲线 / 96
图 6.12　School reentry 尺度的恢复曲线 / 96
图 6.13　Transfers and mobility 尺度的恢复曲线 / 97
图 6.14　Upper extremity function 尺度的恢复曲线 / 97
表 3.1　广义 Gamma 分布簇广义线性混合模型编程算法 1 / 55

表 3.2　广义 Gamma 分布簇广义线性混合模型编程算法 2 / 57
表 6.1　测量 - 再测量信度分析 / 86
表 6.2　烧伤儿童的分布状况 / 88
表 6.3　烧伤儿童康复状况（父母角度）/ 88
表 6.4　烧伤儿童康复状况（儿童角度）/ 88
表 6.5　烧伤儿童康复状况（父母儿童评价之间的差异及检验）/ 89
表 6.6　分布参数估计 / 90
表 6.7　联合得分检验 / 91
表 7.1　固定效应和随机效应估计值分解结果 / 119
表 7.2　回归结果 / 128

第1章
引　言

本章对本书写作的背景进行介绍，对模型发展进行综述，然后是文献综述，对国内外研究现状进行概述；第1.3节指出本书的实际意义和理论意义；最后介绍了本书的写作内容、结构安排和创新点。

1.1　背景与模型综述

在经典回归模型中，响应变量要求满足独立同正态分布，解释变量非随机，并与扰动项无关，模型为线性关系，等等。一般线性模型允许解释变量可以取离散变量（统一了哑变量模型），并把 Type Ⅰ、Type Ⅱ、Type Ⅲ、Type Ⅳ方差分析拿进来搭建了经典线性模型框架。随着统计理论的发展，人们对数据的认识不断深化，逐渐认识到经典线性模型中对响应变量的假设过于苛刻而与现实不符合。

首先是响应变量正态性的假定。一个最常见的例子是 logistic 模型研究的概率与频率问题（logistic 模型对于优势以及优势比的刻画）。另外一种情况是对逻辑变量进行回归，如天气预报中是否下雪、疾病是否会复发、维生素 A 缺乏的孩子是否容易感染某疾病等问题。还有一种情况是计数数据回归，如保险精算中的索赔次数模型（Poisson 回归）以及高维列联表分析（多项分布），这些常常采用多项分布对数线性模型和 Poisson 对数线性模型来实现。事实上这些研究是远远不够的，以 Poisson 回归为例，Poisson 分布要求均值等于方差，但是实际计数数据回归的结果往往是方差大于均值的，这种现象我们称为过度离散（over dispersion），于是有些学者转而使用

负二项分布回归。除此之外还有很多对逆高斯分布、指数分布、Gamma 分布进行回归的情况。广义线性模型（Nelder and Wedderburn，1972）最终统一了这些模型而构建了一套针对指数分布簇响应变量的完整模型框架。广义线性模型的概念最先在 1972 年 Nelder 和 Wedderburn 一篇论文中引进，1983 年 McCullagh 和 Nelder 出版了系统论述此专题的专著并于 1989 年再版，Fisher 在 1919 年曾用过它的名字。作为其特例的最重要的 Logistic 模型，在 20 世纪四五十年代曾由 Berkson，Dyke 和 Patterson 等人使用过。

现在再看独立同分布假定。假定各个不同的个体间彼此相互独立并且满足同一分布在有些情况下未必成立，其一是随机效应，当数据具有系统结构的特征时，如数据来自于不同的群体（层次或水平），数据常存在群体内的相似性或聚集性（Cluster）；其二是重复观测效应，如某一个体被重复观测多次。简言之，响应变量观测值之间出现了异质性或者非齐次性。对于这类分层数据建模，常常有边际模型、随机效应模型、转移模型之分，响应变量之间的相关以及不同分布的特征都可以放到随机效应模型或者重复观测效应模型中研究，于是混合模型最终统一了这个框架。

要拿掉“解释变量非随机，并与扰动项无关”的假定，可以通过构造特殊的加权矩阵或者工具变量来实现，而从线性模型到非线性模型的推广只需要构造特殊的函数就可以实现。一些方法或者模型可以很好地实现非线性化，如光滑样条法或者神经网络模型，本书暂不考虑这些情况。

综合前面分析可知，通过拿掉经典线性回归模型的假定，我们得到了越来越接近现实和越来越能刻画真实数据的模型。模型推广的线路可以分为两条，一条思路是拿掉正态性假定发展广义线性模型，另外一条思路是拿掉同分布假定和独立性假定发展线性混合模型。这看上去是两条路，但是这两条路彼此之间又有千丝万缕的联系：一方面在广义线性模型中，我们可以考虑添加随机效应项来考虑异质性效应，还可以添加扰动项结构来解释响应变量之间的相关；另外一方面，类比从经典线性模型中正态性到广义线性模型中指数分布簇的推广，完全可以把混合模型中的正态分布假定推广到指数分布簇情形。于是这两条线路又最终合并到一个模型，也即广义线性混合模型（Breslow and Clayton，1989）。

广义线性混合模型是一个集大成的模型，一方面它允许响应变量之间存在异质性问题，把随机效应模型与重复观测效应模型统一起来了；另外

一方面还可以在更广义的指数分布簇框架下对响应变量进行建模。经典的广义线性模型或者广义线性混合模型研究的指数分布簇主要包括正态分布、逆高斯分布、二项分布、负二项分布、多项分布、Poisson 分布和 Gamma 分布共七大分布，成功地解决了计数数据回归等问题。

诚然，这些分布都是非常常见的分布，然而这些分布的灵活性并不是很好，这些模型最多有两个参数，一个是尺度参数（或散度参数），另一个是位置参数（或自然参数）。当然，实际建模中往往是取这些参数的特定函数或者连接函数（link function）形式。在广义线性混合模型中常常使用 Breslow-Clayton 方法，其基本思想是沿用广义线性模型的思路把似然函数推广为拟似然函数而使用罚拟似然（penalized quasi-likelihood，PQL）方法，一个特点是假定 φ 参数取值为 1，这一假定很大程度上简化了参数估计，但从另外一方面也意味着模型适用范围上的缩小。前面提到过 Poisson 分布建模中的过度离散现象，事实上，除了用负二项分布替换，还可以通过设置一个过度离散参数（McCullagh and Nelder，1989）来解决问题。再以 Gamma 分布模型为例，在这种情况下，一旦 φ 参数取 1，那么 Gamma 分布实际上已经退化为指数分布了。如果所有两个参数可以取不同值的话，其形状可以很丰富。图 1.1 ~ 图 1.3 分别展示了形状参数为 0.5、1 和 3 下，尺度参数分别取 0.1（对应曲线 f_01）、0.3（对应曲线 f_03）、0.5（对应曲线 f_05）、1（对应曲线 f_1）、3（对应曲线 f_3）和 5（对应曲线 f_5）情况下的概率密度图。

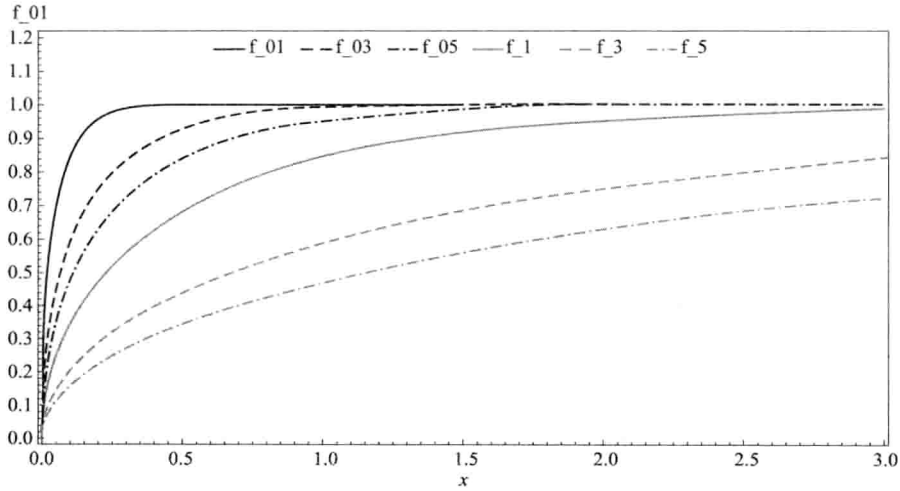

图 1.1　形状参数为 0.5 的 Gamma 分布图

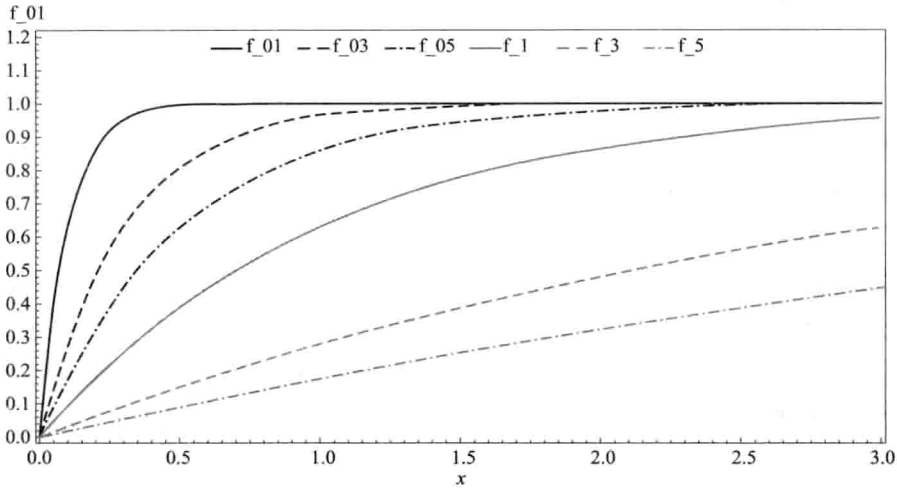

图 1.2　形状参数为 1 的 Gamma 分布图

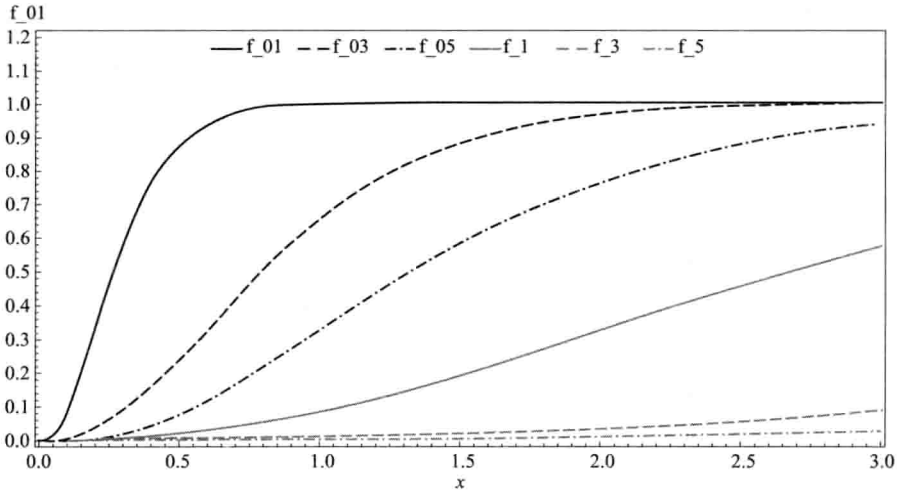

图 1.3　形状参数为 3 的 Gamma 分布图

正如后面将要提到的，Wolfinger-O'Connell 使用了基于伪似然（pseudo-likelihood，PL）函数的方法来估计，发展了伪似然法或者受限伪似然法，这两种方法可以放宽上述罚拟似然（penalized quasi-likelihood，PQL）方法的假定。尽管如此，仅仅这七类分布还是远远不够的，现实数据需要有更多灵活性的分布模型。

在生存分析和保险精算中，应用最广泛的分布簇还是广义 beta 分布簇、广义 Gamma 分布簇、(a，b，0) 分布类、(a，b，1) 分布类和复合分布类。广义 beta 分布簇和广义 Gamma 分布簇主要用来刻画连续变量。广义 beta 分布簇中有两个重要特例，一个是 Burr 分布，一个是逆 Burr 分布；作为 Burr 分布最重要特例的 Pareto 分布具有众数为零的特点，作为逆 Burr 分布最重要特例的逆 Pareto 分布不存在高阶矩，这些都严重限制了其应用。但作为广义 Gamma 分布特例的 Weibull 分布、Gamma 分布以及指数分布在刻画损失风险、生存时间等变量的应用上很广泛，同时还能把对数正态模型作为其极限分布来研究。同时该分布可以轻易变换得到逆幂变换 Gamma 分布簇，进而把逆 Gamma 分布和逆 Weibull 分布作为特例包含进来。

尽管作为特例的 Gamma 分布和指数分布都是指数分布簇，但广义 Gamma 分布本身已经超出指数分布簇的范围，因此与经典的广义线性模型或者广义线性混合模型有一定的差别。

1.2 文献综述

广义线性模型的概念最先在 1972 年 Nelder 和 Wedderburn 一篇论文中引进，他们同时提出了广义线性模型极大似然估计方程。McCullagh and Nelder (1989) 提出一些适用于广义线性模型的一般方法，他们同时还研究了固定效应估计的渐进正态性问题。

另外一个领域中研究人员对异质性数据展开了深入分析，常用的模型有三大类，即边际效应模型、随机效应模型和转移模型 (Diggle，Liang and Zeger，1995)。转移模型在 Korn and Whittemore (1979)，Ware 等 (1988)，Wong (1986)，Zeger and Qaqish (1988) 的研究中出现，但这些研究主要集中在离散变量研究中。

Zeger，Liang and Albert (1988) 集中在纵向数据系统讨论了随机效应模型的建模问题。提出关于随机效应部分建模的两种不同的思路：也即总体平均法 (PA，population-averaged) 和个体设定方法 (SS，subject-

specific)。SS 方法的重要例子是最优线性无偏预测（BLUP，best linear unbiased prediction）以及 Stein 型的收缩估计（Stein-type shrinkage estimation)。

线性随机效应模型的研究成果颇丰，Stefanski and Carroll（1985）发现边际模型的参数估计值小于随机效应模型估计值的绝对值现象。Neuhaus 等（1991)、Zeger 等（1988）进一步研究了三类模型之间的系数关系。随着研究的进展，一些学者开始构建线性混合模型，把上述众多的模型放在一个统一的框架下进行分析，线性混合模型提供了一个统一的模型框架，在这个模型框架中可以同时进行随机效应分析、边际分析、重复观测分析、空间相关分析以及小域估计。

随着广义线性模型和线性随机效应模型的发展，逐渐出现了两类模型之间的结合。一些学者在广义线性模型中引入随机效应项；而线性随机效应模型的研究人员尝试把正态分布推广到其他分布来解释实际数据非正态情况下的应用问题，特别是计数数据的建模问题。后来的很多广义线性混合模型都只是 Harville（1977，P328）线性随机效应模型方法在 GLM 非线性估计方程中的推广。

参数估计等推断的前提是构造似然函数，如果引入随机效应项扩展到广义线性混合模型，那么对随机效应的处理，无论是条件似然法还是极大似然法都会遇到很多麻烦。一些学者选择偏似然方法（Cox，1975）进行边际模型估计，偏似然是边际似然和条件似然的推广。而 Fraser（1968)、Kalbfleisch and Sprott（1974）早已对边际似然和条件似然进行了研究。

完全似然方法也有不少学者研究，其中最关键的一步是对似然函数的推广。Wedderburn（1974）提出拟似然（quasi-likelihood）函数，在 McCullagh（1981)、McCullagh and Nelder（1989）的文献中有系统论述。

拟似然的提出为一大类模型提供了推断的基础。对于大多数非正态模型，需要使用数值积分（如 Crouch and Spiegelman，1990)。在阶数取更高的时候，可以使用 Monte Carlo 积分方法。Li and Raghunathan（1991）在先验分布中使用重要重复抽样或者 Gibbs 抽样来回避数值积分（Zeger and Karim，1991)。对于求解极大似然估计，最常用的是 EM 算法（Dempster 等，1977)。

当然还有似然函数的近似处理方法，这类近似方法包括 Laplace 近似