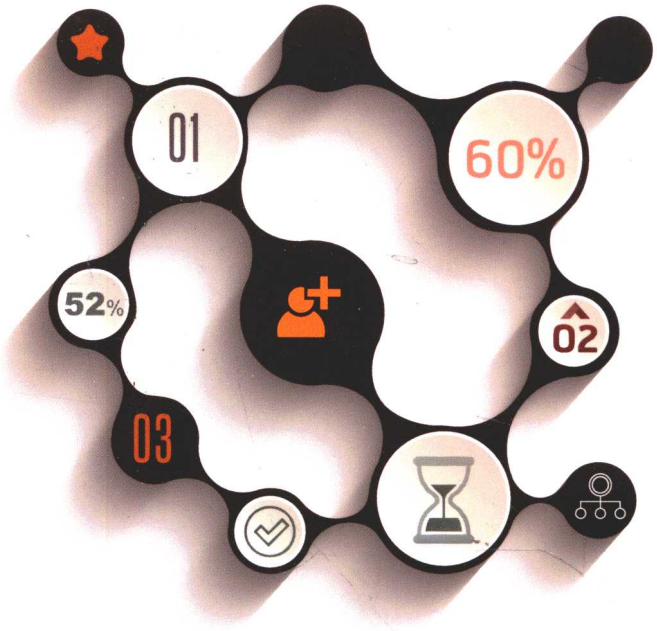

纵向数据与生存数据的半参数联合模型

唐安民 唐年胜 赵 慧 著

Semiparametric Joint Models of Longitudinal and Survival Data

科学出版社

纵向数据与生存数据的半参数联合模型

唐安民　唐年胜　赵　慧　著

科学出版社

北京

内 容 简 介

本书基于半参数贝叶斯方法或极大似然方法，提出几个更具柔性和实践性的纵向数据与生存数据的半参数联合模型，在模型里包含更少的假定. 首先，放松随机效应或个体内测量误差的全参数分布的假设. 用中心化 Dirichlet 随机过程或偏正态分布定义它们的先验分布，由此而推导出的后验分布可以柔性地具有对称、偏态或多峰的分布的特征. 其次，基于惩罚样条方法，用半参数方法建模纵向数据与生存数据共享的轨迹函数，以及建模基本危险函数. 最后，对于建议的半参数联合模型，发展了一些统计诊断方法，包括数据删除影响分析和局部影响分析. 另外，由于联合模型涉及两类复杂数据，并应用半参数方法建模，所以发展高效的半参数联合模型的算法也是本书的亮点.

本书可供统计专业研究生和临床医学研究者参考.

图书在版编目（CIP）数据

纵向数据与生存数据的半参数联合模型/唐安民，唐年胜，赵慧著. —北京：科学出版社，2018

ISBN 978-7-03-053671-6

Ⅰ. ①纵… Ⅱ. ①唐… ②唐… ③赵… Ⅲ. ①半参数模型-研究 Ⅳ. ①O211.3

中国版本图书馆 CIP 数据核字（2017）第 140363 号

责任编辑：张振华 / 责任校对：王 颖

责任印制：吕春珉 / 封面设计：东方人华平面设计部

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

北京虎彩文化传播有限公司 印刷

科学出版社发行 各地新华书店经销

*

2018 年 6 月第 一 版 开本：B5（720×1000）

2018 年 6 月第一次印刷 印张：9 1/2

字数：170 000

定价：79.00 元

（如有印装质量问题，我社负责调换〈虎彩〉）

销售部电话010-62136230 编辑部电话010-62135120-2005

前　言

纵向数据和生存数据联合模型普遍应用于研究纵向数据和生存数据之间的联系，近年来，在临床医学领域的数据分析中有广泛的应用，特别是在癌症临床研究、艾滋病临床研究和生物医药方面. 在以往的文献中，对联合模型的普遍假定是：随机效应或误差项服从全参数分布，如正态分布；轨迹函数是时间的线性函数. 在本书中，基于贝叶斯方法或极大似然方法，我们提出几个更具柔性和实践性的半参数多维纵向数据和生存数据的联合模型，在模型里包含更少的假定.

在贝叶斯框架下，我们首先放松随机效应或随机误差的全参数分布假设，基于中心化Dirichlet随机过程定义它们的先验分布，推导出它们的后验分布，可以拟合对称、偏态或多峰的分布，也建议多维偏正态分布为随机误差的先验分布，并基于偏正态分布的分层分布推导随机误差的后验分布，从而进行联合模型的统计推断和统计诊断；其次，基于贝叶斯惩罚样条方法，部分线性联合建模纵向数据与生存数据共享的轨迹函数，即纵向指标与时间的关系考虑非线性函数，而对其他协变量考虑线性函数；基于逐段常函数的半参数方法或基于贝叶斯惩罚样条非参数方法建模基本危险函数.

为了执行贝叶斯推断，基于数值积分技术近似复杂的生存函数，我们发展了Metropolis-Hastings算法和Gibbs抽样的快速高效的混合算法，从半参数联合模型的后验分布抽样随机观察值，可同时获取未知参数、非参数函数和随机效应的贝叶斯估计. 基于Laplace分布定义感兴趣参数的先验，我们建议了一种半参数联合模型的贝叶斯Lasso（least absolute shrinkage and selection operator）方法，利用该方法在获取未知参数、随机效应和非参数函数的贝叶斯估计的同时，能对感兴趣参数进行变量选择. 在贝叶斯框架下，基于ϕ-距离，我们发展了贝叶斯数据删除影响分析和贝叶斯局部影响分析诊断半参数联合模型里的潜在异常点或影响点.

在极大似然框架下，基于标量对向量或矩阵求导方法，我们建议一个半参数部分线性联合模型，其中纵向数据服从于多维偏正态分布，结合惩罚样条和蒙特

卡罗最大期望算法来估计感兴趣的参数和非参数函数. 在E步，我们基于结合了Gibbs抽样技术和Metropolis-Hastings算法的混合算法，可以从随机效应和潜变量相应条件分布抽取观察值，从而可以近似计算期望得到Q函数；在M步，我们基于数值积分技术近似生存函数，从而大大简化了Q函数的一阶导数和二阶导数计算. 借助于最大期望算法，我们进行了极大似然数据删除影响分析及局部影响分析，执行半参数联合模型的统计诊断.

在本书的出版之际，我要向云南大学数学与统计学院的唐年胜教授表示衷心的感谢，本书的许多成果都是在他的悉心指导下完成的；感谢赵慧对本书所做的校稿工作和给出的一些建设性的意见；感谢云南大学数学与统计学院统计系各位老师的启发和对本书提出的修改建议；也要感谢给予我许多帮助和支持的学长和同学；本书的出版得到了国家自然科学基金（编号：11561074）及云南大学复杂数据统计推断省创新团队基金、云南大学统计建模和数据分析省重点实验室的资助，对此作者向他们表示诚挚的谢意.

由于作者水平有限，书中错误或不当之处在所难免，恳请专家及广大读者批评指正.

唐安民

2017 年 11 月

目　录

第 1 章　绪论 …… 1
1.1　研究背景概述 …… 1
1.2　国内外研究现状 …… 2
1.3　本书主要工作 …… 3
1.4　生存数据分析 …… 6
1.5　纵向数据 …… 8
1.6　统计算法 …… 9
1.6.1　Gibbs抽样 …… 9
1.6.2　MH算法 …… 10
1.6.3　MCEM算法 …… 10
1.6.4　标量或向量对向量求导 …… 11
第 2 章　纵向数据和生存数据半参数联合模型的贝叶斯推断 …… 16
2.1　引言 …… 16
2.2　半参数联合模型 …… 18
2.3　联合模型的贝叶斯分析 …… 23
2.3.1　条件分布及算法 …… 25
2.3.2　贝叶斯估计 …… 30
2.4　半参数联合模型贝叶斯数据删除影响分析 …… 31
2.5　模拟研究及实例 …… 32
2.5.1　贝叶斯推断模拟研究 …… 32
2.5.2　贝叶斯数据删除影响统计诊断模拟试验 …… 41
2.5.3　实例分析 …… 42
第 3 章　偏正态纵向数据和生存数据部分线性半参数联合模型的贝叶斯推断 …… 51
3.1　引言 …… 51
3.2　半参数联合模型 …… 53

3.3 联合模型的贝叶斯分析 59
3.3.1 贝叶斯估计 60
3.3.2 条件分布及算法实现 61
3.4 半参数联合模型贝叶斯局部影响分析 66
3.5 模拟研究及实例 69
3.5.1 贝叶斯统计推断模拟研究 69
3.5.2 贝叶斯局部影响分析模拟研究 75
3.5.3 实例分析 77
第 4 章 半参数联合模型的贝叶斯变量选择 84
4.1 引言 84
4.2 半参数联合模型 86
4.2.1 模型和概念 86
4.2.2 测量误差项分布的设定 87
4.2.3 建模对数基本危险函数 88
4.2.4 生存函数的近似计算 89
4.2.5 先验的设定 89
4.3 变量选择的BLasso方法 90
4.4 半参数联合模型的贝叶斯算法和抽样 93
4.5 模拟研究及实例 96
4.5.1 贝叶斯统计推断模拟研究 96
4.5.2 实例分析IBCSG数据 103
第 5 章 偏正态纵向数据和生存数据联合模型极大似然统计推断和诊断 107
5.1 引言 107
5.2 半参数联合模型 108
5.2.1 建模偏正态纵向数据 109
5.2.2 生存数据子模型 110
5.3 半参数联合模型EM算法 111
5.3.1 E步 112
5.3.2 M步 114
5.3.3 算法执行 114

5.4　半参数联合模型的极大似然统计诊断分析 …… 115
5.4.1　极大似然数据删除影响分析 …… 115
5.4.2　极大似然局部影响分析 …… 116
5.5　模拟研究及实例 …… 117
5.5.1　半参数联合模型极大似然统计推断 …… 117
5.5.2　半参数联合模型极大似然统计诊断 …… 121
5.5.3　实例分析 …… 123
第 6 章　总结及进一步研究 …… 129
附录 …… 132
附录A　Q函数关于未知参数θ的一阶导数和二阶导数 …… 132
附录B　局部影响分析1 …… 135
附录C　局部影响分析2 …… 136
参考文献 …… 138

第1章 绪 论

本章首先阐述联合模型的起源和发展，并综述国内外的研究现状和我们近几年的研究工作，包括几类半参数联合模型及求解等；然后介绍联合模型中的数据结构及常用的研究方法，以及本书要用到的一些统计算法.

1.1 研究背景概述

生存分析一般是指对事件时间的分析，已广泛应用于许多研究领域，包括医药、生物学、工程学、公共卫生、流行病学和经济学等. 这里的事件时间指事件发生的时间，事件可能是死亡、复发和痊愈等，也可能是试验者退出试验. 在医药研究中，为了研究药效，往往要收集大量的纵向数据和生存数据. 例如，在不同观察时间点要收集患者的各方面信息(如血压、心跳及血红蛋白数等)，以及患者从生病到痊愈的恢复时间或痊愈到再次复发的复发时间；对癌症患者的临床研究中，往往要收集事件时间和在不同时间测量的患者各项生理指标或反映患者生活质量(quality of life，QOL)的指标. 这里生存数据就是事件时间和示性指标(确定到底是哪一类事件)，纵向数据可能是在不同观测时间点记录的肿瘤细胞数、对疫菌的免疫反应度量、基因的生物标志和健康指标等. 纵向数据和生存数据通过某些方式联系，最常用的联系方式是纵向轨迹函数. 单独分析纵向数据或生存数据可能会导致有偏差或无效估计，而对生存数据和纵向数据联合建模，同时考虑两类数据的所有信息，则可提供有效的统计推断. Ibrahim等(2010)建议采用以下联合模型建模一维纵向数据和生存数据，并说明了联合建模的一般结构和相比较单独建模的优势.

一维线性模型建模纵向数据一般有如下结构：

$$Y_{ij} = X_i(t_{ij}) + \varepsilon_{ij}. \tag{1.1}$$

其中，Y_{ij}是第i个个体在第j个观察时间点的观察值(纵向指标)；ε_{ij}是误差项，通常假定服从正态分布；$X_i(t_{ij})$ 是联合模型的轨迹函数，在以往的文献中通常

假定：

$$X_i(t_{ij}) = \theta_{0i} + \theta_{1i}t_{ij} + \beta Z_i, \tag{1.2}$$

这里轨迹函数是观察时间t_{ij}和协变量Z_i的线性函数，参数β反映了对纵向指标的处理效应，θ_{0i}、θ_{1i}分别反映了纵向指标随时间变化的强度. 在时刻t，生存模型的危险函数通常定义为

$$\lambda(t) = \lambda_0(t)\exp\{\psi X_i(t) + \gamma Z_i\}. \tag{1.3}$$

其中，$\lambda_0(t)$为基本危险函数，反映了没有个体差异的危险函数；参数γ反映了事件时间的处理效应；参数ψ测度纵向轨迹函数与事件时间之间的关联；$\psi X_i(t) + \gamma Z_i$ 揭示事件时间的总的处理效应. 联合建模的主要目标是估计感兴趣参数ψ、β、γ，从而估计总的处理效应，为临床设计提供有意义的参考. 显然，当$\psi = 0$ 时，说明纵向指标与事件时间并没有必然联系，相比较于单独建模生存模型，联合模型中的纵向指标提供的信息对事件时间的处理效应γ估计不能提供有价值的信息，所以在这种情况下，联合建模是不必要的，对生存数据建模可以忽略纵向指标；而当$\psi < 0$时，危险率下降，即纵向指标的增加会相应地减少危险率.

Tsiatis等(1995)建议分两步建模：第一步，基于假定的线性混合模型建模纵向数据，从而拟合轨迹函数；第二步，视拟合好的轨迹函数为随时间变化的协变量，代入Cox 模型(Cox, 1972)建模生存数据. 相比较于在Cox模型直接代入原始的纵向数据，这种方法有它的优点，但依然可能会导致估计偏差和低效. 在这以后，出现了很多关于联合建模的文献，建模方法是同时建模纵向数据和生存数据，而不是分两个阶段建模. 这些联合模型，相比较于单独对事件时间建模，在探究影响事件时间的因素的过程中结合纵向数据的信息，可以提高估计效率. 所以，当纵向指标与事件时间高度相关时，可以小样本取得较高的功效.

1.2 国内外研究现状

早期联合建模生存数据和纵向数据主要是为了研究来自于HIV/AIDS的临床试验数据，Pawitan等(1993)、De Gruttola等(1994)、Taylor 等(1994)、Tsiatis

等 (1995)、Faucett 等 (1996)、Lavalley 等 (1996)、Wulfsohn 等 (1997)、Wang 等(2001)、Faucett 等(2002)、Brown 等(2005)及Chi等(2007)发表了有关具体建模生存数据和纵向CD4数据方面的论文.在联合模型中，Henderson等(2000)、Xu等(2001a, 2001b)及Song等(2002)发表了多维纵向数据的文献.在癌症临床试验中，Chi等(2006, 2007) 联合建模探究事件时间与生活质量指标之间的关系. Ibrahim等(2004)、Brown等 (2003a, 2003b)、Chen等 (2004). Schluchter (1992)和Hogan等(1997)联合建模研究了生物医药应用. Zhu 等(2012) 建议了贝叶斯局部影响测度；Ibrahim等(2002)、Tsiatis等(2005)和Ibrahim等(2010)在他们的专著或论文中系统回顾和评论了联合模型.

综合上述文献，不难发现：对随机效应和纵向数据误差项假定全参数分布，如正态分布；轨迹函数是关于时间的参数函数，如线性函数或多项式函数. 本书的主要工作就是放松这些参数假定，而建议采用更柔性和实践性更强的联合模型，并给出了联合模型在不同非参数假定下的局部影响分析和数据删除影响分析.

1.3 本书主要工作

我们建议采用一个联合模型，建模K维纵向数据和M维生存数据如下：

$$\begin{aligned} \boldsymbol{Y}_{ij} &= \boldsymbol{\eta}(t_{ij}, \boldsymbol{b}_i) + \boldsymbol{\varepsilon}_{ij} = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{R}_i + \boldsymbol{f}(t_{ij}) + \boldsymbol{W}(t_{ij})\boldsymbol{b}_i + \boldsymbol{\varepsilon}_{ij}, \\ \lambda_m(t|\boldsymbol{b}_i) &= \lambda_{m0}(t)\exp\{\boldsymbol{\eta}^{\mathrm{T}}(t, \boldsymbol{b}_i)\boldsymbol{\psi}_m + \boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_m\}. \end{aligned} \tag{1.4}$$

其中，$i = 1, \cdots, n$, $j = 1, \cdots, n_i$, $m = 1, \cdots, M$；轨迹函数$\boldsymbol{\eta}(t_{ij}, \boldsymbol{b}_i) = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{R}_i + \boldsymbol{f}(t_{ij}) + \boldsymbol{W}(t_{ij})\boldsymbol{b}_i$ 被纵向数据与生存数据共享，在第j个观测时间t_{ij}观察到的纵向数据$\boldsymbol{Y}_{ij}$为K 维，纵向协变量$\boldsymbol{R}_i$对应固定效应参数矩阵$\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_K]$，第$k$列$\boldsymbol{\beta}_k = (\beta_{k1}, \cdots, \beta_{kp})^{\mathrm{T}}$，在观测时间$t_{ij}$ 随机效应设计矩阵为$\boldsymbol{W}(t_{ij})$，随机效应向量为$\boldsymbol{b}_i$，如果$\boldsymbol{W}(t_{ij})$是单位矩阵，那么我们考虑的是随机截距模型，$\boldsymbol{f}(t_{ij}) = (f_1(t_{ij}), \cdots, f_K(t_{ij}))^{\mathrm{T}}$ 为时间效应函数向量，如果所有分量函数都是线性函数，那么轨迹函数$\boldsymbol{\eta}(t, \boldsymbol{b}_i)$是线性的，否则是部分线性的；$\lambda_m(t|\boldsymbol{b}_i)$ 为第i个个体的第m个事件时间的危险函数；$\lambda_{m0}(t)$为基本危险函数；参数$\boldsymbol{\psi}_m$ 测度了纵向

轨迹函数与第m个事件时间之间的关联；$\boldsymbol{Z}_i$ 是与时间独立的生存协变量，对应固定效应向量$\boldsymbol{\gamma}_m$.

在联合模型(1.4)中，在尽可能少的参数假定条件下，我们基于半参数方法建议了一个半参数联合模型，包括中心化Dirichlet过程混合模型(centralization Dirichlet process mixture model, CDPMM) 或中心化Dirichlet过程离散模型(centralization Dirichlet process discrete model，CDPDM) 建模随机效应$\boldsymbol{b}_i$ 或误差项$\boldsymbol{\varepsilon}_{ij}$ 的先验，用惩罚样条(penalized splines，P-splines)或分段常函数拟合基本危险函数$\lambda_{m0}(t)$，用部分线性模型建模轨迹函数$\boldsymbol{\eta}(t, \boldsymbol{b}_i)$，也尝试用偏正态分布(skew normal，SN) 建模随机效应$\boldsymbol{b}_i$ 和纵向数据误差项的分布$\boldsymbol{\varepsilon}_{ij}$. 对我们假定的模型，基于马尔科夫蒙特卡罗(Markov Chain Monte Carlo，MCMC) 算法或蒙特卡罗最大期望(Monte Carlo expectation maximization，MCEM)算法执行了贝叶斯统计推断或极大似然统计推断，并发展了用数据删除影响分析统计量度量删除一个数据点(集)对参数估计或似然函数的影响，以及用局部影响分析评价微小扰动对不同方向的局部影响. 基于我们建议的半参数统计推断或统计诊断方法，用我们建议的模型分析了IBCSG(international breast cancer study group，国际乳腺癌研究组) 数据，并与Chi等(2006) 和Zhu等(2012) 分析结果进行了比较.

联合模型建模涉及两类复杂数据：纵向数据和生存数据. 特别地，如果涉及非平衡纵向数据和右删失生存数据，计算量相当巨大，为了提高计算效率而不损失(或微小损失)估计效率，我们主要做以下3方面工作：①在计算生存数据似然函数时，我们基于矩形积分近似计算生存函数，这样对未知参数求一阶导数还是二阶导数都相当简便；②在用Dirichlet随机过程(Dirichlet process，DP) 建模随机效应或纵向测量误差时，我们用截断Dirichlet随机过程(truncated Dirichlet process，TDP)；③基于MATLAB设计程序时，尽可能用矩阵运算，而减少循环语句.

在第2章，介绍了基于贝叶斯方法的纵向数据和生存数据的半参数联合模型. 首先，中心化Dirichlet随机过程(centralization Dirichlet process，CDP)混合正态模型建模联合模型的随机效应. 具体地，假定第i 个个体的随机

效应$\boldsymbol{b}_i$(均值为0) 的先验分布为混合正态分布$\sum_{g=1}^{\infty} p_g N(\boldsymbol{\mu}_g, \boldsymbol{\Omega}_g)$，即从一系列正态分布$\{N(\boldsymbol{\mu}_g, \boldsymbol{\Omega}_g) | g = 1, 2, \cdots\}$中选取第$L_i$个正态分布，其中$L_i$从多点分布Multinomial$(p_1, p_2, \cdots)$抽取，$p_1, p_2, \cdots$由折棍(stick-breaking)随机过程确定，为了确保随机效应均值为0，还引入以权重$p'_g s$加权中心化$\boldsymbol{\mu}'_g s$；其次，基于半参数方法建模基本危险函数，即分段常函数方法；最后，给定先验分布，推导出感兴趣的未知参数、随机效应和基本危险函数的条件后验分布；最后，还给出这个模型数据删除影响分析的方法.

在第3章，基于半参数贝叶斯部分线性联合模型建模多维偏正态纵向数据和生存数据. 在第2章的基础上，这一章提出了一种更具普遍性和实践性的联合模型. 在这个模型里，假定轨迹函数是时间的非参数函数、协变量的线性函数，即部分线性模型，纵向数据的误差项服从于多维偏正态分布(正态分布是其特例)，中心化Dirichlet随机过程离散模型建模随机效应. 基于贝叶斯惩罚样条(Bayes splines，B-slpines)、多维偏正态分布的分层模型、分段常函数方法及相应的先验分布假定，构建联合模型的贝叶斯推断，并提出了该联合模型的贝叶斯局部影响分析，具体是通过扰动随机效应和误差项，度量对微小扰动的敏感性，从而诊断模型中的异常点或影响点.

在第4章，用中心化Dirichlet随机过程混合正态分布建模纵向随机误差项，用贝叶斯惩罚样条建模对数基本危险函数，并考虑基于Laplace 分布设定感兴趣参数的先验，并给出Laplace 分布的分层模型，由于在分层模型中只涉及熟识的正态分布和指数分布，从而建议采用半参数联合模型的贝叶斯Lasso(Bayesian least absolute shrinkage and selection operator，Blasso) 变量选择，可以高效地同时实现参数估计和变量选择.

在第5章，讨论了纵向数据和生存数据联合模型的极大似然推断. 本章在极大似然框架下分析了联合模型，在这个模型里，假定随机效应服从偏正态分布，误差项服从正态分布，部分线性轨迹函数和基本危险函数用惩罚样条拟合. 建模方法类似于第3章，只是在这一章不是基于贝叶斯方法，而是基于极大似然方法. 为了得到观测数据的对数似然函数，视随机效应和潜变量

为缺失数据，基于MCMC抽样填充，然后在E 步近似计算得Q函数，在M步，极大化Q 函数，得未知参数一步迭代值，不断迭代直至收敛，这一方法就是MCEM 算法. 对建议的联合模型，这一章还提出基于MCEM算法下的数据删除影响分析和贝叶斯局部影响分析.

接下来，我们将介绍本书用到的一些基本知识，包括生存数据分析、纵向数据、一些统计算法(Gibbs抽样、MH算法及MCEM 算法)及标量或向量对向量求导的知识.

1.4 生存数据分析

为了对生存数据建模，我们首先定义一些概念和公式. 令T是连续非负随机变量，用来描述某一总体中的一个个体生存时间，$f(t)$ 是T 的概率密度函数(probability density function，PDF)，那么其累积分布函数$F(t)$(cumulative distribution function，CDF)可表示为

$$F(t) = P(T \leqslant t) = \int_0^t f(u)\mathrm{d}u.$$

那么该个体生存到时间t的概率由以下生存函数定义：

$$S(t) = 1 - F(t) = P(T > t).$$

显然，生存函数$S(t)$是一单调递减函数，并且有$S(0) = 1$及$S(\infty) = \lim\limits_{t\to\infty} S(t) = 0$. 危险函数$\lambda(t)$ 表示在时间t 的瞬时失败率，可以定义为

$$\lambda(t) = \lim_{\Delta t\to 0} \frac{P(t < T \leqslant t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)}. \tag{1.5}$$

特别地，$\lambda(t)\Delta t$表示一个个体存活到时间t，在区间$(t, t + \Delta t]$死亡(或失败)的近似概率. 函数$f(t)$!$F(t)$、$S(t)$ 及$\lambda(t)$ 都可以用来定义生存时间随机变量T的分布. 由$f(t) = -\dfrac{\mathrm{d}}{\mathrm{d}t}S(t)$，基于式(1.5)，容易推导得

$$\lambda(t) = -\frac{\mathrm{d}}{\mathrm{d}t}\log(S(t)). \tag{1.6}$$

容易验证，危险函数有以下性质：$\lambda(t) \geqslant 0$及$\int_0^{+\infty} \lambda(t)\mathrm{d}t = +\infty$. 对式(1.6) 两边关于$t$ 积分，并指数化可得

$$S(t) = \exp\left(-\int_0^t \lambda(u)\mathrm{d}u\right). \tag{1.7}$$

定义累积危险函数$\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$，于是由式(1.7)可得生存函数与累积函数之间的关系为$S(t) = \exp(-\Lambda(t))$. 结合式(1.5)和式(1.7)不难得到随机变量T的概率密度函数为

$$f(t) = \lambda(t)\exp\left(-\int_0^t \lambda(u)\mathrm{d}u\right). \tag{1.8}$$

概率密度函数$f(t)$可以理解为一个个体活到时间t，并在时刻t瞬时死亡的机率，它是构造生存数据似然函数的基础. 当随机变量T 的分布只与时间t 有关时，常见参数分布有指数分布、威布尔分布、极值分布、对数正态分布和伽马分布等. 然而，随机变量T 的分布不仅与时间有关，而且与一些协变量有关，Cox(1972) 提出了一个比例危险模型，具体如下：

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp(G(\boldsymbol{x},\boldsymbol{\gamma})).$$

其中，$\lambda_0(t)$是基本危险函数，反映了不考虑个体间差异的危险函数；$\boldsymbol{\gamma}$是回归系数向量；协变量$\boldsymbol{x}$与时间有关，第二项表示为指数形式是为确保危险函数为非负，也为了计算简便. 通常，假定协变量的影响是线性的，上式可以写成：

$$\lambda(t|\boldsymbol{x}) = \lambda_0(t)\exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\gamma}). \tag{1.9}$$

由于在式(1.9)中，基本危险函数并不是t的参数函数，所以该式可以视为对生存数据的半参数建模. 假定协变量与时间无关，Cox(1972) 建议极大化$\boldsymbol{\gamma}$的部分似然函数估计未知参数$\boldsymbol{\gamma}$. 由式(1.9)得到的生存分布通常不是标准分布，为了产生随机数，需要用到逆变换法，从均匀分布$U(0,1)$抽取一随机数u，然后解关于t的方程$F(t) = u$，得生存时间T的随机数，有时可能要用数值方法求解该方程的近似解.

本节并不考虑上述常见分布或比例危险模型建模生存数据，而是用一种半参数方法，基本危险函数用分段常函数或惩罚样条拟合，并且有部分协变

量与时间有关. 我们考虑建模右删失生存数据，包括事件时间和示性数据. 当生存时间只对一部分个体是已知的，这时的事件时间是生存时间，而其余个体只知道它们的生存时间超越某些固定值(如退出试验时间)，这时的事件时间是右删失时间. 假定一个个体的确切的生存时间为t^*，而右删失时间为c，那么事件时间为$t = \min(t^*, c)$，为了判定事件时间到底是确切生存时间还是右删失时间，还须引入一个示性数据$\delta = \mathbf{1}(t^* \leqslant c)$，$\mathbf{1}(A)$ 是一个示性函数，当A是真时取值为1. 于是，随机数对(t, δ) 就构成这个个体的右删失生存数据，基于式(1.8)，相应于这个个体$\boldsymbol{\theta}$的似然函数为

$$L(\boldsymbol{\theta}|(t,\delta)) = [\lambda(t)]^{\delta} \exp\left\{-\int_0^t \lambda(u)\mathrm{d}u\right\}. \tag{1.10}$$

其中，危险函数$\lambda(t)$类似于式(1.4)中定义，可能涉及与时间有关的纵向预测变量，也包含与时间独立的生存协变量.

1.5 纵向数据

纵向数据(longitudinal data)是指对同一组受试个体在不同时间或空间的重复观测数据(Diggle et al.，2002)，已广泛应用于医学、生物学、社会学和经济学等诸多领域. 在社会学和经济学中，纵向数据也被称为面板数据(panel data). Fitzmaurice等(1993)较为全面地介绍了纵向数据分析中的模型、分析方法及最近的发展. Diggle等(2002)系统探究了纵向数据的统计推断，讨论了基于线性和广义线性模型的纵向数据统计分析；纵向数据与截面数据(cross-sectional data)不同，截面数据指仅仅在某一时间点对同一组个体做一次观测所得的数据. 当只考虑一个个体数，而有很多时间点时，纵向数据就是时间序列数据. 假设有n 个观测个体, 记t_{ij} 为第i 个个体第j 次观测的时间，$\boldsymbol{x}_{ij}$ 和$\boldsymbol{y}_{ij}$ 分别为第i 个个体在时间t_{ij} 的解释变量和响应变量，n_i 为第i 个个体重复观测的数目，则纵向数据集可记为$\{(t_{ij}, \boldsymbol{y}_{ij}^{\mathrm{T}}, \boldsymbol{x}_{ij}^{\mathrm{T}}), 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n_i\}$. 在纵向数据中，研究者的兴趣通常集中在评价时间$t$ 和解释变量x 对响应变量y 的效应. 其中，解释变量x可以依赖于时间t，也可以不依赖于时间t. 尽管在不同个体之间的观测数据是独立的，但由于是对同一个体进行重复观测，因此对同一个体的不同观察数据可能是相关的.

对于截面数据，一般是对响应变量的总体均值进行建模，而对纵向数据，可以有几种不同的方法用于建模响应变量的依赖关系：第一种方法称为边缘模型(marginal model)，类似于截面数据均值建模，基于个体间的独立性将响应变量对解释变量的回归与组内相关性分离开来进行建模，其研究的重点是回归系数的估计，而把内相关视为讨厌参数，回归系数是基于总体的，而不是基于个体的，当个体数目很大，且观测次数较多时，可以考虑此方法；第二种方法称为随机效应模型，假定组内相关性是由于个体间回归系数的变化而产生的，回归系数可理解为随机变量，解释对个体的作用，而不是总体平均的作用，当研究的目的是对个体进行推断，而并非总体平均数时，随机效应模型特别有用，也是建模纵向数据最常用的方法；第三种方法称为转移模型(the transition model)，这种模型是在给定过去响应变量的情况下建模当前响应变量的条件均值，类似于时间序列模型中时滞模型，回归系数的解释依赖于个体的先前观测，因而，同一个体内重复观测值具有相关性时，可以考虑该方法.

1.6 统计算法

在这一节，我们介绍本书用到的一些统计算法，包括Gibbs抽样、Metropololis-Hastings(MH)算法、MCEM算法及标量或向量对向量求导.

1.6.1 Gibbs抽样

首先，我们定义单元素Gibbs抽样. 令$\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_p)^{\mathrm{T}}$是$p$维参数向量，$\pi(\boldsymbol{\theta}|\boldsymbol{D})$是未知参数$\boldsymbol{\theta}$的给定观察数据$\boldsymbol{D}$的条件分布，给出起始点$\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \cdots, \theta_p^{(0)})^{\mathrm{T}}$，假定第$\iota$次迭代开始时的估计值为$\boldsymbol{\theta}^{(\iota-1)}$，那么Gibbs抽样(Geman et al.，1984)第ι步迭代分为如下p步：

(1) 从条件后验分布$\pi(\theta_1|\theta_2^{(\iota-1)}, \cdots, \theta_p^{(\iota-1)}, \boldsymbol{D})$抽取$\theta_1^{(\iota)}$.

……

(i) 从条件后验分布$\pi(\theta_i|\theta_1^{(\iota)}, \cdots, \theta_{i-1}^{(\iota)}, \theta_{i+1}^{(\iota-1)}, \cdots, \theta_p^{(\iota-1)}, \boldsymbol{D})$抽取$\theta_i^{(\iota)}$.

……