

广义线性模型引论

陈 夏 著



科学出版社

广义



陈 夏 著

陕西师范大学出版基金资助出版



科 学 出 版 社

北 京

内 容 简 介

本书系统阐述了广义线性模型的基本理论、方法及其应用. 全书共 8 章: 第 1 章是绪论, 概括广义线性模型的基本理论和方法; 第 2 章介绍广义线性模型极大似然估计理论; 第 3 章讨论自然联系下广义线性模型的拟似然估计理论; 第 4 章介绍二值响应变量自然联系拟似然方程解的相合性; 第 5 章讨论基于广义估计方程的广义线性模型拟似然理论; 第 6 章给出广义线性模型的非参数拟似然估计理论; 第 7 章讨论广义线性模型的方差函数估计理论; 第 8 章介绍广义线性模型的经验似然方法.

本书可作为数学、生物、医学、经济、金融等领域的教师或科研工作者的参考书, 也可作为高等院校数学、概率统计、生物统计和计量经济等相关专业高年级本科生、研究生的教材.

图书在版编目(CIP)数据

广义线性模型引论/陈夏著. —北京: 科学出版社, 2017.3

ISBN 978-7-03-052162-0

I. ①广… II. ①陈… III. ①线性模型—研究 IV. ①O212

中国版本图书馆 CIP 数据核字 (2017) 第 053647 号

责任编辑: 李 萍 / 责任校对: 赵桂芬

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>

北京教图印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2017 年 3 月第 一 版 开本: 720 × 1000 B5

2017 年 3 月第一次印刷 印张: 15 1/4

字数: 310 000

定价: 95.00 元

(如有印装质量问题, 我社负责调换)

谨以此书献给恩师陈希孺先生！

前 言

广义线性模型是经典线性模型的直接推广,可用于连续数据和离散数据统计分析,特别是属性数据和计数数据.这在实际应用上,尤其是医学、生物、经济和社会数据的统计分析上有相当重要的意义.

广义线性模型的个别特例起源很早, Fisher 在 1919 年曾使用过它. Logistic 模型作为广义线性模型最重要的特例之一,在 20 世纪四五十年代曾由 Berkson 等使用过.但抽象成一般形式,用统一的方法去处理,却是在 20 世纪 60 年代末. 1983 年, McCullagh 和 Nelder 出版了专著 *Generalized Linear Models*,并于 1989 年再版.该专著对广义线性模型有较系统的论述,但应用较少,内容面也窄一些. 1994 年, Fahrmeir 和 Tutz 出版了关于广义线性模型的多元统计建模的专著 *Multivariate Statistical Modelling Based on Generalized Linear Models*,并于 2001 年再版.该书的应用举例较多,内容广泛,但缺乏深度.上述两本专著都在理论上着力很少,所有结果都未给出证明.

因此,本书把近些年来研究成果以及广义线性模型的一般理论和方法整理出来,侧重于理论结果的讨论和证明,希望能对读者进入这个领域有一定的指导作用.作者在武汉大学攻读博士学位期间,跟随恩师陈希孺先生系统而深入地学习了广义线性模型理论,并在相应的方向上取得了一定的进展.当年开设的广义线性模型讨论班,除了恩师的弟子之外,还有来自中国科技大学、华中科技大学、华中师范大学等高校的研究生,形成了良好的讨论氛围.本书内容主要取材于这个系列讨论班及其后续的研究成果.

本书是广义线性模型理论的一本介绍性和研究性著作,希望对概率论与数理统计的科研工作者有一定的参考价值.同时,它可作为概率论与数理统计专业研究生的学位课程教材,也可作为大学数学专业或经济专业特别是保险精算专业的选修课教材,并可供需要这方面知识的读者自学.

借本书出版之际,向我的恩师陈希孺先生表示衷心的感谢.不幸的是,陈老师已经于 2005 年 8 月离开了我们.恩师是我国著名的数理统计学家,曾任中国科学技术大学研究生院(北京)教授、中国现场统计研究会理事长、中国统计学会副理

事长, 1997 年当选为中国科学院院士. 恩师学富五车、德高望重, 他严谨治学、提携后辈、平易近人的风范影响了一代统计科研工作者.

本书的出版得到了科学出版社的支持, 在此表示诚挚的谢意. 另外, 感谢武汉大学概率统计科学系刘禄勤教授、邹新堤教授、胡亦均教授、高付清教授、刘妍岩教授和章逸平教授等, 以及中国科学技术大学统计与金融系的赵林城教授, 在作者求学期间的照顾和关心. 感谢作者的同门师妹和同窗好友丁洁丽与岳丽为本书提供的材料, 希望她们一切都好. 经过她们的同意, 本书列出了她们各自部分的研究成果. 还要感谢作者的师兄——中国科学院大学的张三国教授一直以来的支持和鼓励! 感谢国家自然科学基金 (项目编号: 11201276)、陕西师范大学出版基金和中央高校基本科研业务费专项资金 (项目编号: GK201503012, GK201503015) 提供的资助! 感谢陕西师范大学数学与信息科学学院的领导和同事一直以来的支持和关心! 特别感谢家人的理解和支持!

由于作者水平有限, 书中难免有不妥之处, 恳请广大读者批评指正.

陈 夏

2016 年 10 月

于陕西师范大学数学与信息科学学院

目 录

前言

第 1 章	绪论	1
1.1	广义线性模型概论	1
1.2	统计推断	10
1.3	本书主要工作	28
第 2 章	广义线性模型极大似然估计理论	30
2.1	引言	30
2.2	广义线性模型极大似然估计的相合性问题	37
2.3	自然联系下带随机自变量的广义线性模型的大样本理论	53
2.4	非自然联系下带随机自变量的广义线性模型似然估计的渐近理论	68
2.5	广义线性回归参数的“学生化”极限定理	83
第 3 章	自然联系下广义线性模型的拟似然估计理论	96
3.1	引言	96
3.2	弱相合性	98
3.3	强相合性	106
3.4	渐近正态性	109
第 4 章	二值响应变量自然联系拟似然方程解的相合性	117
4.1	引言	117
4.2	一个反例	118
4.3	自变量缓增长的情形	120
4.4	自变量无界情况的一个正面例子	125
第 5 章	基于广义估计方程的广义线性模型拟似然理论	128
5.1	引言	128
5.2	强相合性及收敛速度	130
5.3	渐近正态性	143
第 6 章	广义线性模型的非参数拟似然估计理论	158
6.1	引言	159

6.2	主要结果	163
6.3	若干引理	164
6.4	定理的证明	180
6.5	应用	182
第 7 章	广义线性模型的方差函数估计理论	188
7.1	方差函数估计	188
7.2	一个方差函数估计的确定	208
第 8 章	广义线性模型的经验似然方法	215
8.1	引言	215
8.2	经验似然方法简介	216
8.3	方法和主要结果	219
8.4	模拟研究	221
8.5	定理的证明	223
参考文献		228

第1章 绪 论

广义线性模型 (generalized linear models, GLMs) 的个别特例起源很早. 1919 年, Fisher 曾使用广义线性模型的个别特例来分析社会学中的一些现象. 20 世纪四五十年代, Berkson 等也使用过广义线性模型中最重要的特例 Logistic 模型, 来研究社会和经济中的一些问题. 自 1972 年 Nelder 和 Wedderburn 引入“广义线性模型”一词以来^[1], 研究工作逐渐增加. 1983 年, McCullagh 和 Nelder 出版了系统论述广义线性模型的专著 *Generalized Linear Models*, 并于 1989 年再版^[2]. 该书以介绍基本理论和基本方法为主, 应用较少, 内容广度稍显不足. 1994 年, Fahrmeir 和 Tutz 出版了关于广义线性模型的多元统计建模的专著 *Multivariate Statistical Modelling Based on Generalized Linear Models*, 并于 2001 年再版^[3]. 该书的应用举例较多, 内容广泛, 但缺乏深度. 两本书都在理论上着力不够, 所有结果都未给予证明. Fahrmeir 与 Kaufmann 在 1985 年首次系统讨论了广义线性模型参数极大似然估计的强弱相合性和渐近分布问题^[4], 并从理论上给予了严格证明. 2002 年, 陈希孺在《数理统计与管理》期刊上连载了关于广义线性模型的专题^[5-11], 以介绍基本理论和基本方法为主. 2011 年, 陈希孺出版了专著《广义线性模型的拟似然法》. 该书是一本关于广义线性模型的理论的入门用书, 简单介绍了关于广义线性模型的几种基本统计推断形式 (极大似然估计、假设检验和拟似然估计) 的大样本理论^[12].

下面首先介绍广义线性模型的基本内容和基本理论; 其次讨论广义线性模型的统计推断问题, 包括极大似然估计方法、假设检验、拟似然估计方法和其他非参数方法等; 最后介绍本书的主要工作.

1.1 广义线性模型概论

广义线性模型从形式上直接推广了正态线性模型, 它不仅可以对连续数据进行统计分析, 还可以对属性数据和计数数据等离散数据类型进行分析. 这对于广义线性模型在实际问题中的应用, 如对社会、经济、生物和医学数据的统计分析有重要的意义. 广义线性模型的提出开辟了一个新的领域, 并在应用中发挥了重要的作用.

早期的广义线性模型出现在 19 世纪初, 用来分析人口增长等社会现象, 其中研究较多的是 Logistic 模型和 Probit 模型. 在相当长的时期内, 这两类模型只作为研究个别社会和经济问题的一种工具. 直到 1972 年, Nelder 和 Wedderburn 正式引入广义线性模型的概念, 并利用极大似然方法处理广义线性模型中的估计问题 [1]. 1974 年, Wedderburn 首次考虑了广义线性模型中响应变量的分布未知 (但期望和方差有正确设定的情形), 提出了广义线性模型中的拟似然方法 [13]. 自此研究工作逐渐增加, 广义线性模型也成为实际数据建模的重要工具.

1.1.1.1 一维广义线性回归模型

首先来看经典的线性模型:

$$y_i = z_i^T \beta + e_i, \quad i = 1, \dots, n$$

这里 y_i 为一维的响应变量 (因变量); $z_i = z(x_i)$ 是依赖于协变量 (自变量) x_i 的设计向量, z^T 表示 z 的转置, 本书均以记号 “T” 表示转置; β 是未知的参数向量. 一般地, 假定误差序列 e_i 相互独立且服从正态分布 $N(0, \sigma^2)$.

下面用另一种方式来重新表示线性模型, 并希望在总结线性模型的特点的基础上, 更好地理解引入广义线性模型的意义. 因变量 y_i 相互独立并且服从正态分布, 即

$$y_i \sim N(\mu_i, \sigma^2), \quad i = 1, \dots, n \quad (1.1.1)$$

且 $\mu_i = E(y_i)$. 均值 μ_i 以如下的线性形式给出:

$$\mu_i = z_i^T \beta, \quad i = 1, \dots, n$$

如果协变量 x_i 是随机的, 则假定 (y_i, x_i) 是独立同分布的. 式 (1.1.1) 应该理解为, 在给定 x_i 的条件下, y_i 是条件分布的, 并且 y_i 是条件独立的.

因此, 设有因变量 y , 自变量 x , y 为一维, x 一般为多维. 可以总结通常的线性回归有以下三个特征.

(1) $E(y) = \mu = z^T(x)\beta$ (线性参数 β), $z(x)$ 为 x 的已知 (向量) 函数. 例如, 若 x 为一维, 则 $z^T(x)$ 可以是 $(1, x)$, $(1, x, x^2)$ 等; 若 x 是多维, 则 $z^T(x)$ 可取 $(1, x^T)$ 等. 以下 $z(x)$ 简记为 z .

(2) 变量 $x, z(x), y$ 多取连续值的情形, 如工农业生产中的产品产量、人的身高体重等特征测量.

(3) 响应变量 y 的分布为正态, 或近似于正态分布.

线性回归模型应该说是到目前为止研究得最为广泛的统计模型之一, 理论上已经达到了一个相当完善的程度. 可以说, 上述特点在线性模型的理论和应用发展中起着很重要的作用. 但在很多实际问题中, 所研究的数据可能不具有以上的特点, 因此很难把它们纳入到经典线性模型的框架.

给一组生物体施加一定剂量的刺激剂 (如毒素等), 记 y 为测量到起反应的生物体的个数, 在剂量 x 下, 起反应的生物体期望比例 p 满足

$$p = \Phi(\alpha + \beta x)$$

这里 Φ 是标准的正态分布函数, α, β 为待估参数. 当 y 取 0 或 n 时, $\Phi^{-1}(y/n)$ 为 ∞ , 所以不能取 $\Phi^{-1}(y/n)$ 为因变量, 得到 $\Phi^{-1}(y/n) = \alpha + \beta x$, 而用经典的线性模型去处理. 因此只能是 y/n 的期望与 x 呈线性关系, y (或它的函数) 与 x 不是线性的; 并且 (y/n) 的方差不是常数, 方差是随均值而变化的, $p = 0$ 或 1 时为零, $p = 0.5$ 时达到最大.

基于此, 广义线性回归从线性回归的三个特征分别进行推广.

(1) $E(y) = \mu = h(z^T \beta)$, h 一般假定为已知的非线性函数, 且是一个充分光滑的单调函数. $g = h^{-1}$ (h 的反函数) 称为联系函数 (link function), 且满足

$$g(\mu) = z^T \beta$$

注意此时 $E(y)$ 已不等于 $z^T \beta$, 而是关于 $z^T \beta$ 的某一函数形式.

(2) 变量 $x, z(x), y$ 可取连续或离散值的情形, 在实际问题的具体研究中更多的是针对离散值的情形, 如 $\{0, 1\}$ 和 $\{0, 1, 2, \dots\}$ 等.

(3) 响应变量 y 的分布为指数型分布, 正态分布是其一特例. 其具体形式可表示为

$$C(y) \exp(\theta y - b(\theta)) d\nu(y), \quad \theta \in \Theta \quad (1.1.2)$$

其中, Θ 为参数空间, θ 称为自然参数, $b(\theta)$ 为 θ 的已知函数, ν 为一测度, 常见的有下面两种可能.

(i) 当 y 为连续变量时, $d\nu(y)$ 为 Lebesgue 测度 $d\nu(y) = dy$.

(ii) 当 y 为离散变量时, y 取有限个值 $a_1, \dots, a_m$ 或可列个值 $a_1, a_2, \dots$, 这时

$$\nu(\{a_i\}) = 1, \quad i = 1, \dots, m; \quad \text{或} \quad \nu(\{a_i\}) = 1, \quad i = 1, 2, \dots$$

对于指数型分布族, 首先指出以下事实: 若 y 有分布(1.1.2), 则容易验证

$$E(y) = \dot{b}(\theta) (:= db(\theta)/d\theta), \quad \text{Var}(y) = \ddot{b}(\theta) (:= d^2b(\theta)/d\theta^2). \quad (1.1.3)$$

常见指数型分布的例子如下.

二项分布 设 $Y \sim B(n, p)$, 分布函数为

$$P(Y = y) = C_n^y p^y (1-p)^{n-y} = C_n^y \exp(y \ln p / (1-p) + n \ln(1-p))$$

故二项分布属于指数型分布族, 其中 $\theta = \ln p / (1-p)$, $b(\theta) = n \ln(1 + e^\theta)$, 并且可验证式 (1.1.3) 成立.

Poisson 分布 设 $Y \sim P(\lambda)$ ($\lambda > 0$), 分布函数为

$$P(Y = y) = \frac{1}{y!} e^{-\lambda} \lambda^y = \frac{1}{y!} \exp(y \ln \lambda - \lambda), \quad \lambda > 0$$

令 $\theta = \ln \lambda$, 有

$$P(Y = y) = \frac{1}{y!} \exp(y\theta - \lambda), \quad y = 0, 1, 2, \dots, \quad -\infty < \theta < \infty$$

故 Poisson 分布属于指数型分布族, 且可验证式 (1.1.3) 成立.

正态分布 设 $Y \sim N(\mu, \sigma^2)$, 密度函数为

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{(y-\mu)^2}{2\sigma^2} \right\} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{y^2}{2\sigma^2}} \exp(\theta y - \sigma^2 \theta^2 / 2)$$

其中 $\theta = \mu/\sigma^2$, $b(\theta) = \sigma^2 \theta^2 / 2$. 显然 $\dot{b}(\theta) = \theta \sigma^2 = \mu = E(Y)$, $\ddot{b}(\theta) = \sigma^2 = \text{Var}(Y)$.

注 1.1.1 在指数型分布中, 均值函数 $\mu = h(z^T \beta)$ 决定了特殊的方差函数结构, 即方差是均值的函数. 这是因为由上述事实知 $\ddot{b}(\theta) = \text{Var}(y) > 0$, 所以 $\dot{b}$ 严格单调, 因此有反函数 $\dot{b}^{-1}$. 故由 $E(y) = \dot{b}(\theta)$ 知 $\theta = \dot{b}^{-1}(E(y)) = \dot{b}^{-1}(\mu)$, 从而知

$$\sigma^2 = \text{Var}(y) = \ddot{b}(\theta) = \ddot{b}(\dot{b}^{-1}(\mu)) := \sigma^2(\mu)$$

但在有些实际问题中, 数据可能显示均值与方差之间的关系不符合此要求, 这时就不能使用指数型分布的假设了. 如果有均值和方差函数的正确设定, 则利用下面提到的拟似然方法可以解决这个问题.

注 1.1.2 联系函数的选择依赖于特定的指数型分布族, 即响应变量分布的类型, 以及实际问题本身的情况. 对于指数型分布族而言, 总是存在一个特殊的联系

函数,称之为自然联系函数 (natural or canonical link function),即要求自然参数 θ 满足

$$\theta = \eta = z^T \beta$$

则由 $g(\mu) = \eta = z^T \beta$ 和 $E(y) = \mu = \dot{b}(\theta)$ 知,此时 $g = \dot{b}^{-1}$ 或 $h = \dot{b}$. 可以看到,自然联系函数的选择会使得广义线性模型下统计推断的大样本理论更易处理.

注 1.1.3 如果自变量的可能组合数很少,则样本会呈现分组的状态. 设 $y_1, \dots, y_m$ 是同一 x 值下的 y 样本,这时往往用一个样本. 例如,若用 $\bar{y} := \sum_{i=1}^m y_i / m$ 取代 $y_1, \dots, y_m$, 则即只看到 $\bar{y}$ 和 m , 而不一定见到原始数据记录. 这样做并无损失,原因如下.

(i) $\bar{y}$ 是充分统计量,因此无信息损失.

(ii) $\bar{y}$ 的分布仍为指数型,并且 $E(\bar{y}) = E(y)$, $\text{Var}(\bar{y}) = \text{Var}(y)/m$.

这些讨论都是在 $y_1, \dots, y_m$ 为 i.i.d. 的条件下进行的,实际问题中这可能不完全成立. 例如, ①同一组 x 值上所观察的 y 值有相关性; ②有些未包含在 x 中的因素,如问题中未予考虑或尚未认知的,对每个观察值的影响不同,而使 $y_1, \dots, y_m$ 分布不同. 这两点总的影晌是加大 $\sum_{i=1}^m y_i$ 的方差,即比按公式算的大,称为“超散布性” (overdispersion).

1.1.2 模型建立与联系函数的选择

下面通过一个例子,说明如何对实际问题进行广义线性模型建模,并选择合适的联系函数.

例 1.1.1 研究一些因素,如剖宫事先有无计划、服用抗生素与否和有无危险因子等对“产妇剖宫产后是否有感染”的影响. 引入因变量 Y , 若有感染定义 $Y = 1$, 否则 $Y = 0$, 自变量 $x = (x_1, x_2, x_3)^T$, 其中

$$x_1 = \begin{cases} 1, & \text{剖宫事先未计划} \\ 0, & \text{剖宫事先计划} \end{cases}$$

$$x_2 = \begin{cases} 1, & \text{服用抗生素} \\ 0, & \text{不服用} \end{cases}$$

$$x_3 = \begin{cases} 1, & \text{有危险因子} \\ 0, & \text{剖宫事先计划} \end{cases}$$

记 $\pi = P(Y = 1)$, 则 (对 $y = 0, 1$)

$$P(Y = y) = \pi^y(1 - \pi)^{1-y} = (1 - \pi) \left(\frac{\pi}{1 - \pi} \right)^y = (1 - \pi) \exp \left(y \ln \frac{\pi}{1 - \pi} \right) \quad (1.1.4)$$

令 $\theta = \ln(\pi/(1 - \pi))$, 则 $1 - \pi = 1/(1 + e^\theta)$, 此时式 (1.1.4) 可写为

$$P(Y = y) = \exp(\theta y - \ln(1 + e^\theta)), \quad -\infty < \theta < \infty$$

此相当于式 (1.1.2) 中的 y 为离散变量取值 $\{a_1, a_2\} = \{0, 1\}$, $b(\theta) = \ln(1 + e^\theta)$, $C(y) = 1$ 的情况, 此时有

$$\dot{b}(\theta) = e^\theta / (1 + e^\theta) = \pi (= E(y))$$

$$\ddot{b}(\theta) = e^\theta / (1 + e^\theta)^2 = \pi(1 - \pi) (= \text{Var}(y))$$

与式 (1.1.3) 一致.

此例中取 $z = (1, x^T)^T = (1, x_1, x_2, x_3)^T$, 令 $\eta = z^T \beta$. 设观察了 n 位产妇, 得到样本 (y_i, z_i) , 其中 $z_i = (1, x_{1i}, x_{2i}, x_{3i})^T$. 记 $\eta_i = z_i^T \beta$, 对应的 π, θ 分别为 π_i, θ_i . 设引入了联系函数 $g(\pi_i) = \eta_i$ 或 $\pi_i = h(\eta_i)$, 则 $(y_1, \dots, y_n)$ 的联合分布为

$$\exp \left\{ \sum_{i=1}^n y_i \ln \frac{h(z_i^T \beta)}{1 - h(z_i^T \beta)} + \sum_{i=1}^n \ln(1 - h(z_i^T \beta)) \right\}$$

利用它可对未知参数 β 进行统计推断, 如判断所提到的三个因素对“产后感染”发生概率的影响.

若本例取自然联系函数, 则应有 $z^T \beta = \theta = \ln(\pi/(1 - \pi))$, 因为 π 为均值, 即取

$$g(t) = \ln \frac{t}{1 - t} \quad \text{或} \quad h(t) = \frac{e^t}{1 + e^t}$$

此时

$$\pi = \frac{e^{z^T \beta}}{1 + e^{z^T \beta}}$$

这就是著名的 Logistic 模型.

注 1.1.4 一般地, $\pi = h(z^T \beta)$, h 应满足 $0 < h < 1$. 若 h 为严增函数, 则 $h(-\infty)$ 一般应为 0, $h(+\infty)$ 一般应为 1, 这样 π 可取 $(0, 1)$ 内的任何值. 因此, h 应为一分布函数. 实际中, 主要有:

(i) probit模型: $h_1(t) = \Phi(t)$ (标准正态分布), 联系函数 $g(t) = \Phi^{-1}(t)$;

(ii) ln-ln模型: $h_2(t) = 1 - \exp(-\exp(t))$, 联系函数 $g(t) = \ln(-\ln(1 - t))$.

1.1.3 多维广义线性回归模型

在实际问题建模中, 目标变量取为向量的情况也比较常见, 如研究人的身高和体重. 因此, 有必要引入多维广义线性模型 (此处多维指因变量或响应变量是多维). 设响应变量 Y 为 q 维: $Y = (Y_1, Y_2, \dots, Y_q)$, 它取值于 $\mathbb{R}^q$ 中的 Borel 集 (可能为离散有限或可列集; 也可能是 $\mathbb{R}^q$ 中的连续区间). 与一维广义线性模型三个特征类似, 多维响应变量 Y 有指数型分布:

$$c(y) \exp\{\theta^T y - b(\theta)\} d\nu(y), \quad \theta \in \Theta \quad (1.1.5)$$

其中 $\theta = (\theta_1, \dots, \theta_q)^T$ 为 q 维参数向量. 与一维指数型分布类似, 容易验证

$$\mu := E(Y) = \dot{b}(\theta) := \frac{\partial b(\theta)}{\partial \theta} = \left(\frac{\partial b}{\partial \theta_1}, \dots, \frac{\partial b}{\partial \theta_q} \right)^T \quad (1.1.6)$$

和

$$\text{Cov}(Y) = \ddot{b}(\theta) := \frac{\partial^2 b(\theta)}{\partial \theta \partial \theta^T} = \left(\frac{\partial^2 b}{\partial \theta_i \partial \theta_j} \right)_{i,j=1,\dots,q} \quad (1.1.7)$$

设有协变量 x (向量), 考虑它对响应变量 Y 的影响. 由 x 产生 $p \times q$ 矩阵 $Z = Z(x)$. 例如, 在 x 的多项式回归中, Z 包括 x 各分量的一些幂次以及交互乘积等, 后面称为设计矩阵. β 为 p 维未知参数向量, 记

$$\eta = Z^T \beta$$

定义集合

$$B := \left\{ \mu : \mu = \int_A y c(y) \exp\{\theta^T y - b(\theta)\} d\mu(y), \theta \in \Theta \right\}$$

联系函数 g 是一个定义于 B 上取值为 $\mathbb{R}^q$ 的充分光滑的函数, 满足

$$g(\mu) = \eta = Z^T \beta$$

且为单射. 可以证明, g 的反函数 h 也是同阶光滑, 且有

$$\mu = E(Y) = h(Z^T \beta)$$

注意此时 $g, h, Z^T \beta$ 都是 q 维列向量.

若有了样本 $(y_i, x_i), 1 \leq i \leq n$, 相应地, 有 $Z_i := Z(x_i)$ 和 $\eta_i = Z_i^T \beta$, 以及

$$\theta_i = \dot{b}^{-1}(\mu_i) = \dot{b}^{-1}(h(Z_i^T \beta))$$

$(y_1, \dots, y_n)$ 的联合密度为

$$\prod_{i=1}^n C(y_i) \exp \left\{ \sum_{i=1}^n y_i^T \dot{b}(h(Z_i^T \beta)) - \sum_{i=1}^n b(\dot{b}(h(Z_i^T \beta))) \right\} \quad (1.1.8)$$

利用它可以对未知参数 β 进行统计推断.

若 $g = \dot{b}^{-1}$ 或 $h = \dot{b}$, 则称 g 为自然联系函数, 满足

$$\theta = \dot{b}^{-1}(\mu) = g(\mu) = \eta$$

因为 $\text{Cov}(y) > 0$, 所以

$$\ddot{b}(\theta) = \frac{\partial^2 b(\theta)}{\partial \theta \partial \theta^T} > 0$$

故映射 $\dot{b}(\cdot)$ 为一一映射, 又 $b(\theta)$ 有任意阶偏导数, 故光滑条件满足, 从而 $\dot{b}^{-1}$ 满足联系函数的条件.

1.1.4 多维 Logistic 模型

继续前面的讨论, 目标变量取 $a_1, \dots, a_k$ 这 k 个值. 若取 a_j 的概率记为 $\pi_{(j)}$, $j = 1, \dots, q = k-1$, 而 $\pi = (\pi_{(1)}, \dots, \pi_{(q)})^T$ (取 a_k 的概率为 $1 - (\pi_{(1)} + \dots + \pi_{(q)})$). π 受自变量 x 的影响, 可记为 $\pi = \pi(x)$.

$Y = (y_{(1)}, \dots, y_{(q)})^T$ 的分布为指数型: 如果取 $(y_{(1)}, \dots, y_{(q)})^T$ 为前述的 $a_1, \dots, a_k$ 之一, 则

$$P(Y = (y_{(1)}, \dots, y_{(q)})^T) = (1 - \pi_{(1)} - \dots - \pi_{(q)})^{1 - (y_{(1)} + \dots + y_{(q)})} \prod_{j=1}^q \pi_{(j)}^{y_{(j)}} \quad (1.1.9)$$

令 $\theta = (\ln \tilde{\theta}_{(1)}, \dots, \ln \tilde{\theta}_{(q)})^T = (\theta_{(1)}, \dots, \theta_{(q)})^T$, 其中

$$\tilde{\theta}_{(j)} = \frac{\pi_{(j)}}{1 - (\pi_{(1)} + \dots + \pi_{(q)})}$$

则

$$\pi_{(j)} = \frac{\tilde{\theta}_{(j)}}{1 + \sum_{l=1}^q \tilde{\theta}_{(l)}} = \frac{e^{\theta_{(j)}}}{1 + \sum_{l=1}^q e^{\theta_{(l)}}}, \quad j = 1, \dots, q$$

可将式 (1.1.9) 写为

$$\begin{aligned} (1 - \pi_{(1)} - \dots - \pi_{(q)}) \prod_{j=1}^q \tilde{\theta}_{(j)}^{y_{(j)}} &= \left(1 + \sum_{j=1}^q \tilde{\theta}_{(j)} \right)^{-1} \exp \left(\sum_{j=1}^q y_{(j)} \theta_{(j)} \right) \\ &= \exp(\theta^T y - b(\theta)) \end{aligned}$$

其中

$$b(\theta) = \ln \left(1 + \sum_{j=1}^q \tilde{\theta}_{(j)} \right) = \ln \left(1 + \sum_{j=1}^q e^{\theta_{(j)}} \right)$$

是指数型分布的形状, ν 为集合 $\{a_1, \dots, a_k\}$ 上的计数测度.

自然联系函数为 (注意到 $\theta = \eta = g(\mu), \mu = \pi, \mu := (\mu_{(1)}, \dots, \mu_{(q)})^T, \pi_{(j)} = \mu_{(j)}$)

$$g(\mu) = (g_{(1)}(\mu), \dots, g_{(q)}(\mu))^T, \quad g_{(j)}(\mu) = \ln(\mu_{(j)} / (1 - \mu_{(1)} - \dots - \mu_{(q)}))$$

g 的反函数 h 定义于 $\Theta = \{(\theta_{(1)}, \dots, \theta_{(q)})^T : -\infty < \theta_{(j)} < \infty, j = 1, \dots, q\}, h(\theta) = (h_{(1)}(\theta), \dots, h_{(q)}(\theta))^T$, 其中

$$h_{(j)}(\theta) = \frac{e^{\theta_{(j)}}}{1 + \sum_{l=1}^q e^{\theta_{(l)}}}, \quad j = 1, \dots, q$$

这由 $h = \dot{b}$ 立即得出.

下面要讨论如何构造 $Z(x)$ 及参数 β 的问题. 需要注意: 一个因素对目标变量 Y 处在何种状态的影响, 与该状态有关而非一致. 下面通过一个例子来说明.

例 1.1.2 要调查人们在假期中是否出行及采用何种交通工具受一些因素, 如经济状况、对安全的重视、旅途时间的考虑、性别、从事的工作等的影响. 目标变量 Y 有 4 个状态: 1. 不出行; 2. 坐火车; 3. 坐汽车; 4. 坐飞机.

一般来说, 经济情况不好有利于状态 1, 其次 2, 3, 而不利 4; 安全因素有利于 1 和 2, 而不利 3 和 4, 等等.

这表明, 对不同的状态要设置不同的参数. 现根据问题性质, 由自变量 x 构造出有关的 r 维向量 $l(x)$, 并引入常数项, 则 “ x 对 y 取状态 j 的影响” 应通过

$$\beta_{0j} + l^T(x)\beta_j := \eta_{(j)}(x), \quad j = 1, \dots, q \quad (1.1.10)$$

去表达, 其中 β_j 是 r 维列向量. β_{0j}, β_j 依赖 j , 体现出 x 对各状态的影响不一. 因此 $\eta(x) = (\eta_{(1)}(x), \dots, \eta_{(q)}(x))^T$ 要等于 $Z^T(x)\beta$, 必须取

$$Z^T(x) = \begin{pmatrix} 1, l^T(x) & 0 & \cdots & 0 \\ 0 & 1, l^T(x) & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1, l^T(x) \end{pmatrix}$$