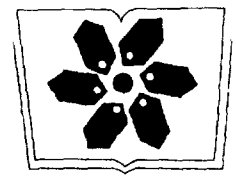
非线性再生散度模型

唐年胜　韦博成 著

科学出版社
www.sciencep.com



中国科学院科学出版基金资助出版

非线性再生散度模型

唐年胜　韦博成　著

科学出版社

北京

内 容 简 介

本书系统介绍非线性再生散度模型和带有随机效应、结构方程以及缺失数据的非线性再生散度模型的理论、方法和若干实际应用,其中包括参数的极大似然估计、Bayes 估计、统计诊断、几何方法、置信域的曲率表示、模型评价和拟合优度、WinBUGS 应用程序等.此外,还介绍这些理论和方法在生物医学、教育心理学和社会学等领域的若干具体应用.

本书可作为统计学、生物医学、教育心理学等专业研究生的教学参考书,也可供相关专业的研究生、教师、科技人员和统计工作者参考.

图书在版编目(CIP)数据

非线性再生散度模型/唐年胜,韦博成著.—北京: 科学出版社,2007
ISBN 978-7-03-019486-2

I.非… II.①唐… ②韦… III.非线性-散度-线性模型 IV.O212

中国版本图书馆 CIP 数据核字(2007) 第 115251 号

责任编辑: 陈玉琢 / 责任校对: 陈玉凤
责任印制: 赵德静 / 封面设计: 王 浩

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

新 蕾 印 刷 厂 印刷

科学出版社发行 各地新华书店经销
*
2007 年 8 月第 一 版 开本: B5(720×1000)
2007 年 8 月第一次印刷 印张: 12 3/4
印数: 1—3 000 字数: 238 000

定价: 32.00 元

(如有印装质量问题, 我社负责调换〈明辉〉)

前　言

再生散度分布族是 Jorgensen 于 1997 年在他的著作《The Theory of Dispersion Models》中提出的, 它是指数族分布的进一步推广. Jorgensen 在该书中曾经预言, 广义线性模型的理论可推广到以再生散度分布族为随机误差的模型. 但是迄今为止, 国内外尚未见到系统介绍这一内容的著作. 本书就是希望弥补这方面的不足. 非线性再生散度模型是经典非线性回归模型、广义线性模型和指数族非线性模型的自然推广和必然发展. 经典非线性回归模型、广义线性模型和指数族非线性模型经过数十年的发展已经有了比较成熟的理论和方法. 近年来, 随着计算机的快速发展和人们对信息需求的提高, 一些更复杂的模型, 如广义线性混合模型、带有缺失数据的广义线性混合模型、指数族非线性混合模型和非线性结构方程模型等得到了国内外统计学者的青睐, 并在国外出现了系统介绍这些复杂模型的专著. 本书则系统介绍非线性再生散度模型的基本理论和若干实际应用方法, 其中包括参数的极大似然估计、Bayes 估计、统计诊断、几何方法、置信域的曲率表示、模型评价和拟合优度、WinBUGS 应用程序等. 本书在详细介绍基本统计理论和方法的同时, 还重点介绍了这些理论和方法在生物医学、教育心理学和社会学等领域的具体应用, 并辅以图表和 WinBUGS 应用程序, 力求做到学以致用. 本书涉及概率论与数理统计的一些基础知识, 假定读者已经熟悉, 不再一一介绍. 我们所使用的概念、符号都是一般教科书上常见的、通用的, 因此也不一一说明. 我们希望本书的出版能引起回归分析、生物医学、教育心理学和社会学等方面学者的兴趣. 特别是, 第 5 章和第 6 章的部分内容还在继续研究和探索之中, 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. 本书共分 6 章. 第 1 章引入非线性再生散度模型的概念, 详细介绍该模型参数的极大似然估计和 Bayes 估计. 第 2 章基于数据删除模型和局部影响分析方法研究模型的统计诊断和影响分析. 第 3 章对非线性再生散度模型在欧氏空间建立几何结构, 基于此几何结构研究模型参数置信域的曲率表示. 第 4 章介绍非线性再生散度随机效应模型, 包括参数的极大似然估计和 Bayes 估计、统计诊断、影响分析、模型的几何结构及参数置信域的曲率表示. 第 5 章研究非线性再生散度结构方程模型, 并介绍了基于 Gibbs 抽样和 Metropolis-Hastings 算法的 Bayes 估计、基于 Path Sampling 的模型选择方法和拟合优度统计量及 WinBUGS 语言的应用. 第 6 章介绍带有缺失数据的非线性再生散度结构方程模型的 Bayes 估计、Bayes 模型比较及 WinBUGS 软件的应用.

本书的出版得到了中国科学院科学出版基金的支持, 同时也得到国家自然科学

基金 (10561008, 10671032)、高等学校博士学科点专项科研基金 (20060673002) 和云南省自然科学基金 (2004A0002M) 及云南大学统计学重点建设专业经费的支持, 特此表示衷心的感谢! 本书在写作过程中, 自始至终得到科学出版社的关心与帮助, 特别要感谢数理分社的陈玉琢同志, 她对本书的写作、审定与出版都给予了大力的支持与帮助, 特此表示衷心的感谢! 同时也对科学出版基金评审过程中审稿老师的厚爱与支持表示衷心的感谢!

由于作者水平有限, 书中难免有不妥之处, 敬请同行专家、学者和广大读者不吝指教.

云南大学　唐年胜
东南大学　韦博成
2007 年 3 月

目　录

第 1 章　非线性再生散度模型

指数族分布是一类重要的统计分布族, 它包括正态分布、二项分布、Poisson 分布、Gamma 分布和逆 Gauss 分布等许多常见分布 (Wei　1998). 广义线性模型的特点是以指数族分布为其随机误差, 从而开辟了非正态回归分析的新领域. 广义线性模型的研究始于 20 世纪 70 年代, 经过近 30 年的发展, 现已广泛地用于工程学、生物学、医学、经济学、教育学和社会学等领域的数据分析. 由于数据的复杂性和人们对数据分析精度要求的提高, 广义非线性模型 (又称指数族非线性模型) 便应运而生. 指数族非线性模型的研究始于 20 世纪 80 年代, 它包括正态非线性回归模型、广义线性模型等, 是一类应用很广的回归模型. Wei (1998) 全面系统地研究了指数族非线性模型的参数估计理论和统计诊断等问题. 然而, 在现实世界中, 仍有许多数据不能用指数族非线性模型来拟合, 为了满足人们对复杂数据分析的需要, 非指数族非线性回归模型的研究也提上了日程. Jorgensen (1997) 在其专著《The Theory of Dispersion Models》中定义了一类比指数族分布更广泛的分布, 他称之为再生散度模型 (reproductive dispersion models, RDM), 并且指出: 广义线性模型的理论可推广到以 RDM 为随机误差的模型. 本书将系统讨论以 RDM 为随机误差的广义非线性模型的统计推断, 这种模型被称为非线性再生散度模型. 本章 1.1 节介绍非线性再生散度模型的定义, 为此, 首先介绍 Jorgensen 提出的再生散度模型. 由于国内外通用分布族 (family) 来表示随机变量的某一类分布, 而不用模型 (即 model), 本书今后都用再生散度分布族来表示 Jorgensen 定义的分布族 (他称之为再生散度模型), 这样也许更确切一些. 在 1.1 节的基础上, 1.2 节和 1.3 节分别介绍非线性再生散度模型的极大似然估计和 Bayes 估计.

1.1　非线性再生散度模型的定义

在定义非线性再生散度模型之前, 首先需要介绍 Jorgensen (1997) 定义的一类分布族 (他称之为再生散度模型) 并讨论其有关的统计推断问题.

1.1.1　再生散度分布族

如果随机变量 Y 的概率密度函数可表示为

$$p(y;\mu,\sigma^2)=a(y;\sigma^2)\exp\Big\{-\frac{1}{2\sigma^2}d(y;\mu)\Big\},\qquad y\in\mathcal{C},\tag{1.1.1}$$

其中 $a(.;.) \geqslant 0$ 为某一合适的已知函数; $d(.;.)$ 为定义在 $\mathcal{C} \times \Theta$ 上的单位偏差度 (unit deviance) 函数, $\Theta \subseteq \mathcal{C} \subseteq \mathcal{R}$, Θ 是一开区间, 凸支撑集 $\mathcal{C}$ 为包含 $\mathcal{S}$ 的最小区间, $\mathcal{S}$ 为概率密度函数的支撑集; $d(.;.)$ 满足正定性条件:

$$d(y;y) = 0, \quad \forall y \in \Theta; \quad d(y;\mu) > 0, \quad \forall y \neq \mu, \tag{1.1.2}$$

其中 $\mu \in \Theta$ 为位置 (position) 参数, $\sigma^2 > 0$ 为散度 (dispersion) 参数, 则称 Y 服从参数为 μ 和 σ^2 的再生散度分布族 (reproductive dispersion family), 简记为 $Y \sim \mathrm{RDF}(\mu, \sigma^2)$. 注意, 函数 $a(\cdot;\cdot)$ 和 $d(\cdot;\cdot)$ 的表达形式不同, 随机变量 Y 的分布亦不一样; 参数 μ 和 σ^2 的意义也不尽相同. 一般说来, $d(\cdot;\cdot)$ 常常可视为平方距离的某种度量, 而式 (1.1.1) 中 $p(y;\mu,\sigma^2)$ 的众数点 (mode) 应在 μ 附近, 而且 σ 的值越小, 峰值就越高、越窄. 因此, 参数 μ 和 σ^2 的意义与通常的位置参数和尺度参数比较类似. 但是, 由于 $a(y;\sigma^2)$ 与 y 有关, 参数 μ 和 σ^2 的意义还与函数 $a(\cdot;\cdot)$ 有关. 因此, 为了区别于通常的位置 (location) 参数和尺度 (scale) 参数, Jorgensen (1997) 建议把 $Y \sim \mathrm{RDF}(\mu, \sigma^2)$ 中的 μ 和 σ^2 分别称为方位 (position) 参数和散度 (dispersion) 参数; 并称相应的分布族为再生散度分布族.

对应于 $a(y;\sigma^2)$ 和 $d(y;\mu)$ 的不同形式, 我们可以得到 (1.1.1) 的如下特殊子分布族:

(1) 如果随机变量 Y 的概率密度函数为

$$p(y;\mu,\sigma^2) = a(\sigma^2)V^{-1/2}(y)\exp\Big\{-\frac{1}{2\sigma^2}d(y;\mu)\Big\}, \tag{1.1.3}$$

其中单位偏差度函数 $d(y;\mu)$ 为正则函数, 即 $d(y;\mu)$ 关于 (y,μ) 在 $\Theta \times \Theta$ 上二次连续可导且满足: 对 $\forall \mu \in \Theta$, $[\partial^2 d(y;\mu)/\partial\mu^2]_{y=\mu} > 0$, 并且 $\mathcal{S} = \mathcal{C} = \Theta$, 则称 Y 服从参数为 μ 和 σ^2 的正则散度分布族 (regular proper dispersion family), 简记为 $Y \sim \mathrm{PD}(\mu, \sigma^2)$. 同时, 正则单位偏差度函数 $d(y;\mu)$ 的单位方差函数定义为

$$V(\mu) = \frac{2}{[\partial^2 d(y;\mu)/\partial\mu^2]_{y=\mu}}.$$

(2) 如果单位偏差度函数可表示为

$$d(y;\mu) = yf(\mu) + g(\mu) + h(y), \tag{1.1.4}$$

其中 $f(.)$, $g(.)$ 和 $h(.)$ 为已知函数, 则称模型 (1.1.1) 为再生指数散度分布族 (reproductive exponential dispersion family), 简记为 $Y \sim \mathrm{ED}(\mu, \sigma^2)$.

(3) 如果 $a(y;\sigma^2) = c(y), \sigma^2 = 1$ 且单位偏差度函数 $d(y;\mu)$ 满足 (1.1.4), 则称模型 (1.1.1) 为自然指数分布族 (natural exponential family), 简记为 $Y \sim \mathrm{NE}(\mu)$. 显然, 当 σ^2 已知时, 再生指数散度分布族即为自然指数分布族.

(4) 如果 $a(y;\sigma^2)=\exp(-\frac{1}{2}s(y,\sigma^2))$ 且单位偏差度函数满足 $d(y;\mu)=d'(y;\mu)-d'(y;y)$, 其中 $d'(y;\mu)=-2(y\mu-b(\mu)-c(y))$, $s(\cdot\ ,\ \cdot)$, $b(.)$ 和 $c(.)$ 为已知函数, 则称 (1.1.1) 为指数分布族 (exponential family) (Wei 1998), 简记为 $Y\sim \mathrm{EX}(\mu,\sigma^2)$.

(5) 如果随机变量 $Y\sim \mathrm{RDF}(\mu,\sigma^2)$, 且二元变换 (duality transformation) $Z=Y/\sigma^2$ 的概率密度函数为

$$p^*(z;\tau,\sigma^2)=a^*(z;\sigma^2)\exp\Big\{-\frac{1}{2\sigma^2}d(z\sigma^2;\tau\sigma^2)\Big\}, \tag{1.1.5}$$

其中 $\tau=\mu/\sigma^2$,

$$a^*(z;\sigma^2)=\begin{cases}\sigma^2 a(z\sigma^2;\sigma^2), & \text{连续型},\\ a(z\sigma^2;\sigma^2), & \text{离散型},\end{cases}$$

则称 (1.1.5) 为加性散度分布族 (additive dispersion family). 如果 $Y\sim \mathrm{ED}(\mu,\sigma^2)$ 且二元变换 Z 的概率密度函数为 (1.1.5), 则称 (1.1.5) 为加性指数散度分布族 (additive exponential dispersion family), 简记为 $Z\sim \mathrm{ED}^*(\theta,\lambda)$, 其中 $\theta=f(\tau\sigma^2)=f(\mu)$ ($f(\mu)$ 的表达式见.(1.1.4)), $\lambda=1/\sigma^2$.

(6) 如果 $a(y;\sigma^2)=a(\sigma^2)$ 且 $d(y;\mu)=d(y-\mu)$, 则称 (1.1.1) 为位置散度分布族 (location dispersion family).

由上面的讨论可以看出, 再生散度分布族是一类较指数族分布更广泛的分布族, 它是指数族分布的推广和发展. 为了方便读者理解再生散度分布族的概念, 下面给出一些常见的例子, 进一步的讨论可参见 Jorgensen (1997) 的专著.

(1) Poisson 分布. 假设随机变量 Y 服从均值为 μ 的 Poisson 分布, 则其概率密度函数为

$$p(y;\mu)=\frac{\mu^y}{y!}\mathrm{e}^{-\mu}=\frac{1}{y!}\exp(y\log\mu-\mu),\quad \forall y\in \mathbf{Z}_+=\{0,1,2,\cdots\}.$$

上式表明 Poisson 分布是一个自然指数族, 其单位偏差度函数为

$$d(y;\mu)=2\Big(y\log\frac{y}{\mu}-y+\mu\Big).$$

单位方差函数为 $V(\mu)=\mu$.

(2) 二项分布. 假设随机变量 Y 服从参数为 m 和 μ 的二项分布 $\mathrm{Bi}(m,\mu)$, 则其概率密度函数为

$$p(y;\mu,m)=\binom{m}{y}\mu^y(1-\mu)^{m-y}=\binom{m}{y}\exp\Big\{y\log\frac{\mu}{1-\mu}+m\log(1-\mu)\Big\},\quad y=0,\cdots,m.$$

由上式可以看出, 对每一固定值 m, 二项分布也是自然指数族.

(3) Gamma 分布. 假设随机变量 Y 服从参数为 $\psi>0$ 和 $\lambda>0$ 的 Gamma 分布, 简记为 $Y\sim\Gamma(\lambda,\psi)$, 则其概率密度函数为

$$p(y;\psi,\lambda)=\frac{\psi^{\lambda}}{\Gamma(\lambda)}y^{\lambda-1}\mathrm{e}^{-\psi y}. \tag{1.1.6}$$

记 $\mu=\lambda/\psi$ 且 $\sigma^2=1/\lambda$, 则 (1.1.6) 可表示为

$$p(y;\mu,\sigma^2)=a(y;\sigma^2)\exp\Big\{-\frac{1}{2\sigma^2}2\Big(\frac{y}{\mu}-\log\frac{y}{\mu}-1\Big)\Big\},$$

其中 $a(y;1/\lambda)=y^{-1}\lambda^{\lambda}\mathrm{e}^{-\lambda}/\Gamma(\lambda)$. 上式表明 Gamma 分布是位置参数为 μ, 散度参数为 σ^2 和单位方差函数为 $V(\mu)=\mu^2$ 的正则散度分布, 其单位偏差度函数为

$$d(y;\mu)=2\Big(\frac{y}{\mu}-\log\frac{y}{\mu}-1\Big).$$

(4) von Mises 分布. 假设随机变量 Y 的概率密度函数为

$$p(y;\mu)=\begin{cases}\dfrac{1}{2\pi I_0(\lambda)}\exp\{\lambda\cos(y-\mu)\}, & \text{若 } y\in[0,2\pi],\\ 0, & \text{其他,}\end{cases} \tag{1.1.7}$$

其中 $\mu\in[0,2\pi),\lambda>0$, $I_0(\lambda)$ 为修正的 Bessel 函数且其表达式为

$$I_0(\lambda)=\frac{1}{2\pi}\int_0^{2\pi}\exp(\lambda\cos y)\mathrm{d}y.$$

从 (1.1.7) 容易看出, von Mises 分布是一个位置参数为 μ, 散度参数为 $\sigma^2=1/\lambda$ 的正则散度分布, 其单位偏差度函数为

$$d(y;\mu)=2\{1-\cos(y-\mu)\},\quad (y,\mu)\in[0,2\pi)\times(0,2\pi).$$

其单位方差函数为 $V(\mu)=1$, 并简记为 $Y\sim\mathrm{vM}(\mu,\sigma^2)$.

(5) 单纯形 (simplex) 分布. 假设 Y 为区间 $(0,1)$ 上的随机变量, 其概率密度函数为

$$p(y;\mu,\sigma^2)=\begin{cases}[2\pi\sigma^2\{y(1-y)\}^3]^{-1/2}\exp\Big\{-\dfrac{1}{2\sigma^2}d(y;\mu)\Big\}, & \text{若 } 0<y<1,\\ 0, & \text{其他,}\end{cases} \tag{1.1.8}$$

其中 $\mu\in(0,1)$, $\sigma^2>0$, 单位偏差度函数为

$$d(y;\mu)=\frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}.$$

由 (1.1.3) 和 (1.1.8) 不难看出, 单纯形分布是一个正则散度分布, 其单位方差函数为 $V(\mu)=\mu^3(1-\mu)^3$. 这一分布可用来刻画连续型的比例数据, 常记为 $Y\sim S^{-1}(\mu,\sigma^2)$.

(6) I 型极值分布 (或 Gumbel 分布). 假设随机变量 Y 的概率密度函数为

$$p(y;\mu)=\mathrm{e}^{-1}\exp\{1+(y-\mu)-\mathrm{e}^{y-\mu}\},\qquad -\infty<y<\infty.$$

这时 $\sigma=1$, $a(y;\sigma^2)=\mathrm{e}^{-1}$, 这也是一个正则散度分布, 其单位偏差度函数为

$$d(y;\mu)=-2[1+(y-\mu)-\mathrm{e}^{y-\mu}].$$

单位方差函数为 $V(\mu)=1$, 常记为 $Y\sim \mathrm{EV}(\mu)$.

(7) 双指数分布. 假设随机变量 Y 的概率密度函数为

$$p(y;\tau,\mu)=\frac{\tau}{2}\exp\{-\tau|y-\mu|\},\qquad -\infty<y<\infty.$$

上式表明双指数分布是一个位置参数为 μ, 散度参数为 $\sigma^2=\tau^{-1}$ 的位置散度分布族, 其单位偏差度函数为

$$d(y;\mu)=2|y-\mu|,$$

简记为 $Y\sim \mathrm{DE}(\mu,\sigma^2)$.

Jorgensen (1997) 系统全面地讨论了单位偏差度和再生散度分布族的性质. 下面只给出一些与本书后面章节有关的内容. 想了解更多内容的读者可以参见 Jorgensen (1997) 的专著.

引理 1.1.1 如果 $d(y;\mu)$ 是一个正则单位偏差度函数, 则

$$\frac{\partial^2 d}{\partial y^2}(\mu;\mu)=\frac{\partial^2 d}{\partial \mu^2}(\mu;\mu)=-\frac{\partial^2 d}{\partial \mu\partial y}(\mu;\mu),\qquad \forall \mu\in\Theta,$$

$$V(\mu)=\frac{2}{\frac{\partial^2 d}{\partial \mu^2}(\mu;\mu)}=\frac{2}{\frac{\partial^2 d}{\partial y^2}(\mu;\mu)}=-\frac{2}{\frac{\partial^2 d}{\partial \mu\partial y}(\mu;\mu)},$$

其中 $(\mu;\mu)$ 表示 $d(y;\mu)$ 计算相应导数后, 变元 $(y;\mu)$ 在 $y=\mu$ 处计值.

证明 由正则单位偏差度函数的定义知

$$d(y;y)=d(\mu;\mu)=0,\qquad d(y;\mu)>0,\qquad \forall y\neq\mu.$$

上式表明: 若视 $d(y;\mu)$ 为 μ 的函数, 则 y 为函数 $d(y;.)$ 的唯一的最小值点; 若视 $d(y;\mu)$ 为 y 的函数, 则 μ 为函数 $d(.;\mu)$ 的唯一的最小值点. 由二元函数的最小值点的性质可知: 对任意 $\mu\in\Theta$, 下面的等式恒成立:

$$\frac{\partial d}{\partial \mu}(\mu;\mu)=0,\quad \frac{\partial d}{\partial y}(\mu;\mu)=0. \tag{1.1.9}$$

对 (1.1.9) 第一式关于 μ 求导数得

$$\frac{\partial^2 d}{\partial \mu^2}(\mu;\mu)+\frac{\partial^2 d}{\partial \mu \partial y}(\mu;\mu)=0. \tag{1.1.10}$$

类似地, 对 (1.1.9) 第二式关于 y 求导数得

$$\frac{\partial^2 d}{\partial y^2}(\mu;\mu)+\frac{\partial^2 d}{\partial \mu \partial y}(\mu;\mu)=0. \tag{1.1.11}$$

由 (1.1.10) 和 (1.1.11) 即得引理 1.1.1 的第一个结论.

由单位方差函数的定义和引理 1.1.1 的第一个结论易知引理 1.1.1 的第二个结论也成立. ∎

由 Jorgensen (1997) 的讨论知: 当 $\sigma^2 \to 0$ 时, 正则散度分布族 (1.1.3) 的鞍点逼近 (saddlepoint approximation) 可表示为

$$p(y;\mu,\sigma^2) \sim \{2\pi\sigma^2 V(y)\}^{-1/2}\exp\Big\{-\frac{1}{2\sigma^2}d(y;\mu)\Big\}. \tag{1.1.12}$$

由上式右边的表达式容易看出, 在一般情况下上式右边的鞍点逼近不一定是某一随机变量在 Θ 上的概率密度函数. 为此, 我们考虑 (1.1.12) 式右边的正态化鞍点逼近:

$$p_0(y;\mu,\sigma^2)=a_0(\mu;\sigma^2)V^{-1/2}(y)\exp\Big\{-\frac{1}{2\sigma^2}d(y;\mu)\Big\}, \tag{1.1.13}$$

其中 $a_0(\mu;\sigma^2)$ 为正态化常数且有表达式:

$$\frac{1}{a_0(\mu;\sigma^2)}=\int_{\Theta} V^{-1/2}(y)\exp\Big\{-\frac{1}{2\sigma^2}d(y;\mu)\Big\}\mathrm{d}y. \tag{1.1.14}$$

Nelder & Pregibon (1987) 研究了指数散度分布族的正态化鞍点逼近, 并称 (1.1.13) 为推广的拟似然 (extended quasi-likelihood); Efron (1986) 称 (1.1.13) 为双指数族 (double exponential family). 于是, 正则散度分布 (1.1.3) 的正态化鞍点逼近可表示为

$$p(y;\mu,\sigma^2) \sim p_0(y;\mu,\sigma^2), \qquad \sigma^2 \to 0.$$

由 (1.1.14) 知, 为了获得正态化常数 $a_0(\mu;\sigma^2)$ 的值, 我们需要求一个积分. 当然, 对一些较复杂的分布, 要获得 $a_0(\mu;\sigma^2)$ 的显示表达式是十分困难的. 为此, 通常可考虑该积分的 Laplace 近似.

引理 1.1.2 (Laplace 近似)　记 $I(\lambda)=\int_{\Theta} b(y)\mathrm{e}^{\lambda t(y)}\mathrm{d}y$. 如果函数 $b(y)(>0)$ 在 $y=\mu\in\Theta$ 处连续, 函数 $t(y)$ 二次可导且在 $y=\mu\in\Theta$ 处达到最大值, $K(\mu)=-\ddot{t}(\mu)>0$, 则当 $\lambda\to\infty$ 时, 有

$$I(\lambda)\sim\sqrt{\frac{2\pi}{\lambda K(\mu)}}b(\mu)\mathrm{e}^{\lambda t(\mu)}.$$

根据积分的 Laplace 近似, (1.1.14) 式中的正态化常数 $a_0(\mu;\sigma^2)$ 有如下结论.

定理 1.1.1 设 $d(.;.)$ 为定义在 $C\times\Theta$ 上的已知正则单位偏差度函数 (但不一定为某一再生散度分布的单位偏差度), 正态化常数 $a_0(\mu;\sigma^2)$ 的定义见 (1.1.14), 则当 $\sigma^2\to 0$ 时, 有

$$a_0(\mu;\sigma^2)\sim(2\pi\sigma^2)^{-1/2}.$$

证明 在引理 1.1.2 中若取 $\lambda=1/\sigma^2$, $b(y)=V^{-1/2}(y)$, $t(y)=-d(y;\mu)/2$, 则比较引理 1.1.2 和 (1.1.14) 式的积分易得 $a_0(\mu;\sigma^2)^{-1}=I(\lambda)$. 由正则单位偏差度函数的性质知 $b(y)=V^{-1/2}(y)$ 关于 y 是连续函数, 且 $t(y)=-d(y;\mu)/2$ 在 $y=\mu$ 处达到最大值. 又由引理 1.1.1 的第二式知 $K(\mu)=-\ddot{t}(\mu)=1/V(\mu)$. 从而由引理 1.1.2 和正则单位偏差度函数的性质可得 $a_0(\mu,\sigma^2)^{-1}=I(\lambda)\sim\sqrt{2\pi\sigma^2}$, 因此定理 1.1.1 的结论成立. ∎

基于上述鞍点逼近可得

定理 1.1.2 假设随机变量 $Y\sim\mathrm{RDF}(\mu_0+\sigma\mu,\ \sigma^2)$, 若 $\sigma^2\to 0$ 时 $\sigma a(y;\sigma^2)$ 关于 y 在 Θ 的紧子集上一致收敛于 $\{2\pi V(y)\}^{-1/2}$, 则当 $\sigma^2\to 0$ 时有

$$\frac{Y-\mu_0}{\sigma}\xrightarrow{\mathcal{L}}N(\mu,V(\mu_0)),$$

其中"$\xrightarrow{\mathcal{L}}$" 表示依分布收敛.

证明 由 (1.1.9) 式以及引理 1.1.1 可知, $d(\mu_0+x\delta;\mu_0+m\delta)$ 在 (μ_0,μ_0) 处的二阶 Taylor 展开可表示为

$$d(\mu_0+x\delta;\mu_0+m\delta)=\frac{\delta^2}{V(\mu_0)}(x-m)^2+o(\delta^2). \tag{1.1.15}$$

因为随机变量 $Y\sim\mathrm{RDF}(\mu_0+\sigma\mu,\sigma^2)$, 所以, 由 (1.1.1) 式得 Y 的概率密度函数为 $p_y(y;\mu_0+\sigma\mu,\sigma^2)$. 记 $Z=(Y-\mu_0)/\sigma$, 则由随机变量函数的概率密度公式以及 $V(y)$ 的连续性、$\sigma a(y;\sigma^2)$ 的一致收敛性和 (1.1.15) 式可得

$$\begin{aligned}p_z(z;\mu,\sigma^2)&=\sigma p_y(\mu_0+\sigma z;\mu_0+\sigma\mu,\sigma^2)\\&=\sigma a(\mu_0+\sigma z;\sigma^2)\exp\Big\{-\frac{1}{2\sigma^2}d(\mu_0+\sigma z;\mu_0+\sigma\mu)\Big\}\\&=\sigma a(\mu_0+\sigma z;\sigma^2)\exp\Big\{-\frac{(z-\mu)^2}{2V(\mu_0)}+o(1)\Big\}\\&\sim\{2\pi V(\mu_0)\}^{-1/2}\exp\Big\{-\frac{1}{2V(\mu_0)}(z-\mu)^2\Big\}.\end{aligned}$$

上式表明随机变量 Z 依分布收敛于正态分布 $N(\mu,V(\mu_0))$. ∎

推论 1　假设随机变量 $Y \sim \mathrm{RDF}(\mu,\ \sigma^2)$, 其他条件与定理 1.1.2 相同, 则当 $\sigma^2 \to 0$ 时有

$$\frac{Y-\mu}{\sigma} \xrightarrow{\mathcal{L}} N(0,V(\mu)) \quad 或 \quad Y \xrightarrow{\mathcal{L}} N(\mu,\sigma^2 V(\mu)).$$

证明　在定理 1.1.2 中取 $\mu_0=\mu;\ \mu=0$ 即可得到以上第一式, 由第一式即可得到第二式. ∎

推论 2　假设随机变量 Y 的概率密度函数为正态化鞍点逼近 (1.1.13) 式, 即 $Y \sim p_0(y;\mu,\sigma^2)$, 则当 $\sigma^2 \to 0$ 时有

$$\frac{Y-\mu}{\sigma} \xrightarrow{\mathcal{L}} N(0,V(\mu)).$$

证明　若 $Y \sim p_0(y;\mu,\sigma^2)$, 则相当于在定理 1.1.2 中 $a(y;\sigma^2)=a_0(\mu;\sigma^2)V^{-1/2}(y)$, 而由定理 1.1.1 可知, $\sigma a_0(\mu;\sigma^2)\to(2\pi)^{-1/2}$(当 $\sigma^2 \to 0$ 时). 因此当 $\sigma^2 \to 0$ 时有 $\sigma a(y;\sigma^2)\to\{2\pi V(y)\}^{-1/2}$, 即 $p_0(y;\mu,\sigma^2)$ 满足定理 1.1.2 的条件. 再取 $\mu_0=\mu;\ \mu=0$ 即可得到上式. ∎

1.1.2　非线性再生散度模型

正如 1.1.1 节所指出, 再生散度分布族是指数族分布的直接推广和发展, 而广义线性模型的特点就是以指数族分布为其随机误差的“线性模型”. 本节将讨论以 RDF 为随机误差的广义非线性模型 —— 非线性再生散度模型.

对于数据集 $\{(\boldsymbol{x}_i,y_i):i=1,\cdots,n\}$, 其中 y_i 为响应变量 Y_i 的观察值, $\boldsymbol{x}_i=(x_{i1},\cdots,x_{iq})^{\mathrm{T}}$ 为 q 个已知固定的解释变量, 如果

(1) 存在一个严增可微的函数 g, 使得

$$\eta_i=g(\mu_i)=f(\boldsymbol{x}_i,\boldsymbol{\beta}),\quad i=1,\cdots,n; \tag{1.1.16}$$

(2) $Y_1,\cdots,Y_n$ 相互独立, 且 $Y_i \sim \mathrm{RDF}(\mu_i,\sigma^2)$, 即 Y_i 的概率密度函数为

$$p(y_i;\mu_i,\sigma^2)=a(y_i;\sigma^2)\exp\Big\{-\frac{1}{2\sigma^2}d(y_i;\mu_i)\Big\},\quad y_i\in\mathcal{C}, \tag{1.1.17}$$

其中, 函数 $g(.)$ 常称为“联系函数” (link function), $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_p)^{\mathrm{T}}$ $(p<n)$ 是定义在子集 $\mathcal{B}\subseteq\mathcal{R}^p$ 上的有兴趣未知参数 (即回归系数), $f(\cdot,\cdot)$ 是某一个已知函数, 则称满足 (1.1.16) 和 (1.1.17) 的模型为非线性再生散度模型 (nonlinear reproductive dispersion model), 并简记为 $\boldsymbol{Y}\sim\mathrm{NRDM}(\boldsymbol{\mu},\sigma^2)$, 其中 $\boldsymbol{Y}=(Y_1,\cdots,Y_n)^{\mathrm{T}}$, $\boldsymbol{\mu}=(\mu_1,\cdots,\mu_n)^{\mathrm{T}}$. 根据 (1.1.16) 式, μ_i 可反解为

$$\mu_i=g^{-1}\circ f(\boldsymbol{x}_i,\boldsymbol{\beta})\triangleq\mu_i(\boldsymbol{x}_i,\boldsymbol{\beta})\triangleq\mu_i(\boldsymbol{\beta}). \tag{1.1.18}$$

由该式代入 (1.1.17) 式可知, $p(y_i;\mu_i,\sigma^2)=p(y_i;\mu_i(\boldsymbol{\beta}),\sigma^2)$ 为 $\boldsymbol{\beta}$ 和 σ^2 的函数.

由前面的讨论知, 若 (1.1.17) 为指数族分布, 即 $d(y_i;\mu_i)=d'(y_i;\mu_i)-d'(y_i;y_i)$, 其中 $d'(y_i;\mu_i)=-2\{y_i\theta_i-b(\theta_i)-c(y_i)\}, \theta_i=\dot{b}^{-1}(\mu_i)$ 且 $a(y_i;\sigma^2)=\exp\{-\frac{1}{2}s(y_i,\sigma^{-2})\}$, 其中 $s(y_i,\sigma^{-2})$ 为某一特定的已知函数, 则非线性再生散度模型即为指数族非线性模型 (Jorgensen 1987, Cordeiro and McCullagh 1991, Cordeiro and Paula 1989b, Wei 1998); 更进一步, 若假设 $\eta_i=g(\mu_i)=\mu_i=f(\boldsymbol{x}_i,\boldsymbol{\beta})=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, 则由 (1.1.16) 和 (1.1.17) 定义的非线性再生散度模型即为典则联系的广义线性模型 (Nelder and Wedderburn 1972). 特别地, 如果 $d(y_i;\mu_i)=-2\{y_i\mu_i-\mu_i^2/2-y_i^2/2\}$, $a(y_i;\sigma^2)=1/\sqrt{2\pi\sigma^2}$ 且 $g(\mu_i)=\mu_i$, 则由 (1.1.16) 和 (1.1.17) 定义的非线性再生散度模型即为正态非线性回归模型 (Ratkowsky 1983, 1990; 韦博成 1989). 如果 $a(y_i;\sigma^2)=a(\sigma^2)b(y_i)$ 且 $g(\mu_i)=\mu_i=f(\boldsymbol{x}_i,\boldsymbol{\beta})=\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}$, 则由 (1.1.16) 和 (1.1.17) 定义的非线性再生散度模型即为 Paula (1996) 曾讨论过的正常散度模型 (proper dispersion models). 由此可知, 非线性再生散度模型是指数族非线性模型、广义线性模型、正态非线性回归等模型的直接推广和进一步发展.

1.2 非线性再生散度模型的极大似然估计

为了对非线性再生散度模型作统计推断, 我们假设

条件 A (1) $d(y;\mu)$ 关于 $(y;\mu)$ 在空间 $\mathcal{C}\times\Theta$ 上存在三阶连续偏导数.

(2) 对任意 $\boldsymbol{x}_i\in\chi, \boldsymbol{\beta}\in\mathcal{B}$, $\mu_i=\mu_i(\boldsymbol{\beta})=g^{-1}\circ f(\boldsymbol{x}_i,\boldsymbol{\beta})\in\Theta$.

(3) χ 是定义在 $\mathcal{R}^q$ 上的紧子集, $\mathcal{B}$ 是定义在 $\mathcal{R}^p$ 上的凸的开子集.

(4) $\boldsymbol{\beta}_0$ 是 $\boldsymbol{\beta}$ 的未知真参数且 $\boldsymbol{\beta}_0$ 为 $\mathcal{B}$ 的内点.

条件 B (1) 对于密度函数 $p(y_i;\mu_i,\sigma^2)$ 的积分, 可以关于参数在积分号下求导数; 因而有 $\mathrm{E}(\dot{d}(\mu_i))=0, \mathrm{E}(\dot{d}(\mu_i))^2=2\sigma^2\mathrm{E}(\ddot{d}(\mu_i)), \mu_i\in\Theta, i=1,\cdots,n$, 其中 $\dot{d}(\mu_i)=\partial d\mu_i/\partial\mu_i, \ddot{d}(\mu_i)=\partial^2 d(\mu_i)/\partial\mu_i^2, d(\mu_i)\triangleq d(\mu_i(\beta))=d(y_i;\mu_i(\beta))$.

因为对积分 $\int p(y_i;\phi_i,\mu_i)dy_i=\int\exp\{l(\phi_i,\mu_i)\}dy_i=1$ 关于 μ_i 求导数可得

$$\int\exp\{l\phi_i,\mu_i)\}\Big\{\frac{1}{2}[\phi_i\dot{d}(y_i,\mu_i)]\Big\}dy_i=0.$$

由此即可得到第一式. 上式再次关于 μ_i 求导数可得

$$\int\exp\{l(\phi_i,\mu_i)\}\Big\{-\frac{1}{2}[\phi_i\dot{t}(y_i,\mu_i)]\Big\}^2dy_i+\int\exp\{l(\phi_i,\mu_i)\}\Big\{-\frac{1}{2}[\phi_i\ddot{d}(y_i,\mu_i)]\Big\}dy_i=0.$$

由此即可得到第二式.

(2) $f(\boldsymbol{x}_i,\boldsymbol{\beta})$ 作为 $\boldsymbol{\beta}$ 的函数至少存在三阶连续的偏导数. $f(\boldsymbol{x}_i,\boldsymbol{\beta})$ 及其关于 $\boldsymbol{x}_i$ 和 $\boldsymbol{\beta}$ 的所有导数在空间 $\chi\times\mathcal{B}$ 上连续.