DOI:10.13860/j.cnki.sltj.2002.05.013

文章编号:1002-1566(2002)05-0054-08

广义线性模型(一)

陈希孺

(中国科学院研究生院,北京 100039)

摘 要:本讲座是广义线性模型这个题目的一个比较系统的介绍。主要分3部分:建模、统计分析与模型选择和诊断。写作时依据的主要参考资料是L. Fahrmeir等人的《Multivariate Statistical Modeling Based on Generalized Linear Models》。

关键词:广义线性模型;建模;统计分析;模型选择和诊断

中图分类号:O212 文献标识码:A

形式上,广义线性模型是常见的正态线性模型的直接推广(见本讲座§1.1,(一))。它可适用于连续数据和离散数据,特别是后者,如属性数据,计数数据。这在实用上,尤其是生物,医学和经济、社会数据的统计分析上,有重要的意义。本讲座是关于这个题目的一个比较系统的介绍。

广义线性模型的个别特例起源很早。Fisher在1919年曾用过它。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。1972年Nelder和Wedderburn在一篇论文中引进广义线性模型一词,自那前后以来研究工作逐渐增加。1983年McCullagh和Nelder出版了系统论述此专题的专著(见下)并于1989年再版,研究论文数以千计。

本讲座是应用取向,分3部分:建模、统计分析与模型选择和诊断。写作时依据的主要参考资料是L.Fahrmeir等《Multivariate Statistical Modeling Based on Generalized Linear Models》,Springer,1994,以及McCullagh等的《Generalized Linear Models》,1989年第2版,Chapman & Hill。此领域的专著一般都不涉及严格的数学推导。本讲座在建模过程及统计方法的导出等方面,力求在数学上交待清楚,但因性质所限,也不涉及一些非常繁琐的证明。对这方面有兴趣的读者应参阅有关的杂志论文,可从下面的论文入手:L. Fahrmeir等:Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models Ann. Statist,1985,342-368。

第一部分 建模

§1.1 一维广义线性回归

(一)定义

设有因变量Y,自变量x。Y为一维,x一般为多维。通常的线性回归有以下几个特征:

1. $E(Y)=\mu=z'(x)\beta$(线性,线性指对β,非X),$z(x)$为x的已知(向量)函数,z'表示转置(本讲义中"′"都表示转置,不是导数),$z'(x)$常简记为z'。

2. x,$z(x)$,Y都是取连续值的变量,如农作物的产量,人的身高体重之类。

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c

3. Y 的分布为正态,或接近正态之分布。

广义线性回归从以下几方面推广:

1. $E(Y)=\mu=h(z'\beta)$, h 为一严格单调,充分光滑的函数。h 已知,$g=h^{-1}$(h 的反函数)称为联系函数(link function)。有 $g(\mu)=z'\beta$。

2. x, $z(x)$, Y 可取连续或离散值,且在应用上更多见的情况为离散值,如{0, 1},{0, 1, 2……}等

*)例如,x 为1维,$z(x)$可以是$(1, x)$, $(1, x, x^2)$, $(1, e^x)$等。若 $x=(x_1, x_2)'$, $z(x)$可以是$(1, x_1, x_2)'$, $(1, x_1, x_2, x_1^2, x_2^2, x_1x_2)$等。

3. Y 的分布属于指数型,正态是其一特例。这里考虑的 Y 为一维,故属于一维指数型。其形式为:

$$c(y)\exp(\theta y-b(\theta))d\mu(y),\ \theta\in\Theta\text{(参数空间)} \tag{1.1}$$

θ 为参数,称为自然参数。$b(\theta)$为 θ 的已知函数。μ 为一测度(不一定是概率测度),常见的有两种可能:

a. 当 Y 为连续变量时,$d\mu(y)$为 Lebesgue 测度:$d\mu(y)=dy$;

b. 当 Y 为离散变量时,Y 取有限个值 $a_1, \cdots, a_m$ 或可列个值 $a_1, a_2, \cdots$,这时

$$\mu(\{a_i\})=1,\ i=1, \cdots, m;\text{或 } \mu(\{a_i\})=1,\ i=1, 2\cdots \tag{1.2}$$

故

$$\int_c^d c(y)\exp(\theta y-b(\theta))dy=1,\text{一切 }\theta\in\Theta\text{(连续情况)} \tag{1.3}$$

$[c, d]$(或(c, d), $[c, d)$等)为 Y 的取值区间,可为$(-\infty, \infty)$, $(0, \infty)$, $(-\infty, 0)$或任何其它有限或无限区间。

或

$$\sum_i c(a_i)\exp(\theta a_i-b(\theta))=1,\text{一切 }\theta\in\Theta\text{(离散情况)} \tag{1.4}$$

在这一情况,$c(a_i)\exp(\theta a_i-b(\theta))$为 Y 取 a_i 的概率(参数为 θ 时)

若 Y 有分布(1.1),则

$$EY=b(\theta)(=db(\theta)/d\theta),\ Var(Y)=\overset{\approx}{b}(\theta)(=d^2b(\theta)/d\theta^2) \tag{1.5}$$

事实上,在(1.3)两边对 θ 求导,有

$$\int_c^d (y-b(\theta))c(y)\exp(\theta y-b(\theta))dy=0 \tag{1.6}$$

注意到在(1.3),以及 $E(Y)=\int_c^d yc(y)\exp(\theta y-b(\theta))dy$,得(1.5)第一式——此处及以下在积分号下求导的合法性没有问题。再在(1.6)时两边对 θ 求导,有

$$\int_c^d (y-b(\theta))^2c(y)\exp(\theta y-b(\theta))dy-\overset{\approx}{b}(\theta)\int_c^d c(y)\exp(\theta y-b(\theta))dy=0$$

此时左边第一项为 $Var(Y)$,第二项为 $\overset{\approx}{b}(\theta)$。故得(1.5)第2式。

例 1.1 研究一些因素(自变量)对"剖腹产后是否有感染"的影响。

$$Y=\begin{cases}1,\text{有感染}\\0,\text{无感染}\end{cases}\qquad x=(x_{(1)}, x_{(2)}, x_{(3)}):$$

$$x_{(1)}=\begin{cases}1,\text{剖腹事先未计划}\\0,\text{剖腹事先计划}\end{cases}$$

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c

$$x_{(2)}=\begin{cases}1, \text{服用抗生素}\\ 0, \text{不服用}\end{cases}$$

$$x_{(3)}=\begin{cases}1, \text{有危险因子(如产妇有高血压, 糖尿病之类)}\\ 0, \text{无}\end{cases}$$

记 $\pi=P(Y=1)$。有(对 $y=1,0$):

$$P(Y=y)=\pi^y(1-\pi)^{1-y}=(1-\pi)(\frac{\pi}{1-\pi})^y=(1-\pi)\exp(y\log\frac{\pi}{1-\pi}) \quad (1.7)$$

令 $\theta=\log\frac{\pi}{1-\pi}$, 则 $1-\pi=\frac{1}{1+e^\theta}$, 而(1.7)可写为

$$P(Y=y)=\exp(\theta y-\log(1+e^\theta)), \quad -\infty<\theta<\infty \quad (1.8)$$

此相当于(1.4)中$\{a_1, a_2\}=\{0, 1\}$, $b(\theta)=\log(1+e^\theta)$, $c(y)=1$ 的情况。有

$b(\theta)=e^\theta/(1+e^\theta)=\pi \quad (=Ey)$

$\overset{\approx}{b}(\theta)=e^\theta/(1+e^\theta)^2=\pi(1-\pi) \quad (=Var(y))$

与公式(1.5)一致。

此例中 z 就取为 x, 引进记号

$$\eta=z'\beta \quad (1.9)$$

观察了 n 位产妇, 第 i 位的 Y 值记为 y_i, z 值记为 z_i(即 $x_{1i}, x_{2i}, x_{3i})'$), $\eta_i=z_i'\beta$, $i=1, \cdots, n$。其 π, θ 值分别为 π_i, θ_i($\theta_i=\log\frac{\pi_i}{1-\pi_i}$)。并引入了联系函数 $g(\pi_i)=\eta_i$(注意 $\mu_i=E(y_i)=\pi_i$), 或 $\pi_i=h(\eta_i)$ ($h=g^{-1}$), 则 $\theta_i=\log\frac{h(\eta_i)}{1-h(\eta_i)}$。代入(1.8)中, 得$(y_1, \cdots, y_n)$的联合概率函数

$$\exp\{\sum_{i=1}^n y_i\log\frac{h(\eta_i)}{1-h(\eta_i)}+\sum_{i=1}^n\log(1-h(\eta_i))\} \quad (1.10)$$

它通过 $\eta_1, \cdots, \eta_n$ 而依赖 β。利用它可对 β 进行统计推断。例如, 判断所提 3 个因素对“产后感染”发生概率的影响。推断方法的讨论见后。

例 1.2 研究两种化学物质 TNF 与 IFN 对引发细胞癌变的影响。$x=(x_1, x_2)'$:

$x_1=$TNF 的剂量 $(0, 1, 2, \cdots)$

$x_2=$IFN 的剂量 $(0, 1, 2, \cdots)$

$Y=$观察到的细胞变异数 $(0, 1, 2\cdots)$

决定取 Poisson 分布作为 Y 的分布:

$$P(Y=y)=\frac{1}{y!}e^{-\lambda}\lambda^y=\frac{1}{y!}\exp(y\log\lambda-\lambda), \lambda>0 \quad (1.11)$$

令 $\theta=\log\lambda$, 有

$$P(Y=y)=\frac{1}{y!}\exp(\theta y-e^\theta), y=0, 1, 2\cdots, \quad -\infty<\theta<\infty \quad (1.12)$$

此相当于(1.4)中的 $c(y)=1/y!$, $b(\theta)=e^\theta$, $a_i=i$, $i=0, 1, 2, \cdots$有

$$b(\theta)=\overset{\approx}{b}(\theta)=e^\theta=\lambda=E(Y)=Var(Y)$$

与公式(1.5)一致。

引进联系函数 $\eta=g(\mu)=g(\lambda)$, (μ 总用于记 $E(Y)$), 或 $\lambda=e^\theta=h(\eta)$ ($h=g^{-1}$), 或 $\theta=\log h(\eta)$。设作了 n 次观察, 第 i 次有 y_i, z_i, 而 λ, θ 值分别为 λ_i, θ_i, 则$(y_1, \cdots, y_n)$的联合概率函数为($\eta_i=z_i\beta$)

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c

$$(y_1!\cdots y_n!)^{-1}\exp\left(\sum_{i=1}^{n} y_i \log h(\eta_i) - \sum_{i=1}^{n} h(\eta_i)\right) \tag{1.13}$$

它通过 $\eta_1, \cdots, \eta_n$ 而依赖 β。利用它可对 β 进行统计推断。以判断两种物质对引发细胞变异的作用如何。

例 **1.3** Y 是某种极值(水文、地震、材料断裂强度之类),采用 Gamma(Γ)分布去描述:Y 有密度

$$f(y \mid \mu, \nu) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1}\exp\left(-\frac{\nu}{\mu}y\right) I(y>0),\ \mu>0,\ \nu>0 \tag{1.4}$$

($\Gamma(s)=\int_0^{\infty} e^{-t}t^{s-1}dt$, $s>0$, $\Gamma(n)=(n-1)!$, $n=1,2\cdots$, $\Gamma(s+1)=s\,\Gamma(s)$)

有

$$E(Y)=\mu \quad Var(Y)=\mu^2/\nu \tag{1.15}$$

此处关心的参数为 μ,而 ν 视为冗余参数。在讨论中,凡有冗余参数(如此处的 ν),则视为已知,当它确为未知时,则必须从样本可以估计,以其估计值代替而视为已知。

令 $\theta=-\frac{\nu}{\mu}$,将(1.14)表为

$$\frac{1}{\Gamma(\nu)}y^{\nu-1}\exp(\theta y-(-\nu\log(-\theta)))I(y>0),\ -\infty<\theta<0 \tag{1.16}$$

得 $E(Y)=b(\theta)=-\nu/\theta=\mu$, $Var(Y)=\overset{\approx}{b}(\theta)=-\nu/\theta^2=\mu^2/\nu$,与公式(1.5)一致。

引进联系函数 g, $h=g^{-1}$,则 $-\nu/\theta=\mu=h(\eta)$,而 $\theta=-\nu/h(\eta)$。若有样本 $y_1, \cdots, y_n$, y_i 相应的 Z 值为 Z_i, $\eta_i=Z_i'\beta$,相应的 θ 值为 $\theta_i=-\nu/h(\eta_i)$。则 $(y_1, \cdots y_n)$ 的联合密度为

$$\prod_{i=1}^{n}\frac{1}{\Gamma(\nu)}y_i^{\nu-1}\exp\left(-\sum_{i=1}^{n}\frac{\nu}{h(\eta_i)}y_i+\sum_{i=1}^{n}\nu\log\left(\frac{\nu}{h(\eta_i)}\right)\right) \tag{1.17}$$

它通过 $\eta_1, \cdots, \eta_n$ 依赖于 β,利用它对 β 进行统计推断。

提醒两点:1. 当 Y 为 1 维时,只能有 1 个未知参数(此例中为 μ)。若有多个参数,剩下的为冗余,它必须已知或可由样本估计,即以估计值为已知值。2. 在各次观察中冗余参数不变。如在此例中,相应 y_i 的 μ 值可变,为 μ_i(与此相应,θ 值则为 $\theta_i=-\nu/\mu_i$),但 ν 则不随 i 变化。

例 **1.4** Y 有正态分布 $N(\mu, \sigma^2)$,密度为

$$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{y^2}{2\sigma^2}}\exp\left(\frac{\mu}{\sigma^2}y-\frac{1}{2}\left(\frac{\mu}{\sigma}\right)^2\right),\ \sigma \text{ 已为冗余参数,已知} \tag{1.18}$$

令 $\theta=\mu/\sigma^2$,则 $\mu^2/\sigma^2=\sigma^2\theta^2$。而(1.18)可写为

$$\frac{1}{\sqrt{2\pi}\sigma}e^{\frac{-y^2}{2\sigma^2}}\exp\left(\theta y-\frac{\sigma^2}{2}\theta^2\right) \tag{1.19}$$

此相当于(1.3)的 $c=-\infty$, $d=\infty$, $c(y)=\frac{1}{\sqrt{2\pi}\sigma}e^{-y^2/2\sigma^2}$, $b(\theta)=\frac{\sigma^2}{2}\theta^2$,有 $b(\theta)=\sigma^2\theta=\mu=E(Y)$, $\overset{\approx}{b}(\theta)=Var(Y)$,与(1.5)一致。

如取联系函数为 $\mu=Z'\beta$,则与通常线性回归一致。若取其它联系函数则属于广义线性回归。

两点注意:

1. 在单参数指数族中,方差是均值的函数(当然反过来也成立):因为方差 $\overset{\approx}{b}(\theta)>0$,故 $b(\theta)$ 严格上升,因此有反函数 b^{-1}。故由 $\theta=b^{-1}(E(Y))$ 有

$$Var(Y)=\overset{\approx}{b}(\theta)=\overset{\approx}{b}(b^{-1}(E(Y))) \tag{1.20}$$

在有些实际问题中,数据显示均值方差之间的关系不符合(1.20)。这时就不可能使用单参指数族的模型。在 Γ 和正态分布的例中包含了一个冗余参数,调整它的值有时可以解决上

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c

述问题。

2. 如在例 1.1 这类例子中, 自变量值的可能组合数很少(在例 1.1 中只有 $2^3=8$ 个)。这时样本呈现分组的状态。设 $y_1, \cdots, y_m$ 是同一 x 值下的 Y 样本。这时往往用一个样本

$$Y=\sum_{i=1}^{m} y_i, \text{ 或 } \bar{Y}=\sum_{i=1}^{m} y_i/m$$

取代 $y_1, \cdots, y_m$(即我们只见到 Y 或 $\bar{Y}$ 和 m, 而不一定能见到原始记录 $y_1, \cdots y_m$)这样做并无损失, 因: a. Y 或($\bar{Y}$)是充分统计量, 因此无信息丧失; b. Y 或($\bar{Y}$)仍为指数型分布: 当 y_i 有分布(1.1)时:

$$Y \text{ 有分布 } c_1(Y)\exp(\theta Y - mb(\theta))d\mu_1(Y) \tag{1.21}$$

$$\bar{Y} \text{ 有分布 } c_2(\bar{Y})\exp(\theta m\bar{Y} - mb(\theta))d\mu_2(\bar{Y}) \tag{1.22}$$

(1.21), (1.22)中的 c_1, c_2 及 μ_1, μ_2 可以与(1.1)中的 c 及 μ 不同, 但不失为指数型分布形状。其中, (1.22)非标准形式。引进新参数 $\theta=m\theta$, 将(1.22)写为

$$c_2(\bar{Y})\exp(\theta\bar{Y} - mb(\frac{\theta}{m}))d\mu_1(Y) \tag{1.23}$$

则成为标准形式。记 $b_1(\theta)=mb(\theta/m)$, 有

$E(\bar{Y})=db_1(\theta)/d\theta=b(\theta/m)=b(\theta)$ (回忆 $b(\theta)=db(\theta)/d\theta$)

$Var(\bar{Y})=d^2b_1(\theta)/d\theta^2=m^{-1}\overset{\approx}{b}(\theta/m)=m^{-1}\overset{\approx}{b}(\theta)$

即 $E(\bar{Y})=E(y_i)$, $Var(\bar{Y})=Var(y_i)/m$, 与常见公式符合

以上的讨论是在 $y_1, \cdots y_m$ 为 iid. 的条件下进行的, 实际问题中这可能不完全成立, 如: a. 同一组 x 值上所观察的 y 值有正相关性。b. 有一些未包含在 x 中的因素(问题中未予考虑或尚未认知)对各观察值的影响不同, 而使 $y_1, \cdots y_m$ 不同分布。这两点总的影响是加大 $Y=\sum_{1}^{m} y_i$ 的方差, 即比按公式算的 $m\overset{\approx}{b}(\theta)$ 大, 称为"超散布性"(overdispersion)。这个问题的处理见后。

(二)哑(或虚)变量(dummy variable)

设有一个因素(自变量之一)有 k 个"状态", 我们固然可以用数字 1, …, k 来标识它, 但不可用于计算, 因为它们无数量意义。例如农业试验中, 品种是一个因素。有 k 类种子, 解决的办法是引进哑变量 $x_1, \cdots, x_q$, $q=k-1$:

$$x_j=\begin{cases}1, \text{若样品处在状态 j, (该试验用种子 j)}\\ 0, \text{其他}\end{cases} \tag{1.24}$$

$$j = 1, \cdots, q$$

故 $x_1=\cdots=x_q=0$, 当样品处在状态 k (1.25)

设这个试验只包括"品种"这一个因素, 模型为

$$E(Y) = \beta_0+\beta_1 x_1+ \cdots+\beta_q x_q \text{ (Y 为产量)} \tag{1.26}$$

则可见

$$E(Y \mid \text{状态 } j) = \beta_0+\beta_j, j = 1, \cdots, q; \tag{1.27}$$

$$E(Y \mid \text{状态 k}) = \beta_0$$

故(1.24)这种取哑变量法, 是以状态 k 作为标准, 而 β_j 衡量(在产量上)状态 j 超出状态 k 之值。

另一种取法是($j=1, \cdots, q$)

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c

$$x_j=\begin{cases}1,\text{若样品处在状态}\quad j\\ -1,\text{若样品处在状态}\quad k,\\ 0,\text{其他}\end{cases}\tag{1.28}$$

这时

$$x_1=\cdots=x_q=-1,\text{当样品处在状态 k}\tag{1.29}$$

因此按(1.26)式有

$$E(Y\mid\text{状态 }j)=\beta_0+\beta_j,\quad j=1,\cdots,q;\tag{1.30}$$
$$E(Y\mid\text{状态 k})=\beta_0-(\beta_1+\cdots+\beta_q)$$

于是 $\frac{1}{k}\sum_{j=1}^{k}E(Y\mid\text{状态 }j)=\beta_0$

故 β_0 为平均效应,而 $\beta_j(j\leqslant q)$衡量状态 j 效应超出平均之值。

(三)联系函数·自然联系函数

联系函数 g:$g(\mu)=\eta=z'\beta$, $\mu=E(Y)$,其反函数 h 也很常用。作为联系函数,g 必须严格单调且充分光滑,即有足够阶数的导数。

有一个特殊的联系函数,即

$$g=b^{-1}\text{或 }h=b(\text{回忆 }b(\theta)=db(\theta)/d\theta)\tag{1.31}$$

起着重要的作用。它称为自然联系函数,这时有

$$z'\beta=g(\mu)=g(b(\theta))=\theta\tag{1.32}$$

因此,指数型分布(1.1)中的自然参数,就是 $z'\beta$。这一重要关系式是“自然联系函数”这一名称的由来。其方便之处,目前我们可以看到一点:若有了样本 $y_1,\cdots y_n$,与 y_i 相对应的 z 值为 z_i,则$(y_1,\cdots y_n)$的联合密度为

$$\prod_{i=1}^{n}c(y_i)\cdot exp\left(\beta'\sum_{i=1}^{n}z_iy_i-\sum_{i=1}^{n}b(z_i'\beta)\right)$$

其形式比在其他联系函数下来得简单,其最重要的优点是:它使广义线性模型下统计推断的大样本理论更易处理。当然,在一个实际问题中选择联系函数,主要应依据问题本身的情况。

例 1.1(续),因本例 $\pi=\mu$,自然联系函数由 $z'\beta=\theta=log\ \frac{\pi}{1-\pi}$确定,即

$$g(t)=\log\frac{t}{1-t}\text{或}h(t)=\frac{e^t}{1+e^t}(=b(t))$$
$$\pi=e^{z'\beta}\backslash(1+e^{z'\beta})\tag{1.33}$$

这就是很知名很重要的logit(或logistic)模型。注意(1.33)右边之值总在(0,1)内,符合 π 作为概率的要求。

一般,$\pi=h(z'\beta)$。故 h 应满足 $0<h<1$。若 h 为严增,则 $h(-\infty)$一般应为 0,$h(\infty)$一般应为 1,这样 π 可取(0,1)内任何值(如果问题的性质限定了 π 只能取(0,1)内某一个子区间中的值,则另当别论)。因此,h 应为一分布函数,有几个选择在实用中用到:

$$h_1(t)=\Phi(t)\quad(N(0,1)\text{的分布}):\text{联系函数 }g=\Phi^{-1}\tag{1.34}$$

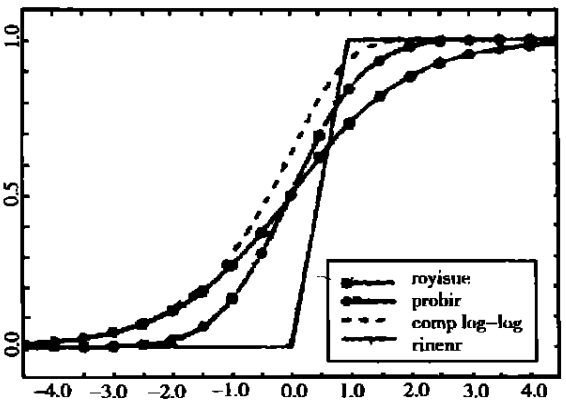
称为 probit 模型:

$$h_2(t)=1-\exp(-e^t):\text{联系函数 }g(\pi)=\log(-\log(1-\pi))\tag{1.35}$$

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c

其联系函数的形式使之有 log－log 模型的名称。

(1.33)－(1.35)这 3 个 $h(t)$的图形如图(1.1)所示。从图形上看,三者颇有些差距,尤其是 log－log 与其它二者的差距。但我们要注意一点:选择 $h(t)$或选择 $h(\frac{t-\alpha}{\sigma})$,使用极大似然法所作的统计分析,结果一致。这里 $\sigma>0$ 和 α 为常数。理由如下:



图(1.1)

令 $h_1(t)=h(\frac{t-\alpha}{\sigma})$,现在我们有两个模型:

$$\pi=h(\beta_0+z'\beta) \qquad (1.36)$$

$$\pi=h_1(\beta_0^*+z'\beta^*)=h\left(\frac{\beta_0^*-\alpha}{\sigma}+z'\beta^*/\sigma\right) \qquad (1.37)$$

此处把常数项 β_0 明确标出(在一般理论探讨中,可设 β_0 已吸入 β 内:令 $z=\begin{pmatrix}1\\ z\end{pmatrix}$, $\beta\begin{pmatrix}\beta_0\\ \beta\end{pmatrix}$ 即可)。当有了样本(z_i, y_i), $1\leqslant i\leqslant n$ 时,按(1.35)和(1.36),分别得出两个联合密度

$$f=\prod_{i=1}^{n}(h(\beta_0+z_i'\beta))^{y_i}(1-h(\beta_0+z_i'\beta))^{1-y_i}$$

$$f_1=\prod_{i=1}^{n}(h(\beta_0+z_i'\beta))^{y_i}(1-h(\beta_0+z_i'\beta))^{1-y_i}$$

其中 $\beta_0=\frac{\beta_0^*-\alpha}{\sigma}$, $\beta=\beta^*/\sigma$

用极大似然估计,即求 f, f_1 的极大值点。因 f, f_1 形式完全一致,故若以 β_0, β 记β_0, β 的极大似然估计,则 β_0, β 的极大似然估计也是β_0, β,因此 β_0^*, β^*的极大似然估计分别为:

$$\beta_0^*=\alpha+\sigma\beta_0,\ \beta^*=\sigma\beta,$$

以此代入(1.36)和(1.37),得出在这两模型下,π 的估计分别为:

按(1.36): $h(\beta_0+z'\beta)$

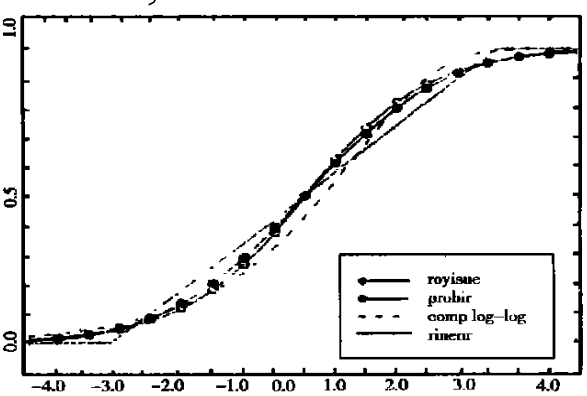
按(1.37): $h_1(\beta_0^*+z'\beta^*)=h\left(\frac{\alpha+\sigma\beta_0-\alpha}{\sigma}+z'\sigma\beta/\sigma\right)=h(\beta_0+z'\beta)$

二者完全一致,故选 h 或 h_1,不影响分析结果。

依这个结果,再考虑前面(1.33)－(1.35)定义的 h, h_1, h_2 之间的差距,我们意识到:这个差距中有一部分是由于"位置"与"刻度"的差异而来,并非真实的有实际意义的差距。因为,这 3 个分布的均值,方差分别为:

h 0, $\pi^2/3$; h_1 0, 1; h_2 －0.5772, $\pi^2/6$

我们看出,它们之间有差别,要调整到同一个数再比较。这样,把 3 者的均值,方差都



图(1.2)

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c

调整为0, 1。这意味着用 $h\left(\frac{\pi}{\sqrt{3}}t\right)$ 取代 h, h_1 不动, 而 $h_2(t)$ 用 $h_2(\pi t/\sqrt{2}-0.5772)$ 取代。经过取代后 3 个分布的形状如图(1.2)所示, 由图上看出, 其差距与图一相比有所接近, 尤其是 h 与 h_1。因为 h_1 较易计算, 故在实用上用得最多。

[参考文献]

[1] L. Fahrmeir.《Multivariate Statistical Modeling Based on Generalized Linear Models》〔M〕. New York, Springer-Verlag, 1994.

[2] McCullagh.《Generalized Linear Models》〔M〕. London/New York, Chapman & Hill, 1989 2nd edition.

[3] L. Fahrmeir. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models〔J〕. Ann. Statist, 1985, 342—368.

Generalized linear models

CHEN Xi-ru

(Graduate School of Chinese Academia of Science, Beijing 100039, China)

Abstract: This set of articles gives an introduction to generalized linear models. They can be divided into three parts: Model building, Statistical inference and Model diagnostics. The presentation is mainly based on L. Fahrmeir et al.《Multivariate Statistical Modeling Based on Generalized Linear Models》.

Key words: generalized linear models; model building; statistical inference; model diagnostics

上接第 36 页

[参考文献]

[1] 刘朝荣试验的设计与分析, 〔M〕武汉: 湖北科学技术出版社, 1990

[2] SAS/QC Software: Reference, Version 6, First〔M〕. SAS Institute Inc, 1991

[3] 高惠璇. SAS 系统 SAS/STAT 软件使用手册〔M〕. 北京: 中国统计出版社, 1997

[4] 盛骤. 概率论与数理统计〔M〕. 北京: 高等教育出版社, 1993.

[5] 李泽慧. 数理统计一基本概念与专题〔M〕. 兰州: 兰州大学出版社, 1991

How to construct a design of factorial experiments by SAS/QC

LI Qin-min, CHEN Zhi-min

(Management college, Shenzhen University, Shenzhen 518060, China)

Abstrat: In this paper, we introduced how to construct a factorial design using factex procedure in SAS/QC. The methods include full factorial designs, fractional factorial designs, mixed—level designs and latin square designs.

Key words: design of experiments; factor; level; effect; SAS/QC

?1994-2015 China Academic Journal Electronic Publishing House. All rights reserved. http://www.c