第八章：广义线性模型(GLMs)

形式上，广义线性模型是常见正态线性模型的直接推广。它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。在生物、医学以及经济社会领域数据的统计分析上有着重要意义。

广义线性模型起源与发展

- Fisher 在 1919 年曾用过。
- 最重要的 Logistic 模型，在 20 世纪四五时年代曾由 Berkson，Dyke 和 Patterson 等使用过。
- 1972 年 Nelder 和 Wedderburn 在一篇论文中引入广义线性模型(Generalized Linear Models，GLMs)一词
- 1983 年 McCullagh 和 Nelder 出版系统论述专著《*Generalized Linear Models*》，London：Chapman and Hall.

8.1 响应变量一维的广义线性模型

设响应变量y(一维)，协变量x(一般多维)，则通常线性回归模型有以下几个特征：

1．$Ey=\mu=x'\beta$（线性指对β而言）；

2．x,y通常取值连续；

3．y的分布为正态(或接近正态)。

广义线性回归模型从以下几方面推广：

1．$Ey=\mu=h(x'\beta)$，h为严格单调充分光滑已知函数。$g=h^{-1}$(h的反函数)称为联系函数(link function)。$g(\mu)=x'\beta$；

2．x,y可取连续或离散值，且在实际应用上更多见的为离散值，如$\{0,1\}$，$\{0,1,2\cdots\}$等；

3．y的分布属于指数型分布，正态是其特例。这里考虑y为一维的，故为一维指数型分布，其密度形式为：

$$f(y;\theta)=\exp\left[\frac{\theta y-b(\theta)}{\phi}+c(y,\phi)\right] \qquad (8.1.1)$$

$b(\cdot),c(\cdot,\cdot)$为已知函数。θ称为自然参数(natural parameter)；ϕ称为额外参数(additional parameter)或者散布参数 dispersion parameter。

定理 8.1.1：若y密度为(8.1.1)，则

$$Ey=\dot{b}(\theta),\quad Var(y)=\phi\ddot{b}(\theta),$$

其中$\dot{b}(\cdot),\ddot{b}(\cdot)$分别为一阶及二阶导数。

例 8.1.1：研究一些因素对“剖腹产后是否有感染”的影响。

$$y=\begin{cases}1,\text{有感染}\\0,\text{无感染}\end{cases},\quad x=(x_{(1)},x_{(2)},x_{(3)})'$$

$$x_{(1)}=\begin{cases}1,\text{剖腹事先未计划}\\0,\text{剖腹事先计划}\end{cases},\quad x_{(2)}=\begin{cases}1,\text{服用抗生素}\\0,\quad\text{不服用}\end{cases}$$

$$x_{(3)}=\begin{cases}1,\text{有危险因子(如有高血压、糖尿病等)}\\0,\quad\text{无风险因子}\end{cases}$$

记$\pi=P(y=1)$，y的密度为$\pi^y(1-\pi)^{1-y}$。令$\theta=\log\frac{\pi}{1-\pi}$，则$y$的密度可以写成(8.1.1)标准形式$\exp(\theta y-\log(1+e^\theta))$, $-\infty<\theta<\infty$。相当于$\phi=1, b(\theta)=\log(1+e^\theta)$。可以验证

$$\dot{b}(\theta)=\frac{e^\theta}{1+e^\theta}=\pi\quad(=Ey),$$

$$\phi\ddot{b}(\theta)=\frac{e^\theta}{(1+e^\theta)^2}=\pi(1-\pi)\quad(=Var(y))。$$

引入联系函数$g(\pi)=x'\beta$，现观察了n位产妇

得$(y_1, y_2 \cdots y_n)$联合密度(似然函数)

$$\exp\left[\sum_{i=1}^{n} y_i \log \frac{h(x_i'\beta)}{1-h(x_i'\beta)} + \sum_{i=1}^{n} \log(1-h(x_i'\beta))\right],$$

其中$h(\cdot) = g^{-1}(\cdot)$。通过似然函数可以对β进行统计推断。

例 8.1.2：研究两种化学物质 TNF 与 IFN 对引发细胞癌变的影响。

$x_{(1)} = TNF$的剂量$(0,1,2\cdots)$,

$x_{(2)} = IFN$的剂量$(0,1,2\cdots)$，$x = (x_{(1)}, x_{(2)})'$

$y =$观测到的细胞变异数$(0,1,2\cdots)$

取 Poisson 分布作为y的分布，密度为

$$\frac{1}{y!} e^{-\lambda} \lambda^y, \lambda > 0。$$

令$\theta = \log \lambda$，则y的密度可以写成(11.1.1)标准形式$\exp(\theta y - e^{\theta} - \log y!)$，$-\infty < \theta < \infty$。相当于$\phi = 1, b(\theta) = e^{\theta}$。可以验证$\dot{b}(\theta) = \phi \ddot{b}(\theta) = e^{\theta} = \lambda \quad (= Ey = Var(y))$。引进联系函数$g(\lambda) = x'\beta$，现作了$n$次观测，

得$(y_1, y_2 \cdots y_n)$联合密度(似然函数)

$$\exp\left[\sum_{i=1}^{n} y_i \log h(x_i'\beta) - \sum_{i=1}^{n} h(x_i'\beta) - \sum_{i=1}^{n} \log y_i!\right],$$

其中$h(\cdot) = g^{-1}(\cdot)$。通过似然函数可以对β进行统计推断。

在指数型分布中，方差是均值的函数，因为$\ddot{b}(\theta) > 0$，因此$\dot{b}(\theta)$严格增，有反函数$\dot{b}^{-1}(\theta)$，因此$\theta = \dot{b}^{-1}(Ey)$，从而$Var(y) = \phi \ddot{b}(\dot{b}^{-1}(Ey))$。在一些实际问题若均值方差之间的关系不符合上式，就不能使用该模型。

联系函数$g(\mu) = x'\beta$，$\mu = Ey$，其必须严单调且充分光滑，既有足够阶数的导数。有一个很特殊的联系函数，即取

$$g(\cdot) = \dot{b}^{-1}(\cdot)，\ (\text{或} h(\cdot) = \dot{b}(\cdot))$$

此时

$$x'\beta = g(\mu) = g(\dot{b}(\theta)) = \theta，$$

称$g(\cdot)$为 canonical link function. 其方便之处在于此时$(y_1, y_2, \cdots y_n)$联合密度(似然函数)为

$$\exp\left[\frac{\beta' \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} b(x_i'\beta)}{\phi} - \sum_{i=1}^{n} c(y_i, \phi)\right],$$

其形式简单，相对于取其它联系函数时的联合密度(似然函数)

$$\exp\left[\frac{\sum_{i=1}^{n} \theta_i y_i - \sum_{i=1}^{n} b(\theta_i)}{\phi} - \sum_{i=1}^{n} c(y_i, \phi)\right],$$

其中

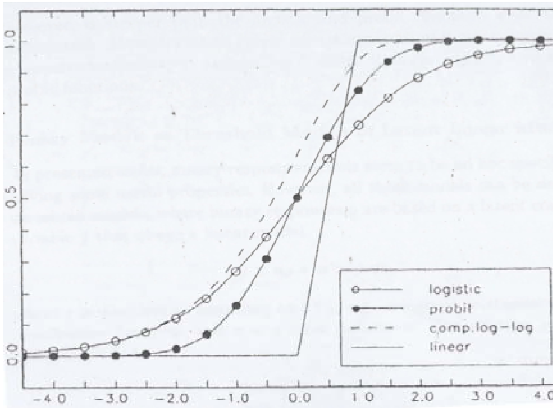
$$\theta_i = \theta_i(\beta) = \dot{b}^{-1}[h(x_i'\beta)]。$$

例 8.1.1(续)：本例$\mu = Ey = \pi$，canonical link 由 $x'\beta = \theta = \log \frac{\pi}{1-\pi}$确定，即取$g(t) = \log \frac{t}{1-t}$或$h(t) = \frac{e^t}{1+e^t} (= \dot{b}(t))$，此时$\pi = \frac{e^{x'\beta}}{1+e^{x'\beta}}$，这就得到知名的很重要的 logit(或 logistic)模型(注意要满足π作为概率的要求)。

一般，$\pi = h(x'\beta)$。故$h(\cdot)$应满足$0 < h(\cdot) < 1$。若严格增，通常$h(-\infty) = 0, h(\infty) = 1$。因此$h(\cdot)$通常为一分布函数，有几个选择在实用中常见。

取$h(t)=\Phi(t)$，即标准正态分布函数，此时联系函数$g(\pi)=\Phi^{-1}(\pi)$，称为 probit 模型；取 $h(t)=1-\exp(-e^t)$ ， 此 时 联 系 函 数 $g(\pi)=\log(-\log(1-\pi))$，称为 log-log 模型。

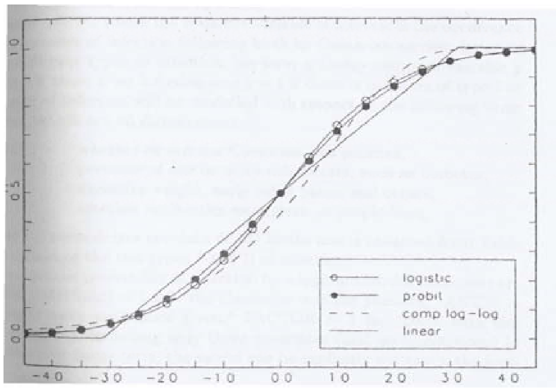
以上几个$h(t)$的图形如下：



相应分布的均值与方差

Response function F	Mean	Variance
linear	0.5	1/12
probit	0.0	1
logistic	0.0	$\pi^2/3$
compl. log-log	−0.5772	$\pi^2/6$

调整为均值为 0 方差为 1 后的图形



8.2 响应变量多维的广义线性模型

前面见过目标变量y取值的情况：

1. 连续取值，如人的身高；
2. 取离散值，但任有数量意义，如细胞变异数目；
3. 变量属性，但只有两个状态，如产后感染与否，用 0-1 描述，并无数量意义。

以上这些情况响应变量可以用单变量描述。

另一些情况，其目标变量y须取向量，如$y=(y_{(1)},y_{(2)})'=(身高，体重)'$，这种取连续向量响应变量，如用多元线性模型，得到多重线性回归模型。

除此之外，还有一种重要情况：响应变量y取k个状态，$k>2$。如在例 8.1.1 中若感染分两种类型，则每个产妇处于 3 个状态之一：无感染、I型感染、II型感染。当然可以用$y=0,1,2$来标识，因此可能会认为此例中响应变量y取值为0,1,2，非向量。**这种看法是错误的，因为此处0,1,2无数量意义，只是一种标签(label)**。

正 确 的 做 法 是 引 入 哑 变 量 $y_{(1)},\cdots,y_{(q)},q=k-1$：

$$y_{(j)}=\begin{cases}1,\text{处在}j\text{状态}\\0,\quad\text{其它}\end{cases},\quad j=1,\cdots,q.$$

而响应变量$y=(y_{(1)},\cdots,y_{(q)})'$，它共取$k$个值 $a_1=(1,0,\cdots,0)',\cdots,a_j=(0,\cdots,0,1,0,\cdots,0)',\cdots,$ $a_k=(0,0,\cdots,0)'$ ；

"$y=a_j$"$\Leftrightarrow$"处在状态j", $j=1,\cdots,k.$

一般，设响应变量y为q维，与响应变量 1 维类似，响应变量多维广义线性模型一个要素是y有指数型分布，其密度形式为：

$$f(y;\theta)=\exp\left[\frac{\theta' y-b(\theta)}{\phi}+c(y,\phi)\right],$$

其中$\theta=\left(\theta_{(1)},\cdots,\theta_{(q)}\right)'$，与模型中一些有实际意义的参数相关联。。

定理 8.2.1：

$$\mu\overset{\Delta}{=}Ey=\dot{b}(\theta)\overset{\Delta}{=}\frac{\partial b(\theta)}{\partial\theta}=\left(\frac{\partial b}{\partial\theta_{(1)}},\cdots,\frac{\partial b}{\partial\theta_{(q)}}\right)',$$

$$Cov(y)=\phi\ddot{b}(\theta),$$

其中$\ddot{b}(\theta)\overset{\Delta}{=}\dfrac{\partial^2 b(\theta)}{\partial\theta\partial\theta'}=\left[\dfrac{\partial^2 b(\theta)}{\partial\theta_{(i)}\partial\theta_{(j)}}\right]_{1\le i,j\le q.}$

另一要素是联系函数(link function)。设协变量x，它影响响应变量y的取值，由x产生$q\times p$的矩阵$Z=Z(x)$，$\beta_{p\times 1}$为未知参数，记$\eta=Z\beta$，称为线性预测子(linear predictor)，联系函数g是一个取值为R^q上的充分光滑函数，满足

$$\mu_1\ne\mu_2\Rightarrow g(\mu_1)\ne g(\mu_2)$$

$$g(\mu)=\eta=Z\beta。$$

记 $h=g^{-1}$，有 $\mu=Ey=h(Z\beta)$，$\theta=\dot{b}^{-1}(\mu)=\dot{b}^{-1}[h(Z\beta)]$。若取$g=\dot{b}^{-1}$，称为 canonical link，此时$\theta=Z\beta$。

若有样本$y_i,x_i,1\le i\le n$，相应有$\mu_i=Ey_i$，$Z_i,\eta_i=Z_i\beta$，得到$y_1,\cdots,y_n$联合密度(似然函数)：

$$\exp\left[\frac{\sum_{i=1}^n\theta_i' y_i-\sum_{i=1}^n b(\theta_i)}{\phi}-\sum_{i=1}^n c(y_i,\phi)\right],$$

其中

$$\theta_i=\theta_i(\beta)=\dot{b}^{-1}[h(Z_i\beta)]。$$

多项分布情形

再继续前面的讨论，响应变量y有$k(>2)$个状态取 $a_1,\cdots,a_k$ 这 k 个值，记取 a_j 的概率为 $\pi_{(j)},j=1,\cdots,q$（$q=k-1$），$\pi=(\pi_{(1)},\cdots,\pi_{(q)})'$(注意：取$a_k$的概率为$1-(\pi_{(1)}+\cdots+\pi_{(q)})\overset{\Delta}{=}1-|\pi|$)，此时 y 的分布密度为 $(1-|\pi|)^{1-|y|}\prod_{j=1}^q\pi_{(j)}^{y_{(j)}}$。令 $\theta_{(j)}=\log\dfrac{\pi_{(j)}}{1-|\pi|}$，则$\pi_j=\dfrac{e^{\theta_{(j)}}}{1+\sum_{i=1}^q e^{\theta_{(i)}}}$。

此时密度函数为

$$\frac{\prod_{i=1}^{q} e^{\theta_{(i)} y_{(i)}}}{1+\sum_{i=1}^{q} e^{\theta_{(i)}}} = \exp(\theta' y - b(\theta)),$$

其中

$$b(\theta) = \log(1+\sum_{i=1}^{q} e^{\theta_{(i)}})。$$

联系函数 $g = (g_{(1)}, \cdots, g_{(q)})'$，

$$g(\pi) = Z\beta。$$

多种选择问题

属性目标变量(响应变量)常见的一个情况是：人们面临有限种选择，可以自由选择其中之一。选择何种，则是根据本人及选择对象条件，依自己判断而定，响应变量是选择结果，而其余为协变量。例如购车者在购车时，目标变量可分为 4 个档次，10 万以下，10～20 万，20～50 万，50 万以上。根据自己财力，对车性能要求，各档次车的本身条件(均为协变量)作出自由选择。此类问题在社会调查和商务调查中有着重要意义，其目的在于哪些因素在决定人们的选择上起多大的作用。

一般根据“利益分析”来看各状态被选择的概率。即假定对一个具体的选择者而言，k 个状态各有一个“利益值”相关联，分别记为 $u_1, \cdots, u_k$。若选择者对 $u_1, \cdots, u_k$ 之值已完全了解，则选择状态 r，使得 $u_r = \max_{1 \le j \le k} u_j$。但一般 u 值并不完全确定，或选择者对其了解存在一定误差。因此，人们估量的利益值为 $U_1, \cdots, U_k$，其中 $U_j = u_j + e_j$，而 $e_1, \cdots, e_k$ 为独立同分布随机变量，密度为 $f(t)$，分布函数为 $F(t)$。

选择者根据“U 值最大”去选择状态，于是

$$\begin{aligned} P(r \ \ is \ \ selected) &= P(U_r > U_j \ \ for \ \ all \ \ j) \\ &= P(e_j < u_r - u_j + e_r \ \ for \ \ all \ \ j) \\ &= \int_{-\infty}^{\infty} \prod_{j=1, j \ne r}^{k} F(t + u_r - u_j) f(t) dt \end{aligned}$$

对 F 的不同选择，可得不同模型。

若 F 选为正态分布 $N(0,1)$ 的分布函数，则得到多维 probit 模型。此模型涉及多维正态分布函数的计算，实施较难。若选择 F 为极值分布 $F(t) = \exp(-e^{-t})$，其密度关于 0 不对称，但结果简单，此时可以得到

$$P(r \ \ is \ \ selected) = \frac{e^{u_r}}{\sum_{i=1}^{k} e^{u_i}}。$$

状态有序的情形

在有些问题中，目标状态有公认的优劣次序，如病情分 I, II, III 期，产品质量分不同等级等。注意，即使在这种场合，序号1,2⋯依然无数量意义。

大多数有序模型是按下述机制产生：有一个(或多个，此处考虑简单情形)明显或潜在变量 U 及门限 $-\infty = \theta_0 < \theta_1 < \cdots < \theta_{k-1} < \theta_k = \infty$，

$$y = r \Leftrightarrow \theta_{r-1} < U \le \theta_r, r = 1, \cdots, k.$$

注：此处 y 表示样品的序值。

例如，学生的考试成绩分为不及格(1)，中(2)，良(3)，优(4)四个等级，U为其考试分数，$\theta_1,\theta_2,\theta_3$可分别取为 59，74 和 84。分析的目的与前面一样，考察一些因素(协变量)对y的影响。例如学生考试等级与其努力程度、学习方法、教师上课质量与考题质量等因素的关系。

设$U=-x'\beta+e$，e的分布函数为F，则$P(y\le r|x)=P(U\le\theta_r|x)=F(\theta_r+x'\beta)$。对$F$不同选择得到不同的模型。

8.3 极大似然估计(MLE)

响应变量一维情形

设有独立样本$y_i,x_i,i=1,\cdots,n$，似然函数为

$$L=\exp\left[\frac{\sum_{i=1}^n\theta_i y_i-\sum_{i=1}^n b(\theta_i)}{\phi}-\sum_{i=1}^n c(y_i,\phi)\right],$$

其中

$$\theta_i=\theta_i(\beta)=\dot{b}^{-1}[h(x_i'\beta)]。$$

先考虑β的估计。找$\beta=\hat{\beta}_n$使L达到最大，即为β的 MLE。

取对数，略去对估计β无影响的量，记$l_i(\beta)=\frac{[y_i\theta_i(\beta)-b(\theta_i(\beta))]}{\phi}$，称

$$l(\beta)=\sum_{i=1}^n l_i(\beta)=\sum_{i=1}^n\frac{[y_i\theta_i(\beta)-b(\theta_i(\beta))]}{\phi}$$

为对数似然函数。记$s_i(\beta)=\frac{\partial l_i(\beta)}{\partial\beta}=\frac{\partial\theta_i(\beta)}{\partial\beta}\frac{[y_i-\dot{b}(\theta_i(\beta))]}{\phi}$，称

$$s(\beta)=\sum_{i=1}^n s_i(\beta)=\sum_{i=1}^n\frac{\partial\theta_i(\beta)}{\partial\beta}\frac{[y_i-\dot{b}(\theta_i(\beta))]}{\phi}$$

为得分函数(score function)。

$\frac{\partial\theta_i(\beta)}{\partial\beta}$取决于联系函数$g$的形式。注意$h$表示$g$的反函数，$\mu_i=h(x_i'\beta)$，记$D_i(\beta)=\frac{dh(t)}{dt}|_{t=x_i'\beta}$，因此$\frac{\partial\mu_i}{\partial\beta}=\frac{dh(t)}{dt}|_{t=x_i'\beta}x_i=D_i(\beta)x_i$。另一方面$\frac{\partial\mu_i}{\partial\beta}=\frac{\partial\mu_i}{\partial\theta_i}\frac{\partial\theta_i}{\partial\beta}$，由于$\mu_i=Ey_i=\dot{b}(\theta_i(\beta))$，$Var(y_i)=\phi\ddot{b}(\theta_i(\beta))\overset{\Delta}{=}\phi\sigma_i^2(\beta)$，从而

$$\frac{\partial\mu_i}{\partial\beta}=\ddot{b}(\theta_i)\frac{\partial\theta_i}{\partial\beta}=\sigma_i^2(\beta)\frac{\partial\theta_i}{\partial\beta}。$$

因此$\frac{\partial\theta_i(\beta)}{\partial\beta}=\sigma_i^{-2}(\beta)\frac{\partial\mu_i}{\partial\beta}=\sigma_i^{-2}(\beta)D_i(\beta)x_i$。为强调$\mu_i$与$\beta$的关系，常写为$\mu_i(\beta)$，从而

$$s_i(\beta)=\phi^{-1}\sigma_i^{-2}(\beta)D_i(\beta)x_i[y_i-\mu_i(\beta)],$$

似然方程

$$s(\beta)=\sum_{i=1}^n\phi^{-1}\sigma_i^{-2}(\beta)D_i(\beta)x_i[y_i-\mu_i(\beta)]=0。$$

其解为$\hat{\beta}_n$为β的 MLE(注意虽然$s(\beta)$的定义中含有ϕ，但上方程求解与ϕ无关)。

当g取 canonical link 时，$h=\dot{b}$，$D_i(\beta)=\ddot{b}(x_i'\beta)=\ddot{b}(\theta_i)=\sigma_i^2(\beta)$，此时

$$s_i(\beta)=\frac{x_i[y_i-\mu_i(\beta)]}{\phi},$$

从而似然方程有简单表示

$$s(\beta)=\sum_{i=1}^n s_i(\beta)=\sum_{i=1}^n\frac{x_i[y_i-\mu_i(\beta)]}{\phi}=0。$$

似然方程可能无解，或者有解不唯一。不过在一定条件下，可以证明以下两点：

1.$\lim_{n\to\infty} P(\text{似然方程有解})=1$；

2.有一个解$P\left(\lim_{n\to\infty}\hat{\beta}_n=\beta_0\right)=1$，其中以$\beta_0$表示$\beta$的真值，与变化的$\beta$区别开来。

Fisher 信息阵

$$F(\beta)=Cov[s(\beta)]=\sum_{i=1}^{n}F_i(\beta),$$

其中

$F_i(\beta)=\phi^{-1}x_i x_i' w_i(\beta)$， $w_i(\beta)=D_i^2(\beta)\sigma_i^{-2}(\beta)$。

在一定条件下有

$$F^{1/2}(\beta_0)(\hat{\beta}_n-\beta_0)\xrightarrow{d.f.}N(0,I_p),$$

其中p为β_0的维数，I_p为p阶单位矩阵。

注意：如果ϕ未知的话，$F(\beta_0)$不仅含有未知的β_0，还含有未知的ϕ。上式还不能用作大样本统计推断。假设ϕ已知，则$F(\beta_0)$有两种估计方法：

1. 用$F(\hat{\beta}_n)$来估计$F(\beta_0)$；

2. $F_{obs}(\beta)=-\dfrac{\partial s(\beta)}{\partial\beta'}=-\dfrac{\partial^2 l(\beta)}{\partial\beta\partial\beta'}$,

则$E[F_{obs}(\beta)]=F(\beta)$. 因此用$F_{obs}(\hat{\beta}_n)$来估计$F(\beta_0)$.

在一定条件下都是相合(consistent)估计。此时

$$F^{1/2}(\hat{\beta}_n)(\hat{\beta}_n-\beta_0)\xrightarrow{d.f.}N(0,I_p) \qquad ***$$

若ϕ未知，由于$\ddot{b}(\theta)=\ddot{b}\left[\dot{b}^{-1}(\mu)\right]\overset{\Delta}{=}V(\mu)$

$$\hat{\phi}_n=\frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

其中$\hat{\mu}_i=\mu(x_i'\hat{\beta}_n)$。在一定条件下，$\hat{\phi}_n$为$\phi$的相合估计。此时用$\hat{\phi}_n$代替(***)式中的$\phi$，渐近形式不变。

响应变量多维情形

原则上与一维情形相似。仿前面记法有

$$l_i(\beta)=\frac{y_i'\theta_i(\beta)-b(\theta_i(\beta))}{\phi},$$

其中y_i，$\theta_i(\beta)$为$q\times1$，β为$p\times1$向量。对数似然

$$l(\beta)=\sum_{i=1}^{n}l_i(\beta)=\sum_{i=1}^{n}\frac{\left[y_i'\theta_i(\beta)-b(\theta_i(\beta))\right]}{\phi}。$$

记$s_i(\beta)=\dfrac{\partial l_i(\beta)}{\partial\beta}=\dfrac{\partial\theta_i(\beta)}{\partial\beta'}\dfrac{\left[y_i-\dot{b}(\theta_i(\beta))\right]}{\phi}$，其中$\dot{b}(\theta)=\dfrac{\partial b(\theta)}{\partial\theta}$。称

$$s(\beta)=\sum_{i=1}^{n}s_i(\beta)=\sum_{i=1}^{n}\frac{\partial\theta_i'(\beta)}{\partial\beta}\frac{\left[y_i-\dot{b}(\theta_i(\beta))\right]}{\phi}$$

为得分函数(score function)。记$g(q\times1)$为联系函数，$g^{-1}=h$，$\mu_i=Ey_i=h(Z_i\beta)$，其中$Z_i=Z_i(x)$为$q\times p$矩阵。

令$D_i(\beta)=\frac{\partial h'(t)}{\partial t}|_{t=Z_i\beta}$，则

$\frac{\partial\mu_i}{\partial\beta'}=\frac{\partial h(t)}{\partial t'}|_{t=Z_i\beta}Z_i=D_i'(\beta)Z_i$ 。另一方面 $\frac{\partial\mu_i}{\partial\beta'}=\frac{\partial\mu_i}{\partial\theta_i'}\frac{\partial\theta_i}{\partial\beta'}$，由于$\mu_i=Ey_i=\dot{b}(\theta_i(\beta))$，$Var(y_i)=\phi\ddot{b}(\theta_i(\beta))\stackrel{\Delta}{=}\phi\Sigma_i(\beta)$，从而

$$\frac{\partial\mu_i}{\partial\beta'}=\ddot{b}(\theta_i)\frac{\partial\theta_i}{\partial\beta'}=\Sigma_i(\beta)\frac{\partial\theta_i}{\partial\beta'}。$$

因此$\frac{\partial\theta_i(\beta)}{\partial\beta'}=\Sigma_i^{-1}(\beta)\frac{\partial\mu_i}{\partial\beta'}=\Sigma_i^{-1}(\beta)D_i'(\beta)Z_i$。为强调$\mu_i$与$\beta$的关系，常写为$\mu_i(\beta)$， 从而

$$s_i(\beta)=Z_i'D_i(\beta)\Sigma_i^{-1}(\beta)\frac{\left[y_i-\mu_i(\beta)\right]}{\phi},$$

似然方程

$$s(\beta)=\sum_{i=1}^{n}Z_i'D_i(\beta)\Sigma_i^{-1}(\beta)\frac{\left[y_i-\mu_i(\beta)\right]}{\phi}=0。$$

其解为$\hat{\beta}_n$为β的 MLE(注意$s(\beta)$的定义中含ϕ，但上方程求解与ϕ无关)。

Fisher 信息阵

$$F(\beta)=Cov[s(\beta)]=\sum_{i=1}^{n}F_i(\beta),$$

其中

$$F_i(\beta)=\phi^{-1}Z_i'W_i(\beta)Z_i,$$

$$W_i(\beta)=D_i(\beta)\Sigma_i^{-1}(\beta)D_i'(\beta)$$

$$=\left(\frac{\partial g(t)}{\partial t'}|_{t=\mu_i}\Sigma_i(\beta)\frac{\partial g'(t)}{\partial t}|_{t=\mu_i}\right)^{-1}。$$

在一定条件下有

1. $\lim_{n\to\infty}P$(似然方程有解)$=1$；

2. 有一个解$\hat{\beta}_n$为β_0的相合估计，其中以β_0表示β的真值，与变化的β区别开来。

3. $F^{1/2}(\beta_0)(\hat{\beta}_n-\beta_0)\xrightarrow{d.f.}N(0,I_p)$，其中$p$为$\beta_0$的维数，$I_p$为$p$阶单位矩阵。

4. 以$\hat{\beta}_n$代替β_0，

$$F^{1/2}(\hat{\beta}_n)(\hat{\beta}_n-\beta_0)\xrightarrow{d.f.}N(0,I_p)。$$

8.4 Iteratively re-weighted least squares

求解极大似然估计可以采用一般优化问题的 Newton-Raphson 迭代方法，也可以采用下面的 Fisher scoring 算法：

$$\beta^{(k+1)}=\beta^{(k)}+\left[-E\frac{\partial^2 l(\beta^{(k)})}{\partial\beta\partial\beta'}\right]^{-1}\frac{\partial l(\beta^{(k)})}{\partial\beta}。$$

由于$l(\beta)$为对数似然函数，在一定正则条件下(例如求导与积分可交换)有

$$F(\beta)=Cov\left[s(\beta)\right]=-E\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta'}。$$

此时，Fisher scoring 算法为：

$$\beta^{(k+1)}=\beta^{(k)}+F(\beta^{(k)})^{-1}s(\beta^{(k)})。$$

以下以响应变量一维为例(多维类似)。此时

$$s(\beta)=\phi^{-1}\sum_{i=1}^{n}\sigma_i^{-2}(\beta)D_i(\beta)x_i\left[y_i-\mu_i(\beta)\right]$$

$$\triangleq\phi^{-1}X'A(\beta)\left[Y-\mu(\beta)\right]$$

这里$X_{n\times p}=\left(x_1,x_2,\cdots,x_n\right)'$，$Y_{n\times 1}=\left(y_1,y_2,\cdots,y_n\right)'$，

$\mu_{n\times 1}(\beta)=\left(\mu_1(\beta),\mu_2(\beta),\cdots,\mu_n(\beta)\right)'$，

$A_{n\times n}(\beta)=diag\left(\sigma_i^{-2}(\beta)D_i(\beta)\right)_{1\le i\le n}$。

由前面记号及推导

$$F(\beta)=Cov\left[s(\beta)\right]=\phi^{-1}\sum_{i=1}^{n}x_i x_i' w_i(\beta),$$

其中$w_i(\beta)=D_i^2(\beta)\sigma_i^{-2}(\beta)$。写成矩阵形式

$$F(\beta)=\phi^{-1}X'W(\beta)X,$$

其中$W_{n\times n}(\beta)=diag\left(w_i(\beta)\right)_{1\le i\le n}$。因此 Fisher scoring 算法为(注参数ϕ正好消去不影响):

$$\beta^{(k+1)}=\beta^{(k)}+\left[X'W(\beta^{(k)})X\right]^{-1}X'A(\beta^{(k)})\left[Y-\mu(\beta^{(k)})\right]。$$

此外，由

$$\beta^{(k+1)}=\beta^{(k)}+\left[X'W(\beta^{(k)})X\right]^{-1}X'A(\beta^{(k)})\left[Y-\mu(\beta^{(k)})\right]$$
$$=\left[X'W(\beta^{(k)})X\right]^{-1}\left\{X'W(\beta^{(k)})X\beta^{(k)}+X'A(\beta^{(k)})\left[Y-\mu(\beta^{(k)})\right]\right\}$$

及

$$A(\beta)=W(\beta)diag\left(D_1^{-1}(\beta),\cdots,D_n^{-1}(\beta)\right),$$

若令

$$z_{n\times 1}(\beta)=X\beta+diag\left(D_i^{-1}(\beta)\right)_{1\le i\le n}\left[Y-\mu(\beta)\right],$$

则 Fisher scoring 算法为

$$\beta^{(k+1)}=\left[X'W(\beta^{(k)})X\right]^{-1}X'W(\beta^{(k)})z(\beta^{(k)})。$$

上计算方法右边形式相当于求解一个加权最小二乘问题，因此该算法称为 Iteratively re-weighted least squares（IRLS 或者 IRWLS）。再来看$z(\beta)$的第i个分量为

$$z_i(\beta)=x_i'\beta+D_i^{-1}(\beta)\left[y_i-\mu_i(\beta)\right],$$

省去β，$z_i=g(\mu_i)+\dot{g}(\mu_i)(y_i-\mu_i)$，为$g(y_i)$在$\mu_i=Ey_i$出的一阶 Taylor 近似。因此$z_i$称为工作变量(working variate)或者称为调整响应(adjusted response)。Fisher scoring 算法可以看成响应变量转化版本的 IRLS 算法：

1. 在每一步在当前β下计算新的工作变量z及新的权重W；
2. 将z对X作权重为W的最小二乘更新β。

注意到$F(\beta)=\phi^{-1}X'W(\beta)X$，因此最后得的$\phi\left(X'WX\right)^{-1}$即为极大似然估计$\hat{\beta}_{MLE}$的渐近方差矩阵。

8.5 Deviance

Deviance 在 GLMs 中是一个很重要的概念。它既可以用来检验联系函数与线性预测子之间拟合效果，又可以用来检验某些协变量是否显著。为理解 Deviance 的含义，需要引入全模型(saturated model)的概念。对 GLMs：$Ey_i=\mu_i$，联系函数为g，$g(\mu_i)=x_i'\beta$，定义其全模型为同样分布及联系函数但$g(\mu_i)=\psi_i$（$1\le i\le n$）为未知参数。意味着，全模型$g(\mu_i)$不像原来的 GLMs 有线性结构的限制。

记$\psi=(\psi_1,\cdots\psi_n)'$，观测数据$Y=(y_1,\cdots y_n)'$全模型的似然函数为$L_S(Y,\psi)$，GLMs 的似然函数为$L(Y,\beta)$，为检验联系函数与线性预测子是否拟合很好，原假设为H_0:存在关系$g(\mu_i)=x_i'\beta$，备则假设H_1:不存在。采用极大似然比检验。原假设下极大似然得到$L(Y,\hat{\beta})$，这里$\hat{\beta}$为极大似然估计，整个空间下极大似然得到$L_S(Y,\hat{\psi})$，似然比为$\dfrac{L(Y,\hat{\beta})}{L_S(Y,\hat{\psi})}$。

由极大似然比检验的理论，一般条件下，若原假设成立，则$-2\log\dfrac{L(Y,\hat{\beta})}{L_S(Y,\hat{\psi})}$服从（或渐近服从）$\chi^2_{n-p}$分布。若用小写的$l$表示对数似然，定义 GLMs 的 deviance 统计量为

$$D=-2\left[l(Y,\hat{\beta})-l_S(Y,\hat{\psi})\right]。$$

由定义可见 deviance 即为上检验问题的极大似然比检验统计量。D越大表明 GLMs 拟合的不好。

例 8.5.1：(Poisson GLMs)设样本$y_1,y_2,\cdots y_n$独立，y_i来自$Poisson(\lambda_i)$分布，$\log\lambda_i=x_i'\beta$。对数似然

$$\begin{aligned}l(Y,\beta)&=\sum_{i=1}^n y_i\log\lambda_i-\sum_{i=1}^n\lambda_i-\sum_{i=1}^n\log y_i!\\&=\sum_{i=1}^n y_i x_i'\beta-\sum_{i=1}^n e^{x_i'\beta}-\sum_{i=1}^n\log y_i!\end{aligned},$$

极大化上式，设在$\beta=\hat{\beta}$达到极大值。

对全模型，其对数似然为

$$l(Y,\psi)=\sum_{i=1}^n y_i\psi_i-\sum_{i=1}^n e^{\psi_i}-\sum_{i=1}^n\log y_i!。$$

令集合$P=\{i|y_i>0\}$，P^c表示其补集。则

$$l(Y,\psi)=\sum_{i\in P}(y_i\psi_i-e^{\psi_i})-\sum_{i\in P^c}e^{\psi_i}-\sum_{i=1}^n\log y_i!。$$

易见，对$i\in P$，$\psi_i=\log y_i$，$i\in P^c$，$e^{\psi_i}=0$时上式有极大值

$$\begin{aligned}l(Y,\hat{\psi})&=\sum_{i\in P}(y_i\log y_i-y_i)-\sum_{i=1}^n\log y_i!\\&=\sum_{i=1}^n(y_i\log y_i-y_i)-\sum_{i=1}^n\log y_i!\end{aligned}。$$

Deviance 为

$$D=2\sum_{i=1}^n\left[y_i(\log y_i-x_i'\hat{\beta})-(y_i-e^{x_i'\hat{\beta}})\right]。$$

注意到$\lambda_i=Ey_i=\exp(x_i'\beta)$，若记$\hat{\lambda}_i=\exp(x_i'\hat{\beta})$，则

$$D=2\sum_{i=1}^n\left[y_i\log\frac{y_i}{\hat{\lambda}_i}-(y_i-\hat{\lambda}_i)\right]。$$

此外，我们可以利用 Deviance 的差异来检验某些协变量是否显著。例如，若检验某$p-q$个协变量$(q<p)$是否对响应变量有影响。设其对应的系数为β_{p-q}，剩余参数记为β_q，即检验下面的假设$H_0:\beta_{p-q}=0\leftrightarrow H_1:\beta_{p-q}\neq 0$。在原假设下，对应的是只有$q$个协变量的子模型，备则假设下是包含其具有$p$个协变量的模型。$\hat{\beta}_q$为原假设下参数的极大似然估计，则极大似然比检验统计量为$-2\left[l(Y,\hat{\beta}_q)-l(Y,\hat{\beta}_p)\right]$。

假设拟合p个协变量时的 GLMs，其 Deviance 记为D_p，由定义有

$$D_p=-2\left[l(Y,\hat{\beta}_p)-l_S(Y,\hat{\psi})\right],$$

则上检验某$p-q$个协变量是否显著的极大似然比检验统计量即为 Deviance 之间的差异

$$D_q-D_p,$$

在原假设下，其服从(渐近服从)χ^2_{p-q}分布。