第10章　离散响应变量模型

吴密霞

北京工业大学统计与数据科学系

E-mail:　*wumixia@bjut.edu.cn*

教材

- 吴密霞, 王松桂. 2024. 线性模型引论（第2版）, 科学出版社.

本章概述

本章内容目录

- 广义线性模型
- 列联表
- Logistic 回归模型
- 多分类Logistic回归
- 泊松回归

本章学习要求

初步了解广义线性模型、列联表、Logistic 回归模型、多分类Logistic回归、泊松回归，重点掌握Logistic 回归模型建模思想，基本估计和检验方法.

广义线性模型

- 当Y的概率分布属于指数分布族时, Nelder 和Wedderburn (1972) 提出了一种更一般方法来拟合线性模型, 即广义线性模型(general linear model, GLM).

该模型由三个部分组成：

1. 随机部分: 给出因变量Y的概率分布;
2. 系统部分: 指定了自变量的线性函数;
3. 连接函数: 描述了系统部分与随机部分期望之间的函数关系.

广义线性模型

1. 随机部分(random component)

给出因变量Y的概率分布: 设 $\boldsymbol{y} = (y_1, y_2, ..., y_n)$为$Y$的观测向量, 其中每个观测值$y_i$ 的分布属于指数分布族, 其概率密度函数为

$$f(y_i; \theta_i) = a(\theta_i)b(y_i)\exp(y_i Q(\theta_i))$$

这里, 参数θ_i 可随$i = 1, 2, \cdots, n$ 变化, 并依赖于同y_i 有关的自变量的值$(x_{i1}, x_{i2}, \cdots, x_{i(p-1)})$

$Q(\theta_i)$ 是该分布的自然参数.

y_i 的分布可为正态分布、泊松分布、二项分布和多项分布等

2 系统部分:

通过线性模型将一个向量$\boldsymbol{\eta} = (\eta_1, \eta_2, \cdots, \eta_n)'$与一组自变量联系起来, 即

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

其中, $\boldsymbol{\eta}$ 被称为线性预测因子, $\mathbf{X}$为$n \times p$ 的自变量矩阵, $\boldsymbol{\beta}$ 是$p \times 1$ 参数向量. 常规假设$\mathbf{X}$的第一列元素皆为1, 并记

$$\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1')' = (\beta_0, \beta_1, \cdots, \beta_{p-1})'.$$

3. 连接函数

将系统部分与随机部分的期望值连接起来.

- 设$\mu_i = E(y_i)$, 则μ_i 通过 $\eta_i = g(\mu_i)$ 与η_i 联结, 即

$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i(p-1)} = \boldsymbol{x}_i' \boldsymbol{\beta}, \quad i = 1, 2, \cdots, n,$$

其中, $g(\cdot)$ 是单调可微分函数, $\boldsymbol{x}_i = (1, x_{i1}, \cdots, x_{i(p-1)})'$ 为$\mathbf{X}$的第i行行向量. 特别地,

- 称$g(\mu) = \mu$为恒等连接(identify link) $\Rightarrow \eta_i = \mu_i$.
- 称$g(\mu) = Q(\theta_i)$为正则连接（canonical link）$\Rightarrow Q(\theta_i) = \boldsymbol{x}_i' \boldsymbol{\beta}$.

列联表

列联表是针对响应变量和自变量都是分类变量的情形.

设X 和Y 表示两个分类变量, X 有I 个水平, Y 有J 个水平.

- 设(X, Y)的联合概率分布为

$$P(X = i, Y = j) = \pi_{ij}, \quad i = 1, 2, \cdots, I, \quad j = 1, 2, \cdots, J, \tag{10.3}$$

- X和Y的边缘概率分布分别为

$$\pi_{i\cdot} = \sum_{j=1}^{J} \pi_{ij}, \quad i = 1, 2, \cdots, I,$$

$$\pi_{\cdot j} = \sum_{i=1}^{I} \pi_{ij}, \quad j = 1, 2, \cdots, J,$$

列联表

- 将X和Y的联合概率和边缘概率汇总表

Table: X和Y的联合概率和边缘概率

		Y				
		1	2	$\cdots$	J	$P(X=i)$
X	1	π_{11}	π_{12}	$\cdots$	π_{1J}	$\pi_{1.}$
	2	π_{21}	π_{12}	$\cdots$	π_{2J}	$\pi_{2.}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	I	π_{I1}	π_{I2}	$\cdots$	π_{1J}	$\pi_{I.}$
$P(Y=j)$		$\pi_{.1}$	$\pi_{.2}$	$\cdots$	$\pi_{.J}$	

列联表

- 若对于所有的i,j, 都有

$$\pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}, \quad i = 1, \cdots, I, \quad j = 1, \cdots, J,$$

则X和Y独立.

- 若$\pi_{i\cdot} \neq 0$, 给定$X = i$条件下, $Y = j$的条件概率为

$$\pi_{j|i} = P(Y = j|X = i) = \frac{\pi_{ij}}{\pi_{i\cdot}},$$

且满足

$$\sum_{j=1}^{J} \pi_{j|i} = 1.$$

- 若X和Y独立, 则 $\pi_{j|i} = \pi_{\cdot j}$, $\pi_{i|j} = \pi_{i\cdot}$

列联表

- 列联表(contingency table)

Table: 样本量为n 的(X, Y)的频数列联表

		Y				
		1	2	$\cdots$	J	$n_{i\cdot}$
X	1	n_{11}	n_{12}	$\cdots$	n_{1J}	$n_{1\cdot}$
	2	n_{21}	n_{12}	$\cdots$	n_{2J}	$n_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	I	n_{I1}	n_{I2}	$\cdots$	n_{1J}	$n_{I\cdot}$
	$n_{\cdot j}$	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot J}$	

n_{ij}为$(X = i, Y = j)$的单元频数, $n = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}$, $n_{i\cdot} = \sum_{j=1}^{J} n_{ij}$, $n_{\cdot j} = \sum_{i=1}^{I} n_{ij}$.

列联表

由列联表立得

- X和Y的联合概率π_{ij}的估计

$$\widehat{\pi}_{ij} = \frac{n_{ij}}{n},$$

- 边缘概率$\pi_{\cdot i}$和$\pi_{\cdot i}$的估计

$$\widehat{\pi}_{i\cdot} = \frac{n_{i\cdot}}{n}, \quad \widehat{\pi}_{i|j} = \frac{n_{\cdot j}}{n},$$

- 条件概率$\pi_{j|i}$和$\pi_{i|j}$的估计

$$\widehat{\pi}_{j|i} = \frac{n_{ij}}{n_{i\cdot}}, \quad \widehat{\pi}_{i|j} = \frac{n_{ij}}{n_{\cdot j}}.$$

条件概率比较指标

在应用中, 感兴趣的往往不是X和Y的联合概率, 而是条件概率, 尤其是比较在X的不同水平下Y的条件概率, 如响应概率的差、相对风险、优势比(odds ratio, OR)等.

假设Y是0, 1二分类响应变量, X为分类变量, $X = i,\ \ i = 1, \cdots, I$.

- 给定$X = i$ 条件下, Y的条件响应概率为

$$\pi_{1|i} = P(Y = 1|X = i),$$

- 给定$X = i$条件下, Y的未响应概率为

$$\pi_{2|i} = P(Y = 0|X = i) = 1 - \pi_{1|i}.$$

条件概率比较指标

1. 条件概率差

- 条件$X=i$和$X=h$下响应概率的差：$\pi_{1|i}-\pi_{1|h}$
- 条件$X=i$和$X=h$下无响应概率的差：$\pi_{0|i}-\pi_{0|h}$

注 这两个差的绝对值相等, 符号相反,

$$\pi_{0|i}-\pi_{0|h}=-(\pi_{1|i}-\pi_{1|h})$$

$$-1\leq\pi_{1|i}-\pi_{1|h}\leq 1,$$

注 当Y与X独立时,

$$\pi_{1|i}-\pi_{1|h}=0,\quad 1\leq i\neq h\leq I.$$

条件概率比较指标

2. 相对风险

设Y 为二分类响应变量, 则称比值$\dfrac{\pi_{1|i}}{\pi_{1|h}}$ 为第i 类响应相对于第h 类响应的相对风险.

- 相对风险是一个非负实数. 对于2×2 的表, 响应的相对风险

$$0 \leq R_1 = \frac{\pi_{1|1}}{\pi_{1|2}} < \infty.$$

相对风险等于1 意味着X和Y独立. 同理, 非响应的相对风险为

$$R_2 = \frac{\pi_{2|1}}{\pi_{2|2}} = \frac{1-\pi_{1|1}}{1-\pi_{1|2}}.$$

3. 几率/优势

X 的同一类(如$X = i$)中的响应概率与非响应概率的比值被称作几率或事件$\{Y = 1\}$的优势.

- 对于2×2 的表, 第一行$(X = 1)$下的几率等于

$$\Omega_1 = \frac{\pi_{1|1}}{\pi_{2|1}}.$$

- 第二行$(X = 0)$下的几率等于

$$\Omega_2 = \frac{\pi_{1|2}}{\pi_{2|2}}.$$

- 当X和Y独立时, $\Omega_1 = \Omega_2$.

条件概率比较指标

4. 优势比

设X和Y皆为二分类变量, 则称

$$\theta = \frac{R_1}{R_2} = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} \tag{10.4}$$

为优势比(odds ratio, OR) 或列联系数(contingency coefficient).

- 此定义下X和Y独立的充要条件为$\theta = 1$.
- 对于任意$I \times J$的表, 其任意两行和两列就构成一个2×2表.
- 通常考虑相邻的2×2表, 其局部优势比为

$$\theta_{ij} = \frac{\Omega_1}{\Omega_2} = \frac{\pi_{i|j}\pi_{(i+1)|(j+1)}}{\pi_{i|(j+1)}\pi_{(i+1)|j}}, \quad i = 1, \cdots, I-1, \ \ j = 1, \cdots, J-1.$$

列联表的抽样分布

介绍列联表的两种抽样: 泊松分布抽样和独立的多项式分布抽样.

1. 泊松分布抽样

   1) 假设n_{ij} 服从参数m_{ij}的泊松分布, $i = 1, 2, \cdots, I, j = 1, 2, \cdots, J$, 则其概率分布为

$$P(n_{ij}) = \frac{e^{-m_{ij}} m_{ij}^{n_{ij}}}{n_{ij}!}, \quad n_{ij} = 0, 1, \cdots,$$

   且 $E(n_{ij}) = \mathrm{Var}(n_{ij}) = m_{ij}$.

   2) 假设所有单元频数n_{ij}相互独立.

列联表的抽样分布

在这两个假设下，我们有

- $n = \sum_i \sum_j n_{ij}$ 仍然服从泊松分布, $E(n) = \mathrm{Var}(n) = \sum_i \sum_j m_{ij}$.
- 给定样本量n的条件下, 则$n_{11}, \cdots, n_{1J}, \cdots, n_{I1}, \cdots, n_{IJ}$ 的条件联合分布为多项式分布(multinomial distribution), 其概率分布为

$$P\left(n_{ij} \middle| \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = n\right) = \frac{n!}{\prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}!} \prod_{i=1}^{I}\prod_{j=1}^{J} \pi_{ij}^{n_{ij}},$$

- n_{ij}的条件边缘概率分布为二项分布$B(\pi_{ij}, n)$, 其中

$$\pi_{ij} = m_{ij} \Big/ \sum_{l=1}^{I}\sum_{h=1}^{J} m_{lh} .$$

2. 独立的多项式分布抽样

- 假设给定 $X = i$ 的条件下, 列联表中Y 的$n_{i\cdot}$ 个观测相互独立, 且

$$P(Y = j|X = i) = \pi_{j|i}, \quad j = 1, 2, \cdots, J,$$

则第i行各单元观测频数$n_{i1}, \cdots, n_{iJ}$ 服从多项式分布,

$$\frac{n_{i\cdot}!}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} \pi_{j|i}^{n_{ij}}.$$

- 假设列联表的不同行的各单元观测频数相互独立, 则

$$\prod_{i=1}^{I} \left(\frac{n_{i\cdot}!}{\prod_{j=1}^{J} n_{ij}!} \left(\prod_{j=1}^{J} \pi_{j|i}^{n_{ij}} \right) \right).$$

故称相应的抽样为乘积多项式抽样或独立多项式抽样.

拟合优度检验

列联表分析主要研究方法: 检验观测频数是否等于理论频数

- 检验列联表的频数是否服从指定的多项式分布:

$$H_0 : \pi_{ij} = \pi_{ij,0}, \quad i = 1, \cdots, I, \ \ j = 1, \cdots, J,$$

其中π_{ij}满足限制条件: $\sum_{i=1}^{I}\sum_{j=1}^{J} \pi_{ij} = 1$. 当$H_0$成立时, 理论频数为

$$m_{ij} = n\pi_{ij,0}, \quad i = 1, 2, \cdots, I, \quad j = 1, 2, \cdots, J.$$

χ^2 检验统计量和其H_0下的渐近分布为

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - m_{ij})}{m_{ij}} \longrightarrow \chi^2_{(I-1)(J-1)}.$$

拟合优度检验

- Y和X 的独立性检验：

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j}, \quad i = 1, 2, \cdots, I, \ \ j = 1, 2, \cdots, J,$$

常用的两种检验方法：

(1) χ^2 检验： 其检验统计量为

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \widehat{m}_{ij})}{\widehat{m}_{ij}}.$$

(2) 似然比检验

例如, 当单元格样本量$\{n_{ij}\}$服从多项式分布时, Y和X 的独立性检验的似然比统计量如下:

拟合优度检验

似然比检验统计量

$$\Lambda = \frac{\prod_{i=1}^{I}\prod_{j=1}^{J}(n_{i\cdot}n_{\cdot j})^{n_{ij}}}{n^n\prod_{i=1}^{I}\prod_{j=1}^{J}n_{ij}^{n_{ij}}} = \prod_{i=1}^{I}\prod_{j=1}^{J}\left(\frac{\widehat{m}_{ij}}{n_{ij}}\right)^{n_{ij}},$$

其中$\widehat{m}_{ij} = n_{i\cdot}n_{\cdot j}/n$是$H_0$下$m_{ij}$的极大似然估计.

在独立假设下, Wilks (1938)证得当$n \to \infty$时,

$$G^2 = -2\ln\Lambda = 2\sum_{i=1}^{I}\sum_{j=1}^{J}n_{ij}\ln\frac{n_{ij}}{\widehat{m}_{ij}} \to \chi^2_{(I-1)(J-1)}.$$

因此, 当样本量较大时, G^2可作为独立假设的检验统计量.

Logistic 回归模型

Logistic回归模型, 又称 “评定模型", 被广泛应用于对事件发生的危险因素探索、发生概率预测以及新样本的判别归类等.

设Y 只取0, 1两个值的分类变量, 描述了事件A发生或未发生两种状态, $(X_1, \cdots, X_{p-1})$ 为对Y有影响的自变量.

记$y_1, \cdots, y_n$ 是随机变量Y在自变量取不同值$(x_{i1}, \cdots, x_{i(p-1)})$, $i = 1, \cdots, n$ 情形下的n次独立观测,

$$P(y_i = 1) = \pi(\boldsymbol{x}_i), \quad P(y_i = 0) = 1 - \pi(\boldsymbol{x}_i).$$

这里$\boldsymbol{x}_i = (1, x_{i1}, \cdots, x_{i(p-1)})'$.

Logistic 回归模型

- 若采用恒等连接函数, $E(y_i) = \pi(\boldsymbol{x}_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$, 便得到线性模型

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i, \quad i = 1, \cdots, n,$$

其中$\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p-1})'$ 为待估参数.

该模型存在结构缺陷:

(1) $0 \leq \pi(\boldsymbol{x}_i) \leq 1$, 而$-\infty < \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta} < -\infty$, 可引起矛盾.

(2) 在实践中概率$\pi(\boldsymbol{x}_i)$往往是$\boldsymbol{x}_i$的非线性单调函数, 而非线性的.

(3) 模型随机误差的方差非常数, 且与自变量$\boldsymbol{x}_i$有关(不合常规)

$$\mathrm{Var}(e_i) = \mathrm{Var}(y_i) = \pi(\boldsymbol{x}_i)(1 - \pi(\boldsymbol{x}_i)) = (\boldsymbol{x}_i'\boldsymbol{\beta})(1 - \boldsymbol{x}_i'\boldsymbol{\beta})$$

Logistic 回归模型

- 采用正则连接函数的广义线性模型.

  由于y_i的概率分布可表示为

$$\begin{aligned} f(y_i) =& \pi^{y_i}(\boldsymbol{x}_i)(1-\pi(\boldsymbol{x}_i))^{1-y_i} \\ =& (1-\pi(\boldsymbol{x}_i))\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right)^{y_i} \\ =& (1-\pi(\boldsymbol{x}_i))\exp\left(y_i \ln\left(Q(\pi(\boldsymbol{x}_i))\right)\right), \end{aligned}$$

  自然参数

$$Q(\pi(\boldsymbol{x}_i)) = \ln\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right),$$

  即事件$\{y_i = 1\}$发生和$\{y_i = 0\}$发生概率比(几率)的对数.

Logistic 回归模型

- 采用正则连接函数, 得Logistic 回归模型:

$$\text{logit}(\pi(\boldsymbol{x}_i)) = \boldsymbol{x}_i'\boldsymbol{\beta}, \quad i = 1, \cdots, n, \tag{10.5}$$

这里

$$\text{logit}(\pi(\boldsymbol{x}_i)) = \ln\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right).$$

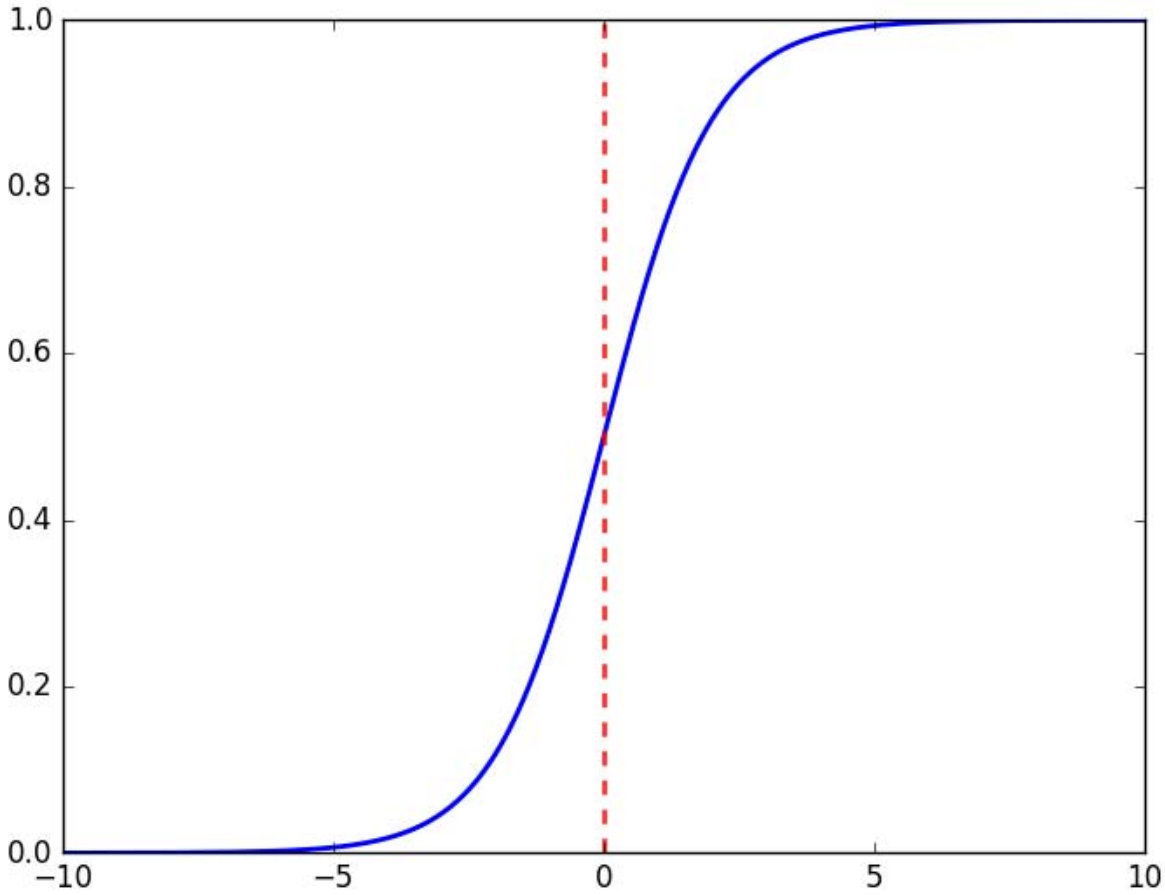
在Logistic 回归模型下, 概率$\pi(\boldsymbol{x}_i)$ 等于

$$\pi(\boldsymbol{x}_i) = h(\boldsymbol{x}_i'\boldsymbol{\beta}) = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})} = \frac{1}{1+\exp(-\boldsymbol{x}_i'\boldsymbol{\beta})}, \tag{10.6}$$

其中，响应函数$h(z)$为Sigmond函数,

$$h(z) = \frac{1}{1+e^{-z}}.$$

Logistic 回归模型



由$h(z)$的函数图像可以看出: Sigmond函数的图像关于点$(0, 0.5)$对称; 其定义域为$(-\infty, \infty)$: 值域为$(0, 1)$.

Logistic 回归模型的优点

- 模型更易解释.
  - $\text{logit}(\pi(\boldsymbol{x}_i)) = \ln\left(\frac{\pi(\boldsymbol{x}_i)}{(1-\pi(\boldsymbol{x}_i))}\right)$ 就是事件$\{y_i = 1\}$相对于事件$\{y_i = 0\}$的优势的对数;
  - 由$\frac{\pi(\boldsymbol{x}_i)}{(1-\pi(\boldsymbol{x}_i))} = \exp\{\boldsymbol{x}_i'\boldsymbol{\beta}\}$ 可得：固定自变量$\boldsymbol{x}_i$中其他元素, 仅变化其第k个元素, 每增加一个单位, 优势就增大$\exp(\beta_k)$倍.
- 自变量的效应可以被估计, 且对于自变量两个不同的取值$\boldsymbol{x}_i$和$\boldsymbol{x}_j$, 则两者的优势比

$$\text{OR}_{ij} = \left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right) \Big/ \left(\frac{\pi(\boldsymbol{x}_j)}{1-\pi(\boldsymbol{x}_j)}\right) = \exp\left\{\boldsymbol{x}_i'\boldsymbol{\beta} - \boldsymbol{x}_j'\boldsymbol{\beta}\right\};$$

- 自变量可以是离散变量, 也可以是连续变量.

Logistic 回归模型、响应函数与线性模型的关系

- 假设二元响应变量Y是基于某连续响应变量Y^c的二分产生的, Y^c和自变量$\boldsymbol{X} = (X_1, \cdots, X_{p-1})'$满足线性关系.

  记$(y_i^c, \boldsymbol{x}_i)$, $i = 1, 2, \cdots, n$ 为Y^c和$\boldsymbol{X}$的n次独立观测,

$$y_i = \begin{cases} 1, & y_i^c \le d, \\ 0, & y_i^c > d, \end{cases} \tag{10.7}$$

  这里, d是某个指定的常数.

$$y_i^c = \boldsymbol{x}_i' \boldsymbol{\beta}^c + e_i^c, \tag{10.8}$$

  其中$\boldsymbol{\beta}^c = (\beta_0^c, \beta_1^c, \cdots, \beta_{p-1}^c)'$, $E(e_i^c) = 0$, $\mathrm{Var}(e_i^c) = \sigma^2$.

Logistic 回归模型

例如随机抽取n名孕妇，记录她们怀孕时间和孕期酗酒程度

$$(y_i^c, x_i), \quad i = 1, \cdots, n.$$

- 若感兴趣的问题是母亲酗酒(X-孕期酗酒程度指数)对其怀孕时间(Y^c)的影响, 可用简单线性模型

$$y_i^c = \beta_0 + \beta x_i + e_i^c$$

- 若感兴趣的问题是母亲酗酒是否会导致婴儿早产, 此时响应变量为

$$y_i = \begin{cases} 1, & y_i^c \leq 38 \text{ 周} \quad (\text{早产}), \\ 0, & y_i^c > 38 \text{ 周} \quad (\text{足月}). \end{cases}$$

Logistic 回归模型

每给定模型误差e_i^c一个分布, 就可导出一个二元响应变量的广义线性模型, 如Logistic 回归模型、probit回归模型、log-log 模型等. .

- 记F_{e^c} 随机误差e_i^c的分布函数. 则响应变量y_i的均值

$$E(y_i) = \pi(\boldsymbol{x}_i) = P(e_i^c \le d - \boldsymbol{x}_i'\boldsymbol{\beta}^c) = F_{e^c}(d - \boldsymbol{x}_i'\boldsymbol{\beta}^c), \tag{10.9}$$

1. Logistic 回归模型

   若e_i^c服从Logistic 分布, 则e_i^c可被表示为

$$e_i^c = \frac{\sigma}{\pi/\sqrt{3}}\varepsilon_i,$$

   其中, ε_i服从标准的Logistic分布, 其均值为0、方差为$\pi^2/3$,

Logistic 回归模型

分布函数为

$$F_\varepsilon(\varepsilon_i) = \frac{\exp(\varepsilon_i)}{1+\exp(\varepsilon_i)} = \frac{1}{1+\exp(-\varepsilon_i)} = h(\varepsilon_i),$$

这里, 响应函数$h(\cdot)$ 为Sigmond函数. 于是

$$\pi(\boldsymbol{x}_i) = F_{e^c}(d - \boldsymbol{x}_i{}'\boldsymbol{\beta}^c) = F_\varepsilon(\boldsymbol{x}_i'\boldsymbol{\beta}) = h(\boldsymbol{x}_i'\boldsymbol{\beta}) = \frac{1}{1+\exp\{-\boldsymbol{x}_i'\boldsymbol{\beta}\}},$$

这里$\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p-1})'$, 其中$\beta_0 = \pi(d - \beta_0^c)/(\sqrt{3}\sigma)$, $\beta_j = -\pi\beta_j^c/(\sqrt{3}\sigma), \quad j = 1, \cdots, p-1$. 由此立得Logistic模型

$$\text{logit}(\pi(\boldsymbol{x}_i)) = \boldsymbol{x}_i'\boldsymbol{\beta}, \quad i = 1, \cdots, n.$$

2. Probit回归模型

若模型误差$e_i^c \sim N(0, \sigma^2)$, 则

$$\pi(\boldsymbol{x}_i) = P(y_i = 1) = P(y_i^c \le d) = \Phi\left(\frac{d - \beta_0^c}{\sigma} - \sum_{j=1}^{p-1} \frac{\beta_j^c}{\sigma} x_{ij}\right).$$

由此立得probit回归模型

$$\text{probit}(\pi(\boldsymbol{x}_i)) = \Phi^{-1}(\pi(\boldsymbol{x}_i)) = \boldsymbol{x}_i'\boldsymbol{\beta},$$

这里, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p-1})'$, 其中

$$\beta_0 = (d - \beta_0^c)/\sigma, \qquad \beta_j = -\beta_j^c/\sigma, \quad j = 1, \cdots, p-1.$$

3. log-log 模型

假设e_i^c 服从Gumbel分布 (一种常用的极(大)值型分布)

$$F_G(e_i^c) = \exp\left(-\exp\left(-e_i^c/\sigma\right)\right).$$

于是 $\pi(\boldsymbol{x}_i) = F_G(d - \boldsymbol{x}_i'\boldsymbol{\beta}_c) = \exp\left(-\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right),$

这里, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p-1})'$, $\beta_0 = -(d - \beta_0^c)/\sigma$, $\beta_j = \beta_j^c/\sigma \quad (j \neq 0)$.

由此得log-log模型：

$$\ln(-\ln(\pi(\boldsymbol{x}_i))) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

其响应函数为Gumbel生存函数：$h_{cG}(z) = \exp\left(-\exp\left(z\right)\right).$

该函数以很快的速度收敛到0, 但相当慢的速度收敛到1.

cloglog 模型

4. 补log-log 模型

在Gumbel分布情形, 往往对补事件$\{Y^c > d\}$更感兴趣. 定义y_i为

$$y_i = \begin{cases} 0, & y_i^c \le d, \\ 1, & y_i^c > d. \end{cases}$$

于是$\pi(\boldsymbol{x}_i) = P(y_i = 1) = P(y_i^c > d) = 1 - \exp\left(-\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right).$

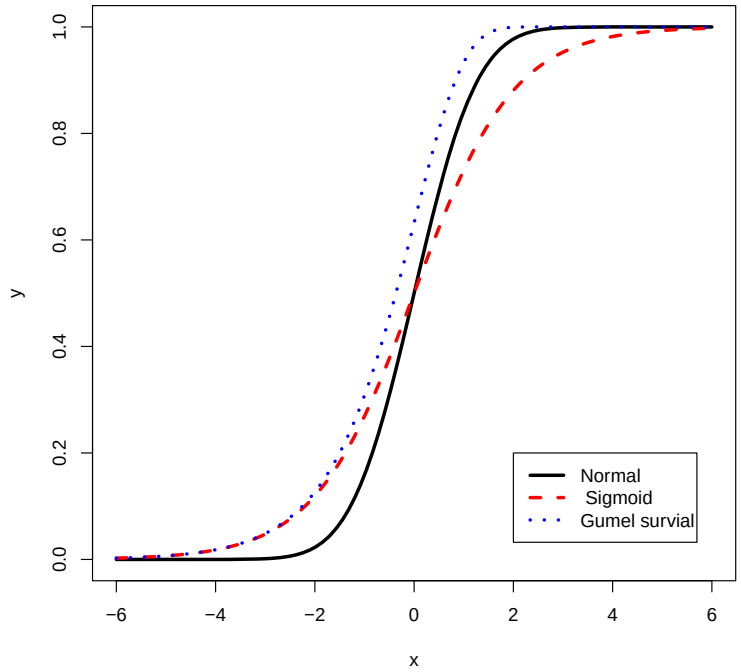
补log-log (complementary log-log, cloglog) 模型：

$$\ln(-\ln(1 - \pi(\boldsymbol{x}_i))) = \boldsymbol{x}_i'\boldsymbol{\beta}. \tag{10.10}$$

其响应函数为Gumbel生存函数：$h_{cG}(z) = 1 - \exp\left(-\exp\left(z\right)\right).$

该函数以很快的速度收敛到1, 但相当慢的速度收敛到0.

Logistic 回归模型



注 Logistic模型、probit模型和cloglog模型是三种最常用的二元响应变量模型.

注 共同特点是取值在0 和1 之间; 图像具有S形, 并随着自变量趋于$-\infty$或$+\infty$逐渐接近0 或1.

注 三个函数分别适用于模型误差重尾分布, 误差正态分布和非对称分布的情形.

Logistic模型参数的估计

设计矩阵不列满秩时, 可以通过模型重新参数化或自变量极大线性无关组等方法将模型设计阵转化为列满秩.

(1). 假设矩阵$\mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)'$列满秩.

(2). 假设$y_1, \cdots, y_n$ 相互独立.

Logistic模型参数的常见的两种估计

- ML估计
- WLS估计 （只适用于重复测量数据）

Logistic 回归模型的ML估计

- 似然函数

$$
\begin{aligned}
L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i) &= \prod_{i=1}^{n} \pi^{y_i}(\boldsymbol{x}_i)(1-\pi(\boldsymbol{x}_i))^{1-y_i} \\
&= \prod_{i=1}^{n} (1-\pi(\boldsymbol{x}_i)) \exp\left(y_i \ln\left(\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)} \right) \right) \\
&= \prod_{i=1}^{n} \left(\frac{1}{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})} \right) \exp\left(y_i(\boldsymbol{x}_i'\boldsymbol{\beta}) \right),
\end{aligned}
$$

对数似然函数

$$
l(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \boldsymbol{x}_i'\boldsymbol{\beta} - \ln\left(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right) \right\}.
$$

$\boldsymbol{\beta}$的ML估计$\widehat{\boldsymbol{\beta}}$ 就是使得$l(\boldsymbol{\beta})$达到最大的$\boldsymbol{\beta}$.

Logistic 回归模型的ML 估计

- 需通过迭代求解$\widehat{\beta}$. 记$\widehat{\beta}$为最终的迭代解. 于是拟合的响应函数

$$\widehat{\pi}(\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}.$$

由ML估计$\widehat{\beta}$的渐近正态性得, 当对应较大的n, 近似有

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \dot{\sim} N_p(0, \left(\mathbf{X}'\mathbf{W}(\boldsymbol{\beta})\mathbf{X}\right)^{-1}).$$

这里, $\mathbf{W}(\boldsymbol{\beta}) = \text{diag}(w_1, \cdots, w_n)$,

$$w_i = \frac{\exp(\boldsymbol{x}_i'\boldsymbol{\beta})}{\left(1+\exp\left(\boldsymbol{x}_i'\boldsymbol{\beta}\right)\right)^2} = \pi(\boldsymbol{x}_i)(1-\pi(\boldsymbol{x}_i)).$$

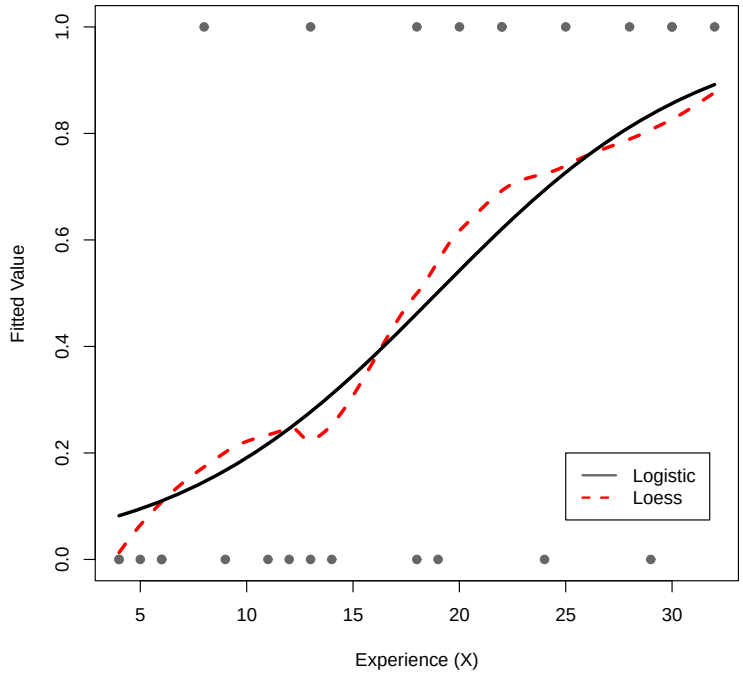
可求参数β及其分量的置信区间和假设检验（类似线性模型）.

例10.3.1 (Kutner et al., 2004)

为了研究计算机编程经验(X)对在规定时间内完成编程任务的能力(Y)的影响，选取了具有不同的编程经验（以经验月数衡量）的25 人，规定时间内完成编程任务的能力采用0 ,1 表示，1表示完成，0表示未完成. 数据如下

i	X	Y	i	X	Y	i	X	Y
1	14	0	10	6	0	19	24	0
2	29	0	11	30	1	20	13	1
3	6	0	12	11	0	21	19	0
4	25	1	13	30	1	22	4	0
5	18	1	14	5	0	23	28	1
6	4	0	15	20	1	24	22	1
7	18	0	16	13	0	25	8	1
8	12	0	17	9	0			
9	22	1	18	32	1			

例10.3.1



- 绘制数据的散点图和局部加权散点平滑回归响应曲线(LOWESS). 因响应变量为0、1 变量, 该散点图的信息量不大, 只是表明成功完成任务的能力似乎随着经验的增加而提高.

例10.3.1

- 由R语言的函数glm()求得β_0和β_1的ML估计

$$\widehat{\beta}_0 = -3.05970, \quad \widehat{\beta}_1 = 0.16149,$$

故拟合的logistic回归函数

$$\widehat{\pi}(x) = \frac{\exp(-3.05970 + 0.16149x)}{1 + \exp(-3.05970 + 0.16149x)}, \tag{10.11}$$

- 对于任意给定的x, 都有

$$\beta_1 = \text{logit}(\pi(x+1)) - \text{logit}(\pi(x)) = \ln\left(\frac{\pi(x+1)}{1+\pi(x+1)} \Big/ \frac{\pi(x)}{1+\pi(x)}\right),$$

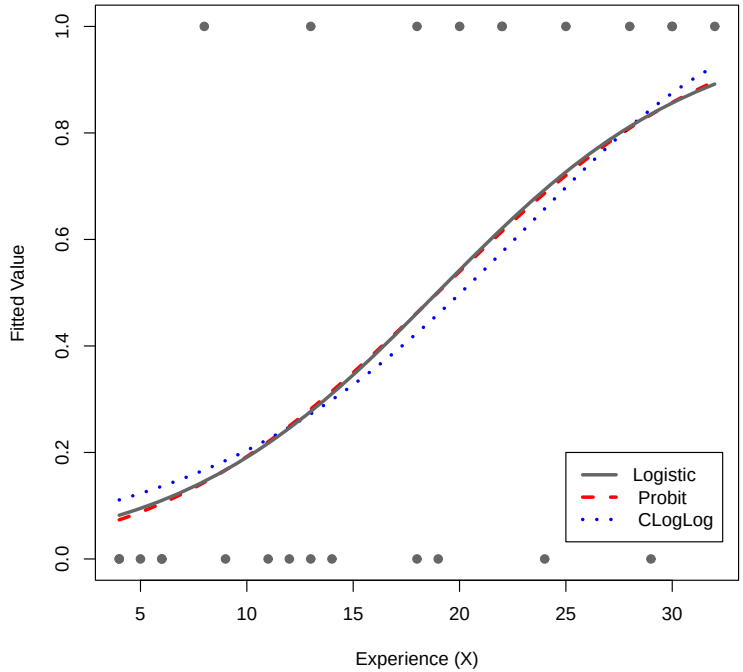
故自变量增加一个单位的优势比OR可通过$\widehat{\beta}_1$估算，即为

$$\widehat{\text{OR}} = \exp(\widehat{\beta}_1) = 1.17526,$$

说明随着计算机编程经验增加，在规定时间内完成编程任务的概率也会提高.

例10.3.1

- 比较Logistic 回归、probit回归及cloglog回归的结果



可看出：Logistic拟合曲线和probit拟合曲线非常接近, 而cloglog拟合曲线稍有不同.

Logistic 回归模型的WLS 估计(重复观测数据)

- 设$y_1, \cdots, y_n$是对m个不同的自变量$\boldsymbol{x}_i$条件下的n次观测, 其中对应于$\boldsymbol{x}_i$观测为n_i次, 记r_i为n_i次观测中感兴趣的事件A发生了次数. 于是, 在自变量取$\boldsymbol{x}_i$条件下, 事件A发生的概率可以用频率

$$\widetilde{\pi}_i = r_i/n_i$$

来估计. 结合(10.5), 令

$$y_i^* = \ln\left(\frac{\widetilde{\pi}_i}{1-\widetilde{\pi}_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \cdots, m. \tag{10.12}$$

假设$\varepsilon_1, \cdots, \varepsilon_m$相互独立, 且$E(\varepsilon_i) = 0$ 和$\mathrm{Var}(\varepsilon_i) = v_i$.

模型(10.12) 就是一个拟合数据$(y_i^*, \boldsymbol{x}_i)$的异方差的线性模型.

Logistic 回归模型的WLS 估计

- 若v_i已知, 由线性模型的知识(本书第4章)知：该模型的BLU估计为

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\boldsymbol{y}^*,$$

其中

$$\boldsymbol{y}^* = (y_1^*, \cdots, y_m^*)', \quad \mathbf{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m)', \quad \mathbf{V} = \mathrm{diag}(v_1, \cdots, v_m)$$

- 对于未知的v_i，可用其估计代替. v_i的一个常用的估计

$$\widehat{v}_i = \frac{1}{n_i\widetilde{\pi}_i(1-\widetilde{\pi}_i).}$$

补充：v_i 的估计的推导

- $\widetilde{\pi}_i$是样本的频率. 由中心极限定理, 当n_i趋于无穷时, 有

$$\sqrt{n_i}(\widetilde{\pi}_i - \pi_i) \to N(0, \pi_i(1-\pi_i)).$$

由于$g(z) = \ln(z/(1-z))$的一阶导数存在, 且

$$g'(z) = \frac{\mathrm{d}g(z)}{\mathrm{d}z} = \frac{1}{z(1-z)},$$

于是, 应用Delta方法可证得：当$n_i \to \infty$时, 有

$$\sqrt{n_i}(\mathrm{logit}(\widetilde{\pi}_i) - \mathrm{logit}(\pi_i)) \to N\left(0, \frac{1}{\pi_i(1-\pi_i)}\right).$$

因此, 当$\min\{n_1, \cdots, n_m\}$充分大时, 可认为y_i^*的方差v_i 近似为$1/(n_i\pi_i(1-\pi_i))$, 于是得到v_i的一个估计$\widehat{v}_i$.

Logistic 回归模型的WLS 估计

注 当$r_i = 0$或n_i时, $\widehat{\pi}_i = 0$ 或$\widehat{\pi}_i = 1$, 此时, 样本logit $= \ln(\widetilde{\pi}_i/(1-\widetilde{\pi}_i))$ 就没有定义. 因此，应用中常采用修正样本logit，即

$$\ln\left((r_i + 0.5)/(n_i - r_i + 0.5)\right).$$

例10.3.2 (Rao和Toutenbuig(1995))

考察牙齿脱落Y与年龄X的关系. 数据见下表.

Table: 5×2的频数列联表

i	年龄组(X)	牙齿脱落Y		$n_{i\cdot}$
		是$(Y=1)$	否$(Y=0)$	
1	< 40	4	70	74
2	$40-49$	28	147	175
3	$50-59$	38	207	245
4	$60-69$	51	202	253
5	≥ 70	32	92	124
	$n_{\cdot j}$	153	718	871

例10.3.2

解 首先对列联表进行χ^2 独立性检验和Wilks基于似然比的G^2 独立性检验. 计算得$\chi^2 = 15.5575$和$G^2 = 17.2489$, 查表得$\chi^2_4(0.05) = 9.4877$, 因此，两个检验都表明牙齿脱落Y和年龄X有关.

- 采用Logistic回归模型拟合数据. 依照公式

$$y_i^* = \text{logit}(\widetilde{\pi}_i) = \ln\left(\frac{\widetilde{\pi}_i}{1-\widetilde{\pi}_i}\right) = \ln\left(\frac{n_{1i}}{n_{2i}}\right),$$

$$\widehat{v}_i = \frac{1}{n_{i\cdot}\widetilde{\pi}_i(1-\widetilde{\pi}_i)} = \frac{1}{n_{i\cdot}(n_{i1}/n_{i\cdot})(1-n_{i1}/n_{i\cdot})} = \frac{n_{i\cdot}}{n_{i1}n_{i2}},$$

例10.3.2

Table: 各组$(y_i^*, \widehat{v}_i)$的值

i	年龄组(X)	$\widetilde{\pi}_i = \frac{n_{i1}}{n_{i\cdot}}$	$y_i^* = \ln\left(\frac{n_{i1}}{n_{i2}}\right)$	$\widehat{v}_i$
1	< 40	0.0541	-2.8622	0.2643
2	$40 - 49$	0.1600	-1.6582	0.0425
3	$50 - 59$	0.1551	-1.6951	0.0311
4	$60 - 69$	0.2026	-1.3764	0.0246
5	≥ 70	0.2581	-1.0561	0.0421

- 对Logistic变换后的数据y^*, 采用方差分析模型：

$$\ln\left(\frac{n_{i1}}{n_{i2}}\right) = y_i^* = \mu + \alpha_i, \tag{10.13}$$

其中α_i为第i组效应, 满足约束条件$\sum_{i=1}^{5} \alpha_i = 0$.

例10.3.2

借助于y_i^*的渐近性质可对模型参数进行推断. 如检验牙齿脱落Y与年龄X是否有关, 就等价于检验假设

$$H_0: \ \alpha_1 = \cdots = \alpha_5 = 0 \tag{10.14}$$

是否成立. 注意到(10.14) 等价于

$$H_0 : \mathbf{C}\boldsymbol{\mu} = \mathbf{0},$$

这里$\boldsymbol{\mu} = (\mu_1, \cdots, \mu_5)$, $\mu_i = \mu + \alpha_i$,

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}.$$

例10.3.2

- 利用$\boldsymbol{y}^*$的渐近正态性和正态向量线性变换的性质, H_0成立时, 对于较大的$\min\{n_{i\cdot}\}$, 可近似认为

$$\mathbf{C}\boldsymbol{y}^* \sim N_4(\mathbf{0}, \mathbf{C}\mathbf{V}\mathbf{C}'),$$

$$\chi^2 = \boldsymbol{y}^{*\prime}\mathbf{C}'(\mathbf{C}\widehat{\mathbf{V}}\mathbf{C}')^{-1}\mathbf{C}\boldsymbol{y}^* \sim \chi^2_4.$$

计算得统计量$\chi^2 = 14.1319 > \chi^2_4(0.05) = 9.4877$, 因此, 拒绝原假设$H_0$, 认为牙齿脱落$Y$与年龄$X$有关. 效应对照$\alpha_2 - \alpha_1$的置信系数为95%的置信区间为

$$y_2^* - y_1^* \mp z_{0.025}(\widehat{v}_1 + \widehat{v}_2) = [0.6027, 1.8053].$$

但从表4不难发现：随着年龄段的上升, $y_i^* = \text{logit}(\widehat{\pi}_i)$ 增大.

例10.3.2

- 按照年龄组中点(或上边界减5或下边界加5)定义组的平均年龄, 记$\boldsymbol{x} = (35, 45, 55, 65, 75)'$. 建立Logistic回归模型

$$\ln\left(\frac{\widehat{\pi}_i(x_i)}{1-\widehat{\pi}_i(x_i)}\right) = \beta_0 + \beta_1 x_i, \quad i = 1, \cdots, 5. \tag{10.15}$$

首先用y^*替换理论logit值, 计算得参数β_0 和β_1 的WLS估计：

$$\widehat{\beta}_0 = -4.4541, \quad \widehat{\beta}_1 = 0.0479.$$

于是$\text{logit}(\pi(x_i))$的拟合值和$\pi(x_i)$的估计值分别为

$$\widehat{y}_i^* = -4.4541 + 0.0479 x_i, \quad \widehat{\pi}(x_i) = \frac{\exp(\widehat{y}_i^*)}{1+\exp(\widehat{y}_i^*)}.$$

进一步得到估计的频数$(n_i.\widehat{\pi}_i(x_i))$.

例10.3.2

Table: 观测结果和拟合结果对比

x_i	y_i^*	$\widehat{y}_i^*$	$n_{i1}/n_{i\cdot}$	$\widehat{\pi}(x_i)$	$n_{i\cdot}\widehat{\pi}(x_i)$	n_{i1}
35	-2.8622	-2.7776	0.054	0.059	4.366	4
45	-1.6582	-2.2986	0.160	0.091	15.925	28
55	-1.6951	-1.8196	0.155	0.139	34.055	38
65	-1.3764	-1.3406	0.202	0.207	52.371	51
75	-1.0561	-0.8616	0.258	0.297	36.828	32

估计logit值$\widehat{y}_i^*$ 与样本logit值y_i^*、估计的概率$\widehat{\pi}(x_i)$与样本频率$n_{i1}/n_{i\cdot}$、估计的频数$n_{i\cdot}\widehat{\pi}_i(x_i)$ 与观测到的频数n_{i1} 两两都很接近，这说明用Logistic回归模型拟合牙齿脱落Y与年龄X的关系是合适的.

Logistic回归模型下的检验

1. 单个回归系数的显著检验: Wald 检验

   检验问题为

$$H_0 : \beta_k = 0 \longleftrightarrow H_1 : \beta_k \neq 0.$$

   可借助于极大似然估计的渐近性质,得到一个近似检验统计量

$$z = \frac{\widehat{\beta}_k}{v(\widehat{\beta}_k)},$$

   其中$v(\widehat{\beta}_k)$为矩阵$\left(\mathbf{X}'\mathbf{W}(\widehat{\boldsymbol{\beta}})\mathbf{X}\right)^{-1}$的第$k+1$个对角元, $k = 1, \cdots, p-1$. 当$|z| \geq z_{\alpha/2}$ 时，拒绝H_0, 否则接受H_0.

- R程序中函数glm() 自带这个检验.

Logistic回归模型下的检验

2. 多个回归系数同时为零的检验: 似然比检验

检验问题为

$$H_0 : \beta_q = \cdots = \beta_{p-1} = 0, \quad H_1 : \beta_q, \cdots, \beta_{p-1}\text{不全为零}.$$

记$\boldsymbol{\beta}_R = (\beta_1, \cdots, \beta_{q-1})', \boldsymbol{X}_R = (1, X_1, \cdots, X_{q-1})'$.

似然比检验的检验统计量

$$G^2 = -2 \ln \left(\frac{L_R}{L_F} \right), \tag{10.16}$$

其中L_R, L_F分别为减模型和全模型下的似然函数最大值.

对于较大的n, 当H_0成立时，G^2近似服从χ^2_{p-q}, 故H_0的拒绝域为

$$G^2 \geq \chi^2_{p-q}(\alpha).$$

Logistic回归模型下的检验

3. Logistic回归模型的拟合优度检验

检验假设

$$H_0 : E(Y) = \frac{1}{1+\exp(-\boldsymbol{X}'\boldsymbol{\beta})} \longleftrightarrow H_1 : E(Y) \neq \frac{1}{1+\exp(-\boldsymbol{X}'\boldsymbol{\beta})}.$$

本节主要介绍三种常见的检验方法：

- 皮尔逊卡方检验　（适用于多个重复观测数据）
- 偏差拟合优度检验　（适用于多个重复观测数据）
- Hosmer-Lemeshow检验　（适用于未重复或少重复观测）

Logistic回归模型下的检验

- Pearson卡方拟合优度检验

  假定自变量有c个组合$\boldsymbol{x}_1, \cdots, \boldsymbol{x}_c$, y_{ij}, $i = 1, \cdots, n_j$, $j = 1, \cdots, c$, 是响应变量Y在自变量$\boldsymbol{x}_j$下的第i 次重复独立观察($y_{ij} = 1$或0), n_j 表示自变量$\boldsymbol{x}_j$下案例数. 于是, 有

$$O_{j1} = \sum_{i=1}^{n_j} y_{ij} = y_{\cdot j} \sim B(n_j, \pi(\boldsymbol{x}_i)),$$

$$O_{j0} = \sum_{i=1}^{n_j} (1 - y_{ij}) = n_j - y_{\cdot j}, \quad j = 1, \cdots, c.$$

  如果Logistic模型合适, 则

$$E(O_{j1}) = n_j \pi(\boldsymbol{x}_j) = \frac{n_j}{1 + \exp(-\boldsymbol{x}_j' \boldsymbol{\beta})}.$$

Logistic回归模型下的检验

于是, 在自变量$\boldsymbol{x}_j$下$y_{ij}=1$和$y_{ij}=0$的平均个数可分别由

$$E_{j1}=n_j\widehat{\pi}(\boldsymbol{x}_j)=\frac{n_j}{1+\exp(-\boldsymbol{x}_j'\widehat{\boldsymbol{\beta}})},\quad E_{j0}=n_j-E_{j1}$$

来估计, 其中, $\widehat{\boldsymbol{\beta}}$为$\boldsymbol{\beta}$的极大似然估计.

Pearson卡方拟合优度检验统计量

$$\chi^2=\sum_{i=1}^{c}\sum_{k=0}^{1}\frac{(O_{j0}-E_{j0})^2}{E_{jk}},$$

当$c>p$且$\min_j\{n_{\cdot j}\}$较大时, 它近似χ^2_{c-p}分布.

注 Pearson 卡方拟合优度检验只适合自变量各组$\boldsymbol{x}_j$下重复观测次数$n_{\cdot j}$都比较大的情形，当样本量不大时要谨慎对待检验结果.

Logistic回归模型下的检验

- 偏差拟合优度检验

  记Logistic回归模型和约减模型为

$$E(y_{ij}) = \frac{1}{1+\exp(-\boldsymbol{x}_j'\boldsymbol{\beta})},$$

$$E(y_{ij}) = \pi_j, \quad j = 1, \cdots, c.$$

  第二个模型又被称为饱和模型(saturated model). 则Logistic回归模型与饱和模型下响应变量的似然比统计量为

$$\begin{aligned} G^2 &= -2(\ln L_R - \ln L_F) \\ &= -2\sum_{j=1}^{c}\left(O_{j1}\ln\left(\frac{\widehat{\pi}(\boldsymbol{x}_j)}{\widetilde{\pi}_j}\right) + (n_{\cdot j} - O_{j1})\ln\left(\frac{1-\widehat{\pi}(\boldsymbol{x}_j)}{1-\widetilde{\pi}_j}\right)\right) \end{aligned}$$

Logistic回归模型下的检验

其中L_R, L_F分别为拟合的Logistic回归模型与饱和模型似然函数,

$$\widehat{\pi}(\boldsymbol{x}_j) = \frac{1}{1+\exp(-\boldsymbol{x}_j'\widehat{\boldsymbol{\beta}})}, \quad \widetilde{\pi}_j = \frac{O_{j1}}{n_{\cdot j}}.$$

这里, $\widehat{\boldsymbol{\beta}}$为$\boldsymbol{\beta}$的极大似然估计.

称统计量G^2为偏差(deviance), 也称Logistic模型偏差拟合优度检验统计量.

由似然比统计量在H_0下的性质：当$c > p$且$\min_j\{n_{\cdot j}\}$较大时, 检验统计量G^2近似服从自由度为$c-p$的χ^2分布, 故H_0的拒绝域为

$$G^2 \geq \chi^2_{c-p}.$$

Logistic回归模型下的检验

- Hosmer-Lemeshow 拟合优度检验

  步骤如下：

  (1) 根据拟合的$\widehat{\pi}(\boldsymbol{x}_j)$ 对自变量进行分组；

  (2) 将$\widehat{\pi}(\boldsymbol{x}_j)$相近的或根据

$$\text{logit}(\widehat{\pi}(\boldsymbol{x}_j)) = \boldsymbol{x}_j\widehat{\boldsymbol{\beta}}$$

  的大小自变量分成5 –10组；

  (3) 然后采用Pearson 卡方拟合优度检验作检验.

该拟合优度检验主要是针对自变量$\boldsymbol{x}_j$下没有重复($n_{\cdot j} = 1$)或重复观测数$n_{\cdot j}$较小的情形提出的.

Logistic回归模型的诊断

假设因变量仅取0和1两个值, 同一自变量$\boldsymbol{x}_i$条件下没有重复测量和观测, 并记$(y_i, \boldsymbol{x}_i)\ i = 1, \cdots, n$ 为观测数据集, $\widehat{\boldsymbol{\beta}}$为Logistic回归模型下$\boldsymbol{\beta}$的极大似然估计.

- Logistic回归普通残差

$$\widehat{e}_i = y_i - \widehat{\pi}_i = \begin{cases} 1 - \widehat{\pi}_i, & y_i = 1, \\ -\widehat{\pi}_i, & y_i = 0, \end{cases} \tag{11.1}$$

其中

$$\widehat{\pi}_i = \widehat{\pi}(\boldsymbol{x}_i) = \frac{1}{1 + \exp(\boldsymbol{x}_i' \widehat{\boldsymbol{\beta}})}.$$

Logistic回归残差

- Logistic回归的Pearson残差

$$r_{\mathrm{p}_i} = \frac{y_i - \widehat{\pi}_i}{\widehat{\pi}_i(1-\widehat{\pi}_i)}. \tag{11.2}$$

即就是普通残差除以y_i的标准差估计$\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)}$.

与Pearson卡方拟合优度检验统计量的关系

$$\begin{aligned}\chi^2 &= \sum_{j=1}^{c}\frac{(O_{j0}-E_{j0})^2}{E_{j0}} + \sum_{j=1}^{c}\frac{(O_{j1}-E_{j1})^2}{E_{j1}} = \sum_{i=1}^{n}\frac{(y_i-\widehat{\pi}_i)^2}{\widehat{\pi}_i(1-\widehat{\pi}_i)} \\ &= \sum_{i=1}^{n} r_{P_i}^2.\end{aligned}$$

Logistic回归残差

- Logistic回归的学生化Pearson残差

$$r_{sP_i} = \frac{r_{P_i}}{1 - h_{ii}}, \tag{11.3}$$

这里w_{ii} 为帽子矩阵

$$\mathbf{H} = \widehat{\mathbf{W}}^{1/2}\mathbf{X}(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}\widehat{\mathbf{W}}^{1/2}$$

的第i对角元, $\widehat{\mathbf{W}} = \text{diag}(\widehat{\pi}_1(1-\widehat{\pi}_1), \cdots, \widehat{\pi}_n(1-\widehat{\pi}_n))$.

考虑到Pearson残差中用$\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)}$估计标准差$\sqrt{\pi_i(1-\pi_i)}$所引起的随机波动, 学生化Pearson残差是一般残差$\widehat{e}_i$的一个更好的标准化, 它是将$\widehat{e}_i$除以其标准差的近似估计$\sqrt{\widehat{\pi}_i(1-\widehat{\pi}_i)(1-w_{ii})}$.

Logistic回归残差

- Logistic回归的偏差残差(deviance residual)

$$dev_i = \text{sign}(\hat{e}_i)\left(-2(y_i \ln \widehat{\pi}(x_i) + (1-y_i)\ln(1-\widehat{\pi}(x_i)))\right)^{1/2}, \quad (11.4)$$

其中$\hat{e}_i = y_i - \widehat{\pi}(x_i)$ 为原始残差或响应残差.

与Logistic 回归的偏差拟合优度检验统计量的关系

$$\begin{aligned} G^2 &= -2\sum_{i=1}^{n}\left[y_i \ln\left(\frac{\widehat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\widehat{\pi}_i}{1-y_i}\right)\right] \\ &= -2\sum_{i=1}^{n}\left[y_i \ln\left(\widehat{\pi}_i\right) + (1-y_i)\ln\left(1-\widehat{\pi}_i\right)\right] = \sum_{i=1}^{n} dev_i^2 = DEV, \end{aligned}$$

Logistic回归残差图

两种残差图:

- 带LOWESS平滑曲线的残差-估计概率散点图

LOWESS: 局部加权回归（Locally Weighted Scatterplot Smoothing）

- 基于偏差残差的半正态(half-normal)概率图.

半正态分布概率图是基于偏差残差绝对值

$$|\mathrm{dev}_1|, \cdots, |\mathrm{dev}_n|$$

的排序和标准正态分布分位数：

$$\Phi^{-1}((k+n-1/8)/(2n+1/2)).$$

带模拟包络(simulation envelope)

不像线性模型, Logistic回归的残差图的作用主要是用于直观评价模型拟合是否合适.

带模拟包络的偏残差绝对值的半正态概率图

具体步骤如下:

(I) 对每个i, 从两点分布$B(1,\hat{\pi}_i)$随机抽取y_{si}, $i=1,\cdots,n$;

(II) 用模拟抽取的n个响应变量和原自变量$\boldsymbol{x}_i$作Logistic回归，计算偏差残差dev_{si}, 并对其绝对值排序：

$$|\mathrm{dev}|_{s(1)},\cdots,|\mathrm{dev}|_{s(n)};$$

(III) 重复步骤(I)和(II) S次, k从1到n依次计算各点在S 次重复中所得的残差绝对值集合$\{|\mathrm{dev}|_{s(k)}, s=1,\cdots,S\}$的最大值$\max|\mathrm{dev}|_{s(k)}$、最小值$\min|\mathrm{dev}|_{s(k)}$，平均值$\overline{|\mathrm{dev}|}_{s(k)}$;

(IV) 画以下散点图:

模拟上边界 $\left(\Phi^{-1}((k+n-1/8)/(2n+1/2)),\ \max|\mathrm{dev}|_{s(k)}\right),\quad k=1,\cdots,n,$

模拟下边界 $\left(\Phi^{-1}((k+n-1/8)/(2n+1/2)),\ \min|\mathrm{dev}|_{s(k)}\right),\quad k=1,\cdots,n,$

模拟均值线 $\left(\Phi^{-1}((k+n-1/8)/(2n+1/2)),\ \overline{|\mathrm{dev}|}_{s(k)},\quad k=1,\cdots,n,\right.$

模型残差散点图 $\left(\Phi^{-1}((k+n-1/8)/(2n+1/2)),\ |\mathrm{dev}|_{(k)}\right),\quad k=1,\cdots,n.$

Logistic回归影响分析

- 识别强影响点的三种度量方法：

$$\Delta\chi_i^2 = \chi^2 - \chi_{(i)}^2,$$

$$\Delta \text{dev}_i = \text{DEV} - \text{DEV}_{(i)},$$

$$D_i = \frac{r_{\text{P}_i} h_{ii}}{p(1-h_{ii})^2}$$

其中χ^2和DEV分别为Pearson 卡方统计量和偏差统计量，$\chi_{(i)}^2$和$DEV_{(i)}$分别为样本去掉第i次观测后所计算的相应的值，D_i 为模型$y_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \varepsilon_i,\ i = 1, \cdots, n$ 全数据和去掉第i次观测后数据拟合值的Cook距离，这里，

$$y_i^* = \ln\left(\widehat{\pi}_i/(1-\widehat{\pi}_i)\right).$$

响应概率的推断

- 在Logistic模型(10.5)下, $\pi(\boldsymbol{x}_i)$的ML 估计为

$$\widehat{\pi}(\boldsymbol{x}_i) = 1/[1 + \exp(-\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})],$$

其中$\widehat{\boldsymbol{\beta}}$为$\boldsymbol{\beta}$的极大似然估计.

为方便应用, 常采用列线图(alignment diagram) 又称诺莫(nomogram) 图将Logistic回归可视化, 即将复杂的回归方程, 转变为可视化的图形

Logistic回归模型下的置信区间

- $\text{logit}(\pi(\boldsymbol{x}_i))$, 即$\boldsymbol{x}_i'\boldsymbol{\beta}$ 的置信系数为$1-\alpha$的置信区间近似为$[L, U]$,

$$L = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} - z_{\alpha/2}\sqrt{\boldsymbol{x}_i'\left(\mathbf{X}'\mathbf{W}(\widehat{\boldsymbol{\beta}})\mathbf{X}\right)^{-1}\boldsymbol{x}_i}, \tag{11.5}$$

$$U = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}} + z_{\alpha/2}\sqrt{\boldsymbol{x}_i'\left(\mathbf{X}'\mathbf{W}(\widehat{\boldsymbol{\beta}})\mathbf{X}\right)^{-1}\boldsymbol{x}_i}. \tag{11.6}$$

- $\pi(\boldsymbol{x}_i))$的置信系数为$1-\alpha$的置信区间近似为

$$[1/(1+\exp(-L)), 1/(1+\exp(-U))]. \tag{11.7}$$

- 关于m个点$\boldsymbol{x}_{i_1}, \cdots, \boldsymbol{x}_{i_m}$处的响应概率$\pi(\boldsymbol{x})$的同时置信区间, 将包含在(11.7)中$L$和$U$内的分位数$z_{\alpha/2}$换成$z_{\alpha/2m}$即可.

Logistic回归判别

- Logistic回归判别就是基于估计的发生概率

$$\widehat{\pi}(\boldsymbol{x}_{new})$$

和一个给定的阈值(cutoff) c 来作判别.

判别方法如下：

$$\text{当 } \widehat{\pi}(\boldsymbol{x}_{new}) > c \text{ 时, 则判 } Y_{new} = 1,$$

$$\text{当 } \widehat{\pi}(\boldsymbol{x}_{new}) \leq c \text{ 时, 则判 } Y_{new} = 0.$$

Logistic回归判别

- 关于阈值c的选择, 主要有以下三种方法：

  (1) 选择$c = 1/2$.

  适用于$Y = 0$和$Y = 1$等可能, 且$Y = 0$和$Y = 1$被判错的代价(cost)相等;

  (2) 基于受试者工作特征(receiver operating characteristic, ROC) 曲线的最佳阈值选取方法.

  适合于样本是从总体中随机抽取的, 且两类错判代价(cost)近似相等的情形;

  (3) 基于先验概率和错判代价的阈值选择.

  适用于当样本不是随机抽取但对总体中$Y = 1$和$Y = 0$可能性有先验信息，对两类错判代价不同度量的情形.

Logistic回归模型下的检验

- 基于ROC曲线的最佳阈值选取方法.

Table: 混淆矩阵

		真实值	
		Positive ($Y=1$)	Negative ($Y=0$)
预测值	Positive ($\widehat{Y}=1$)	TP	FP
	Negative ($\widehat{Y}=0$)	FN	TN

其中TP代表模型分类中真阳性的样本数量;
FP代表模型分类中假阳性的样本数量;
TN代表模型分类中真阴性的性样本数量;
FN代表分类中假阴性的性样本数量.

Logistic回归模型下的检验

- 计算真阳性率(true positive rate, TPR), 真阴性率(true negative rate, TNR)以及假阴性率(false negative rate, FNR)：

$$\mathrm{TPR} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) = \mathrm{sensitivity}(\mathrm{se}),$$

$$\mathrm{TNR} = \mathrm{TN}/(\mathrm{TN}+\mathrm{FP}) = \mathrm{specificity}(\mathrm{sp}),$$

$$\mathrm{FNR} = \mathrm{FN}/(\mathrm{TN}+\mathrm{FP}) = 1-\mathrm{TNR}$$

- TPR和FNR随着阈值的变化而变化；
- ROC曲线就是随着阈值的变化, 以1-sp 和se 分别为横轴和纵轴绘制的曲线；
- 最佳阈值点通常通过Youden指数进行选择，

$$\text{Youden index} = \mathrm{se}-(1-\mathrm{sp}) = \mathrm{se}+\mathrm{sp}-1.$$

案例分析

在一项健康研究中, 为了调查一种由蚊子传播的疾病暴发情况, 在一个城市的两个区域内随机抽取98个人. 通过访谈提问提出相关问题, 评估近期是否出现了与该疾病相关的某些特定症状, 从而确定他们近期是否感染了所研究的疾病. 研究中还包括三个已知或潜在的风险因素(自变量).

- 年龄(X_1);
- 家庭的社会经济地位:高收入、中收入和低收入(三分类变量); 为可估, 家庭收入类型由以下两个虚拟指标变量(X_2 和X_3) 表示.
- 所居住的城市区域(二分类, $X_4 = 0, 1$)

解 采用包含所有变量的Logistic模型

$$\pi_F = 1/[1+\exp(-\beta_0-\beta_1X_1-\beta_2X_2-\beta_3X_3-\beta_4X_4)] \qquad (11.8)$$

拟合数据. 该模型的计算可由R语言中的函数glm()或lrm()完成. 从以上结果可得β的ML估计:

$$\widehat{\boldsymbol{\beta}} = (-2.31293, 0.02975, 0.40879, -0.30525, 1.57475)'.$$

从而得拟合的Logistic响应函数

$$E(Y=1) = \pi(\boldsymbol{X}) = (1+\exp(-\boldsymbol{X}'\widehat{\boldsymbol{\beta}}))^{-1}$$

- 得4个自变量各自的优势比：

$$\widehat{\mathrm{OR}}_1 = \exp(\widehat{\beta}_1) = 1.0302, \quad \widehat{\mathrm{OR}}_2 = \exp(\widehat{\beta}_2) = 1.5050,$$

$$\widehat{\mathrm{OR}}_3 = \exp(\widehat{\beta}_3) = 0.7369, \quad \widehat{\mathrm{OR}}_4 = \exp(\widehat{\beta}_4) = 4.8295,$$

ML估计$\widehat{\boldsymbol{\beta}}$的渐近协方差阵的估计：

$$\mathbf{S} = (\mathbf{X}'\mathbf{W}(\widehat{\boldsymbol{\beta}})\mathbf{X})^{-1}$$

$$= \begin{pmatrix} 0.4129 & -0.0057 & -0.1836 & -0.2010 & -0.1632 \\ -0.0057 & 0.0002 & 0.0011 & 0.0007 & 0.0003 \\ -0.1836 & 0.0011 & 0.3588 & 0.1482 & 0.0129 \\ -0.2010 & 0.0007 & 0.1482 & 0.3650 & 0.0623 \end{pmatrix}.$$

案例分析

```
reg<-glm(Y~X1+X2+X3+X4, data, family = binomial(link = "logit") )
summary(reg)

Call:
glm(formula = Y ~ X1 + X2 + X3 + X4, family = binomial(link = "logit"),
    data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.31293    0.64259  -3.599 0.000319 ***
X1           0.02975    0.01350   2.203 0.027577 *
X2           0.40879    0.59900   0.682 0.494954
X3          -0.30525    0.60413  -0.505 0.613362
X4           1.57475    0.50162   3.139 0.001693 **
```

- 以上结果显示：β_2 和β_3不显著.

案例分析

- 采用似然比检验来检验假设

$$H_{02}:\beta_2=\beta_3=0 \longleftrightarrow H_{12}:\beta_2 \text{ 和 } \beta_3 \text{ 不全为零 }.$$

计算似然比检验统计量

$$G^2=-2(\hat{l}_R-\hat{l}_F)=-2[-51.130-(-50.527)]=1.206,$$

统计量的P值为$P(\chi_2^2\geq G^2)=0.55.$
在检验水平$\alpha=0.05$下接受H_{02}. 又由于X_2和X_3是家庭收入变量的两个水平, 故可认为感染所研究的疾病的概率与年龄和所在城区有关, 与家庭收入级别相关性不显著.
综合Wald检验和似然比检验的结果, 得模型:

$$\pi_R=[1+\exp(-(\beta_0+\beta_1X_1+\beta_4X_4))]^{-1}.$$

案例分析

- 采用AIC 和BIC准则变量选择的结果皆得模型:

$$\pi_R = [1 + \exp(-(\beta_0 + \beta_1 X_1 + \beta_4 X_4))]^{-1}.$$

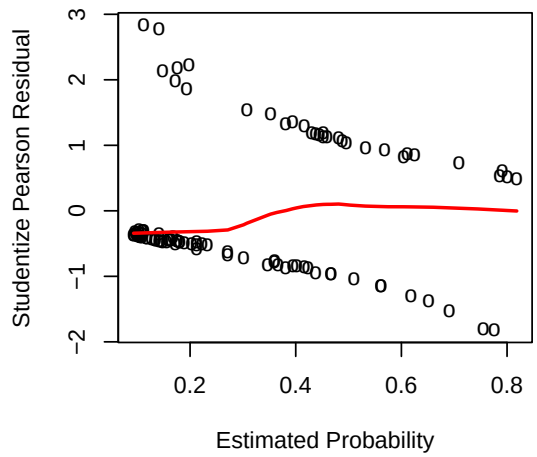
- 基于选模型作带局部加权光滑曲线的四种残差图.

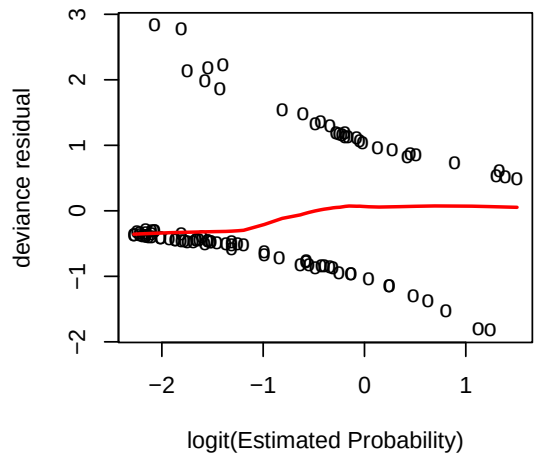
  我们选择更具有实际意义的学生化Pearson残差r_{sP_i}和偏差dev_i,

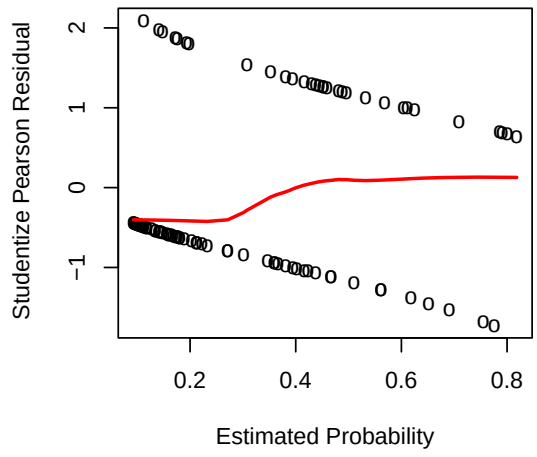
  分别针对$\widehat{\pi}_i$ 和$\text{logit}(\widehat{\pi}_i)$ 作残差图和相应的局部加权光滑曲线.

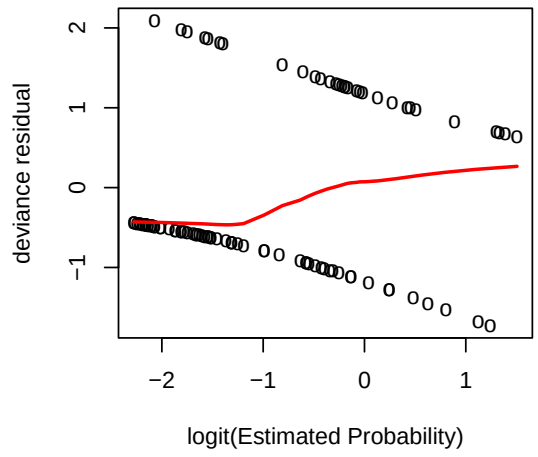
  从中可以看出两类残差无论关于$\widehat{\pi}_i$和还是$\text{logit}(\widehat{\pi}_i)$作局部加权光滑回归, 所得到的Lowess曲线大致都可认为是与斜率截距都为0的直线接近, 说明没有充分的理由拒绝选模型.

案例分析

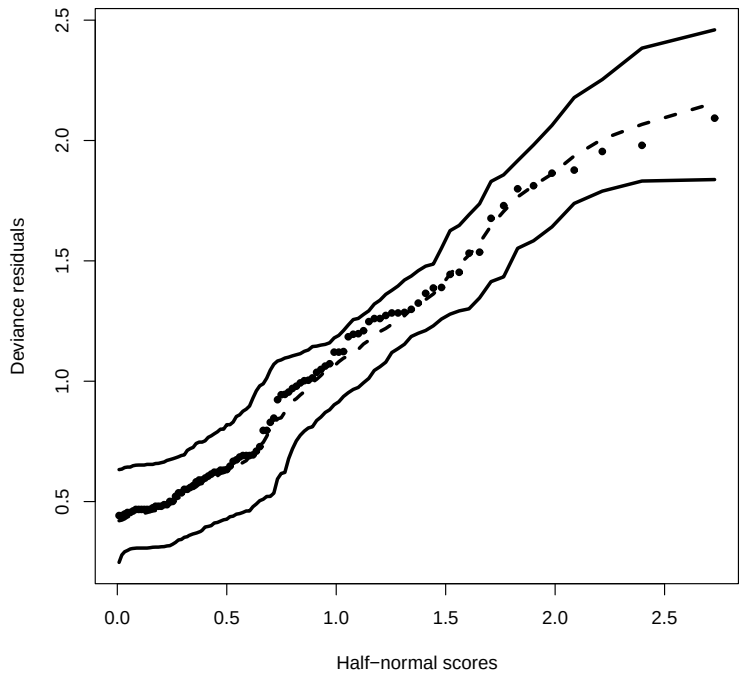

Studentize Pearson Residual
Estimated Probability
0.2
0.4
0.6
0.8
−2
−1
0
1
2
3


deviance residual
logit(Estimated Probability)
−2
−1
0
1
2
3


Studentize Pearson Residual
Estimated Probability
0.2
0.4
0.6
0.8
−1
0
1
2

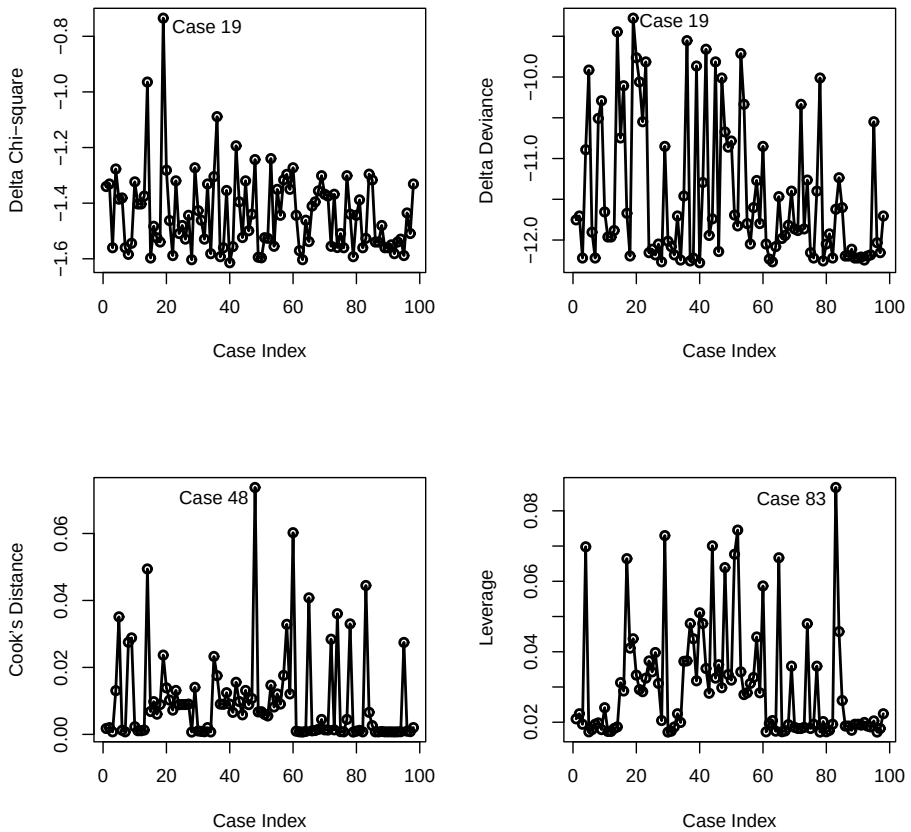

deviance residual
logit(Estimated Probability)
−2
−1
0
1
2

案例分析



- 带模拟包络的偏差残差绝对值的半正态概率图. 由于模型的偏差残差散点大部分靠近模拟均值线, 因此认为所选得Logistic模型是适合的. 尽管存在个别点靠近模拟包络的上边界, 仍在包络内, 因此, 可以认为异常值点不显著.

案例分析



- 不同数据点对应的$\Delta\chi_i^2$, Δdev_i, Cook距离D_i和杠杆值$h_{(ii)}$的差异不大.

案例分析

- 因此最终采用的模型为

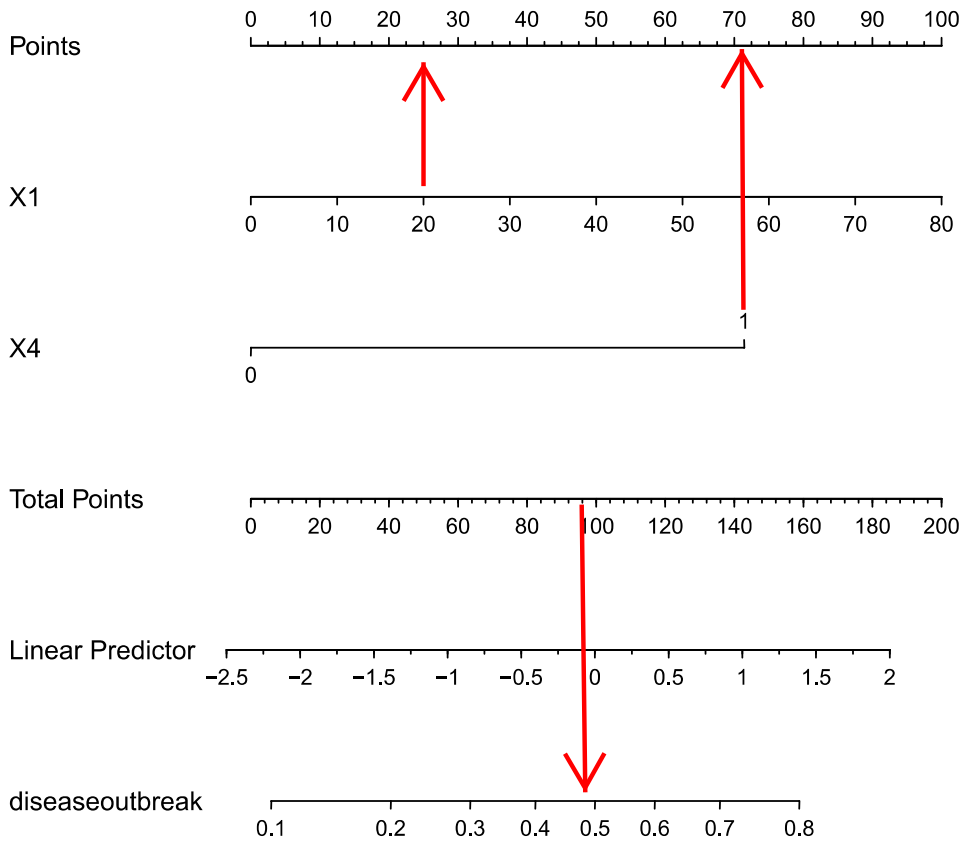
$$\text{logit}(\pi) = -2.33515 + 0.02929X_1 + 1.67345X_4.$$

- 可估算一个居住在区域2的年龄为20岁的居民感染这种疾病的概率为$\hat{\pi} = 0.481$, 线性预测因子(linear predictor)为

$$\text{Logit}(\hat{\pi}) = -0.0759.$$

- 为方便应用, 应用中将Logistic回归模型的概率$\pi(X_1, X_4)$的估算制作成了可视化的列线图. 可轻松估算的响应概率.

案例分析



- $X_1 = 20, X_4 = 1$, 总得分=25+71=96; 疾病暴发预测概率值为0.48左右; 线性预测因子的值在-0.08.

案例分析

- 在点$X_1 = 20$和$X_4 = 1$处, $\text{logit}(\widehat{\pi}) = \boldsymbol{x}'\widehat{\boldsymbol{\beta}}$的方差可被近似估计为

$$s^2 = \boldsymbol{x}'(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\boldsymbol{x} = 0.1193459,$$

其中$\boldsymbol{x} = (1, 20, 1)'$. 由(11.5)和(11.6)可得$\text{logit}(\pi)$ 的置信系数为95%的置信区间上下界：

$$L = -0.0759000 - 1.960 \times \sqrt{0.1193459} = -0.7530109,$$

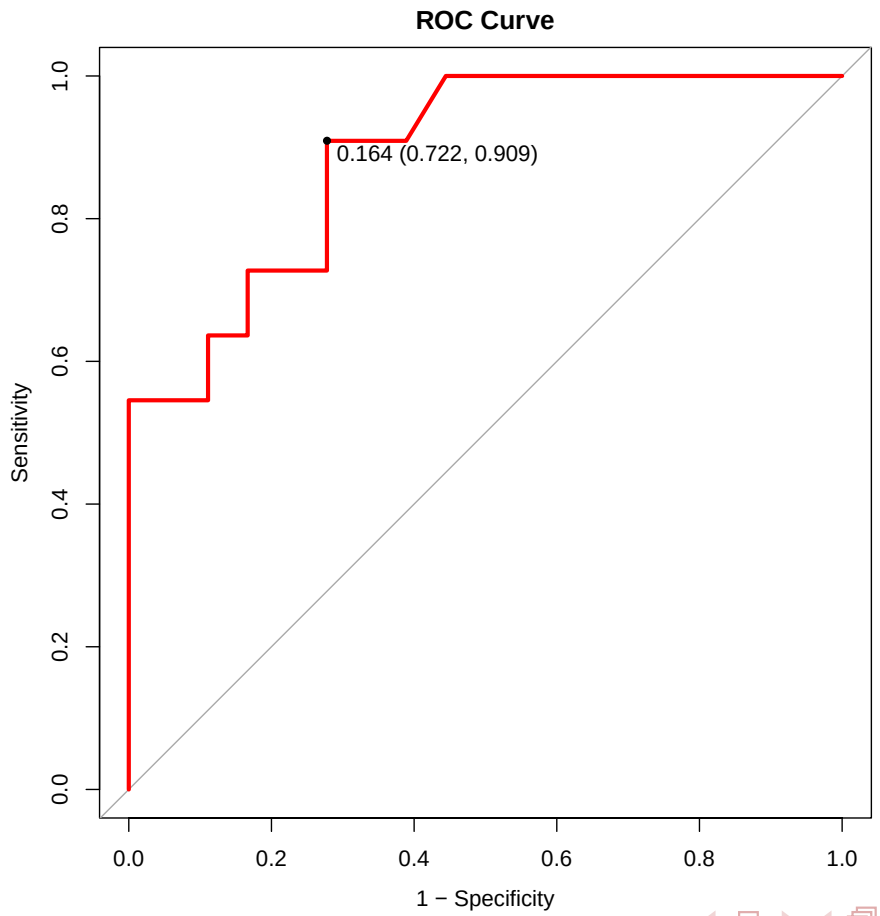
$$U = -0.0759000 + 1.960 \times \sqrt{0.1193459} = -0.6012109.$$

由(11.7) 推得该点响应概率π 的置信系数为95%的置信区间为

$$[0.320,\ 0.646].$$

案例分析

- 基于Logistic模型: $\pi = 1/(1 + \exp(-\beta_0 - \beta_1 X_1 - \beta_4 X_4))$ 作判别. 将数据集随机分成训练集和测试集, 样本量分别为69和29. 得到基于Logistic选模型判别的ROC曲线图



案例分析

- 计算ROC曲线下面积(area under curve, AUC), AUC =0.8864. 基于全模型判别的AUC= $0.846 < 0.8864$. 因此, 从判别的角度说明了选模型优于全模型.
- 最佳阈值点为$c^* = 0.164$,

$$se = P(\hat{y} = 1|Y = 1) = P(\hat{\pi} > 0.164 \mid Y = 1) = \frac{10}{11} = 0.909,$$

$$1 - sp = P(\hat{y} = 1|Y = 0) = P(\hat{\pi} \leq 0.164 \mid Y = 0) = \frac{13}{18} = 0.722,$$

$$\text{Youden index} = se - (1 - sp) = 0.909 - 0.722 = 0.187.$$

错判概率为$1 - 0.187 = 0.813$

注 需要训练样本和需要训练集和测试集的样本量都足够大, 否则基于Logistic模型的预测/判别会由于其错误预测/判别概率过大而变得无用.