第六章　含定性变量的回归模型

Tianxiao Pang

Zhejiang University

December 16, 2018

内容

内容

内容

在许多实际问题中, 人们常常碰到一些定性变量. 譬如性别、教育程度、职业、民族、季节、天气状况等等. 事实上, 定性变量在有关态度和意见调查的社会科学研究中是十分普遍的, 它也常常出现在临床医学研究中. 譬如患者经过手术后是否存活(是/否),受伤害的严重程度(未受伤/轻微伤害/中等伤害/重伤害)等等. 因此, 在对一个实际问题建立回归模型时, 常常需要考虑这些定性变量.

本章主要介绍两种情形: 自变量含有定性变量的统计建模过程; 因变量含有定性变量的统计建模过程.

自变量含有定性变量的回归模型

例6.1.1 在酿酒工艺中, 要将大麦浸在水中吸收一定的水分x_1, 为了提高产量加入某种化学溶剂浸泡一定的时间x_2, 然后测量大麦吸入化学溶剂的份量y, 控制y的量对质量是极为重要的. 由经验知, y和x_1, x_2间有较好的线性关系, 但随着季节不同会有所差异. 现在三个季节各做6次试验, 结果如下表.

表6.1.1:

序号	季节	x_1	x_2	y	序号	季节	x_1	x_2	y
1	冬	130	200	7.5	10	春	138	240	5.6
2	冬	136	200	4.2	11	春	139	220	4.6
3	冬	140	215	1.5	12	春	141	260	3.9
4	冬	138	265	3.7	13	夏	130	205	11.0
5	冬	134	235	5.3	14	夏	140	265	6.0
6	冬	142	260	1.2	15	夏	139	250	6.5
7	春	136	215	6.2	16	夏	136	245	9.1
8	春	137	250	7.0	17	夏	135	235	9.3
9	春	136	180	5.5	18	夏	137	220	7.0

处理这种问题的一种方法是分季度建立回归方程, 然后看看不同季节间方程是否不同.

```
yx=read.table( “***.txt” )
x1=yx[, 1]
x2=yx[, 2]
y=yx[, 3]
wine=data.frame(x1,x2,y)
wine
lm.reg1=lm(y~x1+x2,data=wine,subset=1:6)
summary(lm.reg1)
lm.reg2=lm(y~x1+x2,data=wine,subset=7:12)
summary(lm.reg2)
lm.reg3=lm(y~x1+x2,data=wine,subset=13:18)
summary(lm.reg3)
```

```
> lm.reg1=lm(y~x1+x2,data=wine,subset=1:6)
> summary(lm.reg1)

Call:
lm(formula = y ~ x1 + x2, data = wine, subset = 1:6)

Residuals:
       1        2        3        4        5        6
 0.11071  0.10524 -0.09286  0.06452 -0.25405  0.06643

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 82.659524   2.893225  28.570 9.42e-05 ***
x1          -0.604643   0.024869 -24.313 0.000153 ***
x2           0.016667   0.004159   4.008 0.027870 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1871 on 3 degrees of freedom
Multiple R-squared:  0.9963,    Adjusted R-squared:  0.9938
F-statistic: 399.7 on 2 and 3 DF,  p-value: 0.0002286
```

```
> lm.reg2=lm(y~x1+x2,data=wine,subset=7:12)
> summary(lm.reg2)

Call:
lm(formula = y ~ x1 + x2, data = wine, subset = 7:12)

Residuals:
       7        8        9       10       11       12
-0.27302  0.26290  0.03667 -0.10300  0.21958 -0.14313

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 101.674233  10.846909   9.374  0.00257 **
x1           -0.745615   0.084637  -8.810  0.00308 **
x2            0.028848   0.005676   5.082  0.01472 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2734 on 3 degrees of freedom
Multiple R-squared:  0.9633,    Adjusted R-squared:  0.9388
F-statistic: 39.38 on 2 and 3 DF,  p-value: 0.007028
```

```
> lm.reg3=lm(y~x1+x2,data=wine,subset=13:18)
> summary(lm.reg3)

Call:
lm(formula = y ~ x1 + x2, data = wine, subset = 13:18)

Residuals:
     13      14      15      16      17      18
-0.4056 -0.4648 -0.1066  0.5023  0.3648  0.1099

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 98.14611   13.18155   7.446  0.00501 **
x1          -0.72897    0.12578  -5.796  0.01022 *
x2           0.03915    0.02064   1.897  0.15414
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.513 on 3 degrees of freedom
Multiple R-squared:  0.9585,    Adjusted R-squared:  0.9308
F-statistic: 34.63 on 2 and 3 DF,  p-value: 0.008458
```

分季度求得的回归方程分别是:

冬季: $\hat{y} = 82.66 - 0.605x_1 + 0.0167x_2$,
春季: $\hat{y} = 101.67 - 0.746x_1 + 0.0288x_2$,
夏季: $\hat{y} = 98.146 - 0.729x_1 + 0.0392x_2$.

三个回归方程中, x_1的系数基本相同, x_2的系数也基本相同, 但常数项差异较大. 可对此进行统计检验, 参考例4.1.2(同一模型检验), 即检验在不同的时间段, 某些自变量对因变量的相依关系是否发生了变化.

为了提高精度, 我们将这批数据统一处理, 但需注意季节是定性变量. 我们用引入"虚拟变量"(或称哑变量)的方法来处理此类问题.

当定性变量只取两个可能"值"时, 我们将其中的一个取值所对应的虚拟变量取为1, 而将另一个所对应的虚拟变量取为0. 为了反映冬、春、夏这一季节因素对y的影响, 我们可引入3个$0-1$型虚拟变量:

$$u_1=\begin{cases}1, & \text{冬季},\\ 0, & \text{其它};\end{cases}\quad u_2=\begin{cases}1, & \text{春季},\\ 0, & \text{其它};\end{cases}\quad u_3=\begin{cases}1, & \text{夏季},\\ 0, & \text{其它}.\end{cases}$$

但这样做却产生了一个新的问题: $u_1+u_2+u_3\equiv 1$, 自变量之间具有多重共线性. 解决的方案是去掉一个$0-1$型虚拟变量, 令

$$\begin{aligned}(u_1,u_2)&=(1,0)\text{表示冬季},\\ (u_1,u_2)&=(0,1)\text{表示春季},\\ (u_1,u_2)&=(0,0)\text{表示夏季}.\end{aligned}$$

一般地, 若一个定性变量有k个可能的取值, 则只需引入$k-1$个$0-1$型虚拟变量.

考虑如下的多元线性回归模型:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \delta_1 u_{i1} + \delta_2 u_{i2} + e_i, \quad i = 1, \cdots, 18,$$
$$e_i \overset{i.i.d.}{\sim} N(0, \sigma^2). \qquad (6.1.1)$$

用最小二乘法可求得相应的回归方程:

```
yx=read.table(“***.txt”)
x1=yx[,1]
x2=yx[,2]
y=yx[,3]
u1=c(rep(1,6),rep(0,12))
u2=c(rep(0,6),rep(1,6),rep(0,6))
alcohol=data.frame(x1,x2,u1,u2,y)
alcohol
lm.sol=lm(y~x1+x2+u1+u2,data=alcohol)
summary(lm.sol)
deviance(lm.sol)
```

```
> lm.sol=lm(y~x1+x2+u1+u2,data=alcohol)
> summary(lm.sol)

Call:
lm(formula = y ~ x1 + x2 + u1 + u2, data = alcohol)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37937 -0.19865 -0.03627  0.14627  0.64357

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 90.311317   4.014262  22.498 8.56e-12 ***
x1          -0.644883   0.034059 -18.934 7.57e-11 ***
x2           0.023874   0.004538   5.261 0.000154 ***
u1          -3.828082   0.198332 -19.301 5.95e-11 ***
u2          -1.389680   0.215242  -6.456 2.14e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3388 on 13 degrees of freedom
Multiple R-squared:  0.9863,    Adjusted R-squared:  0.982
F-statistic: 233.4 on 4 and 13 DF,  p-value: 5.826e-12
```

回归方程为:

$$\hat{y} = 90.31 - 0.64x_1 + 0.024x_2 - 3.828u_1 - 1.390u_2.$$

各季度的回归方程是:

冬季: $\hat{y} = \hat{\beta}_0 + \hat{\delta}_1 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 86.48 - 0.64x_1 + 0.024x_2,$
春季: $\hat{y} = \hat{\beta}_0 + \hat{\delta}_2 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 88.92 - 0.64x_1 + 0.024x_2,$
夏季: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 90.31 - 0.64x_1 + 0.024x_2.$

残差平方和(通过deviance(lm.sol)求得)和自由度分别为

$$\mathsf{RSS} = 1.4923,\ n - 4 - 1 = 13.$$

为检验季节对因变量有无显著影响, 这相当于检验假设

$$H:\ \delta_1 = \delta_2 = 0. \tag{6.1.2}$$

这其实是关于未知参数向量$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \delta_1, \delta_2)'$的线性假设:

$$H:\ \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{b},$$

其中

$$\boldsymbol{A} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \boldsymbol{b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

在假设(6.1.2)为真时, 约简模型为

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \quad i = 1, \cdots, 18,$$
$$e_i \ i.i.d. \ \sim N(0, \sigma^2).$$

再求约简模型的回归方程:

```
yx=read.table( “***.txt” )
x1=yx[, 1]
x2=yx[, 2]
y=yx[, 3]
wine=data.frame(x1,x2,y)
wine
lm.reg=lm(y~x1+x2,data=wine)
summary(lm.reg)
deviance(lm.reg)
```

```
> lm.reg=lm(y~x1+x2,data=wine)
> summary(lm.reg)

Call:
lm(formula = y ~ x1 + x2, data = wine)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6217 -1.8208  0.4435  1.2843  2.2519

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 93.29157   19.55451   4.771 0.000248 ***
x1          -0.69127    0.16199  -4.267 0.000675 ***
x2           0.03090    0.02184   1.415 0.177550
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.755 on 15 degrees of freedom
Multiple R-squared:  0.5749,    Adjusted R-squared:  0.5182
F-statistic: 10.14 on 2 and 15 DF,  p-value: 0.001636
```

约简模型的回归方程为:

$$\hat{y} = 93.29 - 0.69x_1 + 0.031x_2.$$

相应的残差平方和(通过deviance(lm.reg)求得)和自由度分别为

$$\mathrm{RSS}_H = 46.1945,\ n - 2 - 1 = 15.$$

由最小二乘法基本定理, 在假设(6.1.2)为真时,

$$F = \frac{(\mathrm{RSS}_H - \mathrm{RSS})/2}{\mathrm{RSS}/(18 - 4 - 1)} \sim F(2, 13).$$

简单计算可得

$$F = \frac{(46.1945 - 1.4923)/2}{1.4923/13} = 194.71 > F_{0.05}(2, 13) = 3.81.$$

因此拒绝原假设(6.1.2), 这表明季节对y有显著影响.

我们用car package中的linearHypothesis()进行假设检验:

```
yx=read.table( “***.txt” )
x1=yx[, 1]
x2=yx[, 2]
y=yx[, 3]
u1=c(rep(1,6),rep(0,12))
u2=c(rep(0,6),rep(1,6),rep(0,6))
alcohol=data.frame(x1,x2,u1,u2,y)
alcohol
lm.sol=lm(y~x1+x2+u1+u2,data=alcohol)
A=matrix(c(0,0,0,1,0,0,0,0,0,1), nrow=2, byrow=T)
b=c(0,0)
library(car)
linearHypothesis(lm.sol,hypothesis.matrix=A,rhs=b,test="F")
```

```
> library(car)
> linearHypothesis(lm.sol,hypothesis.matrix=A,rhs=b,test="F")
Linear hypothesis test

Hypothesis:
u1 = 0
u2 = 0

Model 1: restricted model
Model 2: y ~ x1 + x2 + u1 + u2

  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     15 46.194
2     13  1.492  2    44.702 194.71 2.043e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

因变量含有定性变量的回归模型

在许多社会经济问题或临床医学研究中, 所研究的因变量往往只有两个可能的结果, 这样的因变量可用虚拟变量(取值为0或1)来表示.

例如, 在一次住房展销会上, 与房产商签订初步购房意向书的顾客中, 在随后的三个月的时间内, 只有一部分顾客确实购买了房屋. 我们可将确定购买了房屋的顾客记为1, 没有购买房屋的顾客记为0.

再如, 在一项社会安全问题的调查中, 一个人在家是否害怕陌生人, 因变量$y = 1$表示害怕, $y = 0$表示不怕.

定性因变量的回归函数的意义:

假设因变量y为只取0和1两个值的定性变量, 考虑如下的简单线性回归模型

$$y_i = \beta_0 + \beta_1 x_i + e_i. \tag{6.2.1}$$

我们通常假设$\mathsf{E}(e_i) = 0$. 在因变量只取0和1两个值时, 回归函数$\mathsf{E}(y_i|x_i) = \beta_0 + \beta_1 x_i$有着特殊的意义: 假设

$$\mathsf{P}(y_i = 1) = \pi_i, \quad \mathsf{P}(y_i = 0) = 1 - \pi_i,$$

则$\mathsf{E}(y_i|x_i) = \pi_i$. 所以

$$\mathsf{E}(y_i|x_i) = \pi_i = \beta_0 + \beta_1 x_i.$$

这表明回归函数$\mathsf{E}(y_i|x_i) = \beta_0 + \beta_1 x_i$是给定自变量水平为$x_i$时$y_i = 1$的概率.

定性因变量回归的特殊性:

(1) 离散非正态误差项. 对只取0和1两个值的定性因变量y, 若它关于自变量x_i的回归模型如(6.2.1)所示, 则其误差项e_i也只能取两个值. 即

当$y_i = 1$时, $e_i = 1 - \beta_0 - \beta_1 x_i = 1 - \pi_i$;

当$y_i = 0$时, $e_i = 0 - \beta_0 - \beta_1 x_i = -\pi_i$;

这样, 误差项为两点分布的随机变量, 于是正态误差回归模型的假定就不再适用了.

(2) 误差项仍保持零均值性质

$$\mathsf{E}(e_i) = (1-\pi_i)\pi_i - \pi_i(1-\pi_i) = 0,$$

但是e_i的方差不相等:

$$\mathsf{Var}(e_i) = \mathsf{Var}(y_i) = \pi_i(1-\pi_i) = (\beta_0+\beta_1 x_i)(1-\beta_0-\beta_1 x_i).$$

因此, 方差项为异方差, 不满足线性回归模型的基本假定. 这表明, 对因变量为定性变量的线性回归模型, 最小二乘估计的效果不会很好.

(3) 回归函数的限制. 当因变量y为只取0和1两个值的定性变量时, $\mathsf{E}(y_i|x_i)$受如下限制:

$$0 \leq \mathsf{E}(y_i|x_i) = \pi_i = \beta_0 + \beta_1 x_i \leq 1.$$

然而, 一般回归函数并不具有这种限制. 也就是说, 对定性因变量直接建立回归方程是不可取的而且得不到合理的解释.

下面, 我们介绍用另外的方法来对定性因变量建立回归方程.

Logistic回归模型:

当因变量y为一个二值变量且只取0和1两个值时, 如果我们对影响y的因素$x_1, \cdots, x_p$(这些x_i中可能既有定性变量又有定量变量)建立类似于(6.2.1)的线性回归模型, 则将遇到如下两个问题:

(1) 因变量y的取值最大为1最小为0, 而(6.2.1)右端的取值可能超出区间$[0, 1]$, 甚至可能在$(-\infty, \infty)$上取值;

(2) 因变量y本身只取0和1两个离散值, 而(6.2.1)右端的取值可在一个范围内连续变化.

对于上述的第一个问题, 我们可通过寻找因变量均值的函数, 使得该函数的取值范围在$(-\infty, \infty)$内来解决这一问题. 符合这一要求的函数有很多, 例如, 随机变量的分布函数的反函数就符合这一要求, 其中最常用的是标准正态随机变量的分布函数的反函数. 还有一个很重要的函数是

$$\mathsf{logit}(z) = \log \frac{z}{1-z},$$

其中z在区间$[0, 1]$上取值. 这一函数被称为Logit函数.

对于第二个问题, 由于π_i是定性变量y_i取1的概率, 可在$[0,1]$内连续变化. 因此, 可用下面的模型

$$\mathsf{E}(y_i|\text{自变量}) = \pi_i = \frac{\exp(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij})}{1+\exp(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij})},\ i = 1,\cdots,n \quad (6.2.2)$$

来研究$0-1$型因变量y与自变量$x_1,\cdots,x_p$之间的关系. 模型(6.2.2)被称为Logistic回归模型. Logistic回归模型也可表示为

$$\mathsf{logit}(\pi_i) = \log\frac{\pi_i}{1-\pi_i} = \boldsymbol{x}_i'\boldsymbol{\beta},\ \ i = 1,\cdots,n, \qquad (6.2.3)$$

其中$\boldsymbol{x}_i' = (1, x_{i1},\cdots,x_{ip})$, $\boldsymbol{\beta} = (\beta_0,\beta_1,\cdots,\beta_p)'$. 这个模型有时也被称为"评定模型", 它在社会学、经济学、生物统计学、数量心理学、市场营销学以及交通等领域有着广泛的应用.

$\pi_i/(1-\pi_i)$是"事件发生"与"事件没有发生"的优势比(odds ratio), 因此, Logit变换有很好的统计解释, 它是优势比的对数. 同时, 可知$\pi_i/(1-\pi_i)$是π_i的严格增函数.

我们也可以用分布函数的反函数取代Logit函数, 譬如我们可假设回归模型为

$$\text{Probit}(\pi_i) = \Phi^{-1}(\pi_i) = \boldsymbol{x}_i'\boldsymbol{\beta}, \quad i = 1, \cdots, n, \tag{6.2.4}$$

其中$\Phi^{-1}(z)$表示标准正态分布函数的反函数. 模型(6.2.4)被称为Probit回归模型(又称为多元概率比回归模型).

若我们采用双对数变换$f(\pi_i) = \log(-\log(1-\pi_i))$取代替Logit函数, 则得到如下的回归模型

$$\log(-\log(1-\pi_i)) = \boldsymbol{x}_i'\boldsymbol{\beta}, \quad i = 1, \cdots, n. \tag{6.2.5}$$

Logistic回归模型的参数估计

对于Logistic回归模型的参数估计问题, 我们分两种情况来讨论.

(1) 分组数据情形

假设某一事件A发生的概率π依赖于一些自变量$x_1, \cdots, x_p$, 我们对事件A在m个不同的自变量条件下作了n次观测, 其中对应于$\boldsymbol{x} = (x_1, \cdots, x_p)'$的一个组合$\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})'$观测了$n_i$组结果, $i = 1, \cdots, m$, $\sum_{i=1}^{m} n_i = n$. 假设在这n_i个观测中事件A发生了r_i次, 于是事件A发生的概率可用$\hat{\pi}_i = r_i/n_i$来估计. 我们把这种结构的数据称为分组数据. 用π_i的估计值$\hat{\pi}_i$代替(6.2.3)中的π_i可得关系式

$$y_i^* \triangleq \log \frac{\hat{\pi}_i}{1-\hat{\pi}_i} = \log \frac{\pi_i}{1-\pi_i} + e_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i, \ i = 1, \cdots, m \tag{6.3.1}$$

这是我们常见的线性回归模型.

因此若假设$e_1, \cdots, e_m$互不相关且$\mathsf{E}(e_i) = 0$和$\mathsf{Var}(e_i) = v_i$, 则参数$\boldsymbol{\beta}$的广义最小二乘估计为

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{Y}^*, \tag{6.3.2}$$

其中

$$\boldsymbol{Y}^* = \begin{pmatrix} y_1^* \\ y_2^* \\ \vdots \\ y_m^* \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & \cdots & x_{mp} \end{pmatrix}, \boldsymbol{V} = \begin{pmatrix} v_1 & 0 & \cdots & 0 \\ 0 & v_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_m \end{pmatrix}.$$

要考察某些x_j是否对事件A发生的概率有影响, 也即要检验x_j对应的回归系数$\beta_j = 0$这一假设是否成立. 为了用前面介绍的线性回归模型的理论和方法来探讨这一问题,我们需要假设e_i服从正态分布. 但这一假设是否成立呢?我们来讨论这一问题.

由于$\hat{\pi}_i = r_i/n_i$是样本的频率, 因此由大数定律和中心极限定理可知: 当$n_i \to \infty$时, $\hat{\pi}_i$以概率1收敛到π_i以及

$$\sqrt{n_i}(\hat{\pi}_i - \pi_i) \xrightarrow{d} N(0, \pi_i(1-\pi_i)).$$

现在来推导y_i^*的分布, 这需要用到下面的Delta方法.

引理 (Delta Method: A Generalized CLT)

Let Y_n be a sequence of random variables that satisfies

$$\sqrt{n}(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

For a given function and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{d} N(0, \sigma^2[g'(\theta)]^2).$$

由$f(z)=\log\frac{z}{1-z}$可得

$$f'(z)=\frac{1}{z(1-z)},\quad f'(z)|_{z=\pi_i}=\frac{1}{\pi_i(1-\pi_i)}.$$

于是, 由Delta方法可知: 当$n_i\to\infty$时, 有

$$\sqrt{n_i}\left(\log\frac{\hat{\pi}_i}{1-\hat{\pi}_i}-\log\frac{\pi_i}{1-\pi_i}\right)\xrightarrow{d}N\left(0,\frac{1}{\pi_i(1-\pi_i)}\right).$$

这表明: 当$\min\{n_1,\cdots,n_m\}$充分大时, 我们可以认为y_i^*服从正态分布$N(\boldsymbol{x}_i'\boldsymbol{\beta},v_i)$,其中$v_i=\frac{1}{n_i\pi_i(1-\pi_i)}$. 由于$\pi_i$是未知的, 因此在求$\hat{\boldsymbol{\beta}}$时我们用$\hat{v}_i=\frac{1}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$去代替$\boldsymbol{V}$中的$v_i$.

例6.2.1 在一次住房展销会上, 与房地产商签订初步购房意向书的共有$n = 325$名顾客, 在随后的3个月时间内, 只有一部分顾客确实购买了房屋. 购买了房屋的顾客记为1, 没有购买房屋的顾客记为0. 以顾客的家庭年收入作为自变量x(单位: 万元), 对下表的数据, 试分析家庭年收入的不同对最终购买住房的影响.

表6.2.1: 签订购房意向和最终买房的客户数据

序号	x	签订意向书人数n_i	实际购房人数m_i
1	1.5	25	8
2	2.5	32	13
3	3.5	58	26
4	4.5	52	22
5	5.5	43	20
6	6.5	39	22
7	7.5	28	16
8	8.5	21	12
9	9.5	15	10

```
yx=read.table( “***.txt” )
x=yx[, 1]
n=yx[, 2]
m=yx[, 3]
k=n-m
house=data.frame(x,m,k)
house
glm.sol=glm(cbind(m,k)~x,family=binomial(link=”logit”),data=house)
summary(glm.sol)
```

```
Call:
glm(formula = cbind(m, k) ~ x, family = binomial(link = "logit"),
    data = house)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.47375  -0.30287   0.04138   0.27065   0.45253

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.8518     0.2931  -2.906  0.00366 **
x             0.1498     0.0534   2.805  0.00502 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9.1386  on 8  degrees of freedom
Residual deviance: 1.0467  on 7  degrees of freedom
AIC: 40.092

Number of Fisher Scoring iterations: 3
```

得到的Logistic回归方程为:

$$\hat{y}^* = \log \frac{\hat{\pi}}{1-\hat{\pi}} = -0.8518 + 0.1498x$$

或写成

$$\hat{\pi} = \frac{\exp(-0.8518 + 0.1498x)}{1 + \exp(-0.8518 + 0.1498x)}.$$

因此可知: x越大, 即家庭年收入越高, $\hat{\pi}$就越大, 即签订意书向后真正买房的可能性就越大. 对于一个家庭年收入为9万元的客户, 签订意向书后真正买房的概率估计为

$$\hat{\pi}_0 = \frac{\exp(-0.8518 + 0.1498 \times 9)}{1 + \exp(-0.8518 + 0.1498 \times 9)} = 0.622.$$

```
> new=data.frame(x=c(6,7,8,9))
> glm.pred=predict(glm.sol,new)
> pred=exp(glm.pred)/(1+exp(glm.pred))
> pred
        1         2         3         4
0.5117861 0.5490852 0.5858407 0.6216645
>
```

(2) 未分组数据情形

假设y为$0-1$型随机变量, 即$y_i \sim B(1,\pi_i)$, 而$x_1,\cdots,x_p$是对y有影响的p个确定性变量. 在$(x_1,\cdots,x_p)$的n个不同点$(x_{i1},\cdots,x_{ip})$, $i=1,\cdots,n$, 对y进行了n次独立观测得其观测值$y_1,\cdots,y_n$. 显然, $y_1,\cdots,y_n$是相互独立的Bernoulli随机变量, 概率分布为

$$\pi_i^{y_i}(1-\pi_i)^{1-y_i},\quad y_i=0,1.$$

于是, $y_1,\cdots,y_n$的似然函数为

$$L(\pi_1,\cdots,\pi_n)=\prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$

其对数似然函数为

$$l(\pi_1,\cdots,\pi_n)=\sum_{i=1}^{n}[y_i\log\pi_i+(1-y_i)\log(1-\pi_i)].$$

将(6.2.3)代入上式得

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \boldsymbol{x}_i^{'} \boldsymbol{\beta} - \log(1 + \exp(\boldsymbol{x}_i^{'} \boldsymbol{\beta}))]. \tag{6.3.3}$$

求$\boldsymbol{\beta}$的极大似然估计就是寻找$\boldsymbol{\beta}$使得$l(\boldsymbol{\beta})$达到最大. 为此, 计算(6.3.3)关于$\boldsymbol{\beta}$的一阶导数:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(y_i - \frac{e^{\boldsymbol{x}_i^{'} \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^{'} \boldsymbol{\beta}}} \right) \boldsymbol{x}_i = \boldsymbol{X}^{'} \boldsymbol{\varepsilon},$$

其中

$$\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^{'}, \boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^{'}, \varepsilon_i = y_i - \frac{e^{\boldsymbol{x}_i^{'} \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^{'} \boldsymbol{\beta}}}.$$

令

$$\boldsymbol{X}'\boldsymbol{\varepsilon} = \sum_{i=1}^{n}\left(y_i - \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\right)\boldsymbol{x}_i = \mathbf{0},$$

求解$\hat{\boldsymbol{\beta}}$.

但上述方程是关于参数$\boldsymbol{\beta}$的一个较复杂的非线性函数, 要获得$\boldsymbol{\beta}$的极大似然估计$\hat{\boldsymbol{\beta}}$是不易的. 一般地, 可采用迭代算法, 如Newton-Raphson迭代算法, 求数值解.

例6.3.1 下表6.3.1是对45名驾驶员的调查结果, 其中四个变量的含义为: x_1: 表示视力状况, 它是一个定性变量, 1表示好, 0表示有问题; x_2: 年龄; x_3: 驾车教育, 它也是一个定性变量, 1表示参加过驾车教育, 0表示没有; y: 一个定性变量(去年是否出过事故, 1表示出过事故, 0表示没有). 试考察x_1, x_2, x_3与发生事故的关系.

表6.3.1: 对45名驾驶员的调查结果

x_1	x_2	x_3	y	x_1	x_2	x_3	y	x_1	x_2	x_3	y
1	17	1	1	1	68	1	0	0	17	0	0
1	44	0	0	1	18	1	0	0	45	0	1
1	48	1	0	1	68	0	0	0	44	0	1
1	55	0	0	1	48	1	1	0	67	0	0
1	75	1	1	1	17	0	0	0	55	0	1
0	35	0	1	1	70	1	1	1	61	1	0
0	42	1	1	1	72	1	0	1	19	1	0
0	57	0	0	1	35	0	1	1	69	0	0
0	28	0	1	1	19	1	0	1	23	1	1
0	20	0	1	1	62	1	0	1	19	0	0
0	38	1	0	0	39	1	1	1	72	1	1
0	45	0	1	0	40	1	1	1	74	1	0
0	47	1	1	0	55	0	0	1	31	0	1
0	52	0	0	0	68	0	1	1	16	1	0
0	55	0	1	0	25	1	0	1	61	1	0

```
yx=read.table( “***.txt” )
x1=yx[, 1]
x2=yx[, 2]
x3=yx[, 3]
y=yx[, 4]
accident=data.frame(x1,x2,x3,y)
accident
glm.sol=glm(y∼x1+x2+x3,family=binomial(link=”logit”),data=accident)
summary(glm.sol)
```

```
glm(formula = y ~ x1 + x2 + x3, family = binomial(link = "logit"),
    data = accident)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5636  -0.9131  -0.7892   0.9637   1.6000

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.597610   0.894831   0.668   0.5042
x1          -1.496084   0.704861  -2.123   0.0338 *
x2          -0.001595   0.016758  -0.095   0.9242
x3           0.315865   0.701093   0.451   0.6523
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 62.183  on 44  degrees of freedom
Residual deviance: 57.026  on 41  degrees of freedom
AIC: 65.026

Number of Fisher Scoring iterations: 4
```

得到初步的Logistic回归方程:

$$\text{logit}(\hat{\pi}) = 0.5976 - 1.4961x_1 - 0.0016x_2 + 0.3159x_3,$$

或等价地写成

$$\hat{\pi} = \frac{\exp(0.5976 - 1.4961x_1 - 0.0016x_2 + 0.3159x_3)}{1 + \exp(0.5976 - 1.4961x_1 - 0.0016x_2 + 0.3159x_3)}.$$

但由于参数β_2和β_3没有通过显著性检验, 类似于线性模型, 我们可以用step()做变量筛选.

```
> glm.reg=glm(y~1,family=binomial(link="logit"),data=accident)
> step.model=step(glm.reg,direction="both",scope=(~x1+x2+x3))
Start:  AIC=64.18
y ~ 1

       Df Deviance    AIC
+ x1    1   57.241 61.241
<none>      62.183 64.183
+ x3    1   61.991 65.991
+ x2    1   62.122 66.122

Step:  AIC=61.24
y ~ x1

       Df Deviance    AIC
<none>      57.241 61.241
+ x3    1   57.035 63.035
+ x2    1   57.232 63.232
- x1    1   62.183 64.183
```

```
> summary(step.model)

Call:
glm(formula = y ~ x1, family = binomial(link = "logit"), data = accident)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4490  -0.8782  -0.8782   0.9282   1.5096

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.6190     0.4688   1.320   0.1867
x1           -1.3728     0.6353  -2.161   0.0307 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 62.183  on 44  degrees of freedom
Residual deviance: 57.241  on 43  degrees of freedom
AIC: 61.241

Number of Fisher Scoring iterations: 4
```

得到新的Logistic回归方程:

$$\text{logit}(\hat{\pi}) = 0.6190 - 1.3728x_1,$$

或等价地写成

$$\hat{\pi} = \frac{\exp(0.6190 - 1.3728x_1)}{1 + \exp(0.6190 - 1.3728x_1)}.$$

最后, 我们做个预测分析:

```
> new=data.frame(x1=c(1,0))
> log.pre=predict(step.model,new)
> pi=exp(log.pre)/(1+exp(log.pre))
> pi
   1    2
0.32 0.65
>
```

这说明视力有问题的司机发生交通事故的概率大约是视力正常的司机的两倍(0.65/0.32).