第六章 含定性变量的回归模型

在许多实际问题中, 人们常常会碰到定性变量. 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, 都是定性变量. 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, 以及因变量含有定性变量的统计建模.

6.1 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之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, 结果见下表.

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

表6.1.1

处理这种问题的一种方法是分季度建立回归方程, 然后看看不同季节的回归方程是否有显著差异. R代码及分析结果如下:

```
> yx=read.table("***.txt")
> x1=yx[,1]
> x2=yx[,2]
> y=yx[,3]
> wine=data.frame(x1,x2,y)
> 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
```

```
14
15 Coefficients:
16              Estimate Std. Error t value Pr(>|t|)
17 (Intercept) 82.659524   2.893225  28.570 9.42e-05 ***
18 x1          -0.604643   0.024869 -24.313 0.000153 ***
19 x2           0.016667   0.004159   4.008 0.027870 *
20
21 > lm.reg2=lm(y~x1+x2,data=wine,subset=7:12)
22 > summary(lm.reg2)
23 Call:
24 lm(formula = y ~ x1 + x2, data = wine, subset = 7:12)
25
26 Residuals:
27        7        8        9       10       11       12
28 -0.27302  0.26290  0.03667 -0.10300  0.21958 -0.14313
29
30 Coefficients:
31              Estimate Std. Error t value Pr(>|t|)
32 (Intercept) 101.674233  10.846909   9.374  0.00257 **
33 x1           -0.745615   0.084637  -8.810  0.00308 **
34 x2            0.028848   0.005676   5.082  0.01472 *
35
36 > lm.reg3=lm(y~x1+x2,data=wine,subset=13:18)
37 > summary(lm.reg3)
38 Call:
39 lm(formula = y ~ x1 + x2, data = wine, subset = 13:18)
40
41 Residuals:
42      13      14      15      16      17      18
43 -0.4056 -0.4648 -0.1066  0.5023  0.3648  0.1099
44
45 Coefficients:
46             Estimate Std. Error t value Pr(>|t|)
47 (Intercept) 98.14611   13.18155   7.446  0.00501 **
48 x1          -0.72897    0.12578  -5.796  0.01022 *
49 x2           0.03915    0.02064   1.897  0.15414
```

分季度得到的回归方程分别是:

$$\begin{aligned} &\text{冬季}: \quad \hat{y} = 82.6695 - 0.6046x_1 + 0.0167x_2, \\ &\text{春季}: \quad \hat{y} = 101.6742 - 0.7456x_1 + 0.0288x_2, \\ &\text{夏季}: \quad \hat{y} = 98.1461 - 0.7289x_1 + 0.0392x_2. \end{aligned}$$

在这三个回归方程中, x_1的系数(绝对)差异不大, x_2的系数(绝对)差异也不大, 但常数项的(绝对)差异较大. 既然x_1的系数和x_2的系数在不同的季度差异不大, 但在每个季度都只用了6个样本用于点估计, 因此为了提高估计精度, 考虑将这批数据进行统一处理. 因为常数项在不同的季度差异较大, 所以需要考虑季节因素, 但需注意季节是定性变量. 为此, 我们引入“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, 而将另一个取值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, 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:

$$u_{i1}+u_{i2}+u_{i3}\equiv 1,\quad i=1,\cdots,n,$$

即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.

现在, 考虑如下的多元线性回归模型:

$$\begin{cases} y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\delta_1u_{i1}+\delta_2u_{i2}+e_i,\quad i=1,\cdots,18,\\ e_i\overset{i.i.d.}{\sim}N(0,\sigma^2).\end{cases}\tag{6.1.1}$$

用最小二乘法求相应的回归方程, R代码及分析结果如下:

```
> 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)
> 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 ***

> deviance(lm.sol)  #求残差平方和
[1] 1.492276
```

得到回归方程

$$\hat{y} = 90.311 - 0.645x_1 + 0.024x_2 - 3.828u_1 - 1.390u_2.$$

分季度的回归方程为

$$\begin{aligned}
&\text{冬季：}\quad \hat{y} = \hat{\beta}_0 + \hat{\delta}_1 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 86.483 - 0.645x_1 + 0.024x_2,\\
&\text{春季：}\quad \hat{y} = \hat{\beta}_0 + \hat{\delta}_2 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 88.921 - 0.645x_1 + 0.024x_2,\\
&\text{夏季：}\quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 90.311 - 0.645x_1 + 0.024x_2.
\end{aligned}$$

残差平方和和自由度分别为

$$\mathrm{RSS} = 1.4923, \quad n - p - 1 = 18 - 4 - 1 = 13.$$

接下来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$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \delta_1, \delta_2)'$的如下线性假设:

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

其中

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

在假设(6.1.2)为真时, 约简模型为

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

来求约简模型的残差平方和:

```
> deviance(lm(y~x1+x2,data=alcohol))
[1] 46.19448
```

约简模型的残差平方和和自由度分别为

$$\mathrm{RSS}_H = 46.1945, \ n - p - 1 = 18 - 2 - 1 = 15.$$

由最小二乘法基本定理, 在假设(6.1.2)为真时,

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

经简单计算可得

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

因此拒绝原假设(6.1.2), 这表明季节对y有显著影响.

下面用程序包car中的linearHypothesis命令进行假设检验, R代码及分析结果如下:

```
> 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")
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
```

可以看出分析结果与前面一致.

6.2 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. 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.

先来了解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

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

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

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

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.

下面介绍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的回归模型如(6.2.1)所示, 则模型的误差项e_i也只能取两个值:

$$\text{当}y_i = 1\text{时},\ e_i = 1 - \beta_0 - \beta_1 x_i = 1 - \pi_i,$$
$$\text{当}y_i = 0\text{时},\ e_i = 0 - \beta_0 - \beta_1 x_i = -\pi_i.$$

即, 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) 零均值异方差性. 误差项仍保持零均值:

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

但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, 模型误差为异方差, 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) 回归函数取值范围的限制. 当因变量y为只取0和1两个值的定性变量时, $\mathsf{E}(y_i|x_i)$的大小受如下限制:

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

然而, 一般的回归函数并不受这种限制. 也就是说, 对定性因变量直接建立回归模型是不可取的而且得不到合理的解释.

下面介绍Logistic(逻辑斯蒂)回归模型, 它可对定性因变量建立回归模型. 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

$$y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + e$$

这样的线性回归模型, 则将遇到如下的两个问题: (1) 因变量y本身只取0和1两个离散值, 而$\beta_0 + \sum_{j=1}^{p} \beta_j x_j$的取值可在某个范围内连续变化; (2) 因变量$y$的取值最大为1最小为0, 而$\beta_0 + \sum_{j=1}^{p} \beta_j x_j$的取值可能超出区间$[0, 1]$, 甚至可能在$(-\infty, \infty)$上取值. 如何解决这两个问题呢? 对于上述的第一个问题, 可以考虑因变量的均值, 它的取值是连续的. 对于上述的第二个问题, 可以考虑因变量均值的某函数, 使得该函数的取值范围为$(-\infty, \infty)$. 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Logit函数:

$$\text{Logit}(z) = \ln \frac{z}{1 - z}, \quad z \in [0, 1].$$

因此, 我们考虑如下的模型:

$$\text{Logit}(\pi_i) = \ln \frac{\pi_i}{1 - \pi_i} = \boldsymbol{x}_i' \boldsymbol{\beta}, \quad i = 1, \cdots, n, \tag{6.2.2}$$

其中$\pi_i = \mathsf{E}(y_i|$自变量$)$,

$$\boldsymbol{x}_i' = (1, x_{i1}, \cdots, x_{ip}), \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'.$$

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. 模型(6.2.2)可以等价地写成

$$\pi_i = \frac{\exp(\beta_0 + \sum_{j=1}^p \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^p \beta_j x_{ij})}, \quad i = 1, \cdots, n. \tag{6.2.3}$$

模型(6.2.3)被称为Logistic回归模型.

$\pi_i/(1-\pi_i)$是“事件发生”与“事件没有发生”的概率比, 也称为赔率(胜率对输率的比率, odds)或优劣率. 因此, Logit变换有很好的统计解释, 它是赔率的对数. 同时可知$\pi_i/(1-\pi_i)$是π_i的严格增函数.

也可以用分布函数的反函数取代Logit函数, 譬如可假设回归模型为

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

其中$\Phi^{-1}(z)$表示标准正态随机变量的分布函数的反函数. 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). 若用双对数变换$f(\pi_i) = \ln(-\ln(1-\pi_i))$取代Logit函数, 则得到如下的回归模型

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

6.3 Logistic回归模型的参数估计

这一节考虑Logistic回归模型的参数估计问题. 下面分两种情况来讨论.

(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$x_1, \cdots, x_p$, 且对事件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, $i = 1, \cdots, m$; n_i满足$\sum_{i=1}^m n_i = n$. 假设在这n_i个观测中事件A发生了r_i次, 于是在自变量水平$\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})'$下, 事件$A$发生的概率$\pi_i$可用频率$\hat{\pi}_i = r_i/n_i$来估计. 我们把这种结构的数据称为分组数据. 用$\pi_i$的估计值$\hat{\pi}_i$代替(6.2.2)中的$\pi_i$, 并引入估计误差$e_i$, 可得下面的模型:

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

这是我们熟悉的线性回归模型. 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

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

下面考察某些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, 需要假设e_i服从(或近似服从)正态分布. 但这一假设是否成立呢? 下面来讨论这一问题.

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且

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

现在来推导y_i^*的近似分布, 这需要用到下面的Delta方法.

引理6.3.1(Delta方法)　假设$\{Y_i, i \geq 1\}$是一列随机变量且下面的中心极限定理成立:

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

若$g'(\theta)$存在且不等于0, 则

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

记$f(z) = \ln \frac{z}{1-z}$, 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, 当$n_i \to \infty$时, 有

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

这表明: 当$\min\{n_1, \cdots, n_m\}$充分大时, 可以认为$y_i^* = \ln\frac{\hat{\pi}_i}{1-\hat{\pi}_i}$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6.3.1　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), 对表6.3.1的数据, 分析家庭年收入对最终购买住房的影响.

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

表6.3.1　签订购房意向和最终买房的客户数据

R代码及分析结果如下:

```
> yx=read.table("***.txt")
> x=yx[,1]
> n=yx[,2]
> m=yx[,3]
> k=n-m
> house=data.frame(x,m,k)
> 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
```

由上述分析结果, 得到Logistic回归方程:

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

或写成

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

由回归方程可知: 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的估计值为

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

类似地, 可以预测一个家庭年收入为6(或7, 或8)万元的客户, 签订意向书后真正买房的概率.

```
> 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个自变量. 在$(x_1,\cdots,x_p)$的n个不同水平$\{(x_{i1},\cdots,x_{ip}), i=1,\cdots,n\}$下对$y$进行了$n$次独立观测得到观测值$\{y_1,\cdots,y_n\}$. 显然, $y_1,\cdots,y_n$是相互独立的Bernoulli随机变量, y_i的概率分布为

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

在未分组数据情形下, 我们无法应用最小二乘法进行参数估计(因为模型(6.2.2)中没有误差项), 所以改用极大似然方法. $y_1,\cdots,y_n$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

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

将(6.2.2)代入上式得

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

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:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{n}\Big(y_i-\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\Big)\boldsymbol{x}_i.$$

令

$$\frac{\partial l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}}=\sum_{i=1}^{n}\Big(y_i-\frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}\Big)\boldsymbol{x}_i=\mathbf{0},$$

求解$\hat{\boldsymbol{\beta}}$. 但上述方程组的左边是关于参数$\boldsymbol{\beta}$的一个较复杂的非线性函数, 要获得$\boldsymbol{\beta}$的极大似然估计$\hat{\boldsymbol{\beta}}$的解析表达式是不容易的. 一般地, 可采用迭代算法, 如Newton-Raphson迭代算法, 求数值解.

例6.3.2　表6.3.2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表示年龄; x_3表示参加驾车教育情况, 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是一个定性变量, 表示去年是否出过事故(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.

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

表6.3.2　对45名驾驶员的调查结果

R代码及分析结果如下:

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

$$\ln \frac{\hat{\pi}}{1-\hat{\pi}} = 0.5976 - 1.4961x_1 - 0.0016x_2 + 0.3159x_3,$$

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, 可以用step命令做变量筛选. 采用逐步回归法, R代码及分析结果如下:

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

现在得到新的Logistic回归方程:

$$\ln \frac{\hat{\pi}}{1-\hat{\pi}} = 0.6190 - 1.3728x_1,$$

或等价地写成

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

最后, 用这个新的回归方程做个预测:

```
> 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
```

这个预测结果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).

作业

1. 对自变量含有定性变量的问题, 为什么不对定性变量的不同属性分别建立回归模型, 而是采用引入虚拟变量的方法建立回归模型?

2. 某经济学家想调查文化程度对家庭储蓄的影响, 在一个中等收入的样本中, 随机调查了13户高学历家庭和14户中低收入家庭. 因变量y表示上一年家庭储蓄增加额, 自变量x_1为上一年度家庭总收入, 自变量x_2表示家庭学历, 其中$x_2 = 1$表示高学历家庭, 而$x_2 = 0$表示低学历家庭, 其调查数据如下表所示. 请分析学历对家庭储蓄增加额有无显著影响.

序号	y(元)	x_1(万元)	x_2	序号	y(元)	x_1(万元)	x_2
1	235	2.3	0	15	3265	3.8	1
2	346	3.2	1	16	3265	4.6	1
3	365	2.8	0	17	3567	4.2	1
4	468	3.5	1	18	3658	3.7	1
5	658	2.6	0	19	4588	3.5	0
6	867	3.2	1	20	6436	4.8	1
7	1085	2.6	0	21	9047	5.0	1
8	1236	3.4	1	22	7985	4.2	0
9	1238	2.2	0	23	8950	3.9	0
10	1345	2.8	1	24	9685	4.8	0
11	2365	2.3	0	25	9866	4.6	0
12	2365	3.7	1	26	10235	4.8	0
13	3256	4.0	1	27	10140	4.2	0
14	3256	2.9	0				

题2的数据

3. 调查某个国家的国民是否愿意接种某种疾病的疫苗. 用y表示接种疫苗的意愿($y=1$表示被调查者愿意接种疫苗, $y=0$表示被调查者不愿接种疫苗). 考察接种疫苗的意愿与年龄(用x_1表示)以及疫苗价格(用x_2表示)的关系. 若直接对y和x_1, x_2建立如下的回归模型:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \quad i = 1, \cdots, n,$$

简要回答:
(1) 上述的回归模型为何不合理?
(2) 对y和x_1, x_2建立一个合理的回归模型.

4. 下表记录了煤矿工人患上严重的硅肺病症状的比例及其在井下的作业时间(单位: 年). 因变量y表示有严重症状的矿工的比例, x表示井下的作业时间. 请建立Logistic回归模型.

序号	井下作业年数x	严重症状的人数	矿工总人数	严重症状的比例y
1	5.8	0	98	0
2	15.0	1	54	0.0185
3	21.5	3	43	0.0698
4	27.5	8	48	0.1667
5	33.5	9	51	0.1765
6	39.5	8	38	0.2105
7	46.0	10	28	0.3571
8	51.5	5	11	0.4545

题4的数据

5. 下表展示了不同目标速度下25枚地对空防空导弹的试发射结果, 每次的测试结果为击中($y=1$)或打偏($y=0$).
(1) 对y进行Logistic回归模型拟合;
(2) 解释模型中的参数β_1的含义(β_1指x的系数).

序号	目标速度x(单位: 节)	y	序号	目标速度x(单位: 节)	y
1	400	0	14	330	1
2	220	1	15	280	1
3	490	0	16	210	1
4	210	1	17	300	1
5	500	0	18	470	1
6	270	0	19	230	0
7	200	1	20	430	0
8	470	0	21	460	0
9	480	0	22	220	1
10	310	1	23	250	1
11	240	1	24	200	1
12	490	0	25	390	0
13	420	0			

题5的数据

6. 在一次关于公共交通的社会调查中, 一个调查项目是“是乘坐公共汽车上下班还是骑自行车上下班”. 因变量$y=1$表示主要乘公共汽车上下班, $y=0$表示主要骑自行车上下班. 自变量x_1是年龄, x_2是月收入, x_3是性别(1表示男性, 0表示女性). 调查对象为工薪族群体, 数据见下表. 请建立Logistic回归模型.

序号	x_3	x_1	x_2(元)	y	序号	x_3	x_1	x_2(元)	y
1	0	18	4250	0	15	1	20	5000	0
2	0	21	6000	0	16	1	25	6000	0
3	0	23	4750	1	17	1	27	6500	0
4	0	23	4750	1	18	1	28	7500	0
5	0	28	6000	1	19	1	30	4750	1
6	0	31	4250	0	20	1	32	5000	0
7	0	36	7500	1	21	1	33	9000	0
8	0	42	5000	1	22	1	33	5000	0
9	0	46	4750	1	23	1	38	6000	0
10	0	48	6000	0	24	1	41	7500	0
11	0	55	9000	1	25	1	45	9000	1
12	0	56	10500	1	26	1	48	5000	0
13	0	58	9000	1	27	1	52	7500	1
14	1	18	4250	0	28	1	56	9000	1

题6的数据

7. 若因变量y的取值0和1分别代表两个属性, 问: Logistic回归模型是否可应用于分类问题(即把$y_1, \cdots, y_n$分别归入“0”类或“1”类)? 若可以, 请制定合理的分类规则.