Chapter 9 - Loglinear Model & Generalized Linear Model

张笑竹 / 201618070114

2018年12月22日

第九章实验采用课本例9-1，例9-2和例9-3中给出的数据，分别建立对数线性模型和广义线性模型进行分析。例9-1是产品满意度和人群收入的列联表数据；例9-2是是否购买房屋与收入的关系研究；例9-3是是否乘坐公共交通的社会调查。
将本次的实验任务拆分如下：
1）调用loglm()函数和glm()函数，对例9-1建立对数线性模型，并进行分析；
2）通过logit变换，利用线性模型函数lm()，对例9-2未分组数据建立Logistic回归模型，并进行分析；
3）调用glm()函数，直接对例9-3分组数据建立Logistic回归模型，并进行分析。

1 例9-1的分析

1.1 预分析

首先，在R中读入例9-1的数据集。

```
X1=read.csv("eg9-1.csv")
X1
```

```
##    频数 收入情况 满意情况
## 1   53        1        1
## 2  434        2        1
## 3  111        3        1
## 4   38        1        2
## 5  108        2        2
## 6   48        3        2
```

该数据集是以频数为因变量，收入情况（定性数据）和满意情况（定性数据）的各个水平为自变量构成的。为了方便后续的分析，首先需要生成与该数据集对应的交叉列联表、频率列联表。将交叉列联表中的观测频数记为O_{ij}，将频率列联表中的频率数据记为p_{ij}.

```
#生成列联表
X<-matrix(c(53,434,111,38,108,48),3,2,
          dimnames=list(A=c('高','中','低'),B=c('满意','不满意')))
t<-as.table(X)
t
```

```
##     B
## A    满意 不满意
##   高   53     38
##   中  434    108
##   低  111     48
```

```r
#频率列联表
prop.table(t)
```

```
##      B
## A           满意       不满意
##    高 0.06691919 0.04797980
##    中 0.54797980 0.13636364
##    低 0.14015152 0.06060606
```

此外，为了方便计算统计量，我们还希望能够得到期望频数的列联表，并将期望频数记为E_{ij}.

```r
#计算期望频数
E=matrix(c(1:6),3,dimnames=list(A=c('高','中','低'),B=c('满意','不满意')))
for(i in 1:3){
  for(j in 1:2){
    E[i,j]=rowSums(t)[i]*colSums(t)[j]/sum(t)
  }
}
E
```

```
##      B
## A         满意      不满意
##    高  68.7096  22.29040
##    中 409.2374 132.76263
##    低 120.0530  38.94697
```

基于此，就可以计算出Pearson卡方统计量和似然比卡方统计量：

$$\chi^2_{psn}=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

$$\chi^2_{lhr}=2\sum_{i=1}^{m}\sum_{j=1}^{n}O_{ij}\ln(\frac{O_{ij}}{E_{ij}})$$

它们都服从自由度为$(m-1)(n-1)$的卡方分布。

```r
#似然比统计量
lhr=2*sum(t*log(t/E))
p.value.lhr=pchisq(lhr,2,lower.tail = F)
#perason统计量
psn=sum((t-E)^2/E)
p.value.psn=pchisq(psn,2,lower.tail = F)
STAT<-list("lhr"=lhr,"p.value.lhr"=p.value.lhr,"psn"=psn,"p.value.psn"=p.value.psn)
STAT
```

```
## $lhr
## [1] 22.08692
##
## $p.value.lhr
## [1] 1.59914e-05
##
## $psn
## [1] 23.5675
##
## $p.value.psn
## [1] 7.627506e-06
```

经过计算和检验，拒绝了收入情况和满意情况相互独立的原假设，认为它们是相互影响的，存在交互作用。

1.2 非饱和模型的分析

尽管在上一节中，已经确定了收入情况和满意情况存在相互交互作用，但是首先建立非饱和模型进行分析是有必要的。

非饱和分析模型为：

$$\ln n_{ij} = \mu + \alpha_i + \beta_j$$

调用loglm()函数，指令中“+”表示不存在交互作用，进行分析。

```
library(MASS)
X.m1<-loglm(~A+B,data=t)
X.m1$param
```

```
## $`(Intercept)`
## [1] 4.44784
##
## $A
##          高          中          低
## -0.7808171  1.0035894 -0.2227724
##
## $B
##        满意       不满意
##  0.5628663 -0.5628663
```

```
X.m1$lrt
```

```
## [1] 22.08692
```

```
X.m1$pearson
```

```
## [1] 23.5675
```

模型输出了各参数的估计结果、Pearson卡方统计量和似然比卡方统计量。两个统计量与我们先前计算得到的结果一致。

对于参数，参数值为正，表示正效应，反之为负效应，而零为无效应。基于此，说明接受调查的顾客大部分都

为中等收入，高收入人群和低收入人群较少，而满意的顾客又占主要部分。
对于非饱和模型，一个更加重要的结论，是可以计算出1维主效应的似然比卡方统计量，这需要调用glm()函数计算得到。

```
y=X1$频数
x1=X1$收入情况
x2=X1$满意情况
log.glm<-glm(y~x1*x2,family=poisson(link=log),data=X1)
log.glm$null.deviance
```

```
## [1] 662.8432
```

得到的统计量可以用来检验模型中的主效应是否显著。
事实上，似然比卡方统计量可以进行分解，即各项效应都有对应的似然比卡方值，并且它们的似然比卡方值之和等于整个模型的似然比卡方值。
因此，我们就可以仿照SPSS输出的出K-way and Higher-Order Effects检验表，分析模型各维主效应和交互作用的显著性。

```
#---仿照SPSS输出K-way and Higher-Order Effects---#
LHR<-c(log.glm$null.deviance+lhr,lhr,log.glm$null.deviance,lhr)
df<-c(5,2,3,2)
P.VALUE=c()
for(i in 1:4){
  p=pchisq(LHR[i],df[i],lower.tail = F)
  P.VALUE<-c(P.VALUE,p)
}
K.way<-data.frame(df,LHR,P.VALUE)
rownames(K.way)<-c("K-way & Higher Order Effects 1","K-way & Higher Order Effects 2",
                  "K-Way Effects 1","K-way Effects 2")
K.way
```

```
##                                df       LHR       P.VALUE
## K-way & Higher Order Effects 1  5 684.93017 8.902114e-146
## K-way & Higher Order Effects 2  2  22.08692  1.599140e-05
## K-Way Effects 1                 3 662.84325 2.391755e-143
## K-way Effects 2                 2  22.08692  1.599140e-05
```

观察p-value的值，可以看出模型一维主效应、二维交互作用都十分显著。并且“K-way & Higher Order Effects 1”的取值就等于“K-Way Effects 1”的取值加上“K-Way Effects 2”的取值。

1.3 饱和模型的分析

由于在上述分析中，二维交互作用显著，因此可以建立饱和模型进行分析。

$$\ln n_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

其中γ_{ij}表示交互作用。
调用loglm()函数，指令中“*”表示存在交互作用，进行分析。

```
X.m2<-loglm(~A*B,data=t)
X.m2
```

```
## Call:
## loglm(formula = ~A * B, data = t)
##
## Statistics:
##                  X^2 df P(> X^2)
## Likelihood Ratio   0  0        1
## Pearson            0  0        1
```

```
X.m2$param
```

```
## $`(Intercept)`
## [1] 4.490631
##
## $A
##         高          中          低
## -0.6866918  0.8869570 -0.2002652
##
## $B
##       满意       不满意
##  0.4269914 -0.4269914
##
## $A.B
##     B
## A            满意        不满意
##   高 -0.26063850  0.26063850
##   中  0.26846528 -0.26846528
##   低 -0.00782678  0.00782678
```

此时，Pearson卡方统计量和似然比卡方统计量都为0，这是因为，“饱和”表示资料的数目只够确定一些未知参数，无法确定随机误差的大小，呈“饱满”的状态，模型的拟合优度是无法检验的。
观察参数的估计结果，可以得出结论：
(1)$\beta_{满意} > 0$，说明接受调查的多数顾客对其产品是满意的。 (2)$\alpha_{高收入} < \alpha_{低收入} < \alpha_{中收入}$，说明参与调查的高收入顾客最少，低收入顾客其次，中收入顾客最多。
(3)研究交互作用的影响，发现$\gamma_{高收入满意} < 0, \gamma_{低收入满意} < 0$，只有$\gamma_{中收入满意} > 0$. 该企业产品主要的消费阶层是中等收入者，同时中等收入者对其产品的满意度也最高。

2 例9-2的分析

2.1 预分析

首先，在R中读入例9-2的数据集。

```
Y=read.csv("eg9-2.csv")
Y
```

```
##     x     p   WLS
## 1 1.5 -0.75  5.44
## 2 2.5 -0.38  7.72
## 3 3.5 -0.21 14.35
## 4 4.5 -0.31 12.69
## 5 5.5 -0.14 10.70
## 6 6.5  0.26  9.59
## 7 7.5  0.29  6.86
## 8 8.5  0.29  5.14
## 9 9.5  0.69  3.33
```

在该数据集中，x_i表示年家庭收入，而p_i则是购房比例的logit变换：

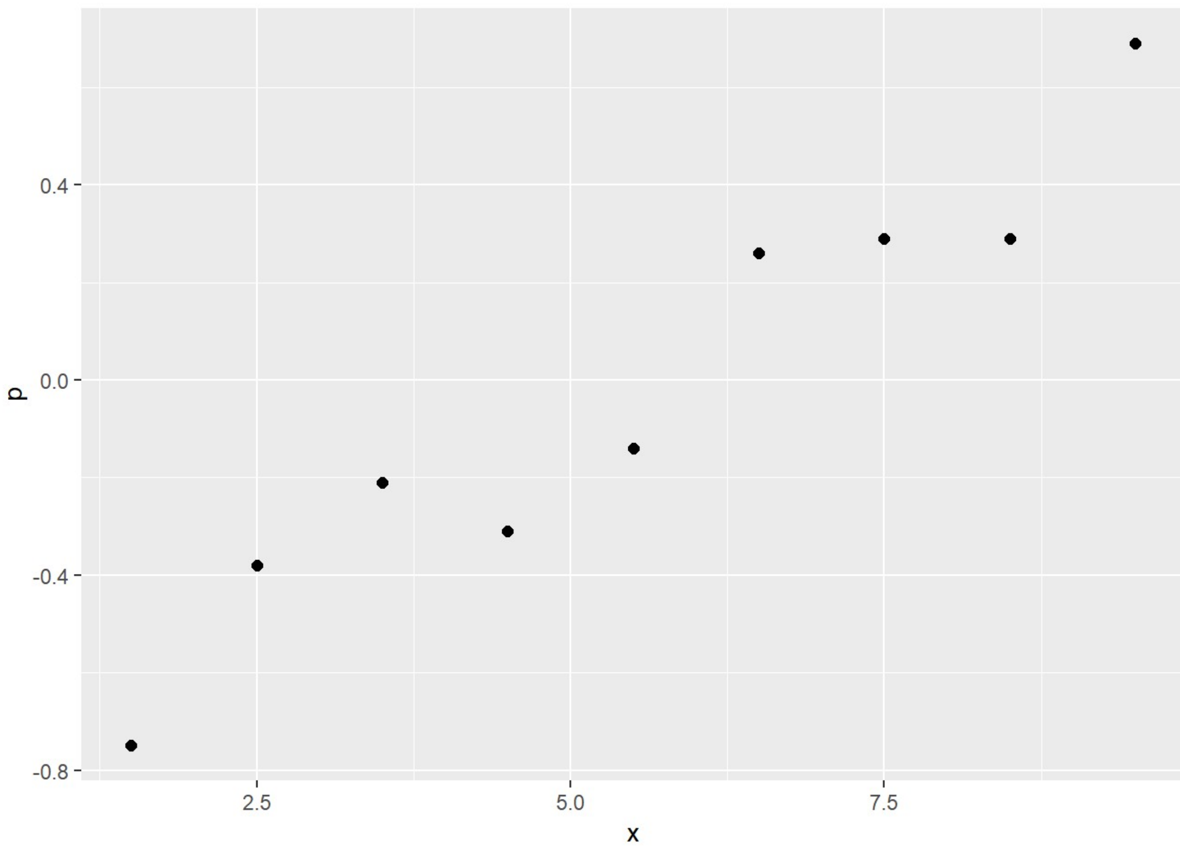
$$p_i = \ln \frac{m_i/n_i}{1 - m_i/n_i}$$

其中m_i表示实际购房人数，n_i表示签订意向书人数。
这种已经进行了logit变换的数据，称为“分组数据”；而变换后的模型就是线性回归模型：

$$p_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

下面，做出logit变换p_i和年家庭收入x_i的散点图，进一步确定这样的线性关系是否成立。

```
#画图观察
attach(Y)
library(ggplot2)
ggplot(Y,aes(x=x, y=p))+
  geom_point(size=2)+
  labs(x = "x", y = "p")
```



可以看出，线性关系非常明显。进行线性回归模型是合适的。

2.2 普通最小二乘回归

基于上述的种种分析，下面利用普通最小二乘法进行回归。

```
#---OLS普通最小二乘回归---
y=Y$p
x=Y$x
Y.m1<-lm(y~x)
summary(Y.m1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.148111 -0.111111  0.007556  0.115889  0.133222
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.8851     0.1015  -8.721 5.23e-05 ***
## x             0.1557     0.0167   9.320 3.40e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1294 on 7 degrees of freedom
## Multiple R-squared:  0.9254, Adjusted R-squared:  0.9148
## F-statistic: 86.87 on 1 and 7 DF,  p-value: 3.396e-05
```

模型整体和回归参数均十分显著。$R^2 = 0.92$, 拟合优度较高。得到的回归方程为：

$$\hat{p_i} = 0.8851 + 0.1557x_i$$

还原为Logistic回归方程为：

$$P(y_i = 1|x_i) = \frac{\exp(0.8851 + 0.1557x_i)}{1 + \exp(0.8851 + 0.1557x_i)}$$

此外，还可以进行预测。例如，令$x_0 = 8$，那么就可以求得预测概率为：

```
#进行预测
pre<-predict(Y.m1,data.frame(x=8))
p_hat=exp(pre)/(1+exp(pre))
p_hat
```

```
##         1
## 0.5891077
```

这说明，在住房展销会上与房地产商签订初步购房意向书的年收入8万元的家庭中，预计实际购房比例为58.9%.

2.3 加权最小二乘回归

尽管可以直接利用普通最小二乘回归对Logit变换做拟合，但是模型中可能会存在异方差的问题。

首先，对异方差问题进行诊断，诊断方法是求得解释变量和残差绝对值的相关系数矩阵。

```
res=abs(residuals(Y.m1))
library(psych)
corr.test(data.frame(x,res),use = "complete")
```

```
## Call:corr.test(x = data.frame(x, res), use = "complete")
## Correlation matrix
##         x   res
## x    1.00 -0.17
## res -0.17  1.00
## Sample Size
## [1] 9
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##        x  res
## x   0.00 0.67
## res 0.67 0.00
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

根据诊断结果，解释变量和残差绝对值的相关性并不显著，相关系数仅有-0.17，而检验的p-value则为0.67.
这从某种程度上说明了，模型并不存在严重的异方差问题。
在这样的判断下，以

$$w_i = n_i \cdot P(y_i = 1|x_i) \cdot (1 - P(y_i = 1|x_i))$$

为权重，重新对模型进行加权最小二乘估计：

```
#---WLS加权最小二乘回归---
w=Y$WLS
Y.m2<-lm(y~x,weights = w)
summary(Y.m2)
```

```
##
## Call:
## lm(formula = y ~ x, weights = w)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47475 -0.29729  0.04803  0.26578  0.44016
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.84932    0.11285  -7.526 0.000134 ***
## x            0.14946    0.02058   7.263 0.000168 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3838 on 7 degrees of freedom
## Multiple R-squared:  0.8828, Adjusted R-squared:  0.8661
## F-statistic: 52.74 on 1 and 7 DF,  p-value: 0.0001681
```

得到的结果并没有太大的差别。同时，再次对$x_0 = 8$进行预测，

```
#进行预测
pre<-predict(Y.m2,data.frame(x=8))
p_hat<-exp(pre)/(1+exp(pre))
p_hat
```

```
##         1
## 0.5857431
```

得到的结果也与之前的58.9%极为接近。

3 例9-3的分析

3.1 调用glm()函数进行估计

在上一节中，我们直接将logit变换的结果作为因变量进行线性回归，并称这样的数据为“分组数据”。然而，分组数据的Logistic回归只适用于大样本的分组数据，对小样本的为分组数据并不适用；并且以组数为回归拟合的样本量，使拟合的精度降低。因此，下面用极大似然估计直接拟合“未分组数据”的Logistic回归模型。

首先，在R中读入例9-3的数据集。

```
Z=read.csv("eg9-3.csv")
head(Z)
```

```
##   SEX AGE   X2 Y
## 1   0  18  850 0
## 2   0  21 1200 0
## 3   0  23  850 1
## 4   0  23  950 1
## 5   0  28 1200 1
## 6   0  31  850 0
```

```
tail(Z)
```

```
##    SEX AGE   X2 Y
## 23   1  38 1200 0
## 24   1  41 1500 0
## 25   1  45 1800 1
## 26   1  48 1000 0
## 27   1  52 1500 1
## 28   1  56 1800 1
```

其中，sex表示性别(定性数据)，age表示年龄,x_2表示月收入，y则是0-1变量.

$$sex = \begin{cases} 0, & \text{女性} \\ 1, & \text{男性} \end{cases}$$

$$y = \begin{cases} 0, & \text{乘坐公交车上下班} \\ 1, & \text{骑自行车上下班} \end{cases}$$

下面，调用glm()函数，直接估计Logistic模型。

```
#---进行估计---#
y=Z$Y
sex=Z$SEX
age=Z$AGE
x2=Z$X2
Z.m1<-glm(y~sex+age+x2,family = binomial(link = logit))
summary(Z.m1)
```

```
##
## Call:
## glm(formula = y ~ sex + age + x2, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1090  -0.7486  -0.2850   0.7011   2.1683
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.655016   2.091218  -1.748   0.0805 .
## sex         -2.501844   1.157815  -2.161   0.0307 *
## age          0.082168   0.052119   1.577   0.1149
## x2           0.001517   0.001865   0.813   0.4160
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38.673  on 27  degrees of freedom
## Residual deviance: 25.971  on 24  degrees of freedom
## AIC: 33.971
##
## Number of Fisher Scoring iterations: 5
```

对各个变量前的参数进行检验，观察p-value，发现age和x_2都是不显著的。因此，利用逐步回归的方法对变量进行筛选。

```
#---逐步回归,进行估计---#
Z.m2<-step(Z.m1)
```

```
## Start:  AIC=33.97
## y ~ sex + age + x2
##
##        Df Deviance    AIC
## - x2    1   26.653 32.653
## <none>      25.971 33.971
## - age   1   28.736 34.736
## - sex   1   32.175 38.175
##
## Step:  AIC=32.65
## y ~ sex + age
##
##        Df Deviance    AIC
## <none>      26.653 32.653
## - sex   1   32.218 36.218
## - age   1   33.446 37.446
```

```
summary(Z.m2)
```

```
##
## Call:
## glm(formula = y ~ sex + age, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1806  -0.6816  -0.3262   0.7890   1.9696
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.6285     1.5537  -1.692   0.0907 .
## sex          -2.2239     1.0476  -2.123   0.0338 *
## age           0.1023     0.0458   2.233   0.0256 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 38.673  on 27  degrees of freedom
## Residual deviance: 26.653  on 25  degrees of freedom
## AIC: 32.653
##
## Number of Fisher Scoring iterations: 5
```

可以看到，模型剔除了变量x_2. 剩下的变量都是显著的。
因此，得到的模型为

$$P(y_i=1|sex,age)=\frac{\exp(-2.6285-0.2239sex+0.1023age)}{1+\exp(-2.6285-0.2239sex+0.1023age)}$$

对所有系数取指数，得到：

```
exp(Z.m2$coefficients)
```

```
## (Intercept)         sex         age
##  0.07218629  0.10818865  1.10767704
```

这说明，男性对女性的优势比(Odds Ratio)为0.108，女性比男性更易乘公交车出行。此外，年龄每增长1岁，增长后与增长前优势比为 1.108，说明年龄越大，乘车的比例也越大。

3.2 多重共线性的检验

进行Logistic回归，一个重要的假设是，自变量之间不存在多重共线性。为此，计算方差膨胀因子(VIF)进行判断。

```
#---多重共线性的诊断(利用方差膨胀因子)---#
library(DAAG)
vif(Z.m2)
```

```
##    sex    age
## 1.1939 1.1939
```

由于变量的方差膨胀因子均没有超过10，可以认为模型不存在严重的多重共线性，模型的设定是正确的。

3.3 计算判别精度

最后，由于Logistic回归在判别上起到的巨大作用，我们将其视作一种判别分析，并计算判别的精度。

首先，根据数据集中的观测值，计算出$P(y=1|sex,age)$的拟合值。

```
#---最后,评价模型,计算判别精度---#
logit=predict(Z.m2,data.frame(sex=Z$SEX,age=Z$AGE))
probibility=exp(logit)/(1+exp(logit))
```

然后，根据拟合得到的$P(y=1|sex,age)$对样本进行判别，规则如下：

$$\begin{cases} P(y_i=1|sex,age)<0.5, & y_i=0 \\ P(y_i=1|sex,age)>0.5, & y_i=1 \\ P(y_i=1|sex,age)=0.5, & \text{无法判断} \end{cases}$$

```
newY=c()
for(i in 1:length(probibility)){
  if(probibility[i]<0.5)
    newY=c(newY,0)
  if(probibility[i]>0.5)
    newY=c(newY,1)
  if(probibility[i]==0.5)
    newY=c(newY,NA)
}
```

将上述计算得到的$P(y=1|sex,age)$拟合值和判别分组加入原数据集中，得到新的数据框。

```
new.Z=cbind(Z,probibility,newY)
new.Z
```

```
##    SEX AGE   X2 Y probibility newY
## 1    0  18  850 0  0.31265544    0
## 2    0  21 1200 0  0.38203047    0
## 3    0  23  850 1  0.43133447    0
## 4    0  23  950 1  0.43133447    0
## 5    0  28 1200 1  0.55846057    1
## 6    0  31  850 0  0.63221055    1
## 7    0  36 1500 1  0.74135770    1
## 8    0  42 1000 1  0.84112567    1
## 9    0  46  950 1  0.88851761    1
## 10   0  48 1200 0  0.90722539    1
## 11   0  55 1800 1  0.95239642    1
## 12   0  56 2100 1  0.95682416    1
## 13   0  58 1800 1  0.96452709    1
## 14   1  18  850 0  0.04690399    0
## 15   1  20 1000 0  0.05694264    0
## 16   1  25 1200 0  0.09147495    0
## 17   1  27 1300 0  0.10995242    0
## 18   1  28 1500 0  0.12036670    0
## 19   1  30  950 1  0.14375673    0
## 20   1  32 1000 0  0.17080937    0
## 21   1  33 1800 0  0.18578463    0
## 22   1  33 1000 0  0.18578463    0
## 23   1  38 1200 0  0.27561630    0
## 24   1  41 1500 0  0.34084827    0
## 25   1  45 1800 1  0.43771124    0
## 26   1  48 1000 0  0.51408105    1
## 27   1  52 1500 1  0.61429427    1
## 28   1  56 1800 1  0.70567280    1
```

不难看出，新的判别分组与原分组还是存在一定的出入。对误判的频数进行统计，得到误判列联表，其对角线上的元素为判断正确的频数，其他格中的元素为误判的频数。

```
table(y,newY)
```

```
##    newY
## y    0  1
##   0 12  3
##   1  4  9
```

基于此，就可以计算得到判断正确的频率。

```
sum(diag(prop.table(table(y,newY))))
```

```
## [1] 0.75
```

判别精度为75%，效率尚可，这也在一定程度上验证了Logistic回归模型的重要作用。