广义线性模型

kaka

2020/11/13

Logistic模型——例子5.1

y：是否发生事故（1：出过事故、0：没有）
x_1:视力情况（1：正常、0：有问题）
x_2:年龄
x_3:驾车教育（1：参加过、0：没有）

```
d5.1=read.xlsx('mvstats5.xlsx','d5.1')
logit.glm<-glm(y~x1+x2+x3,family=binomial,data=d5.1)
summary(logit.glm)
```

```
Call:
glm(formula = y ~ x1 + x2 + x3, family = binomial, data = d5.1)

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
```

Logistic模型估计：

$$log(\frac{\hat{p}}{1-\hat{p}}) = 0.5976 - 1.4961x_1 - 0.0016x_2 + 0.3159x_3$$

只有视力情况情况影响显著，正常视力的发生事故优势比对数较有问题的降低了约1.5。年龄和驾车教育对发生事故优势比的影响分别为负和正，但统计意义上不显著。

手编极大似然估计

```
Y=d5.1[,1];X=as.matrix(cbind(x0=1,d5.1[,2:4]))
maxIterNum <- 20000;  #最大迭代次数
W <- rep(0, ncol(X))  #系数估计向量，初始为0向量
sigmoid <- function(z) { 1 / (1 + exp(-z))}
for (i in 1:maxIterNum){
  P=as.vector(sigmoid(X %*% W))
  grad <- t(X) %*% (P-Y);
  grad2<- t(X) %*% diag(P*(1-P))%*%X
  if (sqrt(as.numeric(t(grad) %*% grad)) < 1e-8){
    print(sprintf('iter times=%d', i));
    break;
  }
  W <- W - solve(grad2)%*%grad  #Newton-Raphson迭代
}
```

```
[1] "iter times=5"
```

```
sd=(diag(solve(grad2)))^0.5 #系估计标准差
cbind(Est=round(W[,1],6),SdEr=round(sd,6))
```

```
         Est     SdEr
x0  0.597610 0.894831
x1 -1.496084 0.704862
x2 -0.001595 0.016758
x3  0.315865 0.701094
```

自编程序迭代5次结束,根据公式计算的结果与函数输出基本一致。利用估计值与标准差相除可得近似正态统计量，在样本容量足够大情况下可以对系数是否为零作检验。

模型评价

```
#AIC值
aic=-2*(t(Y)%*%X %*% W-sum(log(1+exp(X %*% W))))+8
#伪R方
psR=1-logit.glm$deviance/logit.glm$null.deviance
#概率预测值
#logit.glm$fitted.values
#注意不能用predict函数
#y预测值
yh=as.numeric(logit.glm$fitted.values>0.5)
#混淆矩阵
hx=table(yh,Y)
#预测正确比例
pcp=sum(diag(hx))/sum(hx)
```

```
AIC值= 65.02562
伪R方= 0.08294016
混淆矩阵=
   Y
yh   0  1
  0 17  8
  1  7 13
预测正确比例= 0.6666667
```

模型解释

```
#均值边际效应(PEA)
xbat<-(apply(X,2,mean))
pea=dlogis(sum(xbat*W))*W
#平均边际效应(AME)
ame=mean(dlogis(X %*% W))*W
cbind(pea=pea[,1],ame=ame[,1])
```

```
             pea           ame
x0  0.1486280508  0.1320413779
x1 -0.3720822185 -0.3305583878
x2 -0.0003967285 -0.0003524542
x3  0.0785568659  0.0697900347
```

两种边际效数值很接近，应由于x_1为分类变量平均边际效应的解释更为合理，即视力正常比视力有问题发生事故的概率要低约33%。

模型变量选择与预测

```
logit.step<-step(logit.glm,direction="both")#逐步筛选法变量选择
```

```
Start:  AIC=65.03
y ~ x1 + x2 + x3

       Df Deviance    AIC
- x2    1   57.035 63.035
- x3    1   57.232 63.232
<none>      57.026 65.026
- x1    1   61.936 67.936

Step:  AIC=63.03
y ~ x1 + x3

       Df Deviance    AIC
- x3    1   57.241 61.241
<none>      57.035 63.035
+ x2    1   57.026 65.026
- x1    1   61.991 65.991

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(logit.step)#逐步筛选法变量选择结果
```

```
Call:
glm(formula = y ~ x1, family = binomial, data = d5.1)

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
```

```
pre1<-predict(logit.step,data.frame(x1=1))#预测视力正常司机Logistic回归结果
p1<-exp(pre1)/(1+exp(pre1))#预测视力正常司机发生事故概率
pre2<-predict(logit.step,data.frame(x1=0))#预测视力有问题的司机Logistic回归结果
p2<-exp(pre2)/(1+exp(pre2))#预测视力有问题的司机发生事故概率
c(p1,p2)#结果显示
```

```
   1    1
0.32 0.65
```

逐步回归的结果只保留x_1，视力正常的事故发生概率为32%，视力有问题的事故发生鼓励为65%，大约相差33%，与前面分析相同。

对数线性模型——例子5.2

y：是人数
x_1:收入水平（1:高,2:中,3:低）
x_2:是否满意（1：满意,2:不满意）

```
d5.2=read.xlsx('mvstats5.xlsx','d5.2');head(d5.2)
```

```
    y x1 x2
1  53  1  1
2 434  2  1
3 111  3  1
4  38  1  2
5 108  2  2
6  48  3  2
```

```
d5.2$x1=factor(d5.2$x1)
d5.2$x2=factor(d5.2$x2)
#列联表卡方独立性检验
chisq.test(matrix(c(53,434,111,38,108,48),3,2))
```

```

    Pearson's Chi-squared test

data:  matrix(c(53, 434, 111, 38, 108, 48), 3, 2)
X-squared = 23.567, df = 2, p-value = 7.628e-06
```

```
library(MASS)
#对数线性模型检验，检验模型和含有交互作用模型的差异
#原假设模型不存在交互作用，即两变量独立。
#大样本下该检验等价与卡方独立性检验
loglm(y~x1+x2,data=d5.2)
```

```
Call:
loglm(formula = y ~ x1 + x2, data = d5.2)

Statistics:
                      X^2 df     P(> X^2)
Likelihood Ratio 22.08692  2 1.599140e-05
Pearson          23.56750  2 7.627506e-06
```

无论是卡方独立性检验(统计量为23.57)还是对数线性模型检验(统计量为22.09)都拒绝原假设，即认为收入和满意程度不独立，存在交互作用。理论上应该在模型中加入交互作用项，但由于每个单元格只有一个数据，只能建立个体效应模型。

个体效应模型估计

```
log.glm<-glm(y~x1+x2,family=poisson(link=log),data=d5.2)#多元对数线性模型
summary(log.glm)#多元对数线性模型结果
```

```
Call:
glm(formula = y ~ x1 + x2, family = poisson(link = log), data = d5.2)

Deviance Residuals:
     1       2       3       4       5       6
-1.975   1.212  -0.837   3.020  -2.222   1.399

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.22989    0.10676  39.619  < 2e-16 ***
x12          1.78441    0.11329  15.751  < 2e-16 ***
x13          0.55804    0.13145   4.245 2.18e-05 ***
x22         -1.12573    0.08262 -13.625  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 662.843  on 5  degrees of freedom
Residual deviance:  22.087  on 2  degrees of freedom
AIC: 68.072

Number of Fisher Scoring iterations: 4
```

又结果得两个分类变量的人数都有显著差异，中低收入的人数比例都比高收入的显著高；而不满意的人数比例比不满意的显著低。

手编极大似然估计

```
Y=d5.2[,1];X=as.matrix(cbind(x0=1,x12=(d5.2[,2]==2),x13=(d5.2[,2]==3),x22=(d5.2[,3]==2)))
maxIterNum <- 20000;
W <- rep(1, ncol(X))
m = nrow(X)
for (i in 1:maxIterNum){
  la=as.vector(exp(X %*% W))
  grad <- t(X) %*% (la-Y);
  grad2<- t(X) %*% diag(la)%*%X
  if (sqrt(as.numeric(t(grad) %*% grad)) < 1e-8){
    print(sprintf('iter times=%d', i));
    break;
  }
  W <- W - solve(grad2)%*%grad
}
```

```
[1] "iter times=43"
```

```
sd=(diag(solve(grad2)))^0.5 #系估计标准差
cbind(Est=round(W[,1],6),SdEr=round(sd,6))
```

```
          Est     SdEr
x0   4.229889 0.106764
x12  1.784406 0.113287
x13  0.558045 0.131447
x22 -1.125733 0.082625
```

```
#AIC值
s=0
for(i in 1:length(Y)){
  s=s+sum(log(1:Y[i]))
}
-2*(t(Y)%*%X %*% W-sum(exp(X %*% W))-s)+8
```

```
        [,1]
[1,] 68.0724
```

自编程序迭代43次结束,需要先进分类变量虚拟化，根据公式计算的结果与函数输出基本一致，AIC值为68.07。

一般线性模型

—— 5.3.1 完全随机设计模型 ————————

```
d5.3=read.xlsx('mvstats5.xlsx','d5.3')
summary(lm(Y~factor(A),data=d5.3))
```

```
Call:
lm(formula = Y ~ factor(A), data = d5.3)

Residuals:
      Min        1Q    Median        3Q       Max
-0.068333 -0.027500  0.000833  0.030417  0.051667

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.42833    0.01585 153.185  < 2e-16 ***
factor(A)2   0.13167    0.02242   5.873 3.06e-05 ***
factor(A)3   0.19833    0.02242   8.847 2.44e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03883 on 15 degrees of freedom
Multiple R-squared:  0.8439,    Adjusted R-squared:  0.823
F-statistic: 40.53 on 2 and 15 DF,  p-value: 8.94e-07
```

```
anova(lm(Y~factor(A),data=d5.3))
```

```
Analysis of Variance Table

Response: Y
          Df   Sum Sq  Mean Sq F value   Pr(>F)
factor(A)  2 0.122233 0.061117  40.534 8.94e-07 ***
Residuals 15 0.022617 0.001508
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

单因素方差分析等价于自含有一个分类自变量的一般线性模型，其F检验等价，但一般线性模型能估计出各因素水平的效应差异。因素A的第二和第三个水平效应显著高于第一个水平。

5.3.2 随机区组设计模型

```
d5.4=read.xlsx('mvstats5.xlsx','d5.4')
anova(lm(Y~factor(A)+factor(B),data=d5.4))
```

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value Pr(>F)
factor(A)  3  15759    5253  0.4306 0.7387
factor(B)  2  22385   11192  0.9174 0.4491
Residuals  6  73198   12200
```

```
summary(lm(Y~factor(A)+factor(B),data=d5.4))
```

```
Call:
lm(formula = Y ~ factor(A) + factor(B), data = d5.4)

Residuals:
    Min      1Q  Median      3Q     Max
-114.50  -60.42  -26.83   74.08  116.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   634.17      78.10   8.120 0.000187 ***
factor(A)2    -83.00      90.18  -0.920 0.392893
factor(A)3    -31.67      90.18  -0.351 0.737488
factor(A)4     10.00      90.18   0.111 0.915324
factor(B)2     -9.50      78.10  -0.122 0.907158
factor(B)3    -96.00      78.10  -1.229 0.265022
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 110.5 on 6 degrees of freedom
Multiple R-squared:  0.3426,    Adjusted R-squared:  -0.2053
F-statistic: 0.6253 on 5 and 6 DF,  p-value: 0.6885
```

多因素方差分析等价于自含有多个分类自变量的一般线性模型，将不同因素水平的系数联合作F检验。从方差分析中认为因素A和B都没有显著的个体效应。从一般线性模型的结果看，各系数也不显著，认为不同因素水平的影响没有显著差别。原则上模型应该考虑交互效应，但单元格的数据只有一个，不能估计交互效应模型的参数。

5.3.3 析因设计模型

```
d5.5=read.xlsx('mvstats5.xlsx','d5.5')
anova(lm(Y~factor(A)*factor(B),data=d5.5))
```

```
Analysis of Variance Table

Response: Y
                   Df Sum Sq Mean Sq F value    Pr(>F)
factor(A)           1   1600 1600.00  28.402 0.0001795 ***
factor(B)           1   2500 2500.00  44.379 2.321e-05 ***
factor(A):factor(B)  1    729  729.00  12.941 0.0036638 **
Residuals          12    676   56.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(Y~factor(A)*factor(B),data=d5.5))
```

```
Call:
lm(formula = Y ~ factor(A) * factor(B), data = d5.5)

Residuals:
   Min     1Q Median     3Q    Max
-14.50  -3.50  -0.25   3.75  12.00

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            47.000      3.753  12.524 3.00e-08 ***
factor(A)2             -6.500      5.307  -1.225  0.24417
factor(B)2             38.500      5.307   7.254 1.01e-05 ***
factor(A)2:factor(B)2 -27.000      7.506  -3.597  0.00366 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.506 on 12 degrees of freedom
Multiple R-squared:  0.8772,    Adjusted R-squared:  0.8465
F-statistic: 28.57 on 3 and 12 DF,  p-value: 9.511e-06
```

factor(A)*factor(B)表示含有两个因素的饱和模型(即包含个体效应也包含全部的交互效应项)。从方差分析结果看，两个因素的个体效应显著，且交互效应也显著。交互效应显著意味不能单独比较某因素不同水平的高度，而要考另一个因素的具体取值。如47是A1B1的效应，A2B1的效应为47-6.5，A1B2的效应为47+38.5；但A2B2的效应不是47-6.5+38.5，而是47-6.5+38.5-27，即因素A、B同时从1变化2会产生额外的负效应。 ## —— 5.3.4 正交设计模型 ————————

```
d5.6=read.xlsx('mvstats5.xlsx','d5.6')
anova(lm(Y~factor(A)*factor(B)+factor(C)+factor(D),data=d5.6))
```

```
Analysis of Variance Table

Response: Y
                   Df Sum Sq Mean Sq F value  Pr(>F)
factor(A)           1    8.0     8.0     3.2 0.21554
factor(B)           1   18.0    18.0     7.2 0.11535
factor(C)           1   60.5    60.5    24.2 0.03893 *
factor(D)           1    4.5     4.5     1.8 0.31175
factor(A):factor(B) 1   50.0    50.0    20.0 0.04654 *
Residuals           2    5.0     2.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(lm(Y~factor(A)*factor(B)+factor(C)+factor(D),data=d5.6))
```

```
Call:
lm(formula = Y ~ factor(A) * factor(B) + factor(C) + factor(D),
    data = d5.6)

Residuals:
   1    2    3    4    5    6    7    8
-1.0  1.0  0.5 -0.5 -0.5  0.5  1.0 -1.0

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            87.000      1.369  63.536 0.000248 ***
factor(A)2              3.000      1.581   1.897 0.198216
factor(B)2              2.000      1.581   1.265 0.333333
factor(C)2              5.500      1.118   4.919 0.038926 *
factor(D)2              1.500      1.118   1.342 0.311753
factor(A)2:factor(B)2 -10.000      2.236  -4.472 0.046537 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.581 on 2 degrees of freedom
Multiple R-squared:  0.9658,    Adjusted R-squared:  0.8801
F-statistic: 11.28 on 5 and 2 DF,  p-value: 0.08343
```

正交设计是一种实验设计，通常如果要建立包含交互作用的模型需要大量实验收集数据，增加研究成本。而正交设计在小量实验的情况下仍能有效建立交互作用模型，并进行假设检验。由方差分析和一般线性模型的输出结果看，因素C和AB的交互项的效应显著，即C由水平1变为水平2效应增加5.5，但因素A、B同时从1变化2会产生-10的负效应。