多元统计分析及R语言建模

第5章 广义与一般线性模型及R使用

王斌会 教授

5 广义与一般线性模型及R使用

- ★内容与要求
  - 内容
  - 要求
- 5.1 数据分类与模型选择
  - 5.1.1变量的取值类型
  - 5.1.2模型选择方式
- 5.2 广义线性模型
  - 5.2.1广义线性模型概述
  - 5.2.2 Logistic模型
  - 5.2.3对数线性模型
- 5.3 一般线性模型
  - 5.3.1完全随机设计模型
  - 5.3.2随机单位组设计模型
- 思考练习题
  - 思考题
    - 手工解答，上交作业本
  - 练习题
    - 计算机分析，发电子邮件

内容与要求

●基本内容：

数据的分类与模型选择、广义线性模型概述、Logistic回归模型、对数线性模型、一般线性模型的计算。

●基本要求：

要求学生针对因变量和解释变量的取值性质，了解统计模型的类型。掌握数据的分类与模型选择方法，并对广义线性模型和一般线性模型有初步的了解。

1.变量的取值类型：

$$\text{因变量 } y \in \begin{cases} \text{连续变量} \\ \text{“0 – 1”变量或称二分类变量} \\ \text{有序变量(等级变量)} \\ \text{多分类变量} \\ \text{连续伴有删失变量} \end{cases}$$

$$\text{解释变量 } x \in \begin{cases} \text{连续变量} \\ \text{分类变量} \\ \text{等级变量} \end{cases}$$

2.模型选择方式：基本公式

$$
\begin{cases} Y = X\beta + e \\ E(e) = 0,\ \mathrm{cov}(e) = \sigma^2 I \end{cases}
$$

X \ y	连续变量	0-1变量	有序变量	多分类变量	连续伴有删失
连续变量	线性回归方程	logistic回归模型	累积比数模型 对数线性模型	对数线性模型 多分类logistic回归模型	cox比例风险模型
分类变量	实验设计模型（方差分析模型）				
连续变量 分类变量	协方差分析模型				

在广义线性模型中，均假定观察值y具有指数族概率密度函数

$$f(y|\theta,\phi)=exp\{[y\theta-b(\theta)]/a(\phi)+c(y,\phi)\}$$

其中$a(\cdot)$、$b(\cdot)$和$c(\cdot)$是三种函数形式，θ为典则参数。

表5.1 广义线性模型中的常用分布族

分布	函数	模型
正态（Gaussian）	$E(y)=X'\beta$	普通线性模型
二项（Binomial）	$E(y)=\frac{exp(X'\beta)}{1+exp(X'\beta)}$	Logistic 模型和概率模型单位（probit）模型
泊松（Poission）	$E(y)=exp(X'\beta)$	对数线性模型

在广义线性模型中，（5.4）式中的典则参数不仅仅是μ的函数，还是参数$\beta_0, \beta_1, \ldots, \beta_p$的线性表达式。对$\mu$作变换，则可得到这三种分布连接函数的形式

正态分布：$m(\mu) = \mu = \sum \beta_j x_j$

二项分布：$m(\mu) = log(\frac{\mu}{1-\mu}) = \sum \beta_j x_j$

Poisson 分布：$m(\mu) = log(\mu) = \sum \beta_j x_j$

广义线性模型函数 glm()的用法

glm(formula, family = gaussian, data,...)

formula 为公式，即为要拟合的模型

family 为分布族，包括正态分布（gaussian）、二项分布（binomial）、泊松分布（poission）和伽玛分布（gamma），分布族还可以通过选项 link=来指定使用的连接函数

data 为可选择的数据框。

说明与举例

说明：

2、**Logistic模型**：函数形式 $logit(y) = ln\dfrac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = X\beta$

其中参数估计采用极大似然估计。

举例：

对45名驾驶员的调查结果，其中四个变量的含义为：

x1：表示视力状况，它是一个分类变量，1表示好，0表示有问题；

x_2：年龄，数值型；

x3：驾车教育，它也是一个分类变量，1表示参加过驾车教育，0表示没有；

y：分类变量（去年是否出过事故，1表示出过事故，0表示没有）。

程序与结果

（1）建立全变量logistic回归模型：

```
d5.1=read.table("clipboard",header=T)  #读取例5.1数据
logit.glm<-glm(y~x1+x2+x3,family=binomial,data=d5.1)  #Logistic回归模型
summary(logit.glm)  #Logistic回归模型结果
```

```
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
```

得到初步的logistic回归模型：

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

程序与结果

（2）逐步筛选变量logistic回归模型：

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

```
summary(logit.step)
#逐步筛选法变量选择结果
```

```
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
```

```
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
```

由此得到新的logistic回归模型：

$$p=\frac{exp(0.6190-1.3728x_1)}{1+exp(0.6190-1.3728x_1)}$$

程序与结果

（3）：预测发生交通事故的概率

```
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
```

说明与举例

说明：

3、对数线性模型：函数形式 $ln(m_{ij}) = \alpha_i + \beta_j + \varepsilon_{ij}$ $ln(m_{ij}) = \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$

式2含有交叉效应

举例：

某企业想了解顾客对其产品是否满意，同时还想了解不同收入的人群对其产品的满意程度是否相同。

	满意	不满意	合计
高	53	38	91
中	434	108	542
低	111	48	159
合计	598	194	792

数据形式变为：用y表示频数，x1表示收入人群，x2表示满意程度

y	x1	x2
53	1	1
434	2	1
111	3	1
38	1	2
108	2	2
48	3	2

程序与结果

（1）建立Poisson对数线性模型：

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

Deviance Residuals:
      1        2        3        4        5        6
-10.784   14.444   -8.468   -2.620    4.960   -3.142

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  6.15687    0.14196  43.371  < 2e-16 ***
x1           0.12915    0.04370   2.955  0.00312 **
x2          -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: 662.84  on 5  degrees of freedom
Residual deviance: 437.97  on 3  degrees of freedom
AIC: 481.96

Number of Fisher Scoring iterations: 5
```

从检验结果可看出，p1=0.0031<0.01，p2<0.01，说明收入和满意程度对产品有重要影响

说明与举例

说明：

1、完全随机设计模型：函数形式 $y_{ij}=\mu+\alpha_i+e_{ij} i=1,2,\ldots,G j=1,2,\ldots,n_i$

其中μ表示观察结果y_{ij}的总体均值，α_i是哑变量的系数，称为A因素各水平的主效应，e_{ij}是误差项。

举例：

设有3台机器，用来生产规格相同的铝合金薄板。现从3台机器生产出的薄板中各随机抽取5块，测出厚度值，见下表，试分析各机器生产的薄板厚度有无显著差异？

机器1	2.36	2.38	2.48	2.45	2.47	2.43
机器2	2.57	2.53	2.55	2.54	2.56	2.61
机器3	2.58	2.64	2.59	2.67	2.66	2.62

程序与结果

```
d5.3=read.table("clipboard",header=T)  #读取例5.3数据
anova(lm(Y~factor(A),data=d5.3))  #完全随机设计模型方差分析
```

（1）数据格式为：

Y	A
2.36	1
2.38	1
2.48	1
2.45	1
2.47	1
2.43	1
2.57	2
2.53	2
2.55	2
2.54	2
2.56	2
2.61	2
2.58	3
2.64	3
2.59	3
2.67	3
2.66	3
2.62	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
```

P<0.05，说明各机器生产的薄板厚度有显著差异。

说明与举例

○ 说明：

2、随机单位组设计模型：函数形式 $y_{ij} = \mu + \alpha_i + \beta_j + e + iji = 1, 2, \ldots, Gj = 1, 2, \ldots, n$

其中μ为总均数，α_i为处理因素A的第i个水平的效应；β_j为第j个单位组的效应，e_{ij}为误差项。

○举例：

使用4种燃料，3种推进器作火箭射程试验，每一种组合情况做一次试验，则得火箭射程列在下表中，试分析各种燃料A与各种推进器B对火箭射程有无显著影响？

BA	A1	A2	A3	A4
B1	582	491	601	758
B2	562	541	709	582
B3	653	516	392	487

程序与结果

```
d5.4=read.table("clipboard",header=T)  #读取例5.4数据
anova(lm(Y~factor(A)+factor(B),data=d5.4))  #随机单位组设计模型方差分析
```

（1）数据格式为：

Y	A	B
582	1	1
491	2	1
601	3	1
758	4	1
562	1	2
541	2	2
709	3	2
582	4	2
653	1	3
516	2	3
392	3	3
487	4	3

```
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
```

PA>0.05，说明各种燃料A对火箭射程有无显著影响，

PB>0.05，说明各种推进器B对火箭射程也无显著影响。

关于40个不同年龄（age，定量变量）和性别（sex，定性变量，用0和1代表女和男）的人对某项服务产品的观点（y，二水平定性变量，用1和0代表认可与不认可）的数据。

一、数据管理

二、R语言操作

$$\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 x \quad \text{或者等价地} \quad p=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

拟合的模型为：

```
Case4=read.table("clipboard",header=T);Case4
fm=glm(y~sex+age,family=binomial,data=Case4)
fm
summary(fm)
attach(Case4)
Pr=predict(fm,data.frame(list(sex,age)))  #模型预测
p=exp(Pr)/(1+exp(Pr))
cbind(sex,age,y,p)
plot(age,Pr)
detach(Case4)
```

谢谢！