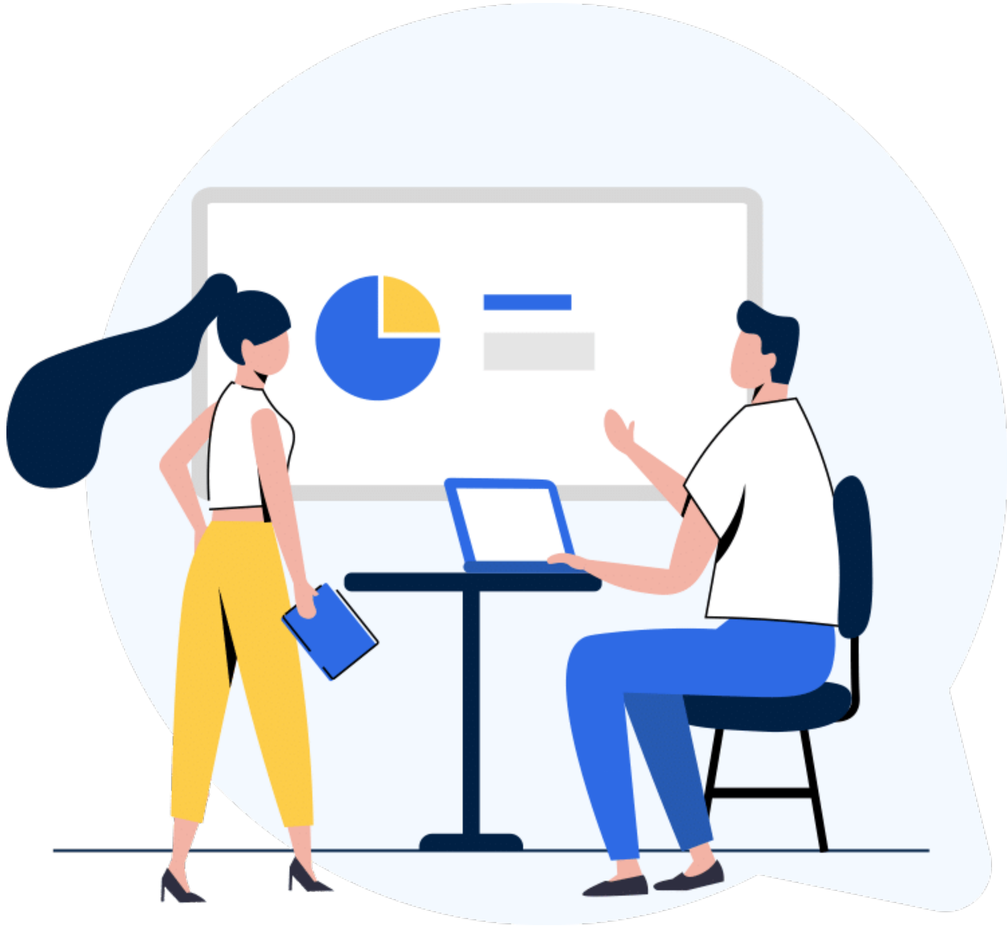
02

定性数据的建模分析



➢学习目标：

1. 掌握对数线性模型的基本原理;
2. 掌握对数线性模型的建模方法;
3. 掌握如何解释Logistic 回归的分析结果;
4. 理解判别分析与Logistic 回归相比的优缺点;
5. 掌握如何通过R 软件实现Logistic 回归。

2.1 对数线性模型的基本理论和方法

表9-1 频数表

A	B		
	B	$\bar{B}$	$\sum j$
A	n_{11}	n_{12}	$n_{1\cdot}$
$\bar{A}$	n_{21}	n_{22}	$n_{2\cdot}$
$\sum j$	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

表9-2 频率表

A	B		
	B	$\bar{B}$	$\sum j$
A	p_{11}	p_{12}	$p_{1\cdot}$
$\bar{A}$	p_{21}	p_{22}	$p_{2\cdot}$
$\sum j$	$p_{\cdot 1}$	$p_{\cdot 2}$	$p_{\cdot\cdot}$

2.1 对数线性模型的基本理论和方法

在对数线性型分析中, 要先将概率取对数, 再分解处理, 用公式表示如下:

$$\eta_{ij} = ln\, p_{ij} = ln\left(p_{i\cdot}p_{\cdot j}\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}}\right)$$

$$= ln\, p_{i\cdot} + ln\, p_{\cdot j} + ln\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}}\ \ i,j = 0,1$$

把上式中的$ln\, p_{i\cdot}$, $ln\, p_{\cdot j}$, $\ln\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}}$分别记为 A_i, B_j 和 $(AB)_{ij}$, 则上式可写成:

$$\eta_{ij} = A_i + B_j + (AB)_{ij}$$

该式的结构与有交互效应且各水平均为2的双因素方差分析模型的结构相似, 因此模仿方差分析, 可以有如下关系式:

$$\eta_{i\cdot} = \sum_{j=1}^{2}\eta_{ij}\ ,\quad \eta_{\cdot j} = \sum_{i=1}^{2}\eta_{ij}\ ,\quad \eta_{\cdot\cdot} = \sum_{i=1}^{2}\sum_{j=1}^{2}\eta_{ij}$$

2.1 对数线性模型的基本理论和方法

若记：

$$\begin{cases} \alpha_i = \bar{\eta}_{i\cdot} - \bar{\eta}_{\cdot\cdot} \\ \beta_j = \bar{\eta}_{\cdot j} - \bar{\eta}_{\cdot\cdot} \\ \gamma_{ij} = \eta_{ij} - \bar{\eta}_{i\cdot} - \bar{\eta}_{\cdot j} + \bar{\eta}_{\cdot\cdot} \end{cases}$$

则：

$$\begin{aligned} \gamma_{ij} &= \eta_{ij} - \bar{\eta}_{i\cdot} - \bar{\eta}_{\cdot j} + \bar{\eta}_{\cdot\cdot} \\ &= \eta_{ij} - (\bar{\eta}_{i\cdot} - \bar{\eta}_{\cdot\cdot}) - (\bar{\eta}_{\cdot j} - \bar{\eta}_{\cdot\cdot}) - \bar{\eta}_{\cdot\cdot} \\ &= \eta_{ij} - \alpha_i - \beta_j - \bar{\eta}_{\cdot\cdot} \end{aligned}$$

移项，可得与有交互效应的双因素方差分析数学模型极为相似的关系式:

$$\begin{cases} \eta_{ij} = \bar{\eta}_{\cdot\cdot} + \alpha_i + \beta_j + \gamma_{ij} \\ \sum_{i=1}^{2} \alpha_i = \sum_{j=1}^{2} \beta_j = \sum_{i=1}^{2} \gamma_{ij} = 0, i = 1,2; j = 1,2 \end{cases}$$

2.1 对数线性模型的基本理论和方法

◆根据γ_{ij}值的正负和相对大小, 可以判断A因素的第i 个水平与B因素的第j个水平间的交互效应。
若$\gamma_{ij} > 0$，表明二者存在正效应;
若$\gamma_{ij} < 0$，则存在负效应;
当γ_{ij}均为0时，A, B因素相互独立。

◆ 若γ_{ij}均为0, 模型称为非饱和模型 (因素间相互独立) , 否则为饱和模型 (因素间有交互效应) 。

◆ 在实际分析中, 概率表中各项值以交叉列联表计算得到的频率表的对应项为无偏估计值。

2.2 对数线性模型的上机实现

【例9-1】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度是否相同。 在随机发放的1 000 份问卷中收回有效问卷792 份，根据收入高低和满意回答的交叉分组数据如表9-3所示。

表 9-3

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

2.2 对数线性模型的上机实现

◆ 首先要准备数据, 上面的交叉列联表的数据要表格里, 具体如表9-4所示。

表 9-4

频数	收入情况	满意情况
53	高	满意
434	中	满意
111	低	满意
38	高	不满意
108	中	不满意
48	低	不满意

2.2 对数线性模型的上机实现

◆ 具体R代码如下：

```
1.> rm(list=ls())
2.> library(MASS)
3.> ex9.1 <- read.table("例9-1.txt", head=TRUE, fileEncoding="utf8")
4.> fit <- MASS::loglm(频数~收入情况+满意情况+收入情况*满意情况, data=ex9.1, param=T,fit=T)
5.>#模型的拟合优度检验
6.> fit
7.Statistics:
8.                      X^2  df  P(> X^2)
9.Likelihood Ratio        0   0         1
10.Pearson                0   0         1
11.>#估计的系数
12.> coef(fit)
13.$'(Intercept)'
14.[1] 4.490631
15.$收入情况
16.        低          高          中
17.-0.2002652 -0.6866918  0.8869570
18.$满意情况
19.        不满意       满意
20.  -0.4269914  0.4269914
21.$收入情况.满意情况
22.满意情况
23.收入情况        不满意        满意
24.        低  0.00782678 -0.00782678
25.        高  0.26063850 -0.26063850
26.        中 -0.26846528  0.26846528
```

2.2 对数线性模型的上机实现

- 由于是饱和模型, 所以模型的拟合优度检验 Likelihood Ratio 方法和Pearson 方法的值和自由度均为0。 我么可以得到各参数为:

$$\alpha_{高收入} = -0.687$$
$$\alpha_{中收入} = 0.887$$
$$\alpha_{低收入} = -0.200$$
$$\beta_{满意} = 0.427$$
$$\beta_{不满意} = -0.427$$
$$\gamma_{高收入满意} = -0.261$$
$$\gamma_{中收入满意} = 0.268$$
$$\gamma_{低收入满意} = -0.007$$
$$\gamma_{高收入不满意} = 0.260$$
$$\gamma_{中收入不满意} = -0.268$$
$$\gamma_{低收入不满意} = 0.008$$

2.2 对数线性模型的上机实现

- 参数值为正, 表示正效应; 反之为负效应; 零为无效应。 分析提供的信息是:

  (1) $\beta_{满意}$为正值, 说明接受调查的多数顾客对其产品还是满意的。

  (2) $\alpha_{高收入} < \alpha_{低收入} < \alpha_{中收入}$, 说明各收入阶层的顾客对其产品的满意度是不同的, 其中, 高收入的顾客满意度最低, 而中等收入的顾客满意度最高。

  (3) $\gamma_{高收入满意}$为负值, 表示高收入对其对产品的满意度有负效应; $\gamma_{中收入满意}$为正值, 表示中等收入对其对产品的满意度有正效应; 同理, 低收入顾客对产品的满意程度也有负效应。

2.3 Logistic 回归的基本理论和方法

◆ 通常我们需要研究某一社会现象发生的概率p的大小, 比如一个公司成功或失败的概率, 以及讨论p的大小与哪些因素有关。但是, 直接处理可能性数值p存在困难:

（1）$0 \leq p \leq 1$, 因此p与自变量的关系难以用线性模型来描述;

（2）当p接近0 或1 时, p值的微小变化用普通的方法难以发现和处理好。

➢ 不处理参数p, 而处理p的一个严格单调函数$Q = Q(p)$

2.3 Logistic 回归的基本理论和方法

2.3.1分组数据的Logistic 回归模型

◆ 针对0-1 型因变量产生的问题, 我们对回归模型应该做两个方面的改进:

1.回归函数应该改用限制在 [0, 1] 区间内的连续曲线, 而不能再沿用直线回归方程。 限制在 [0, 1] 区间内的连续曲线有很多, 我们常用的是 Logistic 函数与正态分布函数。Logistic 函数的形式为:

$$f(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

2.因变量y_i本身只取0,1两个离散值,不适合直接作为回归模型中的因变量。由于回归函数$E(y_i) = \pi_i = \beta_0 + \beta_1 x_i$表示在自变量为xi的条件下$y_i$的平均值, y_i是 0-1 型随机变量,从而$E(y_i) = \pi_i$就是在自变量为x_i的条件下y_i等于1的比例。这提示我们可以用y_i等于1的比例代替y_i本身作为因变量。

.

2.3 Logistic 回归的基本理论和方法

2.3.1分组数据的Logistic 回归模型

【例9-2】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=313名顾客。在随后的3个月内,只有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。以顾客的年家庭收入(万元)为自变量x,对表9-5中的数据建立Logistic回归模型。

表9-5

序号	年家庭收入(万元) x	签订意向书人数 n_i	实际购房人数 m_i	实际购房比例 $p_i = m_i/n_i$	逻辑变换 $p_i' = \log(\frac{p_i}{1-p_i})$	权重 $\omega_i = n_i p_i(1-p_i)$
1	1.5	25	8	0.320 000	-0.753 77	5.440
2	2.5	32	13	0.406 250	-0.379 49	7.719

2.3 Logistic 回归的基本理论和方法

2.3.1分组数据的Logistic 回归模型

(续表)

序号	年家庭收入（万元）x	签订意向书人数 n_i	实际购房人数 m_i	实际购房比例 $p_i = m_i/n_i$	逻辑变换 $p_i' = \log(\frac{p_i}{1-p_i})$	权重 $\omega_i = n_i p_i(1-p_i)$
3	3.5	58	26	0.448 276	-0.207 64	14.345
4	4.5	52	22	0.423 077	-0.310 15	12.692
5	5.5	43	20	0.465 116	-0.139 76	10.698
6	6.5	39	22	0.564 103	0.257 829	9.590
7	7.5	28	16	0.571 429	0.287 682	6.857
8	8.5	21	12	0.571 429	0.287 682	5.143
9	9.5	15	10	0.666 667	0.693 147	3.333

2.3 Logistic 回归的基本理论和方法

2.3.1分组数据的Logistic 回归模型

◆ Logistic回归方程为：

$$p_i = \frac{exp(\beta_0 + \beta_1 x_i)}{1 + exp(\beta_0 + \beta_1 x_i)}, \quad i = 1, 2, \cdots, n$$

将以上回归方程做线性变换，令

$$p_i' = \ln\left(\frac{p_i}{1 - p_i}\right)$$

变换后的线性回归模型为:

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

2.3 Logistic 回归的基本理论和方法

2.3.1分组数据的Logistic 回归模型

◆ 对表9-5中的数据，算出经验回归方程为:$\hat{p}' = -0.886 + 0.156x$
判定系数$r^2 = 0.9242$，显著性检验P值约等于0，高度显著。
Logistic回归方程为：

$$\hat{p} = \frac{exp(-0.886 + 0.156x)}{1 + exp(-0.886 + 0.156x)}$$

利用上式对购房比例做预测。例如对$x_0 = 8$，则有：

$$\hat{p} = \frac{exp(-0.886 + 0.156\times 8)}{1 + exp(-0.886 + 0.156\times 8)}$$
$$= \frac{1.436}{1+1.436} = 0.590$$

➢ 这表明, 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59%。或者说, 一个签订初步购房意向书的年收入8 万元的家庭, 其购房概率为59%。

2.3 Logistic 回归的基本理论和方法

2.3.1分组数据的Logistic 回归模型

◆ 问题：异方差性并没有解决  加权最小二乘估计

当n_i较大时，p_i'的近似方差为：

$$D(p'_i) \approx \frac{1}{n_i \pi_i (1-\pi_i)}$$

其中 $\pi_i = E(y_i)$，因而选取权数：

$$\omega_i = n_i p_i (1-p_i)$$

2.3 Logistic 回归的基本理论和方法

2.3.1分组数据的Logistic 回归模型

- 对例9-2重新用加权最小二乘估计。具体R代码如下

```
1.> rm(list=ls())
2.> ex9.2 <- read.table("E:例9-2.txt", head=TRUE, fileEncoding="utf8")
3.>#未加权
4.> fit <- lm(ex9.2$逻辑变换~ex9.2$年家庭收入)
5.> fit$coefficients
6.         (Intercept) ex9.2$年家庭收入
7.       -0.8862679              0.1557968
8.>#加权
9.> fit_wi <- lm(ex9.2$逻辑变换~ex9.2$年家庭收入, weights=ex9.2$权重)
10.> summary(fit_wi)
11.Coefficients:
12.                          Estimate    Std. Error      t value          Pr(>|t|)
13.(Intercept)              -0.84887        0.11358     -7.474     0.000140 ***
14.ex9.2$年家庭收入    0.14932        0.02071       7.210     0.000176 ***
15.---
16.Signif. codes:    0    ‘***’    0.001    ‘**’    0.01    ‘*’    0.05    ‘.’    0.1    ‘ ’    1
17.
18.Residual standard error: 0.3862 on 7 degrees of freedom
19.Multiple R-squared:    0.8813,        Adjusted R-squared:    0.8644
20.F-statistic: 51.98 on 1 and 7 DF,    p-value: 0.0001759
```

2.3 Logistic 回归的基本理论和方法

2.3.1分组数据的Logistic 回归模型

◆ 用加权最小二程方法得到的Logistic回归方程为：

$$\hat{p} = \frac{exp(-0.849 + 0.149x)}{1 + exp(-0.849 + 0.149x)}$$

利用上式对购房比例做预测。例如对$x_0 = 8$，则有：

$$\hat{p} = \frac{exp(-0.849 + 0.149\times 8)}{1 + exp(-0.849 + 0.149\times 8)}$$
$$= \frac{1.409}{1+1.409} = 0.585$$

所以, 年收入8 万元的家庭预计实际购房比例为58.5%, 这个结果与未加权的结果很接近。

2.3 Logistic 回归的基本理论和方法

2.3.2未分组数据的Logistic 回归模型

◆ 设y是0-1型变量，$x_1, x_2, \cdots, x_p$是与y相关的确定性变量，n组观测数据为$(x_{i1}, x_{i2}, \cdots, x_{ip}; y_i)(i = 1, 2, \cdots, n)$,其中，$y_1, y_2, \cdots, y_n$是取值为0或1的随机变量，$y_i$与$x_{i1}, x_{i2}, \cdots, x_{ip}$ 的关系为：

$$E(y_i) = \pi_i = f(\beta_0 + \beta_1 x_{i1} + \cdots \beta_p x_{ip})$$

其中函数$f(x)$是值域在[0,1]区间内的单调增函数，对于Logistic回归，有

$$f(x) = \frac{e^x}{1 + e^x}$$

于是y_i遵从均值为$\pi_i = f(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})$的0-1型分布，概率函数为：

$$P(y_i = 1) = \pi_i$$
$$P(y_i = 0) = 1 - \pi_i$$

2.3 Logistic 回归的基本理论和方法

2.3.2未分组数据的Logistic 回归模型

可以把y_i的概率函数合写为：

$$P(y_i) = \pi_i(1-\pi_i)^{1-y_i}, \mathrm{y_i} = 0, 1; \mathrm{i} = 1, 2, , \cdots, n$$

于是$y_1, y_2, \cdots, y_n$的似然函数为：

$$L = \prod_{i=1}^{n} P(y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$

对似然函数取自然对数，得

$$ln\, L = \sum_{i=1}^{n} [y_i ln\pi_i + (1-y_i) ln(1-\pi_i)]$$

$$= \sum_{i=1}^{n} [y_i ln \frac{\pi_i}{1-\pi_i} + ln(1-\pi_i)]$$

2.3 Logistic 回归的基本理论和方法

2.3.2未分组数据的Logistic 回归模型

◆ 对于Logistic回归，将

$$\pi_i = \frac{exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}$$

代入, 得

$$ln\, L = \sum_{i=1}^{n} \{y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) - ln[1 + exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]\}$$

极大似然估计就是选取$\beta_0, \beta_1, \cdots, \beta_p$的估计值，使上式达到极大。

2.3 Logistic 回归的基本理论和方法

2.3.2未分组数据的Logistic 回归模型

【例9-3】在一次关于公共交通的社会调查中，一个调查项目为“是乘坐公交车还是骑自行车上下班”。因变量$y=1$表示主要乘坐公交车上下班，$y=0$表示主要骑自行车上下班。自变量x_1是年龄，作为连续型变量; x_2是月收入(元)；x_3是性别，$x_3=1$表示男性，$x_3=0$表示女性。调查对象为工薪族群体，数据见表9-6，试建立y与自变量间的Logistic回归。

2.3 Logistic 回归的基本理论和方法

2.3.2未分组数据的Logistic 回归模型

表 9-6

序号	性别	年龄(岁)	月收入(元)	y	序号	性别	年龄(岁)	月收入(元)	y
1	0	18	850	0	15	1	20	1 000	0
2	0	21	1 200	0	16	1	25	1 200	0
3	0	23	850	1	17	1	27	1 300	0
4	0	23	950	1	18	1	28	1 500	0
5	0	28	1 200	1	19	1	30	950	1
6	0	31	850	0	20	1	32	1 000	0
7	0	36	1 500	1	21	1	33	1 800	0
8	0	42	1 000	1	22	1	33	1 000	0
9	0	46	950	1	23	1	38	1 200	0
10	0	48	1 200	0	24	1	41	1 500	0
11	0	55	1 800	1	25	1	45	1 800	1
12	0	56	2 100	1	26	1	48	1 000	0
13	0	58	1 800	1	27	1	52	1 500	1
14	1	18	850	0	28	1	56	1 800	1

2.3 Logistic 回归的基本理论和方法

2.3.2未分组数据的Logistic 回归模型

◆ 使用glm函数进行Logsitic回归：

```
1.> rm(list=ls())
2.> ex9.3 <- read.table("例9-3.txt", head=TRUE, fileEncoding="utf8")
3.> head(ex9.3)
4.序号 性别 年龄 月收入 y
5.1     1     0    18     850 0
6.2     2     0    21    1200 0
7.3     3     0    23     850 1
8.4     4     0    23     950 1
9.5     5     0    28    1200 1
10.6     6     0    31     850 0
11.> fit9.3_full <- glm(y~性别+年龄+月收入, binomial(link = "logit"), data=ex9.3)
12.> summary(fit9.3_full)
13.Coefficients:
14.                Estimate    Std. Error   z value    Pr(>|z|)
15.(Intercept)    -3.655016    2.091218    -1.748     0.0805.
16.性别           -2.501844    1.157815    -2.161     0.0307*
17.年龄            0.082168    0.052119     1.577     0.1149
18.月收入          0.001517    0.001865     0.813     0.4160
```

2.3 Logistic 回归的基本理论和方法

2.3.2未分组数据的Logistic 回归模型

- 从输出结果可以看到, 月收入不显著, 决定将其剔除。 用y 对性别与年龄两个自变量做回归。

```
1.> fit9.3<- glm(y~性别+年龄, binomial(link = "logit"), data=ex9.3)
2.> summary(fit9.3)
3.Coefficients:
```

4.	Estimate	Std. Error	z value	Pr(>|z|)
5.(Intercept)	-2.6285	1.5537	-1.692	0.0907.
6.性别	-2.2239	1.0476	-2.123	0.0338*
7.年龄	0.1023	0.0458	2.233	0.0256*

可以看到,性别(SEX)、年龄(AGE)两个自变量都是显著的,因而最终的回归方程为：

$$\hat{p}_i = \frac{exp(-2.6285 - 2.2239SEX + 0.1023AGE)}{1 + exp(-2.6285 - 2.2239SEX + 0.1023AGE)}$$

- 以上方程式表明, 女性乘公交车的比例高于男性, 年龄越大, 乘公交车的比例也越高。