第五章　分类数据的关联分析

分类变量与变量之间的关系是统计结构中的重要参数,其中变量的数据类型常常是以计数数据的方式呈现. 本章主要分成三个部分，第一部分主要是分类变量独立性检验，包括χ^2独立性检验和Fisher独立性检验、齐性检验和McNemar 检验。第二部分是变量关联分析的扩展，主要介绍了分层Mantel-Haenszel 检验和关联规则，第三部分是Ridit 检验法和对数线性模型.

§5.1　$r \times s$列联表和χ^2独立性检验

假设有n个随机试验的结果按照两个变量A和B分类, A取值为A_1, A_2, $\cdots$, A_r, B取值为B_1, B_2, $\cdots$, B_s. 将变量A和B的各种情况的组合用一张$r \times s$列联表表示, 称$r \times s$二维列联表, 如表5.1所示. 其中n_{ij}表示A取A_i及B取B_j的频数, $\sum\limits_{i=1}^{r}\sum\limits_{j=1}^{s} n_{ij} = n$, 其中：

$$n_{i\cdot} = \sum_{j=1}^{s} n_{ij}, i = 1, 2, \cdots, r, \quad \text{表示各行之和;}$$

$$n_{\cdot j} = \sum_{i=1}^{r} n_{ij}, j = 1, 2, \cdots, s, \quad \text{表示各列之和;}$$

$$n_{\cdot\cdot} = \sum_{j=1}^{s} n_{\cdot j} = \sum_{i=1}^{r} n_{i\cdot}$$

表5.1　$r \times s$ 二维列联表

	B_1	B_2	$\cdots$	B_s	总和
A_1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r\cdot}$
总和	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot s}$	$n_{\cdot\cdot}$

令$p_{ij} = P(A = A_i, B = B_j), i = 1, 2, \cdots, r; j = 1, 2, \cdots, s.$ $p_{i\cdot}$和$p_{\cdot j}$分别表示A和B的边缘概率. 对于二维$r \times s$列联表, 如果变量A和B独立, 或说没有关联, 则A和B的联合概率应等于A和B的边缘概率之积.

于是分类变量独立性的问题可以描述为以下假设检验问题：

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}, \ 1 \leqslant i \leqslant r; 1 \leqslant j \leqslant s.$$

我们注意到如果两个变量之间没有关系, 那么观测频数与期望频数之间的总体差异应该很小. 反之, 如果观测频数与期望频数之间的差异足够大, 那么就可以推断两个变量之间存在相互依赖关系. 在零假设下, $r \times s$列联表每格中期望值为

$$m_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n_{\cdot\cdot}},$$

则可以定义统计量

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}. \tag{5.1}$$

如果有$m_{ij} > 5$, 则χ^2近似服从自由度为$(s-1)(r-1)$的卡方分布. 如果Pearson χ^2值过大, 或p值很小, 则拒绝零假设, 认为行变量与列变量存在关联. 像这样没有指出两变量之间更细微的相关或其他特殊的关系, 称为一般性关联(general association) .

例5.1 为研究血型与肝病之间的关系, 对295名肝病患者及638名非肝病患者(对照组)调查不同血型的得病情况, 如表5.2所示, 问血型与肝病之间是否存在关联?

表5.2 血型与肝病间的关系

血型	肝炎	肝硬化	对照	合计
O	98	38	289	425
A	67	41	262	370
B	13	8	57	78
AB	18	12	30	60
合计	196	99	638	933

本例中的行和列都是分类变量, 因而可用chisq.test求出Pearsonχ^2值, 如下所示:

```
> blood <-read.table("bloodtyp.txt",header=T)
> chisq.test(blood)
    Pearson's chi-square test with Yates' continuity correction
  data:  blood
  X-square = 15.073, df = 6, p-value = 0.020
```

表中输出了Pearsonχ^2检验结果, 自由度为$(3-1)(4-1) = 6$, χ^2值为15.073, p值为0.020. 由于p值小于0.05, 可以拒绝血型与病种独立的假设, 认为血型与肝病有一定关联.

为达到χ^2检验的效果, 一般需要保证在应用χ^2检验时满足一些特殊的假定条件. 具体而言, 要测量不同类之间是否独立, 频数过小的格点不能太多. 比如, Siegel和Castellan(1988) 指出行数或列数至少其一超过2, 单元格中期望频数低于5的单元格的数目不能超过总单元格个数的20%, 不能允许存在单元格中的期望频数小于1.

当实际观测次数过少时, Pearson卡方检验会有很大偏差, Wilk(1995)建议改用有偏的卡方值公式G^2 :

$$\begin{aligned} G^2 &= -2\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}\ln(n_{ij}/m_{ij}) \\ &= -2\left[\sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}\ln(n_{ij}) - \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}\ln(m_{ij})\right]. \end{aligned}$$

G^2称为似然比卡方值(likelihood ratio chi-square).G^2在零假设下与Pearsonχ^2统计量分布相同, 近似服从自由度为$(s-1)(r-1)$的卡方分布. 如果G^2值过大, 或零假设下p值很小, 则拒绝零假设, 认为行变量与列变量存在强关联.

§5.2 χ^2齐性检验

一般关系说明行与列向量有一定关系,如不同血型的病人患某种疾病较多或较少. 由于行和列的变量都是无序的, 因而它的结果与各行或各列的顺序无关.另外一类问题是行表示不同的区组, 列表示我们感兴趣的问题,我们希望回答列变量比例分布在各个区组之间是否一致,这类检验问题称为齐性检验. 先看下面的例题.

例5.2 简·奥斯汀(1775—1817)是英国著名女作家, 在其短暂的一生中为世界文坛奉献出许多经久不衰的作品, 如《理智与情感》(1811)、《傲慢与偏见》(1813)、《曼斯菲尔德花园》(1814)、《爱玛》(1815)等. 在其身后, 奥斯汀的哥哥亨利主持了遗作《劝导》和《诺桑觉寺》两部作品的出版, 很多热爱奥斯汀的文学爱好者自发研究后面两部作品与奥斯汀本人的语言风格是否一致. 以下是一个例子, 表5.3中收集了代表作《理智与情感》、《爱玛》以及遗作《劝导》前两章(分别以Ⅰ,Ⅱ标记)中常用代表词的出现频数, 希望研究不同作品之间在选择常用词汇的比例上是否存在差异, 并借此为作品真迹鉴别提供证据.

表5.3 不同作品中选词频率统计表

单词	理智与情感	爱玛	劝导 I	劝导 II
a	147	186	101	83
an	25	26	11	29
this	32	39	15	15
that	94	105	37	22
with	59	74	28	43
without	18	10	10	4

齐性检验问题的一般表述为

$$\forall\ i=1,2,\cdots,r, H_0: p_{i1}=\cdots=p_{is}=p_{i\cdot} \leftrightarrow\ H_1: \text{等式不全成立}. \tag{5.2}$$

本例中, p_{ij}是第i个词条在第j部著作中出现的概率, 由节选章节出现该词条的频率估计. 在原假设下, 这些概率应视为与不同著作无关, 因此n_{ij}的期望值为$e_{ij}=n_{\cdot j}p_{i\cdot}$, $p_{i\cdot}$用其零假设下的估计值$\hat{p}_{i\cdot}=n_{i\cdot}/n_{\cdot\cdot}$代替. 这时的观测值为$n_{ij}$, 而期望值为$e_{ij}\equiv\dfrac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$, 于是构造$\chi^2$ 检验统计量反应观测数和期望数的差异为

$$Q=\sum_{ij}\frac{(n_{ij}-e_{ij})^2}{e_{ij}}=\sum_{i,j}\frac{n_{ij}^2}{e_{ij}}-n_{\cdot\cdot}.$$

该χ^2统计量和独立性检验的统计量形式上完全一致, 近似服从自由度为$(r-1)(s-1)$的χ^2分布. 以下是示例程序:

```
Jane=matrix(c(147,186,101,83,25,26,11,29,32,
+39,15,15,94,105,37,22,59,74,28,43,18,10,10,4),byrow=T,,4)
chisq.test(Jane)
Pearson's Chi-squared test
data:  Jane
X-squared = 45.5775, df = 15, p-value = 6.205e-05
```

该例子的$Q=45.58$, p值为6.205×10^{-5}, 于是拒绝零假设, 认为后两部作品未必全部为简·奥斯汀的真迹.

§5.3 Fisher精确性检验

Pearsonχ^2检验要求2维列联表中只允许20%以下格子的期望数小于5. 对于2×2列联表, 如果2×2列联表中有一个格(对$r\times s$列联表实际上是25%以上的格子)期望数小于5, 则R程序会输出警告提示, 此时应当用Fisher精确检验法(Fisher's exact test 或Fisher-Irwin test及Fisher-Yates test; Fisher, 1935a,b; Yates, 1934)。下面我们仅以2×2列联表为例, 介绍Fisher 检验. 假设有2×2列联表如表5.4所示.

表5.4 典型的2 × 2 列联表

	B_1	B_2	总和
A_1	n_{11}	n_{12}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$n_{2\cdot}$
总和	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

如果固定行和和列和, 那么在零假设条件下出现在四格表中的各数值分别为n_{11}, n_{12}, n_{21} 及n_{22}, 假设边缘频数$n_{1\cdot}, n_{2\cdot}, n_{\cdot 1}, n_{\cdot 2}$ 和$n_{\cdot\cdot}$ 都是固定的. 在A和B独立的零假设下, 对任意的i, j, n_{ij}服从超几何分布为

$$P\{n_{ij}\} = \frac{n_{1\cdot}!n_{2\cdot}!n_{\cdot 1}!n_{\cdot 2}!}{n_{\cdot\cdot}!n_{11}!n_{12}!n_{21}!n_{22}!}. \tag{5.3}$$

由于4个格点中只要有一个数值确定, 另外3个也确定了, 因此只要对n_{11}的分布进行分析就足够了. 下面举例说明n_{11}的分布.

比如行总数为5, 3, 列总数为5, 3 时, 所有可能的表为四种,如下所示：

$$\begin{array}{cccccccc} 2 & 3 & \quad & 3 & 2 & \quad & 4 & 1 & \quad & 5 & 0 \\ 3 & 0 & & 2 & 1 & & 1 & 2 & & 0 & 3 \end{array}$$

n_{11}所有的可能取值为2,3,4,5. 但是在独立或没有齐性的零假设下, 出现这些值的可能性是不同的. 第二个较最后一个表更像是独立或没有齐性的情况, 因此$P(n_{11} = 3) > P(n_{11} = 5)$,用上面的公式也容易计算出$n_{11}$ 取这些值的概率为

表5.5n_{11}取值的分布列

2	3	4	5
0.1785714	0.5357143	0.2678571	0.01785714

当然, n_{11}取各种可能值的概率之和为1. 由此很容易得到各种有关的概率, 比如

$$P(n_{11} \leqslant 3) = P(n_{11} = 2) + P(n_{11} = 3) = 0.1785714 + 0.5357143 = 0.7142857.$$

在原假设下(齐性或独立性), n_{ij}的各种取值都不会是小概率事件, 如果n_{11} 过大或过小都可能导致拒绝零假设, 由此可以进行各种检验.

由式(5.3)可得

$$E(n_{11}) = \frac{n_{\cdot 1}n_{1\cdot}}{n_{\cdot 1} + n_{\cdot 2}}, \tag{5.4}$$

$$\mathrm{var}(n_{11}) = \frac{n_{\cdot 1}n_{1\cdot}n_{2\cdot}n_{\cdot 2}}{n_{\cdot\cdot}^2(n_{\cdot\cdot} - 1)}. \tag{5.5}$$

在大样本情况下, 在原假设下, n_{11}近似服从正态分布. 将n_{11}标准化为

$$Z=\frac{\sqrt{n_{..}}(n_{11}n_{22}-n_{12}n_{21})}{\sqrt{n_{1\cdot}n_{2\cdot}n_{\cdot 1}n_{\cdot 2}}}\xrightarrow{\mathcal{L}}N(0,1).$$

我们注意到分子正好是2×2列联表所对应方阵的行列式. 行列式越大表示行列关系越强, 行列式接近零表示方阵降秩, 这正是两变量独立的典型特征.

例5.3 为了解某种药物的治疗效果, 采集药物A与B的疗效数据整理成二维列联表, 如表5.6所示.

表5.6 某病两种药物治疗结果

	疗 效		
药 物	有 效	无 效	合 计
A	8	2	10
B	7	23	30
合计	15	25	40

解 在这个问题中, 某些类别的例数较少, 因而一般的χ^2检验不适用, 只能采用精确检验法.

统计计算：如果固定边缘值(15,25,10,30), 那么在零假设条件下出现在四格表中各数值分别为n_{11},n_{12},n_{21}及n_{22} 的概率按超几何分布为

$$\begin{aligned}P\{n_{11}=8\}&=\frac{n_{1\cdot}!n_{2\cdot}!n_{\cdot 1}!n_{\cdot 2}!}{n_{..}!n_{11}!n_{12}!n_{21}!n_{22}!}\\&=\frac{15!25!10!30!}{40!8!2!7!23!}=0.0023.\end{aligned}\tag{5.6}$$

如果用fisher.test函数可以计算得到$P(n_{11}\geqslant 8)=0.0024$. 作为比较, 我们还用了$\chi^2$检验, 此时Pearson统计量为2.6921, p值为0.1008, 程序和相应的输出如下所示：

```
> fisher.test(medicine)
  Fisher's Exact Test for Count Data
  data:  medicine
  p-value = 0.002429
  alternative hypothesis: true odds ratio is not equal to 1

> chisq.test(medicine)
  Pearson's Chi-squared test with Yates' continuity correction
```

```
data:  medicine
X-squared = 8, df = 1, p-value = 0.004678
Warning message:
Chi-squared asymptotic algorithm may not be correct in: chisq.test(medicine)
```

在上面的程序中, 进行χ^2检验时出现了警告信息,另外也发现格点中数据量较少的时候, 用χ^2 检验近似得到的p 值与Fisher 精确检验的p值相差较大.

1951年Freeman Halton将2×2的情形推广到$r\times s$的情形,此时假设X 变量取值为$j=1,\cdots,r$, Y变量取值为$j=1,\cdots,s, r,s>2$, 有如下列联表:

表5.7 $r\times s$ 二维列联表

	1	2	$\cdots$	s	总和
1	n_{11}	n_{12}	$\cdots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	$n_{r\cdot}$
总和	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot s}$	$n_{\cdot\cdot}$

各交叉处数值的联合分布服从多元超几何分布(multivariate hypergeometric distribution). 那么有$p-$值$=\frac{\prod_i n_{i+}!\prod_i n_{+j}!}{n!\prod_{ij} n_{ij}!}$

例5.4 猩红热是一种儿童急症，常伴随并发3种疾病：急性鼻窦炎、咽部炎症和急性中耳炎，出现6种症状。表5.8 统计了24 位收治病人，分别被病人主诉为6种症状之一，确诊后分别为3种并发症如行显示，标记为1：急性鼻窦炎；2：咽部炎症和3：急性中耳炎。6种症状如列显示，标记为1：剧烈头痛；2：流大量脓涕；3：鼻塞；4：嗅觉减退；5：咽部疼痛；6：扁桃体红肿。分布情况如表所示，分析两者之间的关系如何？

表5.8 $r\times s$ 二维列联表

	1	2	3	4	5	6	总和
1	1	1	0	1	8	0	11
2	0	1	1	1	0	1	4
3	1	0	0	0	7	1	9
总和	2	2	1	2	15	2	$n_{\cdot\cdot}=24$

解5.4 根据公式可以计算出$p-$值是$5.7689E-05$,可以拒绝原假设，认为主诉症状和三种疾病之间有紧密的关系，其中猩红热主要并发病症为急性鼻窦炎和急性中耳炎，咽部疼痛是二者的主诉症状。

§5.4 McNemar检验

McNemar检验,中文译名为麦克尼马尔检验(McNemar Test)。用于配对计数数据的分析，主要分析配对数据中控制组和处理组的频率或比率是否有差异，对于比较同一批观测对象用药前后或实验前后的结果有无差异时非常有效。配对数据中控制组和处理组均为0/1 数据，如“是”或“否”,“阳性”或“阴性”,“有反应”或“无反应”,“有效”或“无效”等。该检验只适用于二分变量，对于非二分变量，应在分析前进行数据变换。

假设配对样本有两个测量$X = 1/0$和$Y = 1/0$，X和Y一共有四种结果分别为：(0,0),(0,1),(1,0) 和(1,1)，每一类的概率用p_{ij}表示,$i = 0, 1, j = 0, 1$ ，麦克尼马尔检验问题为

$$H_0: \ p_{01} - p_{10} = 0 \leftrightarrow H_A: \ p_{01} - p_{10} \neq 0$$

;

有如下四格列联表：

表5.9 典型的2 × 2 列联表

	0	1	总和
0	n_{00}	n_{01}	$n_{1\cdot}$
1	n_{10}	n_{11}	$n_{2\cdot}$
总和	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot\cdot}$

$p_{10} - p_{01}$的估计是$\hat{p}_{10} - \hat{p}_{01} = n_{01}/n - n_{10}/n_{\cdot\cdot}$. 这是两个比例之差，它的标准差是

$$\mathrm{SE}(\hat{p}_{10} - \hat{p}_{01}) = \sqrt{\frac{\hat{p}_{10} + \hat{p}_{01} - (\hat{p}_{10} - \hat{p}_{01})^2}{n_{\cdot\cdot}}}.$$

可以使用Wald统计量，它是用一个正态z得分和其标准差相除得到的比率，这里用这个比率的平方来产生一个度量差异的得分，得到如下χ^2 检验统计量：

$$\chi^2 = \frac{(n_{01} - n_{10})^2}{n_{01} + n_{10}}$$

在H_0检验下，该统计量服从$\chi^2(1)$的分布.可以在$\chi^2(1)$分布下根据χ^2值过大来拒绝原假设。

例5.5 有131份血清样品，每份样品分别进行两种血清学检验A和B，问两种方法阳性检出率是否不同，数据如表5.10 所示：

表5.10 A和B两种检验结果

	A 方法		
B 方法	1	0	合 计
1	80	10	90
0	31	10	41
合计	111	20	131

解5.5 计算$\chi^2 = \frac{(10-31)^2}{10+31} = 10.76$,自由度为1，$p-$值是0.0018，该数据表如果做$\chi^2$拟合优度检验，则发现$p-$ 值为0.0896，无法发现A 和B在不一致上的关联性。

McNemar检验主要利用了非主对角线单元格上的信息，它关注的是行变量和列变量两者之间不一致的评价信息，用于比较两个评价者间各自存在怎样的倾向性。如果对于一致性较好的大样本数据，McNemar检验也可能会失效。例如对10,000例数据进行一致性评价，假设其中9,993例都是完全一致的，一致的评价信息集中在主对角线上。不一致的评价信息共计7例，3例位于左下和4 例位于右上区。显然，此时一致性相当的好。但如果使用McNemar 检验，反而会得出两种评价有差异的结论。

§5.5 Mantel-Haenszel检验

很多研究都涉及分层数据结构, 比如产品研究中,需要根据城市和农村特点分别研究不同人群对产品或服务的满意程度; 不同类型的医院由于收治的病人特征不同,要对不同的医院研究不同治疗方案对病人的恢复效果. 这里城市和农村是问题的两个层, 研究所涉及的不同医院也是不同的层.于是在回答处理与反应结果之间是否独立的问题时,需要首先按层计算差异, 再将各层的差异进行综合比较,从而做出综合的判断.一个较为简单的情况是每层都有一个2×2列联表,于是多个层涉及多个2×2列联表. 例如在3个中心临床试验中,每个医院随机地把病人分为试验组和对照组, 疗效分为有效和无效,每个医院形成一个2×2表数据.

以医院为例, 令分层结构$h = 1, 2, \cdots, k$,n_{hij}表示第h层四格列联表观测频数,h表示多层四格表的第h 层,第h 层观测病案数为n_h, $\sum_{h=1}^{k} n_h = n$.

假设检验问题为

H_0：试验组与对照组在治疗效果上没有差异;

H_1：试验组与对照组在治疗效果上存在差异.

下表是第h层四格表的记号表示.

表5.11 第h层四格列联表各单元格记号

	有效	无效	合计
试验组	n_{h11}	n_{h12}	$n_{h1.}$
对照组	n_{h21}	n_{h22}	$n_{h2.}$
合计	$n_{h.1}$	$n_{h.2}$	n_h

当零假设H_0成立时, 先求出第h层n_{h11}的期望En_{h11}和方差var(n_{h11}):

$$En_{h11}=\frac{n_{h1.}n_{h.1}}{n_h},$$
$$\text{var}\,(n_{h11})=\frac{n_{h1.}n_{h2.}n_{h.1}n_{h.2}}{{n_h}^2\,(n_h-1)}.$$

不同组与疗效之间的关系可用Mantel-Haenszel 1959年提出的Q_{MH}统计量表示

$$Q_{MH}=\frac{\left(\sum_{h=1}^{k}n_{h11}-\sum_{h=1}^{k}En_{h11}\right)^2}{\sum_{h=1}^{k}\text{var}\,(n_{h11})}.$$

式中, k为层数.

定理5.1 $\forall h=1,2,\cdots,k$层, $\forall i=1,2$ 行,$n_{hi.}=\sum_{j=1}^{2}n_{hij}$ 不小于30 时, 统计量Q_{MH}近似服从自由度等于1的卡方分布.

例5.6 对2家医院考察某治癌药的治癌效果,试验组(A)与对照组(B)(安慰剂)对比记录其疗效, 如表5.12所示.

表5.12 不同医院治癌药治癌效果比较

医院	药品	有效	无效	合计
1	A	50	15	65
	B	92	90	182
	合计	142	105	247
医院	药品	有效	无效	合计
2	A	47	135	182
	B	5	60	65
	合计	52	195	247

解 列R程序如下:

```
HA=matrix(c(50,92,15,90),2)
```

```
HB=matrix(c(47,5,135,60),2)
m=c(HA,HB); x=array(m,c(2,2,2))
mantelhaen.test(x)
```

```
Mantel-Haenszel chi-squared test with continuity correction
data:  x Mantel-Haenszel X-squared = 21.9443, df = 1, p-value =
2.807e-06 alternative hypothesis: true common odds ratio is not
equal to 1 95 percent confidence interval:
2.080167 6.099585
sample estimates: common odds ratio
 3.562044
```

以上得到Mantel-Haenszel检验的结果$Q_{MH}=21.9443, p$值为2.807×10^{-6}, 通过检验, 说明治癌药有效果. 进一步比较各层, 发现在第一家医院, 药品A相对于安慰剂疗效显著; 在第二家医院, 无论是药品A还是药品B, 疗效都倾向于不明显.

进一步计算发现, 如果不按分层结构计算分类变量的关系, 则只能出现两分类变量无关的结论, 请见习题5.6.

Mantel-Haenszel方法消除了层次因素的干扰而提高了检验出变量关联性的可靠性.

§5.6 关联规则

前面几节中, 我们给出了两个分类变量的关系度量和检验方法, 这些方法都是针对两个固定变量进行的测量. 实际中, 常常会碰到大规模变量的选择问题. 比如, 超市的购物篮数据中, 哪些物品在选购时相比另一些物品而言, 更倾向于同时被选中, 这是消费者购买行为分析中的核心问题. 比如, 购买面包和牛奶的人, 是否更倾向于购买牛肉汉堡和番茄酱. 如何从为数众多的变量中用最快的方法将关联性最强的两组或更多组变量选出来, 是值得关注的一个技术问题. 该问题自然引发了大规模数据探索分析中的核心技术问题, 即关联规则的有效取得.

§5.6.1 关联规则基本概念

给定一个事务数据表D, 设有m个待研究的不同变量的取值构成有限项集$I=\{i_1,i_2,\cdots,i_m\}$, 其中每一条记录T是I中k项组成的集合, 称为k **项集**, 即$T\subseteq I$, 如果对于I的子集X, 有$X\subseteq T$, 则称该交易T包含X. 一条**关联规则**是一个形如$X\to Y$的形式, 其中$X\subseteq I$, $Y\subseteq I$, 且$X\bigcap Y=\varnothing$. X称关联规则的前项, Y称关联规则的后

项. 我们关注的是两组变量对应的项集X和项集Y之间因果依存的可能性. 关联规则中常涉及两个基本的度量：支持度和可信度.

关联规则的**支持度**S定义为X与Y同时出现在一次事务中的可能性, 由X项和Y项在D中同时出现的事务数占总事务的比例估计, 反映X与Y同时出现的可能性, 即

$$S(X \Rightarrow Y) = |T(X \vee Y)|/|T|.$$

其中, $|T(X \vee Y)|$表示同时包含X和Y的事务数,$|T|$表示总事务数. 关联规则的支持度(support)用于测度关联规则在数据库中的普适程度, 是对关联规则重要性(或适用性)的衡量. 如果支持度高, 表示规则具有较好的代表性.

关联规则的**可信度**(confidence)用于测度后项对前项的依赖程度, 定义为：在出现项目X的事务中出现项目Y的比例, 即

$$C(X \Rightarrow Y) = |T(X \vee Y)|/|T(X)|.$$

其中, $|T(X)|$表示包含X的事务数, $|T(X \vee Y)|$表示同时出现X和Y的事务数. 可信度高说明X发生引起Y发生的可能性高. 可信度是一个相对指标, 是对关联规则准确度的衡量, 可信度高, 表示规则Y依赖于X的可能性比较高.

关联规则的支持度和可信度都是位于$0 \sim 100\%$之间的数.关联规则的主要目的是建立变量值之间的可信度和支持度都比较高的关联规则.最常见的关联规则是最小支持度-可信度关联规则, 即找到支持度-可信度都在给定的最小支持度和最小可信度以上的关联规则, 表示为$X \Rightarrow Y$(支持度S, 置信度C)关联规则. Apriori算法是这类关联规则的代表.

§5.6.2 Apriori算法

常用的关联规则算法有Apriori算法和CARMA算法. 其中Apriori算法是由Agrawal, Imielinski和Swami于1993年设计的对静态数据库计算关联规则的代表性算法,Apriori还是许多序列规则和分类算法的重要组成部分.而CARMA算法则是动态计算关联规则的代表.Apriori 是发现布尔关联规则所需频繁项集的基本算法,即每个变量只取1 或0.

Apriori算法主要以搜索满足最小支持度和可信度的频繁k项集为目的,频繁项集的搜索是算法的核心内容.如果k_1项集A 是k_2 项集B的子集($k_1 < k_2$),那么称B由A生成.我们知道k_1项集A的支持度不大于任何它的生成集k_2项集B. 支持度随项数增加呈递减规律,于是可以从较小的k 开始向下逐层搜索k项集,如果较低的k 项集不满足最小支持度条件,则由该k项集生成的n项集($n > k$) 都不满足最小条件, 从而可能有效地截断大项集的生长,削减非频繁项集的候选项集,有效地遍历满足条件的大项集.

具体而言, 首先从频繁1项集开始,支持度满足最小条件的项集记作L_1.从L_1寻找频繁2项集的集合L_2, 如此下去, 直到频繁k 项集为空,找每个L_k扫描一次数据库.

下表是人为编制的一个购物篮数据, 这个数据有5次购买记录,我们以此为例说明Apriori算法的原理.

表5.13 购物篮数据表

Basket-Id	A	B	C
t_1	1	0	0
t_2	0	1	0
t_3	1	1	1
t_4	1	1	0
t_5	0	1	1

在上表中, t_i表示第i笔购物交易, $A=1$表示某次交易中,用户购买了A, 显然可以将上表转化为项集形式，如表5.14 所示：

表5.14 购物篮交易数据表

Tid	items
t_1	A
t_2	B
t_3	ABC
t_4	AB
t_5	BC

预先将支持度和置信度分别设定为0.4和0.6, 执行Apriori算法如下：

(1) 扫描数据库, 搜索1项集, 从中找出频繁1项集$L_1=\{A,B,C\}$.

(2) 在频繁1项集中将任意二项组合生成候选2项集C_2, 比如,从1项集L_1可生成候选二项集$C_2=\{AB,AC,BC\}$, 扫描数据库找出频繁2项集, $L_2=\{AB,BC\}$.

(3) 从频繁2项集按照第二步的方法构成3项候选集C_3, 找出频繁3项集.因为$s(A\vee B\vee C)=20\%$, 低于设定的最小支持度,所以到第三步算法停止, $L_3=\varnothing$.

找出频繁项集之后将构造关联规则, 继续上面的例子,下面是构造出的一些规则.

规则1：支持度0.4, 可信度0.67,

$$A\Rightarrow B.$$

规则2：支持度0.4, 可信度1,

$$B \Rightarrow A.$$

规则3：支持度0.4, 可信度1,

$$C \Rightarrow B.$$

例5.7 Adult数据取自1994年美国人口调查局数据库,最初是用来预测个人年收入是否超过5万美元. 它包括age (年龄), workclass(工作类型), education(教育), race(种族),sex(性别)等15个变量, 48,842 个观测.我们对这个数据集运用Apriori 算法发现了一些有意义的规则，如表5.15 所示.下面是R软件的程序.

```
install.packages("arules")
library(arules)
library(Matrix)
library(lattice); data("Adult") ## Mine association rules
myrules=apriori(Adult, parameter
= list(supp = 0.7, conf = 0.9,target = "rules"))
write(myrules[1:10])
```

表5.15 关联规则输出结果

ID	rules	support	confidence	coverage	lift	count
1	$\{\} \to \{capital-gain=None\}$	0.917	0.917	1	1	44,807
2	$\{\} \to \{capital-loss=None\}$	0.953	0.953	1	1	46,560
3	$\{race=White\} \to \{native-country=United-States\}$	0.788	0.922	0.855	1.023	38,493
4	$\{race=White\} \to \{capital-gain=None\}$	0.782	0.914	0.855	0.997	38,184
5	$\{race=White\} \to \{capital-loss=None\}$	0.814	0.952	0.855	0.998	39,742
6	$\{native-country=United-States\} \to \{capital-gain=None\}$	0.822	0.916	0.897	0.998	40,146
7	$\{native-country=United-States\} \to \{capital-loss=None\}$	0.855	0.953	0.897	0.999	41,752
8	$\{capital-gain=None\} \to \{capital-loss=None\}$	0.871	0.949	0.917	0.996	42,525
9	$\{capital-loss=None\} \to \{capital-gain=None\}$	0.871	0.913	0.953	0.996	42,525
10	$\{race=White, native-country -United-States\} \to \{capital-gain=None\}$	0.720	0.913	0.788	0.995	35,140

值得注意的是, 并非可信度越高的规则都是有意义的. 比如, 某超市里,80%的女性(A)购买了某类商品(B)($A \to B$), 但这个商品的购买率也是80%, 也就是说, 女

性购买率和男性购买率是一样的, 即$P(B|A) = P(B|\overline{A})$, 通常这类规则实用性不大. 如果$P(B|A) > P(B)$, 则说明由$A$决定的$B$更有意义, 于是就产生了评价关联规则的第三个概念——提升度(lift). 提升度定义为

$$L(A \Rightarrow B) = \frac{C(A \Rightarrow B)}{T(B)},$$

它是关联度量$P(A,B)/(P(A)P(B))$的一个估计. 当$P(B) > \frac{1}{2}$时, 可以证明当提升度$L(A \Rightarrow B) > 1$时, 有$P(B|A) > P(B|\overline{A})$, 这表示$A \Rightarrow B$规则的集中度较好.

§5.7 Ridit检验法

实际中经常需要对某个抽象概念进行测量, 比如, 通过测量病人对几种药物治疗的反应程度, 以判断不同药物的反应程度之间是否存在差异, 如果存在差异, 这些差异的感知顺序是怎样的? 类似的问题在行为学上同样存在, 在几个不同的项目设定量表测量用户对产品或服务的满意程度, 问题是要确定不同项目用户感知差异的顺序. 这类问题的共同特征是采用量表测量受访者的感知, 由于人为和个体差异, 不一定总能理想地测量到真实的数据. 比如, 通过病人对于药物的反应程度进行药物评价或分级时可能会存在一定的级别感知缺陷. 例如, 4级痛感不能代表1级痛感的4倍; 10分钟精神忧郁感也不可认为是1分钟忧郁感的10倍; 药物使4级痛感减轻至3级不会与2级痛感减轻至1级的痛感一致. 总之, 我们只能测量到顺序级别的数据, 这些不同的项目之间不具有完整的事实独立性, 因而单纯应用定距分级或评分进行各处理强弱的比较, 数据的量关系可能与客观实际不符. 一个自然的想法是考虑将不能明显显示顺序的得分合并, 重新计算量表评级, 降低人为干扰, 从而作出更客观的评价.

Bross于1958年提出一种非参数检验Ridit分析方法. Ridit是relative to identified distribution的缩写和Unit的词尾it的组合, 有时也称为参照单位分析法. 它的基本原理是: 取一个样本数较多的组或将几组数据汇总成为参照组, 根据参照组的样本结构将原来各组响应数变换为参照得分——Ridit得分, 利用变换后的Ridit得分进行各处理之间强弱的公平比较.

1. Ridit得分及计算

考虑$r \times s$双向列联表, 如表5.16所示.

表5.16 $r \times s$ 二维列联表

	B_1	B_2	$\cdots$	B_s	总和
A_1	O_{11}	O_{12}	$\cdots$	O_{1s}	$O_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	O_{r1}	O_{r2}	$\cdots$	O_{rs}	$O_{r\cdot}$
总和	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot s}$	$O_{\cdot\cdot}$

行向量$\boldsymbol{A}$是关于不同比较组, 或不同处理的分类变量, $A_1, A_2, \cdots, A_r$表示不同的处理; 列向量$\boldsymbol{B}$是顺序尺度变量, 不失一般性, 一般假定$B_1 < B_2 < \cdots < B_s$. O_{ij}表示回答第i处理(类)在第j个顺序类上的响应数. 需要检验的问题是：$A_1, A_2, \cdots, A_r$个不同处理的强弱程度是否存在差异.

假设检验问题为

$$\begin{aligned} &H_0 : A_1, A_2, \cdots, A_r\text{之间没有强弱顺序;} \\ \leftrightarrow\ &H_1 : \text{至少存在一对}A_i, A_j\text{, 使得}A_i \neq A_j\text{成立.} \end{aligned} \tag{5.7}$$

为比较不同处理之间的强弱顺序, 回想在Kruskal-Wallis检验中,我们用每个处理的秩和或平均秩作为代表值, 参与处理之间的差异的比较.秩和或平均秩可以理解为各不同处理的综合得分,这是多总体位置比较的基础. 假定每个处理的得分分布在不同的s个顺序类上,假设v_j是第j个顺序类的得分, 那么可以如下计算第i个处理的得分：

$$\begin{aligned} R_i &= \sum_{j=1}^{s} v_j p(j|i) \\ &= \sum_{j=1}^{s} v_j \frac{p_{ij}}{p_{i\cdot}}. \end{aligned}$$

式中, $p_{i\cdot}$是第i个处理类的边缘概率, p_{ij}是第i个处理第j个顺序类的联合概率, $p(j|i)$是条件概率. 但是, 一般v_j在很多情况下很不明确, 有时为计算方便, 则以等距数据替代. 比如：在Likert5级量表中,$s = 5, v_j$ 按照$j = 1, 2, \cdots, 5$分别表示非常不重要、不重要、一般、重要和非常重要. 这些顺序常常以$1, 2, 3, 4, 5$ 表示, 1表示弱, 5表示强. 但是, 正如本节起始段落所言, 如此人为指定等距得分进行计算的结果常常与事实不符.

Ridit得分选择用累积概率得分表示各顺序真实的强弱顺序, 假设顺序类别中

第j类的边缘分布是$p_{\cdot j}, j=1,2,\cdots,s$.第$j$类的顺序强度如下定义：

$$\begin{aligned} R_1 &= \frac{1}{2}p_{\cdot 1}, \\ &\vdots \\ R_j &= \sum_{k=1}^{j-1} p_{\cdot k} + \frac{1}{2}p_{\cdot j},\ j=2,3,\cdots,s \\ &= \frac{F_{j-1}^B + F_j^B}{2}. \end{aligned}$$

其中

$$F_j^B = \sum_{k=1}^{j} p_{\cdot k},\ j=2,3,\cdots,s.$$

式中, F_j^B是B的累积概率. 从上面的定义来看, $R_1 < R_2 < \cdots < R_s$, 这符合顺序类别等级度量特征.

定理5.2 如上定义的Ridit得分, 满足如下性质：

$$R = \sum_{j=1}^{s} R_j p_{\cdot j} \equiv \frac{1}{2}.$$

如果定义

$$R_i = \sum_{j=1}^{s} R_j p(j|i), \tag{5.8}$$

则$R = \sum_{i=1}^{r} R_i p_{i\cdot} \equiv \frac{1}{2}$.

证明 为简单起见, 只证明第一个等式, 第二个等式留给读者自己证明.

$$\begin{aligned} R &= \sum_{j=1}^{s} R_j p_{\cdot j} = \sum_{j=1}^{s} \frac{F_{j-1}^B + F_j^B}{2} p_{\cdot j} \\ &= \frac{1}{2}\left(\sum_{j=1}^{s}\sum_{k=1}^{j-1} p_{\cdot j}p_{\cdot k} + \sum_{j=1}^{s}\sum_{k=1}^{j} p_{\cdot j}p_{\cdot k} \right) \\ &= \frac{1}{2}\left(2\sum_{j=1}^{s}\sum_{k=1}^{j-1} p_{\cdot j}p_{\cdot k} + \sum_{j=1}^{s} p_{\cdot j}^2 \right) \\ &= \frac{1}{2}\left(\sum_{j=1}^{s} p_{\cdot j} \right)^2 = \frac{1}{2}. \end{aligned}$$

另外, 注意到Ridit得分是用累积概率F_j^B定义的, 这正是Ridit得分法区别于人为定分的实质所在. 通常的Likert量表采用的是均匀分布, 如果各顺序类响应数均匀, 则这样假设是可能的. 但是, 如果各类响应人数不等, 则如此定级可能就不客观. 在实际计算中, F_j^B需要用样本估计, 为方便计算, 下面给出Ridit计算的步骤, 并将计算过程显示于表5.17中.

(1) 计算各顺序类别响应总数的一半$H_j = \dfrac{1}{2}O_{\cdot j}$, 得到行(1).

(2) 将行(1)右移一格, 第一格为0, 其余为累计前一级$(j-1)$的累积频率$C_j, C_j = \sum\limits_{k=1}^{j-1} O_{\cdot k}$, 得到行(2).

(3) 将行(1)与行(2)对应位置相加, 即得到行(3), 行(3)中$N_j = H_j + C_j$.

(4) 计算各顺序类别的Ridit得分$R_j = \dfrac{N_j}{O_{\cdot\cdot}}$, 得到行(4).

(5) 将R_j的值按照O_{ij}占$O_{i\cdot}$的权重重新配置第i, j位置的Ridit得分：$R_{ij} = \dfrac{O_{ij}}{O_{i\cdot}} R_{\cdot j}$.

(6) 计算第i处理(类)的Ridit得分： $R_i = \sum\limits_{j=1}^{s} R_{ij}$, 这些Ridit得分的期望为0.5.

表5.17 各顺序级别R_j计算表

步骤		B_1	B_2	$\cdots$	B_s	合计
	A_1	O_{11}	O_{12}	$\cdots$	O_{1s}	$O_{1\cdot}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	A_r	O_{r1}	O_{r2}	$\cdots$	O_{rs}	$O_{r\cdot}$
	总和	$O_{\cdot 1}$	$O_{\cdot 2}$	$\cdots$	$O_{\cdot s}$	$O_{\cdot\cdot}$
(1)		$H_1 = \dfrac{1}{2}O_{\cdot 1}$	$H_2 = \dfrac{1}{2}O_{\cdot 2}$	$H_j = \dfrac{1}{2}O_{\cdot j}$	$H_s = \dfrac{1}{2}O_{\cdot s}$	
(2)		0	$C_2 = \sum\limits_{k=1}^{1} O_{\cdot k}$	$C_j = \sum\limits_{k=1}^{j-1} O_{\cdot k}$	$C_s = \sum\limits_{k=1}^{s-1} O_{\cdot k}$	
(3)		N_1	N_2	$N_j = H_j + C_j$	N_s	
(4)		R_1	R_2	$R_j = \dfrac{N_j}{O_{\cdot\cdot}}$	R_s	

2. Ridit得分及假设检验

对假设检验问题(5.7)：

$$H_0 : A_1, A_2, \cdots, A_r\text{之间没有强弱顺序;}$$
$$\leftrightarrow H_1 : \text{至少存在一对}A_i, A_j\text{, 使得}A_i \neq A_j\text{成立.}$$

有了R_j, 如果需要比较几个处理强弱是否存在差异, 可以用Kruskal-Wallis检验方法：

$$W=\frac{12O_{\cdot\cdot}}{(O_{\cdot\cdot}+1)T}\sum_{i=1}^{r}O_{i\cdot}(R_i-0.5)^2,$$

其中, T为打结校正因子. Agresti于1984年指出当样本量足够大时, T的值趋近于1, 所以检验统计量简化为

$$W=12\sum_{i=1}^{r}O_{i\cdot}(R_i-0.5)^2.$$

当H_0成立, W近似服从自由度为$\nu=r-1$的χ^2分布, 当W过大和过小都考虑拒绝零假设.

3. 根据置信区间分组

R_i是按照公式(5.8)计算得到的, Agresti于1984年指出, R_i在大样本情况下服从正态分布, 其95%置信区间为

$$R_i\pm 1.96\hat{\sigma}_{R_i}.$$

如果希望通过置信区间来比较第i处理与参照组之间的差异, 可以用$\hat{\sigma}_{R_i}$的最大值简化上式, 即

$$\max(\hat{\sigma}_{R_i})=\frac{1}{\sqrt{12O_{i\cdot}}}.$$

取$\alpha<0.05$, 因而得到近似公式

$$\bar{R}_i\pm 1/\sqrt{3O_{i\cdot}}. \tag{5.9}$$

其中$O_{i\cdot}$为第i处理的响应数.

由置信区间与假设检验之间的关系, 可以根据参照组的平均Ridit$\bar{R}$与处理组的平均Ridit$\bar{R}_i$得分的差别来进行两两对比检验, 如果Ridit$\bar{R}$与Ridit$\bar{R}_i$ 的置信区间没有重叠, 则说明两组存在显著性差别($\alpha<0.05$).

例5.8 表5.18所示为用头针治疗瘫痪800例的疗效分析, 不同病因的疗效可以不一样, 究竟哪一种病因所引起的瘫痪用头针的治疗效果最佳, 哪些次之, 哪些最差, 是医务人员希望通过数据回答的问题.

表5.18 头针治疗瘫痪800例的疗效分析

组别	总数	基本痊愈	显效	有效	无效	恶化	死亡
1. 脑血栓形成及后遗症	539	194	134	182	28	1	0
2. 脑出血及后遗症	132	9	38	73	11	0	1
3. 脑栓塞及后遗症	59	20	13	20	6	0	0
4. 颅内损失及后遗症	54	4	12	33	5	0	0
5. 急性感染性多发性神经炎	10	4	2	3	1	0	0
6. 脊髓疾病	6	1	3	0	2	0	0
总病例数	800	232	202	311	53	1	1

解 本例中, 从治疗效果看, 各治愈数存在较大差异, 因而不易采用人为定级的方法, 可以考虑使用Ridit分析. 首先将总数800例的疗效结果作为参照组, 而以各病因组(1∼6组)的疗效结果作比较组. 参照组的Ridit得分的计算步骤如表5.19所示, 这里为书写方便采用按列计算的方式排列计算步骤, 其中最后一列表示各顺序类Ridit得分.

表5.19 头针治疗瘫痪800例疗效的Ridit计算步骤

步骤\ 级别	基本痊愈	显效	有效	无效	恶化	死亡
(Ⅰ)(病例数总计)	232	202	311	53	1	1
(Ⅱ)(病例数×1/2)	116	101	155.5	26.5	0.5	0.5
(Ⅲ) 累积	0	232	434	745	798	799
(Ⅱ)+(Ⅲ)	116	333	589.5	771.5	798.5	799.5
$R=\frac{\text{Ⅱ}+\text{Ⅲ}}{800}$	0.145	0.416	0.737	0.964	0.998	0.999
合计	33.64	84.082	229.168	51.11	0.998	0.999

从表5.19最后一行合计项总数为400, 可以证实参照组平均Ridit$\bar{R}=0.5$.

根据公式(5.9)可得出其95%置信限为$0.5\pm0.020=(0.480,0.520)$, 将表5.19的第一组即脑血栓形成及后遗症539例组的疗效结果作比较, 如表5.20所示.

表5.20 脑血栓形成及后遗症疗效结果的Ridit得分

等级	(1)	(2)	(3)
基本痊愈	194	0.145	28.130
显效	134	0.416	55.744
有效	182	0.737	134.134
无效	28	0.964	26.992
恶化	1	0.998	0.998
死亡	0	0.999	0
合计	539		222.246

其余各项Ridit得分计算类似. 得出95%可信限为$(0.431, 0.481)$, 由于$\bar{R}_1 < \bar{R}$, 可认为第1组的治疗效果对总数800例的效果来讲较好. 又由于两置信区间互不相交, 说明第1组与总数800例的疗效差别是显著的$(\alpha < 0.05)$. 用相同方法可得出第2~6组的平均Ridit及95%置信限如下：

$$
\begin{aligned}
R_2 &= 0.63 \pm 0.050 = (0.575, 0.675) \\
R_3 &= 0.49 \pm 0.075 = (0.414, 0.564) \\
R_4 &= 0.64 \pm 0.079 = (0.564, 0.721) \\
R_5 &= 0.46 \pm 0.183 = (0.276, 0.641) \\
R_6 &= 0.55 \pm 0.236 = (0.318, 0.789)
\end{aligned}
$$

Ridit分析的结果也可用图来表示, 图5.1表示了不同组Ridit值置信区间, 中横线是参照单位0.5, 第1组在中横线下方, 说明疗效较参照组(800例)好; 第3组的平均Ridit虽也在参照单位0.5的下方, 但其95%置信限与参照组相交叠, 因此差别不显著; 第2组与第4组皆在上方, 且其95% 置信限皆不与参照组相交叠, 说明疗效较差; 第5, 6组由于病例数较少. 病症的治疗情况分成3 组, 第1组最好, 第3, 5, 6组差异不大, 第2, 4组较差.

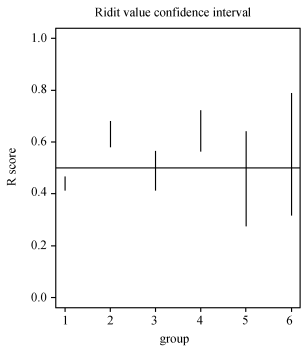

图 5.1 不同组Ridit值置信区间

如果要对各处理组(除参照组)进行比较, 可将比较的两组平均Ridit相减后再加0.5得出. 例如第1组与第4组比较为$\bar{R}_1 - \bar{R}_4 + 0.5 = 0.3$, 这表示第1组病人治疗效果差于第4组的概率为0.3, 或者第1组病人治疗效果优于第4组的概率为0.7, 即在10个病人中, 平均有7人优于第4组, 仅3个病人差于第4组.

从例子中发现, Ridit分析不仅能比较处理之间的优劣, 而且能说明优劣的程度, 这是普通的秩检验难以做到的.

§5.8 对数线性模型

由前面的章节可知, 列联表是研究分类变量独立性和依赖性的重要工具. 列联表主要采用假设检验反映事件发生的相对频率,不能反映事件的相对强度等更多或更深层的信息; 与之相比,定量数据之间的依赖关系多采用模型法, 比如线性模型,它强调参数估计和检验,但是线性模型需要研究者事先确定哪些变量是响应变量, 而哪些变量是解释变量. 但有时, 研究者无需区分响应变量和解释变量,特别对于定性数据而言, 想了解的是变量的哪些取值之间有关联、强度如何等.这就需要一个介于列联分析和线性模型之间的工具, 对数线性模型正是把列联表问题和线性模型统一起来的研究方法.与线性模型相比, 它更强调模型的拟合优度、交互效应和网格频数估计,这些信息可以更好地揭示变量之间的关系强度,也可以像模型一样预测网格点的频数.

这部分首先介绍泊松回归，接着是对数线性模型和参数估计，最后是高维对数线性模型的独立性检验。

§5.8.1 泊松回归

假设计数变量Y表示某类事件的发生频数，Y服从泊松分布，$Y=y,y=0,1,2,\cdots$发生的概率有如下表示：

$$f(y)=\frac{\mu^y e^{-\mu}}{y!},y=0,1,2,\cdots$$

这里$\mathrm{E}(Y)=\mu,\mathrm{Var}(Y)=\mu$. μ是事件的平均发生数。比如要研究一段时间内用户购买商品的件数，可以从每天顾客进店的人数开始研究。顾客光临门店可以理解为有一定的购买商品的倾向性，这里进店的顾客人数就是曝光数。购买事件发生与有多少顾客光临门店有关。如果用曝光率表示单位时间用户进店的人数，那么通过购买率可以知道，购买商品的人数服从泊松分布。

令$Y_1,\cdots,Y_N$独立同分布的随机变量，Y_i 表示在曝光数n_i 基础上的事件发生数，Y_i 的期望可以表示为：

$$\mathrm{E}(Y_i)=\mu_i=n_i\theta_i.$$

比如：Y_i表示保险公司的索赔数，Y_i由每年上保险的车辆数和索赔率两部分决定，而索赔率可能和其他的变量有关系，比如车龄和行驶的路段等等。下标i用来表示车龄和行驶路段的所产生的不同的影响。θ_i 和其他的解释变量之间的关系可以用下面的模型表达出来

$$\theta_i=e^{\mathbf{x}_i^T\beta}.$$

这就是一个一般的广义线性模型的形式

$$\mathrm{E}(Y_i)=\mu_i=n_i e^{\mathbf{x}_i^T\beta},Y_i\sim\mathrm{Pois}(\mu_i).$$

两边取对数

$$\log \mu_i = \log n_i + \mathbf{x}_i^T \beta$$

其中$\log n_i$是个常数项，而$\mathbf{x}_i$和β表达了协变量的影响模式。如果x_j是个二值变量：当这个变量= 车龄超过10 年，$x_j = 1$，如果这个变量= 车龄小于10 年，$x_j = 0$. 定义发生率（rate ratio）如下：

$$R = \frac{\mathrm{E}(Y_i|X = 1)}{\mathrm{E}(Y_i|X = 0)} = e^{\beta}.$$

模型的假设检验是:

$$\frac{\hat{\beta}_j - \beta_j}{s.e.(\hat{\beta}_j)} \sim N(0, 1).$$

拟合值如下：

$$\hat{Y}_i = \hat{\mu}_i = n_i e^{\mathbf{x}_i^T \hat{\beta}_j}, i = 1, \cdots, N.$$

$\hat{Y}_i$是$E(Y_i) = \mu_i$的估计，记作e_i,，由于$\mathrm{Var}(Y_i) = E(Y_i) = e_i, \mathrm{Sd}(Y) = \sqrt{e_i}$, 于是有皮尔逊残差如下：

$$r_i = \frac{o_i - e_i}{\sqrt{e_i}}$$

o_i是Y_i的观测频数，由残差可以引导出拟合优度检验如下：

$$\chi^2 = \sum r_i^2 = \sum \frac{(o_i - e_i)^2}{e_i}.$$

对于泊松分布而言，对数似然比偏差可以表示为

$$D = 2 \sum [o_i \log(o_i/e_i) - (o_i - e_i)].$$

由于$\sum o_i = \sum e_i$ 于是

$$D = 2 \sum [o_i \log(o_i/e_i)].$$

可以证明D 和χ^2 等价,如果定义残差偏差为:

$$d_i = \mathrm{sign}(o_i - e_i)\sqrt{2[o_i \log(o_i/e_i) - (o_i - e_i)]}, i = 1, \cdots, N$$

那么

$$D = \sum d_i^2.$$

运用Taylor展开，近似的有：

$$o \log\left(\frac{o}{e}\right) \approx (o - e) + \frac{1}{2}\frac{(o - e)^2}{e};$$

代入d_i,

$$D=\sum_{i=1}^{N}\frac{(o_i-e_i)^2}{e_i}=\chi^2.$$

例5.9 (Breslow和Day1987提供的数据,安特尼J·杜布森(Anniette J.Dobson)《广义线性模型》教材例9.2.1)英国医生的吸烟习惯和冠状动脉性猝死之间的关系：

表5.21 英国医生的吸烟习惯与冠状动脉性猝死之间的关系

年龄(Age)	吸烟者	(Smoke)	不吸烟者	(Nonsmoke)
分 组	死亡人数 (Deaths)	每年跟踪人数 (Person-Years)	死亡人数 (Deaths)	每年跟踪人数 (Person-Years)
35-44	32	52407	2	18790
45-54	104	43248	12	10673
55-64	206	28612	28	5710
65-74	186	12663	28	2585
75-84	102	5317	31	1462

关心三个问题：

1. 吸烟者的死亡率高于不吸烟者吗?
2. 如果1的结论是对的,高多少?
3. 年龄对死亡率的影响上有差异吗?

解：

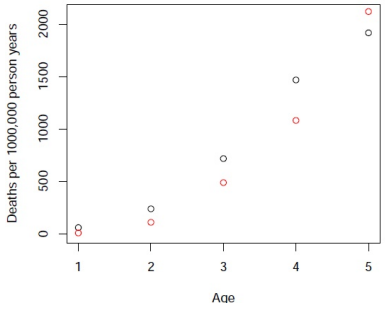

图 5.2 年龄和吸烟习惯对心脏病死亡率的影响

由图5.2可以观察到每10万人群中吸烟者和不吸烟者的死亡率。很明显，死亡率随着观察者年龄的增长而增长，吸烟人的死亡率比不吸烟者的死亡率略高，可以用模型来刻画这些因素的影响的大小如下式所示：

$$\begin{aligned}\log(\text{死亡}_i)=\log(\text{观察数}_i)+\beta_1+\beta_2\text{吸烟者}_i+\beta_3\text{年龄级别}_i\\+\beta_4\text{年龄}_i^2+\beta_5\text{年龄与吸烟的交互因子}_i\end{aligned}$$

观察数(personyears)一项是每年处于冠状动脉性猝死潜在危险中的医生数，吸烟者(smoke) 一项是=1 或0，吸烟记为1，不吸烟记为0;年龄级别(agecat) 一项取1,2,3,4,5分别对应的是年龄组(35-44),(45-54),(55-64),(65-74),(75-84)。年龄^2(agesq) 一项代表的是年龄项的平方，反映二次关系。年龄与吸烟的交互因子(smkage)对于吸烟者而言与年龄等值，对未吸烟者而言表示0, 这样设置可用于表达吸烟人群相对于未吸烟人群与年龄的关系有增加更快的效应,死亡数(deaths) 是响应变量。R 程序及结果输出如下所示:

```
> res.britdoc=glm(deaths~agecat+agesq+smoke+smkage
+offset(log(personyears)),famliy=poisson,data =britdoc)
```

```
Deviance Residuals:
       1         2         3         4         5         6         7         8         9        10
 0.43820  -0.27329  -0.15265   0.23393  -0.05700  -0.83049   0.13404   0.64107  -0.41058  -0.01275

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.79176    0.45008 -23.978  < 2e-16 ***
agecat        2.37648    0.20795  11.428  < 2e-16 ***
agesq        -0.19768    0.02737  -7.223 5.08e-13 ***
smoke         1.44097    0.37220   3.872 0.000108 ***
smkage       -0.30755    0.09704  -3.169 0.001528 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 935.0673  on 9  degrees of freedom
Residual deviance:   1.6354  on 5  degrees of freedom
AIC: 66.703

Number of Fisher Scoring iterations: 4
```

统计模型显示所有的变量都显著，在考虑年龄之后，吸烟是不吸烟者的冠状动脉性猝死率的4($\approx e^{1.4}$) 倍.从输出的第一行中的偏差残差(Deviance Residual) 来看，拟合的效果是比较理想的，所有的残差都很小。事实上，根据这些结果可以很容易得到模型检验结果,可以使用R中的$predict\ fit; predictresidual\ d; deviance; pearson$函数来输出每个观测的模型拟合值，拟合残差，偏差，皮尔逊χ^2 值

```
> fit_p=c(fitted.values(res.britdoc)
> pearsonresid=(britdoc$deaths-fit_p)/sqrt(fit_p)
> chisq=sum(pearsonresid*pearsonresid)
> devres=sign(britdoc$deaths-fit_p)*(sqrt(2*(britdoc$deaths*
log(britdoc$deaths/fit_p)-(britdoc$deaths-fit_p))))
> deviance=sum(devres*devres)
```

如此计算，$\chi^2 = 1.550, D = 1.635$与自由度为$n - p = 10 - 5$分布$p-$值为0.09，展现了较好的拟合度。

§5.8.2 对数线性模型的基本概念

泊松线性模型可用于刻画服从泊松分布的事件发生数与各影响因素（特别是分类变量）之间的关系，它的结构和回归模型十分相似，也称为泊松对数线性模型，

其一般形式为：

$$\log \mu_{ij} = \log n_{ij} + \alpha + \mathrm{x}^T \beta.$$

其中$\log n_{ij}$表示偏移量(offset)，用于去除观察单位数不等的影响。如果单元格中的频数服从多项分布，此时拟合的就是对数线性模型。

简单来看，对数线性模型分析是将列联表的网格频数取对数表示为各个变量（边缘分布）及其交互作用的线性模型形式，从而运用类似方差分析的思想检验各变量及其交互作用的大小,是用于离散数据的列联计数表数据分析方法，把列联分析和线性模型统一起来的研究方法，它强调了模型拟合优度，交互效应和网格频率的估计。

考虑定性变量A和B的联合分布, 其中A取值$A_1, A_2, \cdots, A_r$, B 取值$B_1, B_2, \cdots, B_s$, 根据A 与B 交叉出现的频数统计成$r \times s$ 双向列联表, 如表所示. 令n_{ij}表示(i, j)单元格中的频数,$i = 1, \cdots, r$ and $j = 1, \cdots, s, \sum n_{ij} = n$. 如果$n_{ij}$ 彼此独立服从泊松分布$\mathrm{E}(n_{ij}) = \mu_{ij}$, 那么$\mathrm{E}(n) = \mu = \sum\sum \mu_{ij}$,如果$n_{ij}$ 来自多项分布，那么:$f(\{n_{ij}, i = 1, ..., r, j = 1, ..., s\}|n) = n! \prod_{i=1}^{r} \prod_{j=1}^{s} p_{ij}^{n_{ij}}/n_{ij}!$ 这里$p_{ij} = \mu_{ij}/\mu$. 对于二维列联表，独立性意味着：

$$p_{ij} = p_{i\cdot} p_{\cdot j}.$$

因为$\mu_{ij} = E(n_{ij})$这意味着

$$\log \mu_{ij} = \log n + \log p_{ij}$$

如果独立性成立

$$\log \mu_{ij} = \log n + \log p_{i\cdot} + \log p_{\cdot j}$$

其中, $p_{i\cdot} = \sum_{j=1}^{s} p_{ij}$, $p_{\cdot j} = \sum_{i=1}^{r} p_{ij}$分别表示变量$A$与变量$B$的边缘分布.

表5.22 行变量A和列变量B的联合分布记号表

	B_1	B_2	$\cdots$	B_s	总和
A_1	$p_{11}(n_{11})$	$p_{12}(n_{12})$	$\cdots$	$p_{1s}(n_{1s})$	$p_{1\cdot}(n_{1\cdot})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_r	$p_{r1}(n_{r1})$	$p_{r2}(n_{r2})$	$\cdots$	$p_{rs}(n_{rs})$	$p_{r\cdot}(n_{r\cdot})$
总和	$p_{\cdot 1}(n_{\cdot 1})$	$p_{\cdot 2}(n_{\cdot 2})$	$\cdots$	$p_{\cdot s}(n_{\cdot s})$	$p_{\cdot\cdot}(n_{\cdot\cdot})$

如果两个变量独立, 则有

$$\begin{aligned} p_{ij} &= p_{i\cdot} \cdot p_{\cdot j} = \frac{1}{rs}[rp_{i\cdot}][sp_{\cdot j}] \\ &= \frac{1}{rs}\left[\frac{p_{i\cdot}}{\frac{1}{r}}\right]\left[\frac{p_{\cdot j}}{\frac{1}{s}}\right],\ i = 1, 2, \cdots, r; j = 1, 2, \cdots, s. \end{aligned} \tag{5.10}$$

对两个分类变量的一般情况, p_{ij}有类似的表达形式：

$$p_{ij} = \frac{1}{rs}\left[\frac{p_{i\cdot}}{\frac{1}{r}}\right]\left[\frac{p_{\cdot j}}{\frac{1}{s}}\right]\left[\frac{p_{ij}}{p_{i\cdot}p_{\cdot j}}\right]. \tag{5.11}$$

注意到我们将每个格子的概率p_{ij}分解为四项, $1/rs$是每个格子的期望概率; $\dfrac{p_{i\cdot}}{1/r}$是第i行概率相对于行期望概率的比例; $\dfrac{p_{\cdot j}}{1/s}$是第j列边缘概率相对于列期望概率的比例; 最后一项是联合概率偏离独立性的大小, 如果该值为1, 则表示独立, 大于1或小于1均表示行和列之间有依赖关系. 这与二因子方差分析模型有些相像. 这里也涉及了两个因子, 分别是行和列变量, 各自有r和s个水平. 仿照二因子方差分析模型, 可以将p_{ij}的平均变异原因分解为总体平均效应、行效应、列效应以及行列的交互作用. 但是与方差分析的不同在于, 行和列对p_{ij}的作用不是相加的作用, 而是乘法作用.

p_{ij} =常数×行主效应×列主效应×因子行列交互效应.

两边取对数就可以将乘法模型转换为加法模型：

$\ln(p_{ij}) = \ln$ 常数$+\ln($行主效应$)+\ln($列的主效应$)+\ln($行列交互效应$)$.

上述模型每一项是相对比例, 一般在列联表的不同位置上不均衡, 因此一般使用几何平均数表达各效应的平均情况. 记$r \times s$格子的几何平均概率为$\bar{p}^G_{\cdot\cdot}$, 则

$$\ln \bar{p}^G_{\cdot\cdot} = \frac{1}{rs}\sum_{j=1}^{s}\sum_{i=1}^{r}\ln p_{ij}.$$

行边缘分布的几何平均概率记为$\bar{p}^G_{i\cdot}$, 列边缘分布的几何平均概率记为$\bar{p}^G_{\cdot j}$, 则

$$\ln \bar{p}^G_{i\cdot} = \frac{1}{s}\sum_{j=1}^{s}\ln p_{ij},$$

$$\ln \bar{p}^G_{\cdot j} = \frac{1}{r}\sum_{i=1}^{r}\ln p_{ij}.$$

注意到独立性的表达式如下：

$$\ln p_{ij} = \ln p_{i\cdot} + \ln p_{\cdot j}.$$

将联合概率重新表达成如下的加法形式：

$$\begin{aligned}\ln p_{ij} =& \ln \bar{p}_{\cdot\cdot}^{G} + [\ln \bar{p}_{i\cdot}^{G} - \ln p_{\cdot\cdot}^{G}] + [\ln \bar{p}_{\cdot j}^{G} - \ln \bar{p}_{\cdot\cdot}^{G}] \\ &+ [\ln p_{ij} - \ln \bar{p}_{i\cdot}^{G} - \ln \bar{p}_{\cdot j}^{G} + \ln \bar{p}_{\cdot\cdot}^{G}]. \end{aligned} \tag{5.12}$$

式中

$$\mu = \ln \bar{p}_{\cdot\cdot}^{G} = \frac{1}{rs} \sum_{j=1}^{s} \sum_{i=1}^{r} \ln p_{ij},$$

$$\mu_{A(i)} = \ln \bar{p}_{i\cdot}^{G} - \mu = \frac{1}{s} \sum_{j=1}^{s} \ln p_{ij} - \ln p_{\cdot\cdot}^{G},$$

$$\mu_{B(j)} = \ln \bar{p}_{\cdot j}^{G} - \mu = \frac{1}{r} \sum_{i=1}^{r} \ln p_{ij} - \ln p_{\cdot\cdot}^{G},$$

$$\begin{aligned}\mu_{AB(ij)} =& \ln p_{ij} - \ln \bar{p}_{i\cdot}^{G} - \ln \bar{p}_{\cdot j}^{G} + \ln \bar{p}_{\cdot\cdot}^{G} \\ =& \ln p_{ij} - \mu - \mu_{A(i)} - \mu_{B(j)}. \end{aligned}$$

将式(5.12)改写为

$$\begin{cases} \ln p_{ij} = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{AB(ij)}. \\ \text{其中：} \\ \sum_{i=1}^{r} \mu_{A(i)} = 0; \quad \sum_{j=1}^{s} \mu_{B(j)} = 0; \quad \sum_{i=1}^{r} \mu_{AB(ij)} = 0; \quad \sum_{j=1}^{s} \mu_{AB(ij)} = 0 \end{cases} \tag{5.13}$$

式(5.13)就是二维对数线性模型的一般形式, 如果行变量A和列变量B独立, 那么

$$p_{ij} p_{kl} = p_{kj} p_{il}, \forall i, k = 1, 2, \cdots, r; j, l = 1, 2, \cdots, s.$$

即

$$p_{ij} = \frac{\bar{p}_{i\cdot}^{G} \bar{p}_{\cdot j}^{G}}{\bar{p}_{\cdot\cdot}^{G}}.$$

这相当于

$$\begin{cases} \ln \bar{p}_{\cdot\cdot}^{G} + \ln p_{ij} = \ln \bar{p}_{\cdot j}^{G} + \ln \bar{p}_{i\cdot}^{G}. \\ \mu_{AB(ij)} = 0. \end{cases} \tag{5.14}$$

因而独立性假设下的对数线性模型可以改写为

$$\begin{cases} \ln p_{ij} = \mu + \mu_{A(i)} + \mu_{B(j)} + \varepsilon_{ij}, \\ \sum\limits_{i=1}^{r} \mu_{A(i)} = 0, \quad \sum\limits_{j=1}^{s} \mu_{B(j)} = 0. \end{cases} \tag{5.15}$$

模型(5.15)称为独立性模型, 而式(5.13)称为饱和模型.

例5.10 为研究不同年龄人群对某地区缺水问题的态度, 按年龄调查了该地区部分居民, 要求他们评价缺水问题的严重程度, 得到表5.23所示的数据表.

表5.23 不同年龄居民对缺水情况的态度——联合分布频率表p_{ij}

年龄	不严重	稍严重	严重	很严重	列合计
30岁以下	0.015	0.076	0.121	0.055	0.267
30~40 岁	0.017	0.117	0.111	0.037	0.282
40~50岁	0.012	0.074	0.104	0.032	0.222
50~60岁	0.007	0.034	0.072	0.020	0.133
60岁及以上	0.001	0.027	0.038	0.030	0.096
行合计	0.052	0.328	0.446	0.174	1.000

要求利用该表建立一个对数线性模型.

解 表中x_{ij}为年龄第i组, 回答第j项目的频率, 它是两因子联合概率分布的估计值. 我们的目的是研究不同年龄层对缺水严重程度的回答是否一致, 即不同年龄的回答是否相同(A因子主效应), 同样也要检验不同严重程度之间的回答比例是否相同(B因子主效应), 还要检验年龄与严重程度之间的关系(A、B两因子交互效应). 首先, 计算年龄和对缺水意见的交互作用, 如表5.24所示.

表5.24 联合分布概率$\dfrac{p_{ij}}{p_{i\cdot}p_{\cdot j}}$

年龄水平	不严重	稍严重	严重	很严重
30岁以下	1.08	0.87	1.01	1.19
30~40岁	1.14	1.27	0.88	0.75
40~50岁	1.07	1.01	1.05	0.83
50~60岁	1.03	0.78	1.22	0.85
60岁及以上	0.18	0.85	0.89	1.81

表5.24中表示了缺水意见与年龄的交互作用与1比较的大小. 表中最小值0.18和最大值1.81均显示了偏离独立性的特点. 最小值和最大值都在最大年龄这一层, 这说明高年龄组中, 有少部分人认为缺水的问题不严重, 但相当多的人认为犯罪情况很严重. 在30~50岁的年龄组中, 只有很少人认为当前缺水问题很严重, 这说明, 年龄与对缺水问题的态度是有关系的.

不同年龄组对缺水情况的格子分布概率计算如表5.25所示.

表5.25 不同年龄组对缺水情况的态度——格子分布概率的对数

年龄水平	不严重	稍严重	严重	很严重	列合计	列平均
					$\left(\sum_{j=1}^{s}\ln p_{ij}\right)$	$\frac{1}{s}\sum_{j=1}^{s}\ln p_{ij}=\ln\bar{p}_{i\cdot}$
30岁以下	−4.200	−2.577	−2.112	−2.900	−11.789	−2.947
30∼40岁	−4.075	−2.146	−2.198	−3.297	−11.716	−2.929
40∼50岁	−4.423	−2.604	−2.263	−3.442	−12.732	−3.183
50∼60岁	−4.962	−3.381	−2.631	−3.912	−14.886	−3.722
60岁及以上	−6.908	−3.612	−3.27	−3.507	−17.297	−4.324
行合计	−24.568	−14.320	−12.474	−17.058	−68.420	
行平均	−4.914	−2.864	−2.495	−3.412		−3.421

由表5.25可得

$$\mu=\ln\bar{p}_{\cdot\cdot}^{G}=-3.421,$$

$$\mu_{B(1)}=\frac{1}{r}\sum_{i=1}^{r}\ln p_{i1}-\ln\bar{p}_{\cdot\cdot}^{G}=-4.914-(-3.421)=-1.493,$$

$$\mu_{B(2)}=\frac{1}{r}\sum_{i=1}^{r}\ln p_{i2}-\ln\bar{p}_{\cdot\cdot}^{G}=-2.864-(-3.421)=0.557,$$

$$\mu_{B(3)}=\frac{1}{r}\sum_{i=1}^{r}\ln p_{i3}-\ln\bar{p}_{\cdot\cdot}^{G}=-2.495-(-3.421)=0.926,$$

$$\mu_{B(4)}=\frac{1}{r}\sum_{i=1}^{r}\ln p_{i4}-\ln\bar{p}_{\cdot\cdot}^{G}=-3.412-(-3.421)=0.009,$$

$$\mu_{A(1)}=\frac{1}{s}\sum_{j=1}^{s}\ln p_{1j}-\ln\bar{p}_{\cdot\cdot}^{G}=-2.947-(-3.421)=0.474,$$

$$\mu_{A(2)}=\frac{1}{s}\sum_{j=1}^{s}\ln p_{2j}-\ln\bar{p}_{\cdot\cdot}^{G}=-2.929-(-3.421)=0.492,$$

$$\mu_{A(3)}=\frac{1}{s}\sum_{j=1}^{s}\ln p_{3j}-\ln\bar{p}_{\cdot\cdot}^{G}=-3.183-(-3.421)=0.238,$$

$$\mu_{A(4)}=\frac{1}{s}\sum_{j=1}^{s}\ln p_{4j}-\ln\bar{p}_{\cdot\cdot}^{G}=-3.722-(-3.421)=-0.301,$$

$$\mu_{A(5)}=\frac{1}{s}\sum_{j=1}^{s}\ln p_{5j}-\ln\bar{p}_{\cdot\cdot}^{G}=-4.324-(-3.421)=-0.903.$$

$\ln p_{ij}-\mu-\mu_{A(i)}-\mu_{B(j)}$表示偏离独立性的大小, 可以用交互作用参数$\mu_{AB(ij)}$表示. 将$\mu_{AB(ij)}$ 用表5.26表示, 其中$A(i)$和$B(j)$相交的位置处表示$\mu_{AB(ij)}$.

表5.26 A与B交互作用的期望值

B	$B(1)$	$B(2)$	$B(3)$	$B(4)$
$A(1)$	0.240	−0.188	−0.091	0.038
$A(2)$	0.347	0.225	−0.195	−0.377
$A(3)$	0.253	0.021	−0.006	−0.268
$A(4)$	0.253	−0.217	0.165	−0.199
$A(5)$	−1.091	0.151	0.128	0.808

其中列和与行和都为零. 从交互作用来看, 回答不严重类中,与零差距最大的是$\mu_{AB(51)}$; 在认为最严重的一类中, 与零差距最大的是$\mu_{AB(54)}$.将这些结果代入式(5.13)就得到一个对数线性模型.

上面给出的对数线性模型是以频率或概率对数的形式出现的, 实际上从格点频数对数的角度也可以得到模型. 这里不再赘述过程, 只给出一般的定义, 如下所示：

$$\begin{cases} \ln M_{ij} = \mu + \mu_{A(i)} + \mu_{B(j)} + \mu_{AB(ij)}, i = 1, 2, \cdots, r; j = 1, 2, \cdots, s; \\ \text{其中：} \\ \displaystyle\sum_{i=1}^{r} \mu_{A(i)} = \sum_{i=1}^{r} \left(\ln \bar{p}_{i\cdot}^{G} - \mu\right) = \frac{1}{s} \sum_{i=1}^{r} \left(\sum_{j=1}^{s} \ln p_{ij} - \ln p_{\cdot\cdot}^{G} \right) = 0, \\ \displaystyle\sum_{j=1}^{s} \mu_{B(j)} = \sum_{j=1}^{s} \left(\ln \bar{p}_{\cdot j}^{G} - \mu\right) = \sum_{j=1}^{s} \left(\frac{1}{r} \sum_{i=1}^{r} \ln p_{ij} - \ln p_{\cdot\cdot}^{G} \right) = 0. \end{cases}$$

上式中

$$\begin{aligned} \mu = &= \frac{1}{rs} \sum_{j=1}^{s} \sum_{i=1}^{r} \ln M_{ij}, \\ \mu_{A(i)} &= \frac{1}{s} \sum_{j=1}^{s} \ln M_{ij} - \mu, \\ \mu_{B(j)} &= \frac{1}{r} \sum_{i=1}^{r} \ln M_{ij} - \mu, \\ \mu_{AB(ij)} &= \ln M_{ij} - \mu - \mu_{A(i)} - \mu_{B(j)}. \end{aligned}$$

用频数定义的最大好处是更方便通过参数估计和模型, 直接估计出每个格点的期望频数. 然后可以根据这些期望频数的分布规律, 进一步分析各变量水平之间的关系.

§5.8.3 模型的设计矩阵

和多元线性模型一样, 对数线性模型也有矩阵的表现形式. 利用矩阵形式可以更方便进行参数估计和检验. 这里我们仅以2×2列联表为例, 说明设计矩阵的

表现形式. 在二维对数线性模型中, 令4个参数为$\beta_0,\beta_1,\beta_2,\beta_3$, 用$L_{ij}$ 表示$\ln p_{ij}, i=1,2,\cdots,r, j=1,2,\cdots,s$, 模型可以用以下矩阵表示：

$$\boldsymbol{L}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$$

式中

$$\boldsymbol{L}=\begin{pmatrix} L_{11}\\ L_{12}\\ L_{21}\\ L_{22}\end{pmatrix},\quad \boldsymbol{X}=\begin{pmatrix} 1 & 1 & 1 & 1\\ 1 & 1 & -1 & -1\\ 1 & -1 & 1 & -1\\ 1 & -1 & -1 & 1\end{pmatrix},\quad \boldsymbol{\beta}=\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\end{pmatrix}.$$

实际上, 由式(5.13), 对于$r\times s=2\times 2$列联表数据结构特征, 由$\ln(p_{ij})=\mu+\mu_{A(i)}+\mu_{B(j)}+\mu_{AB(ij)}$,因而有

$$\begin{cases}\ln(p_{11})=\mu+\mu_{A(1)}+\mu_{B(1)}+\mu_{AB(11)},\\ \ln(p_{12})=\mu+\mu_{A(1)}+\mu_{B(2)}+\mu_{AB(12)},\\ \ln(p_{21})=\mu+\mu_{A(2)}+\mu_{B(1)}+\mu_{AB(21)},\\ \ln(p_{22})=\mu+\mu_{A(2)}+\mu_{B(2)}+\mu_{AB(22)}.\end{cases}\tag{5.16}$$

联立方程组(5.16)有9个未知数, 但只有4个观测值, 再加入下列限制条件：
$\mu_{A(1)}+\mu_{A(2)}=0$, $\mu_{B(1)}+\mu_{B(2)}=0$, 即$\sum\mu_{A(i)}=0$, $\sum\mu_{B(j)}=0$ 及

$$\begin{cases}\mu_{AB(11)}+\mu_{AB(21)}=0,\\ \mu_{AB(12)}+\mu_{AB(22)}=0;\end{cases}$$

$$\begin{cases}\mu_{AB(11)}+\mu_{AB(12)}=0,\\ \mu_{AB(21)}+\mu_{AB(22)}=0.\end{cases}$$

因此9个未知参数减少到4个, 式(5.16)改写为

$$\begin{cases}\ln(p_{11})=\mu+\mu_{A(1)}+\mu_{B(1)}+\mu_{AB(11)},\\ \ln(p_{12})=\mu+\mu_{A(1)}-\mu_{B(1)}-\mu_{AB(11)},\\ \ln(p_{21})=\mu-\mu_{A(1)}+\mu_{B(1)}-\mu_{AB(11)},\\ \ln(p_{22})=\mu-\mu_{A(1)}-\mu_{B(1)}+\mu_{AB(11)}.\end{cases}\tag{5.17}$$

注意到$\sum\limits_{ij}p_{ij}=1$, 因此实际上模型还可以化简为

$$\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$$

用矩阵表示如下：

$$\boldsymbol{Y}=\begin{pmatrix} y_1=\ln p_{11}/\ln p_{22} \\ y_2=\ln p_{12}/\ln p_{22} \\ y_3=\ln p_{21}/\ln p_{22} \end{pmatrix}=\begin{pmatrix} 2 & 2 & 0 \\ 2 & 0 & -2 \\ 0 & 2 & -2 \end{pmatrix}\cdot\begin{pmatrix} \mu_{A(1)} \\ \mu_{B(1)} \\ \mu_{AB(11)} \end{pmatrix}+\boldsymbol{\varepsilon}.$$

其中只有3个需要估计的参数.

§5.8.4 模型的估计和检验

建立对数线性模型后, 就可以估计参数$B=\{\beta_1,\beta_2,\beta_3\}$以及它们的方差var$(B)$, 以便检验各效应是否存在. 对于饱和模型, 通常可以采用加权最小二乘法(weighted-least squares estimation)或极大似然估计法, 但对于不饱和模型通常采用极大似然估计算法估计模型参数, 这里不详细介绍.

模型的拟合优度(goodness of fit test)用于检验模型拟合的效果. 以$r\times s$二维列联表为例, 模型的独立参数有3个, 设为β_1,β_2,β_3. 则假设检验问题为

$$H_0:\beta_i=0,i=1,2,3\leftrightarrow H_1:\exists i\ \ \beta_i\neq 0.$$

常用的检验统计量有两个：一个是Pearsonχ^2统计量; 另一个是对数似然比统计量, 分别表示为

$$\chi^2=\sum_{i,j}^{rs}\frac{(n_{ij}-m_{ij})^2}{m_{ij}}; \tag{5.18}$$

$$G^2=-2\sum_{i,j}^{rs}n_{ij}\ln\frac{n_{ij}}{m_{ij}}. \tag{5.19}$$

其中, n_{ij}表示列联表中第i行第j列的观察频数, m_{ij}表示该格的期望频数. 在零假设之下, 两个统计量都近似服从自由度$df=rs-k$的χ^2分布, k是模型中独立参数的个数.

根据对数线性模型(5.13)的数学表达式和限制条件可知, 变量A的主效应有$r-1$个独立参数, 变量B的主效应有$s-1$个独立参数, 变量A和变量B的交互效应有$(r-1)\times(s-1)$个独立参数, 再加上常数项, 应该有$1+(r-1)+(s-1)+(r-1)(s-1)=rs$个独立参数, 而没有交互项的独立模型只有$1+(r-1)+(s-1)=r+s-1$个独立参数. 模型的自由度等于数据提供的信息量减去模型中独立参数的个数. 对列联表数据而言, 所有的格子的个数就是整个信息量, 即rs. 因此模型(5.12)的自由度为0, 独立模型的自由度$df=rs-(r+s-1)=(r-1)(s-1)$.

§5.8.5 高维对数线性模型和独立性

类似二维列联表, 也有高维列联表的对数线性模型. 以$r\times s\times t$三维表为例,假设有三个分类变量A,B,C, A变量有r个水平, B变量有s个水平, C变量有t个水平, 它们构成一个$r\times s\times t$的三层列联表. 令X_{ijk}为第i行j列k层格子的观测值, p_{ijk}为X_{ijk}的理论概率值, 三维对数线性模型的一般形式为

$$\begin{aligned}\ln p_{ij} =& \mu+\mu_{A(i)}+\mu_{B(j)}+\mu_{C(k)}\\ &+\mu_{AB(ij)}+\mu_{BC(ij)}+\mu_{AC(ij)}+\mu_{ABC(ijk)}\\ i=1,2,\cdots,r;& j=1,2,\cdots,s;k=1,2,\cdots,t.\end{aligned}$$

其中

$$\sum_{i=1}^{r}\mu_{A(i)}=\sum_{j=1}^{s}\mu_{B(j)}=\sum_{k=1}^{t}\mu_{C(k)}\equiv 0,$$

$$\sum_{i=1}^{r}\mu_{AB(ij)}=\sum_{j=1}^{s}\mu_{AB(ij)}\equiv 0,$$

$$\sum_{i=1}^{r}\mu_{AC(ik)}=\sum_{k=1}^{t}\mu_{AC(ik)}\equiv 0,$$

$$\sum_{j=1}^{s}\mu_{BC(jk)}=\sum_{k=1}^{t}\mu_{BC(jk)}\equiv 0,$$

$$\sum_{i=1}^{r}\mu_{ABC(ijk)}=\sum_{j=1}^{s}\mu_{ABC(ijk)}=\sum_{k=1}^{t}\mu_{ABC(ijk)}\equiv 0.$$

如果三个变量A,B,C独立, 则对数线性模型为

$$\ln p_{ij}=\mu+\mu_{A(i)}+\mu_{B(j)}+\mu_{C(k)}. \tag{5.20}$$

三维列联表的独立性共有4种情况, 如表5.27所示.

表5.27 三维列联表的独立类型

标记	独立类型	定义说明
Ⅰ型	边缘独立	三维列联表的任意两个变量独立
Ⅱ型	条件独立	当一个变量固定不变, 另外两个变量独立
Ⅲ型	联合独立	将两个变量组合, 形成新变量, 新变量和第三个变量独立
Ⅳ型	相互独立	3个变量中任何一个变量与另外两个变量联合独立

值得注意的是, 四种独立性之间存在如下关系.

(1) (Ⅳ⇒ Ⅲ)：若X,Y,Z相互独立, 则任意两个变量组合成的新变量与剩余的第三个变量独立.

(2) (Ⅲ⇒ Ⅰ, Ⅲ⇒ Ⅱ)：若X与Y,Z联合独立, 则X与Y, X与Z边缘独立; 给定Y, X与Z条件独立, 给定Z, X与Y条件独立.

但是, 条件独立不能得到边缘独立.

(3) (Ⅱ和Ⅰ不能互推)若X与Y条件独立, 不一定有X与Y 边缘独立. 反之, X与Y边缘独立, 也不一定有X与Y条独立.

可以作不同的独立性检验, 如表5.28所示.

表5.28 三维列联表可作的不同独立性检验

模型记号	可作的检验	独立类型
(X,Y,Z)	X,Y,Z相互独立	Ⅳ型
(XY,Z)	(X,Y)与Z独立	Ⅲ型
(Y,XZ)	(X,Z)与Y独立	Ⅲ型
(X,YZ)	X与(Y,Z)独立	Ⅲ型
(XZ,YZ)	给定Z时X与Y独立	Ⅱ型
(XY,YZ)	给定Y时X与Z独立	Ⅱ型
(XY,XZ)	给定X时Y与Z独立	Ⅱ型

为叙述方便,用(XYZ)表示饱和模型, (X,Y,Z)表示独立性模型. 中间一些模型用这三个字母的一些组合来代表. 比如(Y,XZ) 代表模型中包含X,Z的交互作用(没有和Y的交互作用) 及所有出现的字母所代表的主效应的模型, 即

$$\ln m_{ijk}=\mu+\lambda_i^X+\lambda_j^Y+\lambda_k^Z+\lambda_{ij}^{XZ};$$

而(XY,XZ) 代表有X,Y及X,Z 两个交互作用及所有主效应的模型, 即饱和模型去掉λ_{ijk}^{XYZ}和λ_{jk}^{YZ} 项.

在各种模型下, 可以作不同的独立性检验, 对于上面所说的各种变量的独立性, 和二维一样可以用Pearson 统计量, 或似然比统计量进行χ^2检验. 如果真实模型和零假设下的模型不一致, 则这两个统计量会偏大.

例5.11 下面是对三所学校五年级分学生性别统计的近视观察数据:

表5.29 三所学校按性别统计学生近视人数数据表

	学校因素(Y)					
	甲		乙		丙	
性别因素(X) / 近视因素(Z)	男	女	男	女	男	女
近视	55	58	66	85	66	50
不近视	45	41	87	70	41	39

研究的目标是想了解哪些变量独立, 哪些不独立.

解 令X表示性别, Y表示学校, Z表示近视, 下面就3个变量可能感兴趣的独立性问题做出检验, 结果如表5.30 所示（显著性水平$\alpha = 0.10$）.

表5.30 对数线性模型的模型拟合优度检验结果

模型	d.f.	LRT G^2	p值	Pearson Q	p值	结论
(X,Y,Z)	7	12.17	0.0951	12.12	0.0968	X,Y,Z不独立
(XY,Z)	5	10.91	0.0531	10.90	0.0533	(X,Y)和Z不独立
(X,YZ)	5	6.36	0.2727	6.347	0.2739	X和(Y,Z)独立
(XZ,Y)	6	10.85	0.0930	10.93	0.0907	Y和(X,Z)不独立
(XZ,XY)	4	9.59	0.0479	9.538	0.0489	给定X,Y和Z不独立
(XY,YZ)	3	5.09	0.1648	5.088	0.1654	给定Y,X和Z独立
(XZ,YZ)	4	5.04	0.2834	5.025	0.2847	给定Z,X和Y独立

由表中可以看出, 近视、性别和学校之间存在关联性. 到底关联性是怎样产生的? 由具体的独立性分析可知,没有发现不同的学校近视情况(YZ)与性别(X)有关。就近视(Z)而言，不能说学校与性别(X)关系密切;就学校(Y)而言，不能说近视(Z)与性别(X)关系密切. 但是由(X,Y)与Z不独立，可以说，最多的近视是乙校的女生，不近视最多的是乙校的男生。可以看出，学校丙的女生不近视率较低，对数线性模型中应加入Y和Z的交互作用项。

在R中进行对数线性模型独立性检验的示范程序如下：

```
> f=function(x)
   {
    df=x$df  #求自由度
    lrt=x$lrt #似然比检验统计量
    p.lrt=1-pchisq(x$lrt,x$df) #似然比检验统计量的p值
    Q=x$pear  #pearson 检验统计量Q
    p.pear=1-pchisq(x$pear,x$df)#pearson检验统计量Q的p值
    if(p.lrt<0.05|p.pear<0.05){conclusion="不独立"}else{conclusion="独立"}
    list(df,lrt,p.lrt,Q,p.pear,conclusion)
```

```
  }
A=matrix(c(55,58,66,85,66,50),nrow=2)
B=matrix(c(45,41,87,70,41,39),nrow=2)
a=array(c(A,B),dim=c(2,3,2))
m1=loglin(a,list(1,2,3))  #模型(x,y,z)
## iterations:deviation 1.1368e-13
f1=f(m1)
loglin(a,list(c(1,2),3))$lrt
loglin(a,list(c(1,2),c(1,3)))$lrt
```

习题

5.1 在一个有3个主要大型商场的商贸中心, 调查479个不同年龄段的人首先去3个商场中的哪一个, 结果如表5.19所示.

表5.31 不同商场客户的倾向性研究

年龄段	商场1	商场2	商场3	总和
$\leqslant 30$	83	70	45	198
$31 \sim 50$	91	86	15	192
> 50	41	38	10	89
总和	215	194	70	479

问题：不同年龄段人对各商场的购物倾向性是否存在差异?

5.2 美国某年总统选举前, 由社会调查总部(General Social Survey)抽查黑白种族与支持不同政党是否有关, 得到表5.32.

表5.32 黑白种族与支持不同政党之间的关系

种族	民主党	共和党	无党
白人	341	405	105
黑人	103	11	15

问：不同种族(race)与所支持的政党之间是否存在独立性?

5.3 下面是一个医学例子, 研究某类肺炎患者和以前是否曾经患过该类肺炎之间的疾病继承性关系. 下面是30个人按照当前患某类肺炎和曾经患某类肺炎之间的2×2分类表.

表5.33 某类肺炎继承性研究数据表

	以前有过某类肺炎	以前没有某类肺炎	总和
当前有过某类肺炎	6	4	10
当前没有某类肺炎	1	19	20
总和	7	23	30

5.4 对479个不同年龄段的人调查他们对各种不同类型电视节目的喜爱情况, 要求每人只能选出他们最喜欢观看的电视节目类型, 结果如表5.34 所示.

表5.34 不同年龄层次的人与电视节目类型之间的关系

年龄段	体育类1	电视剧类2	综艺类3	总和
$\leqslant 30$	83	70	45	198
$31 \sim 50$	91	86	15	192
> 50	41	38	10	89
总和	215	194	70	479

问：不同观众对三类节目的关注率是否一样?

5.5 有人认为当代学生和20世纪60年代的学生之间存在很大差异. 他在某学校做了一些跟踪调查试验, 问了学生如下问题：以下哪个因素是你选择大学深造的主要原因(单项选择)? (a) 丰富人生哲学; (b) 增强对周围世界的了解; (c) 找到好工作; (d) 不清楚. 同样的问题1965年也向在校学生提问过, 以下是两个调查结果：

表5.35 大学生选择大学深造的原因调查数据表

	1965年	1998年
丰富人生哲学	15	8
增加对周围世界的了解	53	48
找到好工作	25	57
不清楚	27	47

作者能够根据这些数据判断出两代大学生之间的差异吗?

5.6 继续例5.6的分析, 如果不按照分层结构直接计算分类变量, 能得到怎样的结论?

5.7 对三类不同学校, 分别考察学生家庭经济情况与其高考成绩之间的关系, 用经济状况好(A)与经济状况一般(B)对比记录其结果, 如下表：

表5.36 家庭经济情况与学生高考情况关系数据表

学校	经济状况	一类学校	二类学校
1	A	43	65
	B	87	77
2	A	9	73
	B	15	30
3	A	7	18
	B	9	11

试分析学生家庭经济情况与其高考成绩之间的关系.

5.8 令S是一个有限项集.

(1) 令A, B是S的子集, 试定义下列规则的支持度(support)、可信度(confidence)、提升(lift)：

$$A \Rightarrow B$$

(2) 一个强规则的定义是满足最小支持度s_0和最小可信度c_0 的规则. 试对$s_0 = 0.6$和$c_0 = 0.8$, 从下面的数据发现所有形式为：$\{x_1, x_2\} \Rightarrow \{y\}(x_1, x_2, y \in S, x_1 \neq x_2 \neq y)$的强规则.

表5.37 购物篮交易记录表

交易	项集
1	$\{a, b, d, k\}$
2	$\{a, b, c, d, e\}$
3	$\{a, b, c, e\}$
4	$\{a, b, d\}$

5.9 见光盘shopping-basket.xls数据, 是对一个超市的购买记录. 其特征变量为：sex(性别), hometown(是否本地), income(收入), age(年龄), fruitveg(果蔬), freshmeat(鲜肉), dairy(乳品), cannedveg(罐头蔬菜), cannedmeat(罐头肉), frozenmeat(冻肉), wine(酒), softdrink(软饮料), fish(鱼), confectionery (糖果) 共1000个观测. 试用Apriori算法找出这个数据中有意义的规则, 把支持度和可信度都设定为0.8.

5.10 证明定理5.2中关于Ridit得分的第二个等式.

5.11 假设某电信公司调查某款便携式手机的售后产品及服务满意度, 统计得到调查数据表如表5.38所示.

表5.38 手机售后满意度统计表

问 项	总数	非常不满意	不满意	一般	满意	非常满意
1. 对手机信号的满意度	200	90	23	53	21	13
2. 对手机外型的满意度	132	47	34	28	18	5
3. 对手机维修质量的满意度	50	20	13	10	5	2
4. 对手机功能的满意度	154	28	32	33	45	16
5.对手机操作方便的满意度	164	34	28	52	40	10
总数	700	219	130	176	129	46

选择方法分析各个问项满意度之间是否存在差异?

5.12 设春秋两个雨季在某山坡上造林,在栽种的部分土穴中放有机肥, 另外一些土穴中未放有机肥, 结果树种成活数量与不成活的数量如表5.24所示.试用对数线性模式检验春秋雨季与是否填埋有机肥对树的成活数是否存在差异, 以及交互作用是否存在.

表5.39 不同季节施肥和树苗成活情况统计表

是否放有机肥 / 季节	放有机肥	无有机肥
	活 死	活 死
春	385 48	400 115
秋	198 50	375 120

案例与讨论1:数字化运营转化率

案例背景

互联网大数据时代，线上店铺通过接入某数字平台可以瞬间提升流量，常用的流量提升平台有APP，微商等。企业掌握了流量相当于掌握了运营的先机，有效的流量销售额转化率是成功经营的关键。本案例根据电商的经营模式一般可以将转化率分为四种不同的层次：静默转化率、咨询转化率、加购转化率和成交转化率。UV 数表示对店铺进行过访问的入店访客数，而其中的有效访客数是在店铺中访问了至少若干个页面之后才离开的用户，也称为静默访客，即全程没有跟客服沟通的访客; 咨询表示有对产品进行文字询问和明确的购买倾向表态的客户,加购访客数，指所有来访的用户中，点击了添加购物车按钮将商品加入购物车的凡客数量;成交访客数，即提交订单并且成功付款的用户。

数据说明与约定

本次数据来自双十一这一天某电商销售数据，该电商有64个产品页面入口，Log Usernumber表示的是该网页注册人数，ClickSilent 表示注册客户中有连续点击2个及以上页面的人数，InputQuery 表示长时间浏览网页后与客服进行产品咨询以及修改寄送地址的人数，ShoopingCar 表示有购买倾向的人将产品点选进购物车的人数，Paycheck 是成功付款成交的人数，USerSource 表示浏览器页面还是微信页面，LastMonth 表示上个月该网页销售量情况，销售量500 件以上标记为”H”，否则为”L”,CommensLevel 表示该商品上个月的好评数，M表示好评数为”高”,S 表示好评数为”一般”，D表示好评数为”低”。Sales 表示当日销售额（人民币元）。

研讨问题

1.请问接入不同的落地页（浏览器还是微商页面）对销量有影响吗？采用怎样的分析方法?

2.请问上月产品的销量和好评数对本月的销量有影响吗？如果有影响，影响是怎样的？

3.销售额和以上这些因素有怎样的关系？

案例与讨论2:影响婴儿体重的相关因素分析

案例背景

一研究机构获得一组研究数据，数据内容包含1000名婴儿的出生体重和5个相关变量。使用这份数据来探究低体重出生婴儿的影响变量。

数据说明与约定

表5.40中给出了6所关注的变量具体名称及含义。

表5.40 birth_weight数据变量含义

变量名称	变量含义	变量名称	变量含义
weight	婴儿出生体重（g）	ed.hs	母亲的教育程度是否为高中
black	母亲是否为黑人（1：黑人，0：白人）	ed.col	母亲的教育程度是否为大学
married	母亲是否已婚（1：是，0：不是）	m.wtgain	母亲怀孕期间增加的体重（磅）

研讨问题

1. 定义低体重出生婴儿体重为2500g以下。请绘图刻画未婚黑人母亲的婴儿出生体重（weight）经验分布的点估计和置信区间估计（置信度90%），并从图上观察判断婴儿低体重和母亲婚姻状况之间的关系。

2. 用泊松回归模型探究黑人母亲、母亲怀孕期间增加体重与低体重婴儿之间的关系。

参考解答过程

1. 问题1: 根据第一章的Dvoretzky-Kiefer-Wolfowitz不等式，可以得到：

令

$$L(x) = \max\{F_n(x) - \epsilon_n, 0\}, \quad U[x] = \min\{F_n(x) + \epsilon_n, 1\}$$

其中

$$\epsilon_n = \sqrt{\frac{1}{2n} \ln \frac{2}{\alpha}}$$

那么

$$P(L(x) \leqslant F(x) \leqslant U(x)) \geqslant 1 - \alpha$$

根据该公式，绘制出未婚黑人母亲的婴儿出生体重经验分布及其置信区间估计。

```
data <- read.csv("birth\_weight.csv") \#读入数据
weight.sort <- sort(data\$weight[data\$black==1\&data\$married==0])
weight.rank <- rank(weight.sort)
n <- length(weight.sort)
weight.ecd <- weight.rank/n \#得到经验分布函数
plot(weight.sort,weight.ecd,type = "o",xlab = "weight",ylab = "Fn(x)",
main="未婚黑人母亲婴儿的体重的经验分布图及置信区间估计")
band <- sqrt(log(2/0.1)/(2*n))
```

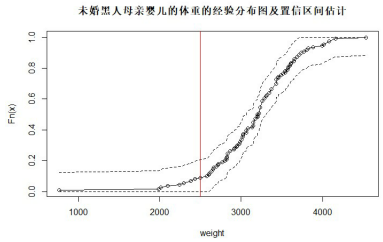

图 5.3　未婚黑人母亲婴儿体重经验分布函数估计

```
#计算得到置信区间在各点处的上下界。
lower.9 <- weight.ecd-band
upper.9 <- weight.ecd+band
lower.9[which(lower.9<0)] <- 0
upper.9[which(upper.9>1)] <- 1
lines(weight.sort,lower.9,lty=2)
lines(weight.sort,upper.9,lty=2)
abline(v=2500,col="red")
```

图中的红色竖线即为判定婴儿是否为低体重的界限（2500g），可以看到在90%的置信度下，未婚黑人母亲的婴儿的低体重率的上限接近20%，与世界平均水平相比这个比率是偏高的；从整体看，婴儿的体重主要分布在2500g-4000g之间，但依然存在一定数量的极小值；从体重角度来看，未婚黑人母亲的婴儿健康状况较为堪忧且整体方差较大。

问题2

```
#将母亲增加的体重按照表格中的级别分类。
data\$classified[data\$m.wtgain>-50 \& data\$m.wtgain<=-20] <- "-50 - -20"
data\$classified[data\$m.wtgain>=-19 \& data\$m.wtgain<=-10] <- "-19 - -10"
data\$classified[data\$m.wtgain>=-9 \& data\$m.wtgain<=0] <- "-9 - 0"
data\$classified[data\$m.wtgain>=1 \& data\$m.wtgain<=10] <- "1 - 10"
data\$classified[data\$m.wtgain>=11 \& data\$m.wtgain<=20] <- "11 - 20"
data\$classified[data\$m.wtgain>=21 \& data\$m.wtgain<=55] <- "21 - 55"
data\$classified <- factor(data\$classified,levels =
c("-50 - -20","-19 - -10", "-9 - 0","1 - 10","11 - 20","21 - 55"))
data\$LowAndBlack <- 0
```

表5.41 birth_weight数据变量含义

母亲增加的体重	黑人母亲低体重婴儿数	黑人母亲数	白人母亲低体重婴儿数	白人母亲数
-50 - -20	2	17	5	51
-19 - -10	29	4	9	119
-9 - 0	6	63	15	276
1 - 10	1	32	5	252
11 - 20	2	22	4	86
21 - 55	1	10	1	43

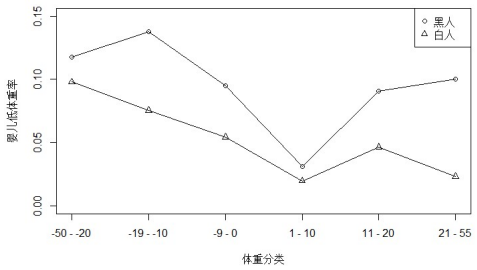

图 5.4 黑人母亲和白人母亲孕期体重增长量与婴儿出生体重之间的关系对比图

```
data\$LowAndBlack[data\$black==1\&data\$weight<2500] <- 1
data\$LowAndWhite <- 0
data\$LowAndWhite[data\$black==0\&data\$weight<2500] <- 1
data\_new <- dplyr::group\_by(data,classified)
block <- dplyr::summarise(data\_new,black = sum(black),black\_low =
sum(LowAndBlack),white=n()-sum(black),white\_low=sum(LowAndWhite))
```

得到填充好的表格如表5.26：根据表5.26可以计算得到不同的母亲怀孕期间体重增长的分类结果，黑人与白人母亲的低体重婴儿比率。 从图5.3中可以看到，母亲孕期间增加体重小于10磅时，婴儿低体重率随母亲增加体重的增加而减少，但当母亲孕期体重增加超过10磅时，婴儿低体重率反而有所上升；另一方面，黑人母亲的低体重婴儿率在各个增加体重级别上均高于白人母亲。
根据这些表象分析，可以建立泊松回归，将婴儿低体重率作为因变量，将母亲孕期体重增加数、母亲的肤色作为自变量，婴儿总数作为基数。

```
dat <- data.frame(loss\_weight=rep(block\$classified,2),
```

```
> summary(res.weight)

Call:
glm(formula = weight_low ~ loss_weight + black + offset(log(total_number)),
    family = poisson, data = dat)

Deviance Residuals:
       1         2         3         4         5         6         7         8         9
-0.37678   0.05534  -0.00003  -0.08945   0.10349   0.50079   0.27073  -0.03640   0.00002
      10        11        12
 0.04146  -0.07058  -0.37314

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -2.4460     0.3937  -6.213 5.19e-10 ***
loss_weight-19 - -10 -0.1238     0.4693  -0.264   0.7919
loss_weight-9 - 0    -0.4664     0.4371  -1.067   0.2860
loss_weight1 - 10    -1.4926     0.5591  -2.670   0.0076 **
loss_weight11 - 20   -0.5870     0.5566  -1.055   0.2916
loss_weight21 - 55   -0.9639     0.8021  -1.202   0.2295
black1                0.5610     0.2984   1.880   0.0601 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 16.95811  on 11  degrees of freedom
Residual deviance:  0.63508  on  5  degrees of freedom
AIC: 51.623

Number of Fisher Scoring iterations: 4
```

```
+black=c(rep(1,6),rep(0,6)),total\_number =
c(block\$black,block\$white),weight\_low =
+c(block\$black\_low,block\$white\_low))
dat\$black <- as.factor(dat\$black)
res.weight <- glm(weight\_low ~loss\_weight+black+
+offset(log(total\_number)),data = dat,family = poisson)
```

从模型来看不存在过度拟合且整体系数较为显著，训练样本的拟合程度也较好。从回归系数分析得到：孕期母亲体重减少会增大婴儿出生低体重的可能性，但母亲的体重增加过多也会对婴儿的健康不利，孕期母亲的体重增幅控制在0-10 磅之间是最为合适的；从母亲的肤色角度看，黑人母亲对较低体重婴儿的概率平均是白人母亲低体重婴儿的1.75 倍，且系数在置信度0.1下是显著的，可见黑人孕妇的总体营养及卫生状况亟待改善。

后续讨论题

1. 尝试对母亲的受教育程度进行分析，并将其与母亲是否已婚纳入回归模型，做进一步的分析。

2. 尝试使用对数线性模型进行分析，分析比较泊松回归于对数线性模型所得结果的异同。