· 计算机应用 ·

应用 R 软件的 poLCA 程序包实现潜在类别分析

武汉大学中南医院心血管内科(430071)　冯雨嘉　李　曙　胡思汗　万　静△

【提　要】 目的　使用 R 软件的 poLCA 程序包进行潜在类别分析。**方法**　本文将以 Dayton《Latent Class Scaling Analysis》中的数据为例来说明如何运用 R 软件的 poLCA 程序包进行潜在类别分析(latent class analysis,LCA),并展示相关代码运行后的结果。**结果**　示例数据被分为两个潜在类别,各类别的概率分别为 0.839、0.161。**结论**　与传统的 SAS、STATA、Mplus 等软件相比, R 软件中的 poLCA 程序包可以在最佳潜在类别数目未知的情况下顺次进行多个潜在类别数目的循环并输出最优模型,同时还可以进行带有协变量的潜在类别分析,为研究者提供了一种快速又简单便捷的潜在类别分析方法。

【关键词】 poLCA 程序包　潜在类别分析　R 软件　潜在聚类分析
【中图分类号】 R54　　【文献标识码】 A　　**DOI**　10.11783/j.issn.1002-3674.2024.02.035

在潜在类别模型(latent class model,LCM)的基础上,出现了潜在类别分析(latent class analysis,LCA)这一统计方法,此方法可用于识别群体中具有某些外部特征的不同亚群,这些亚群被称为潜在类别。潜在类别分析旨在通过潜在类别变量来解释外显变量之间的关联,进而维持外显变量的局部独立性[1]。潜在类别分析在心理学、医学、人文社会科学等领域广泛使用[2-5]。R 软件具备强大的数据处理及绘图功能,其下载及使用均免费[6]。poLCA 程序包是 R 软件专门用来做潜在类别分析的程序包,由 Drew Linzer 和 Jeffrey Lewis 等人研发[7],版本为 1.6.0,该程序包采用的算法是两种基于最大似然估计的迭代算法,分别为 EM(expectation maximization)[8] 和 NR(newton-raphson)算法。本文以 Dayton《Latent Class Scaling Analysis》中的数据为例来演示 poLCA 程序包的使用方法[9]。

模型基本原理

当我们使用其他统计学方法进行分析时,容易忽略变量间的关联性。潜在类别分析的基本假设是,可以用少数互斥的潜在类别变量来解释各外显变量之间的关系,使得外显变量能够维持其局部独立性[10]。假设有 A、B、C 三个外显变量,分别具有 i、j、k 个水平数,其彼此之间不相互独立。若存在一具有 t 个潜在类别的潜变量 X,其不仅可以解释 A、B、C 三者间的关系,且在 X 的每个类别中,A、B、C 这三个外显变量能够维持局部独立性,即为潜在类别分析,其数学模型为:

$$\pi_{ijk}^{ABC}=\sum_{t=1}^{T}\pi_{t}^{X}\pi_{it}^{\bar{A}X}\pi_{jt}^{\bar{B}X}\pi_{kt}^{\bar{C}X}$$

式中 π_{ijk}^{ABC} 表示一个潜在类别模型的联合概率,π_{t}^{X} 为潜在类别概率(class membership probabilities),它表示当观察变量局部独立时,潜变量 X 在第 t 个水平的概率,即各潜类别的人数占总体的比例,各潜在类别概率之和等于 1。$\pi_{it}^{\bar{A}X}$ 为条件概率(item response probabilities),表示属于第 t 个潜在类别的个体对观察变量 A 的第 i 个水平作出反应的概率,即潜类别组内的个体在外显变量上的作答概率,类似于因子分析中的因子载荷,反映潜变量与外显变量间关系的强弱。我们可以根据条件概率,命名和解释各潜类别。各外显变量的条件概率总和等于 1[1, 11]。

poLCA 程序包的加载

在 R 软件中输入下列命令进行程序包的安装及加载:

```
install.packages("poLCA")
install.packages("scatterrplot3d")
library("scatterplot3d")
library("poLCA")
```

此处同时安装 scatterrplot3d 程序包是由于 poLCA 程序包绘图需要 scatterrplot3d 程序包的辅助,在弹出的对话框中选择某个镜像安装(CRAN),安装完成之后使用上述命令加载两个程序包即可。

数据加载

本文的实例分析采用程序包的示例数据 cheating.csv 来进行示例演示[9],该示例数据显示的是 319 名大学生的 GPA 以及有无 4 种作弊行为,该数据集共有 5 个变量,分别为“LIEEXAM”、“LIEPAPER”、“FRAUD”、“COPYEXAM”、“GPA”,均为分类变量,分别表示:①为了逃避考试而撒谎(LIEEXAM);②因不能按时交论文而撒谎(LIEPAPER);③购买一篇论文作为自己的论文上交,或者在考试前获得了一份试卷的副本(FRAUD);④考试时抄袭邻座同学的答案

△通信作者:万静,E-mail:wanjing_zn@163.com

(COPYEXAM)；⑤平均学分绩点(grade point average, GPA)。其中“GPA”分为五类：①≤2.99；②3.00～3.25；③3.26～3.50；④3.51～3.75；⑤3.76～4.00。有4名学生无法获取GPA。具体数据见http://github.com/fengyujia88/poLCA-。需要注意的是，潜在类别分析只针对分类变量，若数据中有连续变量，则需转化为分类变量或使用潜在剖面分析(latent profile analysis, LPA)。另外，数据集中不能包含0、小数及负值，若为二分类变量，则需将“0”、“1”改为“1”、“2”。使用指令data(cheating)加载数据，加载数据后可以使用指令view(cheating)来查看数据，还可以使用指令edit(cheating)来编辑数据，若要导入自己的数据集，可以使用指令data<-read.csv(file.choose())。

数据分析

加载完数据集之后即可开始数据分析，poLCA程序包的数据分析功能主要通过“poLCA()”函数进行，该函数可完成无条件的潜在类别分析以及带有协变量的潜在类别分析。下面以示例演示使用该程序包进行数据分析的过程。

1.无条件潜在类别分析

若潜在类别数量已知，则可使用下列命令进行无条件潜在类别分析：

```
f <- cbind(LIEEXAM, LIEPAPER, FRAUD, COPYEXAM)~1
c2<-poLCA(f, cheating, nclass=2, maxiter=1000, graphs=TRUE, tol=1e-10, na.rm=TRUE, probs.start=NULL, nrep=1, verbose=TRUE, calc.se=TRUE)
```

命令中，“LIEEXAM”、“LIEPAPER”、“FRAUD”、“COPYEXAM”为数据集中用于进行潜在类别分析的外显变量；~1表示此为不带协变量的无条件潜在类别分析；cheating为数据集；nclass为潜在类别数量；maxiter为最大迭代次数；graphs为是否在命令运行结束时对结果进行图形化的展示，默认值为FALSE；tol为一个用于判断何时达到收敛的容错值，当一次迭代的对数似然函数值(log-likelihood)变化小于tol值时，考虑已找到最大对数似然函数值，算法终止；na.rm为处理缺失值的方式，TRUE表示缺失值在进行分析前已从数据集中删除(整行删除)，FALSE表示保留缺失值，na.rm的默认值为TRUE；probs.start为潜在类别的顺序，默认值为NULL，表示潜在类别顺序随机；nrep为模型估计的次数，将nrep设置为大于1的值则会自动寻找对数似然函数的全局最大值，而不仅仅是局部最大值；verbose为是否输出模型拟合的结果，默认值为TRUE；calc.se为是否计算条件概率的标准误，默认值为TRUE。上述命令的结果见图1、图2，其中class 1/2表示潜在类别1/2在各外显变量上的作答概率，即条件概率，Pr(1)表示未经历过外显变量事件的概率，Pr(2)表示经历过外显变量事件的概率，根据条件概率，可对各类别进行命名和解释，例如我们发现类别1中大部分同学未有过作弊行为，类别2中大部分同学曾有过作弊行为，那么我们可以据此将类别1命名为非作弊者组，类别2命名为作弊者组；Estimated class population shares为各潜类别的人数占总体的比例，即潜在类别概率；两个类别潜在类别分析的条件概率见表1。结果中基于信息理论的赤池信息准则(akaike information criteria, AIC)、贝叶斯信息准则(Bayesian information criteria, BIC)、G^2值χ^2值为评价模型拟合程度的指标，AIC、BIC、G^2值χ^2值越小，表示模型拟合程度越好。

```
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)

$LIEEXAM
           Pr(1)  Pr(2)
class 1:  0.4231 0.5769
class 2:  0.9834 0.0166

$LIEPAPER
           Pr(1)  Pr(2)
class 1:  0.4109 0.5891
class 2:  0.9708 0.0292

$FRAUD
           Pr(1)  Pr(2)
class 1:  0.7840 0.2160
class 2:  0.9629 0.0371

$COPYEXAM
           Pr(1)  Pr(2)
class 1:  0.6236 0.3764
class 2:  0.8181 0.1819

Estimated class population shares
 0.1606 0.8394

Predicted class memberships (by modal posterior prob.)
 0.1693 0.8307

=========================================================
Fit for 2 latent classes:
=========================================================
number of observations: 319
number of estimated parameters: 9
residual degrees of freedom: 6
maximum log-likelihood: -440.0271

AIC(2): 898.0542
BIC(2): 931.9409
G^2(2): 7.764242 (Likelihood ratio/deviance statistic)
X^2(2): 8.323399 (Chi-square goodness of fit)
```

图1 无条件潜在类别分析结果图

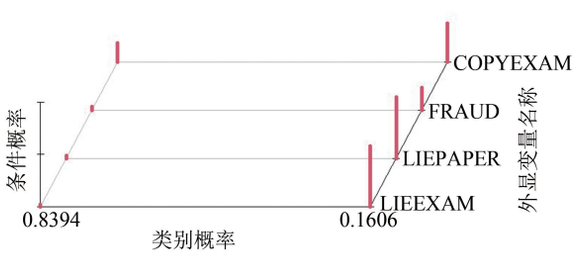

图2 无条件潜在类别分析类别概率及条件概率结果图

表1 两个类别潜在类别分析的条件概率

外显变量	潜变量	
	类别1(%)(非作弊者组)	类别2(%)(作弊者组)
LIEEXAM	57.7	1.7
LIEPAPER	58.9	2.9
FRAUD	21.6	3.7
COPYEXAM	37.6	18.2

若潜在类别数量未知，则需循环多个可能数量的潜在类别，并最终输出最优模型，可使用下列命令进行循环：

```
f <- cbind(LIEEXAM, LIEPAPER, FRAUD, COPYEXAM)~1
max_II <- -100000
min_bic <- 100000
for(i in 1:5){
  lc <- poLCA::poLCA(f, cheating, nclass = i, maxiter = 1000,
      tol = 1e-10, na.rm = TRUE,
      nrep = 10, verbose = TRUE, calc.se = TRUE)
  if(lc$bic < min_bic){
    min_bic <- lc$bic
    LCA_best_model <- lc
  }
}
LCA_best_model
```

命令中各参数所代表含义如前述，*i* in 1:5 表示从一个潜在类别模型自动循环到五个潜在类别模型，每个潜在类别模型拟合结束后均会输出一个的结果，5 个模型的拟合程度指标见表 2。LCA_best_model 表示输出最优模型，需要注意的是，该命令语句是基于 BIC 值最小输出的最优模型，最终模型的确定依然需要综合 AIC 值、G^2 值 χ^2 值以及专业需要等各方面考虑。

表 2　不同潜在类别模型的拟合程度指标

类别数	AIC	BIC	G^2	χ^2	类别概率(%)
1	942.876	957.937	62.586	136.342	-
2	898.054	931.941	7.764	8.323	16.1/83.9
3	900.471	953.184	0.181	0.182	6.4/88.7/4.9
4	910.290	981.829	<0.001	<0.001	12.5/72.0/3.9/11.6
5	920.290	1010.655	<0.001	<0.001	51.6/3.6/14.1/26.2/4.5

注：AIC：赤池信息准则；BIC：贝叶斯信息准则；G^2：似然比卡方；χ^2：Pearson 检验卡方值。

2.带有协变量的潜在类别分析

若要进行带有协变量的潜在类别分析，则可使用下列命令进行分析：

```
f2 <- cbind(LIEEXAM, LIEPAPER, FRAUD, COPYEXAM)~GPA
ch2c <- poLCA(f2, cheating, nclass = 2, maxiter = 1000, graphs = TRUE, tol = 1e-10, na.rm = TRUE, probs.start = NULL, nrep = 1, verbose = TRUE, calc.se = TRUE)
probs.start <- poLCA.reorder(ch2c$probs.start, order(ch2c$P, decreasing = TRUE))
```

命令中各参数含义如前述，其中~GPA 表示协变量为 GPA，若有多个协变量，则可在~之后输入多个协变量，例如 cbind(dv1,dv2,dv3)~iv1+iv2+……+ivn；poLCA.reorder 命令表示固定前一步潜在类别分析的数字顺序(类别 1、类别 2)，ch2c 为上一步潜在类别分析的结果，P 为各类别比例的大小，decreasing 表示降序排列；该命令的结果如图 3 所示，由结果可知 $p=0.030$，$\beta=-0.842$，提示以类别 1 作为参照，GPA 会影响个体的类别归属，且归属到类别 2 的可能性较小。需要注意的是，此命令是使用单步法进行分析，协变量的纳入和剔除会影响模型估计的结果，若协变量较多，则模型估计结果可能会产生偏差。

```
Conditional item response (column) probabilities,
 by outcome variable, for each class (row)

$LIEEXAM
           Pr(1)  Pr(2)
class 1:  0.9903 0.0097
class 2:  0.4389 0.5611

$LIEPAPER
           Pr(1)  Pr(2)
class 1:  0.9647 0.0353
class 2:  0.4858 0.5142

$FRAUD
           Pr(1)  Pr(2)
class 1:  0.9655 0.0345
class 2:  0.7850 0.2150

$COPYEXAM
           Pr(1)  Pr(2)
class 1:  0.8257 0.1743
class 2:  0.5925 0.4075

Estimated class population shares
 0.8219 0.1781

Predicted class memberships (by modal posterior prob.)
 0.8508 0.1492

=========================================================
Fit for 2 latent classes:
=========================================================
2 / 1
             Coefficient  Std. error  t value  Pr(>|t|)
(Intercept)      0.11343     0.50992    0.222     0.833
GPA             -0.84248     0.28131   -2.995     0.030
=========================================================
number of observations: 315
number of estimated parameters: 10
residual degrees of freedom: 5
maximum log-likelihood: -429.6384

AIC(2): 879.2768
BIC(2): 916.8025
X^2(2): 8.641731 (Chi-square goodness of fit)
```

图 3　带有协变量的潜在类别分析结果

进行完带有协变量的潜在类别分析后，可使用下列命令输出协变量 GPA 预测学生作弊行为的图(图 4)：

```
GPAmat <- cbind(1, c(1:5))
exb <- exp(GPAmat %*% ch2c$coeff)
matplot(c(1:5), cbind(1/(1+exb), exb/(1+exb)), type = "l", lwd = 2,
    main = "GPA as a predictor of persistent cheating",
    xlab = "GPA category, low to high",
    ylab = "Probability of latent class membership")
text(1.7, 0.3, "作弊者组")
text(1.7, 0.7, "非作弊者组")
```

命令中，c(1:5)代表 GPA 的五个类别；exp(y)表示 e^y；ch2c 为上一步潜在类别分析的结果；$coeff 表示提取 ch2c 中 coeff 变量的结果；matplot 命令为绘图

命令，其中 type = "l" 表示绘制实线图，lwd 表示线条宽度（默认值为 1），main 表示图的标题，xlab 表示横坐标名称，ylab 表示纵坐标名称。

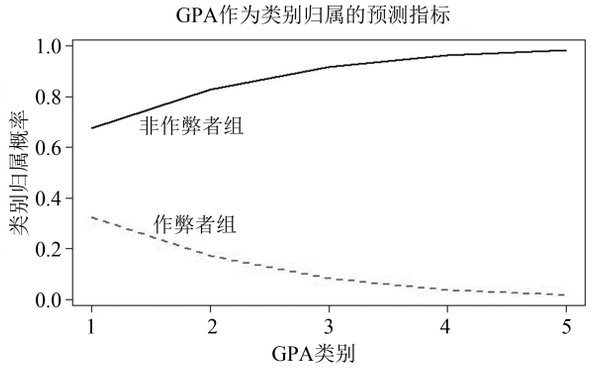

图 4 GPA 预测学生类别归属

3.计算后验概率

可使用如下命令计算每个个体的后验概率：

round(ch2c $ posterior[1:315,],2)

其中 round( ) 命令表示将结果四舍五入，2 表示保留两位小数；ch2c 为之前潜在类别分析的结果；[1:315,] 表示 1～315 个个体，若 na.rm = FALSE，则应有 319 个个体。该命令结果如表 3 所示（由于结果过长，本文只展示前 10 个个体的结果），其中[1,] 表示第一个个体，以此类推，[,1] 表示该个体分到类别 1 的概率，[,2] 表示该个体分到类别 2 的概率。

表 3 后验概率计算结果

	[,1]	[,2]
[1,]	0.94	0.06
[2,]	0.94	0.06
[3,]	0.94	0.06
[4,]	0.94	0.06
[5,]	0.94	0.06
[6,]	0.94	0.06
[7,]	0.94	0.06
[8,]	0.94	0.06
[9,]	0.94	0.06
[10,]	0.94	0.06

4.计算每种组合的人数

可通过下列命令计算出 4 个外显变量应答的不同组合的人数：

ch2c $ predcell

命令中 ch2c 为之前潜在类别分析的结果，该命令结果如表 4 所示，其中 1 表示未经历过外显变量事件，2 表示经历过外显变量事件，“人数”表示该组合的人数。“预期”表示通过密度估计计算出此种组合预期的人数。

讨 论

目前可用于潜在类别分析的软件有 Mplus、STATA、SAS 等[12-13]，R 软件作为一款具有强大统计分析及绘图功能的免费软件，受到研究者们的普遍欢迎。

表 4 四个外显变量应答的不同组合的人数

	LIEEXAM	LIEPAPER	FRAUD	COPYEXAM	人数	预期
1	1	1	1	1	203	202.711
2	1	1	1	2	46	45.457
3	1	1	2	1	7	8.572
4	1	1	2	2	5	2.536
5	1	2	1	1	13	13.112
6	1	2	1	2	4	5.576
7	1	2	2	1	1	1.872
8	1	2	2	2	2	1.164
9	2	1	1	1	10	9.051
10	2	1	1	2	3	5.301
11	2	1	2	1	1	2.017
12	2	1	2	2	2	1.354
13	2	2	1	1	11	7.600
14	2	2	1	2	4	5.193
15	2	2	2	1	1	2.065
16	2	2	2	2	2	1.419

本文使用实例展示了如何使用 R 软件的 poLCA 程序包进行潜在类别分析。通过本文的示例可见 R 软件相比传统的 Mplus、STATA、SAS 等软件具有以下优点：①分析速度快，相较于其他软件，对于较大的数据集，可以快速计算出结果；②操作简单，其他统计分析软件需要手动进行多个数量潜在类别的拟合，而 R 软件可以自动循环多个可能数量的潜在类别，并输出最佳模型；③R 软件具有强大的绘图功能，用户可根据自己的需求对图片进行修改。

另外 poLCA 程序包也存在一些缺点，例如每次分析的潜在类别的数字顺序都是随机的，只能通过 poLCA.reorder( ) 命令固定类别的数字顺序，另外，此程序包也无法进行带有结局变量的潜在类别分析，其功能还需不断完善，才能满足用户的需求。

综上所述，R 软件的 poLCA 程序包分析速度快，操作简单，且能根据用户需求输出优质的结果图，对数据量较大，不希望手动进行多个数量潜在类别拟合的研究者而言显然是更好的选择。

参 考 文 献

[1] 曾宪华，肖琳，张岩波. 潜在类别分析原理及实例分析. 中国卫生统计，2013，30(6)：815-817.

[2] 张洁婷，焦璨，张敏强. 潜在类别分析技术在心理学研究中的应用. 心理科学进展，2010，18(12)：1991-1998.

[3] 刘爱楼，王瑞明. 大学生抑郁情绪特征的潜在类别分析. 中国临床心理学杂志，2022(5)：1208-1212.

[4] 张振香，何福培，张春慧，等. 慢性病共病患者服药依从性潜在类别及其影响因素分析. 中国全科医学，2022，25(31)：3904-3913.

[5] 田超，曹正国，韩庆杰，等. 基于潜在类别分析的不同 PSA 水平前列腺癌患者术后生化复发影响因素分布. 临床泌尿外科杂志，2022，37(7)：537-542.

（下转第 315 页）