变量变换回归分析(Ⅳ)——偏好评分资料的结合分析法

胡良平[1,2*]

(1.军事科学院研究生院,北京 100850;
2.世界中医药学会联合会临床科研统计学专业委员会,北京 100029
*通信作者:胡良平,E-mail:lphu812@sina.com)

【摘要】 本文目的是介绍偏好评分资料的数据结构及其对应的结合分析方法。产生此类资料的场合类似于“多因素析因设计或正交设计”,但计量结果变量的取值在一定程度上受到评价者主观或偏好的影响。结合分析模型是基于各属性(或因素)的“分值效用或水平效用”可以“简单叠加”的假定成立的条件下构造出来的,当实际问题符合此假定时,其分析结果是正确的;否则,要慎重使用。必要时,需要选择其他统计模型。本文通过一个实例,并利用 SAS 中 TRANSREG 过程演示实现结合分析的详细步骤。

【关键词】 属性;析因设计;正交设计;偏好评分;结合分析
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2019.03.004

Regression analysis based on the variable transformation(Ⅳ)——the conjoint analysis method of the data with preference scores

Hu Liangping[1,2*]

(1. *Graduate School, Academy of Military Sciences PLA China, Beijing* 100850, *China*;
2. *Specialty Committee of Clinical Scientific Research Statistics of World Federation of Chinese Medicine Societies, Beijing* 100029, *China*
* *Corresponding author: Hu Liangping, E-mail: lphu812@sina.com*)

【**Abstract**】 The purpose of this paper was to introduce the data structure of the preference data and its corresponding analysis method called the conjoint analysis. The situations which could produce such kind of the data mentioned before were similar to "the factorial design or orthogonal design of the multi-factors". The value of the measurement variables, however, could be affected by the subjectivity or preference of the valuators. The conjoint analysis model was set up under the condition of the assumption that the "Part-Worth Utility or Level-Worth Utility" could be simply superimposed. When the actual problem was conformed the assumption mentioned before, the analyzed results was correct, otherwise the conjoint analysis should be used with caution. The other statistical model should be selected when it was necessary. The paper showed that detailed steps of performing the conjoint analysis by using the TRANSREG procedure in SAS through a real example.

【**Keywords**】 Attribute; Factorial design; Orthogonal design; Preference score; Conjoint analysis

1 基本概念

1.1 偏好评分

在心理、生理和精神卫生等学科领域中,研究者常使用各种量表对患者进行评定[1-4],此类评分一般被视为“计量资料”。然而,在很多其他领域,研究者常基于自己的知识、经验、感受和偏好,来给被评价对象(以下简称为“被评者”)评分,此即“偏好评分”。这里的“被评者”可以是人(如某病患者、参加比赛的歌手、前来应聘的求职者、课题或项目的申请者等)、物品或商品(如电脑、汽车、住宅或服装等)等。评价者进行评分时,通常依据被评者在若干方面或属性或因素的表现或真实情况,并结合自己的“偏好”,给出一个自己认为最合理的分值。由此可知,“偏好评分”不像试验结果那样客观、精准,而在一定程度上带有主观性。这就不难理解,为什么同一位患者的影像学光片由多位放射科医生来阅读,通常会给出不同的评分结果。

1.2 因素及水平

如何区分“被评者 A”与“被评者 B”?通常需要根据具体情况,从几个主要方面或角度来度量或认定被评者,用统计学的术语来描述,就是拟考虑的“因素”。以一位患者的 CT 光片为例,假定要考虑的主要因素有 A、B、C、D,每个因素又可分为好、中、差 3 个档次。于是,这 4 个因素都具有 3 个水平。请一位临床医生基于前述 4 个因素来给 100 例某病

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

患者的 CT 光片进行病情评分,病情由轻到重依次评为 1、2、……、10 分。这样就可以获得对这 100 例患者的“偏好评分”。这里的“偏好”主要反映了该临床医生在阅片方面的经验和技能,显然不可能像定量检测样品中某种物质含量那样精准。

1.3 偏好评分的性质

通常,偏好评分有两种具体形式,其一,仅用很少的几个分值来描述所有的“被评者”,如用“1、2、3、4、5”5 个分值来描述 100 例某病患者;其二,使用 1 ~ 81 共 81 个不相同的分值来描述 81 例某病患者。前一种“偏好评分”只能被视为“有序资料”;而后一种“偏好评分”可以被视为“有序资料”,也可被近似视为“计量资料”。

1.4 偏好评分资料的数据结构

在一项关于轮胎的消费情况的调查中,假定轮胎主要属性(或因素)有 4 个:品牌、价格、使用寿命、有无公路意外保险计划,各属性包含不同水平[5]。见表 1。

表 1 轮胎的属性及其水平

属 性	各属性的具体水平			
	水平代码:	1	2	3
品牌		GOODSTONE	PIROGI	MACHISMO
价格($)		69.99	74.99	79.99
使用寿命(km)		50 000	60 000	70 000
公路意外保险计划		有	无	–

若将各属性的所有水平进行组合,一共有 3 ×3 ×3 ×2 =54 种可能的组合形式(相当于有 54 种不同的轮胎或 54 种不同的试验条件;在多因素试验设计中,每种“组合”被称为一个“试验点”)。最好的做法是采用析因设计,即在 54 个试验点的每个试验点上至少做 2 次独立重复试验。那么总试验次数至少需要 108 次。为了节约成本,研究者拟采用正交设计。利用 SAS 软件产生一个组合数为 18 的正交设计表。见表 2。其中最后两列为两名顾客(即在各试验点上均做了 2 次独立重复试验)给出的偏好评分,用 1 ~18 来表示其偏好(1 表示最愿意购买,18 表示最不愿意购买)。

表 2 具有 18 种水平组合的混合水平正交表

组合编号	品牌	价格($)	使用寿命(km)	公路意外保险计划	偏好评分(顾客 A)	偏好评分(顾客 B)
1	GOODSTONE	69.99	60 000	有	3	4
2	GOODSTONE	69.99	70 000	无	2	1
3	GOODSTONE	74.99	50 000	无	14	15
4	GOODSTONE	74.99	60 000	无	10	10
5	GOODSTONE	79.99	50 000	有	17	17
6	GOODSTONE	79.99	70 000	有	12	14
7	PIROGI	69.99	50 000	无	7	8
8	PIROGI	69.99	70 000	无	1	2
9	PIROGI	74.99	50 000	有	8	7
10	PIROGI	74.99	70 000	有	5	3
11	PIROGI	79.99	60 000	有	13	12
12	PIROGI	79.99	60 000	无	16	16
13	MACHISMO	69.99	50 000	有	6	5
14	MACHISMO	69.99	60 000	有	4	6
15	MACHISMO	74.99	60 000	无	15	11
16	MACHISMO	74.99	70 000	有	9	13
17	MACHISMO	79.99	50 000	无	18	18
18	MACHISMO	79.99	70 000	无	11	9

由表2可知,偏好评分资料的数据结构由两种性质的变量及其取值构成:定性的影响因素(表2中第2~5列)和偏好评分(表2中最后2列)。就整体而言,它是一个多因素析因设计或正交设计(或均匀设计)下所收集的一元或多元有序资料。

由于未找到合适的临床实例,故借用上面关于"商品"的例子。若临床工作者能在临床工作中发现与此实例类似的"临床问题",可以采用本文介绍的"结合分析法"进行数据处理。

1.5 偏好评分资料的分析任务及分析方法

针对每一位顾客(即评价者)给出的"评分",需要回答以下三个问题:其一,4个因素的"重要性(或贡献率)"分别是多少(重要性之和为100.00%)?其二,每个因素各水平的"分值效用"(即每个水平的"重要性",确切的含义为"水平效用")是多少?其三,该顾客最偏爱或喜欢的轮胎是哪一款(即理想试验点)?为实现前述的统计分析任务,需要选择"结合分析法"。

2 结合分析简介[6-8]

2.1 何为结合分析

结合分析也叫联合分析,它用于确定哪些产品(或服务)的属性(或因素)对于消费者(或评价者)来说是最重要的、哪些是中等重要的、哪些是次要的;还可用于估计每种属性(或因素)的每个水平的"效用(即对偏好评分的作用)"大小。

2.2 结合分析的基本思想

结合分析的基本思想是:将偏好评分近似视为计量因变量,将每个属性的每个水平视为一个"二值自变量"。假定每个属性的所有水平对因变量的影响是可以叠加的,进一步假设每个属性所有水平效应之和为0(这在统计学上被称为"约束条件",以保证计算出的"效用值"有正有负,代表不同的作用方向)。在前述假设成立的条件下,构建多重线性回归模型,并基于最小平方法原理求解回归模型中的参数估计值。

2.3 结合分析的基本模型

结合分析中通常采用普通最小平方法估计回归模型中的参数,因素的每个水平被视为一个自变量,并且,每个自变量只有0或1两个可能的取值。模型可用式(1)表示。

$$Y = a + \sum vx \quad (1)$$

在式(1)中,Y表示所有属性(或因素)的一种水平组合条件下被评价对象的总效用,也被称为"轮廓(即一个试验点)"的总效用。a为截距,v为各水平的分值效用(相当于回归系数),x为取值为0或1的哑变量,当它代表的某属性的一个水平出现时,$x = 1$;否则$x = 0$。

若模型中属性水平的分值效用的差值(最大效用与最小效用之差)越大,则该属性的相对重要性越高。一般用百分比来描述各属性的重要性。见式(2)。

$$W_j = \frac{\max(v_j) - \min(v_j)}{\sum_{j=1}^{m}[\max(v_j) - \min(v_j)]} \times 100\% \quad (2)$$

在式(2)中,m表示属性个数,W_j表示第j个属性的相对重要性,$\max(v_j)$和$\min(v_j)$分别表示第j个属性各水平中最大和最小的分值效用。

3 实例的SAS实现

3.1 对实例的进一步解说

在前面的"实例"中,表1列出了拟考察的"属性(或因素)及其水平",表2是将这个实际问题付诸研究所给出的一种"试验设计"(第2~5列)及其两位顾客给出的"偏好评分"(最后两列)。表2中的每一行代表4个因素各取一个水平所对应的一种组合(也叫试验点),也就是一款特定的产品(在本例中为一种轮胎)。统计分析的目的是希望依据某位顾客的"偏好评分",回答前述"偏好评分资料的分析任务"中提及的问题。

3.2 分析实例所需要的SAS程序

```
options validvarname = any;
proc format;
value brandf  1 = ´goodstone´  2 = ´pirogi´  3 = ´machismo´;
value pricef  1 = ´ $69.99´  2 = ´ $74.99´  3 = ´ $79.99´;
value lifef  1 = ´50,000´  2 = ´60,000´  3 = ´70,000´;
value hazardf  1 = ´yes´  2 = ´no´;
run;
data tires;
input brand price life hazard rank1 rank2 @@;
format brand brandf9. price pricef9.
       life lifef6. hazard hazardf3.;cards;
```

1	1	2	1	3	4
1	1	3	2	2	1
1	2	1	2	14	15
1	2	2	2	10	10
1	3	11	17	17	
1	3	3	1	12	14
2	1	1	2	7	8
2	1	3	2	1	2
2	2	1	1	8	7
2	2	3	1	5	3
2	3	2	1	13	12
2	3	2	2	16	16
3	1	1	1	6	5
3	1	2	1	4	6
3	2	2	2	15	11
3	2	3	1	9	13
3	3	1	2	18	18
3	3	3	2	11	9

;

```
run;
proc transreg utilities cprefix=0 lprefix=0;
ods select convergencestatus fitstatistics utilities;
model identity(rank1 rank2/reflect) =
      class(brand price life hazard/zero=sum);
output out=out replace predicted;
run;
proc print label data=out;
var rank1 rank2 prank1prank2 brand price life hazard;
run;
```

【说明】“model 语句”中的选项“reflect”的含义:代表各水平取正的“分值效用”时,对应着最好的“偏好评分”。这样可以免去使用者在下专业结论时,需要顾及结果变量究竟属于高优指标还是低优指标。

若忽略了选项“reflect”,当结果变量为高优指标时,需要选取绝对值最大的正效用值对应的水平组成“理想试验点”;而当结果变量为低优指标时,需要选取绝对值最大的负效用值对应的水平组成“理想试验点”。

3.3 主要输出结果及解释

因篇幅所限,下面仅给出第一位顾客偏好评分对应的结合分析结果:

Root MSE	1.72562	R-Square	0.9385
Dependent Mean	9.50000	Adj R-Sq	0.8955
Coeff Var	18.16446		

以上结果表明:模型对资料的拟合优度较高,均方根误差 = 1.72562、R^2 = 0.9385。与“分值效用(Utility)”和“重要性(Importance)”有关的计算结果见图 1。

Utilities Table Based on the Usual Degrees of Freedom

Label	Utility	Standard Error	Importance (% Utility Range)	Variable
Intercept	9.5000	0.40673		Intercept
goodstone	-0.1667	0.57521	10.986	Class.goodstone
pirogi	1.1667	0.57521		Class.pirogi
machismo	-1.0000	0.57521		Class.machismo
$69.99	5.6667	0.57521	54.085	Class.$69.99
$74.99	-0.6667	0.57521		Class.$74.99
$79.99	-5.0000	0.57521		Class.$79.99
50,000	-2.1667	0.57521	25.352	Class.50,000
60,000	-0.6667	0.57521		Class.60,000
70,000	2.8333	0.57521		Class.70,000
yes	0.9444	0.40673	9.577	Class.yes
no	-0.9444	0.40673		Class.no

图 1 与表 2 中顾客 A 的偏好评分对应的计算结果

由图 1 中第 4 列计算结果可知:轮胎的 4 个属性的相对重要性为:价格 > 使用寿命 > 品牌 > 是否有公路意外保险计划。价格方面,越便宜越受顾客偏好;使用寿命方面,越长越受顾客偏好;品牌方面,顾客最偏好的是 pirogi;是否有公路意外保险计划方面,顾客更偏好有保险计划的轮胎。最受顾客欢迎的轮胎属性组合为“品牌 pirogi + 使用寿命 70 000km + 价格 $69.99 + 有公路意外保险计划”,它们都是各属性中“分值效用”取最大正值的“水平”。

4 讨论与小结

4.1 讨论

4.1.1 式(1)的解读

结合分析的“回归模型”式(1)的真实含义,就是把“偏好评分值”视为“计量因变量”,而把所有的“属性(或因素)”视为“定性自变量”,但需要对每个属性变量产生哑变量。值得注意的是:在对每个属性变量产生哑变量时,不采取通常的方法(以其中一个水平为基准),而是将其每个水平产生一个“0 与 1”的二值变量,但必须限制该属性的所有水平对应的哑变量之和等于 0。例如:对于“品牌”这个属性变量而言,由表 2 的第 2 列可知,第 1 ~ 6 行都是第 1 种品牌“GOODSTONE”,若用 TB1 代表它,则 TB1 = 1,其后的 12 行都不是该品牌,故 TB1 = 0;同理,可用 TB2 代表第 2 种品牌,则在第 7 ~ 12 行令

TB2 = 1,其他行上令 TB2 = 0;可用 TB3 代表第 3 种品牌,则在第 13 ~ 18 行令 TB3 = 1,其他行上令 TB3 = 0。于是,就将一个具有 3 水平的“品牌”转换成 TB1、TB2、TB3 三个“二值变量”了。

类似的,利用以上方法可将“价格”转换成 TP1、TP2、TP3 三个“二值变量”,将“使用寿命”转换成 TL1、TL2、TL3 三个“二值变量”,将“有无公路意外保险计划”转换成 TH1、TH2 两个“二值变量”。于是,可用拟合多重线性回归模型的 REG 过程来实现模型(1)的拟合。

4.1.2 用 REG 过程拟合式(1)

```
/ * 结合模型的构建与参数估计 * /
data abc;
   input a b tb1 tb2 tb3 tp1 tp2 tp3
              tl1 tl2 tl3 th1 th2;
cards;
```

16	15	1	0	0	1	0	0	0	1	0	1	0
17	18	1	0	0	1	0	0	0	0	1	0	1
5	4	1	0	0	0	1	0	1	0	0	0	1
9	9	1	0	0	0	1	0	0	1	0	0	1
2	2	1	0	0	0	0	1	1	0	0	1	0
7	5	1	0	0	0	0	1	0	0	1	1	0
12	11	0	1	0	1	0	0	1	0	0	0	1
18	17	0	1	0	1	0	0	0	0	1	0	1
11	12	0	1	0	0	1	0	1	0	0	1	0
14	16	0	1	0	0	1	0	0	0	1	1	0
6	7	0	1	0	0	0	1	0	1	0	1	0
3	3	0	1	0	0	0	1	0	1	0	0	1
13	14	0	0	1	1	0	0	1	0	0	1	0
15	13	0	0	1	1	0	0	0	1	0	1	0
4	8	0	0	1	0	1	0	0	1	0	0	1
10	6	0	0	1	0	1	0	0	0	1	1	0
1	1	0	0	1	0	0	1	1	0	0	0	1
8	10	0	0	1	0	0	1	0	0	1	0	1

```
;
run;
proc reg data = abc;
model a = tb1 - tb3 tp1 - tp3 tl1 - tl3 th1 th2;
restrict tb1 + tb2 + tb3 = 0, tp1 + tp2 + tp3 = 0,
         tl1 + tl2 + tl3 = 0, th1 + th2 = 0;
run;
```

4.1.3 上述 SAS 程序的主要输出结果

方差分析

源	自由度	平方和	均方	F	Pr > F
模型	7	454.72222	64.96032	21.82	<0.0001
误差	10	29.77778	2.97778		
校正合计	17	484.50000			

均方根误差	1.72562	R^2	0.9385
因变量均值	9.50000	调整 R^2	0.8955
变异系数	18.16446		

以上结果与前面使用“TRANSREG 过程”输出的“拟合优度”结果是相同的。

参数估计值

变量	自由度	参数估计值	标准误差	t	Pr > |t|
Intercept	1	9.50000	0.40673	23.36	<0.0001
tb1	1	-0.16667	0.57521	-0.29	0.7779
tb2	1	1.16667	0.57521	2.03	0.0700
tb3	1	-1.00000	0.57521	-1.74	0.1128
tp1	1	5.66667	0.57521	9.85	<0.0001
tp2	1	-0.66667	0.57521	-1.16	0.2734
tp3	1	-5.00000	0.57521	-8.69	<0.0001
tl1	1	-2.16667	0.57521	-3.77	0.0037
tl2	1	-0.66667	0.57521	-1.16	0.2734
tl3	1	2.83333	0.57521	4.93	0.0006
th1	1	0.94444	0.40673	2.32	0.0426
th2	1	-0.94444	0.40673	-2.32	0.0426
RESTRICT	-1	1.66533E-16	1.28569E-8	0.00	1.0000*
RESTRICT	-1	-9.4369E-16	1.28569E-8	-0.00	1.0000*
RESTRICT	-1	7.12104E-16	1.28569E-8	0.00	1.0000*
RESTRICT	-1	0	0	-	-

以上的输出结果中,除了最后“RESTRICT”所在的4行外,第3列和第4列与前面图1中第2列和第3列是完全一致的。此处,还多出了“t值”和“P值”,但少了关于各属性“重要性”的计算结果。但若利用上面的计算结果代入式(2),就不难计算出“重要性”的数值。例如:

属性	最大值-最小值	重要性(%)
品牌	1.1667-(-1.0000)= 2.1667	2.1667/18.8889=0.114708=11.471%
价格	5.6667-(-5.0000)=10.6667	10.6667/18.8889=0.564707=56.471%
寿命	2.8333-(-2.1667)= 5.0000	5.0000/18.8889=0.264705=26.471%
保险	0.9444-(-0.9444)= 1.8888	1.8888/18.8889=0.099995=10.000%
合计	18.8889	

上面的“重要性”计算结果与前面图1中第4列对应的结果略有出入,可能是每一项“分值效用”输出时仅保留了4位小数,属于“舍入误差”所致。

4.1.4 在第3.2节SAS程序的“model语句”中不使用“reflect”的输出结果

若将前面第3.2节SAS程序“model语句”中的选项“reflect”删除,其他内容不变,与图1对应的输出结果见图2。

将图2与图1对照,仅第2列中各因素的各水平的“分值效用”的正、负号发生了反转,绝对值没有任何改变。此时,若希望找出“理想试验点”,必须弄清楚“偏好评分”属于“高优指标”还是“低优指标”。本例开始就交代了“偏好评分”为“低优指标”,故“理想试验点”应由各属性中“分值效用”取最大绝对值且为负号对应的“水平”组合起来,即品牌pirogi+使用寿命70 000km+价格$69.99+有公路意外保险计划。

Utilities Table Based on the Usual Degrees of Freedom

Label	Utility	Standard Error	Importance (% Utility Range)	Variable
Intercept	9.5000	0.40673		Intercept
goodstone	0.1667	0.57521	10.986	Class.goodstone
pirogi	-1.1667	0.57521		Class.pirogi
machismo	1.0000	0.57521		Class.machismo
$69.99	-5.6667	0.57521	54.085	Class.$69.99
$74.99	0.6667	0.57521		Class.$74.99
$79.99	5.0000	0.57521		Class.$79.99
50,000	2.1667	0.57521	25.352	Class.50,000
60,000	0.6667	0.57521		Class.60,000
70,000	-2.8333	0.57521		Class.70,000
yes	-0.9444	0.40673	9.577	Class.yes
no	0.9444	0.40673		Class.no

图2 与表2中顾客A的偏好评分对应的计算结果(未用reflect选项)

4.2 小结

结合分析模型是基于各属性(或因素)的分值效用可以简单叠加的假定成立的条件下构造出来的,当实际问题符合此假定时,其分析结果是正确的;否则,要慎重使用。必要时,需要选择其他统计模型。

参考文献

[1] 杨慧, 张余, 陈旭义, 等. 三七总皂苷对脊髓损伤大鼠运动功能恢复的作用[J]. 中国应用生理学杂志, 2016, 32(2): 142-145.

[2] 李晶晶, 梅刚, 瞿秋霜, 等. 老年精神分裂症患者血浆脂蛋白磷脂酶 A2 浓度与 BPRS 评分的相关性[J]. 四川精神卫生, 2017, 30(4): 341-344.

[3] 孙磊, 陈清刚, 王莹, 等. 艾司西酞普兰联合正念认知疗法对广泛性焦虑障碍患者临床症状和生活质量的效果[J]. 四川精神卫生, 2018, 31(5): 428-431.

[4] 陈胡丹, 王娟, 王新森, 等. 阻塞性睡眠呼吸暂停患者临床特征与焦虑抑郁的相关分析[J]. 四川精神卫生, 2019, 32(1): 33-37.

[5] SAS Institute Inc. STAT SAS 9.3 User's Guide[M]. Cary, NC: SAS Institute Inc, 2011: 7761-8002.

[6] 李卫东. 应用多元统计分析[M]. 北京: 北京大学出版社, 2008: 289-312.

[7] 何晓群. 多元统计分析[M]. 2 版. 北京: 中国人民大学出版社, 2008: 350-373.

[8] 胡良平. 面向问题的统计学——(2)多因素设计与线性模型分析[M]. 北京: 人民卫生出版社, 2012: 527-540.

(收稿日期:2019-06-12)

(本文编辑:吴俊林)