变量变换回归分析(Ⅲ)——寻找理想试验点的方法

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】 本文目的是介绍一种特殊的处理多因素试验设计一元定量资料差异性分析方法,即结合分析法。通常情况下,处理多因素试验设计一元定量资料应采用方差分析。但是,此方法无法给出各对影响因素"重要性"的评价,也无法给出因素各水平的"效用值",更无法给出"理想试验点"。本文通过对一个实例的全面解析,显示了 SAS 中的 TRANSREG 过程具有很强且多样性的变量变换能力,它集方差分析、回归分析和结合分析于一体,能够很好地处理不符合传统统计学要求的复杂资料,能够实现前述期望达到的目的。

【关键词】 方差分析;回归分析;结合分析;BOX-COX 变换;理想试验点
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2019.03.003

Regression analysis based on the variable transformation(Ⅲ)——the approach of searching the ideal experimental point

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 a special approach of the difference analysis to process the univariate quantitative data collected from the multi-factor experimental design, which was called the conjoint analysis. Normally, ANOVA should be used to deal with the quantitative data of multi-factor design. The variance analysis, however, could not give the evaluation of the "importance" about each of the influence factors and could not output the "Part-Worth Utility" of each level of every factor as well as could not produce the "ideal experimental point". The paper showed that the TRANSREG procedure in SAS had powerful and various abilities of the variable transformation. The procedure could achieved the previous aims, since it assembled the variance analysis and regression analysis and conjoint analysis together, and it could process the complex data of noncompliance with requirements of the traditional statistics.

【**Keywords**】 Analysis of variance; Regression analysis; Conjoint analysis; BOX-COX transformation; Ideal experimental point

1 基本概念

1.1 多因素试验设计类型

在一项试验研究中,通常都会涉及多个试验因素,从每个因素中各取一个水平组合起来,就形成了一个特定的"试验条件"。在有多个试验因素的研究场合中,以不同的方式选取因素的水平组合,就对应着不同的"试验设计类型"。例如,将所有试验因素的水平全面组合且在各种组合条件下进行两次或两次以上独立重复试验,此安排就被称为"析因设计";又例如,依据正交原理从全部水平组合中选取部分水平组合来安排试验,就被称为"正交设计"[1]。以此类推,还有"均匀设计""最优设计"和"正交组合设计"等[2-3]。

1.2 理想试验点

前面所说的每个试验条件常被称为"试验点",也就是说,每个"试验点"实际上就是由拟考察的试验因素各取一个水平的一种组合。若试验结果是定量的,在各试验点上实施试验后,就可以观测到一个或多个具体的数值。在实际问题中,当定量观测结果的取值越大越好时,就称此类定量指标为"高优指标";反之,就称为"低优指标"。若定量指标取中等值为优,这种情况并不多见,不属于本文讨论的

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

范畴。

所谓“理想试验点”,也被称为“最优试验条件”,是指“高优指标”或“低优指标”获得最优取值时所对应的“试验点”或“试验条件”。

1.3 三种分析方法的异同点

本文涉及到三种统计分析方法:方差分析、回归分析、结合分析。一般来说,方差分析的主要目的是考察各因素对定量指标的影响,一方面希望能将全部因素及其交互作用对定量结果的影响分出主次关系,另一方面希望能揭示出每个因素各水平对定量结果影响之间的差异。回归分析的主要目的是构建因变量依赖自变量变化而变化的回归模型,同时筛选出对因变量具有统计学意义的自变量,有时,还需要在给定自变量取不同值的条件下,预测因变量的数值;而结合分析的主要目的是希望给出各因素对“偏好评分”影响大小的“重要性”的度量,同时希望给出每个属性(或因素)的每个水平作用大小的“效用值”的度量。

从上面的介绍似乎可以认为上述三种分析方法是完全不同的,事实上,它们都属于回归模型分析法。因为方差分析模型本质上就是一种简单的回归模型,而结合分析模型实际上是将属性(或因素)的每个水平当作一个“二值自变量”,并基于“效用值”可叠加的假定构建出来的回归模型[4]。

2 一个取自 TRANSREG 过程的样例

2.1 样例的名称与内容

在 SAS/STAT 的 TRANSREG 过程中有一个名为“BOX - COX 变换”的样例:在纺织研究中,纱线的寿命主要受三个试验因素的影响[5]。见表 1。

表 1 影响纱线寿命的三个主要试验因素及其水平

编 号	因素名称	1 水平	2 水平	3 水平
1	A:纱线测试样品的长度(mm)	250(-1)	300(0)	350(1)
2	B:负载循环时的振幅大小(mmd)	8(-1)	9(0)	10(1)
3	C:负载量(g)	40(-1)	45(0)	50(1)

注:表中圆括号之前的数字为因素的“真实水平”,圆括号内为因素的“代码水平”

由表 1 可知,这是一个涉及三个 3 水平试验因素的试验研究,若将 3 个试验因素的水平全面组合,就有 27 种不同的试验条件,每个试验条件下的试验结果为纱线的寿命长短(单位不详),是一个计量变量。研究者在 27 种不同试验条件下都只进行了一次试验,即没有进行重复试验,其试验结果(用 Fail 表示)和 27 种水平组合见表 2。

表 2 三个 3 水平因素水平全面组合条件下纱线寿命数据

编 号	A	B	C	Fail	编 号	A	B	C	Fail
1	-1	-1	-1	674	15	0	0	1	438
2	-1	-1	0	370	16	1	0	-1	442
3	-1	-1	1	292	17	1	0	0	332
4	0	-1	-1	338	18	1	0	1	220
5	0	-1	0	266	19	-1	1	-1	3636
6	0	-1	1	210	20	-1	1	0	3184
7	1	-1	-1	170	21	-1	1	1	2000
8	1	-1	0	118	22	0	1	-1	1568
9	1	-1	1	90	23	0	1	0	1070
10	-1	0	-1	1414	24	0	1	1	566
11	-1	0	0	1198	25	1	1	-1	1140
12	-1	0	1	634	26	1	1	0	884
13	0	0	-1	1022	27	1	1	1	360
14	0	0	0	620					

2.2 结果变量 Fail 的频数分布

表 2 中结果变量 Fail 的频数分布见图 1。由图 1 可知,结果变量 Fail 呈极严重的正偏态分布。

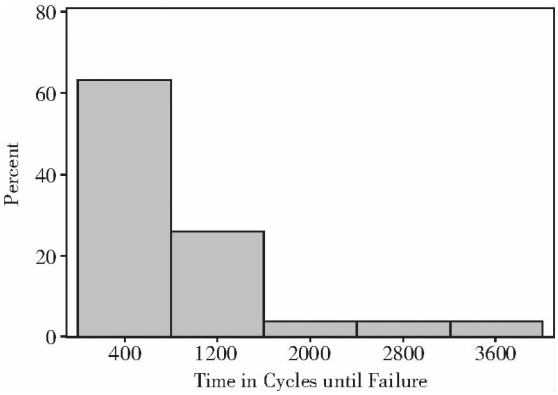

图 1 结果变量 Fail 的频数分布直方图

2.3 需解决的问题及困难

上述样例的专业问题实际上就是一个具有三因素析因设计结构的一元计量资料的统计处理问题。由于在因素各水平组合条件下未进行重复试验,所以,表 2 中的“安排”不能被称为一个“标准的析因设计”,而只能叫做具有“析因结构”。统计处理的困难在于:其一,结果变量偏离正态分布很远;其二,未进行重复试验,样本含量严重不足,无法分析因素之间可能存在的交互作用效应的大小。

2.4 统计分析的任务

一般来说,在多因素试验研究场合,当结果变量为计量变量时,统计分析的任务是研究哪些因素对结果的影响是主要的、哪些是次要的;因素之间各级交互作用的效应大小;有时,研究者还希望求出“理想试验点”,即在多个试验因素分别取什么样的水平组合条件下,所得到的试验结果在专业上是最满意的。就本例而言,在什么样的试验条件下,纱线的寿命最长(它属于“高优指标”)。

2.5 解决上述困难的策略

第一,可以对计量因变量采取 BOX - COX 变换,使其服从或近似服从正态分布[5-6]。第二,可以对定性自变量(即试验因素)采取变量扩展变换,例如“CLASS 变换”或“POINT 变换”或“EPOINT 变换”或“QPOINT 变换”[5]。实际上,前述提及的那些“变量扩展变换”类似于将定性变量数量化,也就是给定性变量的水平重新编码,并引入交互作用项。第三,将原本属于“方差分析的任务”改换为“回归分析任务”,即构建变换后的因变量关于变量扩展变换产生的自变量的回归模型。第四,借助“结合分析”[7-8]的思路和方法,获得各试验因素对结果变量的“重要性”评价及试验因素各水平的“分值效用”大小,得出“理想试验点(即全部因素最佳的水平组合对应的试验条件)”。

3 数据集的形成与上述策略的实现

3.1 数据集的形成

利用以下 SAS 程序,可以形成待分析的 SAS 数据集:

```
proc format;
value  a   -1=80   =9   1=10;
value  l   -1=250  0=300  1=350;
value  o   -1=40   0=45   1=50;
run;
data yarn;
input Fail Amplitude Length Load @@;
format amplitude a. length l. load o.;
label fail = 'Time in Cycles until Failure';
datalines;
 674  -1  -1  -1  370  -1  -1   0   292  -1  -1   1  338   0  -1  -1
 266   0  -1   0  210   0  -1   1   170   1  -1  -1  118   1  -1   0
  90   1  -1   1 1414  -1   0  -1  1198  -1   0   0  634  -1   0   1
1022   0   0  -1  620   0   0   0   438   0   0   1  442   1   0  -1
 332   1   0   0  220   1   0   1  3636  -1   1  -1 3184  -1   1   0
2000  -1   1   1 1568   0   1  -1  1070   0   1   0  566   0   1   1
1140   1   1  -1  884   1   1   0   360   1   1   1
;
run;
```

3.2 显示结果变量 Fail 的频数分布

利用以下 SAS 程序,可以直方图形式呈现结果变量 Fail 的频数分布情况:

```
proc univariate data = yarn normal;
var fail;
histogram fail;
run;
```

以上程序运行的结果如图 1 所示。

3.3 对因变量和自变量进行变量变换

对结果变量 Fail 进行 BOX – COX 变换,同时,对定性自变量进行 QPOINT 变量扩展变换。所需要的 SAS 程序如下:

```
ods graphics on;
proc transreg details data = yarn ss2 utilities
              plots = ( transformation ( dependent )
              obp);
model BoxCox(fail / convenientlambda = -2 to 2 by
      0.05) =
      qpoint(length amplitude load);
output out = aaa approximations;
run;
```

【SAS 程序说明】“proc transreg”调用 TRANSREG 过程;“model 语句”等号左边为对因变量 Fail 进行“BOX – COX 变换”,此变换的一个关键参数叫做“lambda”,经过尝试,需在(-2,2)范围内按步长为 0.05 去选择具体的 lambda 值并代入计算,选择使对数似然函数取最大值时的lambda值。当此值带有很多位小数时,尽可能取一个简约的数值(即“convenient”的含义)。“qpoint 变量扩展变换”是对三个定性变量进行二次反应面变换,即在三个定性变量的基础上,增加它们的平方项和二次交叉乘积项。

3.4 显示对因变量 Fail 进行 BOX – COX 变换的结果

利用以下 SAS 程序,可以直方图形式显示对因变量 Fail 进行 BOX – COX 变换的结果(注:tfail 是对变量 fail 采用 BOX – COX 变量变换后的变量)。

```
proc univariate data = aaa normal;
var tfail;
histogram tfail/normal;
run;
```

以上 SAS 程序的输出结果见图 2:

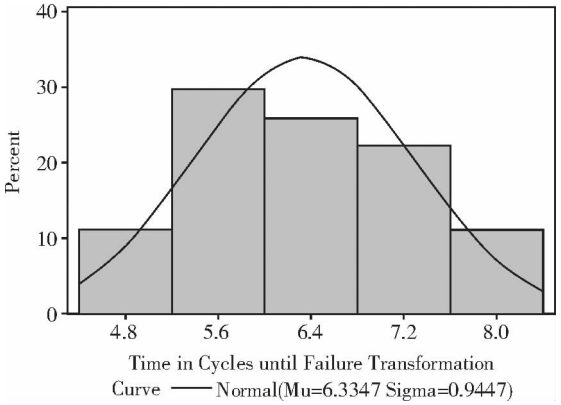

图 2 经过 BOX – COX 变换后的结果变量 tfail 的频数分布直方图

对变换后的因变量 tfail 进行假设检验,所得结果如下:

Goodness – of – Fit Tests for Normal Distribution

检验	统计量		P	
Kolmogorov – Smirnov	D	0.08312402	Pr > D	>0.150
Cramer – von Mises	W – Sq	0.02172925	Pr > W – Sq	>0.250
Anderson – Darling	A – Sq	0.13498929	Pr > A – Sq	>0.250

由图 2 和以上关于正态性检验结果可知,经过 BOX – COX 变换后的结果变量 tfail 服从正态分布。

3.5 上述“model 语句”输出的结果

上述“model 语句”实际上创建了一个经 BOX – COX变换后的因变量 tfail 关于经 QPOINT 变量扩展变换后的三个定性自变量及其所有二次项的“二次反应曲面回归模型”。见图 3。

由图 3 中倒数第 2 列可知,三个试验因素的主效应项(即一次项)都具有统计学意义;而由它们产生的所有二次项都没有统计学意义。

Univariate Regression Table Based on the Usual Degrees of Freedom							
Variable	DF	Coefficient	Type II Sum of Squares	Mean Square	F Value	Liberal p	Label
Intercept	1	6.4206207	159.008	159.008	4232.19	>= <.0001	Intercept
Qpoint.Length	1	0.8323842	12.472	12.472	331.94	>= <.0001	Length
Qpoint.Amplitude	1	-0.6309916	7.167	7.167	190.75	>= <.0001	Amplitude
Qpoint.Load	1	-0.3924940	2.773	2.773	73.80	>= <.0001	Load
Qpoint.Length_2	1	-0.0856974	0.044	0.044	1.17	>= 0.2939	Length_2
Qpoint.Amplitude_2	1	0.0242183	0.004	0.004	0.09	>= 0.7633	Amplitude_2
Qpoint.Load_2	1	-0.0674555	0.027	0.027	0.73	>= 0.4058	Load_2
Qpoint.LengthAmplitude	1	-0.0382414	0.018	0.018	0.47	>= 0.5035	LengthAmplitude
Qpoint.LengthLoad	1	-0.0684146	0.056	0.056	1.49	>= 0.2381	LengthLoad
Qpoint.AmplitudeLoad	1	-0.0208340	0.005	0.005	0.14	>= 0.7142	AmplitudeLoad

图 3 二次反应曲面回归模型的参数估计与假设检验结果

反映回归模型对资料拟合效果的输出结果如下：

Root MSE	0.19383	R-Square	0.9725
Dependent Mean	6.33466	Adj R-Sq	0.9579
Coeff Var	3.05987	Lambda	0.0000

以上结果表明：模型对资料的拟合效果较好，R^2 =0.9725，校正的 R^2 =0.9579。

3.6 寻求更简约的回归模型及结果

利用以下 SAS 程序，可以获得更简约的回归模型。

```
proc transreg details data = yarn ss2 utilities
        plots = ( transformation ( dependent ) obp);
model BoxCox(fail / convenientlambda = -2 to 2 by 0.05) =
        class(length amplitude load/zero = sum);output
out = aaa approximations;
run;
```

简约回归模型的输出结果见图 4。

Univariate Regression Table Based on the Usual Degrees of Freedom							
Variable	DF	Coefficient	Type II Sum of Squares	Mean Square	F Value	Liberal p	Label
Intercept	1	6.3346643	1083.46	1083.46	30195.2	>= <.0001	Intercept
Class.Length250	1	-0.8609500	10.01	10.01	278.88	>= <.0001	Length 250
Class.Length300	1	0.0571316	0.04	0.04	1.23	>= 0.2809	Length 300
Class.Length350	1	0.8038184	8.72	8.72	243.10	>= <.0001	Length 350
Class.Amplitude8	1	0.6390643	5.51	5.51	153.66	>= <.0001	Amplitude 8
Class.Amplitude9	1	-0.0161455	0.00	0.00	0.10	>= 0.7574	Amplitude 9
Class.Amplitude10	1	-0.6229188	5.24	5.24	145.99	>= <.0001	Amplitude 10
Class.Load40	1	0.3700088	1.85	1.85	51.51	>= <.0001	Load 40
Class.Load45	1	0.0449703	0.03	0.03	0.76	>= 0.3934	Load 45
Class.Load50	1	-0.4149791	2.32	2.32	64.79	>= <.0001	Load 50

图 4 仅含主效应回归模型的参数估计与假设检验结果

由图 4 可知，三个试验因素都有 3 个水平，都以中间水平为“基准”，除中间水平无统计学意义外，其他均有统计学意义。

反映回归模型对资料拟合效果的输出结果如下：

Root MSE	0.18942	R-Square	0.9691
Dependent Mean	6.33466	Adj R-Sq	0.9598
Coeff Var	2.99029	Lambda	0.0000

以上结果表明：模型对资料的拟合效果较好，R^2 =0.9691，校正的 R^2 =0.9598。

与前面的结果相比，简约回归模型比复杂回归模型的 R^2 略低，但校正的 R^2 反而略高。

关于各试验因素的“重要性”和“效用值”的输出结果见图 5。由图 5 第 4 列可知，三个试验因素的重要性分别为：纱线测试样品的长度（A）占 44.851%、负载循环时的振幅大小（B）占 34.000%，负载量（C）占 21.149%。由图 5 第 2 列可知，“Utility”为各试验因素的各水平的“效用值”，当结果变量属于“高优指标”时，将各试验因素正的最大效用值对应的“水平”组合在一起，就构成了“理想试验点”。就本例而言，在理想试验点为“length 350”“Amplitude 8”和“Load 40”所构成的试验条件下，即当纱线长度取 350 mm、振幅取 8 mmd 和负载量取 40 g 时，纱线寿命最长。

Utilities Table Based on the Usual Degrees of Freedom				
Label	Utility	Standard Error	Importance (% Utility Range)	Variable
Intercept	6.3347	0.03645		Intercept
Length 250	-0.8609	0.05155	44.851	Class.Length250
Length 300	0.0571	0.05155		Class.Length300
Length 350	0.8038	0.05155		Class.Length350
Amplitude 8	0.6391	0.05155	34.000	Class.Amplitude8
Amplitude 9	-0.0161	0.05155		Class.Amplitude9
Amplitude 10	-0.6229	0.05155		Class.Amplitude10
Load 40	0.3700	0.05155	21.149	Class.Load40
Load 45	0.0450	0.05155		Class.Load45
Load 50	-0.4150	0.05155		Class.Load50
The standard errors are not adjusted for the fact that the dependent variable was transformed and so are generally liberal (too small).				

图 5　各试验因素的“重要性”及其各水平的“效用值”的输出结果

4　讨论与小结

在多因素试验研究中,要了解各因素对试验结果的影响情况,特别是因素之间各级交互作用的效应,最合适的试验设计类型为多因素析因设计。然而,多因素析因设计至少应满足两个特点:第一,全部试验点应该由所有试验因素水平的全面组合而成;第二,在各试验点条件下,至少要做两次独立重复试验。本文中的样例满足了前面提及的第一点,但不满足第二点,严格来说,此样例在试验设计上是存在瑕疵的。

通常的方差分析或多重回归分析对资料都有很高的要求,例如正态性和方差齐性等。样例中的因变量呈严重的正偏态分布,通过采用 BOX - COX 变换,使其偏斜情况得到了很好的校正。将结合分析与回归分析有机地结合在一起,不仅可以获得各试验因素对试验结果影响情况的分析结果,还能获得关于各试验因素重要性的评价以及确定出理想的试验点。

SAS 中的 TRANSREG 过程具有很强且多样性的变量变换能力,它集方差分析、回归分析和结合分析于一体,能够很好地处理不符合传统统计学要求的复杂资料,获得满意的统计分析结果。

参考文献

[1] 姬振豫. 正交设计的方法与理论[M]. 香港: 世界科技出版社, 2001: 1 - 38, 231 - 262.

[2] 方开泰, 马长兴. 正交与均匀试验设计[M]. 北京: 科学出版社, 2001: 1 - 156.

[3] 任露泉. 试验优化设计与分析[M]. 2 版. 北京:高等教育出版社, 2003: 10 - 399.

[4] 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.

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

[6] 胡良平. 回归建模的基础与要领(Ⅱ)——偏态分布计量资料的变换[J]. 四川精神卫生, 2018, 31(6): 493 - 497.

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

[8] 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.

(收稿日期:2019 - 06 - 12)

(本文编辑:吴俊林)