变量变换回归分析(Ⅱ)——拟合近似呈均匀分布资料的方法

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】 本文利用 SAS 帮助数据库中的一个数据集 sashelp.enso,介绍对自变量进行样条变换后的曲线回归分析方法。在 SAS/STAT 的 TRANSREG 过程中,涉及到六种样条变换方法,分别为:B-样条变换、B-样条基函数变换、单调 B-样条变换、非迭代惩罚 B-样条变换、迭代光滑样条变换、非迭代光滑样条变换。获得的结论是:在确保 $R^2 \approx 0.7$ 且回归模型尽可能精简的条件下,“非迭代惩罚 B-样条变换”与“迭代光滑样条变换”两种方法是以上六种方法中最好的曲线回归建模方法,这两种方法的拟合效果几乎完全相同。

【关键词】 曲线回归;非迭代惩罚 B-样条变换;光滑样条变换;节点;光滑参数
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2019.03.002

Regression analysis based on the variable transformation(Ⅱ)——the methods of fitting the data with almost uniform distribution

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**】 This paper was to introduce the approaches of curve regression analysis through the spline transformation of the independent variable by means of using the data set named sashelp.enso in the data base of SAS HELP. In the TRANSREG procedure of the SAS/STAT, six approaches of the spline transformation were involved as below: B-spline transformation, B-spline base transformation, monotonic B-spline transformation, non-iterative penalized B-spline transformation, iterative smoothing spline transformation, non-iterative smoothing spline transformation. The conclusion were as follows: under the conditions of ensuring the R-square to be equal to 0.7 approximately and the regression model streamlining as much as possible, the fourth and fifth approach mentioned above were the best and they had almost the same fitting effects.

【**Keywords**】 Curve regression; Noniterative penalized B-spline transformation; Smoothing spline transformation; Knot; Smoothing parameter

1 概 述

1.1 如何直观分析散布图的表现

在单个自变量的回归分析中,为了选择到合适的回归分析模型,最简单且最有效的方法是先绘制(X,Y)的散布图。通常可依据散布图中所有散点的分布情况,结合初等函数的表达式及其图象[1],尝试拟合几种最可能的曲线类型(也包括直线),并通过拟合优度比较,最终确定一种最合适的曲线回归模型。然而,如图 1 中散点所呈现的“形态”,确实难以作出合理的判断。所有散点几乎在一个长方形区域内“星罗密布”,说它们近似呈“均匀分布”似乎比较合理。

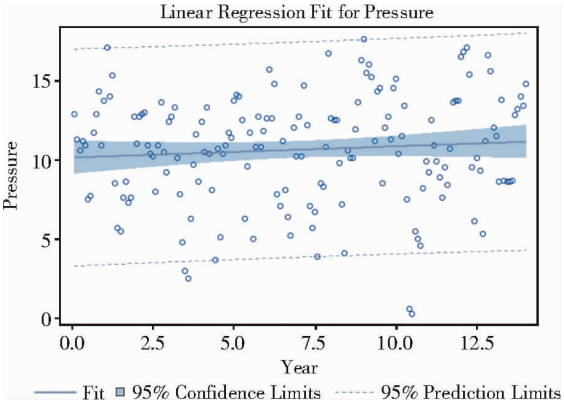

图 1 大气压值(pressure)随年份值(year)变化而变化的散布图

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

1.2 变量变换的必要性

对于图 1 中散点的分布情况,若直接拟合直线回归模型似乎是很合理的,但在统计学上又是毫无价值的(因为 $R^2=0.0071$)。所以,应采用合理的变量变换方法,使变换后的"新因变量"与"新自变量"之间呈现出较好的相关关系,以便尽可能地反映图 1 中大多数散点的变化趋势(即具有较高的拟合优度)。由此可知,"变量变换"是成功实现曲线拟合的有效途径。

2 一个取自 SAS 帮助的数据集

2.1 数据集的名称和数据结构

在 SAS 帮助"数据库"或"文件夹"中,有一个名为"sashelp.enso"的 SAS 数据集,其数据含义与结构如表 1 所示[2]。

表 1 澳大利亚港口城市与东部岛屿之间大气压值的变化数据

月份编号	年份编码值	大气压值
1	0.08333	12.9
2	0.16667	11.3
3	0.25000	10.6
4	0.33333	11.2
5	0.41667	10.9
⋮	⋮	⋮
164	13.6667	12.8
165	13.7500	13.2
166	13.8333	14.0
167	13.9167	13.4
168	14.0000	14.8

因篇幅限制,表 1 中仅列出了该数据集的前 5 行和最后 5 行。该数据集中含有 3 个变量和 168 个观测(即 N=168)。以下 SAS 程序可以显示出完整的 SAS 数据集:

```
proc print data=sashelp.enso noobs;
run;
```

2.2 用散布图呈现完整资料的变化趋势

以下 SAS 程序可以呈现图 1 中的大气压值(pressure)随年份值(year)变化而变化的趋势:

```
proc transreg data=sashelp.enso;
model identity(pressure)=identity(year);
run;
```

以上程序输出结果为图 1。由图 1 可知,除少数点外,绝大多数点几乎是随机地分布在一个长方形区域内,近似呈"均匀分布"。此长方形与 X 轴不完全平行,略呈一个微小的倾斜角。图 1 中的实线是基于最小二乘原理拟合出来的一条直线,按常规的统计学理念,这种表现的散布图提示分析者,对此资料拟合直线回归模型是无可非议的。然而,其 R^2 仅为 0.0071,说明用年份值去预测大气压值是非常不准确的。

2.3 统计分析的任务

针对图 1 中散点的分布或变化趋势,可否找到一个较为合适的统计模型,使 $R^2>0.6$,即用年份值(year)预测大气压值(pressure)的预测结果具有一定程度的准确性。

3 基于 TRANSREG 过程中各种样条变换后建模[2-3]

3.1 基于 B-样条变换(spline)后建模

3.1.1 基本概念与做法

在 SAS 中,实现 B-样条变换的关键词为"spline(自变量名)"。在运用 SAS 的 TRANSREG 过程时,可以对自变量 year 进行"B-样条变换",变换后的结果记为 Tyear。再构建因变量 pressure 关于新自变量 Tyear 的回归模型。所谓"B-样条变换",实际上就是拟合因变量关于自变量的多项式曲线回归模型,一次就是直线回归模型、二次就是抛物线回归模型、三次就是三次多项式回归模型,以此类推。通常,拟合三次多项式回归模型。

问题在于:是在自变量的整个取值区间内拟合一个样条函数曲线模型,还是将区间划分成多个子区间,分别在各子区间上各拟合一个样条函数曲线模型,即"分段拟合"。若在"nkonts="之后写一个"k(k≥0)",就是将自变量的整个区间划分成"k+1"个子区间。"nkonts"代表"节点数"或"断点数","nkonts=k"代表节点数为"k"。显然,随着 k 值增加,分段数目也在增加,在各个很短的子区间上的"多项式曲线"能更好地拟合该子区间上的散点,因而拟合的效果就会提升,直到曲线回归模型完全拟合给定的实际资料。下面给出"nkonts=0""nkonts=5""nkonts=10"等情形下的拟合结果。

3.1.2 SAS 输出结果及解释

基于 B-样条变换且节点数 k 取不同数值时对应的拟合效果(以 R^2 来度量)。见表 2。

表 2 基于 B－样条变换且节点数 k 取不同数值时对应的 R^2 值

编 号	k 值	R^2	编 号	k 值	R^2
1	0	0.0190	14	65	0.7831
2	5	0.0505	15	70	0.8093
3	10	0.1972	16	75	0.8074
4	15	0.2881	17	80	0.8421
5	20	0.3096	18	85	0.8439
6	25	0.3849	19	90	0.8523
7	30	0.5853	20	95	0.8746
8	35	0.7171	21	100	0.8973
9	40	0.7180	22	120	0.9299
10	45	0.7324	23	140	0.9698
11	50	0.7444	24	160	0.9890
12	55	0.7485	25	180	1.0000
13	60	0.7844	26	200	1.0000

由表 2 可知，除了 k＝65 和 k＝75 两行外，其他各行的 R^2 都随着 k 值增大而增大。其根本原因在前面已述及，此处不再赘述。按 R^2 ＞0.6 且回归模型尽可能精简的要求，k＝35 即可。若进一步尝试，发现 k＝31时，R^2 ＝0.6751，此时拟合的图形见图 2。

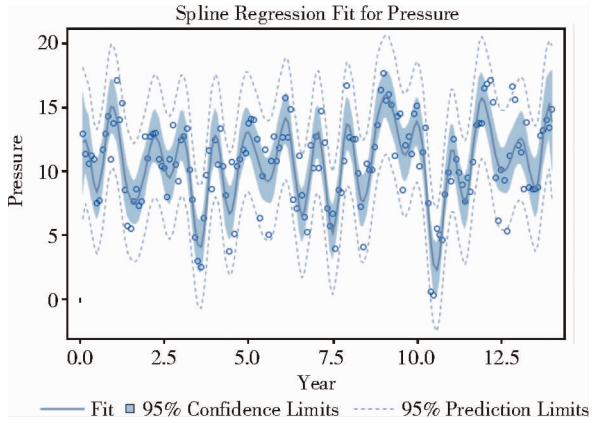

图 2 k＝31 时 B－样条变换对资料的拟合效果图示

3.1.3 “nkonts＝k”所对应的 SAS 程序

注意：“k”必须取一个具体数值，下面给出 k＝31 时对应的 SAS 程序。

```
proc transreg data=sashelp.enso outtest=aaa
ss2 coefficients test;
model identity(pressure)=spline(year/nknots=31);
output out=bbb predicted residuals;
run;
```

3.2 基于 B－样条基函数变换(bspline)后建模

3.2.1 基本概念与做法

在 SAS 中，实现 B－样条基函数变换的关键词为“bspline(自变量名)”。与前面的“B－样条变换”一样，也可以通过给“nkonts＝k”中的“k”赋一个具体值，来获得拟合效果不同的拟合结果。而且，两者的计算结果完全相同。它们之间的区别仅仅在于：使用“bspline(自变量名)”时，可以输出“B－样条基函数”的计算结果，而使用“spline(自变量名)”时，无法输出“B－样条基函数”的计算结果。

3.2.2 SAS 输出结果及解释

基于 B－样条基函数变换且节点数 k 取不同数值时对应的拟合效果(以 R^2 来度量)，与表 2 完全相同，此处从略。下面分别将 k＝0、k＝1 和 k＝2 时对应的“B－样条基函数”的计算结果呈现出来。

第一种情形：k＝0 时对应的“B－样条基函数”的计算结果见图 3。

Univariate Regression Table Based on the Usual Degrees of Freedom

Variable	DF	Coefficient	Type II Sum of Squares	Mean Square	F Value	Pr > F	Label
Bspline.Year_0	1	11.2274782	1383.56	1383.56	117.83	<.0001	Year 0
Bspline.Year_1	1	8.7523540	191.15	191.15	16.28	<.0001	Year 1
Bspline.Year_2	1	10.9625696	299.88	299.88	25.54	<.0001	Year 2
Bspline.Year_3	1	11.6055932	1478.32	1478.32	125.90	<.0001	Year 3

图 3 k＝0 时对应的“B－样条基函数”的计算结果

图 3 中的结果表明：在自变量 year 的整个取值区间内，拟合了一个“三次多项式曲线回归模型”，其表达式见式(1)。

$$\hat{y}=11.2274782t_0+8.7523540t_1+10.9625696t_2+11.6055932t_3 \quad (1)$$

在式(1)中，“$\hat{y}$”代表大气压 pressure 的估计值，而“t_i(i＝0,1,2,3)”分别代表对自变量 year 所做的变量变换结果。

第二种情形：k＝1 时对应的“B－样条基函数”的计算结果见图 4。

Univariate Regression Table Based on the Usual Degrees of Freedom							
Variable	DF	Coefficient	Type II Sum of Squares	Mean Square	F Value	Pr > F	Label
Bspline.Year_0	1	12.1138650	1145.54	1145.54	98.04	<.0001	Year 0
Bspline.Year_1	1	8.1283662	303.38	303.38	25.97	<.0001	Year 1
Bspline.Year_2	1	12.6590431	397.46	397.46	34.02	<.0001	Year 2
Bspline.Year_3	1	9.4318646	411.99	411.99	35.26	<.0001	Year 3
Bspline.Year_4	1	12.5024051	1211.92	1211.92	103.72	<.0001	Year 4

图 4　k = 1 时对应的“B - 样条基函数”的计算结果

图 4 中的结果表明:在前面式(1)基础上,又增加了一项“t_4”及其系数估计值。但前面各项的系数估计值也作了相应的调整。

第三种情形:k = 2 时对应的“B - 样条基函数”的计算结果见图 5。

Univariate Regression Table Based on the Usual Degrees of Freedom							
Variable	DF	Coefficient	Type II Sum of Squares	Mean Square	F Value	Pr > F	Label
Bspline.Year_0	1	11.7519389	793.570	793.570	67.86	<.0001	Year 0
Bspline.Year_1	1	10.1266883	440.097	440.097	37.63	<.0001	Year 1
Bspline.Year_2	1	9.0410189	313.657	313.657	26.82	<.0001	Year 2
Bspline.Year_3	1	12.3669290	591.529	591.529	50.58	<.0001	Year 3
Bspline.Year_4	1	9.0770713	354.714	354.714	30.33	<.0001	Year 4
Bspline.Year_5	1	13.2037228	989.497	989.497	84.61	<.0001	Year 5

图 5　k = 2 时对应的“B - 样条基函数”的计算结果

图 5 中的结果表明,在前面的基础上,又增加了一项“t_5”及其系数估计值。但前面各项的系数估计值也作了相应的调整。

由此可知,自变量 year 经过变量变换后的变量数目随着 k 值的增加而增加,其具体个数为(4 + k)。即 k = 0 时,有 4 个;k = 1 时,有 5 个;而 k = 2 时,有 6 个。值得注意的是,变换后的变量的下角标从“0”开始。

3.2.3　“nkonts = k”所对应的 SAS 程序

注意:“k”必须取一个具体数值,下面给出 k = 5 时对应的 SAS 程序。

```
proc transreg data = sashelp.enso outtest = aaa
ss2 coefficients test;
model identity(pressure) = bspline(year/nknots = 5);
output out = bbb predicted residuals;
run;
```

3.3　基于单调 B - 样条变换后建模

3.3.1　基本概念与做法

在 SAS 中,实现单调 B - 样条变换的关键词为“mspline(自变量名)”。与前面的“B - 样条变换”一样,也可以通过给“nkonts = k”中的“k”赋一个具体值,来获得拟合效果不同的拟合结果。

3.3.2　SAS 输出结果及解释

基于单调 B - 样条变换且节点数 k 取不同数值时对应的拟合效果(以 R^2 来度量)。见表 3。

表 3　基于单调 B - 样条变换且节点数 k 取不同数值时对应的 R^2 值

编　号	k 值	R^2	编　号	k 值	R^2
1	0	0.0116	14	65	0.0464
2	5	0.0177	15	70	0.0466
3	10	0.0263	16	75	0.0475
4	15	0.0342	17	80	0.0467
5	20	0.0388	18	85	0.0473
6	25	0.0389	19	90	0.0473
7	30	0.0414	20	95	0.0482
8	35	0.0432	21	100	0.0484
9	40	0.0440	22	120	0.0487
10	45	0.0444	23	140	0.0492
11	50	0.0452	24	160	0.0495
12	55	0.0452	25	180	0.0495
13	60	0.0457	26	200	0.0495

表 3 与表 2 相比,随着回归模型越来越复杂,表

3 中的 R^2 值增加十分缓慢,且 R^2 最大值均未超过 0.05。

3.3.3 “nkonts = k”所对应的 SAS 程序

注意:“k”必须取一个具体数值,下面给出 k = 160时对应的 SAS 程序。

```
proc transreg data = sashelp.enso outtest = aaa ss2 coefficients test;
model identity(pressure) = mspline(year/nknots = 160);
output out = bbb predicted residuals;
run;
```

3.4 基于非迭代惩罚 B - 样条变换(pbspline)后建模

3.4.1 基本概念与做法

在 SAS 中,实现非迭代惩罚 B - 样条变换的关键词为“pbspline(自变量名)”。与前面的“B - 样条变换”一样,也可以通过给“nkonts = k”中的“k”赋一个具体值,来获得拟合效果不同的拟合结果。

3.4.2 SAS 输出结果及解释

基于非迭代惩罚 B - 样条变换且节点数 k 取不同数值时对应的拟合效果(以 R^2 来度量)。见表 4。

表 4 基于非迭代惩罚 B - 样条变换且节点数 k 取不同数值时对应的 R^2 值

编 号	k 值	R^2	编 号	k 值	R^2
1	0	0.6998	14	65	0.6989
2	5	0.0178	15	70	0.7012
3	10	0.0178	16	75	0.6993
4	15	0.0178	17	80	0.6994
5	20	0.2059	18	85	0.7004
6	25	0.2045	19	90	0.6997
7	30	0.2056	20	95	0.7002
8	35	0.2063	21	100	0.6998
9	40	0.2064	22	120	0.6999
10	45	0.2064	23	140	0.6999
11	50	0.2067	24	160	0.6999
12	55	0.2067	25	180	0.6999
13	60	0.2068	26	200	0.6999

观察表 4,很难总结出其中规律,当 k = 0、k = 65、k ≥ 100 时,都能获得较大的 R^2 值。事实上,在使用 pbspline 变换时,涉及到一个关键的“光滑参数 lambda”。当此参数取不同数值时,回归模型对资料的拟合效果是不同的。有多种评价指标用来衡量回归模型对资料的拟合优度,其中,SBC 统计量是经常被选用的。SBC 取值最小时,对应的“lambda 值”是最合适的光滑参数。此时,对应的 R^2 将取极大值。

3.4.3 寻找使 SBC 取得极小值时所对应的 SAS 程序

运行下面的一段 SAS 程序,可以找到使 SBC = 398.9 为极小值,此时,lambda = 1.14。

```
proc transreg data = sashelp.enso;
model identity(pressure) = pbspline(year/sbc lambda = .1 to 100.1 by 0.01);
run;
```

运行下面的一段 SAS 程序,可以找到使 SBC = 435.7 为极小值,此时,lambda = 1801.1。

```
proc transreg data = sashelp.enso;
model identity(pressure) = pbspline(year/sbc lambda = 100.1 to 10000.1 by 0.01);
run;
```

由于在 lambda 的整个取值范围内,SBC 取最小值时,回归模型对资料的拟合优度最好,故应取 lambda = 1.14。使用下面的 SAS 程序,可以快速获得最大的 R^2 = 0.6999。

```
proc transreg data = sashelp.enso outtest = aaa ss2 coefficients test;
model identity(pressure) = pbspline(year/sbc lambda = 1.14);
output out = bbb predicted residuals;
run;
```

3.5 基于迭代光滑样条变换(sspline)后建模

3.5.1 基本概念与做法

在 SAS 中,实现迭代光滑样条变换的关键词为“sspline(自变量名)”。与前面的“B - 样条变换”不一样,它不能通过给“nkonts = k”中的“k”赋一个具体值,来获得拟合效果不同的拟合结果。但它引入了一个“光滑参数 smooth(简写为 sm)”,通过调整 sm 的数值,可以获得拟合效果不同的拟合结果。sm 的取值从 0 开始,其取值区间为[0,100],即可取此区间内任何一个实数,也包括区间的两个端点。数值越小,表明光滑程度越差,当 sm = 0 时,就是将所有相邻的两点用折线连起来,此时,为完全拟合(也称为“过拟合”);当 sm = 100 时,就是用一条直线来拟合该资料,拟合的效果最差。因此,可以依据 R^2 的大小,尝试寻找合适的 sm 数值。

3.5.2 SAS 输出结果及解释

经尝试,当 sm = 22.2 时,R^2 = 0.6039;当 sm = 22.3 时,R^2 = 0.6008;当 sm = 22.4 时,R^2 = 0.5976;当 sm = 22.5 时,R^2 = 0.5944。通过进一步尝试,发现当 sm = 22.32 时,R^2 = 0.6001,刚好满足 R^2 > 0.6 的基本要求。

3.5.3 寻找使 R^2 略大于 0.6 时的 sm 数值所对应的 SAS 程序

```
proc transreg data = sashelp.enso outtest = aaa
ss2 coefficients test;
model identity(pressure) = sspline(year/sm = 22.32);
output out = bbb predicted residuals;
run;
```

3.6 基于非迭代光滑样条变换(smooth)后建模

3.6.1 基本概念与做法

在 SAS 中,实现非迭代光滑样条变换的关键词为“smooth(自变量名)”。与前面的“迭代光滑样条变换”一样,引入了一个“光滑参数 smooth(简写成 sm =)”,通过调整 sm 的数值,可以获得拟合效果不同的拟合结果。

3.6.2 SAS 输出结果及解释

经尝试,当 sm = 20.8 时,R^2 = 0.6048;当 sm = 20.9 时,R^2 = 0.6018;当 sm = 21.0 时,R^2 = 0.5989;当 sm = 21.1 时,R^2 = 0.5959。通过进一步尝试,发现当 sm = 20.96 时,R^2 = 0.6001,刚好满足 R^2 > 0.6 的基本要求。

3.6.3 寻找使 R^2 略大于 0.6 时的 sm 数值所对应的 SAS 程序

```
proc transreg data = sashelp.enso outtest = aaa
ss2 coefficients test;
model identity(pressure) = smooth(year/sm = 20.96);
output out = bbb predicted residuals;
run;
```

4 讨论与小结

本文图 1 中的全部散点几乎呈“均匀分布”状态,常规的直线回归模型没有任何使用价值。然而,在 SAS/STAT 模块中的“TRANSREG 过程”收录了很多种类的变量变换方法。本文采用了其中的一大类方法,即“样条变换方法”。从思路和算法上又给出了六个彼此稍有区别的具体方法,它们各具优点和缺点。根据分析者的目的不同,可以考虑选用不同的方法来拟合资料。

本文介绍的六种样条变换方法主要作用是曲线拟合,在自变量的取值范围内找到合适的曲线回归模型,即尽可能高的拟合优度(如本文期望 R^2 > 0.6)且尽可能精简的回归模型(可以回归模型误差项的自由度“$DF_{误差}$”来反映,其数值越大越好,它标志着回归模型中被估计的参数数目少,即回归模型较为精简)。

B - 样条变换和 B - 样条基函数变换:节点数为 31 时,R^2 = 0.6751,$DF_{误差}$ = 133。

单调 B - 样条变换:拟合效果极差。

非迭代惩罚 B - 样条变换:Lambda = 1.14 时,R^2 = 0.6999,$DF_{误差}$ = 131.28。

迭代光滑样条变换:Sm = 22.32 时,R^2 = 0.6001,$DF_{误差}$ = 138.34。

非迭代光滑样条变换:Sm = 20.96 时,R^2 = 0.6001,$DF_{误差}$ = 137.09。

由于以上方法所依赖的“调节参数”不同,因此,结果尚缺乏可比性。但可将 R^2 值定为 0.6999,计算出除“单调 B - 样条变换”之外的其他几种情况下的结果:

B - 样条变换和 B - 样条基函数变换:节点数为 34 时,R^2 = 0.6974,$DF_{误差}$ = 130;节点数为 35 时,R^2 = 0.7171,$DF_{误差}$ = 129。

非迭代惩罚 B - 样条变换:Lambda = 1.14 时,R^2 = 0.6999,$DF_{误差}$ = 131.28。

迭代光滑样条变换:Sm = 18.4 时,R^2 = 0.6998,$DF_{误差}$ = 131.34。

非迭代光滑样条变换:Sm = 16.88 时,R^2 = 0.6999,$DF_{误差}$ = 129.18。

综上所述,“非迭代惩罚 B - 样条变换”与“迭代光滑样条变换”的拟合效果基本相同,是以上六种方法中最好的曲线回归建模方法。

参考文献

[1] 胡良平, 高辉. 非线性回归分析[M]. 北京: 电子工业出版社, 2013: 33 - 37.

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

[3] 胡良平. 提高回归模型拟合优度的策略(Ⅳ)——优化计分变换与其他变量变换[J]. 四川精神卫生, 2019, 32(1): 21 - 28.

(收稿日期:2019 - 06 - 12)

(本文编辑:陈 霞)