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

胡良平[1,2]*

(1.军事医学科学院研究生院,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】 本文目的是介绍第四种提高回归模型拟合优度的策略,即优化计分变换与其他变量变换。具体方法包括以下几个方面:①第一,对多值名义自变量采取“优化计分变换”;②对有序自变量分别采取“单调变换”与“优化计分变换”;③对定量自变量分别采取“样条变换”和“单调样条变换”;④对定量因变量分别采取“样条变换”“单调样条变换”和“BOX-COX变换”。全部变量变换方法组合起来共12种,共创建了12个多重非线性回归模型。依据“拟合优度评价指标”的取值,从12个回归模型中挑选出一个,即本文中的“模型1”,其“均方误差平方根=0.30935、R^2=0.9586、调整R^2=0.9527”。结合本期科研方法专题同类文章的结果和结论,得出提高回归模型拟合优度的策略主要在于以下四点:①应对“定量因变量”“定量自变量”和“多值有序自变量”采取合适的变量变换方法;②在拟合回归模型的过程中,应尽可能多地引入派生变量;③应假定回归模型中不含截距项;④在构建回归模型的过程中,应尽可能多地使用筛选自变量的策略,如“前进法”“后退法”和“逐步法”。

【关键词】 优化计分变换;单调变换;样条变换;BOX-COX变换
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2019.01.004

Strategy of improving the goodness of fit of the regression model(Ⅳ)
——the optimal scoring transformation and the other variable transformations

Hu Liangping[1,2]*

(1. *Graduate School, Academy of Military Medical 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 fourth strategy of improving the goodness of fit of the regression models, the optimal scoring transformation and the other variable transformations. The concrete approaches were as follows: ①"The optimal scoring transformation" was adopted to the multi-value nominal independent variable. ②"The monotonic transformation" and "the optimal scoring transformation" were adopted to the multi-value ordered independent variable, respectively. ③"The spline transformation" and "the monotonic spline transformation" were adopted to the quantitative independent variables, respectively. ④"The spline transformation" "the monotonic spline transformation" and "the BOX-COX transformation" were adopted to the quantitative dependent variable, respectively. There were twelve variable transformation ways, so the twelve multiple nonlinear regression models were built. One best regression model, which was "the model one" in this article, was selected from the twelve models mentioned above in terms of the results of the goodness of fit evaluation. The results were as follows: Root MSE=0.30935, R-Square=0.9586, and the adjusted R-Square=0.9527. Combined the results of this article with the other results of the previous three articles in the similar titles in this journal, the final conclusions were acquired as follows: ①"The quantitative dependent variable" "the quantitative independent variables" and "the multi-value ordered independent variables" should be transformed in an appropriate form. ②The derived variables should be introduced as many as possible in fitting the regression model. ③No intercept term should be applied in fitting the regression models. ④The strategies of screening independent variables should be adopted as many as possible during fitting the regression models, such as "forward selection" "backward selection" and "stepwise selection".

【**Keywords**】 Optimal scoring transformation; Spline transformation; Monotonic transformation; Box-cox transformation

1 回归建模前的变量变换[1]

1.1 对多值名义自变量进行“优化计分变换(Opscore)”或“单调变换(Monotone)”

1.1.1 优化计分变换

优化计分变换就是采用Fisher[2]提出的给多值名义变量或多值有序变量各水平赋值的方法。Opscore(x)代表对变量x进行优化计分变换,它可以被用于“多值名义变量”或“多值有序变量”。

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

1.1.2 单调变换

单调变换就是采用 Kruskal 提出的给多值有序变量各水平赋值的方法(被称为次要最小平方单调变换)。Monotone(x)代表对变量 x 进行单调变换,它只能被用于“多值有序变量”。

1.1.3 “优化计分变换”与“单调变换”的具体做法

两者具体的变换方法是相同的,每一个不同的非缺失值形成一个不同的类,例如:“1,1,1,2,2,3”形成三类,即 3 个“1”算作一类、两个“2”算作一类、一个“3”算作一类。具体赋值方法参见下面的实例:

设变量 x 的具体取值为(. . . A . A . B 1 1 1 2 2 3 3 3 4)′;设变量 y 的具体取值为(5 6 2 4 2 1 2 3 4 6 4 5 6 7)′。

【说明】以上“x”与“y”都有 14 个取值,用“向量”形式表示,即这两个向量都有 14 个分量。其中,变量 x 的前两个分量都是缺失值“.”;第 3 和第 4 个分量都是“. A”;第 5 个分量是“. B”;第 6 ~ 14 个分量都是数字。而变量 y 的全部 14 个分量都是数字。

于是,以“y”为“因变量”,对“变量 x”进行“优化计分变换”和“单调变换”。用关键词表示为:Opscore(x)和 Monotone(x)。它们只有 8 个分量,即:

Opscore(x) = (5 6 3 2 2 5 5 7)′

Monotone(x) = (5 6 3 2 2 5 5 7)′

【说明】在“x”的向量中,最前面的两个缺失值各形成一类,即有两类;“. A”出现了两次,形成一类;“. B”形成一类 ;三个“1”形成一类;两个“2”形成一类;三个“3”形成一类;一个“4”形成一类,故总共形成 8 类。

在同一类中,求出“对应于因变量 y 全部数值的算术平均值”,赋值给变换后的“变量 x”,作为其“优化计分变换”或“单调变换”的结果,例如:两个“. A”对应着变量 y 的两个数值为 2 和 4,其算术平均值为 3;同理,三个“1”对应着变量 y 的三个数值为“1、2、3”,其算术平均值为 2;两个“2”对应着变量 y 的两个数值为“4 与 6”,其算术平均值为 5;三个“3”对应着变量 y 的三个数值为“4、5、6”,其算术平均值为 5;一个“4”对应着变量 y 的一个数值为“7”,其算术平均值为 7。

1.2 可对定量变量进行的变量变换的种类

1.2.1 变量扩展

变量扩展就是由“一个变量”变换出“多个变量”,即由原变量产生出“派生变量”[4]。例如:由一个变量 x 可以产生出:x^2、x^3、ln(x)、$\sqrt{x}$等;由一个 x 和一个 y 可以产生出:xy、x^2y、xy^2 和 e^{x+y}等。在派生变量中,可以包含“原变量”。

1.2.2 非优化变换

非优化变换就是用一个变换后的“新变量”取代“原变量”。这个“新变量”在后续的迭代算法中不再被重新变换了(对具有缺失值估计所做的可能的线性变换除外)。常用的“非优化变换”有如下几种。

ARSIN(x):此处的“x”为“百分率”数据,通常其取值区间为“0 < x < 1”,在数学上,其取值区间为“$-1.0 \leqslant x \leqslant 1.0$”,此变换被叫做“反三角正弦变换”。

EXP(x):指数变换,通常为“e^x”变换;若希望底数为任意的正实数 a,即做“a^x”变换,在使用 SAS 的“TRANSREG 过程”中,需要另外再使用“Parameter = ”来指定,即在“ = ”后写出“a”的具体数值。

LOG(x):对数变换,通常取以“e”为底数的对数(e = 2.718281828),即取自然对数变换;若希望底数为任意的正实数 a,即做“$\log_a x$”变换,在使用 SAS 的“TRANSREG 过程”中,需要另外再使用“Parameter = ”来指定,即在“ = ”后写出“a”的具体数值。

LOGIT(x):logit 变换,即做“$\log_a[x/(1-x)]$”变换,这里的“a”通常为“e = 2.718281828”。注意:x 为数值型变量,其定义域为(0.0,1.0)。

POWER(x):幂变换,即做“x^a”变换。例如:希望做“$x^{1.5}$”变换,关键词写为 POWER(x/PARAMETER = 1.5)。显然,当“PARAMETER = ”后数值为“2”时,就是平方变换;为“ -1”时,就是倒数变换;为“0.5”时,就是开平方根变换。

RANK(x):秩变换。将变量 x 由小到大排序,再分别赋值为“1、2、3、4、5、……”。

1.2.3 非线性拟合变换

BOXCOX(x):这是 Box 等[5]提出的变量变换方法,实际应用参见文献[6]。此变换只适用于回归分析中的“定量因变量”。

PBSPLINE(x):这是非迭代惩罚 B 样条变换,使用时的特殊要求参见文献[1]。此变换只适用于回归分析中的“定量自变量”。

SMOOTH(x):这是一种非迭代光滑样条变换,由 Reinsch[7]提出的变量变换方法,使用时的特殊要求参见文献[1]。此变换只适用于回归分析中的

“定量自变量”。

1.2.4 优化变换

优化变换就是采用迭代计算导出的变量变换方法。在 SAS 的“TRANSREG 过程”中,涉及到以下六种优化变换方法,即:LINEAR(x)、MONOTONE(x)、MSPLINE(x)、OPSCORE(x)、SPLINE(x)和 UNTIE(x)。以上六种优化变换在使用时都有具体要求,详见文献[1],此处从略。

1.3 对定量变量进行其他变量变换

1.3.1 恒等变换[IDENTITY(x)]

恒等变换,就是没有改变变量的取值。通常情况下,就是“不做变换”。应注意:在使用 SAS 的“TRANSREG 过程”时,写在“model 语句”中的所有“因变量”和“自变量”之前必需要有“变量变换”的关键词。若不想对“$x_4 - x_{10}$”进行任何变量变换,必须按如下的方式写:IDENTITY($x_4 - x_{10}$)。

使用此种变换,可以产生“交互项”,具体使用方法参见文献[1],此处从略。

1.3.2 迭代光滑样条变换[SSPLINE(x)]

这是对变量 x 进行迭代光滑样条变换,此变量变换方法通常不会使误差平方达到最小值。具体使用方法参见文献[1],此处从略。

2 实际问题与数据结构

沿用本期科研方法专题第一篇文章《提高回归模型拟合优度的策略(I)——哑变量变换与其他变量变换》中的“实际问题与数据结构”[1],此处从略。

3 解决问题的思路和做法

3.1 对多值名义自变量和多值有序自变量进行变量变换

3.1.1 所需要的 SAS 程序

对“燃油种类(fuel)”这个“6 值名义自变量”进行“优化计分变换”;对“压缩比(cpratio)”这个“5 值有序自变量”进行“单调变换”。所需要的 SAS 程序如下:

```
ods graphics on;
title'Gasoline Example';
title2 ' Iteratively Estimate NOx, CpRatio, EqRatio, and Fuel';
 * Fit the Nonparametric Model;
proc transreg data = sashelp. Gas solve test nomiss plots = all;
ods exclude where = (_path_ ? 'MV');
model mspline(NOx / nknots = 9) =
      spline(EqRatio / nknots = 9)
      monotone(CpRatio) opscore(Fuel);
output out = aaa tdprefix = td tiprefix = ti;
run;
proc freq data = aaa;
      tables cpratio ticpratio fuel tifuel;
run;
```

3.1.2 与所需变量变换有关的 SAS 输出结果

Compression Ratio

CpRatio	频数	百分比	累积频数	累积百分比
7.5	93	54.39	93	54.39
9	17	9.94	110	64.33
12	24	14.04	134	78.36
15	20	11.70	154	90.06
18	17	9.94	171	100.00

以上是“压缩比(cpratio)”的原始取值及其各水平的频数分布。

Compression Ratio Transformation

tiCpRatio	频数	百分比	累积频数	累积百分比
7.1598227594	93	54.39	93	54.39
10.953020684	17	9.94	110	64.33
12.944552566	24	14.04	134	78.36

14.20153726	20	11.70	154	90.06
17.433541388	17	9.94	171	100.00

以上是“压缩比(cpratio)”经过“单调变换”后的取值及其各水平的频数分布。

由以上两部分输出结果可知:“单调变换”是将“多值有序变量”的水平做一一对应的变换,各水平出现的“频数”不会发生变化。

Fuel	频数	百分比	累积频数	累积百分比
0	9	5.26	9	5.26
1	25	14.62	34	19.88
2	90	52.63	124	72.51
3	13	7.60	137	80.12
4	22	12.87	159	92.98
5	12	7.02	171	100.00

以上是“燃油种类(fuel)”的“6 种水平代码(分别为 0、1、2、3、4、5)”及其各水平的频数分布。

Fuel	频数	百分比	累积频数	累积百分比
82rongas	9	5.26	9	5.26
94% Eth	25	14.62	34	19.88
Ethanol	90	52.63	124	72.51
Gasohol	13	7.60	137	80.12
Indolene	22	12.87	159	92.98
Methanol	12	7.02	171	100.00

以上是“燃油种类(fuel)”的“6 种水平的真实含义”及其各水平的频数分布。

Fuel Transformation

tiFuel	频数	百分比	累积频数	累积百分比
1.3742357556	12	7.02	12	7.02
1.5572054576	25	14.62	37	21.64
1.5975408831	90	52.63	127	74.27
4.2093356776	9	5.26	136	79.53
4.3664153593	22	12.87	158	92.40
4.4654059829	13	7.60	171	100.00

以上是“燃油种类(fuel)” 经过“优化计分变换”后的取值及其各水平的频数分布。

3.2 对定量自变量进行变量变换

对定量自变量可以进行“样条变换(spline)”“单调样条变换(mspline)”或“变量扩展变换(pspline)(即产生‘派生变量’,例如变量的平方项、立方项等)”。例如:对“eqratio”进行“样条变换”,可以写成 spline(eqratio)或 spline(eqratio/degree = 3 nknots = 0)。后一种写法涉及到两个参数,一是“degree = ”,它是指“样条函数(通常是多项式)”中多项式的“次数”,“degree = 3”代表“三次多项式”;二是“nknots = ”,它是指“结点(或写成‘节点’)个数”,实际上就是“断点个数”。其具体含义是:在该定量变量的取值区间内,确定“几个断点(如 nknots = 9)”,于是,就将该区间划分成“10 段”。在每一段上,拟合一个“多项式”。就整体而言,被称为“分段多项式(a piecewise polynomial)”。

3.3 对定量因变量进行变量变换

若需要变换的定量变量为“因变量”,除了可以进行“定量自变量”的某些变量变换(注意:“非迭代

惩罚 B 样条变换”只能用于“定量自变量”)之外,还可以进行其他一些变量变换,如“BOX - COX 变换(注意:该变换仅适用于‘定量因变量’)”等。

3.4 针对实际问题,寻找合适的变量变换方法

3.4.1 寻找最优回归模型的策略

前面提及的“实际问题”涉及“定量因变量(nox)”“定量自变量(eqratio)”“多值有序自变量(cpratio)”和“多值名义自变量(fuel)”。在拟合多重回归模型的过程中,涉及到如何对上述 4 个变量进行“变量变换”。当然,对每个变量选取不同的“变量变换”方法,将会得到不同的拟合效果。下面将固定“多值名义自变量(fuel)”的变量变换方法为:优化计分变换,即“OPSCORE(fuel)”,对其他变量依次采取各种可能的变量变换,呈现出各种情况下回归模型的拟合优度,最终选择“拟合优度最好”的回归模型。

3.4.2 不同变量变换方法组合下的回归模型拟合优度

在对“多值名义自变量(fuel)”做优化计分变换“OPSCORE(fuel)”的前提条件下,对其他变量依次采取各种可能的变量变换,对应的回归模型拟合优度见表 1。

表 1 在 OPSCORE(fuel)的前提下对“nox”“eqratio”和“cpratio”进行不同变量变换后对应的回归模型的拟合优度

模型编号	Nox_B	Eqratio_B	Cpratio_B	Root MSE	R^2	调整 R^2
模型 1	Spline	Spline	Monotone	0.30935	0.9586	0.9527
模型 2	Mspline	Spline	Monotone	0.31019	0.9584	0.9525
模型 3	Spline	Mspline	Monotone	0.88234	0.6613	0.6155
模型 4	Mspline	Mspline	Monotone	0.95998	0.5990	0.5449
模型 5	Boxcox	Spline	Monotone	0.25591	0.9372	0.9283
模型 6	Boxcox	Mspline	Monotone	0.72093	0.4984	0.4306
模型 7	Spline	Spline	Opscore	0.30935	0.9586	0.9527
模型 8	Mspline	Spline	Opscore	0.31019	0.9584	0.9525
模型 9	Spline	Mspline	Opscore	0.88077	0.6625	0.6169
模型 10	Mspline	Mspline	Opscore	0.95353	0.6044	0.5510
模型 11	Boxcox	Spline	Opscore	0.25591	0.9372	0.9283
模型 12	Boxcox	Mspline	Opscore	0.71511	0.5065	0.4398

注:“Root MSE”代表“均方误差的平方根”,此值越小越好;“Nox_B”代表对“Nox”进行变量变换的方法,表中第 3 列和第 4 列符号代表的含义与“Nox_B”相同,此处从略

由表 1 可知:模型 1 和模型 7 的结果相同且拟合优度最好,对应的 SAS 过程步程序如下:

```
proc transreg data = sashelp. Gas solve test nomiss
     plots = all;
ods exclude where = (_path_ ? ´MV´);
model spline(NOx / nknots = 9)
      = spline(EqRatio / nknots = 9)
      monotone(CpRatio)   opscore(Fuel);
run;
```

3.4.3 模型 1 的主要输出结果

由于“模型 1”是一个“非参数模型”,只能给出总模型的有关信息和与“拟合优度”有关的结果,不能给出“回归系数”等信息。但可以以“图形”方式给出各变量变换的结果,以提示对不同变量采取什么变量变换方法最有效。下面先给出与“拟合优度”有关的结果:

Root MSE	0.30935	R - Square	0.9586
Dependent Mean	2.34593	Adj R - Sq	0.9527
Coeff Var	13.18661		

表 1 中的第 1 行就是摘录了以上的结果。

在前面给出的“SAS 过程步”的“MODEL 语句”中,对定量因变量(nox)、定量自变量(EqRatio)、多值有序自变量(CpRatio)和多值名义自变量(Fuel)分别做了“样条变换”“样条变换”“单调变换”和“优化计分变换”,变换的结果用“图形”呈现出来,见图 1。

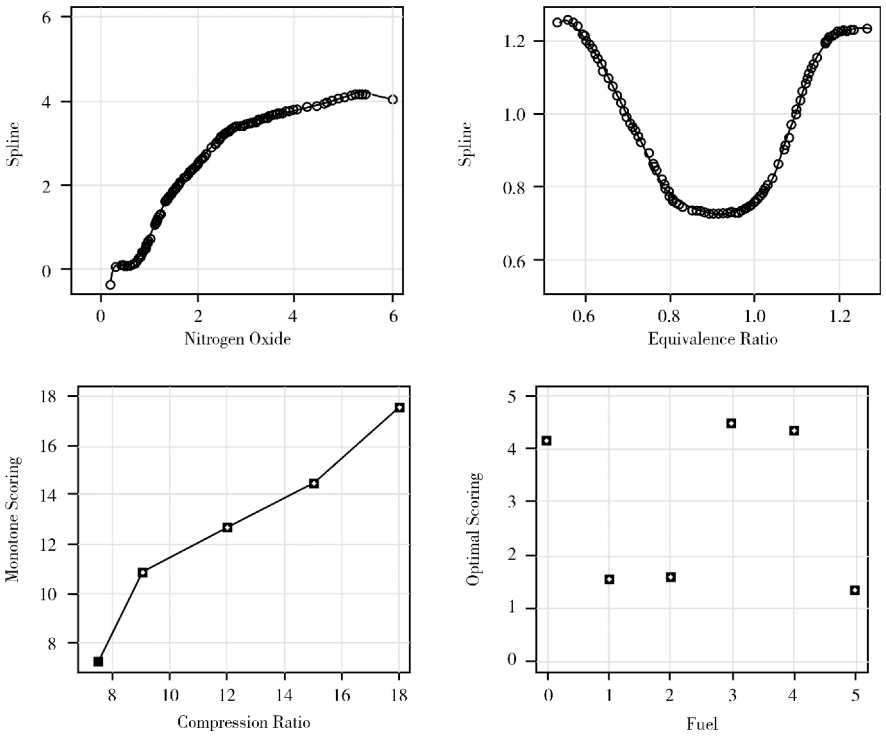

图 1 基于 MODEL 语句中的要求对 4 个变量做的变换结果

在图 1 中,左上方一图表明:定量因变量(nox)呈现出“对数”变化趋势,提示可对其进行“对数变换”;右上角一图表明:定量自变量(EqRatio)呈现出“二次抛物线”变化趋势,提示可对其进行“二次多项式变换”;左下角一图表明:多值有序自变量(CpRatio)呈现出近似“直线”变化趋势,提示可对其进行“恒等变换(即不做变换)”;右下角一图表明:多值名义自变量(Fuel)呈现出“两种级别”,其中,水平代码为“1、2、5”的燃料种类对结果的影响数量大约在“1 与 2”之间;而水平代码为“0、3、4”的燃料种类对结果的影响数量大约在“4 与 5”之间,纵轴上的“数量”就是“优化计分”的结果。

基于以上的“非参数多重非线性回归分析结果(以图 1 反映其变量变换的效果)”,可以构建参数多重非参数回归分析模型如下:

log(nox) = b0 + b1 * EqRatio + b2 * EqRatio * *2 + b3 * CpRatio + Sum b(j) * Fuel(j) + Error

此模型的含义是:对定量因变量(nox)进行自然对数变换,对定量自变量(EqRatio)进行二次多项式变换,对多值有序自变量(CpRatio)进行恒等变换,而对多值名义自变量(Fuel)进行优化计分变换并将其 6 个计分值[Fuel(j)]与其回归系数[b(j)]分别相乘后求和(j = 0、1、2、3、4、5)(注意:相当于把“6 值名义自变量”视为 6 个新变量,就有 6 个回归系数;但它们又属于同一个多值名义变量,因此,其自由度为 6 - 1 = 5,在本质上相当于是 5 个哑变量)。

3.4.4 构建参数多重非线性回归模型

采用 SAS 拟合上述参数多重非线性回归模型所需要的 SAS 过程步程序如下:

```
title2'Now fit log(nox) = b0 + b1 * EqRatio + b2 *
      EqRatio * *2 +';
title3'b3 * CpRatio + Sum b(j) * Fuel(j) + Error';
* - Fit the Parametric Model Suggested by the Non-
parametric Analysis - ;
proc transreg data = sashelp. Gas solve ss2 short
                nomiss plots = all cldetail;
model log(nox) = pspline(EqRatio / deg = 2)
                identity(CpRatio) opscore(Fuel);
run;
```

3.4.5 参数多重非线性回归模型输出结果

Univariate ANOVA Table Based on the Usual Degrees of Freedom

Source	DF	Sum of Squares	Mean Square	*F*	Pr > *F*
Model	8	79.33838	9.917298	213.09	<0.0001

Error	160	7.44659	0.046541
Corrected Total	168	86.78498	

以上是总模型的假设检验结果，$F=213.09$，$P<0.0001$，表明总模型具有统计学意义。

Root MSE	0.21573	R-Square	0.9142
Dependent Mean	0.63130	Adj R-Sq	0.9099
Coeff Var	34.17294		

以上是拟合优度评价结果：均方误差平方根为 0.21573，R^2 为 0.9142，调整 R^2 为 0.9099，相对于表 1 中“模型 1”的三个对应的数值（0.30935、0.9586、0.9527），均方误差平方根略微变小了一点，然而，R^2 和调整 R^2 的数值却下降得比较多（因为模型中自变量的项数减少得很多，现在模型的自由度 $df=8$；而模型 1 的 $df=21$）。

Univariate Regression Table Based on the Usual Degrees of Freedom

Variable	DF	Coefficient	Type II Sum of Squares	Mean Square	F	Pr > F	Label
Intercept	1	-15.274649	57.1338	57.1338	1227.60	<0.0001	Intercept
Pspline. EqRatio_1	1	35.102914	62.7478	62.7478	1348.22	<0.0001	Equivalence Ratio 1
Pspline. EqRatio_2	1	-19.386468	64.6430	64.6430	1388.94	<0.0001	Equivalence Ratio 2
Identity(CpRatio)	1	0.032058	1.4445	1.4445	31.04	<0.0001	Compression Ratio
Opscore(Fuel)	5	0.158388	5.5619	1.1124	23.90	<0.0001	Fuel

以上为回归模型的回归系数与假设检验结果（因输出结果过宽，将其中参数的置信区间等信息省略掉了）。其中，第 2、3 两行上分别为“定量自变量（EqRatio）”的“一次项”与“二次项”的计算结果；第 4 行上为“多值有序自变量（CpRatio）”的计算结果；而第 5 行上为“多值名义自变量（Fuel）”的计算结果，具有 5 个自由度，但只有一个回归系数。

基于上述回归模型计算出定量因变量的预测值作为横坐标的数值，其原始观测值作为纵坐标的数值，绘制出散布图，以直观的方式呈现模型对资料的拟合效果。见图 2。

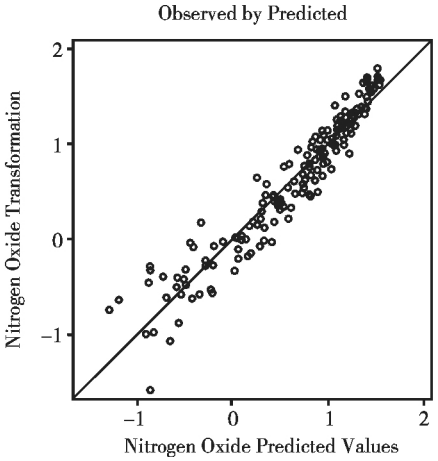

图 2 反映因变量（nox）的观测值与预测值的散布图

由图 2 可知：此回归模型对资料的拟合效果比较令人满意。

4 总 结

4.1 本文方法的小结

本文对“多值名义自变量”进行“优化计分变换（Opscore）”，取代了经典统计学中常规的“哑变量变换”。在此基础上，对“定量因变量（nox）”分别进行了“样条变换（Spline）”“单调样条变换（Mspline）”和“Box-Cox 变换”三种变量变换；对“定量自变量（EqRatio）”分别进行了“样条变换（Spline）”和“单调样条变换（Mspline）”两种变量变换；对“多值有序自变量（CpRatio）”分别进行了“单调变换（Monotone）”和“优化计分变换（Opscore）”。从而，获得“最优模型”的拟合优度评价指标为“均方误差平方根 = 0.30935、R^2 = 0.9586、调整 R^2 = 0.9527”。由于此模型相对比较复杂（自变量有 21 项且写不出具体的回归系数）；简化后的参数多重非线性回归模型仅含 8 个自变量（注：包括派生变量），R^2 为 0.9142，调整 R^2 为 0.9099。

4.2 其他方法的回顾

在本期科研方法专题的前三篇文章中，分别对“多值名义自变量”进行“哑变量变换”“算术均值变换”和“校正均值变换”，对“定量因变量（nox）”进行

了“不变换”“自然对数变换”“平方根变换”“倒数变换”“指数变换”和“Logistic 变换”;对“定量自变量(EqRatio)”和“多值有序自变量(CpRatio)(被视为‘定量的’)”引入了“派生变量”。还在模型中假定“包含截距项”和“不含截距项”两种情形下,分别采取“前进法”“后退法”和“逐步法”筛选自变量。在前三篇文章中的每一篇中,都从大约 72 个模型中选出了 24 个“最优模型”且以“模型 24”为“最佳模型”。

4.3 四组方法的总结

“四组方法”指本期科研方法专题中四篇文章所介绍的方法,即分别对“多值名义自变量”进行“哑变量变换”“算术均值变换”“校正均值变换”和“优化计分变换”。在此基础上,再对其他变量采用多种变量变换方法,最后以“拟合优度评价指标”的取值为判断标准,总结于下表 2。

表 2 四篇文章中“最佳”回归模型拟合优度评价指标汇总

文章编号	自变量数	均方误差平方根	R^2	调整 R^2
第 1 篇	16	0.02757	0.9991	0.9990
第 2 篇	12	0.02449	0.9992	0.9992
第 3 篇	12	0.02470	0.9992	0.9992
第 4 篇	21	0.30935	0.9586	0.9527

注:前三篇文章原先给出的是“均方误差(MSE)”,分别为“0.00076”“0.00060”和“0.00061”,将它们分别开平方根,得到“0.02757”“0.02449”和“0.02470”。由表 2 可知,前三篇文章回归模型的拟合效果非常接近,均优于第 4 篇文章回归模型的拟合效果;然而,第一篇文章回归模型中保留了“部分哑变量”,严格来说,它不是非常理想的结果

由此可以得出如下结论:在对“定量因变量”构建多重非线性回归模型中,对“多值名义自变量”是采取“哑变量变换”“优化计分变换”还是“均值或校正均值变换”,对结果的影响都不是特别大,关键在于以下四点:第一,应对“定量因变量”“定量自变量”和“多值有序自变量”采取合适的变量变换;第二,应尽可能引入较多的“派生变量”;第三,应假定“模型中不含截距项”;第四,应尽可能多采取一些筛选自变量的策略,如前进法、后退法和逐步法。

最后还需要注意:引入派生变量后获得的“最佳”回归模型通常具有“严重多重共线性”,若最终的回归模型中的某些回归系数的“正负号”与专业知识不符合,此时,再基于“最佳回归模型”所确定的“自变量组合”进行“岭回归分析”。从而获得满意的结果[4]。

参考文献

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

[2] Fisher RA. Statistical methods for research workers[M]. 10th Edition. Edinburgh: Oliver & Boyd, 1938: 1 - 124.

[3] Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetrichypothesis[J]. Psychometrika, 1964, 29(1): 1 - 27.

[4] 胡良平. 岭回归分析[J]. 四川精神卫生, 2018, 31(3): 193 - 196.

[5] Box GEP, Cox DR. An analysis of transformations[J]. J R Stati Soc, 1964, 26(2): 211 - 252.

[6] 胡良平. 回归建模的基础与要领(Ⅲ)——变量状态与相互间关系[J]. 四川精神卫生, 2018, 31(6): 493 - 497.

[7] Reinsch CH. Smoothing by spline functions[J]. Numer Math (Heidelb), 1967, 10(3): 177 - 183.

(收稿日期:2019 - 02 - 01)

(本文编辑:陈 霞)