·科研方法专题·

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

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】 本文目的是介绍第一种提高回归模型拟合优度的策略,即哑变量变换与其他变量变换。具体方法包括以下几个方面:①对多值名义自变量采取“哑变量变换”;②对定量和有序自变量引入派生变量,包括“对数变换”“平方根变换”“指数变换”“平方变换”“立方变换”和“交叉乘积变换”的结果;③对定量因变量分别采取“对数变换”“平方根变换”“指数变换”“倒数变换”和“Logistic 变换”;④构建回归模型时,在假定“包含截距项”与“不含截距项”的条件下,分别采取“前进法”“后退法”和“逐步法”筛选自变量。得到了如下几个结论:①对定量因变量和自变量不做变量变换时,回归模型的拟合优度非常差;②根据资料所具备的条件,对定量因变量采取不同的变量变换方法,其回归模型的拟合优度是不尽相同的;③对多值名义自变量进行“哑变量变换”是常规的做法,但存在不足之处;④对定量自变量引入派生变量是非常有价值的;⑤假定回归模型中不含截距项有助于提高回归模型的拟合优度。

【关键词】 变量变换;哑变量变换;Logistic 变换;派生变量;拟合优度
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2019.01.001

Strategy of improving the goodness of fit of the regression model(Ⅰ)
——the transformation of the dummy variable 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 first strategy of improving the goodness of fit of the regression models, the transformation of the dummy variable and the other variable transformations. The concrete approaches were as follows: ①“The transformation of the dummy variable” was adopted to the multi-value nominal independent variable. ②The derived variables were introduced to the quantitative and the ordered independent variables including the results of “logarithmic transformation”“square root transformation”“exponential transformation”“square transformation”“cubic transformation” and “cross product terms transformation”. ③“Logarithmic transformation”“square root transformation”“exponential transformation”“reciprocal transformation” and “Logistic transformation” were adopted to the quantitative dependent variable, respectively. ④During building the regression models, the “forward selection” “backward selection” and “stepwise selection” were used for screening the independent variables under the conditions both with the intercept term and without it, respectively. The several conclusions were achieved as below: ①The goodness of fit of the regression models was very poor when no transformations were applied to the quantitative dependent variable and independent variables. ②The distinct results of the goodness of fit of the regression models could be achieved by using the distinct transformations to the quantitative dependent variable in accordance with the data conditions. ③It was the common measurement to transform the multi-value nominal independent variable into the dummy variables, however, there were disadvantages of the approach mentioned above. ④It was wonderful to introduce the derived variables to the quantitative independent variables in fitting the regression models. ⑤It was helpful to improve the goodness of fit of the regression models by getting rid of the intercept term.

【**Keywords**】 Variable transformation; Transformation of the dummy variable; Logistic transformation; Derived variable; Goodness of fit

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

1 变量变换的必要性及变换方法

1.1 多值名义变量的变量变换

1.1.1 选择合适变量变换方法的必要性

在进行回归分析时,若自变量中有“多值名义变量”(如职业、血型、仪器品牌等),其具体的“表现或水平”不能用“文字”或“字母”表示,也不能简单地赋值“1、2、3……”前者无法参与统计计算,而后者将会导致计算结果错误。那么,究竟应该对“多值名义变量”进行什么样的变量变换呢?本文将介绍常规做法,即进行“哑变量变换”。

在回归分析中应如何处置“多值有序变量”?在统计学上,人们认为:直接采用多值有序变量各水平的数值为其取值,例如:假定 x 代表“肿瘤分级”,依据临床专业知识,已知它可以分为五级,于是,认为 x 的取值就是“1、2、3、4、5”。依据基本常识可知,这样的做法是不妥的。因为当肿瘤处于不同等级,其对结果的影响可能不是“线性关系”,很可能是较复杂的“非线性关系”。因此,应将“多值有序自变量”视为“多值名义自变量”,采用合适的变量变换方法。

1.1.2 对多值名义变量进行“哑变量变换”

所谓哑变量变换,就是将一个具有 k 个水平的多值名义变量转换成($k-1$)个新变量,每个新变量都是一个“二值变量(即仅有两个不同取值的变量)”。这些新变量像“哑巴”一样,其中的每一个都携带着原变量的一部分信息,在计算中发挥一定的作用,但又不能完全取代原变量,故它们都被形象地称为“哑变量”。

实施哑变量变换的方法是:选择一个频率高的水平作为“基准水平”,其他水平都与该基准水平作比较而产生一个“比较变量”(即哑变量)。例如:在 ABO 血型系统中,假定在样本资料中属于 O 型血的人数最多,就可以以“O 型血的人”为“基准水平”,其他三种血型的人相对于 O 型血的人分别产生一个“哑变量”。简化形式呈现如下:

个体编号	血型	$X_{A\|O}$	$X_{B\|O}$	$X_{AB\|O}$
1	A	1	0	0
2	B	0	1	0
3	AB	0	0	1
4	O	0	0	0

在上面的简化形式中,“$X_{A|O}$、$X_{B|O}$、$X_{AB|O}$”这三个变量都是与“血型”这个 4 值名义变量对应的“哑变量”,它们分别代表“是否为 A 型血”“是否为 B 型血”和“是否为 AB 型血”。

1.1.3 对多值名义变量进行“其他变量变换”

在进行回归分析中,上面的“哑变量变换”已经成为统计学界处置“多值名义自变量”的“金标准”。是否还有更合理的“变量变换”方法可以取代“哑变量变换”呢?此问题将在本期“科研方法专题”的另三篇文章中深入讨论。

1.2 定量变量的变量变换

1.2.1 选择合适变量变换方法的必要性

通常情况下,人们在进行回归分析时,对于定量的自变量和/或因变量不作任何变换。然而,由基本常识可知,前述做法是不切实际的,通常情况下,效果是不够好的。因为变量之间的关系往往是错综复杂的,它们之间永远以“一次方”形式存在联系的可能性是非常罕见的。因变量 Y 可能与某个自变量之间是抛物线关系、指数曲线关系或对数曲线关系;因变量 Y 本身可能偏离正态分布很远,而很多统计模型要求因变量必须服从正态分布。因此,需要对定量因变量作合适的变量变换,以使其符合特定统计模型的基本要求;需要对某些定量自变量作合适的变量变换,以更真实地呈现其与定量因变量之间的变化趋势。

1.2.2 对定量自变量进行两方面的变量变换

第一方面的变量变换就是对某定量自变量作了某种变量变换后,丢弃原先的那个自变量,而仅采用变换后的变量。例如:建模时,只用“$\log(x_1)$”,而丢弃“x_1”。第二方面的变量变换就是不仅用变换后的变量,还保留未变换的原变量。这样做的结果会使自变量的数目大大增加,常称为产生“派生变量”。例如:假定有 10 个定量变量,可以给它们都取对数变换,就会增加 10 个新变量;也可以对 10 个变量进行平方变换或平方根变换;还可以基于 10 个定量变量产生交叉乘积项等。

1.2.3 对定量因变量进行变量变换

在通常情况下,人们进行的是“一元多重回归分析”,因此,若对定量因变量进行变量变换,在回归建模时,只使用变换后的因变量,而不会同时使用原先的“因变量”与变换后的因变量(因为这样做已经把“一元”问题转变成“二元”问题了)。

何时需要对定量因变量进行变换呢?通常在以下两种情况之一:其一,已知因变量与自变量之间呈

某种函数关系,就选择相应的变量变换方法。例如:当因变量与自变量之间呈“指数函数”变化关系时,就可以对因变量取对数变换;其二,当定量因变量(严格地说,应该是模型的误差项)偏离正态分布很远时,需要选择一种合适的变量变换方法,目的是使变换后的因变量服从模型所要求的某种概率分布,如正态分布、指数分布或威布尔分布等。

2 实际问题与数据结构

2.1 实际问题

研究者关心的定量结果变量为“氧化氮释放量(nox)”,该定量指标的数值测自单缸发动机。已知影响因素有:燃油种类(fuel)、压缩比(cpratio)和等值比(eqratio)。其中,燃油种类(fuel)是多值名义变量,而氧化氮释放量(nox)、压缩比(cpratio)和等值比(eqratio)都是计量变量。该资料来自 SAS 软件中的“帮助”数据库,数据集名为:sashelp.gas。

试以“氧化氮释放量(nox)”为因变量,以“燃油种类(fuel)、压缩比(cpratio)和等值比(eqratio)”为自变量,创建一元多重回归模型。

【说明】该实际问题和对应的数据来源于“SAS/STAT 的 TRANSREG 过程中的样例及 SASHELP 数据库,其数据集名为 sashelp.gas”[1]。

2.2 数据结构

利用以下 SAS 程序可以显示该例的数据结构:

```
proc print data = sashelp.gas;
run;
```

【燃油资料的数据结构】

Obs	Fuel	CpRatio	EqRatio	NOx
1	Ethanol	12	0.907	3.741
2	Ethanol	12	0.761	2.295
3	Ethanol	12	1.108	1.498
4	Ethanol	12	1.016	2.881
5	Ethanol	12	1.189	0.760

以上显示出数据集的前 5 个观测,全部资料共 171 个观测。其中,在结果变量 nox 上有两个缺失值。

利用如下 SAS 程序可以显示三个自变量(一个为多值名义自变量、一个为多值有序自变量、一个为定量自变量)及定量结果变量(nox)的频数分布情况:

```
proc  freq data = sashelp.gas;
  tables fuel eqratio cpratio nox;
run;
```

【燃油种类的频数分布】

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

以上结果表明:共有 6 种燃油,其中,频数最多的是“Ethanol”,涉及此种燃油的观测共有 90 个。

【压缩比的频数分布】

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

以上结果表明:压缩比只有 5 种,属于“多值有序”变量(注意:以下简称为“定量变量”)。其中,频数最多的是“7.5”,涉及此种压缩比的观测共有 93 个。

等值比(eqratio)与氧化氮释放量(nox)的取值都很多,其频数分布表此处从略;但利用下面的 SAS 程序可以显示这两个变量的频数分布直方图,同时,还可以对它们进行正态性检验:

```
proc univariate data = sashelp.gas normal;
    var eqratio nox;
    histogram eqratio nox/normal;
run;
```

【等值比的正态性检验结果】

正态性检验

检验	统计量		P	
Shapiro – Wilk	W	0.969774	Pr < W	0.0009
Kolmogorov – Smirnov	D	0.063047	Pr > D	0.0941
Cramer – von Mises	W – Sq	0.196943	Pr > W – Sq	0.0058
Anderson – Darling	A – Sq	1.289752	Pr > A – Sq	<0.0050

以上结果表明:等值比不服从正态分布。

等值比的频数分布直方图见图 1。由图 1 可知,等值比呈“负偏态分布”

【氧化氮释放量的正态性检验结果】

正态性检验

检验	统计量		P	
Shapiro – Wilk	W	0.945485	Pr < W	<0.0001
Kolmogorov – Smirnov	D	0.098374	Pr > D	<0.0100
Cramer – von Mises	W – Sq	0.336953	Pr > W – Sq	<0.0050
Anderson – Darling	A – Sq	2.431071	Pr > A – Sq	<0.0050

以上结果表明:氧化氮释放量不服从正态分布。

氧化氮释放量的频数分布直方图见图 2。由图 2 可知:氧化氮释放量呈“正偏态分布”。

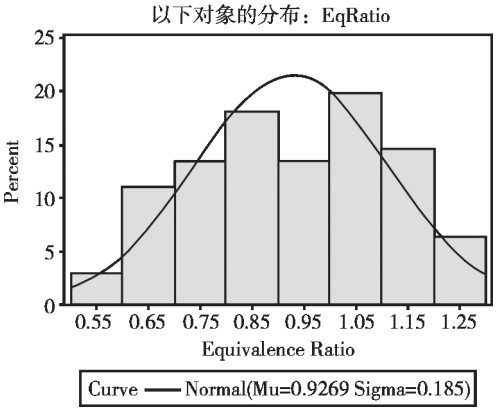

图 1 等值比的频数分布直方图

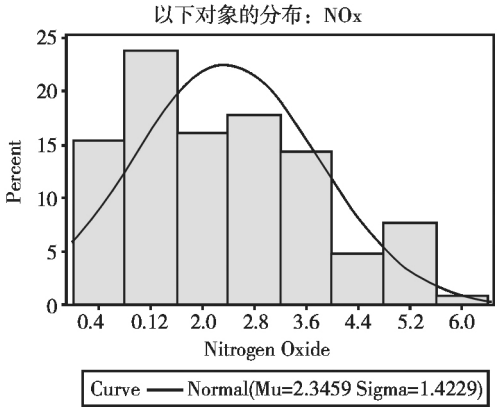

图 2 氧化氮释放量的频数分布直方图

3 变量变换,为回归建模做准备工作

3.1 对“燃油种类(fuel)”这个 6 值名义自变量进行哑变量变换[2]

选择出现频数最多的水平“Ethanol”为“基准”,产生 5 个哑变量:g_1 到 g_5。实现此任务的 SAS 程序如下:

```
data a1;
    set sashelp.gas;
g1=0;g2=0;g3=0;g4=0;g5=0;
if fuel='82rongas' then g1=1;
else if fuel='94% Eth' then g2=1;
else if fuel='Gasohol' then g3=1;
else if fuel='Indolene' then g4=1;
else if fuel='Methanol' then g5=1;
run;
```

g_1 到 g_5 分别代表:“是否为 82rongas 燃油”“是否为 94% Eth 燃油”“是否为 Gasohol 燃油”“是否为 Indolene 燃油”和“是否为 Methanol 燃油”。

3.2 产生派生变量[3]

在数据集 a1 基础上增加由定量自变量派生出来的 13 个新变量,产生数据集 a2。SAS 程序如下:

```
data a2;
    set a1;
x1=eqratio**2;x2=eqratio*cpratio;
x3=cpratio**2;x4=x1*eqratio;
x5=x3*cpratio;x6=x1*cpratio;
x7=x3*eqratio;x8=sqrt(eqratio);
x9=sqrt(cpratio);x10=log(eqratio);
x11=log(cpratio);x12=exp(eqratio);
x13=exp(cpratio);
run;
```

【说明】cpratio 和 eqratio 是资料中两个原始的定量自变量;x_1、x_4、x_8、x_{10}、x_{12} 分别是“eqratio”的平方变换、立方变换、平方根变换、自然对数变换和指数变换的结果;x_3、x_5、x_9、x_{11}、x_{13} 分别是“cpratio”的平方变换、立方变换、平方根变换、自然对数变换和指数变换的结果;x_2 是“eqratio”与“cpratio”的交叉乘积项;x_6 是“eqratio”的平方项与“cpratio”的交叉乘积项;而 x_7 是“cpratio”的平方项与“eqratio”的交叉乘积项。

3.3 对定量因变量进行 5 种变量变换

在数据集 a2 基础上同时增加定量因变量的对

数变换 y_1、平方根变换 y_2、指数变换 y_3、倒数变换 y_4 和 Logistic 变换 y_5，产生数据集 a3。SAS 程序如下：

```
data a3;
    set a2;
    y1 = log(nox);y2 = sqrt(nox);y3 = exp(nox);
    y4 = 1/nox;y5 = exp(nox)/(1 + exp(nox));
run;
```

4 以“哑变量变换”为基础的回归建模

4.1 回归建模策略概述

对一个“多值名义自变量”采取“哑变量变换”，以其为基础，再分别选取定量因变量(nox)的 6 种不同“表现”为每次建模的“因变量”，并对定量自变量在“不做变量变换”和“引入 13 个派生变量”且分别在回归模型中假定“包含截距项”与“不含截距项”的条件下，采取“前进法”“后退法”和“逐步法”筛选自变量。

4.2 定量因变量(nox)的 6 种不同“表现”

定量因变量(nox)的 6 种不同“表现”分别是：①定量因变量(nox)，即对“定量因变量(nox)”不做变量变换；②定量因变量[y_1 = log(nox)]，即对“定量因变量(nox)”做自然对数变换；③定量因变量[y_2 = SQRT(nox)]，即对“定量因变量(nox)”做平方根变换；④定量因变量[y_3 = exp(nox)]，即对“定量因变量(nox)”做指数变换；⑤定量因变量(y_4 = 1/nox)，即对“定量因变量(nox)”做倒数变换；⑥定量因变量{y_5 = exp(nox)/[1 + exp(nox)]}，即对“定量因变量(nox)”做 Logistic 变换。

4.3 在定量因变量(nox)每种“表现”下找出 4 个“最优回归模型”

在定量因变量(nox)每种“表现”且分别在定量自变量“不做变换”与“引入派生变量”的条件下，再在回归模型中假定“包含截距项”与“不含截距项”时，分别采取“前进法”“后退法”和“逐步法”筛选自变量。这实际上就有“2 × 2 × 3 = 12”个回归模型，它们分属于 4 种情形：①“定量自变量不做变换”且假定“包含截距项”；②“定量自变量不做变换”且假定“不含截距项”；③“定量自变量做变换”且假定“包含截距项”；④“定量自变量做变换”且假定“不含截距项”。每种情形都涉及 3 种筛选自变量的方法，最多有 3 种不同的回归模型，从中选取一个拟合最好的回归模型。

所以，在每种特定的因变量条件下，就对应着 4 个“最优回归模型”；故在因变量的 6 种条件下，一共有 24 个“最优回归模型”。见表 1。

表 1 反映 24 个多重回归模型拟合优度的计算结果

模型编号	R^2	调整 R^2	均方误差	Cp 值	自变量个数	有无截距项
第 1 组模型：未对定量因变量做变量变换						
1	0.2275	0.2135	1.59250	3.9363	3	有
2	0.7712	0.7643	1.77176	5.1626	5	无
3	0.8946	0.8887	0.22545	11.2046	9	有
4	0.9724	0.9701	0.22496	14.7623	13	无
第 2 组模型：对定量因变量做自然对数变换						
5	0.1418	0.1262	0.45138	5.3331	3	有
6	0.5050	0.4899	0.46520	4.4310	5	无
7	0.9478	0.9438	0.02903	19.2183	12	有
8	0.9712	0.9686	0.02861	18.8054	14	无
第 3 组模型：对定量因变量做平方根变换						
9	0.1851	0.1703	0.18935	4.6523	3	有
10	0.8932	0.8899	0.25820	7.6867	5	无
11	0.9336	0.9294	0.01612	16.8992	10	有
12	0.9939	0.9933	0.01573	16.8033	14	无
第 4 组模型：对定量因变量做指数变换						
13	0.3240	0.3118	2368.18743	1.1782	3	有
14	0.4738	0.4610	2384.03602	3.0487	5	无
15	0.5406	0.5206	1649.58244	1.7793	7	有
16	0.6487	0.6218	1662.82397	8.1529	12	无

续表 1：

第 5 组模型：对定量因变量做倒数变换						
17	0.0891	0.0781	0.37187	2.5112	2	有
18	0.5830	0.5780	0.37923	2.2523	2	无
19	0.8285	0.8199	0.07265	13.5416	8	有
20	0.9243	0.9185	0.07320	18.7606	12	无
第 6 组模型：对定量因变量做 Logistic 变换						
21	0.0856	0.0746	0.01436	7.06592	2	有
22	0.9545	0.9525	0.03543	7.00000	7	无
23	0.9539	0.9504	0.00077	15.4067	12	有
24	0.9991	0.9990	0.00076	16.1852	16	无

注：第 1 组模型对应的因变量为“氧化氮释放量(nox)”；第 2 组模型对应的因变量为“氧化氮释放量的自然对数变换结果(y_1)”；第 3 组模型对应的因变量为“氧化氮释放量的平方根变换结果(y_2)”；第 4 组模型对应的因变量为“氧化氮释放量的指数变换结果(y_3)”；第 5 组模型对应的因变量为“氧化氮释放量的倒数变换结果(y_4)”；第 6 组模型对应的因变量为“氧化氮释放量的 Logistic 变换结果(y_5)”

5 拟合优度评价标准与评价结果

5.1 回归模型拟合优度高低的评价标准

一般来说，当模型中包含的自变量数目相等且都包含截距项或都不含截距项时，R^2 值越大越好；此时，Cp 值越接近自变量个数越好；当保留在模型中的自变量个数相差较多时，在前述判断方法基础上，再加上“均方误差”(越小越好)和“调整 R^2”(越大越好)，则更好。

5.2 基于“哑变量变换与其他变量变换”回归建模效果的评价

5.2.1 第 1 组模型的拟合效果评价

第 1 组模型对应的因变量为“氧化氮释放量”，模型 1 与模型 2 都是基于“5 个哑变量加上 2 个定量自变量”进行变量筛选，其区别在于模型 1 假定包含截距项，而模型 2 假定不含截距项；模型 3 与模型 4 都是基于“5 个哑变量加上 2 个定量自变量及其 13 个派生变量”进行变量筛选，其区别在于模型 3 假定包含截距项，而模型 4 假定不含截距项。由表 1 中前 4 行结果可知：模型 2 优于模型 1、模型 4 优于模型 3，即在相同情况下，假定不含截距项的拟合结果优于假定包含截距项的拟合结果；进一步比较可知：模型 4 优于模型 2，即引入派生变量的拟合结果优于不引入派生变量的拟合结果。

5.2.2 第 2 组模型的拟合效果评价

第 2 组模型对应的因变量为“氧化氮释放量的自然对数变换结果(y_1)”，模型 5 与模型 6 都是基于“5 个哑变量加上 2 个定量自变量”进行变量筛选，其区别在于模型 5 假定包含截距项，而模型 6 假定不包含截距项；模型 7 与模型 8 都是基于“5 个哑变量加上 2 个定量自变量及其 13 个派生变量”进行变量筛选，其区别在于模型 7 假定包含截距项，而模型 8 假定不包含截距项。由表 1 中第 5～8 行结果可知：模型 6 优于模型 5、模型 8 优于模型 7，即在相同情况下，假定不含截距项的拟合结果优于假定包含截距项的拟合结果；进一步比较可知：模型 8 优于模型 6，即引入派生变量的拟合结果优于不引入派生变量的拟合结果。

5.2.3 第 3 组模型的拟合效果评价

第 3 组模型对应的因变量为“氧化氮释放量的平方根变换结果(y_2)”，模型 9 与模型 10 都是基于“5 个哑变量加上 2 个定量自变量”进行变量筛选，其区别在于模型 9 假定包含截距项，而模型 10 假定不包含截距项；模型 11 与模型 12 都是基于“5 个哑变量加上 2 个定量自变量及其 13 个派生变量”进行变量筛选，其区别在于模型 11 假定包含截距项，而模型 12 假定不包含截距项。由表 1 中第 9～12 行结果可知：模型 10 优于模型 9、模型 12 优于模型 11，即在相同情况下，假定不含截距项的拟合结果优于假定包含截距项的拟合结果；进一步比较可知：模型 12 优于模型 10，即引入派生变量的拟合结果优于不引入派生变量的拟合结果。

5.2.4 第 4 组模型的拟合效果评价

第 4 组模型对应的因变量为“氧化氮释放量的指数变换结果(y_3)”，模型 13 与模型 14 都是基于“5 个哑变量加上 2 个定量自变量”进行变量筛选，其区别在于模型 13 假定包含截距项，而模型 14 假定不包含截距项；模型 15 与模型 16 都是基于“5 个哑变量加上 2 个定量自变量及其 13 个派生变量”进

行变量筛选,其区别在于模型 15 假定包含截距项,而模型 16 假定不包含截距项。由表 1 中第 13 ~ 16 行结果可知:模型 14 优于模型 13、模型 16 优于模型 15,即在相同情况下,假定不含截距项的拟合结果优于假定包含截距项的拟合结果;进一步比较可知:模型 16 优于模型 14,即引入派生变量的拟合结果优于不引入派生变量的拟合结果。

5.2.5 第 5 组模型的拟合效果评价

第 5 组模型对应的因变量为“氧化氮释放量的倒数变换结果(y_4)”,模型 17 与模型 18 都是仅基于“3 个定量自变量”进行变量筛选,其区别在于模型 17 假定包含截距项,而模型 18 假定不包含截距项;模型 19 与模型 20 都是基于“3 个定量自变量及其 18 个派生变量”进行变量筛选,其区别在于模型 19 假定包含截距项,而模型 20 假定不包含截距项。由表 1 中第 17 ~ 20 行结果可知:模型 18 优于模型 17、模型 20 优于模型 19,即在相同情况下,假定不含截距项的拟合结果优于假定包含截距项的拟合结果;进一步比较可知:模型 20 优于模型 18,即引入派生变量的拟合结果优于不引入派生变量的拟合结果。

5.2.6 第 6 组模型的拟合效果评价

第 6 组模型对应的因变量为“氧化氮释放量的 Logistic 变换结果(y_5)”,模型 21 与模型 22 都是仅基于3 个定量自变量进行变量筛选,其区别在于模型 21 假定包含截距项,而模型 22 假定不包含截距项;模型 23 与模型 24 都是基于 3 个定量自变量及其 18 个派生变量进行变量筛选,其区别在于模型 23 假定包含截距项,而模型 24 假定不包含截距项。由表 1 中第 21 ~ 24 行结果可知:模型 22 优于模型 21、模型 24 优于模型 23,即在相同情况下,假定不含截距项的拟合结果优于假定包含截距项的拟合结果;进一步比较可知:模型 24 优于模型 22,即引入派生变量的拟合结果优于不引入派生变量的拟合结果。

5.2.7 各组模型中最优模型拟合优度总评价

从以上的“评价结果”可知:模型 4、模型 8、模型 12、模型 16、模型 20 和模型 24 分别是从 6 组模型中挑选出来的“最优模型”,现将它们从表 1 中摘录出来,以便直观比较和判断。见表 2。

表 2 各组挑选出来的 6 个“最优”多重回归模型拟合优度的计算结果

模型编号	R^2	调整 R^2	均方误差	Cp 值	自变量个数	有无截距项
4	0.9724	0.9701	0.22496	14.7623	13	无
8	0.9712	0.9686	0.02861	18.8054	14	无
12	0.9939	0.9933	0.01573	16.8033	14	无
16	0.6487	0.6218	1662.82397	8.1529	12	无
20	0.9243	0.9185	0.07320	18.7606	12	无
24	0.9991	0.9990	0.00076	16.1852	16	无

由表 2 可知:模型 24 是 6 个“最优”模型中“最佳”的。该模型的因变量为“氧化氮释放量(nox)的 Logistic 变换结果(y_5)”,从全部(5 + 2 + 13 = 20 个)自变量中筛选出了 16 个具有统计学意义的自变量,模型中不含截距项。具体计算结果如下:

方差分析

源	自由度	平方和	均方	F	Pr > F
模型	16	126.03904	7.87744	10431.6	<0.0001
误差	153	0.11554	0.00075515		
未校正合计	169	126.15458			

变量	参数估计值	标准误差	II 型 SS	F	Pr > F
g1	0.05367	0.00999	0.02179	28.86	<0.0001
g3	0.06021	0.00866	0.03650	48.33	<0.0001
g4	0.05957	0.00713	0.05275	69.85	<0.0001
EqRatio	2915.10929	665.61298	0.01448	19.18	<0.0001

CpRatio	-932.92081	221.15537	0.01344	17.79	<0.0001
x1	-591.67642	128.76619	0.01594	21.11	<0.0001
x2	-0.09658	0.04372	0.00369	4.88	0.0287
x3	29.76542	7.05701	0.01343	17.79	<0.0001
x4	93.27029	19.61988	0.01707	22.60	<0.0001
x5	-0.55800	0.13232	0.01343	17.78	<0.0001
x6	0.07200	0.01940	0.01040	13.77	0.0003
x7	-0.00227	0.00112	0.00307	4.07	0.0454
x8	-5597.26653	1310.34945	0.01378	18.25	<0.0001
x9	3191.32640	756.40499	0.01344	17.80	<0.0001
x10	785.94985	188.21633	0.01317	17.44	<0.0001
x13	6.991368E-7	1.657578E-7	0.01343	17.79	<0.0001

输出以上结果的“SAS 过程步程序”如下：

```
/*模型 24:R2 = 0.9991,调整 R2 = 0.9990,MSE = 0.00075515,Cp = 16.1852,niv = 16,无截距项*/
proc reg data = a3;
    model y5 = g1 - g5 eqratio cpratio x1 - x13/noint
    selection = backward sls = 0.05 r;
/*模型 24*/
run;
```

应注意：全部哑变量共有 5 个（它们之间不是互相对立的），采用筛选自变量的方法，保留下来其中的 3 个。严格地说，由一个多值名义自变量产生的全部哑变量应当同时被保留在回归模型中或同时被排除出回归模型，但这两种结局都存在局限性；而将有关联性的 5 个哑变量视为“独立”的，根据假设检验结果保留其中的 3 个，这个结果也存在弊端。如何更妥善地处置“多值名义自变量”，将在本期科研方法专题后续文章中继续讨论。

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] Kleinbaum DG, Kupper LL, Muller KE, et al. Applied regression analysis and other multivariable methods[M]. 3 版. 北京：机械工业出版社, 2006: 317-360.

[3] 谷恒明，胡良平. 基于经典统计思想实现多重线性回归分析[J]. 四川精神卫生, 2018, 31(1): 7-11.

（收稿日期：2019-02-01）

（本文编辑：陈 霞）

科研方法专题策划人——胡良平教授简介

胡良平，男，1955 年 8 月出生，教授，博士生导师，曾任军事医学科学院研究生部医学统计学教研室主任和生物医学统计学咨询中心主任、国际一般系统论研究会中国分会概率统计系统专业理事会常务理事和北京大学口腔医学院客座教授；现任世界中医药学会联合会临床科研统计学专业委员会会长、中国生物医学统计学会副会长，《中华医学杂志》等 10 余种杂志编委和国家食品药品监督管理局评审专家。主编统计学专著 48 部，参编统计学专著 10 部；发表第一作者学术论文260余篇，发表合作论文 130 余篇，获军队科技成果和省部级科技成果多项；参加并完成三项国家标准的撰写工作；参加三项国家科技重大专项课题研究工作。在从事统计学工作的 30 年中，为几千名研究生、医学科研人员、临床医生和杂志编辑讲授生物医学统计学，在全国各地作统计学学术报告 100 余场，举办数十期全国统计学培训班，培养多名统计学专业硕士和博士研究生。近几年来，参加国家级新药和医疗器械项目评审数十项、参加 100 多项全军重大重点课题的统计学检查工作。归纳并提炼出有利于透过现象看本质的“八性”和“八思维”的统计学思想，独创了逆向统计学教学法和三型理论。擅长于科研课题的研究设计、复杂科研资料的统计分析与 SAS 实现、各种层次的统计学教学培训和咨询工作。