提高回归模型拟合优度的策略(Ⅱ)
——算术均值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】 本文目的是介绍第二种提高回归模型拟合优度的策略,即算术均值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"算术均值变换";②对定量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同的;③对多值名义自变量进行"算术均值变换"是合理的,且有助于提高回归模型拟合优度;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;算术均值变换;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.1007-3256.2019.01.002

Strategy of improving the goodness of fit of the regression model(Ⅱ)
——the transformation of the arithmetic mean 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 second strategy of improving the goodness of fit of the regression models, the transformation of the arithmetic mean and the other variable transformations. The concrete approaches were as follows: ①"The transformation of the arithmetic mean" was adopted to the multi-value nominal independent variable. ②The derived variables were introduced to the quantitative independent variables, such as 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. The several conclusions were achieved as below: ①The goodness of fit of the regression models was very low when no transformations was 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 rational to transform the multi-value nominal independent variable by using the arithmetic mean transformation, which was conducive to improving the goodness of fit of the regression models. 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 arithmetic mean; Logistic transformation; Derived variable; Goodness of fit

1 问题的提出

1.1 "哑变量变换"存在一些弊端

在进行回归分析时,对"多值名义(也包括多值有序)自变量"进行"哑变量变换"已成为统计学界公认的"标准做法"。若采用某种筛选自变量的方法,就可能保留一部分哑变量而淘汰掉另一部分哑变量,这样做的前提是"假定它们互相独立",但事实上,源自一个多值名义自变量的多个哑变量之间是有联系的,即它们之间并不满足"互相独立"的要

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

求;按理说,由一个多值名义自变量产生的全部哑变量,要么全部被保留在回归模型中,要么全部被剔除到回归模型之外。那么问题又出现了:当它们全部出现在回归模型中时,有些哑变量可能就没有统计学意义,这样的回归模型就不是最节俭的,因此,回归模型的拟合优度不高;若全部剔除这些哑变量,就相当于没有发挥该多值名义自变量的作用。总之,“哑变量变换”存在一些弊端,是一个值得深究的“统计疑难问题”。

1.2 用“算术均值变换”取代“哑变量变换”

1.2.1 何为“算术均值变换”

所谓“算术均值变换”,就是先求出某个多值名义自变量 x 的第 i(i = 1、2、…、k)个水平条件下定量因变量的算术平均值“$\bar{y}_i$”,采用该“算术均值”代替多值名义自变量 x 的第 i(i = 1、2、…、k)个水平,这实际上就是把一个“多值名义自变量”变换成为一个具有 k 个具体数值的“定量自变量”了。

1.2.2 用“算术均值变换”取代“哑变量变换”的合理性

将一个“多值名义自变量”变换成一个“定量自变量”,这似乎很不合理。若孤立地考察“多值名义自变量”各水平之间的关系,前述的变换的确很不合理。因为“多值名义自变量”的各水平之间只是“名称或符号上的不同(例如 ABO 血型系统中的 A 型、B 型、AB 型和 O 型血型之间的关系)”,而没有“数量或程度上的差别”。然而,既然叫做“自变量”,它一定要影响其他变量,其中,定量结果变量(也叫定量因变量)是进行定量资料差异性分析和回归分析时最关注的“变量”。换言之,无论是进行 t 检验或方差分析,还是进行回归分析或判别分析,永远不可能孤立地去考察“自变量”各水平之间的关系,而不可避免地要建立起“自变量各水平”与因变量的具体取值之间的数量联系。

在对定量资料进行差异性分析(如 t 检验或方差分析)时,统计学上也是这样做的。例如:设“药物种类(因素 A)”有两个水平:试验药(A_1 水平)和对照药(A_2 水平)。两个药物组都有较多数量的受试对象参与试验,假定评价药物疗效的指标为“收缩压下降值”,测定并计算出两组“收缩压下降值”的“算术平均值”。显然,这就是采用了各自的“算术平均值”代替“试验药(A_1 水平)”与“对照药(A_2 水平)”的“疗效”水平。显而易见,“A_1”与“A_2”之间是没有数量大小之分的,但它们各自组中疗效的“算术平均值”是有数量大小之分的。在多因素定量资料的方差分析中,情况也是如此。

由此可知,“算术均值变换”是用“动态思维(利用了因果关系)”建立起自变量各水平与定量结果变量之间数量关系的一种处理方法;而“哑变量变换”是用“静态思维(孤立地看待自变量)”处置自变量各水平之间关系的方法,此法割裂了“多值名义自变量”与“定量因变量”之间的数量联系。因此,“算术均值变换”比“哑变量变换”更具有合理性。

1.3 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],此处不再赘述。

2 解决问题的思路和做法

2.1 对自变量和因变量的处置方法

2.1.1 对“燃油种类(fuel)”这个“6 值名义自变量”进行“算术均值变换”

求出“燃油种类(fuel)”各水平下“氧化氮释放量(nox)”的算术均值所需要的 SAS 程序如下:

```
/* 下面的 SAS 程序计算出多值名义变量各水平下定量因变量的算术均值 */
data aa;
set sashelp.gas;
proc sort data = aa;
    by fuel;
run;
proc univariate data = aa noprint;
    var nox;
    by fuel;
    output   out = aaa means = m_nox;
run;
proc print data = aaa;
run;
```

【SAS 输出结果】

Obs	Fuel	m_nox
1	82rongas	3.52589
2	94% Eth	2.08788
3	Ethanol	1.95738
4	Gasohol	3.34938
5	Indolene	3.54659
6	Methanol	1.55967

以上就是 6 种燃油对应的“氧化氮释放量(nox)”的算术均值。

将“燃油种类(fuel)”的各水平变换成与定量因变量相应的“算术均值”所需要的 SAS 程序如下:

```
/*下面的 SAS 程序将多值名义变量各水平变换成与定量因变量相应的算术均值*/
data a1;
    set sashelp.gas;
if fuel='82rongas' then mfuel=3.52589;
else if fuel='94%Eth' then mfuel=2.08788;
else if fuel='Gasohol' then mfuel=3.34938;
else if fuel='Indolene' then mfuel=3.54659;
else if fuel='Methanol' then mfuel=1.55967;
else if fuel='Ethanol' then mfuel=1.95738;
run;
```

通过运行上面的 SAS 程序,在原数据集 sashelp.gas 中就增加了一个定量自变量 mfuel,将用它取代多值名义自变量“燃油种类(fuel)”。

2.1.2 对定量因变量和自变量不进行任何变换

在进行回归分析时,通常都不对定量因变量和自变量做任何变换。然而,由基本常识和统计学知识可知,这样的建模结果往往不够理想。

2.1.3 仅对定量自变量进行变换

善于思考问题的分析者会依据探索性分析结果,对某些定量自变量进行合适的变量变换,以获得更好的回归建模效果。包括单一变换,如进行“对数变换、平方根变换、指数变换等”之一的变换;或引入派生变量,即同时使用多项变量变换的结果,如引入某变量的“平方项、立方项、交叉乘积项等”。

2.1.4 仅对定量因变量进行变换

在进行简单直线回归分析和/或多重回归分析之前,应考察“误差项应服从正态分布”的前提条件是否成立。若不成立,就需要通过探索性分析了解定量因变量可能的分布规律,从而采取合适的变量变换方法,以使其基本满足或近似满足进行相应回归分析的前提条件。然而,原始数据中并没有“误差项”这个变量,只能先假定资料满足回归分析的前提条件,基于创建的回归模型计算出因变量的预测值,再计算出各观测点上的“残差”,用此“残差”取代统计学理论上所指的“误差”,最后再检验“误差”是否服从正态分布。

由此可知:在很多实际问题中,有必要对定量因变量进行某些可能的变量变换,以获得更好的拟合效果。

2.1.5 同时对定量自变量和因变量进行变换

事实上,在对数据进行回归分析之前,除了对定性自变量进行必要的变量变换以外,还应进行一系列探索性分析,以了解定量因变量与定量自变量各自的分布情况,以及定量因变量与各定量自变量之间的相互关系和变化趋势,以便对它们分别选择不同的变量变换方法。目的是获得在专业上和统计学上都成立的最佳回归模型。

2.2 对定量自变量和因变量进行变量变换的方法

2.2.1 对定量自变量进行多种变量变换,以便产生派生变量

所需要的 SAS 程序如下:

```
/*在数据集 a1 的基础上增加定量自变量的各种派生变量 18 个,形成数据集 a2*/
data a2;
    set a1;
x1=log(cpration);x2=sqrt(cpration);x3=exp(cpration);
x4=cpratio**2;x5=x4*cpratio;
w1=log(eqration);w2=sqrt(eqration);w3=exp(eqration);
w4=eqratio**2;w5=w4*eqratio;
z1=log(mfuel);z2=sqrt(mfuel);z3=exp(mfuel);
z4=mfuel**2;z5=z4*mfuel;
m1=cpration*eqration;m2=cpration*mfuel;
m3=eqration*mfuel;
run;
```

运行以上 SAS 程序后,就创建了数据集 a2,它在数据集 a1 基础上增加了由三个定量自变量“cpratio”“eqratio”和“mfuel”派生出来的 18 个新自变量,它们分别是每个定量自变量的自然对数变换、平方根变换、指数变换、平方变换和立方变换的结果;还有三个定量自变量两两交叉乘积变换的结果。

2.2.2 对定量因变量进行 5 种变量变换

所需要的 SAS 程序如下:

```
/*在数据集 a2 的基础上增加定量因变量的 5 种变量变换结果,形成数据集 a3*/
data a3;
    set a2;
    y1=log(nox);y2=sqrt(nox);y3=exp(nox);
```

y4 = 1/nox;y5 = exp(nox)/(1 + exp(nox));
run;

运行以上 SAS 程序后,就创建了数据集 a3,它在数据集 a2 基础上增加了由定量因变量“nox”派生出来的 5 个新因变量,它们分别是自然对数变换(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)的结果。

2.3 基于“算术均值变换”的建模策略

【说明】在以下的建模策略中,先对多值名义自变量进行“算术均值变换”;然后,在下面的每种情形中都将分别在“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。

2.3.1 以“氧化氮释放量(nox)”为定量因变量

以“cpratio”“eqratio”和“mfuel”为三个定量自变量建模,在“包含截距项”与“不含截距项”的条件下,选取最佳模型编号分别为“模型 1”与“模型 2”。

以“cpratio”“eqratio”“mfuel”和 18 个派生变量为定量自变量建模,在“包含截距项”与“不含截距项”的条件下,选取最佳模型编号分别为“模型 3”与“模型 4”。

2.3.2 以“氧化氮释放量的自然对数变换结果(y_1)”为定量因变量

以“cpratio”“eqratio”和“mfuel”为三个定量自变量建模,在“包含截距项”与“不含截距项”的条件下,选取最佳模型编号分别为“模型 5”与“模型 6”。

以“cpratio”“eqratio”“mfuel”和 18 个派生变量为定量自变量建模,在“包含截距项”与“不含截距项”的条件下,选取最佳模型编号分别为“模型 7”与“模型 8”。

2.3.3 以“氧化氮释放量的平方根变换结果(y_2)”为定量因变量

以“cpratio”“eqratio”和“mfuel”为三个定量自变量建模,在“包含截距项”与“不含截距项”的条件下,选取最佳模型编号分别为“模型 9”与“模型 10”。

以“cpratio”“eqratio”“mfuel”和 18 个派生变量为定量自变量建模,在“包含截距项”与“不含截距项”的条件下,选取最佳模型编号分别为“模型 11”与“模型 12”。

2.3.4 以“氧化氮释放量的指数变换结果(y_3)”为定量因变量

以“cpratio”“eqratio”和“mfuel”为三个定量自变量建模,在“包含截距项”与“不含截距项”的条件下,选取最佳模型编号分别为“模型 13”与“模型 14”。

以“cpratio”“eqratio”“mfuel”和 18 个派生变量为定量自变量建模,在“包含截距项”与“不含截距项”的条件下,选取最佳模型编号分别为“模型 15”与“模型 16”。

2.3.5 以“氧化氮释放量的倒数变换结果(y_4)”为定量因变量

以“cpratio”“eqratio”和“mfuel”为三个定量自变量建模,在“包含截距项”与“不含截距项”的条件下,选取最佳模型编号分别为“模型 17”与“模型 18”。

以“cpratio”“eqratio”“mfuel”和 18 个派生变量为定量自变量建模,在“包含截距项”与“不含截距项”的条件下,选取最佳模型编号分别为“模型 19”与“模型 20”。

2.3.6 以“氧化氮释放量的 Logistic 变换结果(y_5)”为定量因变量

以“cpratio”“eqratio”和“mfuel”为三个定量自变量建模,在“包含截距项”与“不含截距项”的条件下,选取最佳模型编号分别为“模型 21”与“模型 2”。

以“cpratio”“eqratio”“mfuel”和 18 个派生变量为定量自变量建模,在“包含截距项”与“不含截距项”的条件下,选取最佳模型编号分别为“模型 23”与“模型 24”。

3 基于“算术均值变换与其他变量变换”的回归建模结果与评价

3.1 各种回归建模策略下所得主要结果的汇总

以摘要形式呈现选出的 24 个拟合效果较好的回归模型。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	0.2343	0.2297	1.55962	2.53246	1	有
2	0.7950	0.7937	1.55034	0.3533	1	无
3	0.9089	0.9031	0.19613	12.5764	10	有
4	0.9767	0.9745	0.19193	13.2056	15	无
第 2 组模型:对定量因变量做自然对数变换						
5	0.1533	0.1482	0.44001	1.8429	1	有
6	0.5281	0.5224	0.43559	1.1005	2	无
7	0.9639	0.9612	0.02006	14.3512	12	有
8	0.9793	0.9777	0.02031	15.3478	12	无
第 3 组模型:对定量因变量做平方根变换						
9	0.1945	0.1897	0.18493	2.4286	1	有
10	0.9143	0.9133	0.20334	3.0089	2	无
11	0.9522	0.9486	0.01174	10.6073	12	有
12	0.9954	0.9949	0.01187	14.3093	15	无
第 4 组模型:对定量因变量做指数变换						
13	0.2969	0.2927	2433.77730	1.0891	1	有
14	0.4373	0.4306	2503.72137	2.6797	2	无
15	0.5197	0.5019	1713.99290	2.9472	6	有
16	0.6323	0.6067	1729.47348	8.4376	11	无
第 5 组模型:对定量因变量做倒数变换						
17	0.0827	0.0772	0.37226	0.2193	1	有
18	0.5566	0.5512	0.40327	2.9529	2	无
19	0.8452	0.8365	0.06596	7.0803	9	有
20	0.9317	0.9265	0.06606	9.3476	12	无
第 6 组模型:对定量因变量做 Logistic 变换						
21	0.1096	0.1043	0.01390	2.8698	1	有
22	0.9666	0.9660	0.02536	3.0000	3	无
23	0.9651	0.9621	0.00059	15.4619	13	有
24	0.9992	0.9992	0.00060	14.9351	12	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)”

3.2 基于“算术均值变换与其他变量变换”回归建模效果的分组评价

3.2.1 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 都是仅基于 3 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 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 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。

3.2.2 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 都是仅基于 3 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 假定不含截距项;模型 7 与模型 8 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 7 假定包含截距项,而模型 8 假定不含截距项。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。

3.2.3 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 都是仅基于 3 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 假定不含截距项;模型 11 与模型 12 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 11 假定包含截距项,而模型 12 假定不含截距项。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。

3.2.4 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 都是仅基于 3 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 假定不含截距项;模型 15 与模型 16 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 15 假定包含截距项,而模型 16 假定不含截距项。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。

3.2.5 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。

3.2.6 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。

3.3 对各组模型中挑选出来的最优模型再进行拟合优度的总评价

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	0.9767	0.9745	0.19193	13.2056	15	无
8	0.9793	0.9777	0.02031	15.3478	12	无
12	0.9954	0.9949	0.01187	14.3093	15	无
16	0.6323	0.6067	1729.47348	8.4376	11	无
20	0.9317	0.9265	0.06606	9.3476	12	无
24	0.9992	0.9992	0.00060	14.9351	12	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“氧化氮释放量的 Logistic 变换结果(y_5)”,从全部(3+18=21 个)自变量中筛选出了 12 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	12	126.05961	10.50497	17367.6	<0.0001
误差	157	0.09496	0.00060486		
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*
CpRatio	−0.30528	0.15101	0.00247	4.09	0.0449
EqRatio	−294.70778	32.77512	0.04890	80.85	<0.0001
x2	1.51695	0.69247	0.00290	4.80	0.0300
x4	0.00458	0.00206	0.00300	4.97	0.0273
w1	87.09856	9.41143	0.05180	85.65	<0.0001
w2	−349.83874	38.95857	0.04877	80.64	<0.0001
w3	284.68974	31.70312	0.04877	80.64	<0.0001
w5	−130.57047	14.46523	0.04928	81.48	<0.0001
z2	0.34215	0.07843	0.01151	19.03	<0.0001
z5	0.00236	0.00098863	0.00346	5.71	0.0180
m1	−0.02414	0.00295	0.04037	66.74	<0.0001
m3	−0.13225	0.01839	0.03129	51.72	<0.0001

输出以上结果的“SAS 过程步程序”如下：

```
/*模型 24:R2 = 0.9992,调整 R2 = 0.9992,MSE = 0.00060486,Cp = 14.9351,niv = 12,无截距项*/
proc reg data = a3;
  model y5 = cpratio eqratio mfuel x1 - x5 w1 - w5 z1 - z5 m1 - m3/noint selection = backward sls = 0.05 r;
/*模型 24*/
run;
```

3.4 小结

在对定量因变量构建多重回归模型的过程中，摒弃了传统统计思维下的理论和方法（对定量因变量和定量自变量保持一次方形式，即不做任何变量变换，也不产生派生变量；对多值名义自变量进行哑变量变换，构建所谓的“多重线性回归模型”），而引入了动态统计思维下的理论和方法（对定量因变量分别采取不做变量变换和进行对数变换、平方根变换、指数变换、倒数变换和 Logistic 变换）；淘汰了对“定量自变量不做任何变换”和“永远固定为一次方形式”的僵化思维，不仅对其做“对数变换、平方根变换和指数变换”，还引入了“平方项、立方项和交叉乘积项”；本文提出了一种新的变量变换方法，即对“多值名义自变量”进行“算术均值变换”，不仅将其变换成“定量自变量”，还产生出多项派生变量。

由表 1 和表 2 可知：基于传统统计思维创建的回归模型拟合效果非常差（模型编号分别为 1、5、9、13、17、21），而基于动态统计思维创建的回归模型拟合效果很好（表 2 中除模型 16 之外）。其中，“算术均值变换”“引入派生变量”“假定回归模型中不含截距项”和“找到定量因变量合适的变量变换方法（就本文实例而言，除了‘指数变换’外，其他 4 种变量变换方法的拟合效果都相当好，其中，最佳的是 Logistic 变换）”是动态统计思维建模策略中的“核心”。

有兴趣的读者还可以在本文的基础上，对定量自变量增加其他一些变量变换（例如倒数变换、Logistic 变换等），并将它们作为“派生变量”引入回归建模的过程中，有可能获得拟合效果更好的回归模型。

本文构建的是“多重参数非线性回归模型”，从拟合优度上来看，可与现代的“机器学习”[2-4]建模效果媲美；众所周知，“机器学习”回归建模的结果，只有“误差”可以达到足够小的程度这个“唯一优势”，而几乎无法写出其回归模型。即便能写出回归模型，其“参数的个数”可能接近“无穷大”。从模型必须“精简实用”且“便于呈现”的角度来考量，“机器学习”回归建模效果似乎要逊色于本文介绍方法的建模效果。

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] 谷恒明，胡良平. 基于机器学习统计思想实现多重线性回归分析[J]. 四川精神卫生, 2018, 31(1): 15 - 18.

[3] 吴喜之. 复杂数据统计方法——基于 R 的应用[M]. 3 版. 北京：中国人民大学出版社, 2015: 41 - 56.

[4] 薛薇. R 语言数据挖掘方法及应用[M]. 北京：电子工业出版社, 2016: 142 - 225.

（收稿日期：2019 - 02 - 01）
（本文编辑：陈 霞）