# 提高回归模型拟合优度的策略(Ⅲ)
# ——校正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】 本文目的是介绍第三种提高回归模型拟合优度的策略,即校正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“校正均值变换”;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低;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“校正均值变换”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;校正均值变换;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.1007-3256.2019.01.003

## Strategy of improving the goodness of fit of the regression model(Ⅲ)
## ——the transformation of the corrected 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 third strategy of improving the goodness of fit of the regression model, the transformation of the corrected mean and the other variable transformations. The concrete approaches were as follows: ①“The transformation of the corrected 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 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 gotten 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 corrected 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 corrected mean; Logistic transformation; Derived variable; Goodness of fit

## 1 问题的提出

### 1.1 “算术均值变换”有改进的余地

本期科研方法专题的《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》一文(以下简称“前文”)提出了对“多值名义自变量(包括多值有序自变量)”进行“算术均值变换”的处理方法,配合其他变量变换方法(指对定量自变量与因变量的多种变量变换方法),较好地提高了回归模型的拟合优度。然而,“算术均值变换”有改进

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

的余地。因为在计算多值名义自变量各水平组中定量因变量的均值时,其“隐含前提条件”是“没有其他自变量”或“其他自变量的影响完全相同”。事实上,在具有多自变量的回归分析资料中,除了“多值名义自变量或多值有序自变量或二值自变量”外,还有多个“定量自变量”,而且,没有理由认为这些“定量自变量”满足前面述及的“隐含前提条件”。需要借助“协方差分析”的计算方法消除“其他定量自变量”对定量因变量的影响。此时,求出的“多值名义自变量”各水平下“定量因变量的均值”被称为“校正均值”。采用“校正均值”取代“算术均值”的方法被称为“校正均值变换”法。

### 1.2 用“校正均值变换”取代“哑变量变换”

#### 1.2.1 何为“校正均值变换”

所谓“校正均值变换”,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校正平均值“$\bar{y}_{Ci}$”,采用该“校正均值”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。

求“校正均值”的具体方法是借助“协方差分析”,在方差分析模型中,分组变量为“多值名义自变量”,它是一个“定性影响因素”,再将其他的定量自变量一同写入“一般线性模型”,以“定量因变量”为协方差分析的“因变量”。在获得“协方差分析”结果之前,需要计算出“消除了其他定量自变量影响”条件下“定性影响因素”各水平下“定量结果变量”的均值,它们被称为“校正均值”。

#### 1.2.2 用“校正均值变换”取代“哑变量变换”的合理性

前文已用了较大篇幅陈述了用“算术均值变换”取代“哑变量变换”的合理性。由上面的内容可知,“校正均值”比未校正的“算术均值”更合理。故用“校正均值变换”取代“哑变量变换”的合理性,也就不言而喻了。

### 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“校正均值变换”

求出“燃油种类(fuel)”各水平下“氧化氮释放量(nox)”的校正均值所需要的 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校正均值*/
data aa;
set sashelp.gas;
proc glm data=aa;
    class fuel;
    model nox=fuel cpratio eqratio;
    lsmeans fuel;
run;
```

【说明】上面的“lsmeans 语句”就是希望输出“多值名义自变量(fuel)”各水平下的“校正均值”。

【SAS 输出结果】

校正均值

| Fuel | nox LSMEAN |
|---|---|
| 82rongas | 3.51852959 |
| 94% Eth | 2.11619735 |
| Ethanol | 1.95691843 |
| Gasohol | 3.32335605 |
| Indolene | 3.54233497 |
| Methanol | 1.54554003 |

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校正均值*/
data a1;
    set sashelp.gas;
if fuel='82rongas' then mfuel=3.51853;
else if fuel='94% Eth' then mfuel=2.11620;
else if fuel='Gasohol' then mfuel=3.32336;
else if fuel='Indolene' then mfuel=3.54233;
else if fuel='Methanol' then mfuel=1.545540;
else if fuel='Ethanol' then mfuel=1.95692;
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的处置方法

“对定量因变量和自变量不做任何变换”“仅对定量自变量做变换”“仅对定量因变量做变换”和“同时对定量自变量和因变量做变换”，这几种情况的“具体含义”参见“前文”“第 2.1.2 节至第 2.1.5 节”，此处从略。

### 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“校正均值变换”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“校正均值变换”；然后在以下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。

建模策略将由以下六部分组成：①以“氧化氮释放量(nox)”为定量因变量；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；③以“氧化氮释放量的平方根变换结果($y_2$)”为定量因变量；④以“氧化氮释放量的指数变换结果($y_3$)”为定量因变量；⑤以“氧化氮释放量的倒数变换结果($y_4$)”为定量因变量；⑥以“氧化氮释放量的Logistic变换结果($y_5$)”为定量因变量。以上 6 种建模策略的具体内容与“前文第 2.3.1 节至第 2.3.6 节”完全相同，此处从略。

## 3 基于“校正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 个拟合较好的回归模型见表 1。

### 3.2 基于“校正均值变换与其他变量变换”回归建模效果的分组评价

这部分有六个小标题，其内容与“前文”中相应部分完全相同，详见“前文第 3.2.1 节至第 3.2.6 节”，此处从略。

### 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 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 | 0.2342 | 0.2296 | 1.55980 | 2.5393 | 1 | 有 |
| 2 | 0.7949 | 0.7937 | 1.55053 | 0.4311 | 1 | 无 |
| 3 | 0.9082 | 0.9024 | 0.19758 | 13.5328 | 10 | 有 |
| 4 | 0.9767 | 0.9744 | 0.19243 | 13.2981 | 15 | 无 |
| 第 2 组模型:对定量因变量做自然对数变换 | | | | | | |
| 5 | 0.1540 | 0.1489 | 0.43966 | 1.8605 | 1 | 有 |
| 6 | 0.5285 | 0.5229 | 0.43518 | 1.1046 | 2 | 无 |
| 7 | 0.9639 | 0.9611 | 0.02007 | 14.4014 | 12 | 有 |
| 8 | 0.9791 | 0.9775 | 0.02048 | 16.6727 | 12 | 无 |
| 第 3 组模型:对定量因变量做平方根变换 | | | | | | |
| 9 | 0.1948 | 0.1900 | 0.18485 | 2.4588 | 1 | 有 |
| 10 | 0.9146 | 0.9135 | 0.20284 | 2.8796 | 2 | 无 |
| 11 | 0.9519 | 0.9482 | 0.01182 | 11.2034 | 12 | 有 |
| 12 | 0.9954 | 0.9949 | 0.01192 | 14.5316 | 15 | 无 |
| 第 4 组模型:对定量因变量做指数变换 | | | | | | |
| 13 | 0.2953 | 0.2911 | 2439.35922 | 1.1146 | 1 | 有 |
| 14 | 0.4361 | 0.4293 | 2509.25255 | 2.7003 | 2 | 无 |
| 15 | 0.5182 | 0.5003 | 1719.26811 | 3.4300 | 6 | 有 |
| 16 | 0.6310 | 0.6053 | 1735.33667 | 8.9613 | 11 | 无 |
| 第 5 组模型:对定量因变量做倒数变换 | | | | | | |
| 17 | 0.0837 | 0.0782 | 0.37185 | 0.2091 | 1 | 有 |
| 18 | 0.5566 | 0.5512 | 0.40327 | 2.9894 | 2 | 无 |
| 19 | 0.8450 | 0.8362 | 0.06607 | 7.8685 | 9 | 有 |
| 20 | 0.9308 | 0.9255 | 0.06697 | 12.0038 | 12 | 无 |
| 第 6 组模型:对定量因变量做 Logistic 变换 | | | | | | |
| 21 | 0.1104 | 0.1051 | 0.01389 | 2.8846 | 1 | 有 |
| 22 | 0.9668 | 0.9661 | 0.02527 | 3.0000 | 3 | 无 |
| 23 | 0.9650 | 0.9621 | 0.00059 | 12.4498 | 13 | 有 |
| 24 | 0.9992 | 0.9992 | 0.00061 | 15.7079 | 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($y_5$)”

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 | 0.9744 | 0.19243 | 13.2981 | 15 | 无 |
| 8 | 0.9791 | 0.9775 | 0.02048 | 16.6727 | 12 | 无 |
| 12 | 0.9954 | 0.9949 | 0.01192 | 14.5316 | 15 | 无 |
| 16 | 0.6310 | 0.6053 | 1735.33667 | 8.9613 | 11 | 无 |
| 20 | 0.9308 | 0.9255 | 0.06697 | 12.0038 | 12 | 无 |
| 24 | 0.9992 | 0.9992 | 0.00061 | 15.7079 | 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做 5 种变量变换,即对数变换、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“校正均值变换”,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的拟合效果非常差(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的拟合效果很好(表 2 中除模型 16 之外)。其中,“校正均值变换”“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。

有兴趣的读者可以运用文献[2 - 4]中“机器学习”算法,实现本文资料的回归建模,并将其拟合结果与本文进行比较,从而发现不同分析方法的优缺点。

## 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)