基于正交化方法的回归分析

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】 本文目的是介绍基于正交化方法的回归分析的概念、作用以及用软件实现计算的方法。先介绍有关的基本概念,再介绍基本原理,最后通过两个实例并基于 SAS 软件演示如何实施此分析方法。结果表明:①此法不能解决资料中存在多重共线性问题带来的坏影响;②此法能够很好地解决多项式回归分析问题。

【关键词】 正交化方法;多重共线性;病态数据;正交变换
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.03.002

Regression analysis based on the orthogonalization approach

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**】 The purpose of this paper was to introduce the concepts and functions and the calculation methods by using the statistical software of the regression analysis based on the orthogonalization approach. Firstly, the basic concepts of the regression analysis was introduced. Secondly, the basic principle of the regression analysis was given. Finally, the regression analysis based on the orthogonalization approach was demonstrated through two examples by using the SAS software. The two distinct conclusions could be drawn below: ① The method mentioned above could not solve the problem of the regression analysis of the data with the multiple collinearity. ② The method mentioned above could perfectly solve the problem of the polynomial regression analysis.

【**Keywords**】 Orthogonalization approach; Multiple collinearity; Ill-conditioned data; Orthogonal transformation

1 概 述

1.1 何为基于正交化方法的回归分析

基于正交化方法的回归分析是在构建多重线性回归模型时,先对数据进行“Gentleman-Givens 变换”[1-2],但仍基于“线性最小二乘原理”推导出的公式估计回归系数。

1.2 基于正交化方法的回归分析应用的场合

当拟做多重线性回归分析的原始数据存在严重的“病态”时,采用此方法构建多重线性回归模型,可以最大限度地消除“病态数据”对建模结果造成的影响。

1.3 基于正交化方法的回归分析的原理

此法使用“Gentleman-Givens 变换”[3]校正数据,并且在计算数据矩阵的 QR 分解[4]的上三角矩阵 R 时十分谨慎。相对于其他正交化方法(例如 Householder 变换[3]),此法的优点是不需要将数据矩阵存储在计算机的内存中。

2 基于正交化方法的回归分析解决实际问题

2.1 此法可否解决多重共线性问题

2.1.1 问题与数据结构

沿用文献[5]中的“问题与数据”,并基于派生变量得到的“最优回归模型”所决定的“数据集”,来提出下面的“新问题”:即“weight”的回归系数为“-88.00801”,这个“负值”表明:体重越重的人收缩压(SBP)越低,这似乎不符合临床专业知识。尽管计算出来的因变量的预测值在专业上都成立,而且模型的残差方差=122.32418、R^2 = 0.9931,这些结果都提示所构建的多重线性回归模型很好。但毕竟存在回归系数的正负号不符合专业知识的“严重瑕疵”,这是一个需要彻底解决的“疑难问题”!

2.1.2 所需要的 SAS 程序

尝试采用“基于正交化方法”解决上述提及的“疑难问题”。所需要的 SAS 程序如下:

```
data a1;
```

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

```
input id age height weight bmi sbp;
cards;
(此处输入文献[5]表 1 中 50 行 6 列数据)
;
run;
/* 以上程序为了创建数据集 a1 */
data a2;
set a1;
x1 = age * age; x2 = age * height; x3 = age * weight;
x4 = age * bmi; x5 = height * height; x6 = height * weight;
x7 = height * bmi; x8 = weight * weight; x9 = weight * bmi;
x10 = bmi * bmi;
run;
/* 以上程序是在数据集 a1 基础上创建数据集 a2,它增添了 10 个派生变量 */
proc reg data = a2;
model sbp = age weight x3 x6 - x10/noint;
quit;
/* 以上程序是调用 REG 过程并基于数据集 a2 拟合文献[5]中那个"最佳"回归模型 */
proc orthoreg data = a2;
model sbp = age weight x3 x6 - x10/noint;
quit;
/* 以上程序是调用 ORTHOREG 过程并基于数据集 a2 拟合文献[5]中那个"最佳"回归模型 */
```

【SAS 程序说明】在以上的 SAS 程序中,都用"/* …… */"注释语句作了说明。

2.1.3 SAS 输出结果及其解释

以上第 1 个 SAS 过程步(REG 过程)程序的主要输出结果如下:

参数估计值

变量	自由度	参数估计值	标准误差	t 值	Pr > |t|
age	1	1.82182	0.49294	3.70	0.0006
weight	1	-88.00801	26.67636	-3.30	0.0020
x3	1	-0.00971	0.00342	-2.84	0.0069
x6	1	0.64569	0.19305	3.34	0.0017
x7	1	4.32456	1.30917	3.30	0.0020
x8	1	-0.05835	0.01884	-3.10	0.0035
x9	1	0.78530	0.25836	3.04	0.0041
x10	1	-2.62458	0.87715	-2.99	0.0046

以上第 2 个 SAS 过程步(ORTHOREG 过程)程序的主要输出结果如下:

变量	自由度	参数估计值	标准误差	t 值	Pr > |t|
age	1	1.82181715589926	0.4929403378	3.70	0.0006
weight	1	-88.0080067638051	26.676358039	-3.30	0.0020
x3	1	-0.0097085369946	0.0034186421	-2.84	0.0069
x6	1	0.64568801124242	0.1930515929	3.34	0.0017
x7	1	4.32455894833146	1.3091729962	3.30	0.0020
x8	1	-0.05835311867068	0.01883696	-3.10	0.0035
x9	1	0.78529603354244	0.2583605157	3.04	0.0041
x10	1	-2.6245849862356	0.8771470203	-2.99	0.0046

比较上述由"REG 过程"与"ORTHOREG 过程"对同一个具有严重共线性问题的资料进行多重回归分析的结果可知:后者比前者所估计的"参数值"更精细,但仍没有消除多重共线性对回归系数的严重影响(尤其是 weight 前的回归系数为负值且其绝对值还比较大,不符合临床专业知识)。也就是说:SAS/STAT 中的"ORTHOREG 过程"并不能解决"多重共线性问题"。

2.2 此法可否解决高阶多项式曲线拟合问题

2.2.1 问题与数据结构

假定有一个高阶多项式资料,见表 1。

表 1 某资料中首尾各 10 对数据($n=101$)

obs	x	y	obs	x	y
1	0.00	0.00000	92	0.91	-0.46975
2	0.01	0.60835	93	0.92	-0.64040
3	0.02	0.96313	94	0.93	-0.81205
4	0.03	1.12892	95	0.94	-0.96998
5	0.04	1.15844	96	0.95	-1.09421
6	0.05	1.09421	97	0.96	-1.15844
7	0.06	0.96998	98	0.97	-1.12892
8	0.07	0.81205	99	0.98	-0.96313
9	0.08	0.64040	100	0.99	-0.60835
10	0.09	0.46975	101	1.00	0.00000

注:表 1 中省略编号为 11 到 91 的数据

绘出表 1 资料中(x,y)各点的散布图,见图 1。

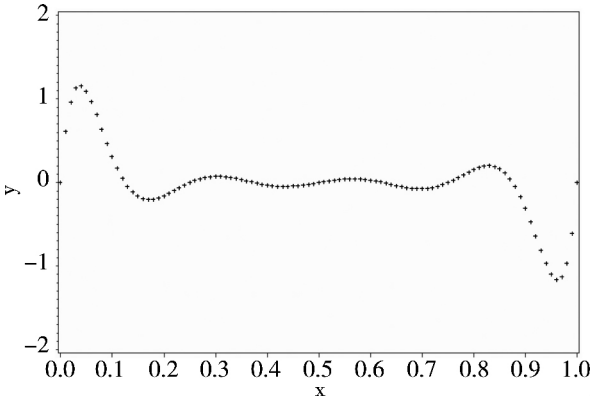

图 1 表 1 资料中(x,y)散布图

【问题】试拟合图 1 中 y 依赖 x 变化的回归模型。

2.2.2 所需要的 SAS 程序

```
data Polynomial;
  do i = 1 to 101;
    x = (i - 1)/(101 - 1);
    y = 10 * * (9/2);
    do j = 0 to 8;
      y = y * (x - j/8);
    end;
    output;
  end;
run;
    /* 以上程序是为了产生表 1 中的全部 101 对数据 */
proc gplot data = Polynomial;
    plot y * x;
run;
    /* 以上程序是为了绘制图 1 中的散布图 */
proc glm data = Polynomial;
  model y = x|x|x|x|x|x|x|x|x;
run;
    /* 以上程序是为了调用 GLM 过程拟合九次多项式曲线回归模型 */
ods graphics on;
proc orthoreg data = Polynomial;
  effect xMod = polynomial(x / degree = 9);
  model y = xMod;
  effectplot fit / obs;
  store OStore;
run;
ods graphics off;
    /* 以上程序是为了调用 ORTHOREG 过程拟合九次多项式曲线回归模型并绘图 */
```

2.2.3 SAS 输出结果及其解释

以下是“GLM 过程步”输出的计算结果:

参数	估计值		标准误差	t 值	Pr > |t|
截距	0.44896		0.125479	3.58	0.0006
x	24.61062		6.054852	4.06	0.0001
x * x	-443.74034		93.388508	-4.75	<0.0001
x * x * x	2626.90806	B	642.972718	4.09	<0.0001
x * x * x * x	-7371.23677	B	2327.022377	-3.17	0.0021
x * x * x * x * x	10697.73145	B	4741.598897	2.26	0.0264
x * x * x * x * x * x	-7749.24904	B	5468.101939	-1.42	0.1598
x * x * x * x * x * x * x	2214.08419	B	3328.355661	0.67	0.5076
x * x * x * x * x * x * x * x	-0.00610	B	830.439899	-0.00	1.0000
x * x * x * x * x * x * x * x * x	0.00000	B	-	-	-

以上结果表明:用 GLM 拟合多项式曲线回归模型并不太合适,其中,标志“B”的那些行上的参数估计值不准确。

以下是基于 GLM 过程拟合结果绘制出的曲线图,见图 2。

图 2 中的实线是基于 GLM 过程计算的结果,而圆圈是实际观测到的结果。不难看出,拟合效果很不理想。

以下是“ORTHOREG 过程步”输出的计算结果:

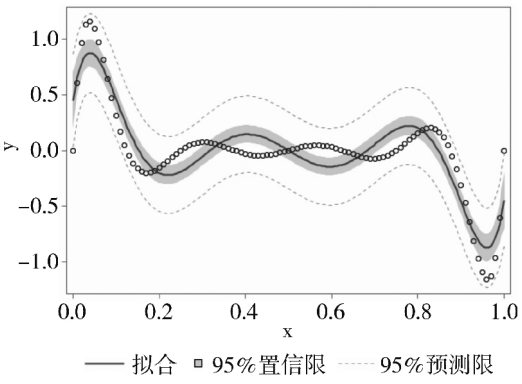

图 2 基于 GLM 过程拟合结果绘制出的曲线图

参数	自由度	参数估计值	标准误差	t 值	Pr > |t|
Intercept	1	−2.17224978239E−11	5.841067E−12	−3.72	0.0003
x	1	75.9977312439284	3.526134E−10	2.16E11	<0.0001
x^2	1	−1652.40781361802	6.8407273E−9	−242E9	<0.0001
x^3	1	14249.4539769373	5.9828349E−8	2.38E11	<0.0001
x^4	1	−64932.4615750127	2.8072627E−7	−231E9	<0.0001
x^5	1	173315.359360303	7.6781594E−7	2.26E11	<0.0001
x^6	1	−280158.036459353	1.2614251E−6	−222E9	<0.0001
x^7	1	269781.812887142	1.2252919E−6	2.2E11	<0.0001
x^8	1	−142302.494709869	6.4807927E−7	−22E10	<0.0001
x^9	1	31622.7766022261	1.4379253E−7	2.2E11	<0.0001

以上结果表明:拟合的九次多项式曲线回归模型中各参数均具有统计学意义。

以下是基于 ORTHOREG 过程拟合结果绘制出的曲线图,见图 3。

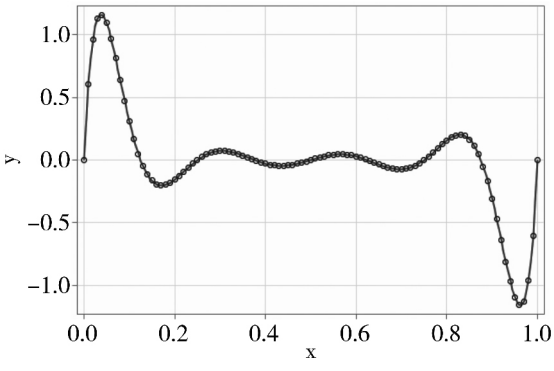

图 3 基于 ORTHOREG 过程拟合结果绘制出的曲线图

图 3 中的实线是基于 ORTHOREG 过程计算的结果,而圆圈是实际观测到的结果。不难看出,拟合效果非常好。

【说明】本例表 1 中的资料属于一种相当严重的“病态”资料,由图 3 可知,当 x 在(0.00,0.18)之间变化时,y 的变化形成了一个“尖峰”;当 x 在(0.18,0.82)之间变化时,y 的变化形成了近似平行于 x 轴的一条“略有起伏的波浪线”;而当 x 在(0.82,1.00)之间变化时,y 的变化形成了一个“低谷”。

参考文献

[1] Gentleman WM. Basic procedures for large, sparse, or weighted linear least squares problems, Technical Report CSRR − 2068 [Z]. Ontario: University of Waterloo, 1972.

[2] Gentleman WM. Least squares computations by givens transformations without square roots[J]. J Inst Math Appl, 1973, 12, 329−336.

[3] 高惠璇. 统计计算[M]. 北京: 北京大学出版社, 2017: 234−248.

[4] 徐士良. 数值方法与计算机实现[M]. 北京: 清华大学出版社, 98−160.

[5] 胡良平. 主成分分析应用——(I)主成分回归分析[J]. 四川精神卫生, 2018, 31(2): 128−132.

(收稿日期:2018−05−03)
(本文编辑:唐雪莉)