回归建模的基础与要领(Ⅳ)
——样品状态与相互间关系

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】 本文目的是介绍回归建模的基础与要领之四,即“样品状态与相互间关系”。首先,介绍“样品状态”,分为“单个体型样品”和“多个体型样品”;其次,介绍“样品间相互关系”,即“全部样品在空间中分布的相对位置”,需要借助“几何方法”来展现。然而,在高维空间中,“几何方法”非常难以实现,取而代之的是“代数方法”,即寻找合适的“权重系数”,以体现各样品(或试验点)在全部资料中的“重要性”。本文的结论是:无论是直线回归分析还是多重线性回归分析,采取“加权最小平方法”建模比采取“普通最小平方法”建模的效果好;而且,若能选取“合适的权重系数”并采取两次加权最小平方法建模,回归模型的拟合效果会更好。

【关键词】 样品;试验点;权重系数;加权最小平方法;拟合优度
中图分类号:R195.1 文献标识码:A doi:10.11886/j.issn.1007-3256.2018.06.004

The basis and essential of the regression modeling(Ⅳ)
——the samples´status and their relationships

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 fourth aspect of the basis and essential of the regression modeling: the samples´status and their relationships. Firstly, "the samples´status" was introduced, which could be divided into two parts, such as "the sample of one single individual type" and "the sample of the multiple individual type". Secondly, "the relationships among the samples": "the relative place of the whole samples which was distributed in the space", were recommended. The relationships among the samples needed to be emerged by means of "the geometric method". In the multidimensional space, however, "the geometric method" was very difficult to realize. So, "the geometric method" should be replaced by "the algebraic method" which was to find the suitable "weighted coefficient". "The weighted coefficient" could reflect the importance of each sample (or an experimental point) in the whole data. In general, the effect of the built models by using "the weighted least square method" was better than the one by using "the common least square method", no matter what in fitting a simple linear regression model or in building a multiple linear regression model. Furthermore, the best regression model would be gotten through the following two steps: one was to select a suitable weighted coefficient, the other was to adopt the weighted least square method to build the regression model for twice.

【Keywords】 Sample; Experimental point; Weighted coefficient; Weighted least square method; Goodness of fit

1 概 述

回归分析是研究因变量如何依赖自变量变化而变化的规律的重要统计分析方法之一,然而,回归分析的基本要素涉及两个方面,其一,变量状态及相互关系;其二,样品(测定变量取值的对象)状态及相互关系。因篇幅所限,本文仅讨论前述的“第二个要素”。

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

2 样品状态

2.1 单个体型样品

通常,适合于采用回归分析的数据结构中的每个“样品”对应着“一个个体”。若受试对象或调查对象是“人”,则每个“人”在统计学上被称为一个“样品”。例如,从 30 例某病患者的血液样品中测得“载脂蛋白 A1、载脂蛋白 B、载脂蛋白 E、载脂蛋白 C、低密度脂蛋白中胆固醇”的含量[1],其数据结构见表 1。

表 1 30 例患者载脂蛋白和低密度脂蛋白中胆固醇含量的测量结果

编 号	载脂蛋白 A1 (mg/dL)	载脂蛋白 B (mg/dL)	载脂蛋白 E (mg/dL)	载脂蛋白 C (mg/dL)	低密度脂蛋白 (mg/dL)
1	173	106	7.0	14.7	137
2	139	132	6.4	17.8	162
3	198	112	6.9	16.7	134
…	…	…	…	…	…
29	173	123	8.7	19.0	188
30	132	131	13.8	29.2	122

又例如，调查某地区某时间段内 685 例年龄≥70 岁老年人的“一般情况、与健康有关的各项指标的取值和生活质量”所得到的资料[2]。显然，若受试对象或调查对象是某种“动物”，则每只“动物”在统计学上也被称为一个“样品”。由此可知，以每个“样品”为观察单位时，就可以称其为“单个体型样品”。

2.2 多个体型样品

有时，观察单位不是一个个体，而是由具有相近条件的多个个体组成。例如，某试验研究药物剂量与有效率之间的关系，数据结构如表 2 所示[1]。

表 2 药物不同剂量与有效数

剂量(mg/kg)	有效数(n)	试验动物数(n)	剂量(mg/kg)	有效数(n)	试验动物数(n)
0.1	0	9	0.6	9	15
0.2	1	10	0.7	13	17
0.3	2	13	0.8	14	18
0.4	4	15	0.9	13	14
0.5	7	17	1.0	16	17

在表 2 中，每个剂量组的“全部动物”被视为一个“观察单位”，共有 10 个剂量组。显然，每个“观察单位”有多只动物。

又例如，某棉纺厂为减轻试验工作量，拟用较易测定的每毫克重纤维的根数 x 估计测定工作量较大的原棉单纤维强力 y。研究者收集到的试验资料见表 3。其中，m_i 为在第 i 个试验点 x_i 上进行的独立重复试验次数；y_i 实际上是第 i 个试验点上 m_i 个 $y_{ij}(j=1,2,\cdots,m_i)$ 的算术平均值(注：若未求平均值，就可求出方差)[3]。

表 3 每毫克重纤维的根数 x 与原棉单纤维强力 y 之间关系的测定结果

编 号	m_i	x_i	y_i	编 号	m_i	x_i	y_i
1	2	188	4.90	9	18	266	3.47
2	3	195	4.58	10	15	275	3.43
3	11	207	4.40	11	12	285	3.19
4	16	217	4.18	12	5	295	3.08
5	18	224	3.90	13	5	312	2.94
6	19	236	3.85	14	4	320	2.79
7	20	246	3.77	15	1	329	2.49
8	22	255	3.54	16	–	–	–

在表 3 中，各“编号”代表一批试验或被称为一个“观察单位”，若对其进行回归分析，各“编号”对应的数据被称为“样品”。则每个“编号”由重复试验次数不等的多个样品构成，被称为“多个体型样品”。

3 样品间相互关系

3.1 概述

与变量间相互关系相比，样品间相互关系比较难理解，因为样品间关系需要借助“几何图形”呈现出来才便于直观判断。通常，可通过在二维直角坐标系中的全部(x，y) 散点分布情况，用目测法得出全部“样品”或“试验点”间实际存在的相互关系(简称为“几何方法”) ; 然而，当自变量数目≥2 时，要在高维空间中直接呈现全部样品间相互关系非常困难。

解决前述困难的办法是: 在二维空间中，找到合适的统计处理方法(简称“代数方法”) ，从而建立起“几何方法”与“代数方法”之间的联系。由此，可将“代数方法”推广到高维空间中去研究“样品”间的相互关系。

根据数学理论和实践结果，上面提及的“代数方法”可归结为给“每个样品”一个“权重系数”，它的作用是反映每个样品在计算中的“分量”或“作用大小”。也就是说，“权重系数”大的“样品”要比“权重系数”小的“样品”发挥更大的作用。对于同一个回归分析资料，选取不同的依据来构造“权重系数”并据此来构建回归模型，其精确度是不同的。因此，可将全部可能的“依据”都用来构造“权重系数”，从而可构建出多种不同的回归模型。于是，可从中选出“最精准的回归模型”。

3.2 样品的同质性与异常点

3.2.1 样品的同质性

在对计量资料进行 t 检验或方差分析时，统计学教科书上都会明确交代: 资料必须满足“独立性”“正态性”和“方差齐性”三个前提条件; 而在对资料进行相关与回归分析，尤其是进行多重回归分析时，统计学教科书上则很少提及极其重要的“前提条件”，即所有样品对于全部变量应满足“同质性”。其含义是: 所有样品或个体在全部变量上的“取值规律”是基本相同的。例如: 研究某地区某时间段内正常成年人的体重是如何随身高变化而变化的依赖关系时，当所有被观测个体(或称为“样品”) 在(身高，体重) 两个变量上的取值对应的“数据点”沿一条直线(或曲线) 变化趋势随机地散布，没有偏离“绝大多数样品”所在“区域”较远或很远的“数据点”，就称该资料中的所有“个体或样品”具有较好的“同质性”。

3.2.2 异常点

在前面的(身高，体重) 例子中，若其中包含了少数特体型个体(例如，身高约为 2. 3 m，但体重约为 50 kg; 体重约为 250 kg，但身高仅 1. 6 m; 身高约为 1. 0 m，但体重为 80 kg) ，那么，这少数特体型人与绝大多数正常成年人就不是“同质的”。于是，那些“特体型个体”在统计学上就被称为“异常点(即异常的个体) ”。之所以说它们是“异常点”，是因为当采用“几何方法”呈现时，它们所处的“空间位置”会偏离其他“数据点”所在的“变化区域”。在二维直角坐标系中，绘制出资料的散布图，数据点的分布情况将一览无余，见图 1。

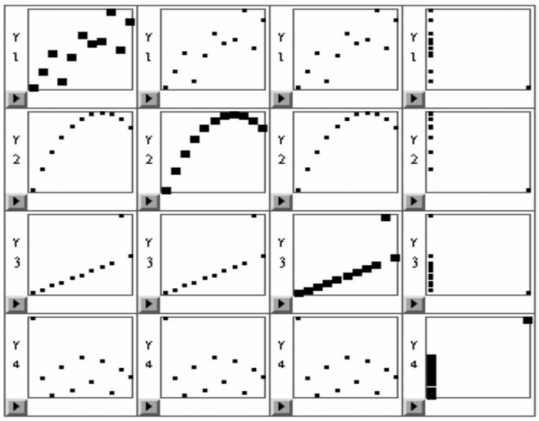

图 1 四组数据的散布图

在图 1 中，从左上角至右下角的对角线上有数据点较粗的 4 幅小图，其中，最后两幅小图中均各有一个“异常点”。

在进行直线回归分析时，若存在“异常点”，但分析者对其视而不见，就很容易得出错误的结果和结论; 在进行多重回归分析时，若存在“异常点”，分析者也同样容易“误入歧途”。在 SAS 的“REG 过程”中，可以通过“学生化残差”和“Cook’s D 距离统计量”来进行“异常点诊断”，淘汰掉资料中“严重的异常点”，将有助于提高回归模型的拟合质量。当一个资料中存在较大比例的“异常点”且又不适合将它们全部删除时，需要找到对“异常点”有一定“耐受性”的回归建模方法，常称为“稳健回归分析法”[4]。经过比较发现 “分位数回归分析法”[5] 比“参数法中的诸多稳健回归分析法”[4] 更加“稳健”。

3.3 可用于构造“权重系数”的常见“依据”[3,6]

3.3.1 某变量在各样品上取值的平方的倒数

选取某个变量(因变量或自变量) ，其在每个

“样品”上会有不同的取值。若在大多数样品上的取值比较接近，而在少数样品上的取值“非常大”，那么，就这个“变量”而言，可能意味着：取值“非常大”的那几个“样品”很可能是“异常点”。若取该“变量平方的倒数”为“权重系数”，则取值“非常大”的“样品”的“权重系数”就比较“小”，从而，它们在回归系数估计中所起的作用就相应地变小了。也就是说，这实际上是间接削弱了“可疑异常点”的影响，使回归系数的估计趋于“稳健”。

3.3.2 因变量在各样品上残差平方的倒数

也可以这样做：先不盲目地寻找任何“依据”来构建“权重系数”，而采取通常的方法构建回归模型，再利用此回归模型计算出各“样品”上因变量的“预测值”；进而可计算出各“样品”上的“残差”。于是，可以求出各“样品”上“残差平方的倒数”。估计回归系数时，以各“样品”上“残差平方的倒数”为“权重系数”。道理如前所述，此处从略。

3.3.3 某变量在每个样品上全部取值的方差的倒数

若在某个变量取每个特定值的条件下都进行了多次重复试验，就会获得因变量的多个观测值，于是，就可计算出多个因变量观测值的方差，进而可计算出“因变量方差的倒数”。若选取各样品上“因变量方差的倒数”为“权重系数”，则“方差大”的样品上的“权重系数”就很小，故它们在回归系数估计中所起的作用就相对变小了。

3.3.4 各样品上重复试验次数的倒数

当各样品上有次数不等的重复试验时，可以取“重复试验次数的倒数”为“权重系数”。因为重复试验次数较多的样品上“因变量”的不同观测值的数目可能就会多一些，也就间接反映了该样品上“因变量的方差可能比较大”。于是，取“重复试验次数的倒数”为“权重系数”，相当于取“因变量方差的倒数”为“权重系数”。道理同上，此处不再赘述。

3.3.5 基于“观测权重”与“抽样权重”构造“综合权重”[7]

对抽样调查资料进行回归建模时，选取合适的“权重系数”至关重要。这里可能涉及到多个有关的概念，如观测权重、抽样权重。

观测权重是基于综合评价中权重系数的思想，在回归分析中引入反映各个体或观测对总体的重要性的度量，表示在其他观测不变的情况下，该观测的变化对结果的影响程度。常用的有经验权重法、试验次数权重法、贡献权重法等。

抽样权重是在抽样研究中，为反映所抽取样本中各个观测在总体中的重要程度，或样本中各个观测代表总体中个体的数目。抽样权重的大小与抽样方法有关，分为基础抽样权重、调整抽样权重与总抽样权重。例如，将某省或州划分为 3 个地区（即“层”），各层总样本量、抽取的样本量和抽样权重的计算方法和结果见表 4。

表 4 美国两个州各地区（层）中农场的数目、抽样数目和抽样权重

层	州	地区	农场数目	抽样数目	抽样权重
1	爱荷华州	1	100	3	33.333
2		2	50	5	10.000
3		3	15	3	5.000
4	内布拉斯加州	1	30	6	5.000
5		2	40	2	20.000
合计			235	19	–

表 4 中，抽样权重 = 农场数目/抽样数目，例如，第 1 行上被抽取的 3 个农场中的每一个代表全部农场（100 个）中的 1/3，即 33.333 个。

综合权重是在对随机抽样所得数据进行统计分析时，不仅考虑抽样权重，还考虑观测权重，计算各个观测对结果的总的重要程度。其计算方法是：综合权重 = 观测权重 × 抽样权重。

在 SAS/STAT 的“SURVEREG 过程”[8] 中，若采用分层随机抽样，则选择“抽样权重”作为“权重系数”。

3.4 实例演示

3.4.1 无重复试验的回归分析问题

3.4.1.1 问题与数据结构

【例 1】某公司对 12 次投标情况进行研究。设

投标规模为 x(单位:百万美元),企业准备投标的费用为 y(单位:千美元)。具体数据见表 5。试建立 y 关于 x 的回归方程[4]。

表 5 某地某年投标规模 x 与企业准备投标费用 y 的数据

编 号	x(百万美元)	y(千美元)
1	2.13	15.5
2	1.21	11.1
3	11.00	62.6
4	6.00	35.4
5	5.60	24.9
6	6.91	28.1
7	2.97	15.0
8	3.35	23.2
9	10.39	42.0
10	1.10	10.0
11	4.36	20.0
12	8.00	47.5

注:表 4 数据摘自文献《应用线性回归模型》(约翰·内特,威廉·沃塞曼,迈克尔·H·库特纳,著,张勇,王国明,赵秀珍,译. 北京:中国统计出版社,1990:178)

若绘制出(x,y)的散布图,各散点随自变量 x 的增加,y 的离散度也变大,为节省篇幅,绘制散布图的 SAS 程序和散布图均省略。

下面,先不考虑“权重系数”拟合直线回归模型,然后再选取不同的“依据”构建“权重系数”,并据此构建直线回归模型。基于模型的拟合优度(R^2、误差等)确定最合适的“权重系数”。

3.4.1.2 所需要的 SAS 程序

```
data a1;
    input x  y @@;
/* 以下产生三个新变量,分别代表不同的权重系数 */
    wx=1/(x**2);
    wy=1/(y**2);
    wxy=1/(x* y);
cards;
 2.13   15.5
 1.21   11.1
11.00   62.6
 6.00   35.4
 5.60   24.9
 6.91   28.1
 2.97   15.0
 3.35   23.2
10.39   42.0
 1.10   10.0
 4.36   20.0
 8.00   47.5
;
run;
/* 以下程序采用普通最小平方法 1 创建含截距项的直线回归模型* /
proc reg data=a1;
    model y=x /r;
quit;
/* 以下程序采用普通最小平方法 2 创建不含截距项的直线回归模型* /
proc reg data=a1;
    model y=x / noint r;
quit;
/* 以下程序采用加权最小平方法 1 创建直线回归模型* /
proc reg data=a1;
    model y=x /r;
    weight wx;
quit;
/* 以下程序采用加权最小平方法 2 创建直线回归模型* /
proc reg data=a1;
    model y=x /r;
    weight wy;
quit;
/* 以下程序采用加权最小平方法 3 创建直线回归模型* /
proc reg data=a1;
    model y=x /r;
    weight wxy;
quit;
/* 以下程序采用普通最小平方法创建直线回归模型,提取各样品点上的残差* /
proc reg data=a1 noprint;
    model y=x / noint r;
    output out=aaa residual=resid;
quit;
/* 以下程序为求取各样品点上残差平方的倒数* /
```

```
data a2;
    set aaa;
    wr = 1/resid * *2;
run;
/* 以下程序采用加权最小平方法 4 创建含截距项的直线回归模型* /
proc reg data = a2;
    model y = x / r;
    weight wr;
quit;
/* 以下程序采用加权最小平方法 5 创建不含截距项的直线回归模型* /
proc reg data = a2;
    model y = x / noint r;
    weight wr;
quit;
```

【SAS 程序说明】

以上 SAS 程序很长，各段 SAS 程序之前都有“注释语句”，这些注释语句解释了其后面程序的作用，此处不再赘述。

3.4.1.3 主要计算结果汇总

“普通最小平方法”和“加权最小平方法”拟合直线回归模型的参数估计结果见表 6。

表 6 普通与加权最小平方法拟合直线回归模型的参数估计值等内容比较

方 法	变 量	参数估计值	标准误	P
普通最小平方法 1	截距	4.22895	3.25174	0.2226
	x	4.51528	0.52847	<0.0001
普通最小平方法 2	x	5.10188	0.28390	<0.0001
加权最小平方法 1	截距	5.65685	0.96524	0.0002
	x	4.19055	0.40366	<0.0001
加权最小平方法 2	截距	5.88021	1.38115	0.0017
	x	3.87436	0.40906	<0.0001
加权最小平方法 3	截距	5.80537	1.15838	0.0005
	x	4.01996	0.40745	<0.0001
加权最小平方法 4	截距	0.33667	1.48925	0.8257
	x	4.94175	0.49532	<0.0001
加权最小平方法 5	x	5.05359	0.05057	<0.0001

在表 6 中，“普通最小平方法 1”对应的结果中，截距项无统计学意义；“普通最小平方法 2”中就没有包含截距项。

在表 6 中，“加权最小平方法”有 5 种，其中，前 4 种对应的“权重系数”分别为“自变量 x 的平方的倒数”“因变量 y 的平方的倒数”“自变量 x 与因变量 y 的乘积的倒数”“因变量 y 的残差平方的倒数”，而“加权最小平方法 5”与“加权最小平方法 4”的“权重系数”相同，都是“因变量 y 的残差平方的倒数”，它们的区别在于“是否保留截距项”。

淘汰掉“普通最小平方法 1”和“加权最小平方法 4”的结果之后，还有 5 种方法对应的结果，分别为“普通最小平方法 2”和“加权最小平方法 1”“加权最小平方法 2”“加权最小平方法 3”“加权最小平方法 5”。那么，这 5 种方法对应的结果哪一个相对更好？

若从假设检验的 P 值来看，很难分辨出孰优孰劣。但可以依据参数“标准误”大小进行比较，标准误小者为好。由此可知，“加权最小平方法 5”给出的斜率的“标准误 0.05057”最小，故该法相对其他 4 种更好。

下面，再列出上述各种方法对应的“R^2”“调整 R^2”“均方根误差”和“预测残差平方和，简称 PRESS”，见表 7。

表 7 普通与加权最小平方法拟合直线回归模型的拟合优度内容比较

方 法	R^2	调整 R^2	均方根误差	PRESS
普通最小平方法 1	0.8795	0.8675	5.86972	586.20863
普通最小平方法 2	0.9671	0.9641	6.05137	563.96447

续表 1:

加权最小平方法 1	0.9151	0.9066	0.88967	10.32205
加权最小平方法 2	0.8997	0.8897	0.16417	0.34635
加权最小平方法 3	0.9068	0.8975	0.38397	1.90498
加权最小平方法 4	0.9087	0.8996	1.04916	109.89517
加权最小平方法 5	0.9989	0.9988	1.00289	13.00709

由统计学基础知识可知: 对于直线回归模型而言, "R^2"和"调整 R^2"的数值越大越好; 而"均方根误差"和"PRESS(预测残差平方和)"越小越好。由此可知,在表 7 中,最后一行的结果是最好的。

3.4.1.4 本例小结

结合表 6 和表 7 的结果以及比较得出的结论可知: 本例以"因变量 y 的残差平方的倒数"构建"权重系数",并采取"加权最小平方法"拟合直线回归模型且不含截距项(因为截距项无统计学意义,需删除),其拟合效果最佳。

```
1   173  106  7.0  14.7  137
3   198  112  6.9  16.7  134
5   139   94  8.6  13.6  138
7   131  154 11.2  21.5  171
9   158  137  7.4  18.2  197
11  162  110  6.0  15.9  145
13  162  137  7.2  20.7  185
15  129  138  6.3  10.1  197
17  185  118  6.0  17.5  156
19  175  111  4.1  27.2  144
21  153  133  8.5  16.9  215
23  160   86  5.3  10.8  118
25  147  110  8.5  18.4  137
27  131  102  6.6  13.4  130
29  173  123  8.7  19.0  188
;
run;
/* 不加权且保留截距项,三种常规筛选方法所得结果相同,仅留一种* /
proc reg data = a1;
    model y = x1 - x4 /selection = backward sls = 0.05 r;
    title1 此处创建的是模型 1;
quit;
/* 不加权且不保留截距项,三种常规筛选方法所得
```

3.4.2 多重线性回归分析问题

3.4.2.1 问题与数据结构

【例 2】沿用前面的"表 1 资料",设 y 代表"低密度脂蛋白",$x_1 \sim x_4$ 分别代表表 1 中第 2 列至第 5 列上的 4 种"载脂蛋白",试建立 y 依赖 4 个自变量的多重线性回归模型。

3.4.2.2 所需要的 SAS 程序

```
data a1;
    input id x1 - x4 y @@;
    wy = 1/y * *2;
cards;
 2  139  132  6.4  17.8  162
 4  118  138  7.1  15.7  188
 6  175  160 12.1  20.3  215
 8  158  141  9.7  29.6  148
10  132  151  7.5  17.2  113
12  144  113 10.1  42.8   81
14  169  129  8.5  16.7  157
16  166  148 11.5  33.4  156
18  155  121  6.1  20.4  154
20  136  110  9.4  26.0   90
22  110  149  9.5  24.7  184
24  112  123  8.0  16.6  127
26  204  122  6.1  21.0  126
28  170  127  8.4  24.7  135
30  132  131 13.8  29.2  122
结果相同,仅留一种* /
/* 数据集 aaa 中包含各样品点上的残差 resid1* /
proc reg data = a1;
    model y = x1 - x4 /noint selection = backward sls
 = 0.05 r;
    output out = aaa residual = resid1;
    title1 此处创建的是模型 2;
quit;
/* 以下程序基于 aaa 数据集,用残差平方的倒数作
```

```
为权重系数* /
/* 这实际上就是做了一次加权多重线性回归分析* /
data a2;
    set aaa;
    wr1 = 1/resid1**2;
run;
proc reg data = a2;
    model y = x1 - x4/noint selection = backward sle
=0.05 r;
    weight wr1;
    title1 此处创建的是模型 3;
quit;
/* 加权且保留截距项,三种常规筛选方法所得结果
相同,仅留一种* /
proc reg data = a1;
    model y = x1 - x4/selection = backward sls = 0.05 r;
    weight wy;
    title1 此处创建的是模型 4;
quit;
/* 加权且不保留截距项,三种常规筛选方法所得结
果相同,仅留一种* /
/* 数据集 bbb 中包含各样品点上的残差 resid2* /
proc reg data = a1;
    model y = x1 - x4/noint selection = backward sls
=0.05 r;
    weight wy;
    output out = bbb residual = resid2;
    title1 此处创建的是模型 5;
quit;
/* 以下程序基于 bbb 数据集,用残差平方的倒数作
为权重系数* /
/* 这实际上就是做了第二次加权多重线性回归分
析* /
data b2;
    set bbb;
    wr2 = 1/resid2**2;
run;
proc reg data = b2;
    model y = x1 - x4/noint selection = backward sle
=0.05 r;
    weight wr2;
    title1 此处创建的是模型 6;
quit;
```

3.4.2.3 主要计算结果汇总

“普通最小平方法”和“加权最小平方法”拟合多重线性回归模型的参数估计结果见表 8。

表 8 普通与加权最小平方法拟合多重线性回归模型的参数估计值等内容比较

方法	变量	参数估计值	标准误	*P*
模型 1	截距	41.84089	31.36117	0.1933
	x_2	1.25446	0.24993	<0.0001
	x_4	-2.34079	0.65023	0.0013
模型 2	x_2	1.55505	0.10969	<0.0001
	x_4	-2.17820	0.64754	0.0022
模型 3	x_2	1.57425	0.01014	<0.0001
	x_4	-2.28049	0.05230	<0.0001
模型 4	截距	57.78330	31.58767	0.0784
	x_2	1.03631	0.25240	0.0003
	x_4	-2.15881	0.50280	0.0002
模型 5	x_1	0.30541	0.12472	0.0211
	x_2	1.11909	0.16777	<0.0001
	x_4	-2.10127	0.47322	0.0001
模型 6	x_1	0.21654	0.03681	<0.0001
	x_2	1.35372	0.07720	<0.0001
	x_3	-2.09195	0.76822	0.0114
	x_4	-1.91615	0.08597	<0.0001

由表 8 可知: 模型 1 和模型 4 均不够理想,因为它们都包含了无统计学意义的截距项; 模型 2 与模型 3 具有可比性,但模型 3 中参数的标准误小于模型 2 中参数的标准误,稍好一些; 同理,模型 6 比模型 5 更好。

那么模型 3 与模型 6 哪一个更好? 为回答这个问题,需要列出与“拟合优度”有关统计量的计算结果,见表 9。

表 9 普通与加权最小平方法拟合多重线性回归模型的拟合优度内容比较

方 法	NI	Cp 值	R^2	调整 R^2	均方根误差	PRESS
模型 1	2	2.5168	0.5377	0.5035	23.7150	18462.00000
模型 2	2	3.3952	0.9776	0.9760	24.0301	18560.00000
模型 3	2	0.2839	0.9997	0.9996	0.9699	31.47106
模型 4	2	3.2168	0.5379	0.5037	0.1712	1.08675
模型 5	3	2.0197	0.9757	0.9730	0.1642	1.01808
模型 6	4	4.0000	0.9995	0.9994	0.9018	57.14586

在表 9 中,NI 代表模型中自变量的个数,Cp 值越接近模型中自变量的个数,表明模型对资料的拟合度越好。模型 3 与模型 6 相比,R^2 和调整 R^2 都比较接近; 模型 6 的均方根误差小于模型 3 的均方根误差; 特别是 Cp 值,说明模型 6 优于模型 3。

3.4.2.4 本例小结

模型 1 和模型 2 都是基于普通最小平方法建模,前者保留截距项,后者不保留截距项; 模型 3 仅采取残差平方的倒数为“权重系数”,进行了一次“加权最小平方法构建回归模型”; 模型 4 和模型 5 都是基于因变量 y 平方的倒数为“权重系数”,进行了第一次“加权最小平方法构建回归模型”; 而模型 6 在模型 5 的基础上,又基于残差平方的倒数为“权重系数”,进行了第二次“加权最小平方法构建回归模型”。最终的结论是: 基于两次加权回归分析得到的模型 6 优于其他模型。其回归模型为:

$$\hat{y} = 0.21654x_1 + 1.35372x_2 - 2.09195x_3 - 1.91615x_4$$

参考文献

[1] 胡良平. 科研设计与统计分析[M]. 北京: 军事科学出版社, 2012: 379-398, 513-551.

[2] 胡良平. Windows SAS 6.12 & 8.0 实用统计分析教程[M]. 北京: 军事医学科学出版社, 2001: 400-412.

[3] 茆诗松. 统计手册[M]. 北京: 科学出版社, 2003: 483-484.

[4] 胡良平. 稳健回归分析[J]. 四川精神卫生, 2018, 31(3): 201-204.

[5] 胡良平. 分位数模型回归分析[J]. 四川精神卫生, 2018, 31(4): 296-301.

[6] 胡良平, 胡纯严, 鲍晓蕾. 应用数理统计[M]. 北京: 电子工业出版社, 2015: 142-184.

[7] 崔壮, 胡良平. 复杂调查资料的特点与统计分析方法概述[J]. 四川精神卫生, 2017, 30(5): 410-414.

[8] SAS Institute Inc. STAT SAS 9.3 User s Guide[M]. Cary, NC: SAS Institute Inc, 2011: 7633-7704.

(收稿日期: 2018-12-12)

(本文编辑: 唐雪莉)