主成分分析应用(Ⅰ)——主成分回归分析

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】 本文目的是介绍主成分回归分析的概念、作用以及用软件实现计算的方法。先对自变量进行主成分分析,然后将主成分变量视为新的自变量,再进行多重线性回归分析。通过不引入和引入派生变量以及采取不同的策略筛选自变量,可以获得多个合格的多重线性回归模型。在回归模型自由度接近相等时,基于残差方差最小、复相关系数最大为评价指标,从众多回归模型中优中选优。得出的经验为:应慎用主成分回归分析。

【关键词】 多重共线性;派生变量;主成分回归分析;残差方差;复相关系数
中图分类号:R195.1 **文献标识码**:A doi:10.11886/j.issn.1007-3256.2018.02.008

Application of the principal components analysis(Ⅰ) ——the principal components regression analysis

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 is to introduce the concepts, functions and the calculation methods based on statistical software of the principal components regression analysis. The calculation process can be divided into three steps as follows: ①implementing the principal components analysis to the independent variables; ②taking the principal components variables as the new dependant variables; ③constructing the multiple linear regression model based on the new independant variables. Several qualified multiple linear regression models can be acquired through the following two measures: ①without and with using the derived variables; ②adopting the different strategies of screening the independent variables. When the regression models with the same or almost the same degree of freedoms, users can select the best regression model from many qualified regression models based on the appraising indexes, such as minimum residual variance and maximum multiple correlation coefficient. The experience drawn from this paper is that the principal components regression analysis should be used very carefully.

【**Keywords**】 Multiple collinearity; Derived variable; Principal components regression analysis; Residual variance; Multiple correlation coefficient

1 概 述

1.1 基本概念

本期《基于 SAS 与 R 软件的主成分分析》一文介绍了“单组设计多元定量资料”这种特殊的数据结构,并指出:分析这种数据结构的统计分析方法占全部多元统计分析方法中的绝大部分,主成分分析法是其中一个最基本的方法,可以被灵活地运用于多重回归分析之中。

1.2 何为主成分回归分析

主成分回归分析(the Principal Components Regression Analysis)是将作为自变量的多个定量因素转换为全部互相独立的综合变量(即主成分变量),构建定量因变量依赖所求得的“全部主成分变量”变化而变化的回归模型,这样完成的多重线性回归分析被称为“主成分回归分析”[1]。

1.3 为何要使用主成分回归分析

由于在经典统计学中,要求自变量互相独立,此时,才可以构建多重线性回归模型。当自变量之间存在多重共线性时,经典统计学理论认为:所建立的多重线性回归模型的质量就不高,甚至可能是不能解决实际问题或违反专业知识的回归模型(指某些回归系数的正负号不符合基本常识和专业知识要求)。消除自变量之间的多重共线性关系,有多种具体方法,本文拟采用“主成分回归分析法”。

项目基金:国家高技术研究发展计划课题资助(2015AA020102)

1.4 问题与数据结构

【例 1】为推算成年人的收缩压(SBP,mmHg),研究者测量并收集了 50 名成年人的年龄(age)、身高(height,inches)、体质量(weight,pounds)和体质量指数(BMI)。结果见表 1[1]。试以 SBP 为因变量、其他四个变量为自变量建立多重线性回归模型。

表 1 50 名成年人的测量数据

id	age	height	weight	BMI	SBP
1	28	68	160	24.33	111
2	26	68	165	25.09	101
3	31	68	175	26.61	120
4	18	76	265	32.26	158
5	50	67	145	22.71	125
6	42	69	247	36.48	166
7	20	66	156	25.18	114
8	29	76	180	21.91	143
9	35	63	166	29.41	111
10	47	66	169	27.28	133
11	20	69	120	17.72	95
12	33	68	133	20.22	113
13	24	71	185	25.80	128
14	28	72	150	20.34	110
15	32	61	126	23.81	117
16	21	68	190	28.89	112
17	28	71	150	20.92	110
18	60	61	130	24.56	117
19	55	66	215	34.70	142
20	74	65	130	21.63	105
21	38	68	126	19.16	94
22	26	66	160	25.82	131
23	52	74	328	42.11	128
24	25	69	125	18.46	93
25	24	67	133	20.83	103
26	26	59	105	21.21	114
27	51	64	119	20.43	130
28	29	62	98	17.92	105
29	26	64	150	25.75	117
30	60	64	175	30.04	124
31	22	70	190	27.26	122
32	19	65	125	20.80	112
33	39	73	210	27.71	135
34	77	62	138	25.24	150
35	39	73	230	30.34	125
36	40	69	170	25.10	126
37	44	62	115	21.03	99
38	27	61	140	26.45	114
39	29	73	220	29.03	139
40	78	63	110	19.49	150
41	62	65	208	34.61	112
42	22	71	125	17.43	127
43	37	64	176	30.21	125
44	38	72	195	26.45	136
45	22	65	140	23.30	108
46	79	61	125	23.62	156
47	24	62	146	26.70	108
48	32	67	141	22.08	105
49	42	70	192	27.55	121
50	42	68	185	28.13	126

1.5 对数据结构的分析

在表 1 中,单从数据角度看,基本满足进行多重线性回归分析的要求;但严格地说,题目中并没有交代清楚:这 50 例受试者是从一个什么样的总体中以什么方式抽取的,更具体的疑问是:他们在收缩压(SBP)这个指标上是否具有同质性? 若有些人属于“严重高血压患者”、有些人属于“中度高血压患者”、有些人属于“轻度高血压患者”、有些人属于“临界高血压状态人群”,还有些人属于“正常血压人群”,而且,这 50 人组成的样本不是从各群体中按相同比例随机抽取的,那么,他们就不具有同质性,此资料也就不值得进行任何统计分析。

在本文中,假定资料具有同质性,且确定样本含量也是有一定依据的。再结合医学专业知识,人的年龄、身高、体质量和体质量指数确实对收缩压有一定的影响(严格地说,应找全所有可能影响收缩压的因素,应确定一个同质的研究总体,按不少于 80% 的效能从此总体中随机或分层随机抽取足够大的样本含量),笔者仅从统计建模角度展开下面的论述。

2 探索性分析

2.1 构建基本的多重线性回归模型

基于本刊2018 年第 1 期《基于经典统计思想实现多重线性回归分析》一文介绍的方法,利用下面的 SAS 程序构建基本的多重线性回归模型:

```
data a1;
input id age height weight bmi sbp;
cards;
(此处输入表 1 中 50 行 6 列数据)
;
run;
proc reg data = a1;
  model sbp = age height weight bmi;
run;
```

因篇幅所限,输出结果从略。从输出的结果中可看到:截距项无统计学意义,需要将其删除后重新拟合多重线性回归模型。在前面的过程步程序的“model 语句”的“;”之前增加一个选择项“/NOINT”,得到的结果表明:weight 和 BMI 两个自变量都无统计学意义,若将它们都从回归模型中删除,则可被利用的自变量就很少了。于是,考虑这 4 个自变量之间是否存在某种线性关系,从而导致不够理想的结果出现。

2.2 对全部自变量进行共线性诊断

采用下面的 SAS 过程步程序对全部自变量进行共线性诊断:

```
proc reg data = a1;
  model sbp = age height weight bmi/noint collin;
run;
```

【SAS 主要输出结果】

共线性诊断

个数	特征值	条件指数	偏差比例			
			age	height	weight	bmi
1	3.82791	1.00000	0.00822	0.00241	0.00086179	0.00058334
2	0.13968	5.23504	0.72118	0.01099	0.01474	0.00314
3	0.02695	11.91819	0.04000	0.93334	0.10601	0.02498
4	0.00547	26.45745	0.23060	0.05326	0.87838	0.97130

由上面的最后一行后 4 列的计算结果可知:weight 和 BMI 的“偏差比例”都大于 0.5,说明它们之间确实存在较严重的共线性关系。

3 主成分回归分析

3.1 对全部自变量进行主成分分析[2]

所需要的 SAS 过程步程序如下:

```
proc princomp data = a1 out = b1 prefix = z;
  var age height weight bmi;
run;
```

此程序输出结果为:产生 4 个主成分变量 z1 - z4 及其取值,这是中间结果(暂不呈现了)存储在名为 b1 的数据集中。

3.2 基于 z1 - z4 作主成分回归分析

所需要的 SAS 过程步程序如下:

```
proc reg data = b1;
  model sbp = z1 - z4;
run;
```

输出结果(此处从略)中包含无统计学意义的变量,需要采取变量筛选方法淘汰掉无统计学意义的主成分变量。所需要的 SAS 过程步程序如下:

```
proc reg data = b1;
  model sbp = z1 - z4/selection = backward sls = 0.05;
run;
```

输出结果(此处从略)表明:此模型的复相关系数 R 平方 = 0.3262,残差方差 = 200.41879。

3.3 仅对有共线性的自变量进行主成分分析

所需要的 SAS 过程步程序如下:

```
proc princomp data = a1 out = b2 prefix = z;
  var weight bmi;
run;
```

此程序输出结果为:产生 2 个主成分变量 z1 - z2

及其取值,这是中间结果(暂未呈现)存储在名为 b2 的数据集中。

3.4 基于独立自变量 age、height 和主成分变量 z1 – z2 作主成分回归分析

所需要的 SAS 过程步程序如下:

```
proc reg data = b2;
  model sbp = age height z1 – z2/selection = backward sls = 0.05;
run;
```

输出结果(此处从略)表明:此模型的复相关系数 R 平方 = 0.3470,残差方差 = 194.24348。

3.5 改变筛选变量的策略

前面筛选自变量时假定需要保留“截距项”,下面假定不需要保留“截距项”,所需要的 SAS 过程步程序如下:

```
proc reg data = b2;
  model sbp = age height z1 – z2/noint selection = backward sls = 0.05;
run;
```

【SAS 主要输出结果】

方差分析

源	自由度	平方和	均方	F 值	Pr > F
模型	3	740877	246959	1285.20	< 0.0001
误差	47	9031.34023	192.15618		
未校正合计	50	749908			

变量	参数估计值	标准误差	II 型 SS	F 值	Pr > F
age	0.52234	0.11439	4006.53478	20.85	< 0.0001
height	1.51626	0.06950	91470	476.02	< 0.0001
z1	3.20908	1.44250	951.01067	4.95	0.0309

此模型的复相关系数 R 平方 = 0.9880,残差方差 = 192.15618。

3.6 基于两种途径得到的主成分回归模型的比较

前面获得两个二重回归模型拟合效果接近,第 2 个比第 1 个稍好,因 R 平方更大、残差方差更小,但它们之间的差别很微小;第 3 个模型为三重线性回归模型,残差方差 192.15618 比第 2 个模型的 194.24348 略小,然而,R 平方 = 0.9880 比第 2 个模型的 0.3470 更大。

4 慎用主成分回归分析

4.1 通过改变筛选自变量的策略来提升回归模型的拟合效果

一般来说,不要急于采取主成分回归分析方法,而应首先考虑改变筛选自变量策略来提升模型拟合效果。就本例而言,直接采取前进法、后退法和逐步法筛选全部自变量,并分别假定模型中包含“截距项”与不包含“截距项”。

经尝试,包含截距项时,三种筛选方法得出的结果一致,结果如下:模型对资料的拟合效果不太好,R 平方 = 0.3723,残差方差 = 186.73058。

经尝试,不包含截距项时,三种筛选方法得出的结果一致,结果如下:模型对资料的拟合效果较好,R 平方 = 0.9883,残差方差 = 186.89456。这个结果比上面基于主成分回归分析得到的最好结果更好。

4.2 通过引入派生变量和改变筛选自变量的策略来提升回归模型的拟合效果

可以引入 4 个自变量的平方项和交叉乘积项(共 10 项),再加上原先的 4 个自变量,总共 14 个自变量参与自变量的筛选。筛选时,仍采取前面提及的策略,即基于包含“截距项”与不包含“截距项”且分别采用前进法、后退法和逐步法筛选,得到最好的结果如下:

方差分析

源	自由度	平方和	均方	F 值	Pr > F
模型	8	744770	93096	761.06	<0.0001
误差	42	5137.61575	122.32418		
未校正合计	50	749908			

变量	参数估计值	标准误差	II 型 SS	F 值	Pr > F
age	1.82182	0.49294	1670.83355	13.66	0.0006
weight	-88.00801	26.67636	1331.38489	10.88	0.0020
x3	-0.00971	0.00342	986.53529	8.06	0.0069
x6	0.64569	0.19305	1368.39348	11.19	0.0017
x7	4.32456	1.30917	1334.75604	10.91	0.0020
x8	-0.05835	0.01884	1173.86687	9.60	0.0035
x9	0.78530	0.25836	1130.12596	9.24	0.0041
x10	-2.62458	0.87715	1095.18837	8.95	0.0046

以上结果表明:模型对资料的拟合效果很好,R 平方 =0.9931,残差方差 =122.32418。

前面的“x3、x6 ~ x10”分别代表:x3 = age * weight、x6 = height * weight、x7 = height * bmi、x8 = weight * weight、x9 = weight * bmi、x10 = bmi * bmi。

这个模型对资料的拟合效果要好于前面未引入派生变量的最好模型的拟合效果。

值得注意的是: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。但应当注意:模型中还包含了“x6 = height * weight”和“x9 = weight * bmi”这两项,它们的系数都为正,且乘积的结果是很大的数值。所以,在整体上,此模型是不违反临床专业知识的。

4.3 小结

显然,上面的最后一个多重线性回归模型中的很多“项”之间存在严重的多重共线性关系,若对它们采取“主成分分析”提取主成分变量,再进行主成分回归分析,最终结果如下:

(1)基于全部 8 个变量(age,weight,x3,x6 - x10)构建主成分回归模型且假定包含截距项,R 平方 =0.5931,残差方差 =129.30182。

(2)基于全部 8 个变量(age,weight,x3,x6 - x10)构建主成分回归模型且假定不包含截距项,无法构建多重线性回归模型。

(3)基于 6 个有共线性关系的变量(weight,x6 - x10)产生的 6 个主成分变量再加上 2 个独立变量(age 和 x3)构建主成分回归模型且假定包含截距项,R 平方 =0.6095,残差方差 =1560.03944。

(4)基于 6 个有共线性关系的变量(weight,x6 - x10)产生的 6 个主成分变量再加上 2 个独立变量(age 和 x3)构建主成分回归模型且假定不包含截距项,R 平方 =0.8892,残差方差 =1731.12908。

以上的结果都不如仅引入派生变量但不采取主成分变量构建的多重线性回归模型的拟合效果好。故笔者建议:一般应慎用主成分回归分析。

类似文献[3]的资料都可能用到多重线性回归分析,当然,也就有可能会用到主成分回归分析。经验表明:应慎用主成分回归分析,而在引入派生变量的前提下,采取不同的筛选自变量的策略,有可能获得比较理想的多重线性回归模型。

参考文献

[1] 胡良平. 面向问题的统计学——(2)多因素设计与线性模型分析[M]. 北京:人民卫生出版社,2012:215 - 228,264 - 272.

[2] 胡良平. 面向问题的统计学——(3)试验设计与多元统计分析[M]. 北京:人民卫生出版社,2012:19 - 39.

[3] 赵巍峰,彭敏,谢博,等. 健康教育对精神分裂症患者病耻感影响的持续性[J]. 四川精神卫生,2017,30(6):519 - 523.

(收稿日期:2018 - 04 - 02)

(本文编辑:陈 霞)