DOI:10.13546/j.cnki.tjyjc.2011.05.003

主成分回归的 SPSS 实现

郭呈全,陈希镇
(温州大学 数学与信息科学学院,浙江 温州 325000)

摘 要:文章结合主成分分析和线性回归分析的原理,利用 SPSS15.0 的 Descriptives、Data Reduction、Linear Regression、Compute Variable 模块的功能,把主成分回归的每一步计算过程用 SPSS 展现出来,并且对结果给出 SAS 验证。这不仅使学生更好地掌握主成分回归的相关知识,而且可以培养学生灵活使用 SPSS 软件。

关键词:共线性;主成分回归;特征值;特征向量;SPSS
中图分类号:O21 **文献标识码**:A **文章编号**:1002-6487(2011)05-0157-03

0 引言

在进行多元线性回归分析时,经常会遇到自变量之间存在近似线性关系的现象,这种现象被称为共线性[1]。当共线性严重时,用最小二乘法建立的回归模型将会增加参数的方差,使得回归方程变的很不稳定,有些自变量对因变量影响的显著性被隐藏起来,某些回归系数的符号与实际意义不相符[2],回归方程和回归系数通不过显著性检验。处理共线性的主要方法有筛选变量法、岭回归法、主成分回归法、偏最小二乘法等。在文献[2]中高惠旋使用 SAS 软件对处理共线性的主成分回归方法进行了实现,但是很多人只熟悉 SPSS 操作,SPSS 没有直接提供主成分回归的模块,文献[3]虽然也提出使用 SPSS 进行主成分回归,但是他首先使用了筛选变量法,没能真正体现主成分回归方法提取主成分的优势,而且其操作过程非常繁琐,没有灵活使用 SPSS 软件模块功能。本文结合主成分分析和线性回归分析的原理,巧用 SPSS15.0 的 Descriptives、Data Reduction、Linear Regression、Compute Variable 模块的功能,把主成分回归的每一步计算过程用 SPSS 展现出来,并且对结果给出了 SAS 验证。不但得出了正确结果,而且把每一步计算过程完整地呈现出来,这样既有利学生掌握有关方面的知识,还能加深学生对统计软件的灵活使用和掌握。

1 基本原理和计算步骤

1933 年,Hotelling 提出主成分分析方法,主成份分析的核心思想就是通过降维,把多个指标化为少数几个综合指标,而尽量不改变指标体系对因变量的解释程度。W.F. Massy 于 1965 年根据主成份分析的思想提出了主成份回归。如今主成份回归方法已经被广泛采用,成为回归分析中解决多重共线性比较有效的方法。

设 $Y=(y_1,y_2,\cdots,y_n)$,假设 X 设计矩阵已经中心化,记 $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p$ 为 X′X 的特征根,$\Phi=(\varphi_1,\varphi_2,\cdots,\varphi_p)$为对应的标准正交化特征向量。主成分回归的计算步骤是:

(1)为了使结果不受量纲的影响,先把原始数据进行标准化;

(2)求 X′X 的特征值和对应的标准正交化特征向量;

(3)做回归自变量选择。最大的特征值对应的特征向量即为第一主成分的系数,第二大的特征值对应的特征向量即为第二主成分的系数,以此类推。取几个主成分取决于主成分对因变量的解释程度。如果前 i 个特征值之和与所有特征值之和的比达到一定的程度比如 85%时,就可以认为这些主分就能代替所有的自变量体系。剔除对应的特征值比较小的那些主成分。

(4)做正交变换 Z=XΦ,获得新的自变量;

(5)将剩余的成分对因变量进行普通最小二乘回归,再返回到原来的参数,便得到因变量对原始变量的主成分回归。

总结这些步骤可以看出:主成份回归解决多重共线性问题是通过求特征值和特征向量达到降维来实现的。因为在降维前指标之间的多重共线性可能是由于某个指标或者少数指标所包含的信息与其他指标所包含的信息之间的相关性引起的,通过降维的处理我们提取出了主成份,就像是把指标体系所包含的信息分了类,某一大类由一个主成份来表现,这样就消除了产生多重共线性问题的根源:信息的交迭[4]。

2 SPSS 对计算过程的实现

利用文献[1]中的外贸数据:因变量 Y 为进口总额,自变

基金项目:国家统计局资助项目(LX08081);浙江省精品课程“统计学概论”和温州大学研究生精品课程“多元统计学分析”资助

表 1

序号	X_1	X_2	X_3	Y
1	149.3	4.2	108.1	15.9
2	161.2	4.1	114.8	16.4
3	171.5	3.1	123.2	19.0
4	175.5	3.1	126.9	19.1
5	180.8	1.1	132.1	18.8
6	190.7	2.2	137.7	20.4
7	202.1	2.1	146.0	22.7
8	212.4	5.6	154.1	26.5
9	226.1	5.0	162.3	28.1
10	231.9	5.1	164.3	27.6
11	239.0	0.7	167.6	26.3

量 X_1 为国内总产值，X_2 为存储量，X_3 为总消费。为了建立 Y 对自变量 X_1，X_2 和 X_3 之间的依赖关系，收集了 11 组数据见表 1。

表 2　描述性统计量表

	样本数	均值	标准差
x1	11	194.5909	29.99952
x2	11	3.3000	1.64924
x3	11	139.7364	20.63440
y	11	21.8909	4.54367

2.1　数据标准化

执行：Analyze→Descriptives Statistics→Descriptives，将变量 y，x_1，x_2，x_3 选入 Variables 的对话框中，选定 Save standardized values as variables，即将标准化后的数据作为变量保存。见表 2。

描述性统计量表中显示各变量的样本数（N）、均数（mean）和标准差（Std.Deviation），以便于对中心化后的自变量进行完主成分回归后还原为原始变量。

2.2　共线性诊断

共线性就是对自变量观测数据构成的矩阵 X′X 进行分析，使用各种指标反映自变量间的相关性。进行共线性诊断的方法有很多种，目前较为常用的诊断方法有：条件数（condition index）、容忍度 Tolerance（或方差膨胀因子（VIF））、特征根（Eigen value）分解法。

（1）条件数：是指 X′X 的最大特征根与最小特征根之比 $k=\lambda_l/\lambda_p$，它刻画了特征值差异的大小。一般情况下 k<100，则认为复共线性很小；100≤k≤1000 认为存在中等程度的复共线性；若 k>1000 则认为存在严重共线性。

（2）容忍度：以每个自变量作为因变量对其他自变量进行回归分析时得到残差比例，用 1 减去决定系数来表示（$1-R^2$），越小说明共线性越重，T<0.1 时共线性非常严重（陈希孺）。由此方差膨胀因子（VIF）：定义 VIF=1/T，VIF 越大，说明共线性越严重。

（3）特征根分解法：对自变量进行主成分分析，若相当多维度的特征根为 0，则共线性严重。

本例共线性诊断操作步骤如下：执行：Analyze→Regression→Linear，在 Dependent 中选择导入，在 Independent 中导入 Zx_1，Zx_2，Zx_3，在 statistics 中选中 Colinearity statistics，其它选项默认，得表 3。

表 3　回归系数和共线性统计量

模型	变量	标准化系数	t	P	共线性统计	
					Tolerance	VIF
1	常数		.000	1.000		
	Zx1	−.339	−.731	.488	.005	185.997
	Zx2	.213	6.203	.000	.981	1.019
	Zx3	1.303	2.809	.026	.005	186.110

表 4　共线性诊断指标

维数	特征植	条件指数	方差百分比			
			常数	Zx1	Zx2	Zx3
1	1.999	1.000	.00	.00	.00	.00
2	1.000	1.414	1.00	.00	.00	.00
3	.998	1.415	.00	.00	.98	.00
4	.003	27.257	.00	1.00	.02	1.00

表 3 给出线性回归方程中回归系数的估计值和共线性统计量，表中 Z_{X1} 和 Z_{X3} 容忍度都为 0.005<0.1，并且其方差膨胀因子 VIF 都很大，说明它们之间存在严重的共线性。

从表 4 可以看出，条件数 1.999/0.003≈666.33，故共线性程度较严重。从方差百分比上看，Z_{X1} 和 Z_{X3} 变量间也存在明显相关性。

2.3　主成分分析

执行：Analyze→Data Reduction→Factor，选定标准化后的变量 Zx_1，Zx_2，Zx_3 进入 Variables 中，Extraction 中的选项，method 选用 principal components，Analyze 选用 covariance matrix，在提取主成分的 Extract 中选用 Number of factor 并在后面的框中填入 3，提取三个主成分。在 Scores 中选择 Save as variables；在 method 中选择 reg；不进行旋转，结果输出如表 5。

表 5 显示三个特征值分别为 $\lambda_1=1.999$，$\lambda_2=0.998$，$\lambda_3=0.003$，前两个特征值的累计贡献率达到 99.91%，因此剔除第三个主成分，相应的因子载荷矩阵如表 6。

2.4　求特征向量和主成分

前两个特征值 $\lambda_1=1.999$，$\lambda_2=0.998$，对应的标准正交化特征向量分别为：

$$\varphi_1'=\left(\frac{0.999}{\sqrt{\lambda_1}},\frac{0.062}{\sqrt{\lambda_1}},\frac{0.999}{\sqrt{\lambda_1}}\right),\varphi_2'=\left(\frac{-0.036}{\sqrt{\lambda_2}},\frac{0.998}{\sqrt{\lambda_2}},\frac{-0.026}{\sqrt{\lambda_2}}\right)$$

下面使用 Compute Variable 模块的功能，计算第一和第二主成分。

执行：Analyze→Transform→Compute Variable，在 Target Variable 中输入 Z_1，在 Numeric Expression 中计算公式为：Z_1=

表 5　主成分提取汇总表

主成分	初始			提取项		
	特征值	方差百分比	累计方差百分比	特征植	方差百分比	累计方差百分比
1	1.999	66.638	66.638	1.999	66.638	66.638
2	.998	33.372	99.910	.998	33.272	99.910
3	.003	.090	100.000			

表 6　得分矩阵

	1	2
Zx1	.999	−.036
Zx2	.062	.998
Zx3	.999	−.026

表 7　主成分表

Z1	−2.13	−1.62	−1.12	−0.89	−0.64	−0.19	0.36	0.97	1.56	1.77	1.93
Z2	0.64	0.56	−0.07	−0.08	−1.13	−0.06	−0.74	1.35	0.96	1.02	−1.66

表 8　回归系数

模型		非标准化系数		标准化系数	t	P	Collinearity Statistics	
		B	Std.Error	Beta			Tolerance	VIF
1	常数	7.07E−017	.036		.000	1.000		
	z1	.690	.027	.976	25.486	.000	1.000	1.000
	z2	.191	.038	.191	4.993	.001	1.000	1.000

FAC1_1*sprt(1.999)，单击 OK 产生新变量 Z_1，同上得：

Z_2=FAC2_1*sqrt(0.998)，于是得：

$$Z_1=\frac{0.999}{\sqrt{\lambda_1}}Z_{x_1}+\frac{0.062}{\sqrt{\lambda_1}}Z_{x_2}+\frac{0.999}{\sqrt{\lambda_1}}Z_{x_3},$$

$$Z_2=\frac{-0.036}{\sqrt{\lambda_2}}Z_{x_1}+\frac{0.998}{\sqrt{\lambda_2}}Z_{x_2}+\frac{0.026}{\sqrt{\lambda_2}}Z_{x_3}$$

输出变量结果如表 7。

2.5 线性回归

对第一主成分 Z_1 和第二主成分 Z_2 做关于中心化因变量 Zy 的最小二乘回归分析。

执行：Analyze→Regression→Linear，在 Dependent 中选择 Zy 导入，在 Independent 中导入 Z_1 和 Z_2，做最小二乘回归。见表 8。

回归系数估计值为：$\hat{\beta}_1$=0.690，$\hat{\beta}_2$=0.191，常数项近似为零。把上面关系式代入：

Z_y=0.69Z_1+0.191Z_2+7.07E-017，求得：

$$Z_y=0.69\times[\frac{0.999}{\sqrt{\lambda_1}}Z_{x_1}+\frac{0.062}{\sqrt{\lambda_1}}Z_{x_2}+\frac{0.999}{\sqrt{\lambda_1}}Z_{x_3}]+0.191\times[\frac{-0.036}{\sqrt{\lambda_2}}Z_{x_1}+\frac{0.998}{\sqrt{\lambda_2}}Z_{x_2}+\frac{0.026}{\sqrt{\lambda_2}}Z_{x_3}]$$

因此，Z_y=0.4806Z_{x_1}+0.2298Z_{x_2}+0.4825Z_{x_3}。

根据 $y=Z_y\sqrt{D_y}+\bar{y}$ 和 $x=Z_x\sqrt{D_x}+\bar{x}$，还原到原始变量的关系为：

y=-9.1057+0.0727x_1+0.6091x_2+0.1062x_3。

3 SAS 验证

使用 SAS 的 REG 过程，对上述数据做主成分分析，SAS 程序如下：

```
Proc reg data=a outset=out1;
Model y=x1-x3/pcomit=1,2 outvif;
Proc print data=out1;
Run
```

后输出如下结果：

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	x1	x2	x3	y
1	MODEL1	PARMS	y	.	.	0.48887	-10.1280	-0.05140	0.58695	0.28685	-1
2	MODEL1	IPCVIF	y	.	1	.	.	0.25083	1.00085	0.25038	-1
3	MODEL1	IPC	y	.	1	0.55001	-9.1301	0.07278	0.60822	0.10626	-1
4	MODEL1	IPCVIF	y	.	2	.	.	0.24956	0.00095	0.24971	-1
5	MODEL1	IPC	y	.	2	1.05208	-7.7458	0.07381	0.08269	0.10735	-1

由 SAS 运行结果可以看出，这个主成分回归中回归系数的符号都是有意义的；各个回归系数的方差膨胀因子均小于 1.1；主成分回归的均方根误差是：RMSE=0.55001，虽然比最小二乘的均方根误差(RMSE=0.48887)有所增加，但增加很小。在删去第三个主成分(PCOMIT=1)后的主成分回归方程为：

y=-9.1301+0.7278x_1+0.960922x_2+0.10626x_3

这一结果与我们 SPSS 处理结果近似相等，进而互相验证了彼此的正确性。

4 结束语

本数据选自文献[1]，在文献[1]中的人工计算结果以及文献[2]通过 SAS 编程得到的计算结果都与此相同，这说明我们利用 SPSS 的计算过程与结果是正确的。另一方面，由计算过程可以看出，一道题的计算过程的实现不只是在一个操作菜单的命令下就可以完成，本例用 SPSS15.0 的 Descriptives、Data Reduction、Linear Regression、Compute Variable 模块的功能，因此对软件 SPSS 的使用要求就上升到能熟练运用的高度。本文说明，如果能在多元统计教学的同时注意有关软件的使用，开动脑筋，灵活使用，不但能很好地实现每一步的计算过程，而且还可用来解决更多新问题。这不但有利于学生掌握有关方面的知识，而且加深了对统计软件的使用和掌握，从而达到培养学生灵活应用统计软件 SPSS 的目的。

参考文献：

[1]王松桂，陈敏，陈立萍.线性统计模型：线性回归与方差分析[M].北京：高等教育出版社，2004.

[2]高惠旋.处理多元线性回归中自变量共线性的几种方法[J].数理统计与管理，2000，20(5).

[3]刘润幸，萧灿培，宫齐等.利用 SPSS 进行主成分回归分析[J].数理医药学杂志，2001,14(2).

[4]周松青.解决多重共线性问题的线性回归方法[J].江苏统计，2000，(11).

(责任编辑/易永生)