主成分分析的应用

《美国数学建模竞赛》

完整课程请长按下方二维码

例：以下是收集整理了的1990-2002年13年间影响中国蔬菜产量的若干因素数据。

请你对这些影响因素作主成分分析，并分析结果。

年份	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	y
1990	6610	4620	792	100.00	121.2	725.95	26.41	22.6	8.49	1510	686	6.21	19519
1991	6916	4749	891	106.10	123.77	812.96	26.94	22.79	8.51	1701	709	6.58	19578
1992	7030	4189	821	116.29	89.00	938.29	27.46	23.43	8.93	2027	784	6.65	19637
1993	8084	5131	861	134.54	127.34	1051.5	27.99	24.58	9.41	2577	922	6.78	19695
1994	8921	6510	923	185.94	140.58	1357.1	28.51	25.72	9.85	3496	1221	6.88	16602
1995	9514	8582	1032	240.42	146.00	1702.4	29.04	26.86	10.2	4283	1578	7.02	25723
1996	10368	9036	795	284.65	104.10	2024.2	30.48	27.89	10.6	4839	1926	7.28	30379
1997	11278	9069	818	283.23	99.70	2208.2	31.91	28.29	10.3	5160	2090	7.41	34473
1998	12291	7464	694	284.08	102.59	2336.7	33.35	28.42	10.2	5425	2162	7.55	38485
1999	13346	7905	699	285.22	115.56	2475.2	34.78	28.32	10.3	5854	2210	7.71	40514
2000	15237	9669	705	303.19	92.72	2694.7	36.22	28.44	10.7	6280	2253	7.93	42400
2001	16339	9794	680	312.89	113.72	2945.7	37.66	28.61	11.0	6860	2366	8.12	48337
2002	17353	10000	580	315.39	121.0	3184.9	39.09	28.72	11.5	7703	2476	8.33	52909
Mean	11022	7440	792	227.07	115.2	1881.4	31.53	26.51	1.00	4410	1645	7.2	
STD	3665	2155	120	85.41	17.32	843.7	4.32	2.38	0.93	2038	688	0.62	

```
data ex;
input x1-x12 y@@;
cards;
6610    4620    792 100.00  121.2   725.95  26.41   22.6    8.49    1510    686 6.21    19519
6916    4749    891 106.10  123.77  812.96  26.94   22.79   8.51    1701    709 6.58    19578
7030    4189    821 116.29  89.00   938.29  27.46   23.43   8.93    2027    784 6.65    19637
8084    5131    861 134.54  127.34  1051.5  27.99   24.58   9.41    2577    922 6.78    19695
8921    6510    923 185.94  140.58  1357.1  28.51   25.72   9.85    3496    1221    6.88    16602
9514    8582    1032    240.42  146.00  1702.4  29.04   26.86   10.2    4283    1578    7.02    25723
10368   9036    795 284.65  104.10  2024.2  30.48   27.89   10.6    4839    1926    7.28    30379
11278   9069    818 283.23  99.70   2208.2  31.91   28.29   10.3    5160    2090    7.41    34473
12291   7464    694 284.08  102.59  2336.7  33.35   28.42   10.2    5425    2162    7.55    38485
13346   7905    699 285.22  115.56  2475.2  34.78   28.32   10.3    5854    2210    7.71    40514
15237   9669    705 303.19  92.72   2694.7  36.22   28.44   10.7    6280    2253    7.93    42400
16339   9794    680 312.89  113.72  2945.7  37.66   28.61   11.0    6860    2366    8.12    48337
17353   10000   580 315.39  121.0   3184.9  39.09   28.72   11.5    7703    2476    8.35    52909
;
proc princomp out=prin;
var x1-x12;
run;
proc print data=prin;
var prin1-prin2;
run;
```

程序中对应运行结果为：

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	9.99891646	8.75163468	0.8332	0.8332
2	1.24728178	0.71487350	0.1039	0.9372
3	0.53240827	0.43460953	0.0444	0.9816
4	0.09779874	0.03612078	0.0081	0.9897
5	0.06167796	0.01106759	0.0051	0.9948
6	0.05061037	0.04491954	0.0042	0.9991
7	0.00569083	0.00209226	0.0005	0.9995
8	0.00359857	0.00192006	0.0003	0.9998
9	0.00167851	0.00145601	0.0001	1.0000
10	0.00022251	0.00010940	0.0000	1.0000
11	0.00011311	0.00011023	0.0000	1.0000
12	0.00000288		0.0000	1.0000

从程序结果可以看出，第一、第二主成分累计解释方差的比率已经达到了93.72%，所以只需要求λ_1、λ_2所对应的正交化特征向量α_i(i=1，2)

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11	Prin12
x1	0.307748	-.029214	-.265672	0.326797	0.117110	0.137092	0.579380	-.119450	0.289679	-.190526	-.376099	0.287987
x2	0.293393	0.225754	0.254769	0.267155	-.596783	0.483835	-.064732	-.281145	-.204468	-.029652	0.040101	-.094355
x3	-.221788	0.521436	0.507545	0.464716	0.415334	-.011294	0.061639	0.153738	-.001988	0.039635	0.017669	0.058315
x4	0.306417	0.106954	0.262881	-.281681	0.017840	0.150417	-.115248	0.118294	0.798551	-.047809	0.201859	-.123626
x5	-.087089	0.773652	-.567568	-.244276	-.020530	0.066172	-.053122	-.057501	0.018142	0.025259	0.002627	-.013087
x6	0.315725	0.006487	-.046116	0.027674	0.099999	0.109399	-.148654	0.403309	-.092239	0.439155	-.572683	-.401657
x7	0.305526	-.093099	-.289268	0.262396	0.233125	0.130281	0.184236	0.106559	-.110135	0.354819	0.685938	-.142491
x8	0.301308	0.128481	0.290595	-.435168	0.273157	-.192401	0.389145	-.443033	-.268772	0.004057	-.026130	-.294291
x9	0.302164	0.186914	0.066559	0.125412	-.433134	-.742267	0.086149	0.153117	0.049051	0.205821	0.053266	0.191216
x10	0.314919	0.066845	-.049385	0.009030	0.061323	-.066831	-.065328	0.458306	-.269976	-.751182	0.095029	-.158453
x11	0.311988	0.033648	0.139363	-.324237	0.170336	0.224950	-.165743	0.151565	-.252817	0.155080	0.003256	0.747625
x12	0.311315	-.023563	-.138311	0.294835	0.312333	-.219064	-.627367	-.488878	0.064630	-.094184	-.069553	0.021270

可知：$z_1 = \alpha_1 X^T, z_2 = \alpha_2 X^T, z_3 = \alpha_3 X^T$

其中：$X = (x_1, x_2, \cdots, x_{13})$

α1=(0.31，0.29，－0.22，0.30，－0.09，0.31，0.30，0.30，0.30，0.31，0.31，0.31)，

α2=（－0.03，0.23，0.52，0.11，0.77，0.01，－0.09，0.13，0.19，0.07，0.03，－0.03）

第一主成分与蔬菜种植面积、每公顷物质费用、蔬菜零售物价指数、市场化程度、城市化水平1、城市化水平2、交通、城镇居民可支配收入、农村居民纯收入、农民文化素质等密切相关，表示的是市场经济综合因素，着重反映的是市场经济的成熟程度与国家现代化水平；**这些都是重要的因素**

第二主成分与每公顷劳动投入、成本纯收益率等密切相关，表示的是劳动者动力因素。

主成分得分

The SAS System

Obs	Prin1	Prin2
1	-4.41530	-0.66608
2	-4.22732	-0.10660
3	-3.53464	-1.88899
4	-2.92766	0.27785
5	-1.90216	1.52193
6	-0.82758	2.67758
7	0.96255	0.00949
8	1.32761	-0.14741
9	1.68568	-0.76619
10	2.10166	-0.13046
11	3.18891	-0.86691
12	3.82051	0.03735
13	4.74773	0.04844

主成分得分一旦得到了，便可以做综合评价和主成分回归

先看综合评价

```
data ex1;input obs prin1 prin2@@;
pingjia=0.8332*prin1+0.1039*prin2;
cards;
1   -4.41530   -0.66608    2   -4.22732   -0.10660
3   -3.53464   -1.88899    4   -2.92766    0.27785
5   -1.90216    1.52193    6   -0.82758    2.67758
7    0.96255    0.00949    8    1.32761   -0.14741
9    1.68568   -0.76619   10    2.10166   -0.13046
11   3.18891   -0.86691   12    3.82051    0.03735
13   4.74773    0.04844
;
proc print;var pingjia;
run;
```

The SAS System

Obs	pingjia
1	-3.74803
2	-3.53328
3	-3.14133
4	-2.41046
5	-1.42675
6	-0.41134
7	0.80298
8	1.09085
9	1.32490
10	1.73755
11	2.56693
12	3.18713
13	3.96084

再看看主成分回归

思想是：

- 1. 得到主成分后，将主成分得分当作自变量。
- 2. 建立因变量和主成分得分的回归方程，并进行检验。
- 3. 调整模型得到最终的以主成分为自变量的回归模型。
- 4. 因主成分是用标准化的变量线性表示的，那么就可以将主成分回代到回归方程中，得到因变量与原始变量的的回归模型，这就是主成分回归模型。

- data ex;
- input x1-x12 y@@;
- cards;
- …
- ;
- proc princomp out=prin;
- var x1-x12;
- run;
- proc print data=prin;
- var prin1-prin2;
- run;
- proc reg;model y=prin1-prin2;
- run;

The SAS System 15:55 Saturday,

The REG Procedure
Model: MODEL1
Dependent Variable: y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1707112793	853556396	74.46	<.0001
Error	10	114636220	11463622		
Corrected Total	12	1821749013			

Root MSE	3385.79710	R-Square	0.9371
Dependent Mean	31404	Adj R-Sq	0.9245
Coeff Var	10.78145		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	31404	939.05116	33.44	<.0001
Prin1	1	3727.28403	309.09632	12.06	<.0001
Prin2	1	-1638.36168	875.16112	-1.87	0.0907

- **proc reg** data=ex outest=out1;
- model y=x1-x12/pcomit=**11** outvif;
- **proc print** data=out1;**run**;

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	1821749013	151812418	.	.
Error	0	.	.		
Corrected Total	12	1821749013			

Root MSE	.	R-Square	.
Dependent Mean	31404	Adj R-Sq	.
Coeff Var	.		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	482141	.	.	.
x1	1	41.10768	.	.	.
x2	1	-24.64181	.	.	.
x3	1	280.91390	.	.	.
x4	1	-955.10710	.	.	.
x5	1	-335.77940	.	.	.
x6	1	-216.18600	.	.	.
x7	1	-15228	.	.	.
x8	1	-72973	.	.	.
x9	1	125796	.	.	.
x10	1	-64.83844	.	.	.

上面的例子不能回代称原始数据，那就再看另外一个例子的结果。

The SAS System 15:55 Saturday, January 17, 2020 87

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RIDGE_	_PCOMIT_	_RMSE_	Intercept	z1	z2	z3	z4	z5	k1
1	MODEL1	PARMS	k1	.	.	0.07600	4.45239	0.4703	-0.17072	-0.16273	-0.0767	-0.0153	-1
2	MODEL1	IPCVIF	k1	.	1	.	.	35.6027	5.22339	1.64911	34.4574	11.5889	-1
3	MODEL1	IPC	k1	.	1	0.07420	4.23322	0.3855	0.02684	-0.26496	-0.1518	0.0473	-1
4	MODEL1	IPCVIF	k1	.	2	.	.	0.8750	4.99467	1.58463	0.4170	11.5210	-1
5	MODEL1	IPC	k1	.	2	0.07721	4.49049	0.0966	0.05750	-0.28954	0.0690	0.0356	-1
6	MODEL1	IPCVIF	k1	.	3	.	.	0.2198	0.05239	1.54331	0.1909	0.0897	-1
7	MODEL1	IPC	k1	.	3	0.07469	4.53062	0.0912	0.03791	-0.28683	0.0665	0.0565	-1
8	MODEL1	IPCVIF	k1	.	4	.	.	0.0489	0.05093	0.02944	0.0494	0.0494	-1
9	MODEL1	IPC	k1	.	4	0.10593	3.60804	0.0336	0.04486	0.05148	0.0261	0.0309	-1

主要看最后两行。倒数第二行的IPCVIF均小于1，说明去掉4个主成分后，已经基本上消除了多重共线的问题。5个原始变量对应的系数分别为0.0336、0.04486、0.05148、0.0261、0.0309，基本上可以说明z3最为重要。