主成分回归分析

1.2 主成分回归

多重共线性解决方法：

主成分回归

岭回归

偏最小二乘回归

1.2 主成分回归---基本思想

主成分分析(principal components analysis,PCA)

- 一种降维的思想，在损失很少信息的前提下把多个指标利用正交旋转变换转化为几个综合指标的多元统计分析方法。
- 转化生成的综合指标称为主成分，其中每个主成分都是原始变量的线性组合，且各个主成分之间彼此独立。
- 根据主成分方差递减排序，选取前几个方差最大的主成分替代原变量，用较少的主成分就能综合反映原变量中所包含的主要信息。

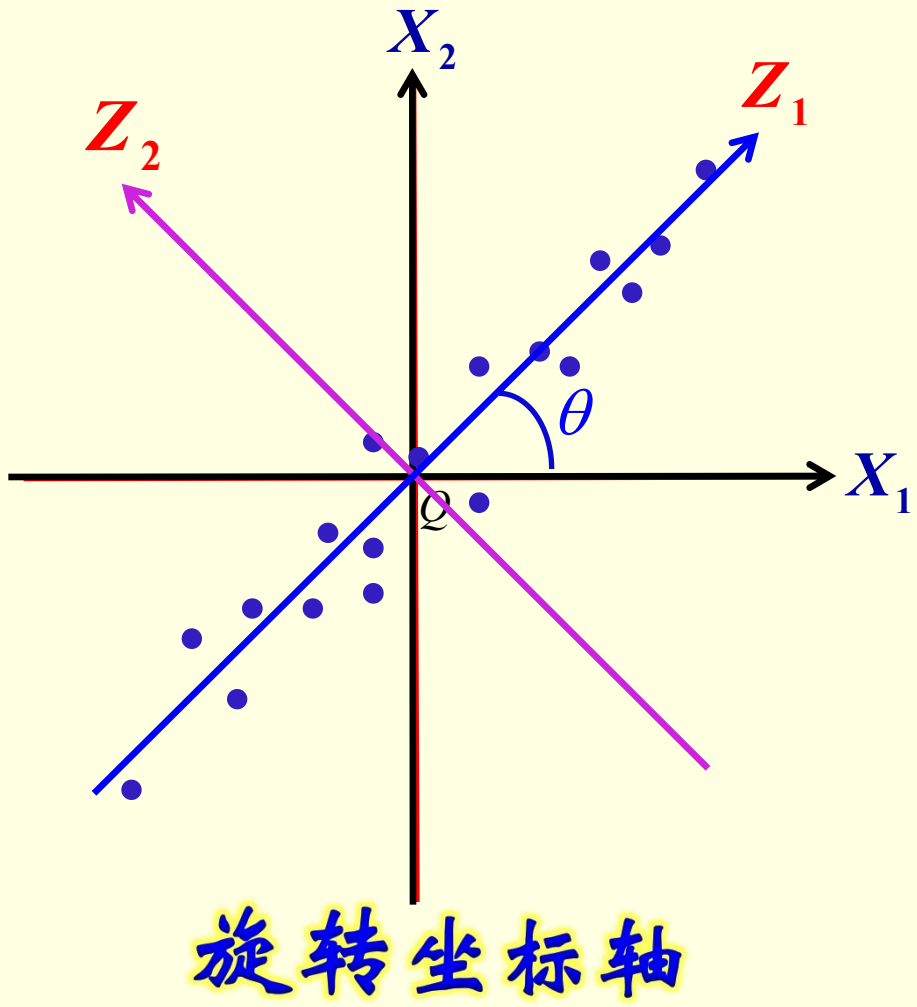
1.2 主成分回归---基本思想

二元情形

$$\begin{cases} Z_1 = X_1\cos\theta + X_2\sin\theta \\ Z_2 = -X_1\sin\theta + X_2\cos\theta \end{cases}$$

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

正交



旋转坐标轴

1.2 主成分的含义

n元情形

假设 $X=(X_1 \quad \cdots \quad X_p)'$ 是原始指标，协方差矩阵为 A

$$\begin{cases} Z_1=u_{11}X_1+u_{21}X_2+\cdots+u_{p1}X_p=u_1'X, \\ Z_2=u_{12}X_1+u_{22}X_2+\cdots+u_{p2}X_p=u_2'X, \\ \vdots \\ Z_p=u_{1p}X_1+u_{2p}X_2+\cdots+u_{pp}X_p=u_p'X, \end{cases} \Rightarrow Z=U'X$$

满足以下条件：

1、$u_{1i}^2+\cdots+u_{pi}^2=1$

2、$u_i'u_j=0(i\neq j)$

3、$\mathrm{Var}(Z_1)\geq \mathrm{Var}(Z_2)\geq\cdots\geq \mathrm{Var}(Z_p)$

➢说明：寻找主成分的过程实际上是寻找使得协方差阵矩阵 A 对角化的正交矩阵。

1.2 主成分的含义

➢例：服装定型分类问题（上衣）

型

领围、肩宽、臂围、
胸围、腰围、臀围

号

领长、袖长、衣长

服装厂制衣 - - - - - 批量生产

9个指标 —降维→ 2个综合指标

1.2 主成分的含义

主成分的选取

$\dfrac{\lambda_k}{\sum_{i=1}^{p}\lambda_i}$ Z_k的方差贡献率

$\dfrac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{p}\lambda_i}$ $Z_1, Z_2, \cdots, Z_k$的累积方差贡献率

进行主成分分析的目的之一是希望用尽可能少的主成分代替原来的个指标，在实际中：

选取原则：累计差贡献率>80%

1.2 主成分的含义

求主成分的步骤

- 列出观测资料矩阵 X；
- 计算样本协方差 S；
- 计算 S 的特征值和单位正交特征向量；
- 计算方差贡献率及累计方差贡献率；
- 确定主成分个数，建立主成分方程；
- 解释主成分的实际意义。

1.2 主成分回归---基本思想

主成分回归

- 将多个彼此相关、信息重叠的指标通过适当的线性组合，使之成为彼此独立或者不相关，而又提取了原指标变异信息的综合变量（即主成分），然后建立因变量与主成分的回归关系式，最后还原为因变量与原自变量的回归方程。

1.2 主成分回归

主成分回归的步骤

- 进行多元线性回归分析及**共线性诊断**；
- 若存在共线性，进行**主成分分析**；
- 求得主成分**得分**，确定主成分**个数**；
- 将因变量对保留主成分进行回归分析；
- **回代**主成分，得到新的回归模型；
- 对回归方程给于**专业解释**。

1.2 主成分回归

例1.1.1 下表是1990-2007年中国棉花单产与要素投入表格。请对5个要素投入做共线性诊断，并做单产对于5个要素投入的主成分回归模型，指出哪个要素投入是最重要的要素？

年份	单产 kg/公顷	种子费 元/公顷	化肥费 元/公顷	农药费 元/公顷	机械费 元/公顷	灌溉费 元/公顷
1990	1017.0	106.05	495.15	305.1	45.9	56.1
1991	1036.5	113.55	561.45	343.8	68.55	93.3
1992	792.0	104.55	584.85	414	73.2	104.55
1993	861.0	132.75	658.35	453.75	82.95	107.55
1994	901.5	174.3	904.05	625.05	114	152.1
1995	922.5	230.4	1248.75	834.45	143.85	176.4
1996	916.5	238.2	1361.55	720.75	165.15	194.25
1997	976.5	260.1	1337.4	727.65	201.9	291.75
1998	1024.5	270.6	1195.8	775.5	220.5	271.35
1999	1003.5	286.2	1171.8	610.95	195	284.55
2000	1069.5	282.9	1151.55	599.85	190.65	277.35

1.2 主成分回归

设 y 表示因变量指标棉花每公顷的产量，x_1, x_2, x_3, x_4, x_5 分别表示自变量指标每公顷的种子费、化肥费、农药费、机械费、灌溉费，令 t 表示年份。下面用主成分回归分析的方法进行求解。

① 多元线性回归的结果

方差来源	自由度	平方和	均方	F 值	p 值
回归	5.0000	231088.9692	46217.7938	6.8580	0.0031
残差	12.0000	80871.5308	6739.29423		
总计	17.0000	311960.5000			

均方根误差	82.0932	R 方	0.7408
因变量均值	1039.8333	调整 R 方	0.6327

1.2 主成分回归

变量	估计值	标准误差	t 值	p 值	方差膨胀因子
常数项	947.0456	95.5299	9.9136	0.0000	0.0000
x_1	0.7762	1.1522	0.6737	0.5133	68.2329
x_2	-0.0929	0.2742	-0.3388	0.7406	35.8803
x_3	-0.2550	0.3051	-0.8359	0.4195	4.9573
x_4	-0.1556	0.7688	-0.2024	0.8430	32.3848
x_5	0.6378	0.6866	0.9290	0.3712	16.3291

说明：自变量对因变量的影响均不显著，自变量间的多重共线性

1.2 主成分回归

②共线性诊断的结果

序号	特征值	条件指数	x_1	x_2	x_3	x_4	x_5
1	4.0945	1.0000	0.0008	0.0016	0.0039	0.0017	0.0034
2	0.7927	2.2728	0.0011	0.0002	0.2165	0.0028	0.0011
3	0.0817	7.0814	0.0001	0.0385	0.0001	0.0805	0.5055
4	0.0207	14.0644	0.0000	0.7706	0.5817	0.5474	0.0030
5	0.0104	19.7974	0.9981	0.1891	0.1979	0.3676	0.4869

③ 简单统计量及相关系数矩阵

	x_1	x_2	x_3	x_4	x_5
均值	290.2250	1188.5083	580.0667	224.2917	251.9750
标准差	142.7373	435.0308	145.3030	147.3854	117.1903

1.2 主成分回归

	x_1	x_2	x_3	x_4	x_5
x_1	1.0000	0.9435	0.3808	0.9802	0.9588
x_2	0.9435	1.0000	0.6156	0.9318	0.9232
x_3	0.3808	0.6156	1.0000	0.3518	0.4629
x_4	0.9802	0.9318	0.3518	1.0000	0.9205
x_5	0.9588	0.9232	0.4629	0.9205	1.0000

④ 主成分分析的结果

序号	特征值	差分	贡献率	累积贡献率
1	4.0945	3.3019	0.8189	0.8189
2	0.7927	0.7110	0.1585	0.9774
3	0.0817	0.0610	0.0163	0.9938
4	0.0207	0.0103	0.0041	0.9979
5	0.0104	0.0000	0.0021	1.0000

1.2 主成分回归

	Z_1	Z_2	Z_3	Z_4	Z_5
x_1^*	0.4810	-0.2384	-0.0178	-0.0039	0.8435
x_2^*	0.4875	0.0792	-0.3359	-0.7565	-0.2662
x_3^*	0.2814	0.9224	-0.0045	0.2443	0.1012
x_4^*	0.4732	-0.2683	-0.4613	0.6058	-0.3527
x_5^*	0.4773	-0.1185	0.8210	0.0321	-0.2882

可选择前两个主成分替代原来的5个自变量。

$$Z_1 = 0.4810x_1^* + 0.4875x_2^* + 0.2814x_3^* + 0.4732x_4^* + 0.4773x_5^*,$$

$$Z_2 = -0.2384x_1^* + 0.0792x_2^* + 0.9224x_3^* - 0.2683x_4^* - 0.1185x_5^*$$

$$Z_3 = -0.0178x_1^* - 0.3359x_2^* - 0.0045x_3^* - 0.4613x_4^* + 0.8210x_5^*$$

$$Z_4 = -0.0039x_1^* - 0.7565x_2^* + 0.2443x_3^* + 0.6058x_4^* + 0.0321x_5^*$$

$$Z_5 = 0.8435x_1^* - 0.2662x_2^* + 0.1012x_3^* - 0.3527x_4^* - 0.2882x_5^*,$$

1.2 主成分回归

第一主成分大约等于这五项投入要素和的一个常数倍，可以简称为总投入主成分；第二主成分Z_2的表达式中系数为正的是化肥费和农药费，系数为负的是种子费、机械费和灌溉费，其中农药费的符号是正的，且系数为0.9224，相比其他投入占有重要优势，可以简称为药物主成分

⑤ 标准化因变量对主成分Z_1, Z_2进行多元回归分析

表 1.1.11 方差分析表

方差来源	自由度	平方和	均方	F 值	p 值
回归	2.0000	11.9879	5.9939	17.9384	0.0001
残差	15.0000	5.0121	0.3341		
总计	17.0000	17.0000			

表 1.1.12 拟合优度检验

均方根误差	0.5780	R 方	0.7052
因变量均值	0.0000	调整 R 方	0.6659

1.2 主成分回归

变量	估计值	标准误差	t 值	p 值	方差膨胀因子
常数项	0.0000	0.1362	0.000	1.0000	0.0000
Z_1	0.3544	0.0693	5.1147	0.0001	1.0000
Z_2	-0.4909	0.1575	-3.1172	0.0071	1.0000

由上表，可得到标准化因变量对主成分的回归方程为：

$$y^* = 0.3544Z_1 - 0.4909Z_2$$

⑥ 因变量与自变量的回归方程

$$y^* = 0.2875x_1^* + 0.1339x_2^{*-} - 0.3531x_3^* + 0.2994x_4^* + 0.2273x_5^*$$

再将标准化变量还原为原始变量：

$$y = 974.1162 + 0.2728x_1 + 0.0417x_2 - 0.3292x_3 + 0.2752x_4 + 0.2627x_5$$

1.2 主成分回归

小结：

主成分回归的方法是将原来的回归自变量变换到主成分，选择其中重要的主成分作为新的自变量，丢弃了影响不大的自变量，实际上达到了降维的目的，然后用原来最小二乘法对选取主成分后的模型参数进行估计，然后再变换回原来的模型求出参数的估计。

优点：简化结构、消除变量间的相关性
缺点：回归方程的解释比较复杂

我们通常仅将主成分回归作为分析多重共线性问题的一种方法。根据实际分析效果来评价方法的适用性。