**主成分分析**

(principal components analysis)

**主成分分析的理论框架**

假设我们原始的变量有个，我们用，这样这个变量组成了一个向量，其中的均值为，协方差矩阵为。

协方差矩阵的表示形式：



表示的方差，表示的协方差。

对进行线性变换的时候，形成的综合变量我们用来表示，那么我们可以得到下面的方程组：



其中：

那么我们希望的方差尽可能大(因为方差越大，主成分的贡献率越大)，并且要求和之间是互不相关的，换言之相互独立。即：。

**主成分的几条重要的理论****性质**

⑴第个主成分与原始变量的相关系数称为因子负荷量，这个因子负荷量在软件操作中可以显示出来。如果我们用协方差矩阵来求解主成分，那么，因此，我们在解释主成分和某个变量的重要性的时候，要根据因子负荷量而不是简单的变换系数；我们用相关矩阵来求解主成分，则(一般当变量的数据数量级差别比较大，我们要进行标准化处理，就会用到用相关矩阵来求解主成分)。

⑵

⑶

**性质2和3在结合后面例子来讲述。**

**主成分的几个很重要的用途**

**⑴进行分类**

我们可以通过主成分分析得到主成分得分，通过计算出总得分来进行分类，或者将前两个主成分得分放到四象限图中来进行分类。

**⑵进行排名**

通过主成分分析得到主成分得分，通过一定的手段计算出总得分来进行对样本的排名。

**⑶ 主成分回归**

由于在实际问题中，我们尽可能多的选取变量，这样会导致多重共线性问题的出现。主成分分析可以用少数几个综合变量来代替原始的变量，很有效的消除多重共线性。

**实际例子(分类、排名)**

在企业经济效益的评价中，设计的指标往往很多。为了可以简化系统结构，抓住经济效益评价的主要问题，我们用百元固定资产原值实现值、百元固定资产原值实现利税、百元资产实现利税、百元工业总产值实现利税、百元销售收入实现利税、每吨标准煤实现工业产值、每千瓦时电力实现工业产值、全员劳动生产率、百元流动资金实现产值，涉及到9项指标，28个样本数。

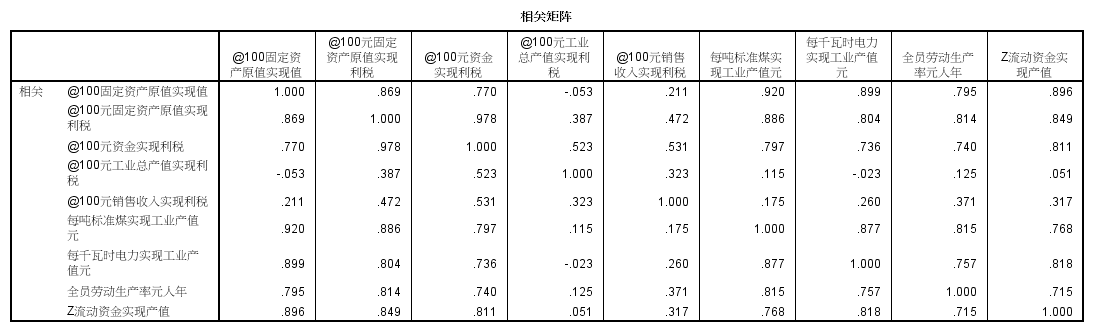


图1：相关矩阵

应该有这个意识，变量之间的存在着较强的相关性，主成分分析才会更加适用。而且如果原始大部分变量间的相关系数都小于0.3，运用主成分分析不会得到的很好的效果。经过图1，我们可以看到这9个变量之间的相关系数矩阵，我们发现变量间的相关性较强，适合运用主成分分析来进行后续的工作。



图2：信息提取率

上图是在主成分分析过程中从每一个变量中提取的信息。例如：百元固定资产原值实现值(96.7%)、百元固定资产原值实现利税(97.8%)、百元资产实现利税(97%)、百元工业总产值实现利税(79.9%)、百元销售收入实现利税(54.3%)、每吨标准煤实现工业产值(89.2%)、每千瓦时电力实现工业产值(87.9%)、全员劳动生产率(76.3%)、百元流动资金实现产值(83.2%)。

我们可以看到，除了百元销售收入实现利税(54.3%)信息损失的较多外，在其余变量提取的信息还是可以的。

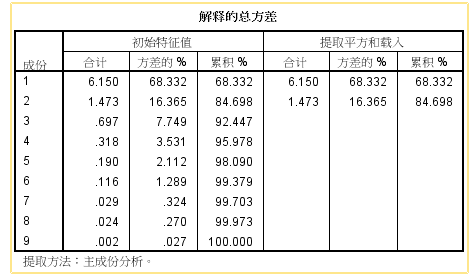


图3：方差贡献率

在提取主成分的时候，我们一般保留特征根大于1的主成分。在这个案例中，我们只保留前两个主成分，这样我们就可以在损失较少信息的前提下，用两个综合指标(主成分)来代替原始的9个变量，达到降维的效果。

(96.7%+97.8%+97%+79.9%+54.3%+89.2%+87.9%+76.3%+83.2%)/9=84.7%

我们也可以通过下图的碎石图来进行主成分数量的选取。

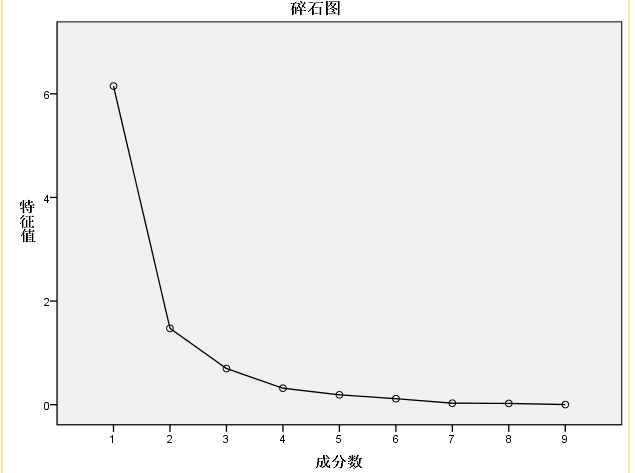


图4：碎石图

观察碎石图，我们发现第二个和第三个特征值的变化已经趋于平稳，则说明只提取两个或者三个主成分即可。

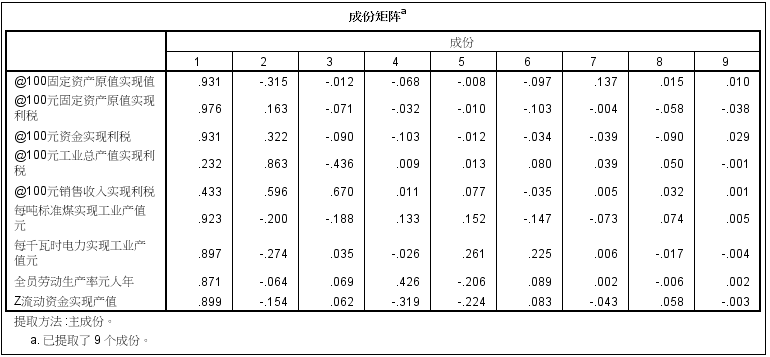


图4：载荷矩阵

虽然我们在选取主成分的时候，只需要选取前两个主成分即可，但是为了说明我们前面的主成分的性质，我们将这9个主成分全部列出。

⑴.

满足性质3

⑵.

说明信息提取率

⑶. 

满足性质2

在软件输出的结果中，我们得到的是因子载荷矩阵，而不是主成分的系数矩阵，由前面得到的，因此我们要对因子载荷矩阵中的每一列除以对应的特征根的平方根，就可以得到主成分分析的系数。(用相关矩阵来求解主成分，则)。

表1：主成分列表

|  |  |  |  |
| --- | --- | --- | --- |
| 第一主成分 | 第二主成分 | 修改的第一主成分 | 修改的第二主成分 |
| 0.93100 | -0.31500 | 0.37542 | -0.25954 |
| 0.97600 | 0.16300 | 0.39356 | 0.13430 |
| 0.93100 | 0.32200 | 0.37542 | 0.26531 |
| 0.23200 | 0.86300 | 0.09355 | 0.71107 |
| 0.43300 | 0.59600 | 0.17460 | 0.49107 |
| 0.92300 | -0.20000 | 0.37219 | -0.16479 |
| 0.89700 | -0.27400 | 0.36171 | -0.22576 |
| 0.87100 | -0.06400 | 0.35122 | -0.05273 |
| 0.89900 | -0.15400 | 0.36251 | -0.12689 |

那么我们可以得到前两个主成分的线性组合：





我们用上面的公式可以得到第一主成分和第二主成分的得分。

表2：主成分得分

|  |  |  |  |
| --- | --- | --- | --- |
| 序号 | 省区 | 第一主成分得分prin1 | 第二主成分得分prin2 |
| 1 | 北京（1） | 2.81627 | 2.42753 |
| 2 | 天津（2） | 3.73587 | .62388 |
| 3 | 河北（3） | -.48683 | -.06118 |
| 4 | 山西（4） | -2.02201 | .37284 |
| 5 | 内蒙（5） | -2.97632 | -.81937 |
| 6 | 辽宁（6） | -.41802 | 1.29095 |
| 7 | 吉林（7） | -1.61357 | -.83425 |
| 8 | 黑龙讲（8） | -1.04155 | .65784 |
| 9 | 上海（9） | 7.03779 | 1.53631 |
| 10 | 江苏（10） | 3.94431 | -2.66205 |
| 11 | 浙江（11） | 4.36848 | -1.70452 |
| 12 | 安徽（12） | .07225 | -.53732 |
| 13 | 福建（13） | .51241 | -.51349 |
| 14 | 江西（14） | -1.18986 | -1.27711 |
| 15 | 山东（15） | .98439 | -.31534 |
| 16 | 河南（16） | -1.02755 | .07585 |
| 17 | 湖北（17） | .35485 | -.28115 |
| 18 | 湖南（18） | -.04354 | .40715 |
| 19 | 广东（19） | 1.81895 | -1.26344 |
| 20 | 广西（20） | .13750 | .72212 |
| 21 | 四川（21） | -1.40995 | -.67580 |
| 22 | 贵州（22） | -2.24420 | 1.20697 |
| 23 | 云南（23） | -.02115 | 2.40220 |
| 24 | 陕西（24） | -1.62420 | -.50005 |
| 25 | 甘肃（25） | -1.64508 | 1.68816 |
| 26 | 青海（26） | -3.40828 | -.74210 |
| 27 | 宁夏（27） | -3.06816 | -.90320 |
| 28 | 新疆（28） | -1.54278 | -.32145 |

我们利用以第一主成分为x轴，第二主成分为y轴，建立平面直角坐标系。并且用象限图来进行表示。

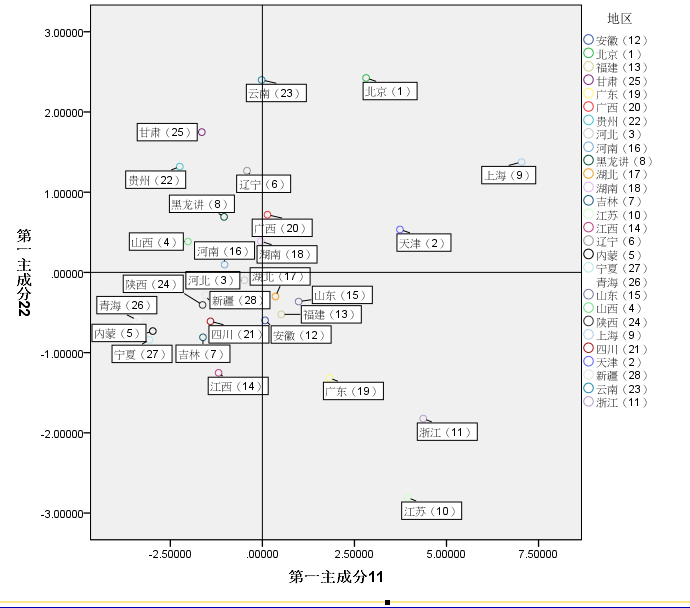


图5：四象限分布图

首先，分布在第一象限的是上海、北京、天津和广西四个省，这四个省的效益在全国属于比较好的(个人觉得不好，广西觉得不在这个范围里面)。第四象限的是湖北、山东、福建、安徽、广东、江苏、浙江7个省区，由于第四象限的主要特征第一主成分，第一主成分所占的信息较大，效益也不错。分布在第二三象限的为一类，效益不好。

下面我们来计算总得分并且来对上面28个省区进行排名。

⑴利用得到下面的排名

表3：法1计算排名

|  |  |  |
| --- | --- | --- |
| 序号 | 省区 | 得分 |
| 9 | 上海（9） | 5.03468 |
| 11 | 浙江（11） | 2.68699 |
| 2 | 天津（2） | 2.64059 |
| 1 | 北京（1） | 2.32140 |
| 10 | 江苏（10） | 2.23589 |
| 19 | 广东（19） | 1.02776 |
| 15 | 山东（15） | .61328 |
| 23 | 云南（23） | .37835 |
| 13 | 福建（13） | .26453 |
| 20 | 广西（20） | .21161 |
| 17 | 湖北（17） | .19334 |
| 18 | 湖南（18） | .03369 |
| 12 | 安徽（12） | -.04855 |
| 6 | 辽宁（6） | -.07818 |
| 3 | 河北（3） | -.34866 |
| 8 | 黑龙讲（8） | -.59842 |
| 16 | 河南（16） | -.68615 |
| 25 | 甘肃（25） | -.83795 |
| 14 | 江西（14） | -1.01810 |
| 21 | 四川（21） | -1.06332 |
| 28 | 新疆（28） | -1.10098 |
| 24 | 陕西（24） | -1.17631 |
| 7 | 吉林（7） | -1.23509 |
| 22 | 贵州（22） | -1.31773 |
| 4 | 山西（4） | -1.31874 |
| 5 | 内蒙（5） | -2.15366 |
| 27 | 宁夏（27） | -2.23435 |
| 26 | 青海（26） | -2.42592 |

⑵由于第一主成分占用的信息量很大，因此我们可以第一主成分来进行排名，前提是第一主成分的系数必须全部为正数。

表4：法2计算排名

|  |  |  |
| --- | --- | --- |
| 序号 | 省区 | 得分 |
| 9 | 上海（9） | 7.03779 |
| 11 | 浙江（11） | 4.36848 |
| 10 | 江苏（10） | 3.94431 |
| 2 | 天津（2） | 3.73587 |
| 1 | 北京（1） | 2.81627 |
| 19 | 广东（19） | 1.81895 |
| 15 | 山东（15） | .98439 |
| 13 | 福建（13） | .51241 |
| 17 | 湖北（17） | .35485 |
| 20 | 广西（20） | .13750 |
| 12 | 安徽（12） | .07225 |
| 23 | 云南（23） | -.02115 |
| 18 | 湖南（18） | -.04354 |
| 6 | 辽宁（6） | -.41802 |
| 3 | 河北（3） | -.48683 |
| 16 | 河南（16） | -1.02755 |
| 8 | 黑龙讲（8） | -1.04155 |
| 14 | 江西（14） | -1.18986 |
| 21 | 四川（21） | -1.40995 |
| 28 | 新疆（28） | -1.54278 |
| 7 | 吉林（7） | -1.61357 |
| 24 | 陕西（24） | -1.62420 |
| 25 | 甘肃（25） | -1.64508 |
| 4 | 山西（4） | -2.02201 |
| 22 | 贵州（22） | -2.24420 |
| 5 | 内蒙（5） | -2.97632 |
| 27 | 宁夏（27） | -3.06816 |
| 26 | 青海（26） | -3.40828 |

我们从上面的排名其实可可以进行聚类，以分数的形式呈现。

**实际例子(主成分回归)**

为了研究我国民航客运量的变化趋势及其成 因，我们以民航客运量作为因变量y(万人)，以国民收 入x1(亿元)、消费额x2(亿元)、铁路客运量x3(万人)、

民航航线里程x4(万里程)、来华旅游人数x5(万人)为影响民航客运量的主要因素，建立回归模型。

对于上面的例子，我们如果做简单的线性回归，并进行共线性诊断，我们可以发现：

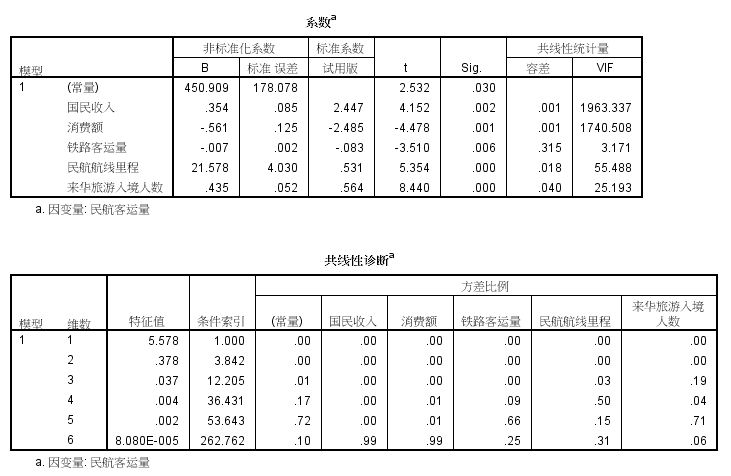


图6：共线性诊断图

通过上图，根据容忍度和方差扩大因子，我们可以知道该问题出现很大的多重共线性。可以运用主成分回归的手段来消除多重共线性。

首先我们运用主成分分析找到两个综合因素作为两个主成分prin1和prin2，然后先以民航客运量y为因变量，以prin1和prin2作为自变量来建立回归方程(**具体做法和前面一样，数据见SPSS软件**)，y和prin1、prin2进行普通最小二乘：

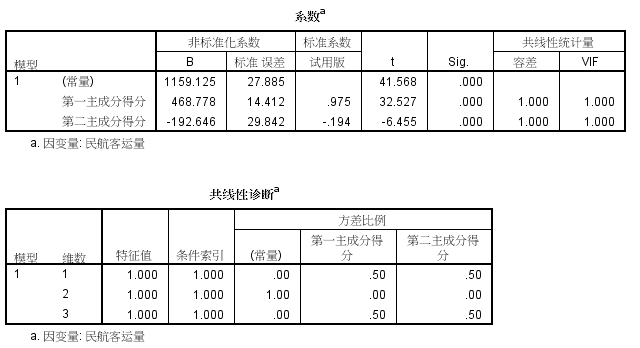
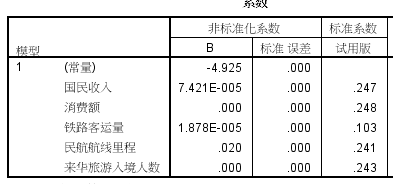


图7：主成分回归步骤一



然后，以prin1、prin2作为因变量，x1(亿元)、消费额x2(亿元)、铁路客运量x3(万人)、民航航线里程x4(万里程)、来华旅游人数x5(万人)做自变量进行最小二乘：



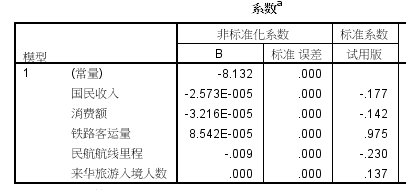


图8：主成分回归步骤二





最后把步骤二的带到步骤一中去，得到最后的方程：

