* **主成分分析的基本思想**

**主成分分析就是把原有的多个指标转化成少数几个代表性较好的综合指标，这少数几个指标能够反映原来指标大部分的信息（85%以上），并且各个指标之间保持独立，避免出现重叠信息。**

**主成分分析主要起着降维和简化数据结构的作用。**

**主成分分析是把各变量之间互相关联的复杂关系进行简化分析的方法。**

**主成分分析试图在力保数据信息丢失最少的原则下，对这种多变量的截面数据表进行最佳综合简化，也就是说，对高维变量空间进行降维处理。**

* **要讨论的问题**

1. **基于相关系数矩阵/协方差矩阵做主成分分析？**
2. **选择几个主成分？**
3. **如何解释主成分所包含的经济意义？**

* **数学模型与集合解释**

**假设我们的实际问题中，有个指标，记为**

**主成分分析就是要把这个指标的问题，转变为讨论个新的指标：**

****

**按照保留主要信息量的原则，充分反映原指标的信息，并且相互独立。**

****

**其中，**

**这种由讨论多个指标降为少数几个综合指标的过程在数学上就叫做降维。主成分分析通常的做法是：寻求原指标的线性组合。**

****

**满足如下的条件：**

**每个主成分的系数平方和为1，即：**

**主成分之间相互独立，即无重叠的信息，即：**

**主成分的方差依次递减，重要性依次递减，即：**

* **求解主成分的步骤**

1. **求样本均值和样本协方差矩阵；**
2. **求的特征根**

**求解特征方程，其中是单位矩阵，解得2个特征根，，其中**

1. **求特征根所对应的单位特征向量**
2. **写出主成分的表达式**

* **主成分的含义**

**通过分析主成分的表达式中原变量前的系数来解释各主成分的含义。**

* **比较主成分重要性**

**第一主成分的方差为，方差贡献率为：**

****

**第二主成分的方差为，方差贡献率为：**

****

**总方差保持不变。**

* **主成分分析的步骤**

**在一般情况下，设有个样本，每个样本观测个指标，将原始数据排成如下矩阵：**

****

1. **求样本均值和样本协方差矩阵**
2. **求解特征方程，其中是单位矩阵**

****

**解得个特征根，且**

1. **求所对应的单位特征向量，**

**即需求解方程组，其中**

****

**再加上单位向量的条件**

**解得**

1. **写出主成分的表达式**

****

* **主成分个数的选取原则**

**根据累积贡献率的大小取前面个（）主成分**

**选取原则：**

**且**

* **主成分求解步骤**

1. **对数据进行标准化变换**
2. **求样本协方差矩阵**

**（求相关系数矩阵）**

1. **求特征根**
2. **求单位特征向量**
3. **写出主成分的表达式**

****

* **主成分的相关结构**

**主成分的方差**

**主成分的方差贡献率为**

**主成分与每个变量之间的相关系数**

**主成分对每个原变量的方差贡献**

* **主成分的性质**

1. **主成分的协差阵为对角阵**

****

1. **总方差保持不变**

****

**若进行型分析，则**

1. **与的相关系数**

****

**若进行型分析，则**

1. **对的方差贡献率为**

**从横行看有：**

**从纵向看有：**

**若进行型分析，则：**

****

* **主成分分析用于系统评估**

**第一种方法，通过主成分分析得到综合指标**

****

**利用作为评估指标，根据得分对样本点进行排序比较，但是有3个前提条件：**

1. **与全体原变量都正相关，即**
2. **各在数值上的分布较为均匀。**
3. **的方差贡献率较大**

**第二种方法，通过主成分分析，取前面个主成分，以每个主成分的方差贡献率为权，构造综合评价函数，按值得大小对样品进行排序比较或分类。**

**【注意：实际上，这一方法不合理，的含义违背了综合评价的本意。】**

* **SAS程序**

**proc princomp data=economic out=prin;**

**run;**

**proc princomp data=economic out=prin n=3 plot=pattern(ncomp=2) plot=score(ncomp=2);**

**var x1-x10;**

**id region;**

**run;**

* **主成分回归**

**解决回归分析中的多重共线性问题**

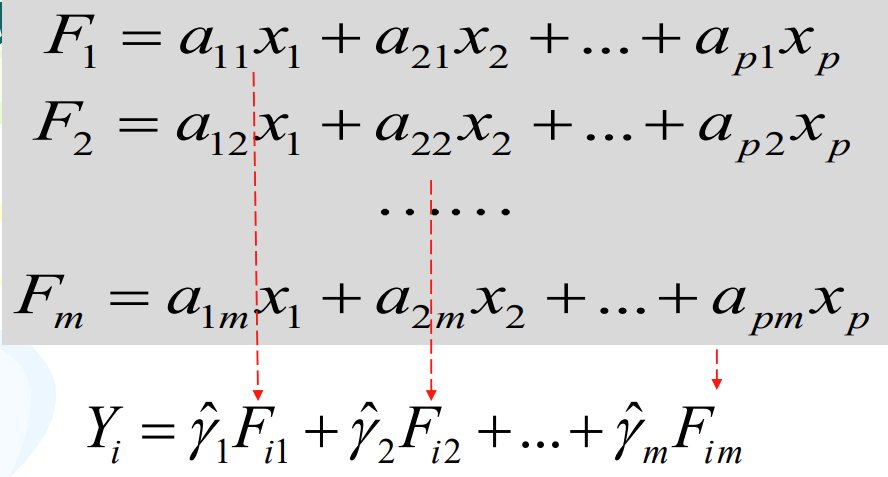
* **主成分回归方法**

1. **对自变量进行主成分分析**

****

**取个主成分，反映原个变量95%以上的信息。**

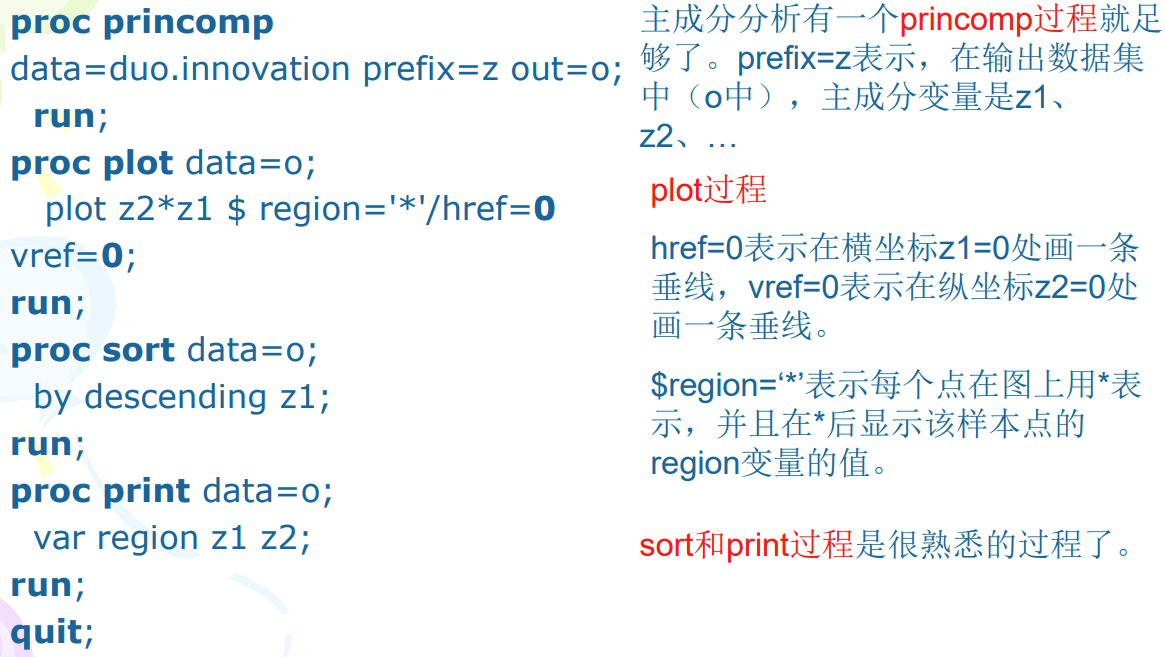
1. **得到主成分的表达式：**

****

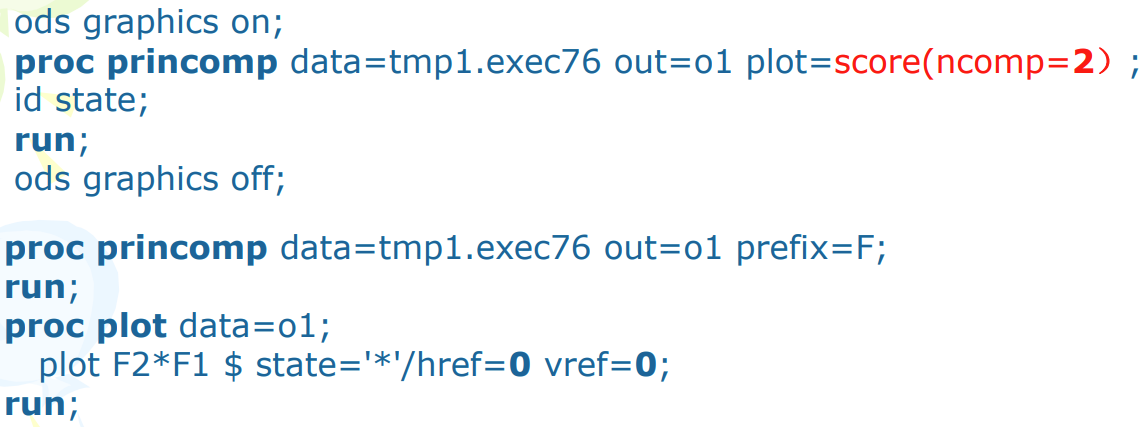
1. **把主成分的表达式代入，得到最终的回归模型**

****

* **主成分回归在SAS中用princomp过程**

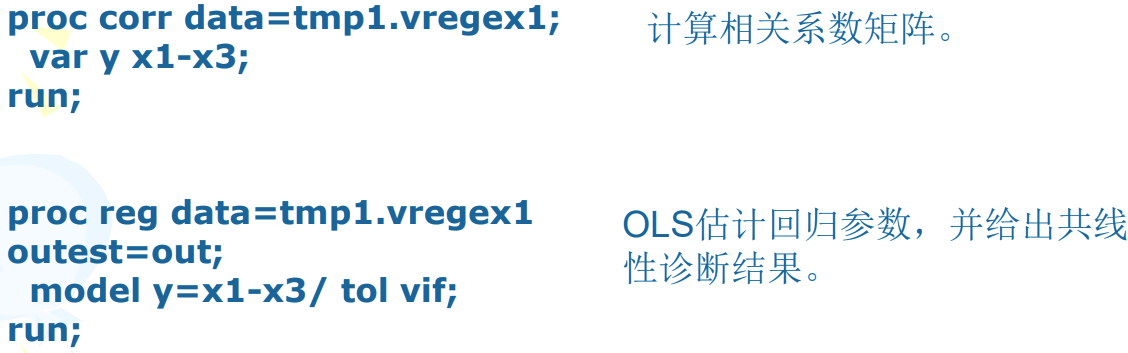
****

* **主成分分析图解样品**

****

* **主成分分析图解变量**

****

****

****

**因子分析**

* **因子分析的基本思想**

**因子分析是根据相关矩阵内部的依赖关系，把一些具有错综复杂关系的变量综合为数量较少的几个因子。通过不同因子来分析决定某些变量的本质及其分类的一种统计方法。**

**简单的说，就是根据相关性大小把变量分组，使得同组内的变量之间相关性较高，不同组的变量相关性较低。每组变量代表一个基本结构，这个基本结构称为因子。**

**因子分析是一种数据简化的技术，它通过研究众多变量之间的内部依赖关系，探求观测数据中的基本结构，并用少数几个假想变量来表示其基本的数据结构。这几个假想变量能够反映原来众多变量的主要信息。原始变量是可观测的显在变量，而假想变量是不可观测的潜在变量，称为因子。**

* **因子分析与其他分析的区别**

**与回归分析：因子分析中的因子是一个比较抽象的概念，而回归因子有非常明确的实际意义。**

**与主成分分析：主成分分析仅仅是变量变换，而因子分析需要构建因子模型。**

**主成分分析：原始变量的线性组合表示新的综合变量，即主成分。**

**因子分析：潜在的假想变量和随机影响变量的线性组合表示原始变量。**

* **因子分析模型**

**设个变量，如果表示为：**

****

****

**我们得到：**

**因子模型：**

****

**称为公共因子，是不可观测的变量，他们的系数称为因子载荷。是特殊因子，是不能被前个公共因子包含的部分。并且满足：**

**，即和不相关。**

****

****

****

**即互不相关，方差为1。**

**，即互不相关，方差不一定相等，**

* **用矩阵的表达方式**

****

1. ****
2. ****
3. ****

* **因子载荷的统计意义**

**因子载荷是第个变量与第个公共因子的相关系数模型为：**

****

****

**根据公共因子的模型性质，有（载荷矩阵中第行，第列的元素）反映了第个变量与第个公共因子的相关性。绝对值越大，相关的的密切程度就越高。**

* **因子载荷不是唯一的**

**设是一个的正交矩阵，令，，则模型可以表示为：**

****

**且满足因子模型的条件：**

****

* **变量共同度**

**定义：变量的共同度是因子载荷矩阵的第行的元素的平方和。记为**

**统计意义：**

**对两边求方差，得**

**即：**

**所有的公共因子和特殊因子对变量的贡献为1。**

**如果非常靠近1，非常小，则因子分析的效果好，从原变量空间到公共因子的转化性质好。**

* **公共因子方差贡献的统计意义**

**因子载荷矩阵中各列元素的平方和，称为对所有的的方差贡献和。衡量的相对重要性。**

* **因子载荷矩阵的估计方法**

1. **主成分分析法**

**设随机向量的均值为，协方差为，为的特征根，为对应的标准化特征向量，则：**

****

****

**上式给出的表达式是精确的，然而，它实际上是毫无价值的，因为我们的目的是要寻找用少数几个公共因子解释，故省略去后面的项的贡献，有：**

****

**其中**

**上式有一个假定，模型中的特殊因子是不重要的，因而从的分解中忽略了特殊因子的方差。**

**总结一下：**

****

* **为什么要旋转因子**

**因子分析的数学目的不仅仅要找出公共因子以及对变量进行分组，更重要的要知道每个公共因子的含义，以便进行进一步的分析。如果每个公共因子的含义不清，则不便于进行实际背景的解释。由于因子载荷阵是不惟一的，所以应该对因子载荷阵进行旋转。目的是使因子载荷阵的结构简化，使载荷矩阵每列或行的元素平方值向0和1两极分化。主要的正交旋转法有方差最大法和四次方最大法。**

* **因子得分的概念**

**前面我们主要解决了用公共因子的线性组合来表示一组观测变量的有关问题。如果我们要使用这些因子做其他的研究，比如把得到的因子作为自变量来做回归分析，对样本进行分类或评价，这就需要对公共因子进行测度，即得出公共因子的值。**

**因子分析的数学模型为：**

****

**因子得分函数：**

****

**可见，要求得每个因子得得分，必须求得分函数得系数，而由于，所以不能得到精确的得分，只能通过估计。**

* **因子分析的步骤**

1. **选择分析的变量**

**用定性分析和定量分析的方法选择变量，因子分析的前提条件是观测变量间有较强的相关性，因为如果变量之间无相关性或相关性较小的话，他们不会有共享因子,所以**

**原始变量间应该有较强的相关性。**

1. **计算所选原始变量的相关系数矩阵**

**相关系数矩阵描述了原始变量之间的相关关系。可以帮助判断原始变量之间是否存在相关关系，这对因子分析是非常重要的，因为如果所选变量之间无关系，做因子分析是不恰当的。并且相关系数矩阵是估计因子结构的基础。**

1. **提供公共因子**

**这一步要确定因子求解的方法和因子的个数。需要根据研究者的设计方案或有关的经验或知识事先确定。因子个数的确定可以根据因子方差的大小。只取方差大于1(或特征值大于1)的那些因子，因为方差小于1的因子其贡献可能很小；按照因子的累计方差贡献率来确定，一般认为要达到70％才能符合要求**

1. **因子旋转**

**通过坐标变换使每个原始变量在尽可能少的因子之间有密切的关系，这样因子解的实际意义更容易解释,并为每个潜在因子赋予有实际意义的名字。**

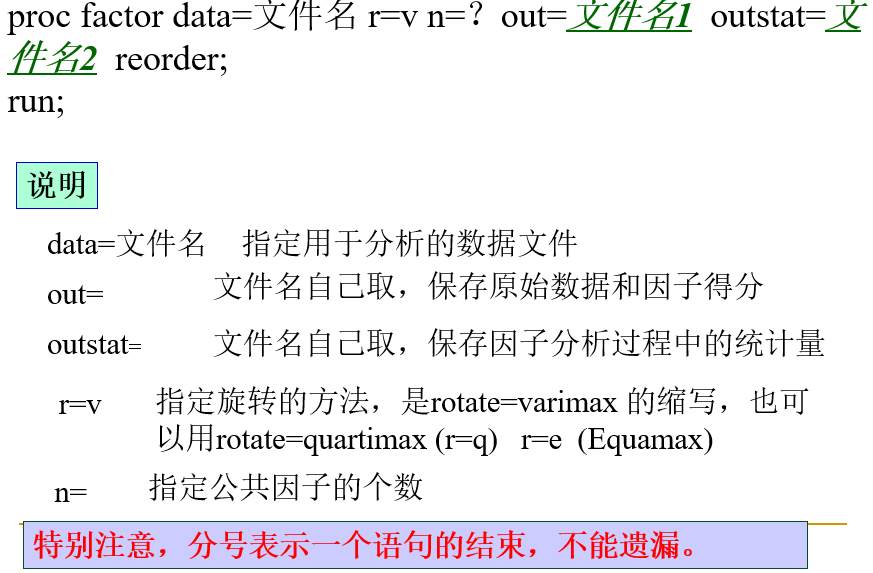
1. **计算因子得分**

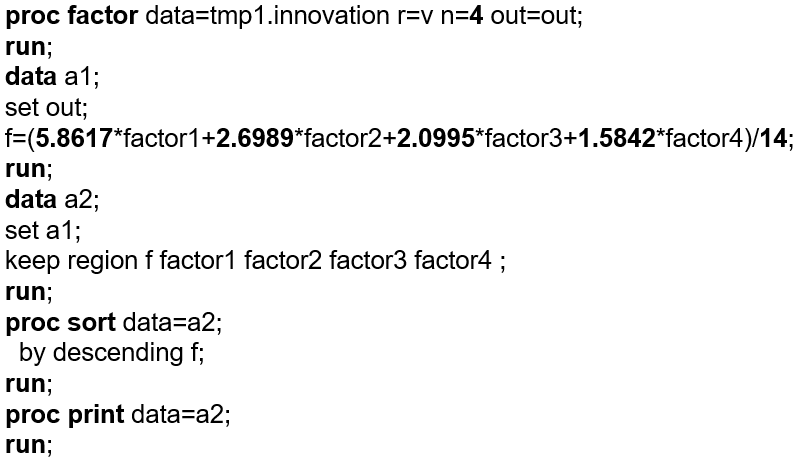
**求出各样本的因子得分，有了因子得分值，则可以在许多分析中使用这些因子，例如以因子的得分做聚类分析的变量，做回归分析中的回归因子。**

* **用因子分析方法进行综合评价**

**通过因子分析，取个公共因子，以每个公共因子的方差贡献率为权，构造综合评价函数，按值得大小对样品进行排序比较或分类。**

* **SAS程序**





**聚类分析**

* **什么是聚类分析**

**聚类分析是研究分类问题的一种多元统计方法，所谓类，就是指相似元素的集合。**

* **聚类分析的研究目的**

**把相似的东西归成类，根据相似的程度将研究目标进行分类。**

* **聚类分析的研究对象**

**R型分析—对变量进行分类**

**Q型分析—对样品进行分类**

* **相似性的测度**

1. **距离：测度样品之间的亲疏程度。将每一个样品看作维空间的一个点，并用某种度量测量点与点之间的距离，距离较近的归为一类，距离较远的点应术语不同的类。**
2. **相似系数：测度变量之间的亲疏程度**

* **明氏距离**

**设原始数据为**

**令表示样品与的距离：**

****

**特别地，当时，即为绝对值距离**

**当时，即为欧式距离**

**当时，即为切比雪夫距离**

* **明氏距离的两个缺点**

1. **明氏距离的数值与指标的量纲有关**
2. **没有考虑各个变量之间相关性的影响**

* **其他常用距离**
* **变量间相关系数的计算**

**变量和的相关系数：**

****

**夹角余弦：**

****

* **系统聚类法的基本思想**

**先将个样品各自看成一类，然后规定样品之间的“距离”和类与类之间的距离。选择距离最近的两类合并成一个新类，计算新类和其他类（各当前类）的距离，再将距离最近的两类合并。这样，每次合并减少一次类，直至所有的样品都归成一类为止。**

* **系统聚类法的基本步骤**

1. **计算个样品两两间的距离，记作**
2. **构造个类，每个类只包含一个样品**
3. **合并距离最近的两类为一新类**
4. **计算新类与各当前类的距离**
5. **重复步骤3、4，合并距离最近的两类为新类，直到所有的类并为一类为止**
6. **画聚类谱系图**
7. **决定类的个数和类**

* **系统聚类方法**

1. **最短距离法**
2. **最长距离法**
3. **中间距离法**
4. **重心法**
5. **类平均法**
6. **离差平方和法**

* **最短距离法**

**定义类与类之间的距离为两类最近样品的距离，即：**

****

**设类与合并成一个新类，记为，则与任一类的距离是：**

****

* **最长距离法**

**定义类与之间的距离为两类最远样品的距离，即：**

****

**设类与合并成一个新类，记为，则与任一类的距离是：**

****

* **中间距离法**

**定义类与类之间的距离既不采用两类之间最近的距离，也不采用两类之间最远的距离，而是采用介于两者之间的距离，故称为中间距离法。**

* **重心法**

**类与类之间的距离就考虑用重心之间的距离表示。设和的重心分别是和，则类和的距离为**

**设聚类到某一步，类与分别有样品和个**

**典型相关分析**

* **典型相关分析**

**当研究两个变量和之间的相关关系时，相关系数是最常用的度量。**

****

**如何进一步确定两组变量在整体上的相关程度呢？**

**通常情况下，为了研究两组变量的相关关系，可以用最原始的方法，分别计算两组变量之间的全部相关系数，一共个简单相关系数，这样既繁琐又不能抓住问题的本质。**

**如果能够采用类似于主成分的思想，分别找出两组变量的各自的某个线性组合，讨论线性组合之间的相关关系，则更简捷。**

* **典型相关分析的定义**

**研究两组变量之间相关关系的一种多元统计方法。它能够揭示两组变量之间的内在联系。**

* **典型相关分析的目的**

**识别并量化两组变量之间的联系，将两组变量相关关系的分析，转化为一组变量的线性组合与另一组变量线性组合之间的相关关系分析。**

* **典型相关分析的思想**

**首先分别在每组变量中找出第一对线性组合，使其具有最大相关性，**

****

**然后再在每组变量中找出第二对线性组合，使其分别与本组内的第一线性组合不相关，第二对本身具有次大的相关性。**

****

**和与和相互独立，但和相关。如此继续下去，直到进行到第步，两组变量的相关性被提取完为止。，可以得到组变量。**

* **典型相关的数学描述**

1. **想法**

**考虑两组变量的向量，其协方差阵为：**

****

**其中是第一组变量的协方差矩阵，是第二组变量的协方差矩阵，是和的协方差矩阵。**

**如果我们记两组变量的第一对线性组合为：**

****

**其中：**

****

****

**所以，典型相关就是求和，使得达到最大。**

1. **典型相关系数和典型变量的求法**

**在约束条件：**

****

**下，求和，使得达到最大。**

**第一对典型变量提取了原始变量和之间相关的主要部分，如果这部分还不能足以解释原始变量，可以在剩余的相关中再求出第二对典型变量和它们的典型相关系数。**

**在剩余的相关中再求出第二对典型变量和它们的典型相关系数。设第二对典型变量为：**

****

**在约束条件：**

****

**求使，达到最大的和。**

* **典型变量的性质**

1. **同一组变量的典型变量之间互不相关；**

**组的典型变量之间是相互独立的：**

**组的典型变量之间是相互独立的：**

1. **不同组变量的典型变量之间的相关性**

**不同组内一对典型变量之间的相关系数为：**

****

**同对相关系数为，不同对则为0.**

1. **原始变量与典型变量之间的相关系数**

**原始变量相关系数矩阵为：**

**典型变量系数矩阵为：**

****

**典型变量系数矩阵为：**

****

**这里直接放上公式：**

****

****

****

****

1. **各组原始变量被典型变量所解释的方差（典型冗余分析）**

**组原始变量被解释的方差比例：**

****

**组原始变量被解释的方差比例：**

****

**组原始变量被解释的方差比例：**

****

**组原始变量被解释的方差比例：**

****

* **从相关矩阵出发计算典型相关**

1. **为消除量纲影响，对数据先做标准化变换，然后再做典型相关分析。**
2. **显然，经标准化变换之后的协差阵就是相关系数矩阵，因而，也即通常应从相关矩阵除法进行典型相关分析。**

* **SAS程序总结**

**proc cancorr data=SAS数据集 out=SAS数据集 outstat=** **SAS数据集 all;**

**with 变量名;（列出被分析的两组变量中的第二组变量）**

**var 变量名；（列出被分析的两组变量中的第一组变量）**

**run;**

**（OUT= 包含原始数据和典型变量得分的数据集，OUTSTAT= 包含分析中用到的各种统计量）**

**proc cancorr data=SAS数据集 out= SAS数据集**

**outstat= SAS数据集 all vdep / vreg;**

**with 变量名;**

**var 变量名;**

**run;**

**（对于vdep来说：以Var变量为因变量，以with变量为自变量，进行多元回归分析。**

**对于vreg来说：以with变量为因变量，以Var变量为自变量，进行多元回归分析。）**