输出 3-1.1(续) **相关系数及 p 值**

Pearson 相关系数, N = 31 Prob > |r| under H0: Rho=0							
	x1	x2	x3	x4	x5	x6	x7
x1	1.00000	-0.23354	-0.30459	0.18875	-0.14157	-0.33787	-0.43292
年龄		0.2061	0.0957	0.3092	0.4475	0.0630	0.0150
x2	-0.23354	1.00000	-0.16275	0.14351	0.02270	0.18152	0.24938
体重	0.2061		0.3817	0.4412	0.9035	0.3284	0.1761
x3	-0.30459	-0.16275	1.00000	-0.86219	-0.34641	-0.39797	-0.23674
肺活量	0.0957	0.3817		<.0001	0.0563	0.0266	0.1997
x4	0.18875	0.14351	-0.86219	1.00000	0.40054	0.31365	0.22610
1.5英里跑的时间	0.3092	0.4412	<.0001		0.0256	0.0858	0.2213
x5	-0.14157	0.02270	-0.34641	0.40054	1.00000	0.31797	0.25750
休息时的脉搏	0.4475	0.9035	0.0563	0.0256		0.0813	0.1620
x6	-0.33787	0.18152	-0.39797	0.31365	0.31797	1.00000	0.92975
跑步时的脉搏	0.0630	0.3284	0.0266	0.0858	0.0813		<.0001
x7	-0.43292	0.24938	-0.23674	0.22610	0.25750	0.92975	1.00000
跑步时的最大脉搏	0.0150	0.1761	0.1997	0.2213	0.1620	<.0001	

输出 3-1.2 **偏协方差矩阵和偏相关系数及 p 值**

2 偏变量:	x1 x2
2 WITH 变量:	x3 x4
3 变量:	x5 x6 x7

偏协方差矩阵, 自由度 = 28				
		x5	x6	x7
x3	肺活量	-18.56248467	-27.82442337	-17.36162850
x4	1.5英里跑的时间	5.29333437	5.44150536	3.79244371

Pearson 偏相关系数, N = 31 Prob > |r| under H0: Partial Rho=0			
	x5	x6	x7
x3	-0.42979	-0.55166	-0.40538
肺活量	0.0200	0.0019	0.0291
x4	0.45042	0.39649	0.32543
1.5英里跑的时间	0.0142	0.0332	0.0850

输出说明:

在上述输出中,检验的 p 值列于其相应的相关系数或偏相关系数的下面,该检验将在下一章的§4.7 中介绍。

二、用交互式数据分析菜单系统作相关分析

选菜单过程如下:

解决方案(见图 3-1.1)⇒**分析▶**⇒**交互式数据分析**⇒**SASUSER**(见图 3-1.2)⇒**EXAMP342**⇒**打开**⇒**分析**(见图 3-1.3)⇒**多元(Y X)** ⇒选 **x1**,**x2**,**x3**,**x4**,**x5**,**x6**,**x7**(见图 3-1.4)⇒**Y**⇒**确定**⇒**表**(见图 3-1.5)⇒选**协方差矩阵**和**相关系数 P 值**(缺省时已选**单变量**和**相关系数矩阵**)

得到的结果与输出 3-1.1 相同(只是界面不同)。

图 3-1.1　SAS9.3 的主界面

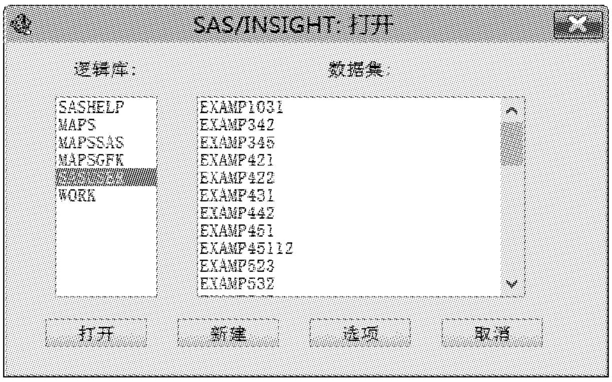


图 3-1.2　打开库成员对话框

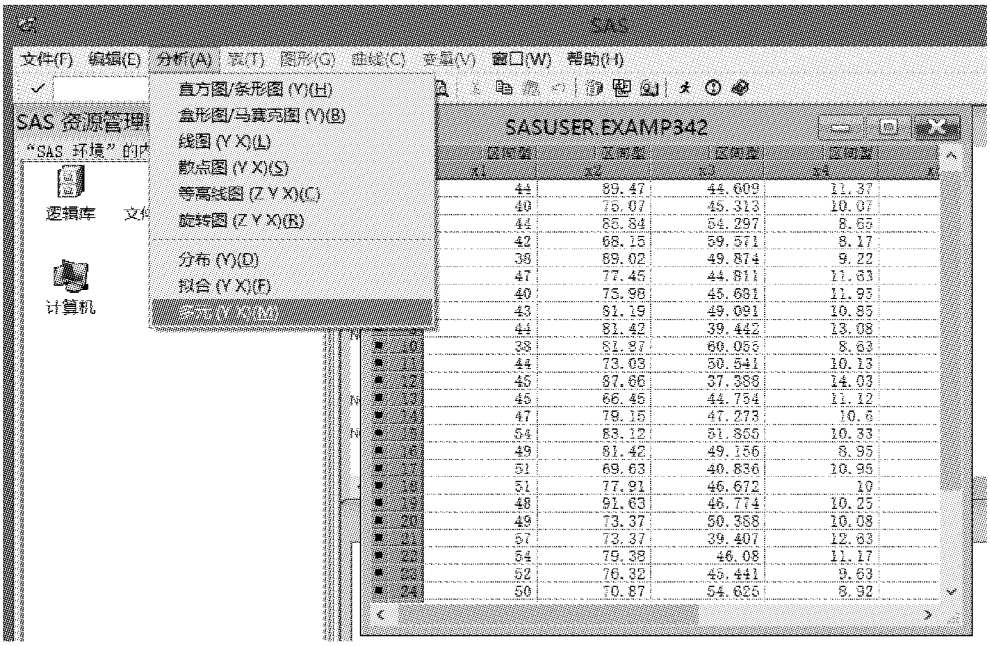


图 3-1.3　交互式数据分析菜单系统

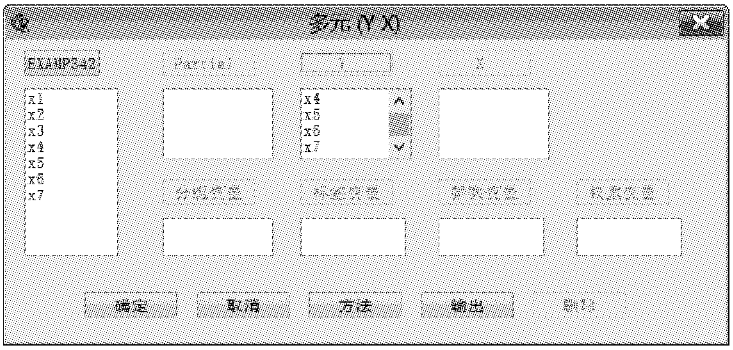


图 3-1.4　多元分析对话框

为获得输出 3-1.2 中的结果，在图 3-1.4 中，选菜单如下：

选 Y 框中的 **x1**，**x2**，**x3**，**x4**⇒**删除**⇒选 **x3**，**x4**→**X**；选 **x1**，**x2**→**Partial**⇒**确定**⇒**表**⇒选**相关系数矩阵**和**相关系数 P 值**

三、用交互式数据分析菜单系统作散点图矩阵

在图 3-1.3 中，选菜单如下：

散点图(Y X)⇒选 **x1**，**x2**，**x3**，**x4**，**x5**，**x6**，**x7**(见图 3-1.6)→**Y**；选 **x1**，**x2**，**x3**，**x4**，**x5**，**x6**，**x7**→**X**⇒**确定**

得到如图 3-1.7 所示的散点图矩阵。

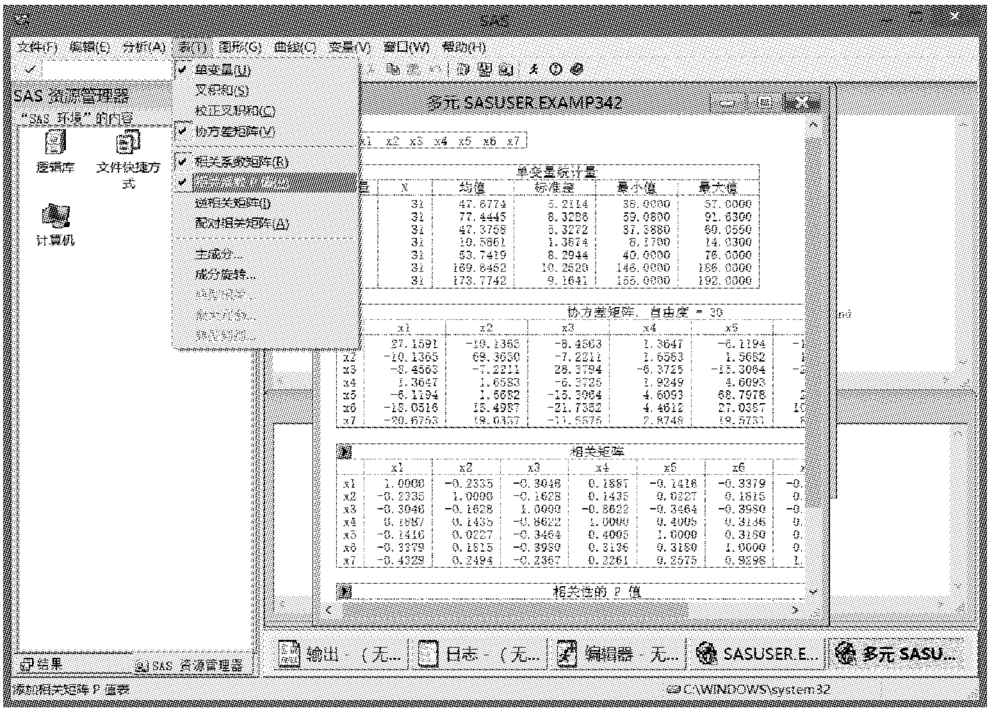


图 3-1.5　结果输出菜单

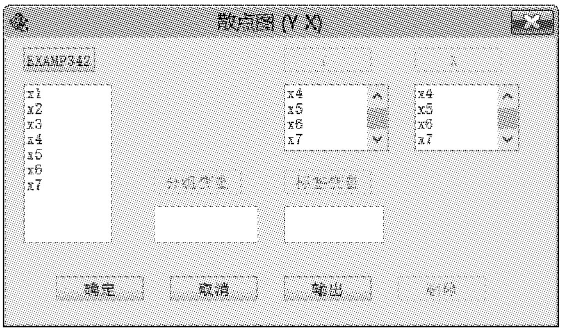


图 3-1.6　散点图对话框

在图 3-1.7 里，变量名及其观测数据的最小值和最大值写在对角线的格子中。在同一行的格子中，该行所含的(对角线上的)变量为该行所有格子的纵轴变量；在同一列的格子中，该列所含的变量为该列所有格子的横轴变量。观测散点图矩阵，可以得到关于异常数据以及变量间关系的直观印象。

需要指出以下几点：

(1) 在图 3-1.7 中，上三角和下三角格子中的散点图是对称的。

(2) 可将格子中的小散点图放大，在散点图矩阵所在的界面中选菜单如下：**编辑**⇒**窗口 ▶**⇒**工具**，在弹出的工具窗中选“放大镜图标”，鼠标即变为放大镜形状。大的放大镜(位于中间时出现)点击时可使散点图慢慢地放大，小的放大镜(位于旁边时出现)点击时可使散点图慢慢地缩小，直至回到图 3-1.7 中的最初状态。在工具窗中，还可选择点的颜色和形状等。

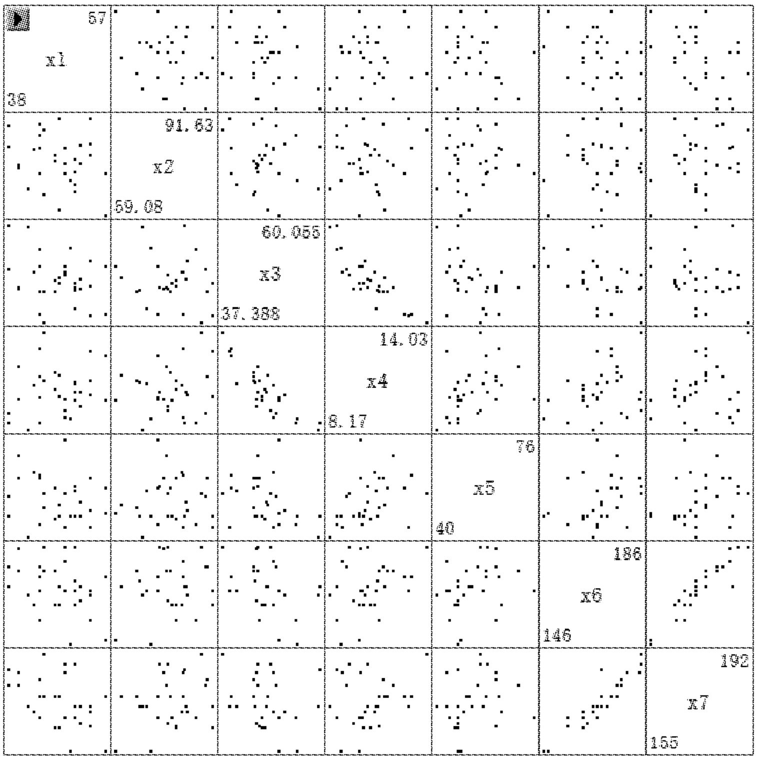


图 3-1.7 散点图矩阵

(3) 若只需画两个变量的散点图，则在图 3-1.6 里，可将横轴变量置于 X 框中，将纵轴变量置于 Y 框中，然后点击确定即可。

(4) 在图 3-1.3 里，也可以先在数据窗中选中 x1 至 x7 这七个变量(可点击左上角的“7”)，然后再点击<u>**散点图(Y X)**</u>，同样可得到图 3-1.7。

四、用交互式数据分析菜单系统作旋转图

旋转图是一个三维图，我们不妨对 x1，x2，x3 作旋转图。在图 3-1.3 中，选菜单如下：

<u>**旋转图(Z Y X)**</u>⇒选 <u>**x3**</u>(见图 3-1.8)→<u>**Z**</u>；选 <u>**x2**</u>→<u>**Y**</u>；选 <u>**x1**</u>→<u>**X**</u>⇒<u>**确定**</u>

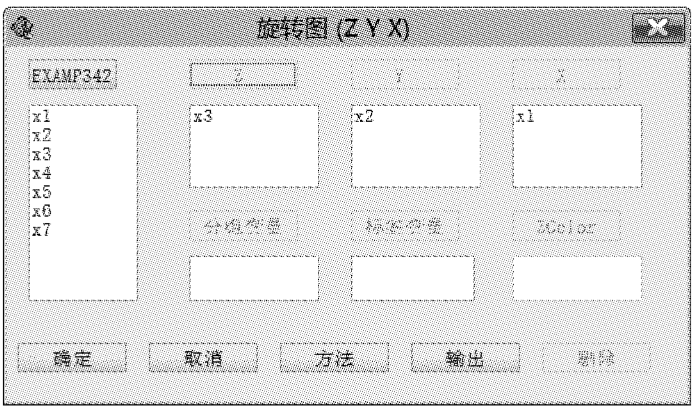


图 3-1.8 旋转图对话框

随即得到旋转图。在工具窗中选择“手图标”或将光标移至图的角上使它变为“手”，然后按住鼠标器左键拖动鼠标器，图就按拖动方向旋转。此外，点击旋转图左下角的“三角符”或在旋转图框中任意处点击右键，弹出一上托菜单，选“立方块”可使旋转看起来更具立体感，如图 3-1.9 所示。

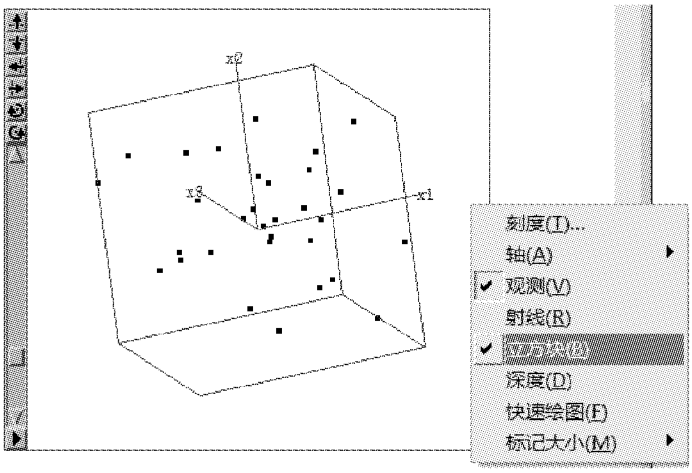


图 3-1.9 旋转图及其上托菜单

观测旋转图，可以动态、持续地进行直至获得足够的信息，这样的图可以使我们从三维角度观测变量之间的关系，并可识别出在平面散点图中发现不了的异常值。如果三个变量之间存在线性关系，则散点将分布在一个平面上，旋转到某一角度所有的散点会呈现在一直线上，此时提醒我们可能存在“多余”的变量。

附录 3-2 §3.2 中若干性质的数学证明

一、性质(1)的证明

由(2.5.3)式知，标准正态变量 u_j 的特征函数

$$\varphi_{u_j}(t)=\mathrm{e}^{-t^2/2},\quad j=1,2,\cdots,q$$

从而由 $u_1,u_2,\cdots,u_q$ 相互独立知，$\boldsymbol{u}=(u_1,u_2,\cdots,u_q)'$ 的特征函数

$$\begin{aligned}\varphi_{\boldsymbol{u}}(\boldsymbol{t})&=E(\mathrm{e}^{i\boldsymbol{t}'\boldsymbol{u}})=E\left[\exp\left(i\sum_{j=1}^{q}t_ju_j\right)\right]=E\left(\prod_{j=1}^{q}\mathrm{e}^{it_ju_j}\right)\\&=\prod_{j=1}^{q}E(\mathrm{e}^{it_ju_j})=\prod_{j=1}^{q}\varphi_{u_j}(t_j)=\prod_{j=1}^{q}\mathrm{e}^{-t_j^2/2}\\&=\exp\left(-\frac{1}{2}\sum_{j=1}^{q}t_j^2\right)=\exp\left(-\frac{1}{2}\boldsymbol{t}'\boldsymbol{t}\right)\end{aligned}$$

所以，$\boldsymbol{x}$ 的特征函数

$$\begin{aligned}\varphi_{\boldsymbol{x}}(\boldsymbol{t})&=E(\mathrm{e}^{i\boldsymbol{t}'\boldsymbol{x}})=E[\mathrm{e}^{i\boldsymbol{t}'(\boldsymbol{\mu}+\boldsymbol{A}\boldsymbol{u})}]=\mathrm{e}^{i\boldsymbol{t}'\boldsymbol{\mu}}E[\mathrm{e}^{i(\boldsymbol{A}'\boldsymbol{t})'\boldsymbol{u}}]=\mathrm{e}^{i\boldsymbol{t}'\boldsymbol{\mu}}\varphi_{\boldsymbol{u}}(\boldsymbol{A}'\boldsymbol{t})\\&=\exp\left[i\boldsymbol{t}'\boldsymbol{\mu}-\frac{1}{2}(\boldsymbol{A}'\boldsymbol{t})'(\boldsymbol{A}'\boldsymbol{t})\right]=\exp\left(i\boldsymbol{t}'\boldsymbol{\mu}-\frac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\boldsymbol{t}\right)\end{aligned}$$

前者。

5.当 n 很大且 n 相对于 p 也很大时，对总体均值向量 $\boldsymbol{\mu}$（或 $\boldsymbol{\mu}$ 的线性组合）的推断可不依赖于总体正态性的假定。在多元正态假定下的均值推断结论中，只需将其中的分位点 $T_\alpha^2(p,n-1)$ 替换为 $\chi_\alpha^2(p)$，或 $T_\alpha(p,n-1)$ 替换为 $\sqrt{\chi_\alpha^2(p)}$，或 $t_{\alpha/2k}(n-1)$ 替换为 $u_{\alpha/2k}$，相应推断结论（近似）成立。

6.在实际应用中，一旦 $H_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$ 被拒绝了，则可以考虑对这两个均值向量的每一对均值分量是否相等再分别进一步检验，以判断是否有分量及（若有）具体是哪些分量对拒绝 $H_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2$ 起了较大作用，这样做常常是有益的。这对单总体和多总体的情形也都同样是适用的。

7.对于成对试验的数据，两个样本一般并不相互独立。数据的成对出现避免了作为抽样误差来源之一的两个样本个体之间的差异，从而减少了抽样误差，以致往往可以得到比独立样本方法更精确的统计推断结论。

8.单总体轮廓分析是对轮廓的水平性进行检验。两总体轮廓分析依次检验两轮廓的平行性、在平行条件下的两轮廓的重合性、在重合条件下的共同轮廓的水平性，只有在前面的检验不拒绝原假设后方可进行后面的检验。对比矩阵 $\boldsymbol{C}$ 的选择一般不是唯一的，但检验结果不会因 $\boldsymbol{C}$ 的不同选择而改变。

9.多元方差分析是一元方差分析的直接推广，但多元方差分析中的检验统计量并不唯一，而是可以有多个。

10.偏相关系数与简单相关系数的检验方法是完全类似的，主要区别是检验统计量分布的自由度有所不同。

11.当对 k 个多元正态总体使用联合协方差矩阵或对其均值向量进行比较检验时，常常可考虑先对 k 个总体协方差矩阵的相等性进行博克斯的 M 检验。$k=2$ 时的 M 检验用于对两总体协方差矩阵的相等性进行检验。

附录 4-1 SAS 的应用

一、用交互式数据分析菜单系统作置信椭圆和预测椭圆

在图 3-1.5 中，选：

曲线⇒散点图置信椭圆▶⇒ 均值:95%，预测:95%

即可获得每两个变量均值的 0.95 置信椭圆和新观测值的 0.95 预测椭圆，如图 4-1.1 所示。

预测椭圆的说明：

设 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 是取自总体 $N_p(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 的一个样本，$\boldsymbol{x}_0$ 是来自该总体的一个新 p 维观测值，且独立于样本，则 $\boldsymbol{x}_0$ 的置信度是 $1-\alpha$ 的预测区域为[①]

---

① 由于 $\sqrt{\frac{n}{n+1}}(\bar{\boldsymbol{x}}-\boldsymbol{x}_0)\sim N_p(\boldsymbol{0},\boldsymbol{\Sigma})$，$(n-1)\boldsymbol{S}\sim W_p(n-1,\boldsymbol{\Sigma})$，且 $\bar{\boldsymbol{x}}-\boldsymbol{x}_0$ 和 $\boldsymbol{S}$ 相互独立，于是

$$T^2=\frac{n}{n+1}(\bar{\boldsymbol{x}}-\boldsymbol{x}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{x}_0)$$
$$=(n-1)\left[\sqrt{\frac{n}{n+1}}(\bar{\boldsymbol{x}}-\boldsymbol{x}_0)\right]'[(n-1)\boldsymbol{S}]^{-1}\left[\sqrt{\frac{n}{n+1}}(\bar{\boldsymbol{x}}-\boldsymbol{x}_0)\right]\sim T^2(p,n-1)$$

故

$$P\left[\frac{n}{n+1}(\bar{\boldsymbol{x}}-\boldsymbol{x}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{x}_0)\leqslant T_\alpha^2(p,n-1)\right]=P\left[\frac{n-p}{p(n-1)}T^2\leqslant F_\alpha(p,n-p)\right]=1-\alpha$$

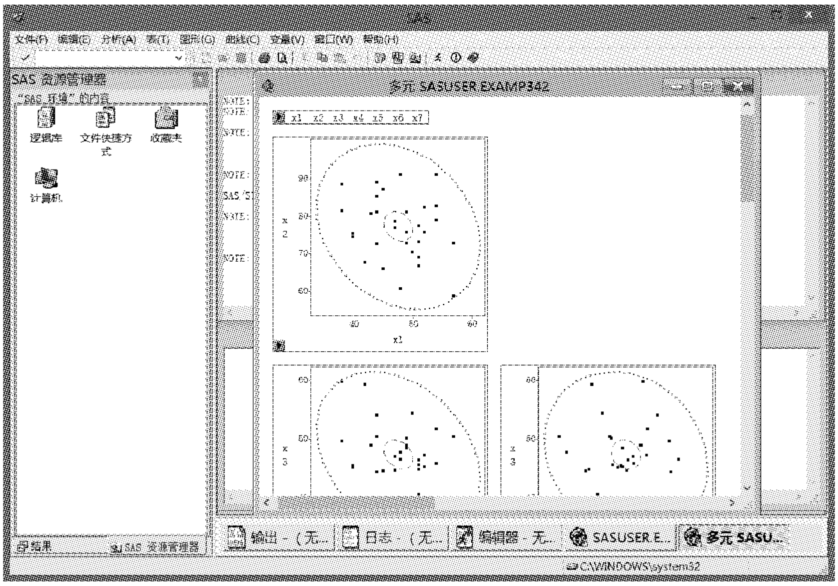


图 4-1.1　两两变量之间的散点图、95%置信椭圆和 95%预测椭圆

$$\left\{\boldsymbol{x}_0:\frac{n}{n+1}(\bar{\boldsymbol{x}}-\boldsymbol{x}_0)'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{x}_0)\leqslant T_\alpha^2(p,n-1)\right\} \tag{4-1.1}$$

当 $p=1$ 时，该预测区域是一个区间；当 $p=2$ 时，它是一个椭圆，正如图 4-1.1 所示；当 $p=3$ 时，它是一个椭球；当 $p>3$ 时，它是一个超椭球。正像图 4-1.1 中所看到的，在同样 $1-\alpha$ 置信度下，预测区域(4-1.1)要比置信区域(4.2.14)大得多。关于这一点，我们也可在 $p=2$ 的情形下从图形的直观角度去理解。总体均值 $\boldsymbol{\mu}$ 是平面上的一个固定点，因而在同样置信度下只需相对较小的(椭圆)圈就能套住它；而 $\boldsymbol{x}_0$ 是平面上的一个随机点，其活动范围基本与样本点的散布范围相近，要在同样置信度下套住这样一个飘忽不定的点自然需要较大的(椭圆)圈才行。图 4-1.1 中，预测椭圆大约能套住 $100(1-\alpha)\%$ 的样本点。

二、用编程对例 4.5.1 作多元方差分析

SAS9.3 的菜单系统只能作一元方差分析，要作多元方差分析就只有编程，可以使用 ANOVA(或 GLM)过程并加上 MANOVA 语句。对于一个和两个总体的均值向量检验，SAS9.3 并没有给出相应的过程步，可以使用 IML 过程直接编写矩阵运算的相对较原始程序。但对于两个总体的情形，较为简便的方法是，将两总体的假设检验(4.3.1)看成是多元方差分析(4.5.1)的一个特例，从而使用 ANOVA(或 GLM)过程，其检验结果是完全一致的。

SAS 程序：

```
proc anova data=sasuser.examp451;
    class g;
    model x1-x4=g;
    manova h=g;
run;
```

程序说明：

“proc anova”是一个对均衡数据的方差分析过程(对非均衡数据的方差分析可使用一般

输出 5-1.9 的说明：

(1) 全(或总)样本标准化判别(或典型)系数。在对各原始变量作标准化变换后再进行判别分析，每一变量的标准化变换是依该变量的全样本(即各组样品混合在一起)数据进行的。标准化判别系数可用来比较各变量对判别函数的重要性。

(2) 联合组内(或合并类内)标准化的判别(或典型)系数。每一变量标准化变换的分母采用的是该变量的联合组内方差(从 $\boldsymbol{S}_p$ 的对角线上获得)的算术平方根，其分子是以该变量的全样本均值为中心作中心化。该标准化判别系数同样可用来比较各变量对判别函数的重要性。

(3) 原始判别(或典型)系数。从原始变量出发，系数向量 $\boldsymbol{t}_i$ 满足 $\boldsymbol{t}_i'\boldsymbol{S}_p\boldsymbol{t}_i=1$，$i=1,2$。

(4) 中心化的费希尔判别函数(或典型变量)的组(或类)均值。

输出 5-1.10　　用费希尔判别函数得分作出的观测散点图

见图 5.4.2。

三、用交互式数据分析菜单系统对例 5.4.2 作费希尔判别

在 SAS/INSIGHT 环境下打开数据集 SASUSER.EXAMP542，选菜单如下：

多元(Y X)(参考图 3-1.3)⇒选 **x1**，**x2**，**x3**，**x4**(见图 5-1.1)→**X**；选 **g**→**Y**⇒**输出**⇒选**典型判别分析**，**典型判别选项**(见图 5-1.2)⇒选**典型相关**，**标准得分系数**，**原始得分系数**(见图 5-1.3)⇒**确定**⇒**确定**⇒**确定**

可得到输出 5-1.8 以及输出 5-1.9 中的全样本标准化判别系数和原始判别系数的结果。

图 5-1.1　多元分析对话框

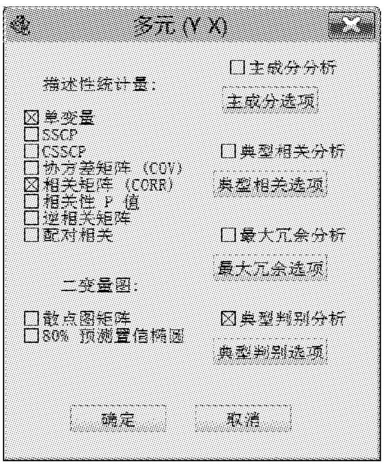


图 5-1.2　多元分析选项对话框

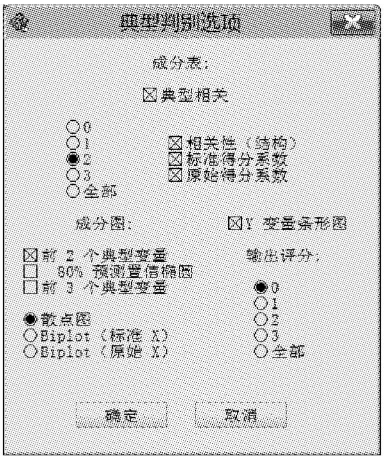


图 5-1.3　典型判别选项对话框

四、对例 5.5.1 编程作逐步判别

SAS 程序：

```
proc stepdisc data=sasuser.examp542;
    class g;
    var x1-x4;
run;
```

程序说明：

“proc stepdisc”是一个逐步判别过程，选择方法选项：method＝fw(或 forward，前进法)，bw(或 backward，后退法)及 sw(或 stepwise，逐步判别法)，缺省值为 sw。选择准则选项：sle(或 slentry)$=\alpha$ 是指定选入变量的显著性水平，$0\leqslant\alpha\leqslant 1$，缺省值为 0.15；sls(或 slstay)$=\alpha$ 是指定保留变量的显著性水平，$0\leqslant\alpha\leqslant 1$，缺省值为 0.15。class 语句规定了分组变量是 g。

SAS 输出：

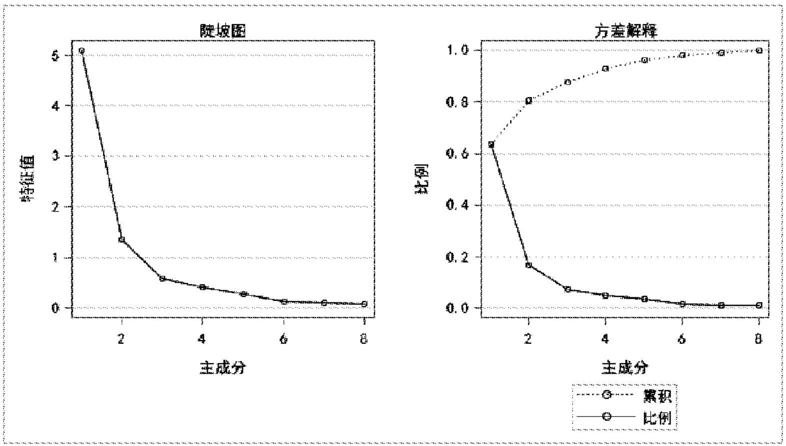
输出 5-1.11　　　　选择变量的方法为逐步判别

选择变量的方法为 STEPWISE			
总样本大小	150	分析中的变量	4
分类水平	3	将包括的变量	0
		输入变量的显著性水平	0.15
		保留变量的显著性水平	0.15

读取的观测数	150
使用的观测数	150

分类水平信息				
g	变量名称	频数	权重	比例
1	_1	50	50.0000	0.333333
2	_2	50	50.0000	0.333333
3	_3	50	50.0000	0.333333

输出 7-1.1(续)　　　　**陡坡图和(累计)贡献率图**



输出 7-1.1(续)的说明：

位于左边的是**陡坡图**(或称**碎石图**,scree plot),右边的是贡献率及累计贡献率图,这两张图是能帮助我们确定主成分个数的视觉工具。在陡坡图中,从第三个主成分起,线段开始变得平坦,意味着从λ_3起特征值的变化开始明显变小,单从该图来看,似乎倾向于只取前面两个主成分。但主成分个数的最终确定还得要进一步看累计贡献率的大小以及主成分能否得到有效的解释(如果所作的分析需要这种解释)。

输出 7-1.2　　　　**按第一主成分排序的全国 31 个地区**

略,结果参见表 7.3.5。

输出 7-1.3　　　　**按第二主成分排序的全国 31 个地区**

略,结果参见表 7.3.6。

输出 7-1.4　　　　**按第一和第二主成分输出的散点图**

略,参见图 7.3.3。

二、使用交互式数据分析菜单系统作主成分分析

在 SAS/INSIGHT 环境下打开数据集 SASUSER.EXAMP633,选菜单过程如下：

分析(参考图 3-1.3)⇒**多元(Y X)**⇒在变量框中,选 **x1**,**x2**,**x3**,**x4**,**x5**,**x6**,**x7**,**x8**(见图 7-1.1)⇒**Y**⇒**输出**⇒选**主成分分析**(参考图 5-1.2),点击**主成分选项**⇒在图 7-1.2 中作图中的

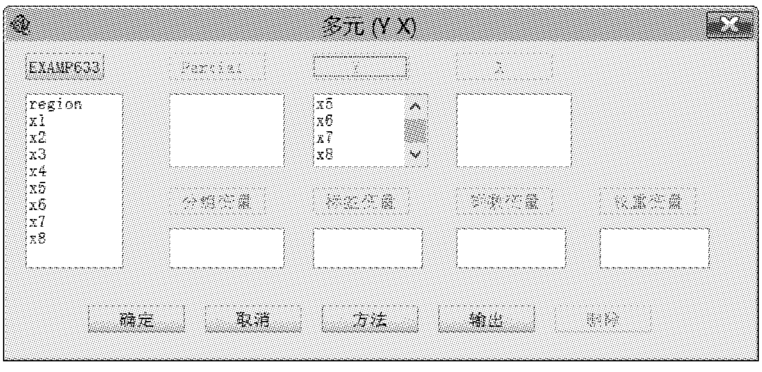


图 7-1.1　多元分析对话框

图 7-1.2　主成分选项对话框

选择(主成分个数缺省时为“自动”选项,此时只输出特征值大于 1 的主成分)⇒<u>**确定**</u>⇒<u>**确定**</u>⇒<u>**确定**</u>

随即得到相应的一些结果(略)。

三、使用交互式数据分析菜单系统制作图 7.3.3

继续以上的操作,选菜单过程如下:

<u>**分析**</u>⇒<u>**散点图(Y X)**</u>⇒选 <u>**PCR1**</u>→<u>**X**</u>;<u>**PCR2**</u>→<u>**Y**</u>;<u>**region**</u>→<u>**标签变量**</u>(见图 7-1.3)⇒<u>**确定**</u>

随即出现如图 7-1.4 所示的散点图,接下来我们对该图进行设置、调整,使之符合我们的要求。

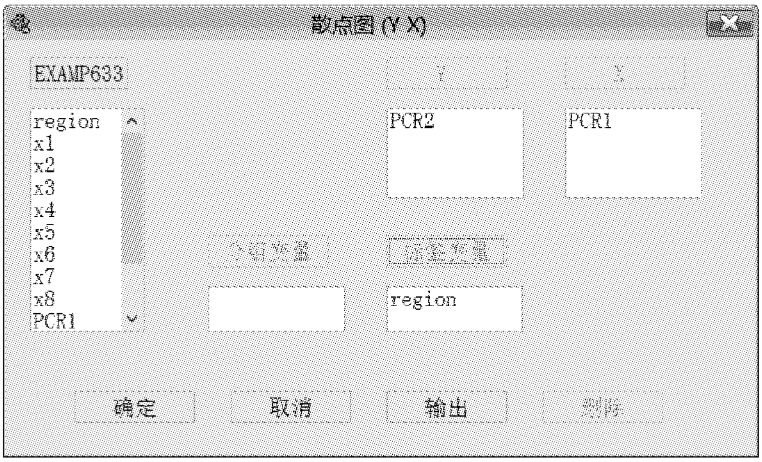


图 7-1.3　散点图对话框

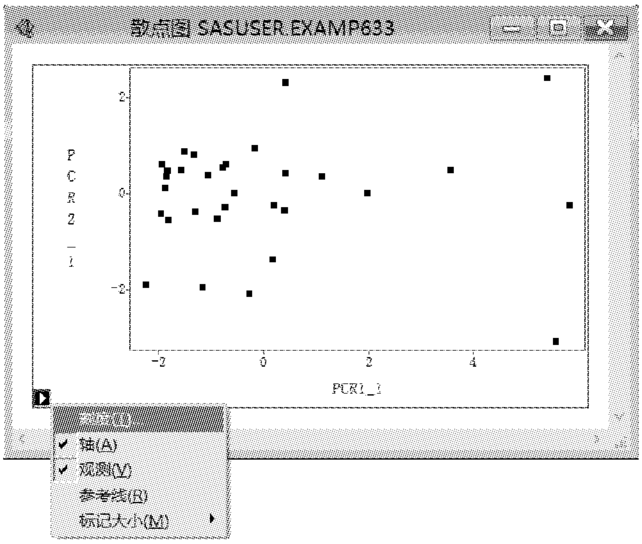


图 7-1.4　两个主成分的散点图

在图 7-1.5 的数据窗口中点击左上角的 **31**，以使所有观测都选中⇒将鼠标移至观测号中，点击右键，选**在图中加标签**⇒在图 7-1.4 中点击左下角的小三角，出现上托菜单⇒**刻度…**⇒作图 7-1.6 中的选择⇒**确定**⇒作图 7-1.7 中的选择⇒**确定**⇒调整好散点图：拉大图外围方框，并使得长和宽如图 7.3.3 中的比例；图 7-1.4 中点击**标记大小 ▶**，选择 6 号字；下拉横坐标，左拉纵坐标；图 7-1.4 中点击**参考线**，产生图 7.3.3（图中的三个圆圈和若干虚线是使用画图工具加上去的，只是用来说明聚类的）。

12	列名型	区间型	区间型	区间型	区间型	区间型	区间型	区间型	区间型	区间型	区间型
31	region	x1	x2	x3	x4	x5	x6	x7	x8	PCR1	PCR2
1	北京	2959.19	730.79	749.41	513.34	467.87	1141.82	478.42	457.64	5.4264	2.4666
2	天津	2459.77	495.47	697.33	302.87	284.19	735.97	570.84	305.08	2.0064	0.0449
3	河北	1495.63	515.90	362.37	285.32	272.95	540.58	364.91	188.63	-0.7696	0.5805
4	山西	1406.33	477.77	290.15	208.57	201.50	414.72	281.84	212.10	-1.8487	0.4044
5	内蒙古	1303.97	524.29	254.83	192.17	249.81	463.09	287.87	192.96	-1.8267	0.5099
			[illegible]	246.91	279.81	239.18	445.20	330.24	163.86	-1.3136	0.8448
			[illegible]	200.49	218.36	220.69	459.62	360.48	147.76	-1.8598	0.1514
			[illegible]	211.88	277.11	224.65	376.82	317.61	152.85	-1.9276	0.6367
			[illegible]	893.37	346.93	527.00	1034.98	720.33	462.03	5.8666	-0.1956
			[illegible]	572.40	211.92	302.09	585.23	429.77	252.54	0.4072	-0.3120

图 7-1.5　sasuser.examp633 数据集

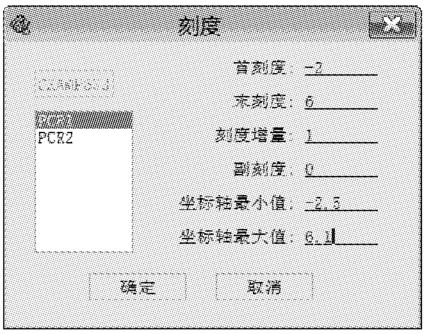


图 7-1.6　PCR1 的刻度对话框

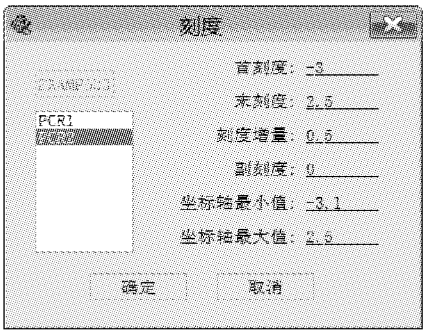


图 7-1.7　PCR2 的刻度对话框

二、使用交互式数据分析菜单系统

该菜单系统可以在主成分分析的基础上进一步进行(参数估计为)主成分法的正交因子分析,只能使用最大方差旋转法对两个因子进行旋转。

在 SAS/INSIGHT 环境下打开数据集 SASUSER.EXAMP633,选菜单过程如下:

多元(Y X)(参考图 3-1.3)⇒在变量框中,选 **x1**,**x2**,**x3**,**x4**,**x5**,**x6**,**x7**,**x8**(参考图 7-1.1)⇒**Y**⇒**输出**⇒选**主成分分析**,点击**主成分选项**(参考图 5-1.2)⇒在主成分选项对话框(参考图 7-1.2)的"成分表"一栏中,选择 **2**;在"输出成分"一栏中,选**成分旋转**,点击**旋转选项**⇒作图 8-1.1 中的选择⇒**确定**⇒**确定**⇒**确定**⇒**确定**
随即得到相应的一些结果(略)。

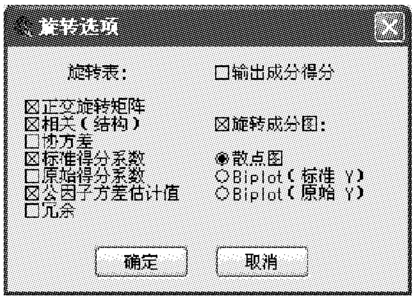


图 8-1.1　主成分选项对话框

习　题

8.1　试证(8.3.2)式。

[提示:$(\boldsymbol{S}-\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}'-\hat{\boldsymbol{D}})$的元素平方和≤$(\boldsymbol{S}-\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}')$的元素平方和,

$$\boldsymbol{S}-\hat{\boldsymbol{A}}\hat{\boldsymbol{A}}'=\hat{\lambda}_{m+1}\hat{\boldsymbol{t}}_{m+1}\hat{\boldsymbol{t}}'_{m+1}+\cdots+\hat{\lambda}_p\hat{\boldsymbol{t}}_p\hat{\boldsymbol{t}}'_p=(\hat{\boldsymbol{t}}_{m+1},\cdots,\hat{\boldsymbol{t}}_p)\begin{pmatrix}\hat{\lambda}_{m+1} & & 0\\ & \ddots & \\ 0 & & \hat{\lambda}_p\end{pmatrix}\begin{pmatrix}\hat{\boldsymbol{t}}'_{m+1}\\ \vdots\\ \hat{\boldsymbol{t}}'_p\end{pmatrix}$$

利用(1.6.10)式]

8.2　比较(8.5.15)与(8.5.14)两式,试证$(\boldsymbol{A}'\boldsymbol{D}^{-1}\boldsymbol{A})^{-1}-(\boldsymbol{I}+\boldsymbol{A}'\boldsymbol{D}^{-1}\boldsymbol{A})^{-1}$是正定矩阵。

8.3　在例 8.1.1 中,十项全能运动得分的样本相关矩阵为

$$\hat{\boldsymbol{R}}=\begin{pmatrix}1.00 \\ 0.59 & 1.00 \\ 0.35 & 0.42 & 1.00 \\ 0.34 & 0.51 & 0.38 & 1.00 \\ 0.63 & 0.49 & 0.19 & 0.29 & 1.00 \\ 0.40 & 0.52 & 0.36 & 0.46 & 0.34 & 1.00 \\ 0.28 & 0.31 & 0.73 & 0.27 & 0.17 & 0.32 & 1.00 \\ 0.20 & 0.36 & 0.24 & 0.39 & 0.23 & 0.33 & 0.24 & 1.00 \\ 0.11 & 0.21 & 0.44 & 0.17 & 0.13 & 0.18 & 0.34 & 0.24 & 1.00 \\ -0.07 & 0.09 & -0.08 & 0.18 & 0.39 & 0.00 & -0.02 & 0.17 & -0.00 & 1.00\end{pmatrix}$$

略，参见图 9.5.1。

2.原始数据形式的数据集

SAS 程序：

```
proc corresp data=sasuser.examp921 out=results rp cp short;
    tables row,column;
    weight f;
proc plot data=results;
    plot dim1 * dim2=_name_ /box vspace=6 hspace=13 haxis=-.30 to .30 by .15 vaxis
=-.30 to .30 by .15;
run;
```

程序说明：

数据集 sasuser.examp921 中的数据是原始数据形式，须使用 tables 语句。Tables 语句规定用于构造列联表的行和列分类变量，该语句的第一个变量规定为行变量，逗号后面的第二个变量规定为列变量。用 tables 语句时不能使用 id 语句。Weight 语句用来读入类别组合的频数。

二、用交互式数据分析菜单系统生成马赛克图

在 SAS/INSHIGT 环境下打开数据集 SASUSER.EXAMP921，见图 9-1.1。选菜单过程如下：

在图 9-1.1 的界面中，**分析**⇒**盒形图/马赛克图(Y)(B)**⇒选 **row**→**X**(见图 9-1.2)；**column**→**Y**；**f**→**频数变量**⇒**输出**⇒作图 9-1.3 中的选择⇒**确定**⇒**确定**

生成的行轮廓及行密度图见图 9.2.1 中的(1)。

在图 9-1.1 的界面中，**分析**⇒**盒形图/马赛克图(Y)(B)**⇒选 **column**→**X**(参考图 9-1.2)；**row**→**Y**；**f**→**频数变量**⇒**输出**⇒在出现的图 9-1.3 中，再选择**垂直的 Y 轴**⇒**确定**⇒**确定**

随即出现的列轮廓及列密度图见图 9.2.1 中的(2)。

图 9-1.1 交互式数据分析菜单系统

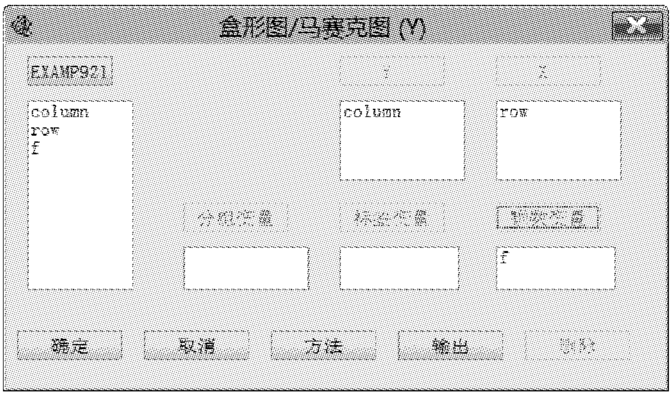


图 9-1.2　盒形图/马赛克图对话框

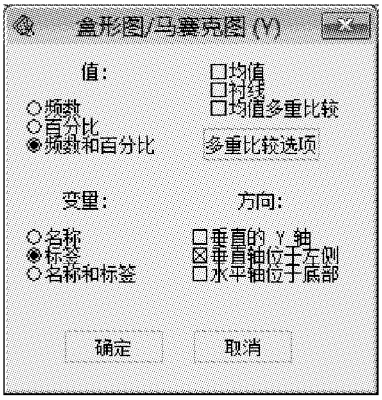


图 9-1.3　盒形图/马赛克图输出选项对话框

附录 9-2　若干推导

一、(9.5.4)式的推导

证明　令 p 维向量 $\boldsymbol{e}_i=(0,\cdots,0,1,0,\cdots,0)'$，这里 1 位于第 i 个位置上，p 维向量 $\boldsymbol{e}_j=(0,\cdots,0,1,0,\cdots,0)'$，这里 1 位于第 j 个位置上，则在对应分析图上第 i 个行点(x_{i1},x_{i2})与第 j 个行点(x_{j1},x_{j2})之间的平方欧氏距离为

$$d_{ij}^2(r)=(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2=(x_{i1}-x_{j1},x_{i2}-x_{j2})\begin{pmatrix}x_{i1}-x_{j1}\\x_{i2}-x_{j2}\end{pmatrix}$$

$$=(\boldsymbol{e}_i'\boldsymbol{X}_1-\boldsymbol{e}_j'\boldsymbol{X}_1)(\boldsymbol{e}_i'\boldsymbol{X}_1-\boldsymbol{e}_j'\boldsymbol{X}_1)'=(\boldsymbol{e}_i-\boldsymbol{e}_j)'\boldsymbol{X}_1\boldsymbol{X}_1'(\boldsymbol{e}_i-\boldsymbol{e}_j)$$

其中 $\boldsymbol{X}_1=(\lambda_1\boldsymbol{D}_r^{-1}\boldsymbol{a}_1,\lambda_2\boldsymbol{D}_r^{-1}\boldsymbol{a}_2)=\begin{pmatrix}x_{11}&x_{12}\\x_{21}&x_{22}\\\vdots&\vdots\\x_{p1}&x_{p2}\end{pmatrix}$，而

输出 10-1.4　　**典型结构**

典型结构

VAR 变量 及其典型变量之间的相关性				
		u1	u2	u3
x1	体重	0.6206	-0.7724	-0.1350
x2	腰围	0.9254	-0.3777	-0.0310
x3	脉搏	-0.3328	0.0415	0.9421

WITH 变量 及其典型变量之间的相关性				
		v1	v2	v3
y1	引体向上	-0.7276	0.2370	-0.6438
y2	起坐次数	-0.8177	0.5730	0.0544
y3	跳跃次数	-0.1622	0.9586	-0.2339

VAR 变量 和 WITH 变量 的典型变量之间的相关性				
		v1	v2	v3
x1	体重	0.4938	-0.1549	-0.0098
x2	腰围	0.7363	-0.0757	-0.0022
x3	脉搏	-0.2648	0.0083	0.0684

WITH 变量 和 VAR 变量 的典型变量之间的相关性				
		u1	u2	u3
y1	引体向上	-0.5789	0.0475	-0.0467
y2	起坐次数	-0.6506	0.1149	0.0040
y3	跳跃次数	-0.1290	0.1923	-0.0170

二、使用交互式数据分析菜单系统

在 SAS/INSIGHT 环境下打开数据集 SASUSER.EXAMP1031，选菜单过程如下：

在图 3-1.3 中，**多元(Y X)** ⇒在变量框中(见图 10-1.1)，选**x1**，**x2**，**x3**→**X**；**y1**，**y2**，**y3**→**Y**⇒**输出**⇒在图 5-1.2 中，选**典型相关分析**，**典型相关选项**⇒按图 10-1.2 中的选项选择⇒**确定**⇒**确定**⇒**确定**

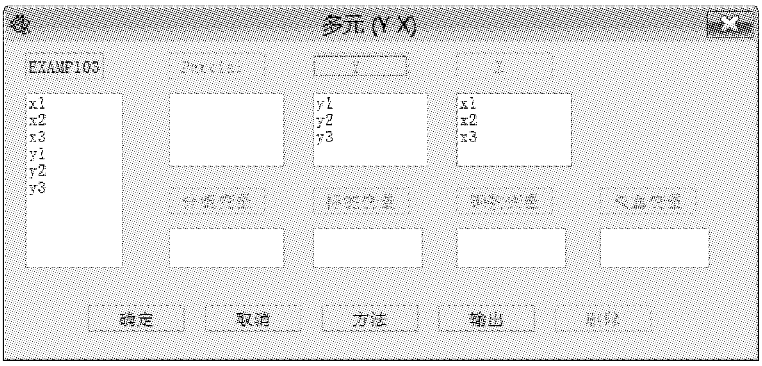


图 10-1.1　多元分析对话框

图 10-1.2　典型相关选项对话框

习　题

10.1　设 $\boldsymbol{x}=(x_1,x_2,\cdots,x_p)'$ 和 $\boldsymbol{y}=(y_1,y_2,\cdots,y_q)'$ 是两个随机向量，且 $V(\boldsymbol{x})=\boldsymbol{\Sigma}_{11}(>0)$，$V(\boldsymbol{y})=\boldsymbol{\Sigma}_{22}(>0)$，$\mathrm{Cov}(\boldsymbol{x},\boldsymbol{y})=\boldsymbol{\Sigma}_{12}$，$\mathrm{rank}(\boldsymbol{\Sigma}_{12})=m$，$\boldsymbol{\Sigma}_{21}=\boldsymbol{\Sigma}_{12}'$，$\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$ 的 m 个正特征值分别为 $\rho_1^2\geqslant\rho_2^2\geqslant\cdots\geqslant\rho_m^2$，相应的单位特征向量分别为 $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\cdots,\boldsymbol{\beta}_m$，试证：

(1) 若令 $\boldsymbol{\alpha}_i=\dfrac{1}{\rho_i}\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\beta}_i$，$\boldsymbol{a}_i=\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\alpha}_i$，$\boldsymbol{b}_i=\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\beta}_i$，这里 ρ_i 是 ρ_i^2 的算术平方根，则 $\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\cdots,\boldsymbol{\alpha}_m$ 为 $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$ 的相应于 $\rho_1^2,\rho_2^2,\cdots,\rho_m^2$ 的正交单位特征向量，$\boldsymbol{a}_1,\boldsymbol{a}_2,\cdots,\boldsymbol{a}_m$ 为 $\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$ 的相应于 $\rho_1^2,\rho_2^2,\cdots,\rho_m^2$ 的特征向量，$\boldsymbol{b}_1,\boldsymbol{b}_2,\cdots,\boldsymbol{b}_m$ 为 $\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ 的相应于 $\rho_1^2,\rho_2^2,\cdots,\rho_m^2$ 的特征向量；

(2) 试证(10.2.20)和(10.2.21)式。

10.2　对 $n=140$ 名初一学生进行四项测试：阅读速度(x_1)，阅读才能(x_2)，数学运算速度(y_1)，数学运算才能(y_2)。这四项测试的样本相关矩阵为

$$\hat{\boldsymbol{R}}=\left(\begin{array}{c:c}\hat{\boldsymbol{R}}_{11} & \hat{\boldsymbol{R}}_{12}\\ \hdashline \hat{\boldsymbol{R}}_{21} & \hat{\boldsymbol{R}}_{22}\end{array}\right)=\left(\begin{array}{cc:cc}1.0000 & 0.6328 & 0.2412 & 0.0586\\ 0.6328 & 1.0000 & -0.0553 & 0.0655\\ \hdashline 0.2412 & -0.0553 & 1.0000 & 0.4248\\ 0.0586 & 0.0655 & 0.4248 & 1.0000\end{array}\right)$$

试对阅读与数学测试成绩之间进行典型相关分析。

10.3　下表列出了 25 个家庭的成年长子和次子的头长和头宽：

x_1=长子头长，　x_2=长子头宽，　y_1=次子头长，　y_2=次子头宽

可以想象，长子和次子之间有相当的相关性。试对长子和次子之间作出典型相关分析。