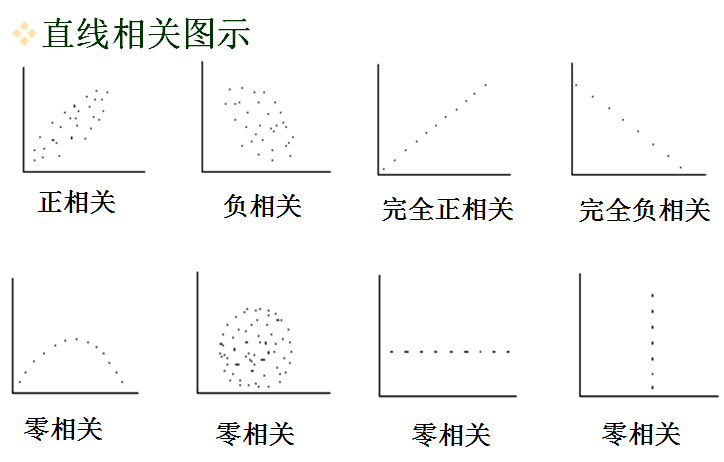
**23. 相关分析**

相关分析和回归分析是研究变量与变量间相互关系的重要方法。相关分析是研究两个或两组变量之间的线性相关情况，回归分析是拟合出变量间的表达式关系。

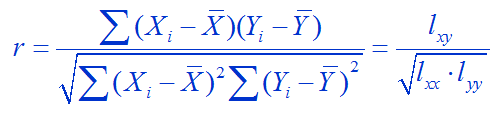
**（一）Pearson直线相关**

**一、**适用于两个变量均为**服从正态分布**，每对数据对应的点在直角坐标系中（即散点图）**呈现直线趋势**。

做相关分析时，要注意剔除异常值；相关关系不一定是因果关系。



二、用**相关系数**r∈[-1,1]来表示相关程度的大小：



r>0: 正相关；r<0: 负相关；r=0: 不相关；

r=1: 完全正相关；r=-1: 完全负相关。

相关程度的判断标准：看相关系数的平方r2，若r2<0.5，结果无实际价值。

注：相关系数只是刻画直线相关（Y=X2相关系数≠1）。

三、**假设检验**

1. H0: 总体相关系数ρ=0；H1: ρ≠0；

计算r值，P值，若P值≤α，则在显著水平α下拒绝H0;

2. 若H0成立，从ρ=0的总体中抽样，所得到的样本相关系数 r 呈对称分布（近似正态分布），此时可用 t 检验。

3. 必要时对相关系数做区间估计

从相关系数ρ≠0的总体中抽样，样本相关系数的分布是偏态的。用Z变换后，服从某种正态分布，估计z，再变换回r.

**（二）Spearman等级相关，也称Spearman秩相关**

对于不符合正态分布的资料，不用原始数据计算相关系数，而是将原始观察值由小到大编秩，然后根据秩次来计算秩相关系数rs, 以此来说明两个变量间相关关系的密切程度。

适用于某些不能准确地测量指标值而只能以严重程度、名次先后、反映大小等定出的等级资料；也适用于某些不呈正态分布或难于判断分布的资料。

**关于编秩**

将各Xi由小到大编秩得RXi（1,…n），当遇到相等的值时要用平均秩，例如X2=X4，按编秩为3和4，应该取平均秩

Rx2=Rx4=(3+4)/2=3.5

**假设检验**

H0: 总体相关系数ρs=0；H1: ρs≠0；

计算r值，P值，若P值≤α，则在显著水平α下拒绝H0;

另外，Kendall等级相关系数τ∈[-1,1]，也可以对两个变量作等级相关分析，而且可对多个变量作等级相关分析。

**（三）典型相关分析**

实际问题中经常遇到研究两组变量间的线性相关情况，例如，考察q个质量指标与p个原材料指标之间的内在联系和相关关系，这就需要用到典型相关分析。

其思想类似于主成分分析（降维思想），分别找出两组变量的一对线性组合V和V的相互关系，既可以使变量个数简化，又可以达到分析相关性的目的。如果一对线性组合不够，可以继续找下一对线性组合之间的关系，直到找不到相关变量对时为止。至于选取多少对典型相关可通过检验来确定。

**注：**（1）第一对典型相关含有最多的有关两组变量间相关的信息，第二对其次，其他对依次递减，各对典型相关所含的信息互不重复；

（2）经标准化的两组变量间的典型相关系数与原始的两组变量间的相应典型相关系数是相同的；

（3）每个典型变量除在另一组里与其配对的那个典型变量外，它同所有其他典型变量变量均不相关；

（4）第一对典型相关的大小至少同任一变量与对应的那组变量间的多重相关一样大；

（5）典型相关必须是定量数据。

**假设检验：**

1. 整体检验

H0: ρ1=…=ρr=0; H1: ρ1, …,ρr中至少ρ1≠0;

检验的统计量，若小，则拒绝H0, 接受H1.

在原假设为真的情况下，检验的统计量Q1（公式略），近似服从自由度为pq的χ2分布。在给定的显著性水平α下，若χ2≥χ2 (pq)，则拒绝原假设，认为至少第一对典型变量之间的相关性显著。再检验下一对典型变量之间的相关性。直至相关性不显著为止。

我们希望使用尽可能少的典型变量对数，为此需要对一些较小的典型相关系数是否为零进行假设检验。H0经检验被拒绝，则应进一步做下面的检验假设：

2. 部分检验

H0: ρ2＝…＝ρr＝0; Hl: ρ2,…,ρr至少有一个不为零;

若原假设H0被接受，则认为只有第一对典型变量是有用的；若原假设H0被拒绝，则认为第二对典型变量也是有用的，并进一步检验假设：

H0: ρ3＝…＝ρr＝0; H1: ρ3,…,ρr至少有一个不为零;

依次进行下去，直到对某个k，

H0: ρk十1＝…＝ρr＝0; H1: ρk十1, …,ρr至少有一个不为零。

检验的统计量Q（公式略），近似服从自由度为(p-k)(q-k)的χ2分布。在给定的显著性水平α下，如果χ2≥χ2 [(p-k)(q-k)]，则拒绝原假设，认为至少第k+1对典型变量之间的相关性显著。

**（四）PROC CORR过程步**

SAS提供的相关分析过程步是PROC CORR，可以计算：Pearson相关系数、Spearman等级相关系数、Kendall’s tau-b统计量、Hoeffding’s独立性分析统计量D。此外，还可以计算偏相关系数（固定其它变量，看两个变量的相关性）等。

**基本语法**：

PROC CORR data = 数据集 <options>;

VAR *variable-list*;

WITH *variable-list*;

<PARTIAL *variable-list*;>

**说明：**

（1）默认计算Pearson相关系数，并进行显著性检验，以及计算简单统计量；若要计算Spearman等级相关系数，需要加上可选参数“SPEARMAN”：

**PROC** **CORR** data = dataset SPEARMAN;

其它还有“HOEFFDING”计算Hoeffding's D 统计量，“KENDALL”计算Kendall's tau-b系数；

（2）VAR和WITH语句指定要做相关分析的变量，其中VAR变量是自变量，WITH变量是因变量；

（3）PARTIAL语句指定偏变量（视为常数），做偏相关分析。

**绘制相关性图形**：

PROC CORR过程步默认没有图形输出，可以加上绘图选项绘制散点图和矩阵图。基本语法：

PROC CORR data = 数据集 PLOTS = (*图形类型*);

可选图形类型有：

SCATTER——散点图，默认会加上置信椭圆；

MATRIX——所有变量的散点图矩阵；

注：（1）若不加置信椭圆用“ELLIPSE = NONE”，若要对散点图加上均值的置信椭圆：

**PROC** **CORR** data = dataset PLOTS = SCATTER(ELLIPSE = CONFIDENCE);

（2）若没有WITH语句，MATRIX选项将绘制各变量两两组合的对称散点矩阵图；默认对角线位置是空图，可以加上参数HISTOGRAM绘制直方图：

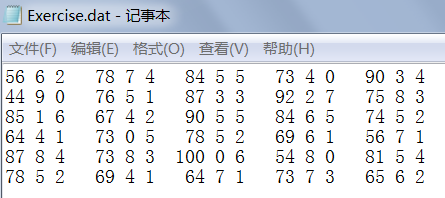
**PROC** **CORR** data = dataset PLOTS = MATRIX(HISTOGRAM);

注：默认只输出5个变量，若更多变量使用

PLOTS = MATRIX(nvar=all)

**例1（Pearson直线相关）**

某班学生考试成绩的数据（C:\MyRawData\Exercise.dat），变量包括考试成绩、考前一周看电视的时间和做练习的时间：



读入数据，用PROC CORR过程步做相关性分析。

**代码**：

**data** class;

infile 'c:\MyRawData\Exercise.dat';

input Score Television Exercise @@;

**run**;

**proc** **corr** data = class;

var Television Exercise;

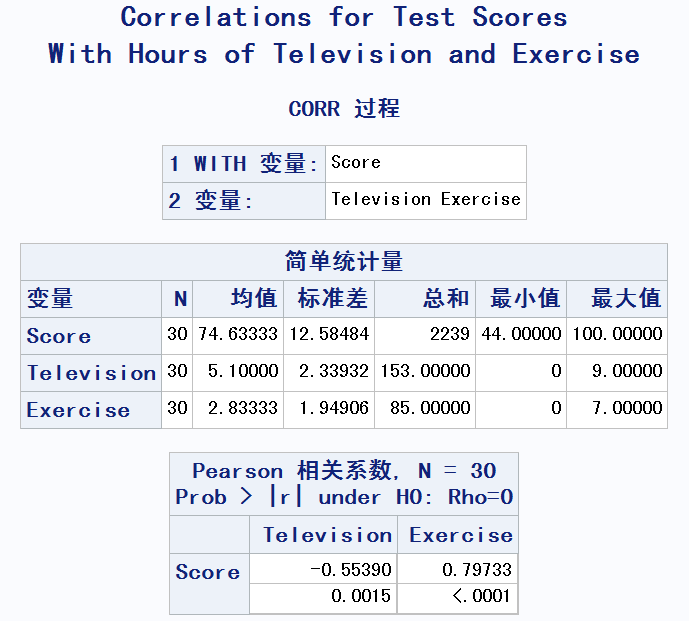
with Score;

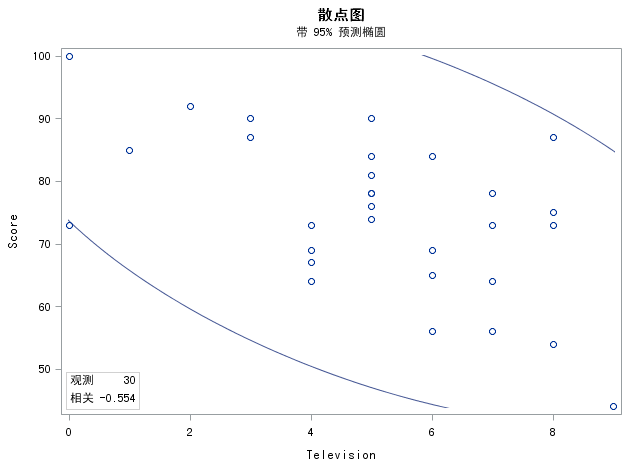
title 'Correlations for Test Scores';

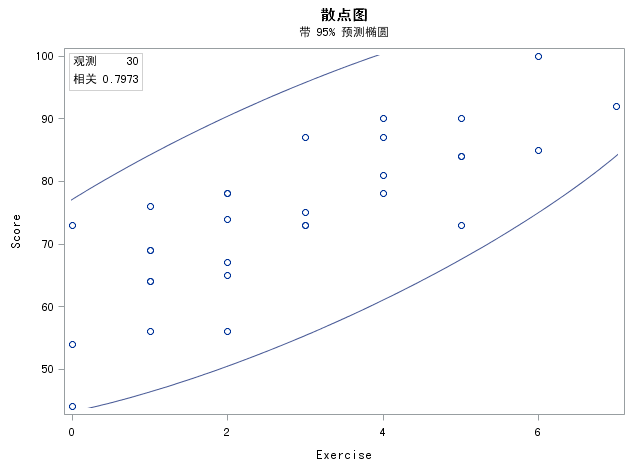
title2 'With Hours of Television and Exercise';

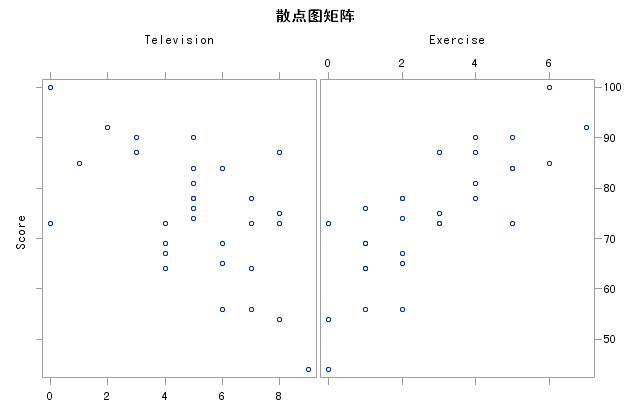
**run**;

运行结果：









**程序说明：**

从Pearson相关系数的表格可知，看电视时间与考试成绩是负相关，相关系数为-0.55390，P值= 0.0015 < α=0.05，说明两变量有相关关系，其结果具有统计学意义；做练习的时间与考试成绩是正相关，相关系数为0.79733，P值<.0001 < α=0.05，说明两变量有很大的相关关系，其结果具有统计学意义。

**例2 （Spearman秩相关）**

某销售公司想要知道，职工入职时的能力评级是否实际销售成绩一致？为了调查这个问题，公司副总裁根据10个职工的初始面试摘要、学科成绩、推荐信等材料给出了职工入职时的能力评级；根据2年后他们的实际销售成绩，得到了第二份等级评分：

|  |  |  |  |
| --- | --- | --- | --- |
| 职工编号 | 能力等级 | 销售成绩 | 成绩等级 |
| 1 | 2 | 400 | 1 |
| 2 | 4 | 360 | 3 |
| 3 | 7 | 300 | 5 |
| 4 | 1 | 295 | 6 |
| 5 | 6 | 280 | 7 |
| 6 | 3 | 350 | 4 |
| 7 | 10 | 200 | 10 |
| 8 | 9 | 260 | 8 |
| 9 | 8 | 220 | 9 |
| 10 | 5 | 385 | 2 |

注：“1”表示能力最强。

**代码：**

**data** persons;

input abilities performance;

performance=**400**-performance;

datalines;

2 400

4 360

7 300

1 295

6 280

3 350

10 200

9 260

8 220

5 385

;

**proc** **corr** data=persons spearman;

var abilities;

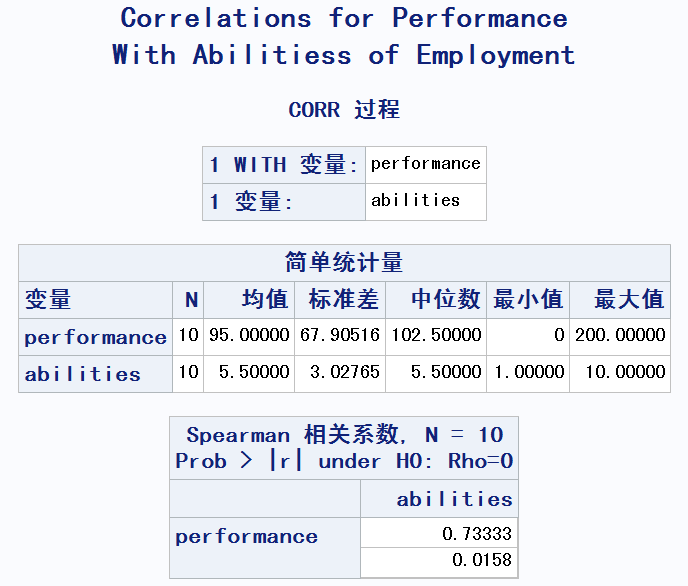
with performance;

title 'Correlations for Performance';

title2 'With Abilitiess of Employment';

**run**;

运行结果：



**程序说明：**

（1）销售成绩数值越大则销售能力越强（等级分越小），所以用最大销售值400- performance做一下颠倒；

（2）Spearman等级相关系数为0.73333（有较大的相关性），P值=0.0158<α=0.05，说明其结果具有统计学意义。

**例3**某康复俱乐部对20名中年人测量了三项生理指标：体重(weight)、腰围(waist)、脉搏(pulse)，以及三项训练指标：引体向上(chins)、起坐次数(situps)、跳跃次数(jumps)。其数据列表如下，试分析这两组变量间的相关性。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Obs* | *weight* | *waist* | *pulse* | *chins* | *situps* | *jumps* |
| 1 | 191 | 36 | 50 | 5 | 162 | 60 |
| 2 | 189 | 37 | 52 | 2 | 110 | 60 |
| 3 | 193 | 38 | 58 | 12 | 101 | 101 |
| 4 | 162 | 35 | 62 | 12 | 105 | 37 |
| 5 | 189 | 35 | 46 | 13 | 155 | 58 |
| 6 | 182 | 36 | 56 | 4 | 101 | 42 |
| 7 | 211 | 38 | 56 | 8 | 101 | 38 |
| 8 | 167 | 34 | 60 | 6 | 125 | 40 |
| 9 | 176 | 31 | 74 | 15 | 200 | 40 |
| 10 | 154 | 33 | 56 | 17 | 251 | 250 |
| 11 | 169 | 34 | 50 | 17 | 120 | 38 |
| 12 | 166 | 33 | 52 | 13 | 210 | 115 |
| 13 | 154 | 34 | 64 | 14 | 215 | 105 |
| 14 | 247 | 46 | 50 | 1 | 50 | 50 |
| 15 | 193 | 36 | 46 | 6 | 70 | 31 |
| 16 | 202 | 37 | 62 | 12 | 210 | 120 |
| 17 | 176 | 37 | 54 | 4 | 60 | 25 |
| 18 | 157 | 32 | 52 | 11 | 230 | 80 |
| 19 | 156 | 33 | 54 | 15 | 225 | 73 |
| 20 | 138 | 33 | 68 | 2 | 110 | 43 |

**代码：**

**data** fit ;

input weight waist pulse chins situps jumps;

datalines;

191 36 50 5 162 60

189 37 52 2 110 60

193 38 58 12 101 101

162 35 62 12 105 37

189 35 46 13 155 58

182 36 56 4 101 42

211 38 56 8 101 38

167 34 60 6 125 40

176 31 74 15 200 40

154 33 56 17 251 250

169 34 50 17 120 38

166 33 52 13 210 115

154 34 64 14 215 105

247 46 50 1 50 50

193 36 46 6 70 31

202 37 62 12 210 120

176 37 54 4 60 25

157 32 52 11 230 80

156 33 54 15 225 73

138 33 68 2 110 43

;

**run**;

**proc** **cancorr** data=fit all

vprefix=PHYS vname='PhysiologicalMeasurements'

wprefix=EXER wname='Exercises';

var weight waist pulse;

with chins situps jumps;

**run**;

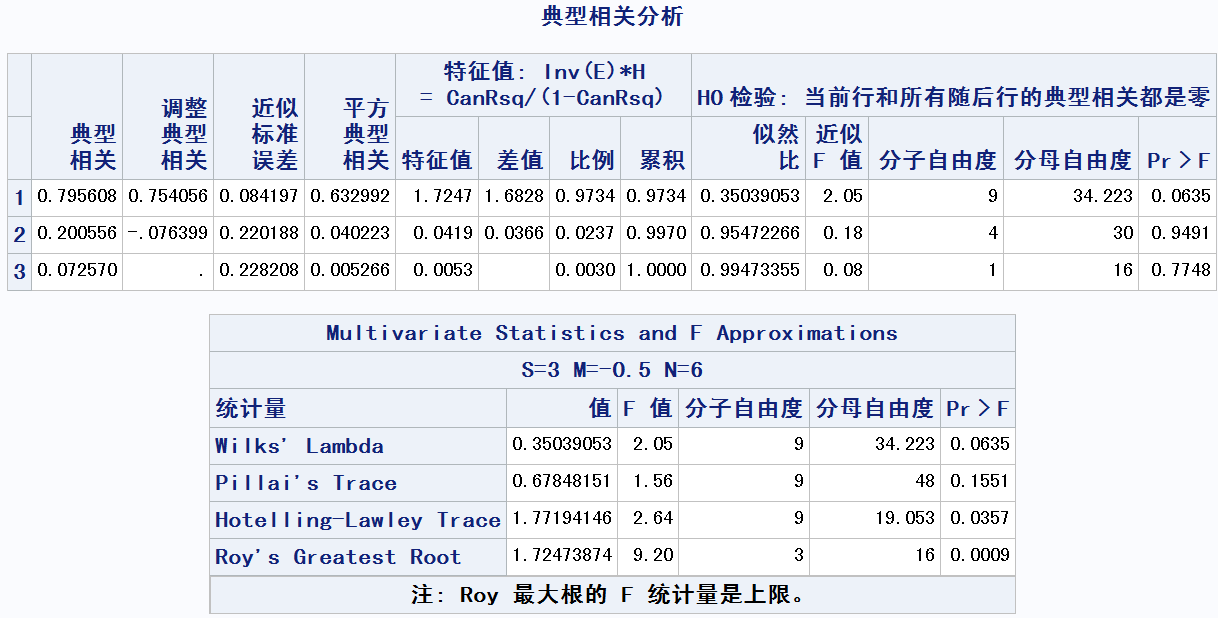
**运行结果及结果说明**：



两组变量各有3个指标变量，共20条观测；列出了各个变量的均值和标准差。



列出组内和组间各变量的相关系数。生理指标和训练指标之间的相关系数是中等的，其中waist和situps的相关系数最大为-0.6456。组内较大的相关是weight和waist的相关系数为0.8702，chins和situps的相关系数为0.6957，situps和jumps的相关系数为0.6692。



第一对典型变量（u，v）之间的典型相关系数为0.795608（应该比每个组间相关系数都大才正确）。校正值为0.754056，标准误差为0.084197，典型相关系数的平方（CanRsq）为0.632992。

矩阵Inv(E)\*H的特征值等于CanRsq/(1－CanRsq)，相对应的三个特征值依次为：

1.7247=0.632992/（1－0.632992）

0.0419=0.040223/（1－0.040223）

0.0053=0.005266/（1－0.005266）

用似然比法检验典型相关系数与零的差别是否显著，其零假设为小于此对典型变量的典型相关系数的所有典型相关系数都为0，其P值依次为0.0635, 0.9491, 0.7748，在显著水平取α=0.1时， 0.0635<0.1，说明第一对典型相关系数具有显著意义。

与用Wilks统计量进行多元分析的结果是等价的（F值和P值都相等），测验结果为：

Λ=0.35039053

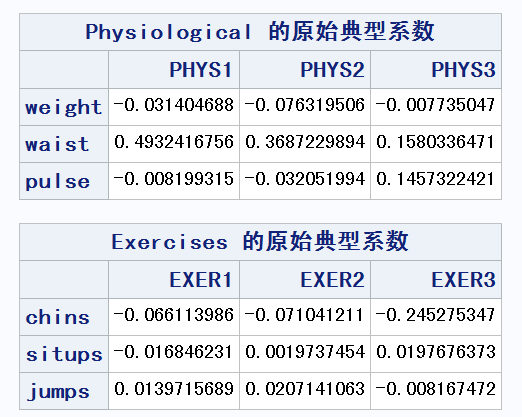
近似F统计量＝2.05

Num DF（分子自由度）=9

Den DF（分母自由度）=34.2223

P值＝0.0635

其余典型相关系数明显不显著。另外，从第2个修正典型相关系数为-0.076399，也可以得出以上只取一对典型相关系数的结论。





用原指标和表中的系数，来线性组合表示第一对典型变量：

PHYS1=－0.031404688weight＋0.4932416756waist－0.008199315pulse

EXER1=－0.066113986chins－0.016846231situps＋0.0139715689jumps

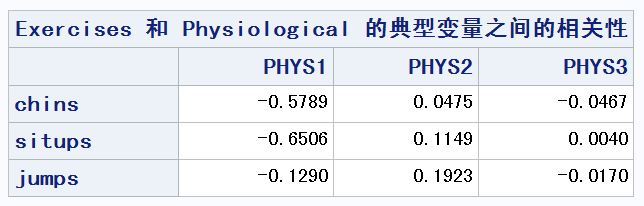
由于变量的单位不一致，应考虑用标准化的典型系数：

PHYS1= －0.7754weight ＋1.5793 waist－0.0591pulse

EXER1= －0.3495chins－1.0540situps＋0.7164jumps

来自生理指标的第一典型变量主要为waist（1.5793）和weight（-0.7754）的加权差，在waist上的权数更大些，在pulse 上的系数近似为0。来自训练指标的第一典型变量在situps上的系数最大。





在给出的典型结构（原始变量与典型变量之间的相关系数）表中，waist和weight与第一典型变量PHYS1的相关系数皆为正值，分别为0.9254和0.6206。因而weight为一抑制变量，因为它在典型变量PHYS1线性表示中的系数符号（-0.7754，负号）与它跟典型变量PHYS1的相关系数符号（0.6206，正号）相反。

所谓的抑制变量是指，它的增加或减少会引起目标变量间的相关系数的绝对值的减少或增加，即相关性减弱或变强。一个变量成为抑制变量是因为它同典型变量的相关系数符号与系数符号相反。

为了理解这是为什么，以抑制变量体重（wight）为例来说明情况。一般来说，瘦的人比胖的人起坐次数（situps）多，而来自训练指标的典型变量EXER1主要与起坐次数（situps）相关。来自生理指标的典型变量PHYS1主要由体重（wight）和腰围（waist）的加权差构成，而体重和腰围之间有很强的正相关性（0.8702），我们可以近似认为体重=腰围×身高，因此典型变量PHYS1实质上是反映一个人的胖瘦程度的某一种指标。

综上所述，我们可以通过肥胖性与起坐次数（situps）的基本相关关系得出体重（wight）和腰围（waist）与起坐次数（situps）的基本相关关系。显然，体重（wight）或腰围（waist）大的人比小的人要肥胖，因此与起坐次数（situps）为负相关。

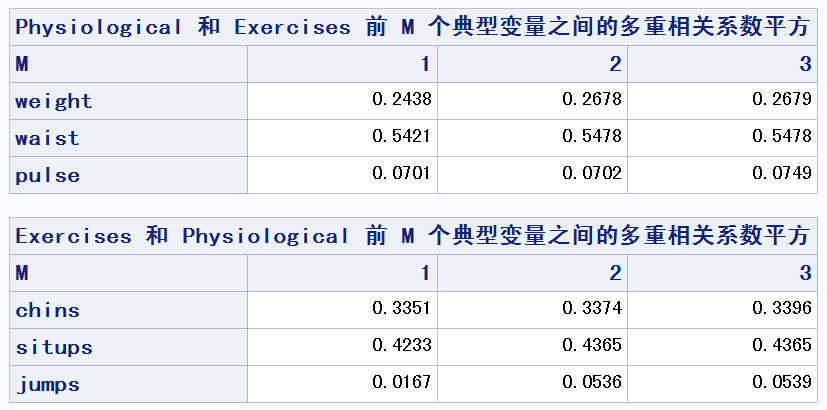
我们考虑简单的情况，把寻找典型变量的线性组合，假设用多元回归方法产生，即由体重（wight）和腰围（waist）来预测起坐次数（situps）。如果固定体重（wight）的值，腰围（waist）大的人，身高就矮，因此人倾向于较肥胖，所以预测起坐次数（situps）应该较小，这样，腰围（waist）在多元回归中的回归系数一定是负值才能使起坐次数（situps）减小。如果固定腰围（waist）的值，体重（wight）大的人，身高就高，因此人倾向于较瘦，所以预测起坐次数（situps）应该较大，这样，体重（wight）在多元回归中的回归系数一定是正值才能使起坐次数（situps）增加。

从上分析我们看到，体重（wight）在预测的起坐次数（situps）时的回归系数是正号，而体重（wight）与起坐次数（situps）的相关性是负号，两者符号相反。

综合结论，第一对典型相关的一般解释为以体重（wight）和跳跃次数（jumps）作为抑制变量来增强或减弱腰围（waist）和起坐次数（situps）之间的相关。







典型冗余分析（Canonical Redundancy Analysis）。我们略过原始变量的方差分析，而来分析标准化的方差。第一典型变量PHYS1和EXER1皆不能很好全面地用于对应的那组变量的预测。

来自生理指标的标准差被自己的第一个典型变量PHYS1解释的方差比例为0.4508，而被对方第一个典型变量PHYS1解释的方差比例为0.2854。来自训练指标的标准差被对方第一典型变量PHYS1解释的方差比例为0.2584。第二对和第三对典型变量实际上没有对标准差有什么贡献，因为三个典型变量的累计比例仅为0.2969和0.2767。

由多重相关系数的平方（多元判定系数）表明，生理指标的第一典型变量PHYS1对运动变量chins（多元判定系数为0.3351）和situps（多元判定系数为0.4233）有一定的预测作用，但对变量jumps（0.0167）几乎没有预测作用。训练指标的第一典型变量EXER1对生理变量waist（0.5421）有相当好的预测作用，对生理变量weight（0.2438）预测作用较差，而对生理变量pulse（0.0701）几乎没有预测作用。