著性水平 α 下是显著的。

8.2　典型相关分析的步骤及逻辑框图

典型相关分析的步骤有以下 6 个：(1) 确定典型相关分析的目标；(2) 设计典型相关分析；(3) 检验典型相关分析的基本假设；(4) 推导典型函数，评价整体拟合情况；(5) 解释典型变量；(6) 验证模型。更详细的内容可见参考文献 [5]。它实现的逻辑框图如图 8—1 所示。

图 8—1　典型相关分析的逻辑框图

第 1 步：确定典型相关分析的目标

典型相关分析所适用的数据是两组变量。我们假定每组变量都能赋予一定的理论意义，通常一组可以定义为自变量，另一组可以定义为因变量。典型相关分析可以达到以下目标：

（1）确定两组变量相互独立，或者相反，确定两组变量间存在关系的强弱。

（2）为每组变量推导出一组权重，使得每组变量的线性组合达到最大程度相关。最大化余下的相关关系的其他的线性函数与前面的线性函数是独立的。

（3）解释自变量与因变量组中存在的相关关系，通常是通过测量每个变量对典型函数的相对贡献来衡量。

第 2 步：设计典型相关分析

典型相关分析作为一种多元分析方法，与其他多元分析技术有共同的基本要求。其他方法（尤其是多元回归、判别分析和方差分析）所讨论的测量误差的影响、变量类型及变换也与典型相关分析有很大关系。

样本大小的影响和每个变量需要足够的观测，都是典型相关分析经常遇到的。研究者容易使自变量组和因变量组包含很多的变量，而没有认识到样本量的含义。小的样本不能很好地代表相关关系，这样掩盖了有意义的相关关系。建议研究者至少保持每个变量 10 个观测，以避免数据的“过度拟合”。

第 3 步：检验典型相关分析的基本假定

线性假定影响典型相关分析的两个方面。首先，任意两个变量间的相关系数是基于线性关系的。如果这个关系不是线性的，则一个或者两个变量需要变换。其次，典型相关是变量间的相关。如果关系不是线性的，典型相关分析将不能测量到这种关系。

典型相关分析能够包容任何没有严格正态性假定的度量变量。正态性是有意义的，因为它使分布标准化，允许变量间更高程度的相关。但在严格意义上，如果变量的分布形式（比如高度偏态）不会削弱与其他变量的相关关系，典型相关分析是可以包含这种非正态变量的，这就允许使用非正态变量。然而，对于每个典型函数的多元正态性的统计检验是必要的。由于多元正态性检验不一定可行，流行的准则是保证每个单变量的正态性。这样，尽管不严格要求正态性，但建议所有变量都检验正态性。如有必要，对变量进行变换。

第 4 步：推导典型函数、评价整体拟合情况

每个典型函数都包括一对变量，通常一个代表自变量，另一个代表因变量。可从变量组中提取的典型变量（函数）的最大数目等于最小数据组中的变量数目。比如，一个研究问题包含 5 个自变量和 3 个因变量，可提取的典型函数的最大数目是 3。

（1）推导典型函数。典型函数的推导类似于没有旋转的因子分析的过程（参见前面推导）。典型相关分析集中说明两组变量间的最大相关关系，而不是一组变量。结果是第一对典型变量在两组变量中有最大的相关关系。第二对典型变量得到第一对典型变量没有解释的两组变量间的最大相关关系。简言之，随着典型变量的提

取，接下来的典型变量是基于剩余残差提取的，并且典型相关系数会越来越小。每对典型变量是正交的，并且与其他典型变量是独立的。

典型相关程度是通过相关系数的大小来衡量的。典型相关系数的平方表示一个典型变量通过另外一个典型变量所解释的方差比例，也可称作两个典型变量间共同方差的比例。典型相关系数的平方称作典型根或者特征根。

（2）典型函数的解释。一般来讲，实际提取的典型函数都是典型相关系数在某个水平（比如 0.05）上显著的函数。对显著的典型变量的解释是基于这样的假设，即认为相关的函数中，每组中的变量都对共同方差有较大贡献。

海尔（Hair，1984）等人推荐结合使用三个准则来解释典型函数。这三个准则是：1）函数的统计显著性水平；2）典型相关的大小；3）两个数据集中方差解释的冗余测量。

通常认为，一个有统计显著性的相关系数的可接受显著性水平是 0.05（也有 0.01 的水平）。统计软件所提供的最常见的检验是基于 Rao 近似的 F 统计量。除了对每个典型函数分别进行检验以外，全部典型根的多元检验也可以用来评价典型根的显著性。许多评价判别函数显著性的测量，包括 Wilks' Lambda、Hotelling 迹、Pillai 迹和 Roy's gcr，这里也可以给出。

典型函数的实际重要性是由典型相关系数的大小代表的。当决定解释哪些函数时，应当考虑典型相关系数。

前面讲到典型相关系数的平方可以提供典型变量间共同方差的一个估计。尽管这是对共同方差的一个简单明了的估计，它还是可能引起一些误解，因为典型相关系数的平方表示由因变量组和自变量组的线性组合所共享的方差，而不是来自两组变量的方差。这样，即使两个典型变量可能并没有从它们各自的变量组中提取显著方差，但这两个典型变量（线性组合）间仍可能得到一个相对较强的典型相关系数。为了克服在使用典型根（典型相关系数平方）作为共同方差的测量中可能出现的有偏性和不稳定性，提出了冗余指数。它等价于在整个自变量组与因变量组的每一个因变量之间计算多元相关系数的平方，然后将这些平方系数平均得到一个平均的 R^2。这个指数提供了一组自变量（取整个组）解释因变量（每次取一个）变化的能力的综合测量。这样，冗余测量就像多元回归的 R^2 统计量，作为一个指数的值也是类似的。Stewart-Love 冗余指数计算一组变量的方差能被另一组变量的方差解释的比例。请注意，典型相关不同于多元回归之处在于，它不是处理单个因变量，而是处理因变量的组合，而且这个组合只有每个因变量的全部方差的一部分。由于这个原因，我们不能假定因变量组中 100%的方差能由自变量组解释。自变量组期望能够解释的只是因变量组的典型变量的共同方差。这样，计算冗余指数分三步：（1）共同方差的比例。在典型相关分析中，我们关心因变量组的典型变量与每个因变量的相关关系，这可以从典型载荷（L_1）中获得，表示每个输入变量与它的典型变量间的相关系数。通过平方每个因变量的载荷（L_i^2），可以得到每个因变量通过因变量组的典型变量解释的方差比例。为了计算典型变量所解释的共同方差的比例，将典型载荷平方进行

简单平均。(2) 解释的方差比例。第二步是要计算通过自变量典型变量能够解释的因变量典型变量的方差比例。这也就是自变量典型变量与因变量典型变量间相关系数的平方，即典型相关系数的平方。(3) 冗余指数。一个典型变量的冗余指数就是这个变量的共同方差比例乘以典型相关系数平方，得到每个典型函数可以解释的共同方差部分。要得到较高的冗余指数，必须有较高的典型相关系数和由因变量典型变量解释的较高的共同方差比例。研究者应注意，虽然在典型函数中两个典型变量的典型相关系数是相同的，但是两个典型变量的冗余指数却有可能差异很大，因为每个都有不同的共同方差比例。已有人提出关于冗余指数的检验，但还没有得到广泛应用。

第 5 步：解释典型变量

即使典型相关系数在统计上是显著的，典型根和冗余系数大小也是可接受的，研究者仍需对结果做大量的解释。这些解释包括研究典型函数中原始变量的相对重要性。主要使用以下三种方法：(1) 典型权重（标准化系数）；(2) 典型载荷（结构系数）；(3) 典型交叉载荷。

(1) 典型权重。传统的解释典型函数的方法包括观察每个原始变量在它的典型变量中的典型权重的符号和大小。有较大的典型权重，则说明原始变量对它的典型变量贡献较大，反之则相反。原始变量的典型权重有相反的符号，说明变量之间存在一种反向关系，反之则有正向关系。但是，这种解释遭到了很多批评。因此，在解释典型相关的时候应慎用典型权重。

(2) 典型载荷。由于典型权重的缺陷，典型载荷逐步成为解释典型相关分析结果的基础。典型载荷，也称典型结构相关系数，是原始变量（自变量或者因变量）与它的典型变量间的简单线性相关系数。典型载荷反映原始变量与典型变量的共同方差，它的解释类似于因子载荷，也就是每个原始变量对典型函数的相对贡献。

(3) 典型交叉载荷。它的提出是作为典型载荷的替代。计算典型交叉载荷包括使每个原始因变量与自变量典型变量直接相关。交叉载荷提供了一个更直接地测量因变量组与自变量组关系的指标。

第 6 步：验证模型

与其他多元分析方法一样，典型相关分析的结果应该验证，以保证结果不是只适合样本，而是适合总体。最直接的方法是构造两个子样本（如果样本量允许），对每个子样本分别做分析，这样可以比较典型函数的相似性、典型载荷等。如果存在显著差别，研究者应深入分析，保证最后结果是总体的代表，而不只是单个样本的反映。

另一种方法是测量结果对于剔除一个因变量或自变量的灵敏度，保证典型权重和典型载荷的稳定性。

另外，还必须看到典型相关分析的局限性。这些局限性中，对结论和解释影响最大的是：

(1) 典型相关反映变量组的线性组合所共享的方差，而不是从变量提取的方差。

（2）计算典型函数推导的典型权重有较大的不稳定性。

（3）推导的典型权重是最大化线性组合间的相关关系，而不是提取的方差。

（4）典型变量的解释可能会比较困难，因为它们是用来最大化线性关系的，没有类似于方差分析中变量旋转的有助于解释的工具。

（5）难以识别自变量和因变量的子集间有意义的关系，只能通过一些不充分的测量，比如载荷和交叉载荷。

8.3　典型相关分析的上机实现

典型相关分析可以通过 SPSS 和 SAS 实现。在 SPSS 中使用宏命令语句可以执行典型相关分析，但由宏命令得到的结果往往不能满足全部的分析需要，还需要调用其他的命令，参见参考文献 [5]。这里我们使用功能强大的 SAS 软件来实现典型相关分析。

例 8—1

这里采用 SAS 软件中的一个生理指标与运动关系的样本程序来说明。SAS/STAT 中的 CANCORR 模块是用来实现典型相关分析的。样本程序如表 8—1 所示。

表 8—1

```
data fit;
    input weight waist pulse chins situps jumps;
    cards;
191  36  50  5  162  60
…    …   …   …  …    …
run;

proc cancorr data=fit all
    vprefix=PHYS vname='Physiological Measurements'
    wprefix=EXER wname='Exercises';
    var weight waist pulse;
    with chins situps jumps;
    run;
```

data 步读取数据，这里有 6 个变量——3 个自变量、3 个因变量。一组变量表示生理指标（physiological measurements），有体重（weight）、腰围（waist）和脉搏（pulse）3 个变量；另一组变量表示运动指标（exercise），有引体向上（chins）、仰卧起坐（situps）和跳跃次数（jumps）。为了研究这两组变量之间的相关关系，我们使用典型相关分析。proc cancorr 表示调用典型相关程序，选项 all 表示输出所有的结果。vprefix=PHYS 表示这组变量的典型变量前缀为 PHYS，它的第一个典型变量为 PHYS1，第二个典型变量为 PHYS2，……。wprefix=

EXER表示另一组变量的典型变量前缀为EXER，它的第一个典型变量为EXER1，第二个典型变量为EXER2，……。var后面接一组变量，with后面接另一组变量。

得到的部分输出结果如输出结果8—1所示。

输出结果 8—1

①

correlations Among the original variables

correlations Among the physiological Measurements

	weight	waist	pulse
weight	1.000 0	0.870 2	−0.365 8
waist	0.870 2	1.000 0	−0.352 9
pulse	−0.365 8	−0.352 9	1.000 0

correlations Among the Exercises

	chins	situps	jumps
chins	1.000 0	0.695 7	0.495 8
situps	0.695 7	1.000 0	0.669 2
jumps	0.495 8	0.669 2	1.000 0

correlations Between the physiological Measurements and the Exercises

	chins	situps	jumps
weight	−0.389 7	−0.493 1	−0.226 3
waist	−0.552 2	−0.645 6	−0.191 5
pulse	0.150 6	0.225 0	0.034 9

②

The CANCORR Procedure

Canonical Correlation Analysis

	canonical correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.795 608	0.754 056	0.084 197	0.632 992
2	0.200 556	−0.076 399	0.220 188	0.040 223
3	0.072 570		0.228 208	0.005 266

Test of HO: The canonical correlations in the current row and all that follow are zero

Eigenvalues of Inv (E) * H = CanRsq/ (1 − CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative	Likelihood Approximate Ratio	F	Value	Num DF	Den DF	Pr>F
1	1.724 7	1.682 8	0.973 4	0.973 4	0.350 390 53	2.05	9	34.223	0.063 5	
2	0.041 9	0.036 6	0.023 7	0.997 0	0.954 722 66	0.18	4	30	0.949 1	
3	0.005 3		0.003 0	1.000 0	0.994 733 55	0.08	1	16	0.774 8	

Multivariate Statistics and F Approximations

S = 3 M = −0.5 N = 6

statistic	value	F Value	Num DF	Den DF	Pr>F
Wilks' Lambda	0.350 390 53	2.05	9	34.223	0.063 5
Pillai's Trace	0.678 481 51	1.56	9	48	0.155 1
Hotelling-Lawley Trace	1.771 941 46	2.64	9	19.053	0.035 7
Roy's Greatest Root	1.724 738 74	9.20	3	16	0.000 9

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

Canonical correlation Analysis ③

Standardized Canonical Coefficients for the Physiological Measurements

	PHYS1	PHYS2	PHYS3
weight	-0.775 4	-1.884 4	-0.191 0
waist	1.579 3	1.180 6	0.506 0
pulse	-0.059 1	-0.231 1	1.050 8

Standardized Canonical Coefficients for the Exercises

	EXER1	EXER2	EXER3
chins	-0.349 5	-0.375 5	-1.296 6
situps	-1.054 0	0.123 5	1.236 8
jumps	0.716 4	1.062 2	-0.418 8

Canonical Structure ④

Correlations Between the Physiological Measurements and Their Canonical Variables

	PHYS1	PHYS2	PHYS3
weight	0.620 6	-0.772 4	-0.135 0
waist	0.925 4	-0.377 7	-0.031 0
pulse	-0.332 8	0.041 5	0.942 1

Correlations Between the Exercises and Their Canonical Variables

	EXER1	EXER2	EXER3
chins	-0.727 6	0.237 0	-0.643 8
situps	-0.817 7	0.573 0	0.054 4
jumps	-0.162 2	0.958 6	-0.233 9

Correlations Between the Physiological Measurements and the Canonical Variables of the Exercises

	EXER1	EXER2	EXER3
weight	0.493 8	-0.154 9	-0.009 8
waist	0.736 3	-0.075 7	-0.002 2
pulse	-0.264 8	0.008 3	0.068 4

Correlations Between the Exercises and the Canonical Variables of the Physiological Measurements

	PHYS1	PHYS2	PHYS3
chins	-0.578 9	0.047 5	-0.046 7
situps	-0.650 6	0.114 9	0.004 0
jumps	-0.129 0	0.192 3	-0.017 0

Canonical Redundancy Analysis ⑤

Raw Variance of the Physiological Measurements Explained by

	Their Own Canonical Variables		The Opposite Canonical Variables		
Canonical Variable Number	Proportion	Cumulative Proportion	Canonical R-Square	Proportion	Cumulative Proportion
1	0.371 2	0.371 2	0.633 0	0.234 9	0.234 9
2	0.543 6	0.914 8	0.040 2	0.021 9	0.256 8
3	0.085 2	1.000 0	0.005 3	0.000 4	0.257 3

Raw Variance of the Exercises Explained by

	Their own Canonical Variables		The Opposite Canonical Variables		
Canonical Variable Number	Proportion	Cumulative Proportion	Canonical R-Square	Proportion	Cumulative Proportion
1	0.411 1	0.411 1	0.633 0	0.260 2	0.260 2
2	0.563 5	0.974 6	0.040 2	0.022 7	0.282 9
3	0.025 4	1.000 0	0.005 3	0.000 1	0.283 0

The CANCORR Procedure ⑥

Canonical Redundancy Analysis

Standardized Variance of the Physiological Measurements Explained by

	Their Own Canonical Variables		The Opposite Canonical Variables		
Canonical Variable Number	Proportion	Cumulative Proportion	Canonical R-Square	Proportion	Cumulative Proportion
1	0.450 8	0.450 8	0.633 0	0.285 4	0.285 4
2	0.247 0	0.697 8	0.040 2	0.009 9	0.295 3
3	0.302 2	1.000 0	0.005 3	0.001 6	0.296 9

Standardized Variance of the Exercises Explained by

	Their own Canonical Variables		The Opposite Canonical Variables		
Canonical Variable Number	Proportion	Cumulative Proportion	Canonical R-Square	Proportion	Cumulative Proportion
1	0.408 1	0.408 1	0.633 0	0.258 4	0.258 4
2	0.434 5	0.842 6	0.040 2	0.017 5	0.275 8
3	0.157 4	1.000 0	0.005 3	0.000 8	0.276 7

Canonical Redundancy Analysis ⑦

Squared Multiple Correlations Between the Physiological Measurements and the First M Canonical Variables of the Exercises

M	1	2	3
weight	0.243 8	0.267 8	0.267 9
waist	0.542 1	0.547 8	0.547 8
pulse	0.070 1	0.070 2	0.074 9

Squared Multiple Correlations Between the Exercises and the First M Canonical Variables of the Physiological Measurements

M	1	2	3
chins	0.335 1	0.337 4	0.339 6
situps	0.423 3	0.436 5	0.436 5
jumps	0.016 7	0.053 6	0.053 9

输出结果 8—1 中第①张表表示原始变量间的相关关系。体重与腰围有较强的正相关关系。在生理测量与运动的相关关系中，我们可以看到体重和腰围与三个运动指标的相关系数为负数，说明体重和腰围较大对运动能力有负面影响。

第②张表是对典型相关系数的检验。(1) 有三个典型相关系数。(2) 对这三个典型相关系数的检验。特征根与典型相关系数的关系是：特征根＝(典型相关系数^2)/(1－典型相关系数^2)。由该检验认为，只有第一个典型相关系数在 0.1 的水平上是显著的。(3) 多元统计检验是用来检验典型根的显著性。结果显示，Hotelling 迹和 Roy's gcr 在 0.05 的显著性水平上，认为典型根是显著的。

第③张表是各组的标准化典型相关系数（典型权重）。这里我们只提取第一个典型变量。第④张表是对典型结构（典型载荷和交叉载荷）的分析。根据前面对这三种方法的介绍，我们可以结合第③张表和第④张表对变量间的关系进行分析。在原始变量与它的典型变量的相关关系（典型载荷）中，生理测量的第一个典型变量与腰围的相关系数最大，说明这个典型变量主要反映人的体形肥胖程度；运动因素的第一个典型变量与仰卧起坐次数和引体向上次数有较强的负相关关系，说明这个典型变量主要反映人不适合运动的程度。在原始变量与另一组原始变量的典型变量

的相关关系（典型交叉载荷）的分析结果中，腰围与运动的第一典型变量的相关性较强，这也说明了腰围大（体形较胖）则运动能力差；仰卧起坐和引体向上与生理测量的第一典型变量呈一定的负相关关系，说明人的体形肥胖程度对这两种运动能力有负面影响。

第⑤～⑦张表是典型冗余分析。第⑤张表是变量的原始方差通过它的典型变量和配对的典型变量所解释的方差比例，第⑥张表是变量的标准化方差通过它的典型变量和配对的典型变量所解释的方差比例，这里我们一般使用第⑥张表，它消除了量纲和单位的影响。由第⑥张表，生理测量通过它的第一个典型变量解释的共同方差的比例是 45.08%，而通过配对的另一个典型变量解释的方差比例是 63.3%，这样，冗余指数应为 28.54%（即 45.08%×63.3%），说明运动指标的第一个典型函数可以解释生理测量（因变量组）的总方差的比例是 28.54%。这个解释方差的比例不高，其具体原因有待进一步的研究。第⑦张表是典型变量与原始变量的平方相关系数，类似于回归分析中的复相关系数 R^2，结果显示运动指标的第一典型变量对生理测量中的腰围指标的解释能力最强，这也可以说明运动对体形影响较大，比如通过体育锻炼可以减肥。

8.4　社会经济案例研究

例 8—2

城市竞争力与城市基础设施关系研究。

基础设施是城市经济、社会活动的基本载体，是城市竞争力的重要组成部分。本例选取中国 20 个城市的统计数据，对城市基础设施和竞争力间的关系进行典型相关分析。

表 8—2 为城市竞争力与城市基础设施数据（数据读者可在网站 www.ruc-6sigma.com《应用多元统计分析》第 14 章下载）。

表 8—2

	城市	劳动生产率	市场占有率	居民人均收入	经济增长率	对外设施指数	对内设施指数	百人电话数	技术设施指数	文化设施指数	卫生设施指数
1	上海	45623.05	2.50	8439.00	16.27	1.09	0.42	50.00	2.15	1.23	1.64
2	深圳	52256.67	1.30	18579.00	21.50	1.34	0.13	131.00	0.33	-0.27	-0.64
3	广州	46551.87	1.13	10445.00	11.92	1.07	0.40	48.00	1.31	0.49	0.09
4	北京	28146.76	1.38	7813.00	16.00	-0.43	0.19	20.00	0.87	3.57	1.80
5	厦门	38670.43	0.12	8980.00	26.71	-0.53	0.26	32.00	-0.09	-0.33	-0.84
6	天津	26316.96	1.37	6609.00	11.07	-0.11	0.07	27.00	0.68	-0.12	0.87
7	大连	45330.53	0.56	6070.00	12.40	0.35	0.06	31.00	0.26	-0.30	-0.16
8	杭州	45853.89	0.26	7896.00	13.93	-0.50	0.27	38.00	-0.78	-0.12	1.61
9	南京	35964.64	0.74	6497.00	8.97	0.31	0.26	43.00	0.49	-0.09	-0.06
10	珠海	55832.61	-0.12	13149.00	9.22	-0.28	0.84	37.00	-0.79	-0.49	-0.98
11	青岛	33334.92	0.63	6222.00	11.63	0.01	-0.14	24.00	0.37	-0.40	-0.49
12	武汉	24633.27	0.59	5573.00	16.39	0.02	-0.47	28.00	0.03	0.15	0.26
13	温州	39298.79	-0.88	9034.00	22.43	-0.47	0.03	45.00	-0.76	-0.46	-0.75
14	福州	38201.47	-0.34	7089.00	18.53	-0.45	-0.20	34.00	-0.45	-0.34	-0.52
15	重庆	16524.32	0.44	5323.00	12.22	0.72	-0.83	13.00	0.05	-0.09	0.58
16	成都	31866.63	-0.02	6019.00	11.88	0.37	-0.54	21.00	-0.11	-0.24	-0.02
17	宁波	22528.80	-0.16	9069.00	16.70	0.01	0.38	40.00	-0.17	-0.40	-0.71
18	石家庄	21831.94	-0.15	5497.00	13.56	-0.81	-0.49	22.00	-0.38	-0.21	-0.59
19	西安	19966.34	-0.15	5344.00	12.43	-0.24	-0.91	18.00	-0.05	-0.27	0.61
20	哈尔滨	19225.71	-0.16	4233.00	10.16	-0.53	-0.77	27.00	-0.45	-0.16	1.08

SAS软件程序读者可在网站 www.ruc-6sigma.com《应用多元统计分析》第14章下载。

（1）变量间的相关性。如输出结果8—2所示。

输出结果 8—2

The CANCORR Procedure

Correlations Among the Original Variables

Correlations Among the VAR Variables

	x1	x2	x3	x4	x5	x6
x1	1.0000	0.1600	0.5556	0.6502	0.0603	0.0499
x2	0.1600	1.0000	0.3899	0.2399	0.1716	-0.1531
x3	0.5556	0.3899	1.0000	0.1622	-0.1032	-0.2386
x4	0.6502	0.2399	0.1622	1.0000	0.6564	0.4261
x5	0.0603	0.1716	-0.1032	0.6564	1.0000	0.6969
x6	0.0499	-0.1531	-0.2386	0.4261	0.6969	1.0000

Correlations Among the WITH Variables

	y1	y2	y3	y4
y1	1.0000	0.2509	0.7107	0.2099
y2	0.2509	1.0000	0.2253	-0.0755
y3	0.7107	0.2253	1.0000	0.3880
y4	0.2099	-0.0755	0.3880	1.0000

Correlations Between the VAR Variables and the WITH Variables

	y1	y2	y3	y4
x1	0.8374	0.6350	0.4819	-0.0595
x2	0.7592	0.2459	0.6182	0.1407
x3	0.5998	0.3181	0.8649	0.3969
x4	0.1295	0.9088	0.0682	-0.0871
x5	-0.0561	0.5722	-0.0215	-0.0103
x6	-0.2286	0.6266	-0.3490	-0.2965

变量：y1＝劳动生产率；y2＝市场占有率；y3＝居民人均收入；y4＝经济增长率；x1＝对外设施指数；x2＝对内设施指数；x3＝百人电话数；x4＝技术设施指数；x5＝文化设施指数；x6＝卫生设施指数。

城市基础设施指标变量相关性相对较强的有：x1与x3，x4；x4与x5；x5与x6，相关系数都超过了0.5。

城市竞争力变量相关性相对较强的有：y1与y3。

城市基础设施指标和城市竞争力变量相关性相对较强的有：y1与x2，x3；y2与x1，x4，x5，x6；y3与x2，x3；y4与x3。

（2）典型相关系数、特征根及多变量检验。如输出结果8—3所示。

输出结果 8—3

The CANCORR Procedure

Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.960103	.	0.017941	0.921797
2	0.949937	.	0.022395	0.902381
3	0.646990	0.557318	0.133383	0.418598
4	0.357136	0.273201	0.200155	0.127546

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)

Test of H0: The canonical correlations in the current row and all that follow are zero

	Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	11.7873	2.5433	0.5383	0.5383	0.00387236	5.89	24	36.096	<.0001
2	3.2439	0.5239	0.4221	0.9604	0.04951694	4.04	15	30.768	0.0005
3	0.7200	0.5738	0.0329	0.9933	0.50724765	1.21	8	24	0.3337
4	0.1462		0.0067	1.0000	0.87245371	0.63	3	13	0.6065

Multivariate Statistics and F Approximations

S=4 M=0.5 N=4

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00387236	5.89	24	36.096	<.0001
Pillai's Trace	2.37032110	3.15	24	52	0.0003
Hotelling-Lawley Trace	21.89735334	8.32	24	16.316	<.0001
Roy's Greatest Root	11.78726242	25.54	6	13	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

样本第一组典型方程 V1 和 W1 的典型相关系数为 0.960 103；第二组为 0.949 937。

注意：这些典型相关系数是越来越小的。

从 4 组典型方程的特征根看，第一、二组的特征根明显较大，其贡献分别为 53.83%，42.21%，累计贡献为 96.04%，足以代表变量组间的相关。

从典型方程的显著性检验的 P 值看，在 0.001 甚至更小的显著性水平上，第一、二组典型方程显著，表明能够用城市基础设施变量组来解释城市竞争力变量组。

（3）基础设施和竞争力变量在 4 个典型方程中的标准化系数。如输出结果 8—4 所示。

输出结果 8—4

The CANCORR Procedure

Canonical Correlation Analysis

Standardized Canonical Coefficients for the VAR Variables

		infrastructure1	infrastructure2	infrastructure3	infrastructure4
x1	对外设施指数	0.1535	0.2134	-0.6966	-1.6312
x2	对内设施指数	0.3423	0.2637	-1.0577	0.2266
x3	百人电话数	0.4913	0.3953	0.9142	0.4179
x4	技术设施指数	0.3372	-0.8690	0.4921	1.3411
x5	文化设施指数	0.1149	0.2429	0.6687	-0.5082
x6	卫生设施指数	0.1419	-0.3856	-0.5884	-0.2076

Standardized Canonical Coefficients for the WITH Variables

		competitive1	competitive2	competitive3	competitive4
y1	劳动生产率	0.1395	0.1322	-1.1342	0.8639
y2	市场占有率	0.7185	-0.7361	0.2058	0.0628
y3	居民人均收入	0.4270	0.7720	0.6261	-1.0788
y4	经济增长率	0.0285	0.0059	0.6004	0.9294

由上表可知，典型相关模型为：

第一组典型模型

$$\text{infrastructure1}=0.153\,5x1+0.342\,3x2+0.491\,3x3+0.337\,2x4+0.114\,9x5+0.141\,9x6$$

$$\text{competitive1}=0.139\,5y1+0.718\,5y2+0.427\,0y3+0.028\,5y4$$

第二组典型模型

$$\text{infrastructure2}=0.213\,4x1+0.263\,7x2+0.395\,3x3-0.869\,0x4+0.242\,9x5-0.385\,6x6$$

$$\text{competitive2}=0.132\,2y1-0.736\,1y2+0.772\,0y3+0.005\,9y4$$

特别说明：标准化系数的性质与回归分析系数类似，但由于受到变量间相关的影响，并不能作为解释各个变量在典型方程上相对重要性的依据，但很多教材在这方面都存在误解。以标准化系数的大小来解释变量的相对重要性肯定是不妥的，除非各变量间不相关，而这又与典型相关分析相悖。

另外需注意，SPSS 的输出结果与 SAS 的结果，相关系数的符号恰好相反。

（4）各变量的典型载荷与交叉典型载荷。如输出结果 8—5 所示。

输出结果 8—5

The CANCORR Procedure

Canonical Structure

Correlations Between the VAR Variables and Their Canonical Variables

		infrastructure1	infrastructure2	infrastructure3	infrastructure4
x1	对外设施指数	0.7145	-0.0945	-0.0269	-0.5317
x2	对内设施指数	0.6373	0.3442	-0.4838	0.3949
x3	百人电话数	0.7190	0.5426	0.2860	-0.0806
x4	技术设施指数	0.7232	-0.6320	0.0551	0.0314
x5	文化设施指数	0.4102	-0.4688	0.2500	0.0032
x6	卫生设施指数	0.1968	-0.7252	-0.0438	-0.1756

Correlations Between the WITH Variables and Their Canonical Variables

		competitive1	competitive2	competitive3	competitive4
y1	劳动生产率	0.6292	0.4974	-0.5116	0.3080
y2	市场占有率	0.8475	-0.5295	0.0169	-0.0336
y3	居民人均收入	0.6991	0.7024	0.0993	-0.0900
y4	经济增长率	0.1693	0.3887	0.5897	0.6874

Correlations Between the VAR Variables and the Canonical Variables of the WITH Variables

		competitive1	competitive2	competitive3	competitive4
x1	对外设施指数	0.6860	-0.0887	-0.0174	-0.1899
x2	对内设施指数	0.6119	0.3270	-0.3169	0.1410
x3	百人电话数	0.6903	0.5154	0.1721	-0.0288
x4	技术设施指数	0.6944	-0.6004	0.0356	0.0112
x5	文化设施指数	0.3938	-0.4453	0.1617	0.0012
x6	卫生设施指数	0.1889	-0.6889	-0.0283	-0.0627

Correlations Between the WITH Variables and the Canonical Variables of the VAR Variables

		infrastructure1	infrastructure2	infrastructure3	infrastructure4
y1	劳动生产率	0.6041	0.4725	-0.3310	0.1100
y2	市场占有率	0.8137	-0.5030	0.0109	-0.0120
y3	居民人均收入	0.6712	0.6672	0.0642	-0.0322
y4	经济增长率	0.1625	0.3693	0.3815	0.2455

x1 至 x4 都与基础设施组的第一典型变量 infrastructure1 高度相关，相关系数均在 0.6 以上，说明对外基础设施、对内基础设施、百人电话数、技术设施指数在反映城市基础设施水平方面占主导地位。同时 x1 至 x4 与城市竞争力变量组的第一典型变量 competitive1 也高度相关，说明上述城市基础设施指标是反映城市竞争力大小的主要因素。x5 和 x6 虽然对城市竞争力的影响相对较小，但其影响效应是正向的，这也与现实相符合。

y1 至 y3 与竞争力变量组的第一典型变量 competitive1 高度相关，相关系数均在 0.6 以上，说明劳动生产率、市场占有率、居民人均收入在反映城市竞争力水平方面占主导地位。同时，y1 至 y3 与城市基础设施变量组的第一典型变量 infrastructure1 也高度相关，体现了基础设施变量的主要因素对城市竞争力的本质影响，与指标的实际经济联系相吻合。

对第二组典型变量的结构解释与此类似，但由于第二组典型变量反映的信息量相对较小，故分析时主要看第一组典型变量。

抑制变量：如果出现变量的典型系数符号与典型相关系数符号相反的情况，则称此变量为抑制变量。由（4）中的典型负荷和（3）中基础设施变量的典型相关系数对比可知，对于 infrastructure2 来讲，x1 与 x5 为抑制变量。对于 infrastructure1，没有抑制变量（见表 8—3）。

表 8—3

变量	x1	x2	x3	x4	x5	x6
相关系数	−	+	+	−	−	−
典型系数	+	+	+	−	+	−

（5）典型冗余分析。如输出结果 8—6 所示。

输出结果 8—6

The CANCORR Procedure

Canonical Redundancy Analysis

Canonical Variable Number	Raw Variance of the VAR Variables Explained by Their Own Canonical Variables: Proportion	Cumulative Proportion	Canonical R-Square	Raw Variance of the VAR Variables Explained by The Opposite Canonical Variables: Proportion	Cumulative Proportion
1	0.5159	0.5159	0.9218	0.4755	0.4755
2	0.2944	0.8103	0.9024	0.2657	0.7412
3	0.0706	0.8809	0.4186	0.0296	0.7708
4	0.0067	0.8876	0.1275	0.0009	0.7716

Canonical Variable Number	Raw Variance of the WITH Variables Explained by Their Own Canonical Variables: Proportion	Cumulative Proportion	Canonical R-Square	Raw Variance of the WITH Variables Explained by The Opposite Canonical Variables: Proportion	Cumulative Proportion
1	0.4026	0.4026	0.9218	0.3711	0.3711
2	0.2651	0.6678	0.9024	0.2392	0.6104
3	0.2436	0.9114	0.4186	0.1020	0.7124
4	0.0886	1.0000	0.1275	0.0113	0.7237

SAS 输出了第一组变量的变异量被 4 个 infrastructure 方程解释的百分比及与 4 个 competitive 典型方程的冗余系数。

由上表看，对于 VAR 组，infrastructure1 解释了本组变量变异值的 51.59%；而 competitive1 能解释 VAR 组变量变异的 47.55%。

注意：原始变量的冗余分析表只有在进行典型分析时，是以方差协方差矩阵取代相关矩阵，进行案例分析才有意义。否则，应看标准化的典型系数的相关报表。

（6）标准化变量的冗余分析。如输出结果 8—7 所示。

输出结果 8—7

The CANCORR Procedure

Canonical Redundancy Analysis

Canonical Variable Number	Standardized Variance of the VAR Variables Explained by Their Own Canonical Variables: Proportion	Cumulative Proportion	Canonical R-Square	Standardized Variance of the VAR Variables Explained by The Opposite Canonical Variables: Proportion	Cumulative Proportion
1	0.3606	0.3606	0.9218	0.3324	0.3324
2	0.2612	0.6218	0.9024	0.2357	0.5681
3	0.0631	0.6849	0.4186	0.0264	0.5945
4	0.0795	0.7644	0.1275	0.0101	0.6046

Canonical Variable Number	Standardized Variance of the WITH Variables Explained by Their Own Canonical Variables: Proportion	Cumulative Proportion	Canonical R-Square	Standardized Variance of the WITH Variables Explained by The Opposite Canonical Variables: Proportion	Cumulative Proportion
1	0.4079	0.4079	0.9218	0.3760	0.3760
2	0.2930	0.7009	0.9024	0.2644	0.6404
3	0.1549	0.8558	0.4186	0.0648	0.7053
4	0.1442	1.0000	0.1275	0.0184	0.7237

由上表看，对于基础设施变量组，infrastructure1 解释了本组变量变异值的 36.06%；而 competitive1 也能解释本组变量变异的 33.24%。infrastructure2 解释了本组变量变异值的 26.12%；而 competitive2 也能解释本组变量变异的 23.57%。其累计解释程度分别为 62.18%和 56.81%。

同理，对于城市竞争力变量组，两个典型变量对竞争力变量变异的累计解释程度分别为 70.09%和 64.04%，说明基础设施变量组和竞争力变量组相互解释能力较强。关系：第二典型冗余＝第一典型冗余×典型相关系数的平方。

例 8—3

为了分析影响生猪养殖的原因，我们选取以下代表生猪生产的主要指标：Y1：肉猪出栏头数（万头）；Y2：生猪年底存栏头数（万头）；Y3：猪肉产量（万吨）；Y4：出口活猪数量（万头）。对生猪生产有影响的指标有：X1：猪（毛重）生产价格指数（1977 年为 100）；X2：粮食产量（万吨）；X3：粮食零售价格指数（1977 年为 100）；X4：农村居民人均纯收入（元）；X5：乡村总人口数（万人）；X6：全国人均猪肉消费量（斤）。

数据读者可在网站 www.ruc-6sigma.com《应用多元统计分析》第 14 章下载。

注：指标 X1 在《中国统计年鉴》里，2002 年以前统一使用的是肥猪收购价格指数，但 2002 年以后改为猪（毛重）生产价格指数，两个阶段使用了不同的统计标准，但不影响问题的分析。

资料来源：中经网统计数据库、万方中国年鉴资源全文数据库、《中国统计年鉴（2008）》。

使用 SAS/STAT 软件中的 CANCORR 过程来完成典型相关分析。样本程序如表 8—4 所示。

表 8—4

```
data pig;
input X1-X6 Y1-Y4;
cards;
100.30   30476.50   101.30   133.60   79014.00   15.34   16110.00   30129.00
975.30   246.28
…   …   …   …
run;
proc cancorr data = pig all;
var X1-X6; with Y1-Y4;
```

data 步创建了生猪养殖数据的 SAS 数据集（名为 pig），它有 30 个观测、10 个变量。

proc cancorr 表示调用典型相关程序，选项 all 表示输出所有的结果。

var 语句列出第一组变量的名字，with 列出第二组变量的名字。部分计算结果如输出结果 8—8 所示。

输出结果 8—8

Correlations Among the VAR Variables

	X1	X2	X3	X4	X5	X6
X1	1.0000	0.8805	0.9633	0.8756	−0.0982	0.5379
X2	0.8805	1.0000	0.8613	0.7915	0.0300	0.6921
X3	0.9633	0.8613	1.0000	0.9390	−0.2639	0.5256
X4	0.8756	0.7915	0.9390	1.0000	−0.5262	0.6036
X5	−0.0982	0.0300	−0.2639	−0.5262	1.0000	−0.1327
X6	0.5379	0.6921	0.5256	0.6036	−0.1327	1.0000

Correlations Among the WITH Variables

	Y1	Y2	Y3	Y4
Y1	1.0000	0.9516	0.9982	−0.7665
Y2	0.9516	1.0000	0.9580	−0.6202
Y3	0.9982	0.9580	1.0000	−0.7369
Y4	−0.7665	−0.6202	−0.7369	1.0000

Correlations Between the VAR Variables and the WITH Variables

	Y1	Y2	Y3	Y4
X1	0.9065	0.9057	0.9156	−0.6423
X2	0.8588	0.8642	0.8780	−0.4928
X3	0.9478	0.8972	0.9466	−0.7745
X4	0.9511	0.8545	0.9409	−0.8579
X5	−0.3192	−0.1170	−0.2748	0.6768
X6	0.6887	0.6599	0.7011	−0.2556

The CANCORR Procedure Canonical Correlation Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.987904	0.984801	0.004465	0.975955
2	0.916487	0.899610	0.029721	0.839948
3	0.762148	0.736296	0.077830	0.580870
4	0.244095	0.123398	0.174631	0.059583

Test of H0: The canonical correlations in the current row and all that follow are zero

Eigenvalues of Inv(E) * H = CanRsq/(1 − CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr>F
1	40.5891	35.3411	0.8584	0.8584	0.00151688	16.05	24	70.982	<.0001
2	5.2480	3.8621	0.1110	0.9694	0.06308546	6.70	15	58.373	<.0001
3	1.3859	1.3225	0.0293	0.9987	0.39415704	3.26	8	44	0.0053
4	0.0634		0.0013	1.0000	0.94041748	0.49	3	23	0.6955

输出结果 8—8 列出了两组变量各指标之间的相关系数。由结果可以看出，X1 与 X3、X1 与 X4 均有较高的相关系数，说明这几个指标在很大程度上都是影响生猪养殖的指标，且指标间包含的信息有重叠的部分。另外，Y1 与 Y2、Y1 与 Y3、Y2 与 Y3 之间的相关系数也较高，它们可以反映生猪养殖的情况。

输出结果 8—8 还给出了典型相关相关系数及各相关系数的检验。第一对典型变量的典型相关系数为 0.987 904，第二对典型变量的典型相关系数为 0.916 487。检验结果表明，两个典型系数均是显著的。因此，两组变量相关性的研究可以转化

为研究第一对和第二对典型变量的相关性。

输出结果 8—9 和输出结果 8—10 是典型系数和典型结构的分析。

输出结果 8—9

The CANCORR Procedure Canonical Correlation Analysis

Standardized Canonical Coefficients for the VAR Variables

	V1	V2	V3	V4
X1	-0.5619	0.8970	0.4774	-1.5071
X2	-0.2882	0.2028	0.2887	2.7474
X3	0.3725	-1.5061	2.8455	-0.8289
X4	1.5623	0.6221	-5.3347	-0.0088
X5	0.5460	0.9031	-1.9598	-0.6186
X6	0.1113	0.2346	1.5978	-0.6004

Standardized Canonical Coefficients for the WITH Variables

	W1	W2	W3	W4
Y1	-0.5005	-9.8843	20.0432	-14.3389
Y2	-0.1387	0.8085	-2.3337	-3.0278
Y3	1.5230	9.5616	-15.9830	16.8028
Y4	-0.1418	0.5254	2.3860	-0.4936

输出结果 8—10

The CANCORR Procedure Canonical Structure

Correlations Between the VAR Variables and Their Canonical Variables

	V1	V2	V3	V4
X1	0.9173	0.2070	-0.1466	-0.1564
X2	0.8679	0.3774	-0.0158	0.2653
X3	0.9644	0.0017	-0.0982	-0.0751
X4	0.9718	-0.1798	-0.0205	0.0309
X5	-0.3427	0.8602	-0.1539	-0.0849
X6	0.6759	0.3215	0.5903	0.1315

Correlations Between the WITH Variables and Their Canonical Variables

	W1	W2	W3	W4
Y1	0.9965	0.0271	0.0391	-0.0686
Y2	0.9320	0.2370	-0.0525	-0.2692
Y3	0.9950	0.0821	0.0312	-0.0477
Y4	-0.7944	0.5542	0.2484	-0.0071

Correlations Between the VAR Variables and the Canonical Variables of the WITH Variables

	W1	W2	W3	W4
X1	0.9062	0.1897	-0.1117	-0.0382
X2	0.8574	0.3459	-0.0120	0.0648
X3	0.9527	0.0016	-0.0749	-0.0183
X4	0.9600	-0.1648	-0.0156	0.0075
X5	-0.3385	0.7883	-0.1173	-0.0207
X6	0.6677	0.2947	0.4499	0.0321

Correlations Between the WITH Variables and the Canonical Variables of the VAR Variables

	V1	V2	V3	V4
Y1	0.9845	0.0248	0.0298	-0.0167
Y2	0.9207	0.2172	-0.0400	-0.0657
Y3	0.9830	0.0752	0.0238	-0.0116
Y4	-0.7848	0.5079	0.1893	-0.0017

前两对典型变量的线性组合为：

$$V1=-0.561\,9X1-0.288\,2X2+0.372\,5X3+1.562\,3X4+0.546\,0X5+0.111\,3X6$$

$$W1=-0.500\,5Y1-0.138\,7Y2+1.523\,0Y3-0.141\,8Y4$$

在第一对典型变量 V1 和 W1 中，V1 是对生猪养殖有影响的各因素的线性组合。可以看出，V1 主要代表 X1（猪（毛重）生产价格指数），X4（农村居民人均纯收入），X5（乡村总人口数）。W1 是对生猪养殖指标的线性组合。W1 主要代表 Y1（肉猪出栏头数），Y3（猪肉产量）。这说明肉猪出栏头数、猪肉产量与猪（毛重）生产价格指数、农村居民人均纯收入、乡村总人口数有较密切的关系。

$$V2=0.897\,0X1+0.202\,8X2-1.506\,1X3+0.622\,1X4+0.903\,1X5+0.234\,6X6$$

$$W2=-9.884\,3Y1+0.808\,5Y2+9.561\,6Y3+0.525\,4Y4$$

在第二对典型变量 V2 和 W2 中，在 V2 的线性组合中，X1（猪（毛重）生产价格指数），X3（粮食零售价格指数），X4（农村居民人均纯收入），X5（乡村总人口数）的载荷系数较大，不过 X3 在这里起负面作用。W2 的线性组合中，Y1（肉猪出栏头数），Y2（生猪年底存栏头数），Y3（猪肉产量）的载荷系数较大。这说明了猪（毛重）生产价格指数、粮食零售价格指数、农村居民人均纯收入、乡村总人口数是影响农民养猪的主要因素。从第二对典型变量中进一步看出，生猪年底存栏头数与猪（毛重）生产价格指数及粮食产量同方向增长，与粮价方向相反。从输出结果 8—10 可以看出，V1 与指标 X1，X2，X3，X4 有较强的相关性；W1 与指标 Y1，Y2，Y3 有较强的相关性。

这些情况表明，年底存栏头数是农民养猪积极性的反映。而农民养猪成本受粮食价格的制约，粮食产量下降，粮食价格升高，饲料必然上涨，农民养猪的成本投入增大，养猪利润较少，农民致富的经营目标当然转移，这是商品生产的必然规律，改革开放初期我国生猪养殖的波动已充分说明了这方面的问题。近几年政府相继出台生猪养殖的各项好政策，落实对养猪户的补贴政策，以维护生猪养殖的稳定发展。

输出结果 8—11 和输出结果 8—12 是典型冗余分析。输出结果说明了典型函数可以解释的变量方差的比例。第一典型变量 V1 可以解释组内变差的 67.39%，并解释另一组（生猪养殖指标）变差的 65.77%；典型变量 W1 可以解释组内变差的 87.07%，并解释另一组（生猪养殖的影响因素指标）变差的 84.98%。可见，第一对典型变量都能较全面地预测另一组变量。而第二和第三对典型变量实际上都没有给出较充分的信息。

输出结果 8—12 是典型冗余分析的复相关系数分析，类似于回归分析中的 R^2，读者可尝试给出相应的解释。

输出结果 8—11

The CANCORR Procedure Canonical Redundancy Analysis

Canonical Variable Number	Standardized Variance of the VAR Variables Explained by Their Own Canonical Variables: Proportion	Cumulative Proportion	Canonical R-Square	Standardized Variance of the VAR Variables Explained by The Opposite Canonical Variables: Proportion	Cumulative Proportion
1	0.6739	0.6739	0.9760	0.6577	0.6577
2	0.1768	0.8507	0.8399	0.1485	0.8062
3	0.0673	0.9180	0.5809	0.0391	0.8453
4	0.0210	0.9390	0.0596	0.0013	0.8466

Canonical Variable Number	Standardized Variance of the WITH Variables Explained by Their Own Canonical Variables: Proportion	Cumulative Proportion	Canonical R-Square	Standardized Variance of the WITH Variables Explained by The Opposite Canonical Variables: Proportion	Cumulative Proportion
1	0.8707	0.8707	0.9760	0.8498	0.8498
2	0.0927	0.9634	0.8399	0.0779	0.9276
3	0.0167	0.9801	0.5809	0.0097	0.9373
4	0.0199	1.0000	0.0596	0.0012	0.9385

输出结果 8—12

The CANCORR Procedure Canonical Redundancy Analysis

Squared Multiple Correlations Between the VAR Variables and the First M Canonical Variables of the WITH Variables

M	1	2	3	4
X1	0.8213	0.8573	0.8697	0.8712
X2	0.7351	0.8547	0.8549	0.8590
X3	0.9077	0.9077	0.9133	0.9137
X4	0.9217	0.9488	0.9491	0.9491
X5	0.1146	0.7361	0.7498	0.7502
X6	0.4458	0.5327	0.7350	0.7361

Squared Multiple Correlations Between the WITH Variables and the First M Canonical Variables of the VAR Variables

M	1	2	3	4
Y1	0.9692	0.9698	0.9707	0.9709
Y2	0.8477	0.8949	0.8965	0.9008
Y3	0.9662	0.9719	0.9724	0.9726
Y4	0.6159	0.8739	0.9098	0.9098

利用 SPSS 软件也可以进行典型相关分析。有关 SPSS 典型相关分析的输出，本书不再详细讨论，有兴趣的读者请参考相关指导书。

□ 参考文献

[1] 王国梁，何晓群. 多变量经济数据统计分析. 西安：陕西科学出版

社，1993

[2] 张尧庭，方开泰. 多元统计分析引论. 北京：科学出版社，1982

[3] 方开泰. 实用多元统计分析. 上海：华东师范大学出版社，1989

[4] 王学仁，王松桂. 实用多元统计分析. 上海：上海科学技术出版社，1990

[5] Joseph F. Hair，Rolph E. Anderson，Ronald L. Tatham，William C. Black. *Multivariate Data Analysis*. Fifth Edition. Prentice-Hall，1998

□ 思考与练习

1. 试述典型相关分析的统计思想及该方法在研究实际问题中的作用。
2. 典型相关分析中的冗余度有什么作用？
3. 典型变量的解释有什么具体方法？实际意义是什么？
4. 运用 SPSS 或 SAS 软件试对一个实际问题的研究应用典型相关分析。

第9章

Chapter 9 定性数据的建模分析

学习目标

1. 掌握对数线性模型的基本原理；
2. 掌握对数线性模型的建模方法；
3. 掌握如何解释 Logistic 回归的分析结果；
4. 理解判别分析与 Logistic 回归相比的优缺点；
5. 掌握如何通过 SPSS 软件实现 Logistic 回归。

前面讨论过有关定性数据的列联表分析，对数线性模型是进一步用于离散型数据或整理成列联表格式的数据的统计分析工具。它可以把方差分析和线性模型的一些方法应用到对交叉列联表的分析中，从而对定性变量间的关系做进一步的描述和分析。列联表分析无法系统地评价变量间的联系，也无法估计变量间交互作用的大小，而对数线性模型是处理这些问题的最佳方法。当被解释变量是非度量变量时，判别分析是合适的。然而当被解释变量只有两组时，Logistic 回归由于多种原因更受欢迎。首先，判别分析依赖于严格的多元正态性和相等协方差阵的假设，这在很多情况下是达不到的。Logistic 回归没有类似的假设，而且这些假设不满足时，结果非常稳定。其次，即使满足假定，许多研究者仍偏好 Logistic 回归，因为它类似于回归分析。两者都有直接的统计检验，都能包含非线性效果和大范围的诊断。再者，Logistic 回归对于自变量没有要求，度量变量或者非度量变量都可以进行回归。这样，很多情况下，Logistic 回归等同于两组的判别分析，而且可能更加实用。本章仅介绍定性数据建模的对数线性模型和 Logistic 回归方法。

9.1　对数线性模型基本理论和方法

本节将利用 2×2 维的交叉列联表来说明对数线性模型的基本理论和方法，同时利用 SPSS 软件对真实的经济定性数据做分析。

从 2×2 维的交叉列联表的概率表，介绍对数线性模型的基本理论和方法（见表 9—1 和表 9—2）。

表 9—1　　频数表

A \ B	B	$\bar{B}$	$\sum_j$
A	n_{11}	n_{12}	$n_{1\cdot}$
$\bar{A}$	n_{21}	n_{22}	$n_{2\cdot}$
$\sum_i$	$n_{\cdot 1}$	$n_{\cdot 2}$	n

表 9—2　　概率表

A \ B	B	$\bar{B}$	$\sum_j$
A	p_{11}	p_{12}	$p_{1\cdot}$
$\bar{A}$	p_{21}	p_{22}	$p_{2\cdot}$
$\sum_i$	$p_{\cdot 1}$	$p_{\cdot 2}$	1

在对数线性模型分析中，要先将概率取对数，再分解处理，用公式表示如下：

$$\begin{aligned}\eta_{ij} &= \ln p_{ij} \\ &= \ln\left(p_{i\cdot}\, p_{\cdot j}\frac{p_{ij}}{p_{i\cdot}\, p_{\cdot j}}\right) \\ &= \ln p_{i\cdot} + \ln p_{\cdot j} + \ln\frac{p_{ij}}{p_{i\cdot}\, p_{\cdot j}}, \quad i,j=1,2\end{aligned} \tag{9.1}$$

若把上式中的 $\ln p_{i\cdot}$，$\ln p_{\cdot j}$，$\ln\dfrac{p_{ij}}{p_{i\cdot}\, p_{\cdot j}}$ 分别记为 A_i，B_j 和 $(AB)_{ij}$，则上式可写成

$$\eta_{ij} = A_i + B_j + (AB)_{ij}$$

该式的结构与有交互效应且各水平均为 2 的双因素方差分析模型的结构相似，因此模仿方差分析，可以有如下关系式：

$$\eta_{i\cdot} = \sum_{j=1}^{2}\eta_{ij}, \quad \eta_{\cdot j} = \sum_{i=1}^{2}\eta_{ij}, \quad \eta_{\cdot\cdot} = \sum_{i=1}^{2}\sum_{j=1}^{2}\eta_{ij}$$

对上面三式各取其平均数为：

$$\bar{\eta}_{i\cdot} = \frac{1}{2}\eta_{i\cdot}, \quad \bar{\eta}_{\cdot j} = \frac{1}{2}\eta_{\cdot j}, \quad \bar{\eta}_{\cdot\cdot} = \frac{1}{4}\eta_{\cdot\cdot}$$

若记

$$\begin{cases}\alpha_i = \bar{\eta}_{i\cdot} - \bar{\eta}_{\cdot\cdot} \\ \beta_j = \bar{\eta}_{\cdot j} - \bar{\eta}_{\cdot\cdot} \\ \gamma_{ij} = \eta_{ij} - \bar{\eta}_{i\cdot} - \bar{\eta}_{\cdot j} + \bar{\eta}_{\cdot\cdot}\end{cases}$$

则　　$\gamma_{ij} = \eta_{ij} - \bar{\eta}_{i\cdot} - \bar{\eta}_{\cdot j} + \bar{\eta}_{\cdot\cdot}$

$$=\eta_{ij}-(\bar{\eta}_{i\cdot}-\bar{\eta}_{\cdot\cdot})-(\bar{\eta}_{\cdot j}-\bar{\eta}_{\cdot\cdot})-\bar{\eta}_{\cdot\cdot}$$
$$=\eta_{ij}-\alpha_i-\beta_j-\bar{\eta}_{\cdot\cdot}$$

移项，可得与有交互效应的双因素方差分析数学模型极为相似的关系式：

$$\begin{cases}\eta_{ij}=\bar{\eta}_{\cdot\cdot}+\alpha_i+\beta_j+\gamma_{ij}\\ \sum_{i=1}^{2}\alpha_i=\sum_{j=1}^{2}\beta_j=\sum_{i=1}^{2}\gamma_{ij}=\sum_{j=1}^{2}\gamma_{ij}=0,\quad i=1,2;\ j=1,2\end{cases}\tag{9.2}$$

为与方差分析保持一致，可称 α_i，β_j 分别是 A，B 因素的主效应，γ_{ij} 是 A，B 因素的交互效应。到此，定性数据的数据变换和变换后的模型关系已清楚地完成，接下来就是对模型的参数估计及检验。这里主要是估计 γ_{ij} 的值，根据 γ_{ij} 值的正负和相对大小，可以判断 A 因素的第 i 个水平与 B 因素的第 j 个水平间的交互效应。若 $\gamma_{ij}>0$，表明二者存在正效应；若 $\gamma_{ij}<0$，则存在负效应；当 γ_{ij} 均为 0 时，A，B 因素相互独立。若 γ_{ij} 均为 0，模型称为非饱和模型（因素间相互独立），否则为饱和模型（因素间有交互效应）。

在实际分析中，概率表中各项值以交叉列联表计算得到的频率表的对应项为无偏估计值。公式为：

$$\hat{p}_{ij}=\frac{n_{ij}}{n},\quad \hat{p}_{i\cdot}=\frac{n_{i\cdot}}{n},\quad \hat{p}_{\cdot j}=\frac{n_{\cdot j}}{n}$$

将其代入 $\eta_{ij}=\ln p_{ij}$ 等式，有

$$\hat{\eta}_{ij}=\ln\hat{p}_{ij}=\ln n_{ij}-\ln n$$
$$\hat{\bar{\eta}}_{i\cdot}=\frac{1}{2}\sum_{j=1}^{2}\eta_{ij}=\frac{1}{2}\sum_{j=1}^{2}(\ln\frac{n_{ij}}{n})=\frac{1}{2}\sum_{j=1}^{2}(\ln n_{ij})-\ln n$$
$$\hat{\bar{\eta}}_{\cdot j}=\frac{1}{2}\sum_{i=1}^{2}\eta_{ij}=\frac{1}{2}\sum_{i=1}^{2}(\ln\frac{n_{ij}}{n})=\frac{1}{2}\sum_{i=1}^{2}(\ln n_{ij})-\ln n$$
$$\hat{\bar{\eta}}_{\cdot\cdot}=\frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}\eta_{ij}=\frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}(\ln\frac{n_{ij}}{n})=\frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}(\ln n_{ij})-\ln n$$

将以上三式代入公式

$$\begin{aligned}\hat{\gamma}_{ij}&=\hat{\eta}_{ij}-\hat{\bar{\eta}}_{i\cdot}-\hat{\bar{\eta}}_{\cdot j}+\hat{\bar{\eta}}_{\cdot\cdot}\\&=\ln n_{ij}-\frac{1}{2}\sum_{j=1}^{2}(\ln n_{ij})-\frac{1}{2}\sum_{i=1}^{2}(\ln n_{ij})+\frac{1}{4}\sum_{i=1}^{2}\sum_{j=1}^{2}(\ln n_{ij})\end{aligned}\tag{9.3}$$

即可得 γ_{ij} 的估计值 $\hat{\gamma}_{ij}$。实际分析中，二维数据表中并非每个因素都是双水平的，调整公式中的 i，j 的取值上限即可。

9.2 对数线性模型的上机实现

可以使用 SPSS 软件来实现对数线性模型分析。这里举一个例子，是 3×2 维的

交叉列联表的分析。我们用 SPSS 软件中的 Loglinear 模块实现分析。

例 9—1

某企业想了解顾客对其产品是否满意，同时还想了解不同收入的人群对其产品的满意度是否相同。在随机发放的 1 000 份问卷中收回有效问卷 792 份，根据收入高低和满意回答的交叉分组数据如表 9—3 所示。

表 9—3

	满意	不满意	合计
高	53	38	91
中	434	108	542
低	111	48	159
合计	598	194	792

首先要准备数据，上面的交叉列联表的数据要输入 SPSS 的表格里，具体如表 9—4 和图 9—1 所示。

表 9—4

频数	收入情况	满意情况
53	1	1
434	2	1
111	3	1
38	1	2
108	2	2
48	3	2

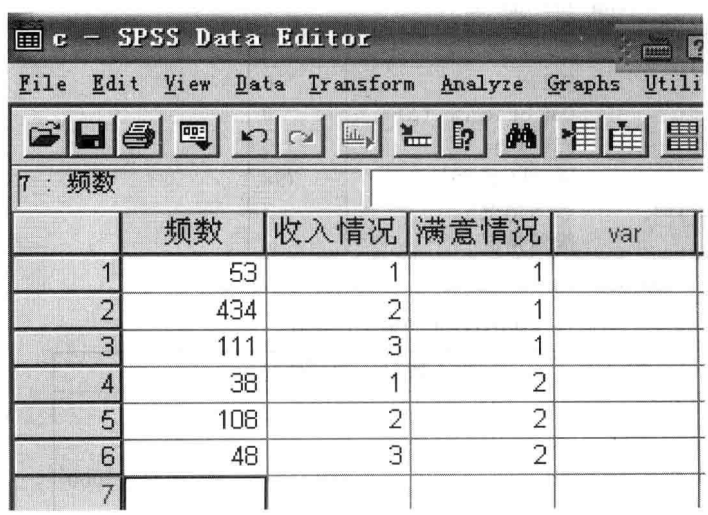


图 9—1

按上面的形式输入数据后，还不能马上进行对数线性模型分析，必须先激活频数，即让频数有效。具体步骤是：使用 SPSS 软件，从主菜单中，以 Data→Weight Cases…顺序，打开 Weight Cases 对话框，选中 Weight cases by 单选框，从变量列表中选出“频数”变量，点击▷按钮，使之进入 Frequency Variable 框，然后点击 OK 按钮，回到数据表格，这时分析前的准备工作就完成了。这一步很重要，如果频数没有被激活，对数线性模型的模块仍会执行命令，但得出的结果是错误的，所以使用时一定要小心。

数据准备工作完成后，就可以进行下一步的分析了。从主菜单中，按 Analyze→Loglinear→Model Selection…的流程可打开 Model Selection Loglinear Analysis 对话框，从左侧变量栏里选中“收入情况”，点击▷按钮进入 Factor(s) 框，这时该框下面的 Define Range…按钮就会从灰色变为黑色，点击弹出 Loglinear Analysis：Define Range 对话框，可以定义变量的范围，即该变量的水平范围。本例中“收入情况”共有三种类型，代号分别是 1，2，3，所以在 Minimum 处键入 1，在 Maximum 处键入 3，点击 Continue 按钮，返回 Model Selection Loglinear Analysis 对话框。按同样方法，把“满意情况”变量选入，并定义其范围为 1，2；然后选中“频数”变量，点击▷按钮使之进入 CellWeights 框。最后，点击 Options 按钮，进入 Loglinear Analysis：Options 对话框，选择 Display for Saturated Model 栏下的 Parameter estimates 项，点击 Continue 按钮返回 Model Selection Loglinear Analysis 对话框，其他选项保持默认值，最后点击 OK 按钮，即可完成分析步骤。

得到的主要输出结果及解释如下。见输出结果 9—1。

输出结果 9—1

Hierarchical Loglinear Analysis

For Design 1, .500 has been added to all observed cells for this saturated model. This value may be changed by using the CRITERIA = DELTA subcommand.

Data Information

–		N
Cases	Valid	6
	Out of Range (a)	0
	Missing	0
	Weighted Valid	792
Categories	收入情况	3
	满意情况	2

a. Cases rejected because of out of range factor values.

Convergence Information

Generating Class	收入情况 * 满意情况
Number of Iterations	1
Max. Difference between Observed and Fitted Marginals	.000
Convergence Criterion	188.356

输出结果 9—1 首先提示在饱和模型中采用的 Delta 校正值为 0.5，该数值可在对话框中更改，接着列出了模型的一般信息，显示系统对 792 例资料进行分析，这 792 例资料可分为 6 类（3×2）。模型中共有两个分类变量，其中“收入情况”变量为 3 水平，“满意情况”变量为 2 水平；分析的效应除了两个分类变量，还有两者的交互作用（收入情况 * 满意情况）。系统经一次迭代后，即达到相邻两次估计之差不大于规定的 0.001。见输出结果 9—2。

输出结果 9—2　　Cell Counts and Residuals

收入情况	满意情况	Observed		Expected		Residuals	Std. Residuals
		Count (a)		%		Count	%
高	满意	53.500	6.8%	53.500	6.8%	.000	.000
	不满意	38.500	4.9%	38.500	4.9%	.000	.000
中	满意	434.500	54.9%	434.500	54.9%	.000	.000
	不满意	108.500	13.7%	108.500	13.7%	.000	.000
低	满意	111.500	14.1%	111.500	14.1%	.000	.000
	不满意	48.500	6.1%	48.500	6.1%	.000	.000

a. For saturated models, .500 has been added to all observed cells.

由于本例对模型采用系统默认的饱和模型，因而实际例数（Observed）与期望例数（Expected）相同，进而残差（Residuals）和标准化残差（Std. Residuals）均为 0。

输出结果 9—3 是模型的拟合优度检验，由于是饱和模型，所以 χ^2 值和自由度均为 0，而 Sig. 无信息显示。

输出结果 9—3　　Goodness-of-Fit Tests

	Chi-Square	Df	Sig.
Likelihood Ratio	.000	0	.
Pearson	.000	0	.

输出结果 9—4 是对模型是否有高阶效应进行检验，原假设是高阶效应为零，即没有高阶效应。表中检验分为两部分。第一部分：K-way and Higher Order Effects (a)，是分别利用 Likelihood Ratio 方法和 Pearson 方法检验模型中 K 维交互作用以及 K 维以上交互作用是否显著。两种检验方法的结果均表明，应拒绝原假设，即一维交互作用以及一维交互以上（即主效应）均极其显著。第二部分：K-way Effects (b)，是检验模型中 K 维交互作用自身是否显著，同样有 Likelihood Ratio 方法和 Pearson 方法。结论与第一部分类似，但这里读者应注意到，由于第二部分是检验 K 维交互作用自身是否显著，因此在检验一维主效应时不再包含二维交互，因而其 χ^2 值减小，减小的值恰好为二维交互的值。

输出结果 9—4 **K-Way and Higher-Order Effects**

	K	df	Likelihood Ratio		Pearson		Number of Iterations
			Chi-Square	Sig.	Chi-Square	Sig.	
K-way and Higher	1	5	662.843	.000	866.318	.000	0
Order Effects (a)	2	2	22.087	.000	23.567	.000	2
K-way Effects (b)	1	3	640.756	.000	842.751	.000	0
	2	2	22.087	.000	23.567	.000	0

a. Tests that k-way and higher order effects are zero.

b. Tests that k-way effects are zero.

输出结果 9—5 是对模型参数的估计，以及对参数的检验结果。

输出结果 9—5 **Parameter Estimates**

Effect Parameter		Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
收入情况 * 满意情况	1	−.260	.078	−3.332	.001	−.413	−.107
	2	.269	.058	4.651	.000	.156	.382
收入情况	1	−.683	.078	−8.737	.000	−.836	−.530
	2	.883	.058	15.267	.000	.769	.996
满意情况	1	.425	.049	8.703	.000	.329	.520

由式（9.2），对数线性模型是固定模型效应的分析，应满足各效应和为零的约束条件，故根据上表结果可推得各参数为：

$$\alpha_{\text{高收入}} = -0.683$$

$$\alpha_{\text{中收入}} = 0.883$$

$$\alpha_{\text{低收入}} = 0 - (-0.683) - 0.883$$

$$= -0.200$$

$$\beta_{\text{满意}} = 0.425(\text{是满意情况的 1 水平})$$

$$\beta_{\text{不满意}} = -0.425$$

$$\gamma_{\text{高收入满意}} = -0.260$$

$$\gamma_{\text{中收入满意}} = 0.269$$

$$\gamma_{\text{低收入满意}} = 0 - (-0.260) - 0.269$$

$$= -0.009$$

$$\gamma_{\text{高收入不满意}} = 0.260$$

$$\gamma_{\text{中收入不满意}} = -0.269$$

$$\gamma_{\text{低收入不满意}} = 0.009$$

参数值为正，表示正效应；反之为负效应；零为无效应。分析提供的信息是：

（1）$\beta_{\text{满意}}$ 为正值，说明接受调查的多数顾客对其产品还是满意的。

（2）$\alpha_{\text{高收入}} < \alpha_{\text{低收入}} < \alpha_{\text{中收入}}$，说明各收入阶层的顾客对其产品的满意度是不同

的，其中，高收入的顾客满意度最低，而中等收入的顾客满意度最高。

（3）通过对企业顾客的收入情况和满意情况交互效应的研究，$\gamma_{\text{高收入满意}}$ 为负值，表示高收入对其对产品的满意度有负效应；$\gamma_{\text{中收入满意}}$ 为正值，表示中等收入对其对产品的满意度有正效应；同理，低收入对其对产品的满意程度也有负效应。该企业产品主要的消费阶层是中等收入者，同时中等收入者对其产品的满意度也最高。

9.3　Logistic 回归基本理论和方法

通常我们需要研究某一社会现象发生的概率 p 的大小，比如一个公司成功或失败的概率，以及讨论 p 的大小与哪些因素有关。但是，直接处理可能性数值 p 存在困难，一是 $0\leqslant p\leqslant 1$，因此 p 与自变量的关系难以用线性模型来描述；二是当 p 接近 0 或 1 时，p 值的微小变化用普通的方法难以发现和处理好。这时，不处理参数 p，而处理 p 的一个严格单调函数 $Q=Q(p)$ 就会方便得多。要求 $Q(p)$ 对在 $p=0$ 或者 $p=1$ 附近的微小变化很敏感，即 $\frac{\mathrm{d}Q}{\mathrm{d}p}$ 应与 $\frac{1}{p(1-p)}$ 成比例，于是令

$$Q=\ln\frac{p}{1-p}$$

将 p 换成 Q，这一变换称为 Logit 变换。从 Logit 变换可看出，当 p 从 0→1 时，Q 的值从 $-\infty\rightarrow+\infty$，因此 Q 的值在区间 $(-\infty,+\infty)$ 上变化。这一变换完全克服了一开始所提出的两点困难，在数据处理上带来很多方便。如果自变量的关系式是线性的、二次的或多项式的，通过普通的最小二乘就可以处理，然后从 p 与 Q 的反函数关系式中求出 p 与自变量的关系。例如 $Q=b'x$，则有 $p=\frac{\mathrm{e}^{b'x}}{1+\mathrm{e}^{b'x}}$，这就是 Logit 变换所带来的方便。

根据上述思想，当因变量是一个二元变量，只取 0 与 1 两个值时，因变量取 1 的概率 $p(y=1)$ 就是要研究的对象。如果有很多因素影响 y 的取值，这些因素就是自变量，记为 x_1，x_2，…，x_k，这些 x_i 中既有定性变量，也有定量变量。最重要的一个条件是：

$$\ln\frac{p}{1-p}=b_0+b_1x_1+\cdots+b_kx_k$$

即 $\ln\frac{E(y)}{1-E(y)}$ 是 x_1，x_2，…，x_k 的线性函数。满足上述条件的称为 Logistic 线性回归。由于上式所确定的模型相当于广义线性模型，因此可以系统地应用线性模型的方法，在处理时比较方便。

在判别分析中，是通过判别 Z 得分来预测所属类的，这就需要计算临界得分和

将观测归类。Logistic 回归完成这一目的颇似回归分析，不同于回归分析的地方在于它直接预测出了事件发生的概率。尽管这个概率值是个度量尺度，Logistic 回归与多元回归还是有很大的差异。概率值可以是 0～1 之间的任何值，但是预测值必须落入 0～1 区间。这样，Logistic 回归假定解释变量与被解释变量之间的关系类似于 S 形曲线。而且，不能从普通回归的角度来分析 Logistic 回归，因为这样做会违反几个假定。首先，离散变量的误差形式遵从贝努里分布，而不是正态分布，从而使得基于正态性假设的统计检验无效。其次，二值变量的方差不是常数，会造成异方差性。Logistic 回归是专门处理这些问题的。它的解释变量与被解释变量之间的独特关系使得在估计、评价拟合度和解释系数方面有不同的方法。

估计 Logistic 回归模型与估计多元回归模型的方法是不同的。多元回归采用最小二乘估计，使解释变量的真实值与预测值差异的平方和最小化。而 Logistic 变换的非线性特征使得在估计模型的时候采用极大似然估计的迭代方法，找到系数的“最可能”的估计。这样在计算整个模型拟合度的时候，就采用似然值而不是离差平方和。

Logistic 回归的另一个好处就是我们只需要知道一件事情（有没有购买、公司成功还是失败）是否发生了，然后再用二元值作为解释变量。从这个二元值中，程序预测出事件发生或者不发生的概率。如果预测概率大于 0.5，则预测发生，反之则不发生。需要注意的是，Logistic 回归系数的解释与多元回归的解释不同。程序计算出 Logistic 系数，比较事件发生与不发生的概率比。假定事件发生的概率为 p，优势比率可以表示为：

$$\frac{p}{1-p}=\mathrm{e}^{b_0+b_1x_1+\cdots+b_nx_n}$$

估计的系数（b_0，b_1，b_2，…，b_n）反映优势比率的变化。如果 b_i 是正的，它的反对数值（指数）一定大于 1，则优势比率会增大；反之，如果 b_i 是负的，则优势比率会减小。

前面已提到 Logistic 回归在估计系数时，采用的是极大似然估计法。就像多元回归中的残差平方和，Logistic 回归对模型拟合好坏通过似然比值来测量。（实际上是用－2 乘以似然比值的自然对数即－2Ln(似然值)，简记为－2LL。）一个好的模型应该有较小的－2LL。如果一个模型完全拟合，则似然比值为 1，这时－2LL 达到最小，为 0。Logistic 回归对于系数的检验采用的是与多元回归中 t 检验不同的统计量，称为 Wald 统计量。有关 Logistic 回归的参数估计和假设检验详见参考文献［1］。

9.3.1 分组数据的 Logistic 回归模型

针对 0—1 型因变量产生的问题，我们对回归模型应该做两个方面的改进。

第一，回归函数应该改用限制在［0，1］区间内的连续曲线，而不能再沿用直

线回归方程。限制在［0，1］区间内的连续曲线有很多，例如所有连续型随机变量的分布函数都符合要求，我们常用的是 Logistic 函数与正态分布函数。Logistic 函数的形式为：

$$f(x)=\frac{e^x}{1+e^x}=\frac{1}{1+e^{-x}} \tag{9.4}$$

Logistic 函数的中文名称是逻辑斯蒂函数，或简称逻辑函数。

这里给出两个 Logistic 函数的图形（见图 9—2、图 9—3）。

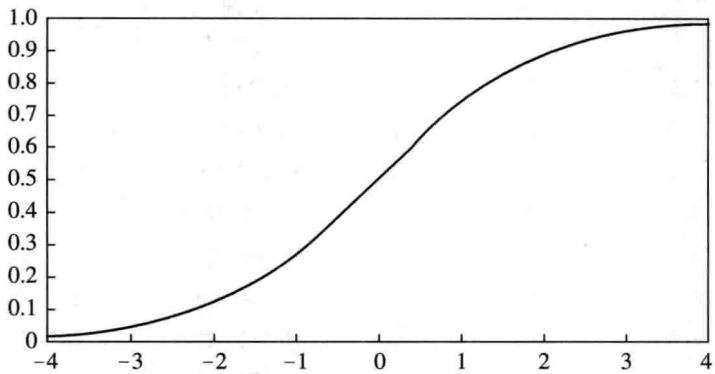


图 9—2　$f(x)=\frac{1}{1+e^{-x}}$的图形

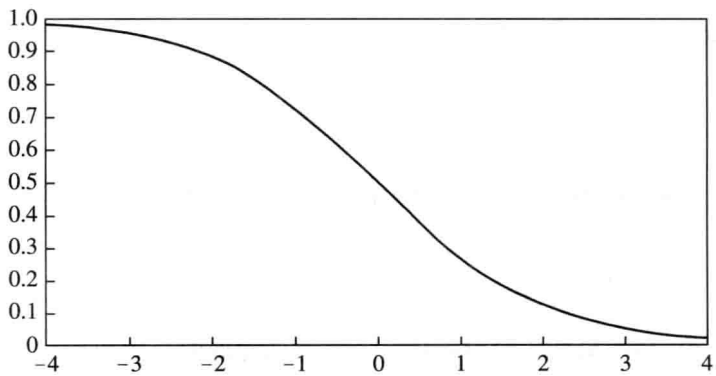


图 9—3　$f(x)=\frac{1}{1+e^{x}}$的图形

第二，因变量 y_i 本身只取 0，1 两个离散值，不适合直接作为回归模型中的因变量。由于回归函数 $E(y_i)=\pi_i=\beta_0+\beta_1 x_i$ 表示在自变量为 x_i 的条件下 y_i 的平均值，而 y_i 是 0—1 型随机变量，从而 $E(y_i)=\pi_i$ 就是在自变量为 x_i 的条件下 y_i 等于 1 的比例。这提示我们可以用 y_i 等于 1 的比例代替 y_i 本身作为因变量。

下面举例说明 Logistic 回归模型的应用。

例 9—2

在一次住房展销会上，与房地产商签订初步购房意向书的共有$n=313$名顾客。在随后的3个月内，只有部分顾客确实购买了房屋。购买了房屋的顾客记为1，没有购买房屋的顾客记为0。以顾客的年家庭收入（万元）为自变量x，对表9—5中的数据，建立Logistic回归模型。

表 9—5

序号	年家庭收入（万元） x	签订意向书人数 n_i	实际购房人数 m_i	实际购房比例 $p_i=m_i/n_i$	逻辑变换 $p_i'=\ln\left(\frac{p_i}{1-p_i}\right)$	权重 $w_i=n_ip_i(1-p_i)$
1	1.5	25	8	0.320 000	−0.753 77	5.440
2	2.5	32	13	0.406 250	−0.379 49	7.719
3	3.5	58	26	0.448 276	−0.207 64	14.345
4	4.5	52	22	0.423 077	−0.310 15	12.692
5	5.5	43	20	0.465 116	−0.139 76	10.698
6	6.5	39	22	0.564 103	0.257 829	9.590
7	7.5	28	16	0.571 429	0.287 682	6.857
8	8.5	21	12	0.571 429	0.287 682	5.143
9	9.5	15	10	0.666 667	0.693 147	3.333

Logistic回归方程为：

$$p_i=\frac{\exp(\beta_0+\beta_1x_i)}{1+\exp(\beta_0+\beta_1x_i)},\quad i=1,2,\cdots,c \tag{9.5}$$

式中，c为分组数据的组数。本例中，$c=9$。将以上回归方程做线性变换，令

$$p_i'=\ln\left(\frac{p_i}{1-p_i}\right) \tag{9.6}$$

式（9.6）的变换称为逻辑变换，变换后的线性回归模型为：

$$p_i'=\beta_0+\beta_1x_i+\varepsilon_i \tag{9.7}$$

式（9.7）是一个普通的一元线性回归模型。式（9.7）没有给出误差项的形式，我们认为其误差项的形式就是做线性变换所需要的形式。对表9—5中的数据，算出经验回归方程为：

$$\hat{p}'=-0.886+0.156x \tag{9.8}$$

判定系数$r^2=0.924\ 3$，显著性检验P值≈0，高度显著。将式（9.8）还原为式（9.5）的Logistic回归方程为：

$$\hat{p}=\frac{\exp(-0.886+0.156x)}{1+\exp(-0.886+0.156x)} \tag{9.9}$$

利用式（9.9）可以对购房比例做预测，例如对 $x_0=8$，则有

$$\hat{p}=\frac{\exp(-0.886+0.156\times 8)}{1+\exp(-0.886+0.156\times 8)}$$
$$=\frac{1.436}{1+1.436}=0.590$$

这表明，在住房展销会上与房地产商签订初步购房意向书的年收入 8 万元的家庭中，预计实际购房比例为 59%。或者说，一个签订初步购房意向书的年收入 8 万元的家庭，其购房概率为 59%。

我们用 Logistic 回归模型成功地拟合了因变量为定性变量的回归模型，但是仍然存在一个不足之处，即异方差性并没有解决。式（9.7）的回归模型不是等方差的，应该对式（9.7）用加权最小二乘估计。当 n_i 较大时，p_i'的近似方差为：

$$D(p_i')\approx\frac{1}{n_i\pi_i(1-\pi_i)} \tag{9.10}$$

其中 $\pi_i=E(y_i)$，因而选取权数：

$$w_i=n_ip_i(1-p_i) \tag{9.11}$$

对例 9—2 重新用加权最小二乘做估计。在 SPSS 软件操作中，点选 Analyze→Regression→linear Regression，Dependent：逻辑变换；Independent：年家庭收入；WLS Weight：权重 $[w_i]$。计算结果如输出结果 9—6 所示。

输出结果 9—6

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.939[a]	0.881	0.864	0.386 2

(1)

a. Predictors: (Constant), x.

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7.754	1	7.754	51.983	0.000
	Residual	1.044	7	0.149		
	Total	8.798	8			

(2)

c. Weighted Least Squares Regression-Weighted by W.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
(Constant)	−0.849	0.114		−7.474	0.000
X	0.149	0.021	0.939	7.210	0.000

(3)

用加权最小二乘法得到的 Logistic 回归方程为：

$$\hat{p}_i=\frac{\exp(-0.849+0.149x)}{1+\exp(-0.849+0.149x)} \tag{9.12}$$

利用式（9.12）可以对 $x_0=8$ 时的购房比例做预测，有

$$\begin{aligned}\hat{p}_i&=\frac{\exp(-0.849+0.149\times 8)}{1+\exp(-0.849+0.149\times 8)}\\&=\frac{1.409}{1+1.409}=0.585\end{aligned}$$

所以，年收入 8 万元的家庭预计实际购房比例为 58.5%，这个结果与未加权的结果很接近。

上例是只有一个自变量的情况，分组数据的 Logistic 回归模型可以很方便地推广到多个自变量的情况，在此就不举例说明了。

分组数据的 Logistic 回归只适用于大样本的分组数据，对小样本的未分组数据不适用，并且以组数 c 为回归拟合的样本量，使拟合的精度降低。实际上，我们可以用极大似然估计直接拟合未分组数据的 Logistic 回归模型，以下就介绍这种方法。

9.3.2 未分组数据的 Logistic 回归模型

设 y 是 0—1 型变量，x_1，x_2，…，x_p 是与 y 相关的确定性变量，n 组观测数据为（x_{i1}，x_{i2}，…，x_{ip}；y_i）（$i=1$，2，…，n），其中 y_1，y_2，…，y_n 是取值 0 或 1 的随机变量，y_i 与 x_{i1}，x_{i2}，…，x_{ip} 的关系为：

$$E(y_i)=\pi_i=f(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip})$$

其中函数 $f(x)$ 是值域在［0，1］区间内的单调增函数，对于 Logistic 回归，有

$$f(x)=\frac{e^x}{1+e^x}$$

于是 y_i 是均值为 $\pi_i=f(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip})$ 的 0—1 型分布，概率函数为：

$$\begin{aligned}&P(y_i=1)=\pi_i\\&P(y_i=0)=1-\pi_i\end{aligned}$$

可以把 y_i 的概率函数合写为：

$$P(y_i)=\pi_i^{y_i}(1-\pi_i)^{1-y_i},\quad y_i=0,1;\ i=1,2,\cdots,n \tag{9.13}$$

于是 y_1，y_2，…，y_n 的似然函数为：

$$L=\prod_{i=1}^{n}P(y_i)=\prod_{i=1}^{n}\pi_i^{y_i}(1-\pi_i)^{1-y_i} \tag{9.14}$$

对似然函数取自然对数，得

$$\begin{aligned}\ln L &= \sum_{i=1}^{n}[y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)] \\ &= \sum_{i=1}^{n}\left[y_i \ln \frac{\pi_i}{(1-\pi_i)} + \ln(1-\pi_i)\right]\end{aligned}$$

对于 Logistic 回归，将

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}$$

代入，得

$$\begin{aligned}\ln L = \sum_{i=1}^{n}\{&y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \\ &- \ln[1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})]\}\end{aligned} \tag{9.15}$$

极大似然估计就是选取 β_0，β_1，β_2，…，β_p 的估计值 $\hat{\beta}_0$，$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_p$，使式（9.15）达到极大。求解过程需要用数值计算，SPSS 软件拥有求解功能。

例 9—3

在一次关于公共交通的社会调查中，一个调查项目为“是乘坐公交车上下班，还是骑自行车上下班”。因变量 $y=1$ 表示主要乘坐公交车上下班，$y=0$ 表示主要骑自行车上下班。自变量 x_1 是年龄，作为连续型变量；x_2 是月收入（元）；x_3 是性别，$x_3=1$ 表示男性，$x_3=0$ 表示女性。调查对象为工薪族群体，数据见表 9—6，试建立 y 与自变量间的 Logistic 回归。

表 9—6

序号	性别	年龄(岁)	月收入(元)	y	序号	性别	年龄(岁)	月收入(元)	y
1	0	18	850	0	15	1	20	1 000	0
2	0	21	1 200	0	16	1	25	1 200	0
3	0	23	850	1	17	1	27	1 300	0
4	0	23	950	1	18	1	28	1 500	0
5	0	28	1 200	1	19	1	30	950	1
6	0	31	850	0	20	1	32	1 000	0
7	0	36	1 500	1	21	1	33	1 800	0
8	0	42	1 000	1	22	1	33	1 000	0
9	0	46	950	1	23	1	38	1 200	0
10	0	48	1 200	0	24	1	41	1 500	0
11	0	55	1 800	1	25	1	45	1 800	1
12	0	56	2 100	1	26	1	48	1 000	0
13	0	58	1 800	1	27	1	52	1 500	1
14	1	18	850	0	28	1	56	1 800	1

依次点选 SPSS 软件的 Analyze→Regression→Binary Logistic 命令，进入 Logistic 回归对话框，将 y 选入 Dependent 框，将性别、年龄、月收入选入 Covariate 框，点选 OK 运行，以下是部分运行结果，见输出结果 9—7。

输出结果 9—7

Variable	B	S.E.	Wald	df	Sig.	R	Exp (B)
SEX	-2.501 6	1.157 8	4.668 9	1	0.030 7	-0.262 7	0.082 0
AGE	0.082 2	0.052 1	2.485 3	1	0.114 9	0.112 0	1.085 6
X2	0.001 5	0.001 9	0.661 3	1	0.416 1	0.000 0	1.001 5
Constant	-3.654 7	2.091 1	3.054 5	1	0.080 5		

输出结果 9—7 中，SEX（性别），AGE（年龄），X2（月收入）是 3 个自变量，Wald 是回归系数检验的统计量值：

$$\text{Wald}=\left(\frac{B}{S.E.}\right)^2=\left(\frac{\beta_j}{\sqrt{D(\beta_j)}}\right)^2 \tag{9.16}$$

Sig. 是 Wald 检验的显著性概率，R 是偏相关系数。可以看到，X2（月收入）不显著，决定将其剔除。用 y 对性别与年龄两个自变量做回归，见输出结果 9—8。

输出结果 9—8

Variable	B	S.E.	Wald	df	Sig.	R	Exp (B)
SEX	-2.223 9	1.047 6	4.505 9	1	0.033 8	-0.254 6	0.108 2
AGE	0.102 3	0.045 8	4.985 6	1	0.025 6	0.277 8	1.107 7
Constant	-2.628 5	1.553 7	2.862 0	1	0.090 7		

可以看到，SEX，AGE 两个自变量都是显著的，因而最终的回归方程为：

$$\hat{p}_i=\frac{\exp(-2.628\ 5-2.223\ 9\text{SEX}+0.102\ 3\text{AGE})}{1+\exp(-2.628\ 5-2.223\ 9\text{SEX}+0.102\ 3\text{AGE})}$$

以上方程式表明，女性乘公交车的比例高于男性，年龄越大，乘车的比例也越高。

SPSS 软件没有给出 Logistic 回归的标准化回归系数，对于 Logistic 回归，回归系数也没有普通线性回归那样的解释，因而标准化回归系数并不重要。如果要考虑每个自变量在回归方程中的重要性，不妨直接比较 Wald 值（或 Sig. 值），Wald 值大者（或 Sig. 值小者）显著性高，也就更重要。当然，这里假定自变量间没有强的复共线性，否则回归系数的大小及其显著性概率都没有意义。

该例中的性别变量严格来说是一个分类变量，读者可以尝试利用 Categorical 按钮将年龄定义为 Categorical 变量，观察输出的结果有什么不同，并给予解释。

9.4 Logistic 回归的方法及步骤

鉴于 Logistic 回归与判别分析的相似性，我们可以对比两种方法的相似性和不

同点。Logistic 回归的自变量可以是定量变量或定性变量（需要编码），这样可以检验自变量对于 Logistic 回归模型的贡献、自变量的显著性以及 Logistic 模型的判别精度。Logistic 回归一般有以下几个步骤。(1) 选择自变量和因变量。这里因变量为分组变量（限于篇幅，我们仅介绍因变量分两组的情况），自变量可以是定量变量和定性变量。Logistic 回归对于资料数据有较强的稳健性(robustness)，无须各组自变量的协方差阵相等的假定。(2) 将一部分样品用于估计 Logistic 函数（分析样品），另一部分样品用于检验模型的判别精度（保留样品）。(3) 模型中假定自变量之间不存在高度相关，因变量发生概率的模型为 Logistic 模型，这样我们可以进行 Logistic 回归估计。(4) 估计模型参数，评估拟合情况。我们选择回归估计的方法对回归参数进行估计并检验回归参数的显著性，对模型的拟合程度进行检验。(5) 解释所得到的模型结果。通过参数的显著性和符号、大小来解释自变量对因变量的意义。(6) 通过保留样本来验证模型的判别精度。

Logistic 回归的逻辑框图如图 9—4 所示。

图 9—4　Logistic 回归逻辑框图

分组数据的 Logistic 回归首先要对频率做 Logit 变换，变换公式为 $p_i'=\ln\left(\frac{p_i}{1-p_i}\right)$，这个变换要求 $p_i=m_i/n_i\neq 0$ 或 1，即要求 $m_i\neq 0$，$m_i\neq n_i$。当存在 $m_i=0$ 或 $m_i=n_i$ 时，可以用如下的修正公式计算样本频率：

$$p_i=\frac{m_i+0.5}{n_i+1} \tag{9.17}$$

分组数据的 Logistic 回归存在异方差性，需要采用加权最小二乘估计。除了式 (9.11) 给出的权函数 $w_i=n_ip_i(1-p_i)$ 之外，也可以通过二阶段最小二乘法确定权函数。

第一阶段是用普通最小二乘拟合回归模型。

第二阶段是从第一阶段的结果估计出组比例 $\hat{p}_i$，用权数 $w_i=n_i\hat{p}_i(1-\hat{p}_i)$ 做加权最小二乘。具体参见参考文献 [2]。

Logistic 回归的应用非常广泛。我们将 Logistic 回归建模方法用于标准化试题的评价也得到了很有意义的结果。详见参考文献 [4]。

因变量为多组（大于两组）的情况下，也可以使用 Logistic 回归模型。Logistic 回归分析大部分用于构建二元（dichotomous）因变量与一组解释变量之间的关系，不过有时候因变量多于两水平，Logistic 回归仍可使用，称为多元（polytomous）Logistic 回归，它用在很多研究领域，如构建疾病的轻、中、重的严重性与患者的年龄、性别及其他感兴趣的解释变量的关系。多元 Logistic 回归模型是二元 Logistic 回归模型的推广，这种推广使问题变得很复杂，由于模型的构建基础、偏差的使用及统计推断不同，可以利用逼近法配合几个二元 Logistic 回归模型做多元 Logistic 回归。这里不做详细介绍。详见参考文献 [5]、[8]。

□ 参考文献

[1] 张尧庭. 定性资料的统计分析. 桂林：广西师范大学出版社，1991
[2] 王国梁，何晓群. 多变量经济数据统计分析. 西安：陕西科学出版社，1993
[3] 约翰·内特. 应用线性回归模型. 北京：中国统计出版社，1990
[4] 何晓群等. 多元统计分析在考试评价中的应用. 国家教育部课题报告，2000
[5] 何晓群，刘文卿. 应用回归分析（第三版）. 北京：中国人民大学出版社，2011
[6] 黄登源. 应用回归分析. 台北：华泰文化事业公司，1998
[7] 张文彤. SPSS 统计分析高级教程. 北京：高等教育出版社，2004
[8] Alan Agresti. 分类数据分析. 重庆：重庆大学出版社，2012

□ 思考与练习

1. 简述对数线性模型应用的原理。
2. 试建立一个实际问题的对数线性模型。
3. Logistic 回归模型在处理问卷调查数据中有何应用？
4. 试用 SPSS 软件建立一个实际问题的 Logistic 回归模型。

第 10 章

Chapter 10 路径分析

学习目标

1. 了解路径分析和回归分析的区别，了解路径分析的假设条件；
2. 理解路径分析所涉及的基本概念；
3. 理解路径系数的求解原理，并能使用软件求出路径系数；
4. 能够检验中间变量的中间作用；
5. 能够使用 Wright 规则对路径图中的相关系数进行分解；
6. 理解对模型进行调试的意义，并能对模型进行检验；
7. 能够运用合适的软件，采用路径分析解决实际问题。

20 世纪初，Pearson 原理在生物遗传学（在过去几乎就是我们现在所称的统计学）中占统治地位。Pearson 原理的一个基本内容就是相关关系是现实生活中最基本的关系，而因果关系仅是完全相关的（理论）极限。这种理论认为没必要寻找变量之间的因果关系，只需计算相关系数。然而，相关分析逐渐暴露出自身的很多局限：一是仅反映变量之间的线性关系；二是反映的变量之间的关系是对称的，而很多变量之间的关系是非对称的；三是只有在正态假设下，相关思想才是有效的。

在遗传学中，很多现象具有明显的因果关系，如父代与子代的基因关系，父代在前，子代在后，二者的关系只能是单向的，而非对称的。对这种变量结构进行思考，遗传学家休厄尔·赖特（Sewall Wright）于 1918—1921 年提出路径分析（path analysis），用来分析变量间的因果关系。现代的路径分析由于生物遗传学家、心理测验学家、计量经济学家以及社会学家的推进，引入隐变量（latent variable，又称不可观测变量（unmeasured variable）），并允许变量间具有测量误差，并用极大似然估计代替了最小二乘法，成为路径系数主流的估计方法。路径分析现在已成为多元分析的一种重要方法，广泛应用于遗传学、社会学、心理学、经济问题和市场调研

领域。然而，习惯上把基于最小二乘的传统的路径分析称作路径分析，而把基于极大似然的路径分析称作结构方程模型（structural equation modeling，SEM）。本章主要介绍传统的路径分析，不进行特别说明。本章所提到的路径分析均指基于最小二乘的路径分析，结构方程模型放在下章介绍。

10.1 基本概念和理论

关于基本概念如路径图、直接作用、间接作用的理解对于掌握路径分析非常重要，这些概念共同构成了路径分析的基本理论。

10.1.1 路径图

路径分析的主要工具是路径图，它采用一条带箭头的线（单箭头表示变量间的因果关系，双箭头表示变量间的相关关系）表示变量间预先设定的关系，箭头表明变量间的关系是线性的，很明显，箭头表示一种因果关系发生的方向。在路径图中，观测变量一般写在矩形框内，不可观测变量一般写在椭圆框内。对于简单的路径模型，可以直接用字母表示变量，绘出路径图。图 10—1 是一个简单的路径图，A 是父亲的智商，B 是母亲的智商，C_1，C_2 是两个成年子女的智商，e_1，e_2 是与 A，B 不相关的其他原因变量。一般来说，父母亲的智商之间不存在关系；父母亲的智商与子女的智商存在因果关系，用单箭头表示；子女之间存在相关关系，用双箭头表示。箭头上的字母表示路径系数，路径系数反映原因变量对结果变量的相对影响大小。在路径分析中，一般采用经过标准化后的变量，没有特别说明，均指经过标准化后的变量。可以把图 10—1 写为方程式的形式：

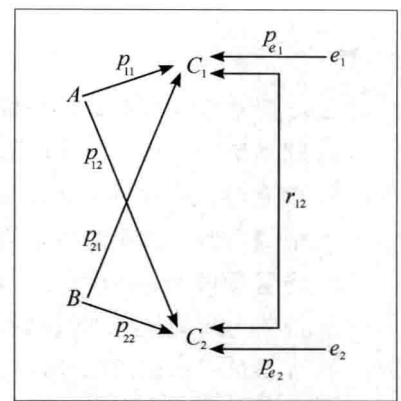


图 10—1

$$
\begin{aligned}
C_1 &= p_{11}A + p_{21}B + p_{12}r_{12}AC_2 + p_{22}r_{12}BC_2 + p_{e_1}e_1 \\
C_2 &= p_{12}A + p_{22}B + p_{11}r_{12}AC_1 + p_{21}r_{12}BC_1 + p_{e_2}e_2
\end{aligned} \tag{10.1}
$$

式（10.1）实际上是普通的多元回归方程。多元回归方程是因果关系模型的一种，但它是一种比较简单的因果关系模型，各个自变量对因变量的作用并列存在，它仅包含一个环节的因果结构。路径分析的优势在于它可以容纳多环节的因果结构，通过路径图把这些因果关系很清楚地表示出来，据此进行更深层次的分析，如比较各种因素之间的相对重要程度，计算变量之间的直接与间接影响，这在后面会涉及。

图 10—2 是关于一种消费性电子产品（如手机）路径分析的例子（这里省略了路径系数），四个变量中，耐用性、使用的简单性、通话效果和价格两两相关，决定感知价值，同时通过感知价值决定客户忠诚度。相对于图 10—1，它具有两层因果关系。接下来主要以图 10—2 为例，说明路径图中的一些基本概念。

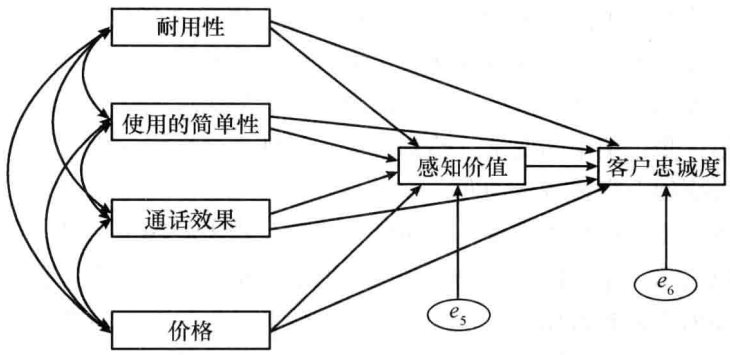


图 10—2

10.1.2　内生变量和外生变量

路径图上的变量分为两大类：一类是外生变量（exogenous variable，又称独立变量或源变量），它不受模型中其他变量的影响，如图 10—2 中的耐用性、使用的简单性、通话效果和价格；与此相反，另一类是内生变量（endogenous variable，又称因变量或下游变量），在路径图上至少有一个箭头指向它，它被模型中的其他变量决定，如图 10—2 中的感知价值由耐用性、使用的简单性、通话效果和价格四个变量和随机误差 e_5 决定，忠诚度取决于四个外生变量、感知价值和随机误差 e_6。此外，我们可以将路径图中不影响其他变量的内生变量称为最终结果变量（ultimate response variable），最终结果变量不一定只有一个。图 10—2 中忠诚度是最终结果变量。

10.1.3　直接作用和间接作用

其他变量（A）对内生变量（B）的影响有两种情况：若 A 直接通过单向箭头对 B 具有因果影响，称 A 对 B 有直接作用（direct effect）；若 A 对 B 间接地通过其他变量（C）起作用，称 A 对 B 有间接作用（indirect effect），称 C 为中间变量（mediator variable）。变量间的间接作用常常由多种路径最终综合而成。图 10—2 中，四个外生变量耐用性、使用的简单性、通话效果和价格既对忠诚度有直接作用，同时又通过感知价值对忠诚度具有间接作用。

10.1.4 间接作用的检验

如果模型中包含中间变量，首先从理论角度考虑，这个中间作用是否有理论依据，其次实际工作者会提出这样的问题："模型中中间变量的中间影响显著吗?"这些问题涉及对中间变量的间接作用的检验。巴伦和肯尼（R. M. Barron，D. Kenny，1986）提出了检验中间变量间接作用是否统计显著的一种做法。他们利用基于普通最小二乘的多元回归进行，下面以图 10—2 为例说明这种做法。

第一步：用中间变量（感知价值）对外生变量耐用性、使用的简单性、通话效果和价格四个变量进行回归；

第二步：用内生变量（忠诚度）对第一步中的四个变量进行回归；

第三步：用忠诚度对第一步中的四个变量以及中间变量感知价值进行回归。

阿加沃尔和蒂斯（S. Agarwal，R. K. Teas，1997，见参考文献［1］）的工作表明，"如果（a）在第一步的估计中解释变量统计显著；（b）在第二步的估计中解释变量统计显著；（c）在第三步的估计中中间变量统计显著，则说明中间变量的间接作用显著"。

假设对图 10—2 进行间接作用检验，得到表 10—1，见参考文献［4］。

表 10—1　　间接作用的检验结果

因变量 / 自变量	第一步 感知价值	第二步 忠诚度	第三步 忠诚度	说明
耐用性	0.26	0.65	0.62	部分间接作用
使用的简单性	0.08	0.07	0.06	部分间接作用
通话效果	0.15	0.14	0.12	部分间接作用
价格	0.39	0.08	Ns	完全间接作用
感知价值			0.12	

说明：所有的间接作用参数均为统计显著的。

对每个外生变量，存在三种可能的中间结果：没有间接作用（no mediation）、部分间接作用（partial mediation）和完全间接作用（full mediation）。如果第一步中外生变量的回归系数不是统计显著的或者第三步中（中间变量）感知价值的回归系数不显著，说明该外生变量不存在间接作用；如果某一外生变量（如耐用性、使用的简单性和通话效果）在第一步和第三步中的回归系数都是统计显著的，说明该外生变量存在部分间接作用；如果某外生变量（价格）的回归系数在第一步显著，而在第三步不显著，说明该外生变量存在完全的间接作用。

10.1.5 递归路径模型和非递归路径模型

广义的路径模型有两种基本类型：递归模型和非递归模型。两种模型在分析时有所不同，递归模型可以直接通过最小二乘求解，而非递归模型的求解比较复杂。

尽管本章主要介绍基于最小二乘的路径分析（即递归路径模型），但同时也要求读者能够预先正确判断一个模型的所属类型，这样才能保证应用路径分析不会出错。

因果关系结构中，全部为单向链条关系、无反馈作用的模型称为递归模型（recursive model）。无反馈作用意味着，各内生变量与其原因变量的误差项之间或每两个内生变量的误差项之间必须相互独立。与递归模型相对的另一类模型称为非递归模型（nonrecursive model）。一般来说，非递归模型相对来说容易判断，如果一个模型不具有非递归模型的特征，它便是递归模型。

如果一个路径模型包括以下四种情况，便是非递归模型。

情况一：模型中任何两个变量之间存在直接反馈作用，在路径图上表示为双向因果关系，如图 10—3（a）所示。

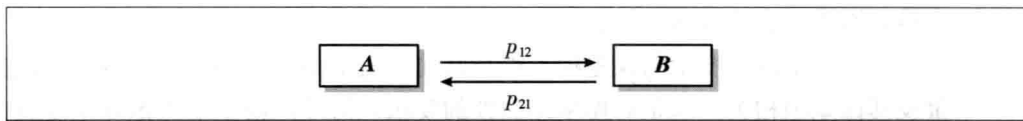


图 10—3（a）

情况二：某变量存在自身反馈作用，即该变量存在自相关，如图 10—3（b）所示。

情况三：变量之间虽然没有直接反馈，但是存在间接反馈作用，即顺着某一变量及随后变量的路径方向循序前进，经过若干变量后，又能返回这一起始变量，如图 10—3（c）所示。

情况四：内生变量的误差项与其他有关项相关，如结果变量的误差项与其原因项相关（见图 10—3（d）），或者不同变量之间的误差项之间存在相关（见图 10—3（e））。

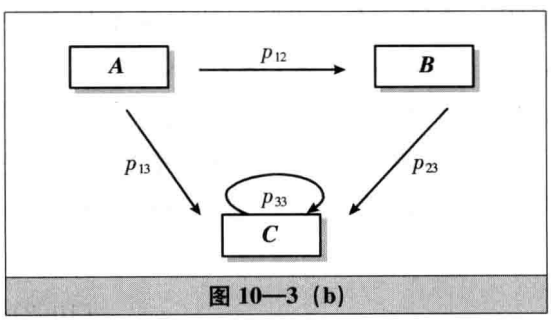


图 10—3（b）

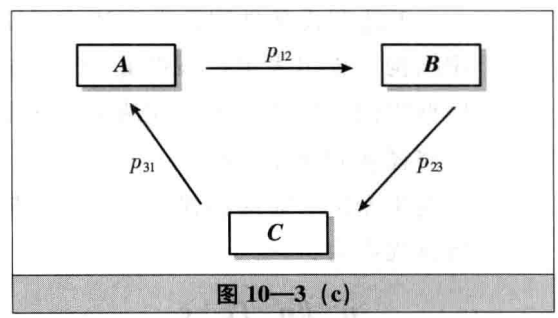


图 10—3（c）

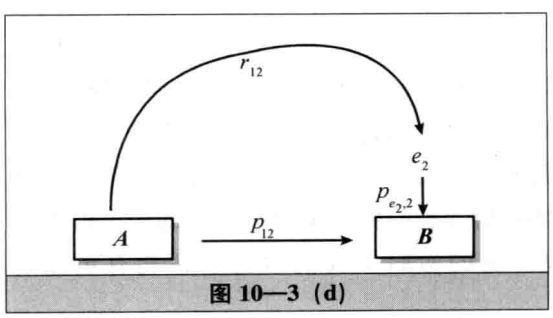


图 10—3（d）

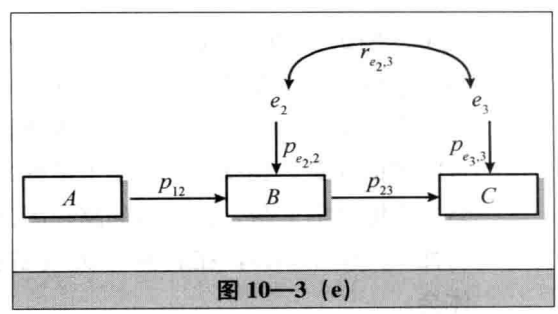


图 10—3（e）

10.1.6 （递归）路径模型的假设条件

使用最小二乘的估计方法要求路径模型具有一些假设要求和限制，现在总结如下：

（1）模型中各变量的函数关系为线性、可加，否则不能采用回归方法估计路径系数。如果变量之间存在交互作用，把交互项看作一个单独的变量，此时它与其他变量的函数关系同样满足线性、可加。

（2）模型中各变量均为等间距测度。尽管路径分析中通常会使用二分数据（dichotomies data）或者顺序数据（ordinal data），但是不能使用超过一个值的虚拟变量，因为这会违反递归性要求。

（3）每个内生变量的误差项不得与其前置变量相关，同时也不得与其他内生变量及其误差项相关。这是对模型递归性的要求。另外，模型不考虑外生变量的相关性，即不对外生变量的相关性进行分析。

（4）模型中的因果关系必须为单向，不得包括各种形式的反馈作用。这同样是对模型递归性的要求。

（5）各变量均为可观测变量，并且各变量的测量不能存在误差。这两个弱点在SEM技术中得到了克服，已经发展了一套成熟的处理隐变量和测量误差的技术。

（6）变量间的多重共线性程度不能太高，否则路径系数估计值的误差将会很大。

（7）需要有足够的样本量。克兰（Kline，1998）建议样本量的个数应该是需要估计的参数个数的10倍（20倍更理想）。

上述假设条件用数学表达式可以说明。任何一个（递归）路径模型都可以用结构方程组表示，假设 $\boldsymbol{\eta}=(\eta_1, \eta_2, \cdots, \eta_m)'$ 和 $\boldsymbol{\xi}=(\xi_1, \xi_2, \cdots, \xi_n)'$ 分别为模型中的内生和外生可观测变量向量，$\boldsymbol{B}_{m\times m}$ 是 $\boldsymbol{\eta}$ 的参系数矩阵，可以证明，若为路径递归模型，则 $\boldsymbol{B}_{m\times m}$ 总可以写为上三角矩阵，$\boldsymbol{\Gamma}_{m\times n}$ 是 $\boldsymbol{\xi}$ 的参系数矩阵，$\boldsymbol{e}$ 为内生变量所对应的误差项，满足期望为零，两两不相关。则该路径模型的结构方程组为：

$$\boldsymbol{\eta}=\boldsymbol{B}\boldsymbol{\eta}+\boldsymbol{\Gamma}\boldsymbol{\xi}+\boldsymbol{e} \tag{10.2}$$

式中，$E(\boldsymbol{e})=\boldsymbol{0}$；$E(\boldsymbol{e}'\boldsymbol{e})=\text{diag}(\boldsymbol{e}'\boldsymbol{e})$。在上述假设下，采用最小二乘法可以很容易地求解出各个参数值，见参考文献[3]，并且可以单独对其中一个方程求解。

作为本节的结束，我们需要提醒读者：一个好的路径图并不意味着一定包含尽可能多的箭头。相反，统计学上最感兴趣的情形是：应该寻找尽可能少的箭头去联结尽可能少的变量，而这时的路径图又能对所代表的样本拟合得好，即所谓模型简约性（parsimony），在后面有关模型拟合度的检验中我们对这段话会有更深的体会。

10.2　分解相关系数

路径分析技术是从分解相关系数发展而来的，因此分解相关系数在路径分析中具有一般性意义，并且是路径分析中很重要的一部分。通过对原因变量和结果变量的相关系数的分解，我们可以很清楚地看出造成相关关系的各种原因。有时也涉及对回归系数的分解，这里不进行介绍，有兴趣的读者可参阅参考文献 [7]。

下面举例说明相关系数的分解过程。图 10—4 为一假想的六个变量的路径图：A，B，C 为三个两两相关的外生变量，A，B 和误差项 e_4 共同决定 D；B，C，D 和误差项 e_5 决定 E；最后，D，E 和误差项 e_6 影响最终结果变量 F，共具有三层因果关系。对应于路径图，我们写出结构方程组：

$$D=p_{14}A+p_{24}B+p_{e_4,4}e_4 \tag{10.3}$$

$$E=p_{25}B+p_{35}C+p_{45}D+p_{e_5,5}e_5 \tag{10.4}$$

$$F=p_{46}D+p_{56}E+p_{e_6,6}e_6 \tag{10.5}$$

外生变量的相关关系在图中体现，内生变量的误差项之间独立，内生变量的误差项与其前置变量之间独立。

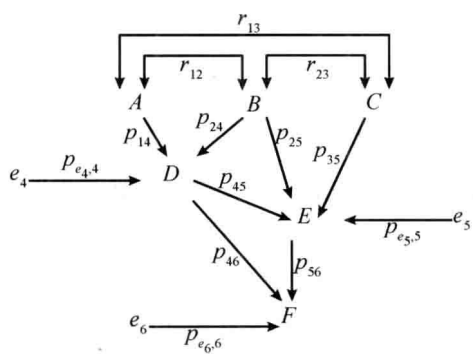


图 10—4

在式 (10.3) 中，如果路径系数 p_{14}，p_{24} 已知，则 D 的方差

$$\begin{aligned}\mathrm{var}(D)&=E[(p_{14}A+p_{24}B+p_{e_4,4}e_4)^2]\\&=p_{14}^2+p_{24}^2+2r_{12}p_{14}p_{24}+p_{e_4,4}^2=1\end{aligned}$$

可以从上式计算出 $p_{e_4,4}$ 的大小。容易看出，对其他的表达式存在同样的结果。这里只是提醒读者，误差项的路径系数由其他路径系数决定，并且该内生变量与其误差项的相关系数即为误差项的路径系数；另外，误差项的路径系数可由多元回归的决定系数计算出，它们之间的关系为：$p_e=\sqrt{1-R^2}$，详细的证明见参考文献 [8]。

下面考虑相关系数的分解。首先分解 A，D 之间的相关系数，由于各变量均经

过标准化处理，所以 A，D 的相关系数 r_{AD} 等于 A，D 乘积的期望值，即变量 D 用式（10.3）代替。

$$\begin{aligned} r_{AD}=r_{14}&=E[AD]\\ &=E[A(p_{14}A+p_{24}B+p_{e_4,4}e_4)]\\ &=p_{14}+r_{12}\times p_{24} \end{aligned}$$

式中，A 的方差为 1，A 与 B 的协方差为 r_{12}，A 与 e_4 独立。可以看出，A 与 D 的相关系数可以分解为两部分：p_{14} 是 A 对 D 的直接作用，$r_{12}\times p_{24}$ 的存在是由于 A 与 B 之间的相关性引入了对 D 有直接影响的 B 的作用。然而，从因果分析的角度看，$r_{12}\times p_{24}$ 并未得到分解，它既不是直接作用，也不是间接作用，仅是由于原因变量之间的相关而引入的一项，我们一般称该项为未析部分（unanalyzed part）。由于 A，B 对 D 作用的对称性，所以很容易写出

$$r_{BD}=r_{24}=p_{24}+r_{12}\times p_{14}=p_{24}+r_{21}\times p_{14}$$

另外也很容易写出 C，D 相关系数的分解式：

$$r_{CD}=r_{34}=r_{13}\times p_{14}+r_{23}\times p_{24}=r_{31}\times p_{14}+r_{32}\times p_{24}$$

下面考虑分解 B，E 之间的相关系数，变量 E 用式（10.4）代替：

$$\begin{aligned} r_{BE}=r_{25}=E[BE]&=E[B(p_{25}B+p_{35}C+p_{45}D+p_{e_5,5}e_5)]\\ &=p_{25}+p_{35}r_{23}+p_{45}r_{24} \end{aligned}$$

把 r_{24} 代入上式，整理后得

$$r_{BE}=r_{25}=p_{25}+r_{23}p_{35}+p_{24}p_{45}+r_{21}\times p_{14}\times p_{45}$$

此时，B，E 之间的相关系数分为四部分，第一部分 p_{25} 是 B 对 E 的直接作用；第二部分 $r_{23}p_{35}$ 是未析部分，理由和分解 A，D 相关系数一样；第三部分 $p_{24}p_{45}$ 是 B 通过中间变量 D 对 E 的间接作用；第四部分是未析部分和间接作用综合的结果。由于 B 与 A 的相关，B 对 E 的作用有一部分通过相关变量 A，通过中间变量 D，最终对 E 产生影响。

接下来考虑分解 D，E 之间的相关系数，采用式（10.4）代替 E，得到如下的关系式：

$$\begin{aligned} r_{DE}=r_{45}=E[DE]&=E[D(p_{25}B+p_{35}C+p_{45}D+p_{e_5,5}e_5)]\\ &=r_{BD}p_{25}+r_{CD}p_{35}+p_{45} \end{aligned}$$

把 $r_{BD}=p_{24}+r_{21}\times p_{14}$ 和 $r_{CD}=r_{31}\times p_{14}+r_{32}\times p_{24}$ 代入，重新整理后为：

$$r_{DE}=r_{45}=p_{45}+p_{24}p_{25}+r_{21}p_{14}p_{25}+r_{31}p_{14}p_{35}+r_{32}p_{24}p_{35}$$

这里，第一项 p_{45} 为 D 对 E 的直接作用，第二项 $p_{24}p_{25}$ 是前面尚未涉及的分解内容，对应路径图，既找不到间接作用的路径链条，也找不到涉及相关的路径图，这一部分产生的原因是相关系数所涉及的两个变量 D，E 有一个共同的作用因子 B。由于

共同原因变量 B 的存在，B 的变化引起 D，E 的同时变化，而使 D，E 的样本数据表现出相关关系，这种相关关系称为伪相关（spurious correlation）。很多情况下均存在伪相关问题，特别是对于时间序列。菲利普斯（Phillips，1986）在理论上证明了不相关的单位根（unit root）变量之间会存在伪相关现象，现在伪相关已经成为时间序列的一个研究专题。第三项 $r_{21}\,p_{14}\,p_{25}$，第四项 $r_{31}\,p_{14}\,p_{35}$ 和第五项 $r_{32}\,p_{24}\,p_{35}$ 的意义相同，均由 D，E 的原因变量 A，B 和 B，C 之间的相关所致，既包括未析部分，又包括伪相关部分。

分解 A，E；A，F；B，F 等任意两个变量间的相关系数与上面的步骤相似，有兴趣的读者可以自己动手做一下，这里不再赘述。

通过上面对相关系数的分解，我们可以总结出，相关系数的分解可能产生四种类型的组成部分：（1）直接作用；（2）间接作用；（3）由于原因变量相关而产生的未析部分；（4）由于共同原因的存在而产生的伪相关部分。

路径系数分解的结果一般通过报表的形式把各种作用展现出来，10.5 节的实例分析会给读者提供一个报表的形式。

然而，如果按照上面的步骤，相关系数的分解将是非常烦琐的。赖特提供了从路径图直接分解的规则。赖特认为，对于一个递归性的路径模型，任何两个变量的相关系数都可以表示成连接这两点之间的所有复合路径之和；而这个复合路径是按下述三个规则选取的（Wright 规则）：

（1）这个复合路径没有闭合环路。

（2）在这个复合路径中的箭头取向不可“先向前，再向后”，也就是说，该路径链上不止两个箭头时，要“先向后”尽可能多的次数，“再向前”尽可能少的次数。

（3）对于有多个双箭头的链，只可以取最远距离的一个双箭头，即一条路径中不可以包含两个双向箭头。

结合 Wright 规则，在图 10—5（a）中，若计算 D 和 E 的相关系数，路径 DCE 是合理的，而路径 $DCABCE$ 则不可以（规则（1））；在图 10—5（b）中，若计算 B 和 C 的相关系数，路径 BAC 是合理的，而路径 BDC 则不可以（规则（2））；在图 10—5（c）中，若计算 D 和 F 的相关系数，路径 $DACF$ 是合理的，而路径 $DABCF$ 不可以（规则（3））。

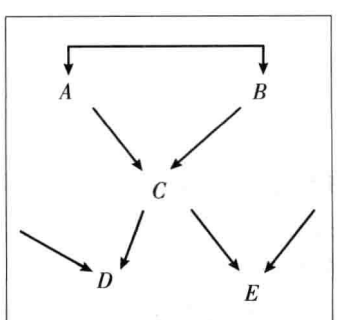


图 10—5（a）

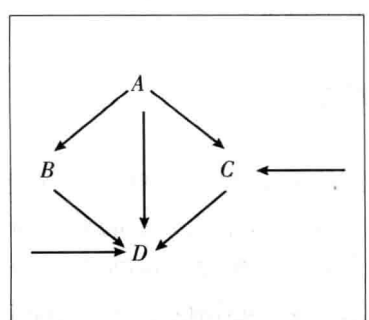


图 10—5（b）

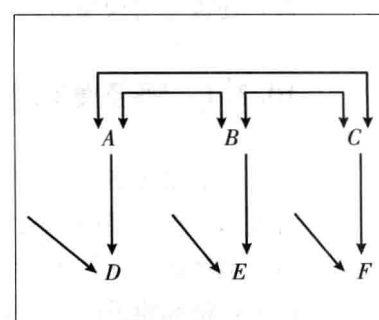


图 10—5（c）

以图 10—4 为例，为方便读者阅读，我们把图 10—4 复制到此（见图 10—6）。下面采用 Wright 规则对两个变量的相关系数进行分解。如果要对 A，D 的相关系数进行分解，从图上看出，A，D 之间存在两条链，AD 及 ABD，所以 $r_{AD}=r_{14}=p_{14}+r_{12}\times p_{24}$。如果对 C，D 之间的相关系数进行分解，C，D 之间同样存在两条链：CAD 和 CBD，链 CED 违反规则（2），链 $CBAD$ 则违反规则（3），所以 $r_{CD}=r_{34}=r_{13}p_{14}+r_{23}p_{24}$。对于 D，F 的相关系数 $r_{DF}(r_{46})$，首先有直接作用 p_{46}，其次 D 通过 E 对 F 的间接作用 $p_{45}p_{56}$，再次存在路径 $DBEF$，$DABEF$，$DACEF$ 以及 $DBCEF$，反映伪相关部分和未析部分。这样，可以把 D，F 的相关系数分解为：

$$r_{DF}=r_{46}=p_{46}+p_{45}p_{56}+p_{24}p_{25}p_{56}+p_{14}r_{21}p_{25}p_{56}$$
$$+p_{14}r_{31}p_{35}p_{56}+p_{24}r_{32}p_{35}p_{56}$$

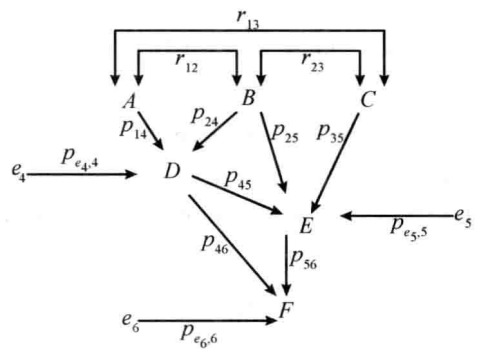


图 10—6

读者可以使用 Wright 规则对其他相关系数进行分解，看一看是否与直接计算的结果一致。

10.3 路径模型的调试和检验

根据理论建立路径模型之后，除了对路径系数进行估计、分解变量间的相关系数，还需要对模型进行调试和检验。

10.3.1 路径模型的调试

一般情况下，路径模型的调试和分析往往是先从饱和模型（即所有变量之间都有表示因果关系的单向箭头或表示相关关系的双向箭头联结）的建立开始的。然而，饱和模型往往不是我们实际上想要的最终模型，它只是作为一个起点或基准。但是，饱和模型的因果关系的建立必须依据一定的理论基础，如根据变量间的逻辑关系、时间关系来设置因果结果。如果饱和模型不符合逻辑关系，我们完全可以采

用非饱和模型作为模型检验的起点，但这个非饱和模型和我们所关注的模型应该具有包含或者说嵌套（nested）关系。

对模型的调试过程有些类似多元回归过程的调试，如果某一变量的路径系数（回归系数）统计上不显著，则考虑是否将其对应的路径从模型中删去；如果多个路径系数同时不显著，则首先删除最不显著的路径后继续进行回归分析，根据下一步的结果再决定是否需要删除其他原因变量。

上面所说的仅是对路径模型进行调试的一般原则，实际进行调试时，还必须考虑模型的理论基础。因为一些先验信息和定义准确的理论概念对于路径分析非常重要，如果忽略理论概念，即使采用路径分析得到一个统计上拟合很好的模型，它的实际意义也不大，因为有可能存在逻辑顺序上无意义的因果关系。路径分析很大程度上是证实性技术，而非探索性技术。作为研究焦点的因果联系必须有足够的理论根据，即使其统计上不显著，仍然应当加以仔细考虑，并寻找其统计不显著的原因：是多重共线性还是其他路径假设不合理，影响了该路径的显著性。在多元回归中碰到的很多问题在这里都可能碰到，我们可以参照相应的方法处理。

如果经过调试的模型与事先已设置的模型有所不同，此时可以采用拟合度对这两个模型进行检验。如果统计检验不显著，说明调试后对模型的修改并不妨碍“接受”原假设模型，更准确地说，统计检验不能拒绝原假设模型；反之，说明所得到的调试模型已经与原假设模型大不相同了。可以看出，路径分析的模型检验不是检验原模型是否符合观测数据，而是检验调试以后的模型是否与原模型一致，这正是路径模型检验的意义所在。

另外需要提醒读者一点，对每个方程所进行的回归分析检验并不等同于对整个模型的检验。对每个回归方程的检验和对整个路径模型的检验虽然有联系，但毕竟不同，因为整个路径模型并不是各个回归方程的简单叠加。即使各个回归方程中所有的路径系数都显著，整个路径模型的检验也有可能通不过。

10.3.2　路径模型的识别

所谓模型的识别，就是判断模型中的参数是否可以被估计出来。计量经济学经常涉及模型识别问题，模型识别的结果按不同情形分为以下几类：

$$\text{模型的识别}\begin{cases}\text{不可识别（under-identified）}\\ \text{可以识别（identifiable）}\begin{cases}\text{恰好识别（just-identified）}\\ \text{过度识别（over-identified）}\end{cases}\end{cases}$$

不可识别的模型就是人们掌握的信息不足以得到模型的确定解。在路径分析中，三个变量只能产生三个相关系数，如果路径模型中除了两两变量的路径之外，在其中一对变量之间还设置了反馈路径，这时需要估计四个路径系数。在这种情况下，路径系数是无法确定的，这个路径模型就是不可识别的。恰好识别的模型指人

们掌握的信息可以对需估计的参数提供唯一解。过度识别的模型指只需要较少的信息便可以得出参数的唯一解，如果一个方程组中方程的个数多于未知数的个数，则这是一个过度识别的模型。

任一递归饱和模型都是恰好识别的，因为任一路径都是两两变量形成的，回归系数也是两两变量形成的，路径系数的个数和相关系数的个数相等，知道样本的相关系数后，我们就可以求出路径系数，并且它们是可以彼此互相转化的。

饱和的递归模型是不可检验的，因为从相关系数的角度而言，它是完全拟合的，我们进行检验的是过度识别模型，在饱和的递归路径模型中删去了某些路径，这些被删去的路径反映了研究人员关于某些变量对于其他变量没有直接作用的假设。模型检验的正是这些假设。

10.3.3 对过度识别的路径模型的整体检验方法

前面提到，饱和模型能够完全“拟合”数据，则可以把饱和模型作为评价过度识别模型的基准，通过对过度识别模型中估计的相关系数与饱和模型估计的相关系数进行比较分析，从而对过度识别模型进行检验。

在介绍检验统计量之前，我们首先定义模型的整体拟合指数，它类似多元回归方程的决定系数 R^2，不妨仍以 R^2 标记这一指数，表示具体对应的路径模型。对一个作为基准的路径模型，不妨设为饱和模型。我们知道，对应每一路径模型，都可以写出其结构方程组，并且方程的个数和内生变量的个数相等，不妨设有 m 个内生变量，则对于这 m 个方程，设其回归后的决定系数分别为 $R^2_{(1)}$，$R^2_{(2)}$，…，$R^2_{(m)}$，每个 R^2 都代表相应内生变量的方差中由回归方程所解释的比例，$1-R^2$ 则表示回归方程未能解释的残差比例。定义整个路径模型的拟合指数为：

$$R_c^2=1-(1-R^2_{(1)})(1-R^2_{(2)})\cdots(1-R^2_{(m)}) \tag{10.6}$$

它指由路径模型已经解释的广义方差（generalized variance）占需要得到解释的广义方差的比例，它是一个值域为 [0，1] 的指数，也可以理解为一个百分比。这个公式中每个括号中表示的是所有内生变量方差中未被其原因变量解释的残差比例，将它们连乘便得到广义的残差比例，它代表没有包括在模型之内的各种因素的影响。用 1 减去广义的残差比例便得到饱和因果模型内部对广义的总方差所解释的比例 R_c^2。对饱和模型计算该指数是为了给检验与其嵌套的其他模型提供评价基准，所以一般称 R_c^2 为基准解释指数，$1-R_c^2$ 为基准残差指数。

同理，还可以计算出与此饱和模型嵌套的非饱和模型的相应指数：

$$R_t^2=1-(1-R^2_{(1)})(1-R^2_{(2)})\cdots(1-R^2_{(t)}) \tag{10.7}$$

它的意义和 R_c^2 类似，只不过是针对非饱和模型计算的。因为对非饱和模型计算该指数是为了检验该模型，所以也称 R_t^2 为待检解释指数，因为非饱和模型比饱和模型少了一些路径，所以两个指数满足关系：$R_t^2 \leqslant R_c^2$。将基准残差指数与待检残差指

数相比，便得到一个关于检验模型拟合度的统计量 Q。即

$$Q=\frac{1-R_c^2}{1-R_t^2} \tag{10.8}$$

Q 的分布很难求出，但可以根据 Q 构造统计量 W：

$$W=-(n-d)\ln Q=-(n-d)\ln\left(\frac{1-R_c^2}{1-R_t^2}\right) \tag{10.9}$$

式中，n 为样本大小；d 为检验模型与饱和模型的路径数目之差。大样本情况下，W 渐进遵从自由度为 d 的 χ^2 分布。

如果作为基准的模型不是饱和模型，同样可以进行模型的检验，只要待检模型与基准模型存在嵌套关系。设基准解释指数为 R_c^2，待检模型的拟合指数为 R_t^2，则所用的检验统计量 W 与上面的一样。同样，在大样本的情况下，W 渐进遵从自由度为 d 的 χ^2 分布。

10.4　路径分析流程图及 SPSS 指令

10.4.1　分析流程图

对前三节的内容进行总结，我们给出路径分析的流程框图，如图 10—7 所示，它给读者提供了路径分析步骤的参考。

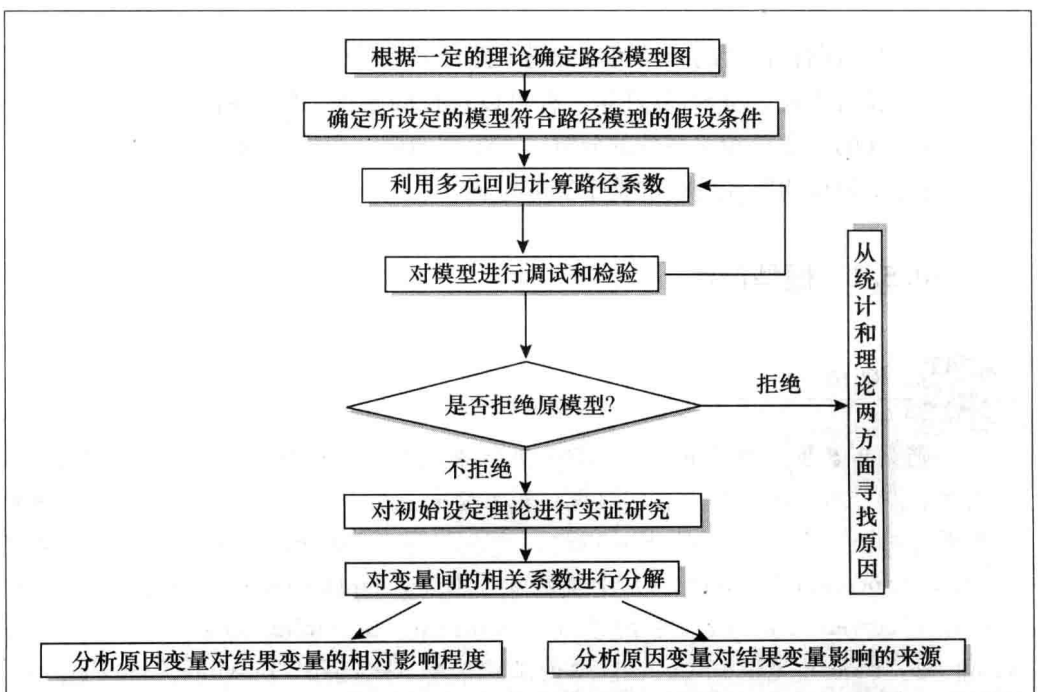


图 10—7　路径分析流程

10.4.2 进行路径分析所使用的 SPSS 指令

下面以 SPSS Amos 模块为例说明路径分析的实现过程。根据上面的介绍，路径分析也可以利用回归分析手工完成。

在 Amos Graphics 模块中，首先需要选择数据文件，在 File 菜单下，选择“Data Files”给出需要进行分析的文件名。

然后绘出路径分析图：在 Diagram 菜单下，选择“Draw Observed”绘制观测变量；选择“Draw Unobserved”绘制不可观测变量，在路径分析中是误差项；选择“Draw Path”绘制两变量的因果关系；选择“Draw Covariance”绘制两变量的相关关系；然后对绘出的各个变量指定变量名。

接着指定误差项方差为 1：选定某个误差项后，点击右键，选择“Object Properties”后，在“Parameters”下设定方差为 1。在菜单中 View/Set 下选择“Analysis Properties”，在“Estimation”一项中选择估计方法为“Scale-free least square”，关闭该窗口。

最后就可以点击 Analyze 菜单下的选项“Calculate Estimates”计算路径系数了。可以通过三种方式查看结果：文字法、表格法和图表法。

10.5 案例分析

从对路径模型的介绍可以知道，路径系数的估计并不复杂，用普通的多元回归方法就可实现。SPSS 软件是一种可以选择的软件，然而路径分析又不完全是靠软件实现的，变量相关系数的分解、对模型的调试和检验通过手工就可完成。下面以具体实例说明路径分析的整个实现过程。

10.5.1 模型设定

例 10—1

我们采用数据文件 Employee data 进行路径分析。该数据共有 474 个观测值，473 个有效，标号为 434 的出生日期缺失，在下面的分析中，不考虑该样品；该数据包含 10 个变量：标号（Id）、性别（Gender）、出生日期（Bdate，date of birth）、教育水平（Educ，educational level）、工作类别（Jobcat，employment category）、当前工资（Salary，current salary）、初始工资（Salbegin，beginning salary）、已经工作时间（Jobtime，months since hire）、以前的工作经验（Preexp，previous experience）、是否少数族裔（Minority）。性别为属性变量，用“f”表示女性，“m”表示男性。教育水平

使用受教育的年数衡量。工作类别分为三类：公务员（“1”）、监督人（“2”）以及经理人员（“3”）。当前工资和初始工资以实际额为准。已经工作的时间和以前的工作经验均以发生的月份衡量。是否少数族裔为 0，1 变量，1 表示是少数族裔，0 表示非少数族裔。假设数据的采集时间为 1997 年，则用 1997 减出生日期的年份数作为年龄（Age）的衡量指标。例如若某人在 1952 年出生，则年龄的测度为 1997－1952＝45。表 10—2 为样本相关系数矩阵。

表 10—2　　样本相关系数矩阵

样本相关系数（r）	已经工作时间	以前工作经验	年龄	初始工资	当前工资	教育水平	工作类别
已经工作时间	1.000						
以前工作经验	0.002	1.000	**		*	**	
年龄	0.051	0.803**	1.000		**	**	
初始工资	−0.018	0.045	−0.011	1.000	**	**	**
当前工资	0.084	−0.097*	−0.146**	0.880**	1.000	**	**
教育水平	0.050	−0.252**	−0.282**	0.633**	0.661**	1.000	**
工作类别	0.004	0.062	0.009	0.755**	0.780**	0.515**	1.000

* 表示在 5%的显著性水平下统计显著；** 表示在 1%的显著性水平下统计显著。

对标号、性别、民族不进行区分，关注其余 7 个变量之间的因果关系。表 10—2 为这 7 个变量的样本相关系数。根据时间和逻辑顺序，我们得到几条因果路径。首先，教育水平影响初始工资和当前工资，因为大量统计结果表明，个人的教育水平越高，所获得工资也越高；同时也认为，一个人教育水平越高，以前的工作经验越多，他从事的工作类别应该越高。另外，初始工资会影响工作类别，在相关系数矩阵中，我们已经看到二者的相关系数较大。年龄影响已经工作的时间以及以前的工作经验，因为年龄越大，（在本职位）已经工作的时间或者以前的工作经验会越长。其次，年龄和教育水平应该存在负相关，这里不关注二者的因果关系，仅简单假设二者相关。最后，初始工资、工作类别、已经工作的时间以及以前的工作经验都影响当前工资，一般来说，初始工资越高，工作类别越高（按 1，2，3 的顺序），以前工作的经验越多，时间越长，当前的工资越高，这些变量间均应有正的因果关系。根据这些逻辑理由，我们假设的路径模型如图 10—8 所示，不妨称此模型为模型 1。很显然，模型 1 为递归的路径模型，各外生变量不存在测量误差。假设各路径的因果关系均为线性、可加，并进一步假设各内生变量之间不存在相关关系。

10.5.2　路径系数估计

采用 Amos 模块对图 10—8 进行估计，结果如图 10—9 所示。

根据图 10—9，我们发现年龄对已经工作时间的路径系数仅为 0.03，R^2 为 0.001（图中只保留两位小数），方程拟合效果不好；同时，以前工作经验对当前工

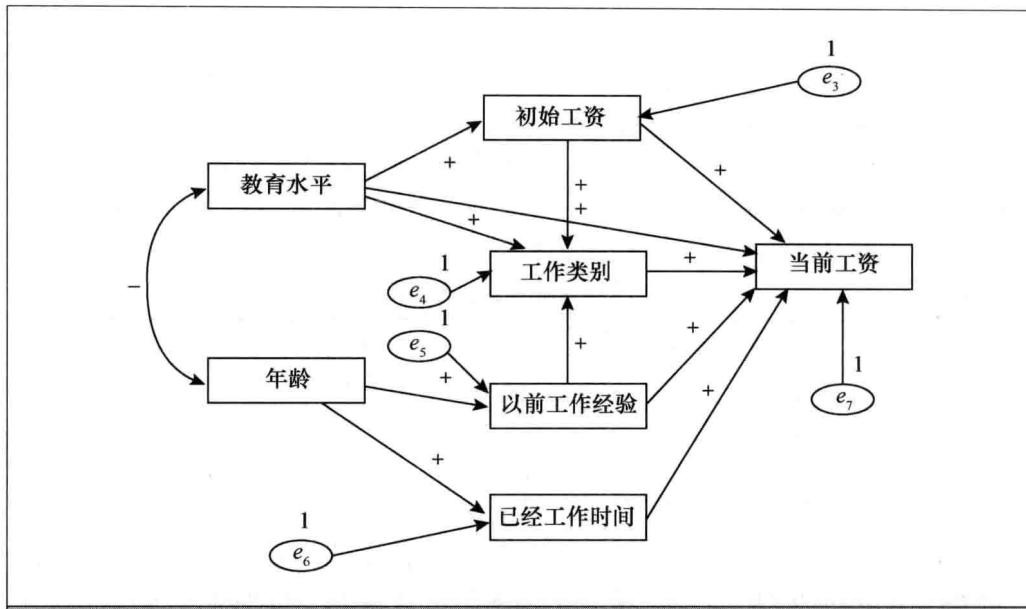


图 10—8

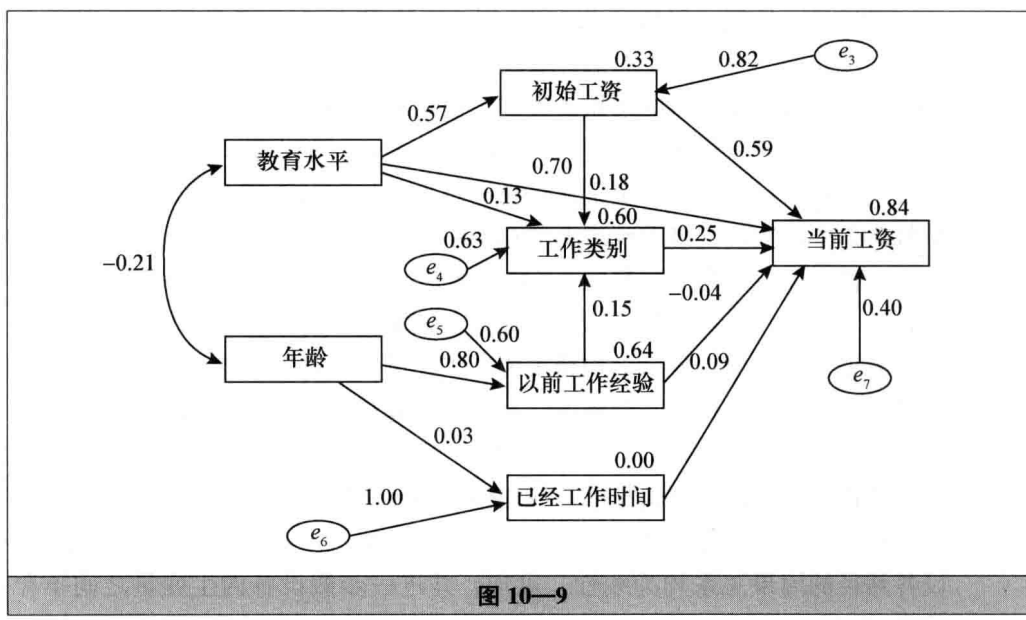


图 10—9

资的路径系数也很小。考虑删除上面的两条路径以及误差项 e_6，并重新估计模型，结果如图 10—10 所示。

10.5.3 模型的调试和检验

假设图 10—9 对应的模型是基准模型，图 10—10 对应的模型为待检模型。下

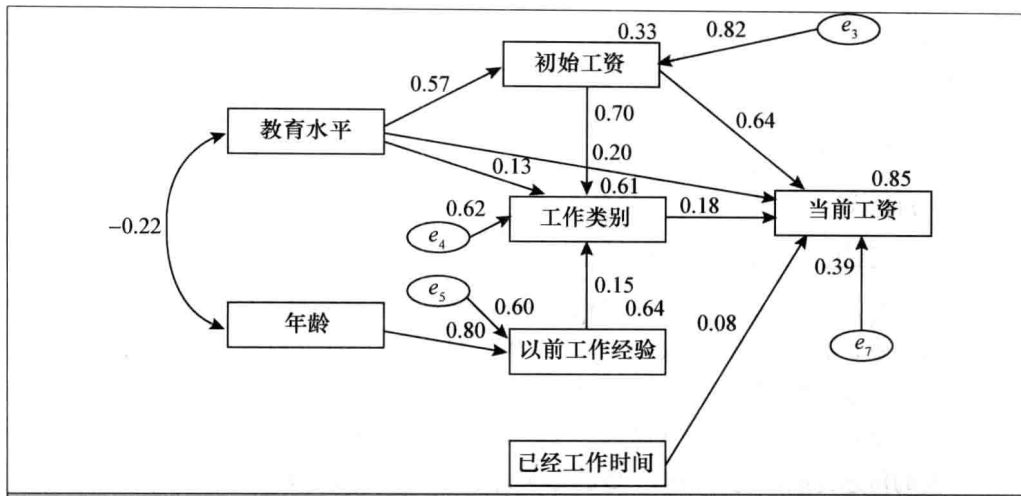


图 10—10

面分别计算基准模型和待检模型的拟合指数 R_c^2 和 R_t^2，对模型进行调试。

$$\begin{aligned}R_c^2 &= 1-(1-R_{(c3)}^2)(1-R_{(c4)}^2)(1-R_{(c5)}^2)(1-R_{(c6)}^2)(1-R_{(c7)}^2)\\&=1-(1-0.325\ 9)(1-0.600\ 7)(1-0.643\ 1)(1-0.000\ 7)\\&\quad(1-0.844\ 0)\\&=0.985\ 1\\R_t^2 &= 1-(1-R_{(t3)}^2)(1-R_{(t4)}^2)(1-R_{(t5)}^2)(1-R_{(t7)}^2)\\&=1-(1-0.325\ 1)(1-0.608\ 2)(1-0.636\ 0)(1-0.845\ 5)\\&=0.985\ 0\end{aligned}$$

从而 W 统计量为：

$$W=-(n-d)\ln\left(\frac{1-R_c^2}{1-R_t^2}\right)=-(473-2)\ln\left(\frac{1-0.985\ 1}{1-0.985\ 0}\right)=3.150\ 5$$

若基准模型正确，W 遵从自由度为 2 的 χ^2 分布。这里 W 的 p 值为 0.189 1，统计不显著，可以认为图 10—10 对应的模型正确。

10.5.4　路径系数分解

表 10—3 是 Amos 模块总效应的分解报表。

表 10—3　　**路径系数的分解报表**

原因变量	结果变量	总影响	直接影响	间接影响
教育水平	初始工资	0.570	0.570	0.000
	工作类别	0.530	0.129	0.401
	当前工资	0.658	0.196	0.462
年龄	以前工作经验	0.801	0.801	0.000
	工作类别	0.123	0.000	0.123
	当前工资	0.022	0.000	0.022

续前表

原因变量	结果变量	总影响	直接影响	间接影响
已经工作时间	当前工资	0.084	0.084	0.000
以前工作经验	工作类别	0.154	0.154	0.000
	当前工资	0.028	0.000	0.028
初始工资	工作类别	0.705	0.705	0.000
	当前工资	0.769	0.640	0.129
工作类别	当前工资	0.183	0.183	0.000

可以看出，教育水平对当前工资的影响主要是通过工作类别和初始工资传递的间接影响，教育水平对初始工资（工作）具有很大的影响作用，但对当前工资的（直接）影响便较弱（0.196），这与我们的常识相一致，初始工作可能取决于学历，以后则主要看工作经历及个人能力。年龄对当前工资的影响主要通过工作类别和以前工作经验的传递完成，它对当前工资的影响为正。其他的分析类似，读者不妨自己动手分析。

下面再举一个例子。

例 10—2

一家大型商业银行在多个地区设有分行，其业务主要是进行基础设施建设、国家重点项目建设、固定资产投资等项目的贷款。近年来，该银行的贷款额平稳增长，但不良贷款额也有较大比例的增长，这给银行业务的发展带来较大的压力。为弄清楚不良贷款形成的原因，管理者希望利用银行业务的有关数据做些定量分析，以便找出控制不良贷款的办法。表 10—4 就是该银行所属的 25 家分行 2002 年的有关业务数据（数据见参考文献 [10]）。

表 10—4　　商业银行所属的 25 家分行 2002 年的有关业务数据

分行号	不良贷款	贷款余额	应收贷款	项目数	固定资产投资额
1	0.90	67.30	6.80	5.00	51.90
2	1.10	111.30	19.80	16.00	90.90
3	4.80	173.00	7.70	17.00	73.70
4	3.20	80.80	7.20	10.00	14.50
5	7.80	199.70	16.50	19.00	63.20
6	2.70	16.20	2.20	1.00	2.20
7	1.60	107.40	10.70	17.00	20.20
8	12.50	185.40	27.10	18.00	43.80
9	1.00	96.10	1.70	10.00	55.90
10	2.60	72.80	9.10	14.00	64.30
11	0.30	64.20	2.10	11.00	42.70
12	4.00	132.20	11.20	23.00	76.70
13	0.80	58.60	6.00	14.00	22.80
14	3.50	174.60	12.70	26.00	117.10
15	10.20	263.50	15.60	34.00	146.70
16	3.00	79.30	8.90	15.00	22.90
17	0.20	14.80	0.60	2.00	42.10
18	0.40	73.50	5.90	11.00	25.30

续前表

分行号	不良贷款	贷款余额	应收贷款	项目数	固定资产投资额
19	1.00	24.70	5.00	4.00	13.40
20	6.80	139.40	7.20	28.00	64.30
21	11.60	368.20	16.80	32.00	163.90
22	1.60	95.70	3.80	10.00	44.50
23	1.20	109.60	10.30	14.00	67.90
24	7.20	196.20	15.80	16.00	39.70
25	3.20	102.20	12.00	10.00	97.10

根据经验可知，各项贷款余额越高则不良贷款越高，但同时，各项贷款余额也会受其他变量的影响，因此综合考虑之下，本例应该建立如下的路径分析模型：

$$\begin{cases} x_1 = \alpha_1 + \beta_{11} x_2 + \beta_{12} x_3 + \beta_{13} x_4 \\ y = \alpha_2 + \beta_{21} x_2 + \beta_{22} x_3 + \beta_{23} x_4 + \beta_{24} x_1 \end{cases}$$

下面考虑对该模型加以拟合，由于整个模型是一个递归模型，可以在 SPSS 中使用分别拟合回归方程的方法来实现对模型中各参数的估计。首先对各项贷款余额回归方程进行估计，见输出结果 10—1 和输出结果 10—2。

输出结果 10—1　　**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.901 (a)	.812	.786	37.202 03

a. Predictors: (Constant).

输出结果 10—2　　**Coefficients[a]**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	−14.069	16.072		−8.75	.391
	应收贷款	3.300	1.482	.260	2.226	.037
	项目数	4.360	1.453	.465	3.001	.007
	固定资产	.620	.285	.310	2.17	.041

a. Dependent Variable: 贷款余额.

可见，应收贷款、项目数、固定资产均对各项贷款余额有影响。应收贷款越高、项目数越多，则各项贷款余额越高。

下面对第二个方程进行估计，参数结果见输出结果 10—3 和输出结果 10—4。

输出结果 10—3　　**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.893 (a)	.798	.757	1.778 75

a. Predictors: (Constant).

输出结果 10—4 **Coefficients[a]**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.022	.782		-1.306	.206
	应收贷款	.148	.079	.260	1.879	.075
	项目数	.015	.083	.034	.175	.863
	固定资产	-.029	.015	-.325	-1.937	.047
	贷款余额	.040	.010	.891	3.837	.001

a. Dependent Variable：不良贷款.

固定资产和贷款余额对不良贷款有影响，而应收贷款、项目数对其影响不显著。

由上面的分析可知，如果只是拟合第二个方程，所得结果其实就是一个简单的多重回归方程，而且自变量间存在共线性。显然，对于不良贷款而言，使用路径分析并不会使模型对最终结果变量预测得更加精确。但通过对自变量间复杂关联的刻画，路径分析模型可以很精确地估计出每一个自变量究竟是通过哪些方式作用于最终因变量的，从而使研究者对问题的理解更加深入和全面。

通过上面的分析，可以将上述模型加以简化，去除那些无统计意义的变量后重新加以拟合，分析结果见输出结果 10—5 和输出结果 10—6。

输出结果 10—5 **Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.872 (a)	.761	.739	1.842 79

a. Predictors：(Constant).

输出结果 10—6 **Coefficients[a]**

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.443	.697		-.636	.531
	固定资产	-.032	.015	-.355	-2.133	.044
	贷款余额	.050	.007	.920	6.732	.000

a. Dependent Variable：不良贷款.

可见，方程的决定系数基本未变，自变量均有统计学意义。显然，简化后的路径分析模型对数据的解释程度与前一个模型相比无显著差别，但更加简洁。

本例所拟合的路径分析模型可以使用标准化系数绘制出路径图，如图 10—11 所示。

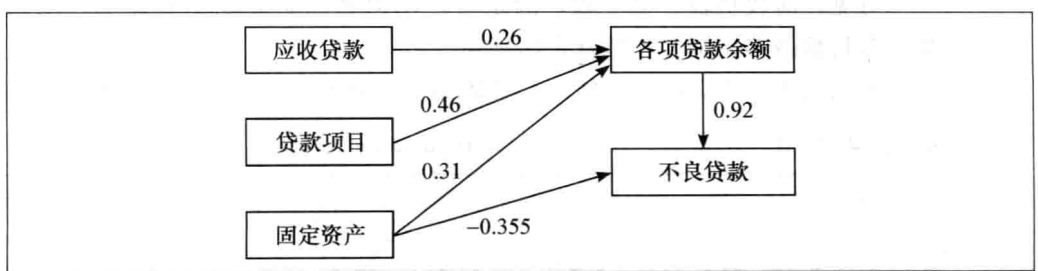


图 10—11

□ 参考文献

[1] Agarwal, Sanjeev & Teas, R. Kenneth. Quality Signals and Perceptions of Quality, Sacrifice, Value and Willingness-to-buy: An Examination of Cross-national Applicability. *Lowa State University Working Paper*, 1997, pp. 37-16

[2] Baron, Ruben M. & David A. Kenny. The Moderator-mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic and Statistical Considerations. *Journal of Marketing Research*, 1986, Vol. 19, pp. 229-239

[3] Gifi, Albert. *Nonlinear Multivariate Analysis*. John Wiley & Sons, Inc., 1990

[4] Grapentine, Terry. Path Analysis VS. Structural Equation Modeling. *Marketing Research*, 2000, pp. 10-20

[5] Kline, Theresa J. B. and Klammer, Joy D. Path Model Analyzed with Ordinary Least Squares Multiple Regression Versus LISREL. *Journal of Psychology*, 2001, 135 (2), pp. 210-225

[6] Teas, R. Kenneth. Path Analysis and LISREL. *Marketing Research*, 2000, pp. 20-22

[7] 郭志刚. 社会统计分析方法——SPSS 软件应用. 北京：中国人民大学出版社，1999

[8] 王学仁，王松桂. 实用多元统计分析. 上海：上海科学技术出版社，1990

[9] 袁志发，宋世德. 多元统计分析. 北京：科学出版社，2009

[10] 贾俊平，何晓群，金勇进. 统计学（第四版）. 北京：中国人民大学出版社，2009

□ 思考与练习

1. 路径分析和回归分析有什么异同之处？
2. 路径系数的计算应注意什么问题？
3. 试对一个实际问题画出路径图，写出相应的结构方程，并作出路径分析。

第 11 章

Chapter 11 结构方程模型

学习目标

1. 了解结构方程模型与路径分析的联系；
2. 理解结构方程模型所涉及的基本概念；
3. 理解结构方程模型分析的过程；
4. 能够借助路径分析建立结构方程模型；
5. 了解 LISREL 软件；
6. 能够运用合适的软件，采用结构方程模型研究实际问题。

结构方程模型（structural equation modeling，SEM）是近 30 年应用统计学领域中发展最为迅速的一个分支。它是一种实证分析模型，通过寻找变量间内在的结构关系，验证某种结构关系或模型的假设是否合理，模型是否正确，并且如果模型存在问题，可以指出如何加以修改。结构方程模型的另一大特点是可以对隐变量进行分析。多元回归分析、因子分析和路径分析等方法都可看成结构方程模型的特例。现实生活中，有许多变量诸如健康、优秀、乐观、智力、满意、公正等概念虽然是客观存在的，但由于人的认识水平或事物本身的抽象性、复杂性等原因，人们是无法直接测量的，我们称这样的变量为隐变量。结构方程可以通过一些可观测变量对这些隐变量的特征及其相互之间的关系进行描述，因此，有时也称结构方程模型为隐变量分析模型。

结构方程模型的应用始见于 20 世纪 60 年代发表的论文中，1987 年洛林（Loehlin）用路径分析模型和结构方程模型对隐变量模型做了出色的介绍，两年之后博伦（Bollen）提出了处理测量误差模型的更专门化的统计方法。到了 90 年代，结构方程模型得到了广泛的应用。目前，结构方程模型已发展成一个内容非常丰富的重要领域。在此，仅介绍结构方程模型的一些基本内容，有兴趣的读者可以进一步参阅相关书籍。

11.1　结构方程的基本思想及模型设定

11.1.1　结构方程模型的基本思想

结构方程模型是反映隐变量和显变量的一组方程，其目的是通过显变量的测量推断隐变量，并对假设模型的正确性进行检验。结构方程模型是模型验证技术，即利用结构方程模型分析的过程实际上是对假定模型的验证过程。对于某个领域的专业人员根据本领域的知识或常识建立的反映结构关系的模型，由于专业人员的认识水平和各种原因的限制，这一模型未必是客观现实的反映，有可能存在偏差和主观性。如何发现模型的问题，如何根据分析结果进一步修正模型，这些正是结构方程模型可以处理的问题。具体来说，结构方程模型分析的过程是：在设定结构模型的基础上，为证实模型的准确性，首先要判断这些方程是否为可识别模型，对于可识别模型，通过收集显变量的数据，利用最大似然估计或广义最小二乘估计等估计方法对未知参数进行估计。对于模型的结果，需要对模型与数据之间的拟合效果进行评价。如果模型与数据拟合得不好，就需要对模型进行修正，重新设定模型。要得到一个拟合较好的模型，往往需要反复试验多次。

在进行模型估计之前，研究者需要根据专业知识或经验设定假设的初始模型。而结构方程模型的主要用途即为确定该假定模型是否合理。

结构方程模型通常借助路径图将初始模型描述出来，对于复杂的模型尤其如此。这里从与结构方程结合的角度，对上一章的内容做简单回顾，并在此基础上看怎样得出结构方程模型。路径图中的变量可以是不同的类型。按能否直接测量，路径图中的变量可以分为显变量和隐变量，通常前者是可以直接测量的，在图中用方框来标识；而后者虽然是客观存在的，但由于人的认识水平或事物本身的抽象性、复杂性等原因，无法直接测量，通常用椭圆形框来标识。按照变量之间的关系，又可分为外生变量和内生变量，内生变量是由隐变量决定的变量，外生变量是由显变量决定的变量。变量之间的关系用线条表示，可以是直接作用，也可以是间接作用。当二者之间有直接连线时，称为直接作用。如果变量之间没有直接连线但可以通过其他变量发生联系，则假设变量之间没有直接联系，称其为间接联系。线条既可以加单箭头，也可以加双箭头。单箭头表示存在因果关系，双箭头表示具有相关关系。

下面用一个具体的实例来看一下路径图，然后在此基础上写出结构方程模型。这是惠顿（Wheaton）等人在 1977 年给出的一个广为人知的例子，是一个测度“神经错乱平稳性”的例子，给出的数据集中使用了伊利诺伊州农村地区 932 个人的调查数据，调查了 6 个变量：

y_1：1967 年的异常程度；

y_2：1967 年的软弱程度；

y_3：1971 年的异常程度；

y_4：1971 年的软弱程度；

x_1：受教育情况（上学年数）；

x_2：当地的社会经济指数。

假设这六个显变量中，异常程度与软弱程度是测量“神经错乱因子”的指标，而受教育情况与社会经济指数是测量“社会经济状况因子”的指标，则此分析包含三个隐变量：

η_1：1967 年的神经错乱因子；

η_2：1971 年的神经错乱因子；

ξ_1：社会经济状况。

图 11—1 是反映这些变量关系的路径图。这一模型是由本特勒（Bentler）首先分析，后又由乔瑞斯考格（Joreskog）和索尔波姆（Sorbom）略作修改的结构模型。

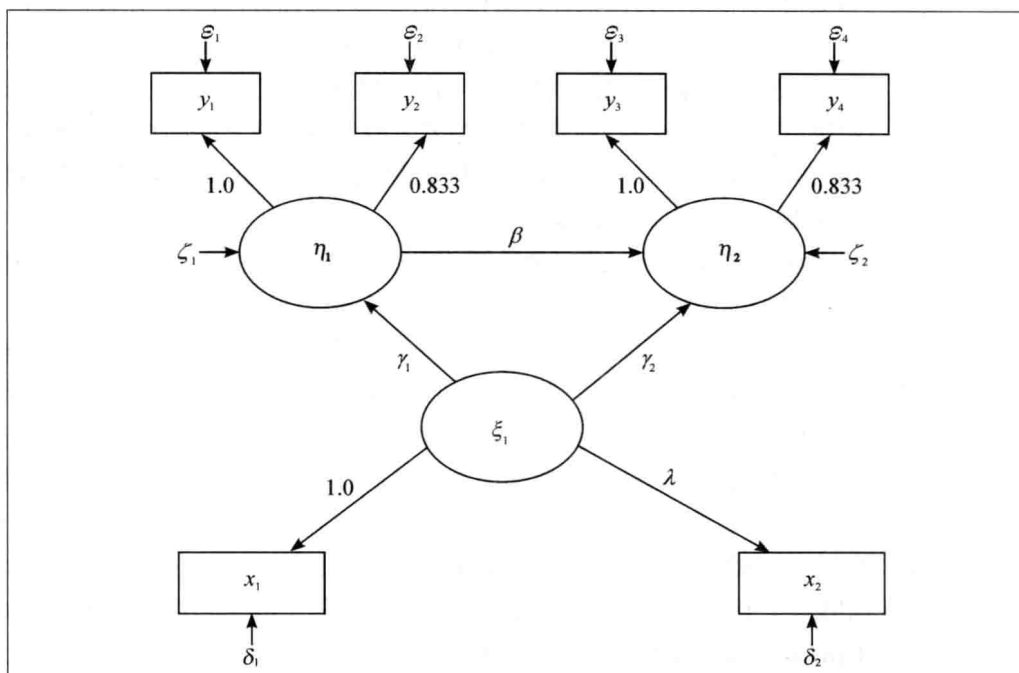


图 11—1

路径图实际上提供了一个假设模型，它体现了隐变量与隐变量之间、隐变量与显变量之间（包括内生隐变量与显变量和外生隐变量与显变量之间）可能存在的关系，而且这种关系的具体程度可以通过路径系数来反映。在这些变量中，显变量是可以观测的，而每个隐变量都对应着几个显变量，如图 11—1 中的隐变量“社会经济状况”就对应着两个显变量“受教育情况”和“社会经济指数”。给出路径图后，我们就可以对这些假设的结构关系，利用显变量的数据，通过建立结构方程模型，进一步检验模型假设的合理性并确定模型中的路径系数。

11.1.2　结构方程模型的结构

结构方程模型一般由测量方程（measurement equation）和结构方程（structural equation）两部分构成。测量方程描述隐变量与指标之间的关系；结构方程则反映隐变量之间的关系。指标含有随机误差和系统误差。前者指测量上的不准确性行为，后者反映指标同时测量隐变量以外的特性。随机误差和系统误差统称为测量误差，但隐变量却不含这些误差。

（1）测量模型。对于指标与隐变量之间的关系，通常写成如下测量方程：

$$\boldsymbol{y}=\boldsymbol{\Lambda}_y\boldsymbol{\eta}+\boldsymbol{\varepsilon} \tag{11.1}$$

$$\boldsymbol{x}=\boldsymbol{\Lambda}_x\boldsymbol{\xi}+\boldsymbol{\delta} \tag{11.2}$$

方程（11.1）和（11.2）为测量模型，表示隐变量与显变量之间的关系，即由显变量来定义隐变量。其中，方程（11.1）将内生隐变量 $\boldsymbol{\eta}$ 连接到内生标识，即显变量 $\boldsymbol{y}$；方程（11.2）将外生隐变量 $\boldsymbol{\xi}$ 连接到外生标识，即显变量 $\boldsymbol{x}$。矩阵 $\boldsymbol{\Lambda}_x$ 和 $\boldsymbol{\Lambda}_y$ 分别为反映 $\boldsymbol{x}$ 对 $\boldsymbol{\xi}$ 和 $\boldsymbol{y}$ 对 $\boldsymbol{\eta}$ 关系强弱程度的系数矩阵，可以理解为相关系数，也可以理解为因子分析中的因子载荷。$\boldsymbol{\varepsilon}$ 和 $\boldsymbol{\delta}$ 分别是 $\boldsymbol{y}$ 和 $\boldsymbol{x}$ 的测量误差。在结构方程模型中，测量误差满足假设：1）均值为 0，方差为常数。2）不存在序列相关。3）与外生、内生隐变量不相关。4）与结构方程误差不相关。

（2）结构模型。对于隐变量之间的关系，可写成如下结构方程：

$$\boldsymbol{\eta}=\boldsymbol{B}\boldsymbol{\eta}+\boldsymbol{\Gamma}\boldsymbol{\xi}+\boldsymbol{\zeta} \tag{11.3}$$

方程（11.3）为结构方程模型，反映了隐变量之间的关系。内生隐变量和外生隐变量之间通过系数矩阵 $\boldsymbol{B}$ 和 $\boldsymbol{\Gamma}$ 以及误差向量联系起来，其中，$\boldsymbol{\Gamma}$ 代表外生隐变量对内生隐变量的影响，$\boldsymbol{B}$ 代表内生隐变量之间的相互影响，$\boldsymbol{\zeta}$ 为结构方程的误差项。结构方程的误差项应满足：1）均值为 0，方差为常数。2）不存在序列相关。3）与外生隐变量不相关。

根据结构方程的定义，上例中路径图的结构方程模型可以表示如下：

$$y_1=1.0\eta_1+\varepsilon_1$$

$$y_2=0.833\eta_1+\varepsilon_2$$

$$y_3=1.0\eta_2+\varepsilon_3$$

$$y_4=0.833\eta_2+\varepsilon_4$$

$$x_1=1.0\xi_1+\delta_1$$

$$x_2=\lambda\xi_1+\delta_2$$

$$\eta_1=\gamma_1\xi_1+\zeta_1$$

$$\eta_2=\beta\eta_1+\gamma_2\xi_1+\zeta_2$$

11.1.3 结构方程模型的优点

（1）能同时处理多个因变量。结构方程模型可同时考虑并处理多个因变量。而回归分析中，只能处理一个因变量，如果有多个因变量需要处理，则需要分别计算，这样在计算一个因变量时，就忽略了其他因变量的存在及影响。

（2）允许自变量和因变量均包含测量误差。从测量方程中可看到，很多变量如学业成绩、社会经济地位等隐变量的观察值不能用单一指标来测量，往往还包含大量的测量误差。从结构方程模型的特点看出，结构方程分析允许自变量和因变量均含有测量误差。而回归分析只允许因变量存在测量误差，假定自变量没有误差。

（3）估计整个模型的拟合程度。在传统的路径分析中，我们只估计每条路径变量间关系的强弱。在结构方程分析中，可以通过结构方程软件 LISREL 计算出的多个拟合参数值，判断不同模型对同一个样本数据的整体拟合程度，从中选取最精确的模型描述样本数据呈现的特征。

11.2 结构方程模型的构建

由上一节介绍的结构方程模型的结构模式可以看出，结构方程模型一般由测量方程和结构方程两部分构成。要很好地完成这两部分的构造，关键是利用结构方程模型中分析变量（包括显变量和隐变量）的关系，根据相关领域的专业知识和研究目的构建理论模型，然后用测得的数据去验证这个理论模型的合理性。下面仍以惠顿等人研究的“神经错乱平稳性”这一经典实例来说明模型的建立过程。

惠顿等人在 1977 年研究了“神经错乱平稳性”问题，在其给出的数据集中使用了伊利诺伊州农村地区 932 个人的调查数据，调查了 6 个变量：y_1，y_2，y_3，y_4，x_1，x_2，上述 6 个变量可以实际观测到，因此为显变量。此分析包含 3 个隐变量：η_1，η_2，ξ_1。关于以上 9 个变量的含义参见 11.1 节。隐变量社会经济状况（ξ_1，这里用 ξ 表示）影响着隐变量 1967 年的神经错乱因子（η_1）和 1971 年的神经错乱因子（η_2）；隐变量 1967 年的神经错乱因子受随机因素（ζ_1）的影响；隐变量 1971 年的神经错乱因子不仅受隐变量 1967 年的神经错乱因子（η_1）的影响，同时受随机因素（ζ_2）的影响。我们可以根据这些假设的结构关系，利用显变量的数据，建立结构方程模型。

（1）结构方程的建立。根据模型的假设条件可以建立反映隐变量间关系的路径图（见图 11—2）。

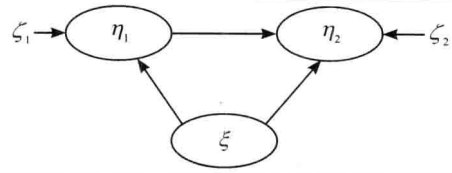


图 11—2

根据变量间的路径图，可以用数学方程表述这种关系：

$$\eta_1 = \gamma_1 \xi + \zeta_1$$

式中，γ_1 代表社会经济状况（ξ）对 1967 年的神经错乱因子（η_1）的影响；ζ_1 代表随机因素。同理，可以得出

$$\eta_2 = \beta\eta_1 + \gamma_2 \xi + \zeta_2$$

式中，β 代表 1967 年的神经错乱因子（η_1）对 1971 年的神经错乱因子（η_2）的影响；γ_2 代表社会经济状况（ξ）对 1971 年的神经错乱因子（η_2）的影响；ζ_2 代表随机因素。

若方程中含有多个内生变量，同样可以写出其他内生变量的理论数学表达式。各个内生变量的结构关系可以写成以下矩阵方程：

$$\boldsymbol{\eta} = \boldsymbol{B\eta} + \boldsymbol{\Gamma\xi} + \boldsymbol{\zeta}$$

上面就是“神经错乱平稳性”理论模型的结构方程表达形式，也就是结构方程模型中的结构模式。

（2）测量方程的建立。根据模型的假设条件可以建立反映显变量和隐变量关系的路径图，如图 11—3、图 11—4、图 11—5 所示。

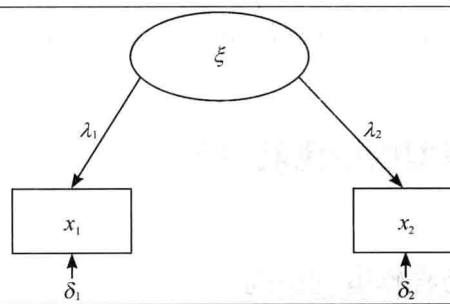


图 11—3

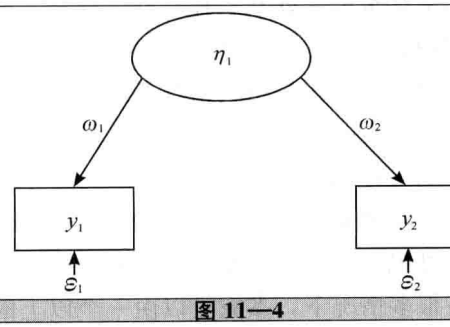


图 11—4

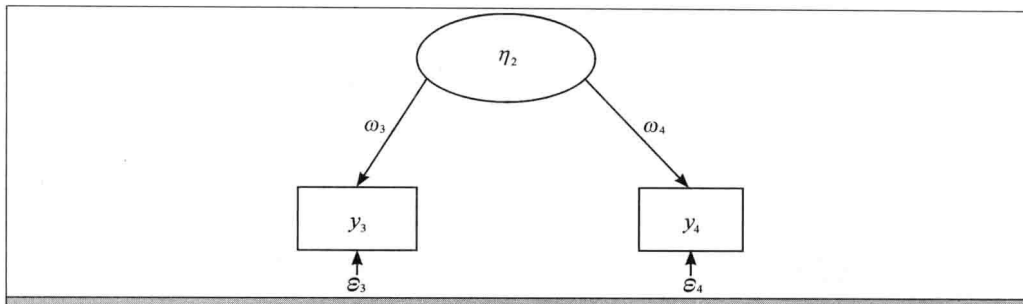


图 11—5

其中，δ 和 ε 分别为 x 和 y 的测量误差。

以上测量关系可用以下方程表示：

$$x_1=\lambda_1\xi+\delta_1$$
$$x_2=\lambda_2\xi+\delta_2$$
$$y_1=\omega_1\eta_1+\varepsilon_1$$
$$y_2=\omega_2\eta_1+\varepsilon_2$$
$$y_3=\omega_3\eta_2+\varepsilon_3$$
$$y_4=\omega_4\eta_2+\varepsilon_4$$

这种测量关系可以写成以下矩阵方程形式：

$$\boldsymbol{x}=\boldsymbol{\Lambda}_x\boldsymbol{\xi}+\boldsymbol{\delta}$$
$$\boldsymbol{y}=\boldsymbol{\Lambda}_y\boldsymbol{\eta}+\boldsymbol{\varepsilon}$$

至此，测量方程和结构方程都已建立，整个结构方程模型也得以建立。当然，初始建立的理论模型有可能不是较理想模型，需要在数据的拟合过程中反复修改，直到建立较理想模型。关于模型的识别、估计、修正将在后面章节中逐一介绍。

11.3 结构方程模型的识别和估计

11.3.1 结构方程模型的识别

模型识别的主要任务就是在初始模型建立之后，考虑模型中的每一个未知参数是否能由观测数据得到唯一解。根据结构方程组的个数与未知参数个数之间的关系，模型可分为恰好识别结构模型（just-determined constructural model）、识别不足结构模型（under-determined constructural model）和过度识别结构模型（over-determined constructural model）。

一个未知参数可以由显变量的协方差矩阵的一个或多个元素的代数函数来表达，就称这个参数可识别。如果模型中的所有未知参数都是可识别参数，这个模型就是可识别的。在很多情况下，参数可以由一个以上的不同函数来表达，这种参数

称为过度识别参数。过度识别参数可以由不同函数来求解，如果模型正确的话，该参数应该有唯一解。当可识别模型不存在过度识别参数时，称模型为恰好识别结构模型；当可识别模型至少存在一个过度识别参数时，称模型为过度识别结构模型。识别不足结构模型指的是模型中至少有一个不能识别的参数。该模型无论样本量多大，仍然不能识别。

识别不足结构模型和恰好识别结构模型是不能令人满意的，因为我们无法得到确定解，即使能得到唯一解，也无法识别模型在统计上合理与否。只有当结构方程个数多于未知参数时，人们才可以在待估参数上附加不同的条件以使所求得的参数满足统计学要求。

11.3.2　结构方程模型的估计

当判断出一个模型可识别之后，下一步工作就是根据显变量的方差和协方差对参数进行估计。传统的统计方法，如回归分析，分析问题的着眼点在于追求尽量缩小每一个观测的真实值与拟合值之间的差异。结构方程模型的不同之处在于，其目标是尽量缩小样本协方差阵与由模型估计出的协方差阵之间的差异。

结构方程模型的估计是从样本的协方差矩阵出发，该矩阵是未知参数的一套函数。将固定参数值和自由参数值的估计值代入结构方程，从中推导出理论的协方差矩阵 $\boldsymbol{\Sigma}$。如果模型正确的话，推导出的协方差矩阵应该十分近似于样本协方差矩阵。

结构方程模型最常见的估计方法有：没有加权的最小二乘法（ULS）、广义最小二乘法（GLS）和最大似然估计（ML）。每种计算方法都是要找到参数估计，以使拟合损失函数达到最小。拟合损失函数是度量观测的样本协方差阵和参数估计给出的预测协方差阵之间差异程度的函数。ML 方法对于多数应用问题特别是考虑到统计问题时，是首选方法。GLS 通常得出与 ML 方法类似的结论。ML 和 GLS 这两种方法在不考虑协方差阵的尺度时是适用的，而且需要显变量是连续的和多元正态的。这是因为变量的偏态或峰度会导致很差的估计、极其不正确的标准误和较高的 χ^2 值。ULS 方法适用于仅当这些变量在可比较的尺度上被测量时得到的协方差阵，否则 ULS 方法使用相关阵。若预测或观测的协方差阵是奇异的，则不能使用 ML 和 GLS 这两种方法，这时要么去掉线性相关变量，要么用 ULS 方法。

11.4　结构方程模型的评价和修改

11.4.1　结构方程模型的评价

在实际工作中，专业人员提出了一个模型，模型的正确与否必须经过检验。结

构方程模型进行模型检验的主要思路就是将实际收集到的样本值运用于假设的模型，通过建立结构方程组解出未知参数，并根据未知参数求解各个显变量之间的模型相关系数矩阵；而同时通过样本可直接算出这些显变量间的样本相关系数矩阵。理论上，上述两个相关系数矩阵应该相等，因此，我们构造统计量或指标来检验其拟合程度。

结构方程模型中，样本量不能过少。各种研究表明，样本量小于 100 时，即使正态分布严格满足，也很容易出现不收敛，或计算结果很反常，或解的精确度很差，等等。因此，足够的样本量是必需的。另一个值得注意的问题就是，χ^2 检验要求样本量在 100～200 之间，样本量太小或过大都不适合。模型拟合的好坏主要通过以下指标来衡量：

- 拟合准则 F（fit criterion，越接近 0，说明拟合越好）。
- 拟合优度指标 GFI（goodness of fit index，最大值为 1，越接近 1 越好）。
- 调整自由度的 GFI 的指标 AGFI（adjusted goodness of fit index，此值越大越好）。
- 均方根残差 RMR（root mean square residual，此值越小越好）。
- 本特勒的比较拟合指数 CFI（comparation fit index，越接近 1，说明拟合越好）。
- AIC 准则（Akaike's information criterion，AIC 达到最小值时最好）。
- CAIC 准则（consistent Akaike's information criterion，同 AIC 一样，达到最小值时最好）。
- SBC 准则（Schwarz's Bayesian criterion，此值越小越好）。
- 正规指数 NI（normed index，越接近 1，说明拟合越好）。
- 非正规指数 NNI（non-normed index，越接近 1，说明拟合越好）。
- 节俭指数（parsimonious index，越大，说明拟合越好）。
- 临界指数 CN（critical n，越大，说明拟合越好）。

11.4.2 结构方程模型的修正

对模型进行评价的目的不是简单地接受或拒绝一个假设的理论模型，而是根据评价的结果来寻求一个理论上和统计上都有意义的相对较好的模型。一个好的模型应具备以下几个条件：（1）测量模型中的因子负荷和因果模型中的结构系数的估计值都有实际意义和统计学意义。（2）模型中所有固定参数的修正指数（MI）不要过高。（3）几种主要的拟合指数达到了一般要求。（4）测量模型和因果模型中的主要方程的决定系数（coefficient of determination）R^2 应足够大。（5）所有的标准拟合残差都小于 1.96。

如果我们希望看到的上述情况中的一种或几种没有实现，可以根据具体的结果做出如下改变：（1）当模型评价结果中含有没有实际意义或统计学意义的参数时，可以将这些参数固定为零，即删除相应的自由参数。（2）当模型的某个或某几个固

定参数的修正指数（MI）比较大时，原则上每次只将最大或较大 MI 的参数改为自由参数。理由是：假设某一固定路径的 MI 原本很大，需要自由估计，但当修改其他路径后，该 MI 可能已变小，对应的路径无须再改动。因此，每次只修改一个固定路径，然后重新计算所有固定路径的 MI。但 MI 受样本容量的影响，因此，不能把 MI 的数值作为修改的唯一根据。（3）当评价结果中有较大的标准残差时，分两种情况：一是当有较大的正标准残差时，需要在模型中添加与残差对应的一个自由参数；二是当有较大的负标准残差时，则需要在模型中删除与残差对应的一个自由参数。不断添加与删除自由参数，直到所有的标准残差均小于 2 为止。（4）如果主要方程的决定系数很小，则可能是以下某个或某几个方面的原因：一是缺少重要的观察变量；二是样本量不够大；三是设定的初始模型不正确。

11.5　结构方程模型的上机实现

目前，国际上一些著名的软件公司都推出了利用结构方程模型进行统计分析的计算机应用软件和模块，例如瑞典阿帕萨拉大学的乔瑞斯考格和索尔波姆专门为进行结构方程模型分析所编写的 LISREL 软件以及我们所熟悉的 SAS 软件中的 CALIS 和 SPSS 的 Amos 等，这样，我们就可以很方便地运用结构方程模型来解决各领域的问题。根据我国统计软件的应用情况，在这里主要对国内比较流行的 SAS 软件中的 CALIS 和 LISREL 软件进行介绍。

11.5.1　结构方程模型分析流程

根据前面关于结构方程模型分析过程的分析，利用结构方程模型进行分析的结构流程图如图 11—6 所示，这是我们进行计算机实现的基础。

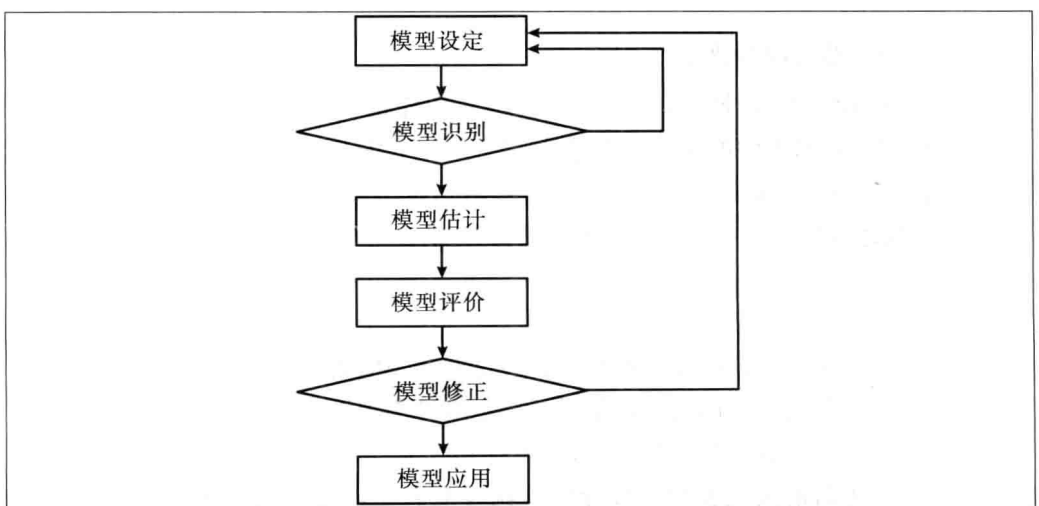


图 11—6

从这一流程图可以看出，在利用结构方程模型进行分析时，首先要对分析的实际问题进行模型设定，这一模型建立的好坏直接影响进一步的分析。要想建立一个好的模型，往往需要研究人员对研究对象有比较透彻的了解，弄清指标之间的关系。为了使模型的建立更直观，进一步写结构方程模型，往往借助于路径图。

接下来，就需要对建立的模型进行识别，看建立的模型是属于恰好识别结构模型、识别不足结构模型还是过度识别结构模型，只有可识别的模型才可以进入下一个环节。

对于可识别模型，就可以利用下面将要介绍的有关软件估计结构方程模型中的有关参数并给出有关的检验参数。

利用这些参数，就可以对模型进行评价，根据有关指标的标准，评价模型是否需要进一步修正。

如果模型不需要修正，就可以对模型进行应用。对于需要修正的模型，需要回到模型设定阶段，再按上面的过程逐步进行，直到模型不需要修正、可以应用为止。

11.5.2 SAS 中的 CALIS 过程简介

SAS/STAT 软件 6.06 以后版本新增加了线性结构方程组的协方差过程（covariance analysis of linear structural equation，CALIS），该过程可以通过协方差结构分析来估计参数并检验结构方程的正确性。

在 SAS 的 CALIS 过程中，还提供了多种模型来建立结构方程模型，如 RAM 模型、LINEQS 模型、LISREL 模型等。在此介绍 RAM 模型，其他模型请参看有关书籍。

1. 路径图的描述

CALIS 中的 RAM 语句可以方便地描述路径图，假定有 n 个显变量，按它们在 SAS 数据集中的顺序用整数 1，2，…，n 编号。每个箭头便可由路径图中它所连接的两个变量的编号确认。RAM 语句包括路径图中所有箭头的说明，说明之间用逗号隔开，每一说明项包括 3 个或 4 个数字或可选择的一个名字，顺序如下：

（1）箭头有几个。

（2）箭头指向的变量的编号，若是双箭头则为任意变量的编号。

（3）箭头出发的变量的编号，若是双箭头则取另一个变量的编号。

（4）箭头表示的系数、（协）方差值。

（5）若箭头代表的参数待估，就写上名字，这时前一个数字表示参数的初始值。若箭头代表一个常量，可以省略名字；若用名字，第 4 个数字可以省略。

2. 选择估计方法

CALIS 提供三种估计方法，可以用选项“METHOD=”来规定：

ULS　　没有加权的最小二乘估计

GLS　　广义最小二乘估计

ML　　多元正态分布的最大似然估计

没有特别规定时（即缺省时），估计方法使用 METHOD=ML，因为 ML 对于多数统计问题是首选方法。例如，对于前面已经给出的例子，如果选择 GLS 作为估计方法，可使用语句 proc calis cov data=Wheaton method=gls tech=lm edf=931，其中，proc calis 是调用 SAS 中的 CALIS 过程，选项 cov 要求对协方差阵进行分析，没有 cov 选项时则计算和分析相关阵；而data=Wheaton 是调用我们分析所用的数据库（注：这是 SAS 自带的数据库）；选项 tech=lm 代表的是使用 Levenberg-Marquandt 或 Newton-Raphson 的最优化方法，这里使用的是后者；选项 edf=931 指明了自由度的个数，即 932 个样本数据。

根据这些规定，表 11—1 中的语句给出了上例中“神经错乱”数据的 RAM 结构模型（注意表 11—1 中第五列给出的字母与图 11—1 中稍有不同）。

表 11—1

```
proc calis cov data=Wheaton tech=nr edf=931;
  Ram
            1       1       7        1.000          ,
            1       2       7        0.833          ,
            1       3       8        1.000          ,
            1       4       8        0.833          ,
            1       5       9        1.000          ,
            1       6       9        0.500    Lamb  ,
            1       7       9       -0.500    Gam1  ,
            1       8       7        0.500    Beta  ,
            1       8       9       -0.500    Gam2  ,
            2       1       1        3.000    The1  ,
            2       2       2        3.000    The2  ,
            2       3       3        3.000    The1  ,
            2       4       4        3.000    The2  ,
            2       5       5        3.000    The3  ,
            2       6       6        3.000    The4  ,
            2       1       3        0.200    The5  ,
            2       2       4        0.200    The5  ,
            2       7       7        4.000    Psi1  ,
            2       8       8        4.000    Psi2  ,
            2       9       9        6.000    Phi   ;
  Vnames 1 F1-F3,
         2 E1-E6 D1-D3
run;
```

其中，Vnames 1 F1－F3，2 E1－E6 D1－D3 给出了隐变量和误差变量的名字。

以 Ram 语句的第一行为例来说明路径图的描述：该箭头为单箭头，从变量 7

出发，指向变量 1，箭头表示的系数值为 1。

运行上面的程序，得到输出结果 11—1（这里仅给出了检验结果，读者运行时还会看到反映变量关系的其他结果，在此不一一列出），可以通过这一结果对模型的正确性进行判断。

输出结果 11—1

The CALIS Procedure

Covariance Structure Analysis: Maximum Likelihood Estimation

Fit Function	0.014 5
Goodness of Fit Index (GFI)	0.995 3
GFI Adjusted for Degrees of Freedom (AGFI)	0.989 0
Root Mean Square Residual (RMR)	0.228 1
Parsimonious GFI (Mulaik, 1989)	0.597 2
Chi-Square	13.485 1
Chi-Square DF	9
Pr > Chi-Square	0.141 9
Bentler's Comparative Fit Index	0.997 9
Normal Theory Reweighted LS Chi-Square	13.280 4
Akaike's Information Criterion	−4.514 9
Schwarz's Bayesian Criterion	−48.050 9
McDonald's (1989) Centrality	0.997 6
Bentler & Bonett's (1980) Non-normed Index	0.996 5
Bentler & Bonett's (1980) NFI	0.993 7
James, Mulaik, & Brett (1982) Parsimonious NFI	0.596 2
Z-Test of Wilson & Hilferty (1931)	1.075 4
Bollen (1986) Normed Index Rho1	0.989 5
Bollen (1988) Non-normed Index Delta2	0.997 9
Hoelter's (1983) Critical N	1 170

11.5.3 LISREL 软件简介

LISREL（linear structural relations）是专门为进行结构方程分析而编写的统计分析软件。与 SAS 软件中的 CALIS 不同的是，LISREL 的路径图可以在输出结果中直观给出，并能够在图形窗口进行编辑和修改。

LISREL 能够在图形窗口进行路径图编辑和修改。用光标点击命令行的 Pathdiagram 或相应的图标，即进入图形窗口。在图形窗口命令行点击 Model，可以选择显示不同的图形。对模板图形进行修改，可以得到所需的路径图。图形窗口命令行的其他命令的用途分别是：Exit 退出该窗口，Kind 调出其他统计结果以便对路径图进行修改，Options 修改统计数值的小数位显示长度，Print 打印路径图，Zoom 对路径图进行放大和缩小，Re-estimate 是根据统计分析结果对路径图进行修改之后再次运行估计程序。

首先，需要编写并运行程序命令。LISREL 程序包含一个子程序 PRELIS，该

子程序对结构方程模型数据进行预处理。该程序包括多个指令，指示原始数据的出处以及变量信息和结果的存入。表 11—2 以程序的形式简略地给出了 PRELIS 的基本指令。

表 11—2

LISREL MODEL EXAMPLE——PRELIS	定义标题
DA NI=6 NO=821	数据 6 个输入变量，821 个观测
LA X1 X2 Y1 Y2 Y3 Y4	给出观测变量名
CO ALL	变量均为连续变量
RA FI=C：\MY DOCUMENT\AA. DAT	指出原始数据文件名称位置
OU MA=CM CM=AA. CM	输出为协方差矩阵并命名为 AA

原始数据经过预处理可以得到其协方差矩阵。根据协方差矩阵开始编写 LISREL 程序，如表 11—3 所示。

表 11—3

LISREL MODEL EXAMPLE	定义标题
DA NI=6 NO=821 MA=CM	数据 6 个输入变量，821 个观测
CM FI=C：\MY DOCUMENT\AA. CM	给出协方差阵位置
LA X1 X2 Y1 Y2 Y3 Y4	给出观测变量名
MO NY=4 NX=2 NE=0 NK=1	模型设定 4 个 Y 变量、2 个 X 变量 0 个内生潜在变量、1 个外生潜在变量
PATH DIAGRAM	输出路径图
OU ME=ML MI ND=3 IT=80	估计方法 ML、打印修正指数、小数位数 3、迭代次数 80

为了清楚起见，表 11—3 所给出的程序并没有对结构方程模型矩阵 $\boldsymbol{\Lambda}_x$，$\boldsymbol{\Lambda}_y$，$\boldsymbol{B}$，$\boldsymbol{\Gamma}$，$\boldsymbol{\Phi}_\epsilon$，$\boldsymbol{\Theta}_\delta$，$\boldsymbol{\Psi}$ 进行设定，在 LISREL 中这些矩阵的默认格式均为自由矩阵，即每个元素均为自由变量，而在实际情况下，要想使模型可识别，是不可能的。LISREL 分别用下列符号来表示上述矩阵：LA，LY，BE，GA，PH，TE，PS。在 MO 指令行可把某个矩阵设定为 **0** 矩阵或自由矩阵，如 MO NY=4 NX=2 NE=0 NK=1 LY=FI LA=FR，即将 $\boldsymbol{\Lambda}_y$ 矩阵所有元素设定为 0，将 $\boldsymbol{\Lambda}_x$ 矩阵所有元素设定为自由变量。也可以在 MO 指令行后面加上语句 GA=SD，定义 $\boldsymbol{\Gamma}$ 矩阵为下三角矩阵。如果想对某个矩阵的单个元素进行定义，例如将 $\boldsymbol{\Lambda}_x$ 的第二行第二列元素定义为 0，则采用语句 FI LA 2 2。同样，将 $\boldsymbol{\Lambda}_y$ 的第一行第二列元素定义为自由元素可采用语句 FR LY 1 2。编辑好 LISREL 命令后，只要点击命令行中的 run lisrel，就可以执行。

由于篇幅所限，对 LISREL 的介绍只限于此。有兴趣的读者请参阅有关的 LISREL 软件书籍或该软件命令行中的“帮助”窗口。

11.6 一个实例[①]

为了使读者对结构方程模型的实际应用过程有一个总体的把握，本章引用了一个运用结构方程模型解决问题的真实案例供读者参考。此案例是里查德·罗森费尔德（Richard Rosenfield）、斯蒂文·麦斯内尔（Steven F. Messner）、埃里克·鲍默（Eric P. Baumer）等人将社会资本（social capital）作为隐变量，考察社会资本和社会谋杀率（homicide rate）之间的结构关系。

社会资本与谋杀率之间的假定关系如图 11—7 所示。根据罗伯特·帕特南（Robert Putnam）对社会资本的定义，社会资本指的是社会组织特点，如社会网络、社会标准、信任、为共同利益而采取的行动和合作。因此罗森费尔德等人对社会资本的评价着重考虑了两个方面：人与人之间的信任（trust）和公众对公共事业的参与（civic engagement）。为什么社会资本与社会谋杀率之间存在联系呢？主要从经典犯罪学的三个方面来考虑：社会控制力度（informal and formal social control）、道德水准下降程度（anomie）和社会压力（strain）。社会资本的变化对上述三个因素产生影响，从而影响了谋杀率的高低。对模型中人与人之间的信任和公众对公共事业的参与两个外生变量的量化采用了 GSS（general social survey）的数据，如表 11—4 所示。

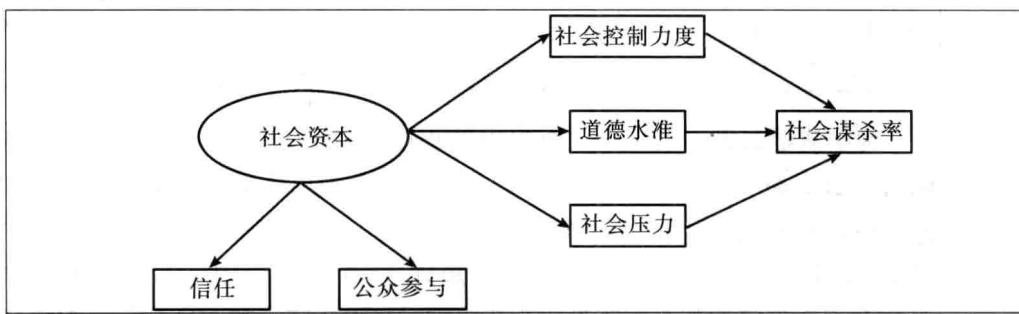


图 11—7 社会资本和社会谋杀率的假定关系

表 11—4

人与人之间的信任：
（1）Trust：“总体来说，您是否认为大多数人都是可信的，与他们打交道不必太小心？”
（2）Fair：“您认为大多数人是想占便宜，还是想公平交往？”
（3）Helpful：“您认为大多数人是乐于助人的，还是只考虑自己？”
公众对公共事业的参与：
（1）公民选举率（electoral participation）
（2）参加慈善组织（FLKs）人数

① Richard Rosenfield, Steven F. Messner, Eric P. Baumer, *Social Forces*, Chapel Hill, Sep 2001.

为了得到无偏估计，必须在模型内考虑影响谋杀率的其他因素。根据以前关于谋杀率模型的研究[①]并通过主成分分析，得到了两个主要因子。第一个因子定义为剥夺（deprivation），相应的人口结构（population structure）、年龄构成（age composition）、失业率（unemployment）、男性离婚率（male divorce）和南方州（south）构成第二个因子。

根据 LISREL 8.14，并采用极大似然估计对上述所讨论的模型进行估计。比较了两个模型：模型 1 没有将隐变量社会资本加入模型；模型 2 则考虑了社会资本。结果如表 11—5 所示。

表 11—5

解释变量	模型 1	模型 2
社会资本	—	0.219**
		(0.104)
人口结构	1.93**	1.77**
	(0.463)	(0.462)
剥夺	3.72**	3.29**
	(0.619)	(0.637)
年龄构成	0.108	−0.134
	(0.092)	(0.091)
男性离婚率	0.446*	0.462*
	(0.249)	(0.245)
失业率	−0.133	0.040
	(0.386)	(0.381)
南方州	2.01**	0.619
	(0.951)	(1.11)
调整	0.631	0.661
(N=99)		

Model 1 is saturated. Model 2: $\chi^2=62.16$ ($p<0.001$), $\chi^2/df=2.00$, RMR=0.101, GFI=0.911, CFI=0.936.

* Coefficient 1.5 times its standard error. Standard errors in parentheses.

** Coefficient 2.0 times its standard error.

— Indicates parameter not estimated.

从模型 2 的拟合指数可以看出模型拟合较好，均方根残差 RMR 也表明模型拟合得很好。将社会资本加入模型提高了谋杀率解释的方差。接下来要做的是考虑一个递归模型，即认为社会资本和谋杀率之间的作用是相互的，并且在新模型中对老模型做了一些修改，将不显著的变量年龄构成和失业率去掉，结果如表 11—6 所示。路径图如图 11—8 所示。

① 详见 Messner & Rosenfield (1998)。

表 11—6

解释变量	谋杀率	社会资本
谋杀率	—	−0.170
社会资本	−0.147*	—
	(0.096)	
人口结构	1.84**	—
	(0.447)	
剥夺	3.39**	−1.03
	(0.642)	(1.31)
年龄结构	−0.128	−1.150
	(0.090)	(0.155)
男性离婚率	0.463*	0.619
	(0.242)	(0.439)
失业率	−0.061	0.406
	(0.377)	(0.323)
南方州	0.965	−6.54**
	(1.12)	(1.80)
读报率	—	0.227**
	—	(0.055)
国外出生率	—	−0.236**
	—	(0.102)
调整 R^2	0.662	0.734
(N=99)		

χ^2=92.83 (p<0.001)，χ^2/df=2.26，RMR=0.114，GFI=0.891，CFI=0.905.

* Coefficient 1.5 times its standard error. Standard errors in parentheses.

** Coefficient 2.0 times its standard error.

— Indicates parameter not estimated.

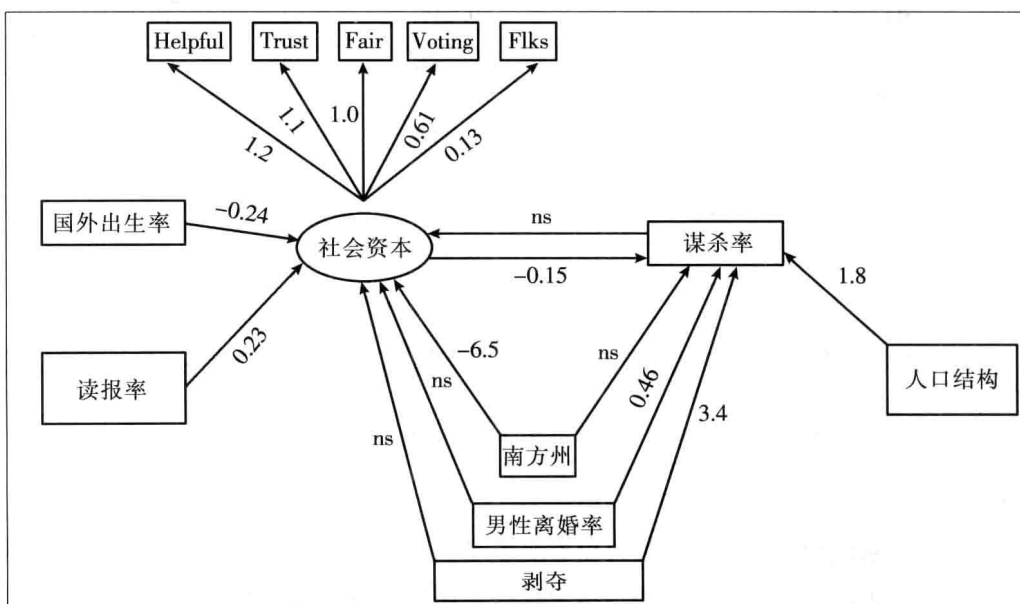


图 11—8

Standardized parameters in parentheses. Social Capital is a latent construct measured by the indicators helpful, trust, fair, Flks membership and voting rates, respectively. For clarity of presentation, the nonsignificant paths to social capital and homicide from divorce, age composition, and unemployment are not shown; correlations between exogenous variables also are omitted.

* Coefficient 1.5 times its standard error.

** Coefficient 2.0 times its standard error.

"ns" indicates a nonsignificant coefficient.

□ 参考文献

[1] Jan-Bernd Lohmollor. *Latent Variable Path Modeling With Partial Least Squares*. Physica-Verlag Heidelberg, 1989

[2] 郭志刚. 社会统计分析方法——SPSS 软件应用. 北京：中国人民大学出版社，1999

[3] Bollen, Kenneth and J. Scott Long. *Testing Structural Equation Models*. New Bury Park: Sage, 1993

[4] Joreskog, K. G. and D. Sorbom. *LISREL* 8: *Structural Equation Modeling with the SIMPLIS Command Language*. Chicago: Scientific Software, 1993

[5] 许健，何晓群. 中国经济增长源泉的量化研究. 经济经纬，2000 (6)

[6] 侯杰泰，温忠麟，程子娟. 结构方程模型及其应用. 北京：教育科学出版社，2004

[7] 张家放. 医用多元统计方法. 武汉：华中科技大学出版社，2002

□ 思考与练习

1. 简述结构方程模型与路径分析的联系与区别。
2. 简述结构方程模型的建模分析过程。
3. 试运用现成软件建立某个实际问题的结构方程模型，并进行分析。

第 12 章
Chapter 12 联合分析

学习目标

1. 解释联合分析的多种实际用途；
2. 掌握为简单的联合分析构造试验设计方案；
3. 评价预测变量及其每个水平在影响消费者决策上的相对重要性；
4. 掌握将联合分析结果应用于消费者对新的属性组合的决策。

联合分析（conjoint analysis），早期称为联合衡量（conjoint measurement），是 1964 年由数理心理学家 R. 卢斯（R. Luce）和统计学家 J. 图基（J. Tukey）首先提出的。1971 年由 P. 格林（P. Green）和劳（Rao）引入市场营销领域，成为描述消费者在多个属性的产品或服务中作出决策的一种重要方法。1978 年F. 卡莫恩（F. Carmone）等人将联合衡量改称为联合分析。20 世纪 80 年代联合分析在许多领域得到了广泛的认可和应用。20 世纪 90 年代应用更加深入，涉及许多研究领域。本章将对联合分析的理论及其在市场研究中的应用做一些探讨。

12.1 联合分析的基本理论和方法

联合分析与方差分析、列联表、因子分析等统计方法既有联系又有区别。方差分析中影响试验指标的因素通常是人们可以控制的因素，如反应温度、溶液浓度等。方差分析的试验指标的值应该是定量数据。它可以提供作用于因变量的各种因素效用的检验，另外它还包括不可解释的误差部分。列联表主要是对有序类数据或定性性质数据（或计数数据）进行分析，研究者通常对列联表变量的独立性或关联程度感兴趣。而因子分析是利用降维思想，用少数几个因子描述多个变量间的协方

差关系。联合分析对数据要求较低，定性数据和定量数据均可使用。联合分析的目的在于分解出各个成分的效用或重要性。

联合分析是在已知受测者对某一受测体集合（a set of stimuli）整体评估结果（overall evaluation）的情形下，通过分解的方法去估计其偏好结构的一种分析方法。在联合分析中，受测体是由研究人员事先依照某种因子结构加以设计的。联合分析的目的在于将受测者的整体反应加以分解，从受测者对受测体的整体评估结果中估计每一受测体成分的效用。联合分析是多变量分析技术中的一种相依方法。M. 安蒂拉（M. Anttila）等人曾指出联合分析具有以下优点：（1）联合分析既可以分析度量属性（如价格）的重要性，又可以分析非度量属性（如品牌名称）的重要程度。（2）资料收集的程序简单易行，受测者只需要对受测体进行排序（rank）或者评分（score）。联合分析对受测者只做很少的要求，就可得到相当可靠的资料。（3）联合分析要求受测者考虑各个属性之间的兑换（trade-off），比直接询问受测者其理想点（ideal-point）的属性水平及属性重要性更切合实际。（4）联合分析所求出的成分效用值可供进行尺度不同的属性或是更基本的非度量属性的直接比较，而这些比较因素正是人们选购决策所面临的真实问题。

首先来看联合分析在公寓调查中应用的一个实例。

例 12—1

联合分析在消费者偏好结构的调查分析中有很重要的用途。这里使用联合分析方法分析消费者对出租公寓的偏好结构。在对出租公寓进行调查时，选择了 6 个属性，每个属性有 3 种水平，如表 12—1 所示。

表 12—1　　公寓的属性和各属性的水平描述

属性 \ 水平	1	2	3
1. 从公寓到公司的乘车时间	15 分钟以内	15～30 分钟	30 分钟以上
2. 公寓周围的噪音水平	非常安静	一般	极其嘈杂
3. 公寓所在地的安全情况	非常安全	一般安全	不安全
4. 公寓情况	全部粉刷过	仅厨房粉刷过	条件不好
5. 居住/进餐房间大小	7/9 平方米	5/7 平方米	3/5 平方米
6. 月租金（包括用具）	150～300 元	300～500 元	500 元以上

这里首先介绍数据收集的几种常用方法：（1）二因素法，又称兑换法（trade-off approach）。受测者每次只对一对属性各水平的不同组合进行评估，排列好顺序，然后再考虑评估另一对属性。比如，本例中可先考虑对乘车时间和噪音水平的 9 种组合的偏好顺序，再考虑对乘车时间与安全情况的 9 种组合的偏好顺序……二因素法每次只评估一对属性，需要评估的次数较多，它仅适用于属性和水平均较少

的情况。(2) 整体轮廓法 (full-profile approach)，它是最常用的一种表现方法，因为它较接近现实，还可以通过部分因子设计减少比较的个数。它在受测体卡片中列举所有的重要属性，并由各属性中的某一水平共同组成一个受测体。受测者对由此构成的受测体排列偏好顺序或者评分。(3) 两两比较法 (pairwise comparison approach)，它将前两种方法结合起来。两两比较是指两个轮廓 (profile) 的比较。这里轮廓并不包含所有的属性，而是一次选择一些属性。两两比较法将一部分属性提出，根据提取属性的水平形成一些轮廓，与二因素法相似，对轮廓组合进行比较。但是，在二因素法中评估的组合是属性，而在两两比较法中评估的组合是具有多重属性的轮廓。在本例中，我们可以使用这三种方法收集数据如下：

兑换法：

		因子 1：房间大小		
		水平 1：7/9 平方米	水平 2：5/7 平方米	水平 3：3/5 平方米
因子 2：月租金	水平 1：150～300 元			
	水平 2：300～500 元			
	水平 3：500 元以上			

整体轮廓法：

从公寓到公司的乘车时间：30 分钟
公寓周围的噪音水平：非常安静
公寓所在地的安全情况：一般安全
公寓情况：条件不好
房间大小：7/9 平方米
月租金：150～300 元

两两比较法：

从公寓到公司的乘车时间：30 分钟
公寓所在地的安全情况：一般安全
月租金：150～300 元

与

从公寓到公司的乘车时间：15 分钟以内
公寓所在地的安全情况：非常安全
月租金：500 元以上

根据表 12—1 中对属性和水平的描述，调查中若采取析因设计，将有 729(即 $3^6=729$)种组合，受测者无法对 729 种组合作出理性判断并一一排序。这里需要找到一个合适的子集来代替全集，并且保持全集的某些性质。当属性个数或水平数较多时，析因设计会产生大量的组合，令受测者无法对其一一排序。部分析因设计 (fractional factorial design) 是最常用的定义受测体子集的方法。部分析因设计选择可能的受测体的一个样本，受测体的数目取决于受测者使用的合成原则。通常可以采用对称正交设计（一个因子中的每个水平出现相同的次数、水平与属性之间没有相关关系）。有关试验设计的部分参见参考文献 [6] 和 [7]。本例中的对称正交设计的结果如表 12—2 所示。

表 12—2　　公寓属性研究的对称正交设计

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
2	1	2	3	1	2	3	2	3	1	3	1	2	2	3	1	3	1	2
3	1	2	3	2	3	1	1	2	3	3	1	2	3	1	2	2	3	1
4	1	2	3	3	1	2	3	1	2	1	2	3	2	3	1	2	3	1
5	1	2	3	2	3	1	3	1	2	2	3	1	1	2	3	3	1	2
6	1	2	3	3	1	2	2	3	1	2	3	1	3	1	2	1	2	3

说明：最左列指公寓的 6 个属性。最上边共 18 栏，指 18 个受测体。其余的部分表示每个受测体对应的每一属性的水平。如第一个受测体的 6 个属性的水平均为 1。

根据调查受测者所得到的偏好顺序或评分的数据，可以对公寓的各个属性的重要程度进行分析。篇幅所限，本例只详细分析一个受测者偏好的顺序数据，其他数据的计算及分析与此相同。受测者 1 的调查数据如表 12—3 所示。

表 12—3　　受测者 1 的偏好顺序

受测体	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
偏好顺序	1	4	18	3	14	12	2	11	17	16	5	9	13	7	10	6	15	8

下面介绍成分效用（part-worth，又称为分值贡献）的含义。从经济学角度来看，商品（即我们讨论的受测体）会给人们带来满足，经济学家用“效用”（utility）这个词来描述这种满足程度。这里假设由于商品的各种属性（或重要属性，如价格、外观等）给人们带来满足，才使得商品具有效用，于是衡量各种属性（或因子）的水平的效用就用“成分效用”一词。假设一种产品或服务有 m 种属性，每种属性有 n 种水平，则通常所用的模型可表示为：

$$\begin{aligned}\text{产品}_{i,j,\cdots,n}\text{的总效用}=&\text{因子 1 水平 } i \text{ 的效用}+\text{因子 2 水平 } j \text{ 的效用}\\&+\cdots+\text{因子 } m \text{ 水平 } n \text{ 的效用}\end{aligned}\tag{12.1}$$

这里假设产品是因子 1 的水平 i，因子 2 的水平 j，…，因子 m 的水平 n 组合而成。

接下来计算受测者 1 各因子水平的成分效用值。成分效用值通过以下四步计算：（1）计算各因子水平的秩（rank）的偏差及其平方和；（2）计算标准化值，标准化值＝全部水平数/偏差平方和；（3）计算标准偏差平方，标准偏差平方＝偏差平方×标准化值；（4）计算成分效用值，成分效用值＝$\pm\sqrt{\text{标准偏差平方}}$，其符号视偏差符号而定，与偏差符号相反。表 12—4 是对受测者 1 的偏好顺序做出的分析。

表 12—4　　受测者 1（即偏好顺序 1）的效用分析

	A	B	C	D	E	F	G	H	I	J	K	L
因	水平 1	1	3	2	16	13	6	6.833	−2.67	7.111	1.627	1.276
子	水平 2	4	14	11	5	7	15	9.333	−0.17	0.028	0.006	0.08
1	水平 3	18	12	17	9	10	8	12.33	2.833	8.028	1.837	−1.36
因	水平 1	1	3	17	5	10	15	8.5	−1	1	0.229	0.478
子	水平 2	4	14	2	9	13	8	8.333	−1.17	1.361	0.311	0.558
2	水平 3	18	12	11	16	7	6	11.67	2.167	4.694	1.074	−1.04

续前表

	A	B	C	D	E	F	G	H	I	J	K	L
因	水平 1	1	12	2	5	7	8	5.833	−3.67	13.44	3.076	1.754
子	水平 2	4	3	11	9	10	6	7.167	−2.33	5.444	1.246	1.116
3	水平 3	18	14	17	16	13	15	15.5	6	36	8.237	−2.87
因	水平 1	1	14	11	16	10	8	10	0.5	0.25	0.057	−0.239
子	水平 2	4	12	17	5	13	6	9.5	0	0	0	0
4	水平 3	18	3	2	9	7	15	9	−0.5	0.25	0.057	0.239
因	水平 1	1	12	11	9	13	15	10.17	0.667	0.444	0.102	−0.32
子	水平 2	4	3	17	16	7	8	9.167	−0.33	0.111	0.025	0.159
5	水平 3	18	14	2	5	10	6	9.167	−0.33	0.111	0.025	0.159
因	水平 1	1	14	17	9	7	6	9	−0.5	0.25	0.057	0.239
子	水平 2	4	12	2	16	10	15	9.833	0.333	0.111	0.025	−0.16
6	水平 3	18	3	11	5	13	8	9.667	0.167	0.028	0.006	−0.08

对以上各列的解释是：

A 列指 6 个因子的各种水平。

B～G 列是在不同受测体中相同因子水平的秩，比如因子 1 水平 1 在受测体 1，4，7，10，13，16 中的秩分别为：1，3，2，16，13，6。

H 列是各因子水平的平均秩，比如因子 1 水平 1 的平均秩为（1＋3＋2＋16＋13＋6)/6＝6.833。

I 列是各因子水平平均秩的偏差。由于我们一共选取了 18 个水平，因此各因子水平的期望秩应为 9.5（即（1＋2＋…＋18)/18)，也就是说全部因子水平的期望秩为 9.5，则 I＝H－9.5。

J 列是偏差平方，即 $J=I^2$，并由此计算所得偏差平方和 $\sum_{i=1}^{18} J=78.67$。

K 列是标准偏差平方，计算的标准化值＝$18/\sum_{i=1}^{18} J=0.228\ 8$，于是 K＝J×0.228 8。

L 列是成分效用值，$L=\pm\sqrt{K}$，符号与 I 栏相反（由于数字舍入误差，可能 L 列的每一因子的成分效用值之和与零有些出入）。

根据计算的 6 个因子共 18 个水平的效用，可以分析属性的重要性及其间的兑换关系（trade-off relation)。属性的效用值比较如图 12—1 所示。

根据表 12—4 和图 12—1 中的各因子水平的成分效用值，可以发现因子 3，1，2 的数值较大（绝对值)，说明了这三种属性在消费者心目中的重要地位。依据 P. 格林（P. Green）和 Y. 温德（Y. Wind）的讨论，计算所得的成分效用值既为间隔尺度，各属性间又为共同尺度（common scale)，因此各属性中成分效用值最大者减去最小者，成为相对重要性的比较基础。由表 12—4 计算可得各属性重要性的结果，如图 12—2 所示。从图 12—2 中可以明显看出第 3 个因子非常重要，即受测者 1 对安全情况远比对其他情况关心，对于交通情况（乘车时间）也比较关注，噪音情况次之，对另外三种属性即公寓情况、房间大小和月租金则不是很重视。

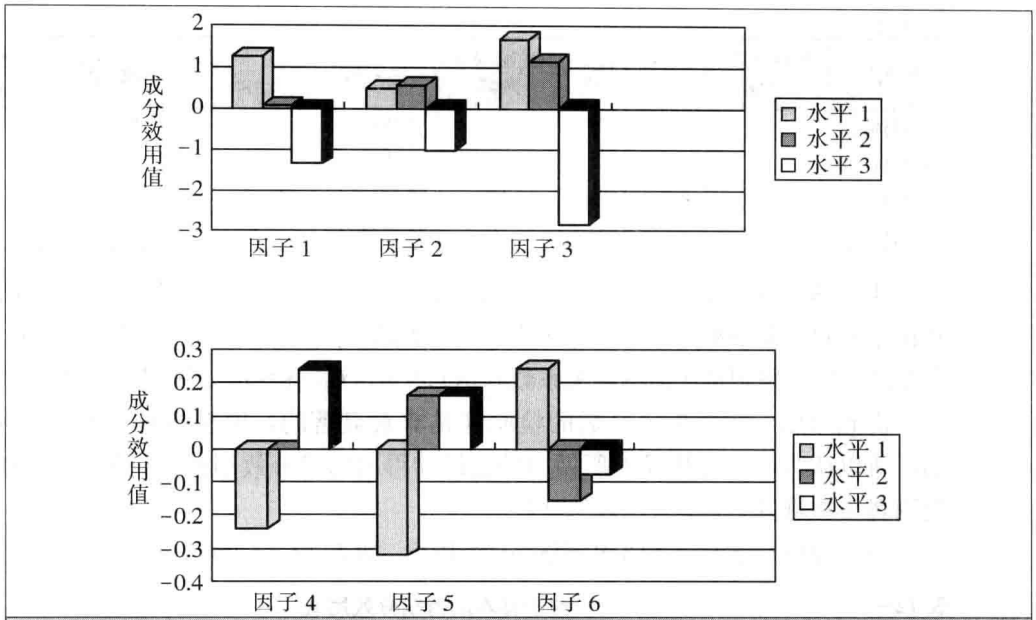


图 12—1　各因子的成分效用图

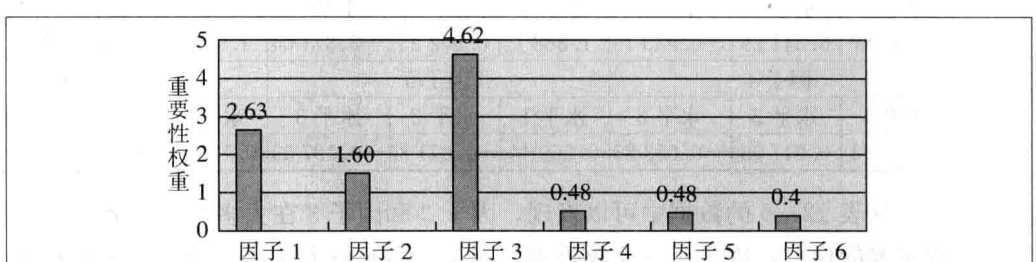


图 12—2　因子相对重要性比较

根据各因子的效用值，还可以得出各属性水平之间的兑换关系。比如对于因子 1，由水平 2 改为水平 3，效用值将下降 0.08－（－1.36）＝1.44，而将因子 2，由水平 3 上升至水平 2，效用值将提高 0.558－（－1.04）＝1.598，两者大致相当，可以互相弥补。

根据计算出的各因子水平的成分效用值，可以对 18 个受测体的偏好顺序进行估计。使用式（12.1）产品的总效用模型，用实际的秩和预测的秩相比较，可以对联合分析的模型进行拟合优度检验。这里可以采用斯皮尔曼（Spearman）的 rho 检验和肯德尔（Kendall）的 tau 检验，所得结果如表 12—5 所示。

表 12—5　　受测体效用预测值

受测体	实际偏好顺序	总效用值	预测偏好顺序	受测体	实际偏好顺序	总效用值	预测偏好顺序
受测体 1	1	3.188 96	2	受测体 6	12	－1.116 1	12
受测体 2	4	1.753 93	5	受测体 7	2	3.826 76	1
受测体 3	18	－4.942 9	18	受测体 8	11	－0.478 3	11
受测体 4	3	3.188 96	3	受测体 9	17	－3.348 4	17
受测体 5	14	－2.072 8	14	受测体 10	16	－2.870 1	16

续前表

受测体	实际偏好顺序	总效用值	预测偏好顺序	受测体	实际偏好顺序	总效用值	预测偏好顺序
受测体 11	5	2.391 72	4	受测体 15	10	1.7E−15	10
受测体 12	9	0.478 34	9	受测体 16	6	1.753 93	6
受测体 13	13	−1.435	13	受测体 17	15	−2.551 2	15
受测体 14	7	1.435 03	7	受测体 18	8	0.797 24	8

相应的斯皮尔曼和肯德尔秩检验的结果如下：肯德尔 tau 检验的预测排序值与实际排序值的相关系数高达 0.974，双尾检验显著性水平为 0.000。斯皮尔曼 rho 检验的实际排序值与预测排序值的相关系数高达 0.996，双尾检验显著性水平为 0.000。

由此可见，两个相关系数的检验都是非常显著的，模型拟合的精度是相当高的，所以认为联合分析模型所做出的假设和得出的成分效用值是合理的，可以说明受测者 1 在选择公寓时的偏好结构。

最后分析全部样本数据得到成分效用值，如表 12—6 所示。

表 12—6　　全部样本所得成分效用值

因子 1			因子 2			因子 3		
水平 1	水平 2	水平 3	水平 1	水平 2	水平 3	水平 1	水平 2	水平 3
0.538 87	0.341 29	−0.880 16	1.868 1	0.502 95	−2.371 04	1.580 7	0.538 87	−2.119 57
因子 4			因子 5			因子 6		
水平 1	水平 2	水平 3	水平 1	水平 2	水平 3	水平 1	水平 2	水平 3
0.035 93	0.017 96	−0.053 89	0.269 44	−0.071 85	−0.197 59	0.125 74	0.053 89	−0.179 62

从表 12—6 的数据中可以发现，因子 2 和因子 3 在大多数消费者心目中占有非常重要的地位，因子 1 也比较重要，因子 5 和 6 不太重要，因子 4 最不重要。所以认为本例中消费者选择公寓时考虑最多的是安全问题和周围的噪音问题，对于公寓的翻新情况不大关心。

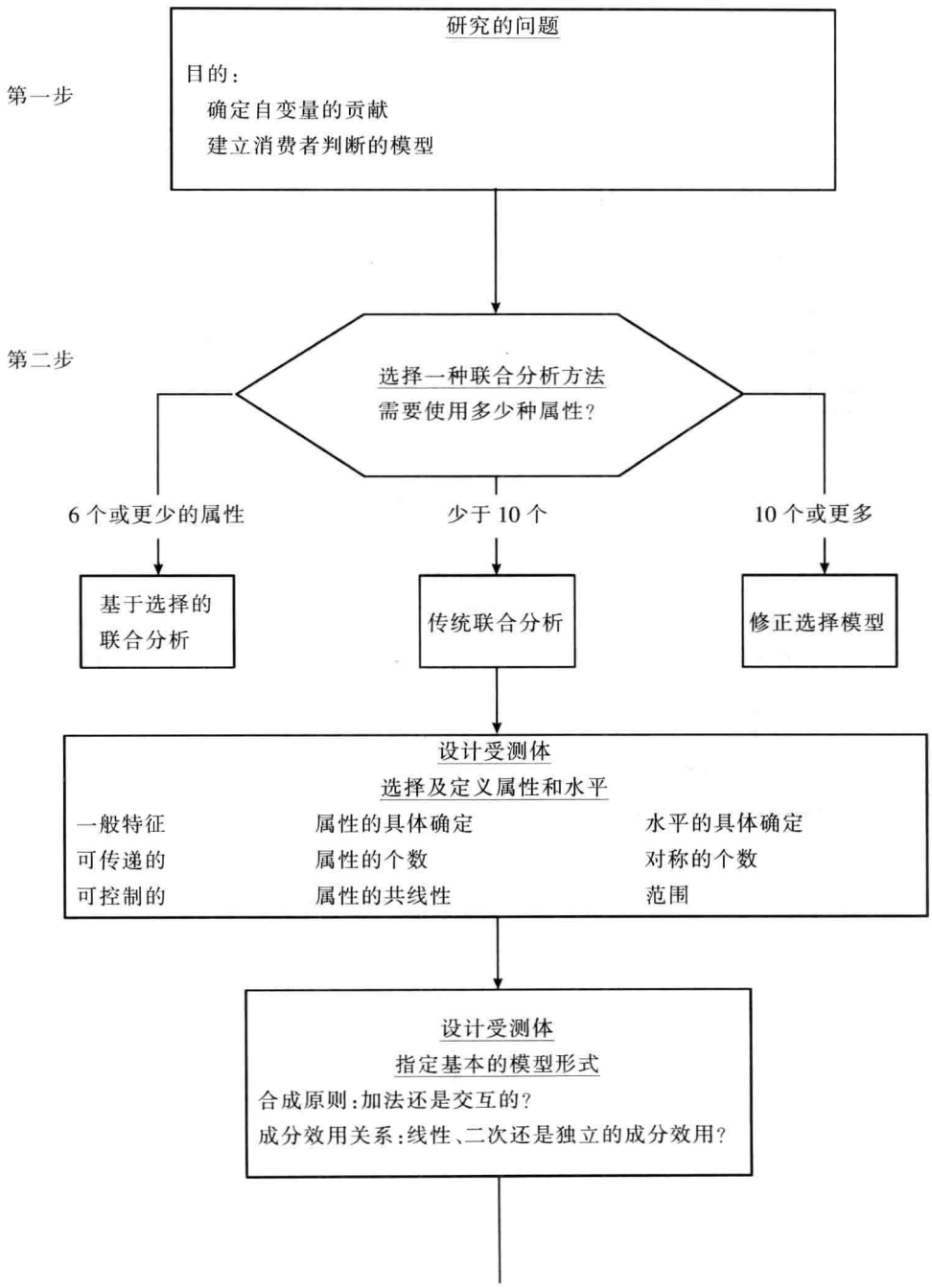
12.2　联合分析的步骤及框图

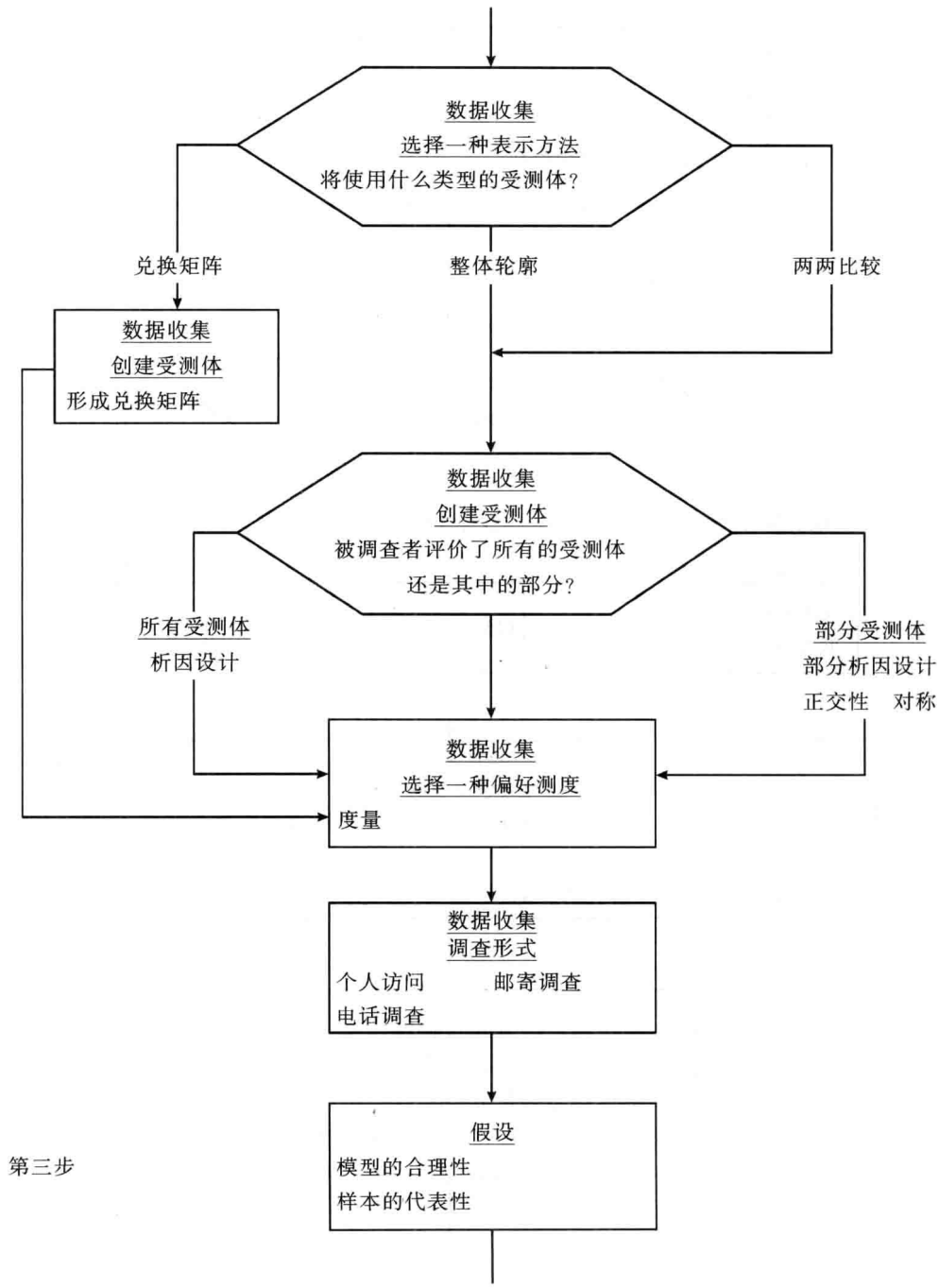
联合分析主要有以下七个步骤，逻辑框图如图 12—3 所示。

第一步：联合分析的目标

在联合分析中，试验设计在消费者决策分析中有两个目标：（1）确定自变量和它们的水平在确定消费者偏好上的贡献；（2）建立一个消费者判断的有效模型。

在此应该主要考虑以下两点：（1）定义物品的总效用。为了精确反映受测者的判断过程，所有构成或者减掉产品或服务的总效用都应考虑在内。很关键的一点是正反两方面的因素都应考虑在内。（2）指定有决定性的属性。这些属性应能最好地区分物品。某些属性虽然非常重要，但是在对物品做出选择时差别并不大，因此不应考虑在内。比如，汽车的安全性是一个很重要的属性，但是多数情况下它并不重


第一步
研究的问题
目的：
确定自变量的贡献
建立消费者判断的模型
第二步
选择一种联合分析方法
需要使用多少种属性？
6 个或更少的属性
少于 10 个
10 个或更多
基于选择的
联合分析
传统联合分析
修正选择模型
设计受测体
选择及定义属性和水平
一般特征
可传递的
可控制的
属性的具体确定
属性的个数
属性的共线性
水平的具体确定
对称的个数
范围
设计受测体
指定基本的模型形式
合成原则：加法还是交互的？
成分效用关系：线性、二次还是独立的成分效用？


数据收集
选择一种表示方法
将使用什么类型的受测体？
兑换矩阵
整体轮廓
两两比较
数据收集
创建受测体
形成兑换矩阵
数据收集
创建受测体
被调查者评价了所有的受测体
还是其中的部分？
所有受测体
析因设计
部分受测体
部分析因设计
正交性 对称
数据收集
选择一种偏好测度
度量
数据收集
调查形式
个人访问 邮寄调查
电话调查
假设
模型的合理性
样本的代表性
第三步

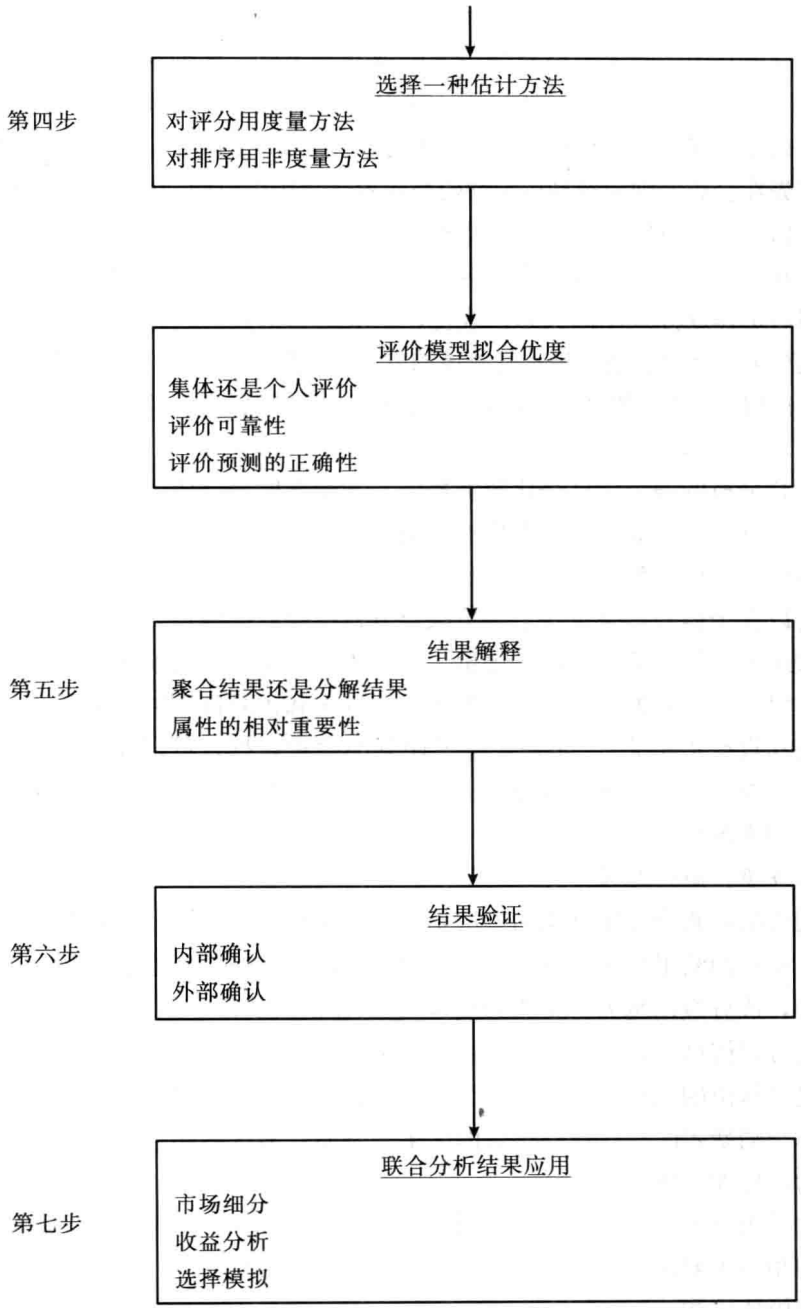


图 12—3　联合分析逻辑框图

要，因为所有汽车都达到行业标准，被认为是安全的。

第二步：联合分析的设计

联合分析的设计主要有以下几个方面：(1) 选择一种联合分析的方法。联合分析的方法有传统联合分析（traditional conjoint）、修正联合分析（adaptive conjoint）和基于选择的联合分析（choice-based conjoint）。在属性个数不同的时候，我们可以选择合适的分析方法。(2) 设计受测体，选择并定义因子和水平。在定义因子和水平时应注

意到，它们的测量应该是可被告知的（communicable）和可实行的（actionable）。可被告知的就是因子和水平容易通过实际评估来表达。比如很难描述一种香水的实际“香味”或者一种护手液的“感觉”。可实行的就是因子和水平必须可以在现实中实施，也就是说属性必须是不同的，代表可以精确实现的概念。（3）规定基本的模型形式。首先在合成原则中选择可加性模型或者交互作用模型。可加性模型简单地将每个属性的值（成分效用）加总，获得属性组合（产品或服务）的总值。交互作用模型的不同点在于它允许某些水平的组合多于或者少于它们的和，因为因子间存在交互作用。然后选择成分效用的关系：线性、二次型和单独的成分效用。（4）数据收集过程。这里需要选择合适的展示受测体的方法，主要有兑换法、整体轮廓方法、两两比较法。在构建受测体的时候，如果因子数和水平数都比较大，需要考虑部分析因设计。

第三步：联合分析的假设

联合分析的假定条件是比较宽松的。如前面所提到的，研究者必须指定模型的一般形式（可加性模型与交互作用模型）。

第四步：估计联合模型和评价总体拟合

可以在个体和集合水平上评价联合分析结果的准确性，目的在于确认模型是如何一致地预测每人给出的偏好评价。评价既可以是度量的响应，也可以是非度量的响应。对于排序数据，可以使用基于实际排序和预测排序的相关关系（比如斯皮尔曼相关和肯德尔相关）。如果获得了评分的度量数据，可以使用简单的 Pearson 相关系数。同时还可以将样本分割为分析样本和保留样本，保留样本用来验证联合分析结果的准确性。

第五步：解释结果

通常解释联合分析结果的方法是在单个水平上，也就是说，对每个受测者单独建模，模型的结果是研究每一个受测者的。最常见的解释方法是每个因子的成分效用估计，成分效用越大（正或者负），它对总效用的影响越大。

也可以解释集合的结果。不论是在个体水平上估计模型，然后汇总，还是对一组受测者做出的估计，分析可以为响应的集合拟合一个模型。但是，这个结果用于预测单个消费者时往往得到很差的结果。然而，多次的集合分析能更精确地预测集合行为，比如市场份额。

除了用成分效用估计刻画每个水平的影响，联合分析还能评估每个因子（属性）的相对重要性。每个因子的重要性可以转化为百分数，使用每个因子的极值除以全部极值之和。

第六步：联合分析的验证

联合分析的结果既可以通过内部验证，又可以通过外部验证。内部验证包括确认选择的合成准则（可加性模型与交互作用模型）是否合适。研究者通常根据经验评估选择的模型的有效性。如果分析是在集合水平上进行的，我们可以用前面已经讨论过的保留样本来评估每个个体或者受测者的一个保留样本的预测精度。

外部验证通常包括联合分析的结果预测实际选择的能力，或者说是样本的代表性的问题。在过去 20 多年中，联合分析用于很多研究，但针对其外部有效性的研

究却很少。尽管在个体水平的模型上没有抽样误差的测定，但是必须保证样本能够代表总体。当联合分析用于市场细分或者选择模拟时，这一点尤为重要。

第七步：联合分析在管理上的应用

联合分析及其对消费者偏好结构的刻画最常见的应用包括市场细分、收益分析和联合分析模拟。

个体水平上的联合分析结果最常见的应用就是用相似的成分效用或重要性对受测者分类，进行市场细分。估计的联合分析的成分效用可以单独使用，或者与其他变量（比如人口统计）一起去推导受测者的分类（在偏好结构上相似）。比如在消费者对清洁剂的偏好研究中，我们应找到某一组最关注品牌，某一组最看重成分，等等。

产品设计决策的一个补充就是对假定产品的边际收益分析。如果每个特征的成本是已知的，那么每个“产品”就可以结合市场份额和销售量来预测它的生存能力。

12.3　联合分析的上机实现

近年来，联合分析方法应用非常广泛，市场经济领域经常用它进行产品的价格性能比较和市场占有率的调查分析。

联合分析可根据产品的价格性能比分为几大类，每一类称为一个卡片（Card），然后在卡片上请顾客根据自己的偏好对这些产品打分，最后采用联合分析方法计算出各类产品的效度（utility，即效用值），从而测评出产品重要程度（average important），即比重。

联合分析可以通过 SPSS 的 Conjoint 模块和 SAS 的 MRA（market research application）模块实现。这里只介绍 SPSS 实现的方法。

下面介绍对例 12—1 通过 SPSS 的 Conjoint 模块实现的方法（见图 12—4、图 12—5 和图 12—6）。由于程序计算方法不同，其输出结果与例 12—1 有差异，但趋势是一致的，因此对结果的解释类似，读者可以自己进行比较，关于结果的解释这里不再重复。

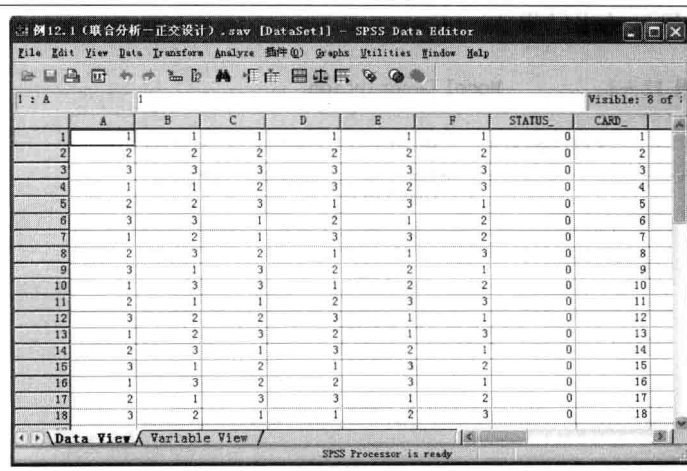
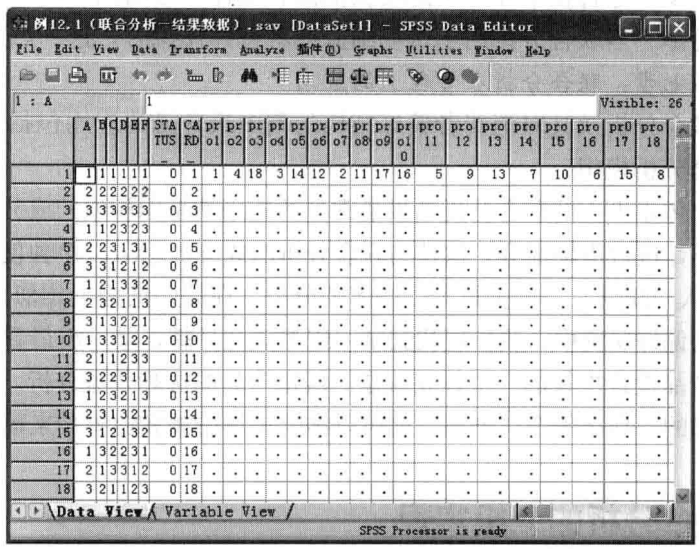


图 12—4


例12.1（联合分析一结果数据）.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze 插件(U) Graphs Utilities Window Help

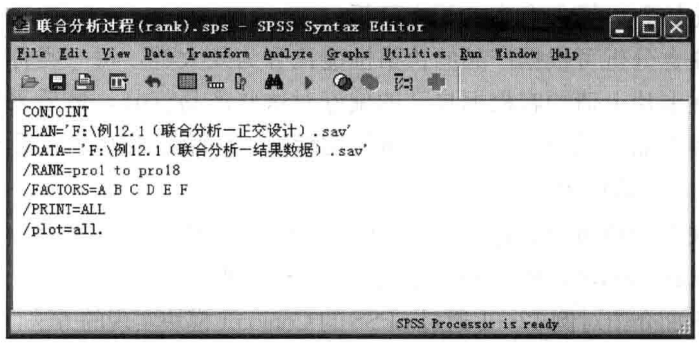
1 : A 1 Visible: 26

	A	B	C	D	E	F	STATUS_	CARD_	pro1	pro2	pro3	pro4	pro5	pro6	pro7	pro8	pro9	pro10	pro11	pro12	pro13	pro14	pro15	pro16	pro17	pro18
1	1	1	1	1	1	1	0	1	1	4	18	3	14	12	2	11	17	16	5	9	13	7	10	6	15	8
2	2	2	2	2	2	2	0	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	3	3	3	3	3	3	0	3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	1	1	2	3	2	3	0	4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
5	2	2	3	1	3	1	0	5	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	3	3	1	2	1	2	0	6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
7	1	2	1	3	3	2	0	7	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
8	2	3	2	1	1	3	0	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	3	1	3	2	2	1	0	9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
10	1	3	3	1	2	2	0	10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
11	2	1	1	2	3	3	0	11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
12	3	2	2	3	1	1	0	12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
13	1	2	3	2	1	3	0	13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
14	2	3	1	3	2	1	0	14	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
15	3	1	2	1	3	2	0	15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
16	1	3	2	2	3	1	0	16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
17	2	1	3	3	1	2	0	17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
18	3	2	1	1	2	3	0	18	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Data View / Variable View /

SPSS Processor is ready

图 12—5


联合分析过程(rank).sps - SPSS Syntax Editor

File Edit View Data Transform Analyze Graphs Utilities Run Window Help

```
CONJOINT
PLAN='F:\例12.1（联合分析一正交设计）.sav'
/DATA=='F:\例12.1（联合分析一结果数据）.sav'
/RANK=pro1 to pro18
/FACTORS=A B C D E F
/PRINT=ALL
/plot=all.
```

SPSS Processor is ready

图 12—6 Conjoint 过程

结果见输出结果 12—1。

输出结果 12—1 Model Description

	N of Levels	Relation to Ranks or Scores
A	3	Discrete
B	3	Discrete
C	3	Discrete
D	3	Discrete
E	3	Discrete
F	3	Discrete

Utilities

		Utility Estimate	Std. Error
A	15 分钟以内	2.667	.527
	15～30 分钟	.167	.527
	30 分钟以上	−2.833	.527
B	非常安静	1.000	.527
	一般	1.167	.527
	极其嘈杂	−2.167	.527
C	非常安全	3.667	.527
	一般安全	2.333	.527
	不安全	−6.000	.527
D	全部粉刷过	−.500	.527
	仅厨房粉刷过	.000	.527
	条件不好	.500	.527
E	7/9 平方米	−.667	.527
	5/7 平方米	.333	.527
	3/5 平方米	.333	.527
F	150～300 元	.500	.527
	300～500 元	−.333	.527
	500 元以上	−.167	.527
(Constant)		9.500	.373

Importance Values

A	25.781
B	15.625
C	45.313
D	4.687
E	4.687
F	3.906

Averaged Importance Score.

Correlations[a]

	Value	Sig.
pearson's R	.987	.000
Kendall's tau	.967	.000

a. Correlations between observed and estimated preferences.

根据输出的效用估计值还可以画出类似图 12—1 的各因子成分效用图，此处从略。

例 12—2

在对电脑市场的调查中，假设被调查的微型计算机的主要技术属性有：（1）品牌（如方正、联想、惠普）；（2）内存容量（分为 256MB，512MB，1 024MB）；（3）CPU（2GHz，3GHz，4GHz）；（4）每台价格（分别为 4 200 元，7 200 元，10 039 元）。构成了 4 个因子，每个因子有 3 个水平，一共有 81 种排列组合的联合模型，如表 12—7 所示。

表 12—7　　电脑品牌的价格性能

品牌	内存容量（MB）	CPU（GHz）	单价（元）
方正	256	2	4 200
联想	512	3	7 200
惠普	1 024	4	10 039

说明：这里主要介绍统计方法，数据有不符实际之处请自行调查。

如果让顾客对这 81 种组合进行选择打分，将会很麻烦，所以必须经过正交设计进行筛选。

首先，点选 SPSS→Data→Orthogonal Design→Generate，得出一种正交设计方案。表 12—8 为正交转换后的数据。

表 12—8　　正交转换后的数据

品牌	内存	CPU	单价	Status	Card
惠普	512MB	4GHz	4 200 元	Design	1
惠普	1024MB	2GHz	7 200 元	Design	2
联想	256MB	4GHz	7 200 元	Design	3
联想	1024MB	3GHz	4 200 元	Design	4
联想	512MB	2GHz	10 039 元	Design	5
方正	1024MB	4GHz	10 039 元	Design	6
方正	256MB	2GHz	4 200 元	Design	7
惠普	256MB	3GHz	10 039 元	Design	8
方正	512MB	3GHz	7 200 元	Design	9

建立产品卡片之后，每一个卡片就成了一种精品购物的选择，同时要通过问卷调查的方式了解顾客对产品的偏好选择。选择的范围应尽量广泛，如表 12—9 所示的“1～9”九个意向。

请给下面的电脑型号打分，依次为：

1. 坚决不买；2. 不买；3. 不想买；4. 说不清；5. 可能想买；6. 较想买；7. 很想买；8. 坚决要买；9. 已买。

表 12—9 所示的打分数据只是某一位顾客对产品的打分数据。应将所有顾客对

这些产品的打分数据输入表中，这里不再一一列出。

表 12—9　　**顾客购物打分表**

电脑序号	1#	2#	3#	4#	5#	6#	7#	8#	9#
打分	9	5	4	8	2	6	4	6	8

若进行联合分析，主要步骤如下：

（1）建立正交数据文件（如表 12—8）："F:\联合. sav"。

（2）建立顾客对产品的打分数据文件："F:\dafen. sav"。

（3）由于目前不能采用对话框进行联合分析，所以建立以上两个数据文件之后，需要输入以下命令编写程序：

```
File→new→syntax
```

（4）按以下形式书写联合分析程序：

程序 1（建立打分数据文件"F:\dafen. sav"）：

```
datalist/x1 to x9 1 - 9.
begin data.
954826468
end data.
```

程序 2（联合分析程序）：

```
conjoint plan = F:\联合. sav/data = F:\dafen. sav/score = x1 to x9
/plot = summary.
```

（5）运行程序：单击 Menue 中的 Run－All 命令。将运行结果整理成表 12—10。

从表 12—10 可以看出，水平值的效度表示该水平值对于顾客而言的效度。效度越高，表示该技术特性越受欢迎。从本例看，对该消费者而言，CPU 的快慢即电脑的工作频率是他所关心的，此因素的相对重要程度为 37.93%；其次为产品的价格，重要程度为 24.14%；再次为产品的品牌，重要程度为 20.69%；而对产品的内存不是太重视，只占 17.24%。那么针对此种类型的消费者，生产商可考虑在价格基本相同的情况下，着重考虑提高 CPU 的配置。

表 12—10　　**第一位顾客对产品特性的重要性及效度的测评**

产品特性 (factor)	因子比重 (averaged importance)	特性的水平值	水平值的效度 (utility)
价格	24.14%	4 200 元	1.222 2
		7 200 元	−0.111 1
		10 039 元	−1.111 1
		方正	0.222 2

续前表

产品特性 (factor)	因子比重 (averaged importance)	特性的水平值	水平值的效度 (utility)
品牌	20.69%	联想	−1.111 1
		惠普	0.888 9
		2GHz	−2.111 1
CPU	37.93%	3GHz	1.555 6
		4GHz	0.555 6
		256MB	−1.111 1
内存	17.24%	512MB	0.555 6
		1 024MB	0.555 6
Pearson's R = 1.000		Significance = 0.000	
Kendall's tau = 1.000		Significance = 0.000 2	

对联合分析进行拟合优度检验，这里使用皮尔逊（Pearson）的 R 检验和肯德尔（Kendall）的 tau 检验，所得结果如下所示：

```
Pearson's R =1.000          Significance=0.000
Kendall's tau = 1.000       Significance = 0.000 2
```

由此可见，两个相关系数的检验都非常显著，模型拟合的精度相当高，所以认为联合分析模型所做出的假设和得出的成分效用值是合理的，可以说明第一个顾客在选择电脑时的偏好结构。

下面是各属性水平的效用图（由 SPSS 软件直接输出的是各个属性水平的效用图，太占篇幅，故本例用 Excel 软件做成如下形式，见图 12—7）。

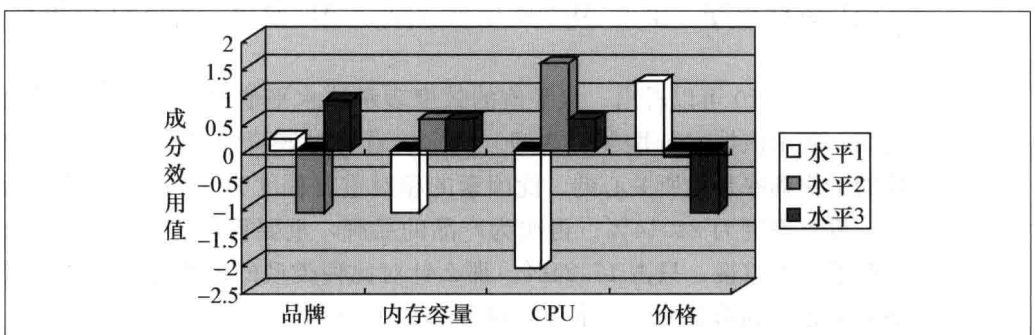


图 12—7 各属性水平的效用图

根据表 12—10 和图 12—7 中的各因子水平的成分效用值，可以发现第三个因子（CPU）和第四个因子（价格）的数值较大（绝对值），说明这两种属性在消费者心目中具有重要地位。

各属性相对重要性比较分析图如图 12—8 所示。

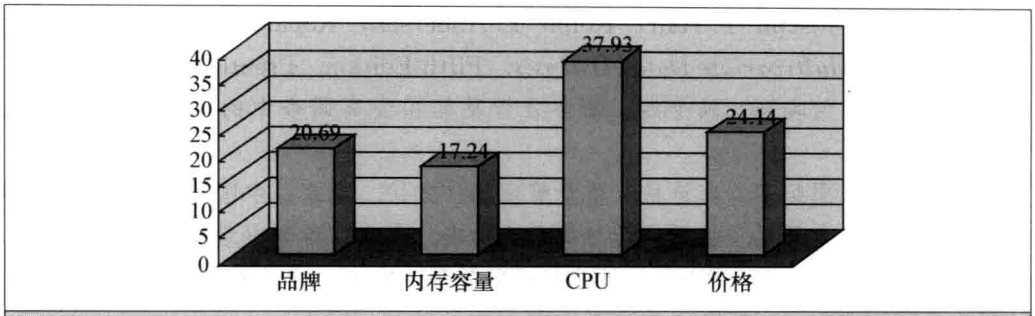


图 12—8　各属性相对重要性比较分析图

由图 12—8 可以看出各个因子的重要程度，消费者非常关注 CPU 的性能，其次是价格的高低，再次是品牌和内存容量。

(6) 市场预测与决策。联合分析的特殊功能在于它可以预测产品的前景，在得到特性水平值的效度后，可以对产品的各种技术特性进行配置组合。

假设该消费者想购买价格为 7 200 元的电脑，那么可以考虑下面三种配置：

(1) 惠普　　CPU 2 GHZ　　内存 1 024MB

(2) 联想　　CPU 4 GHZ　　内存 256MB

(3) 方正　　CPU 3 GHZ　　内存 512 MB

分别计算三种产品对消费者的效用，计算结果如下：

$$U_1 = U(\text{品牌}+\text{CPU}+\text{内存}+\text{价格})$$
$$= 0.888\,9+(-2.111\,1)+0.555\,6+(-0.111\,1)=-0.777\,7$$
$$U_2 = U(\text{品牌}+\text{CPU}+\text{内存}+\text{价格})$$
$$= -1.111\,1+0.555\,6+(-1.111\,1)+(-0.111\,1)=-1.777\,7$$
$$U_3 = U(\text{品牌}+\text{CPU}+\text{内存}+\text{价格})$$
$$= 0.222\,2+1.555\,6+0.555\,6+(-0.111\,1)=2.222\,3$$

由以上结果可知 $U_3 > U_1 > U_2$，因此在该消费者的心目中，产品 (3) 的效用值最大，即该产品应具有的属性水平为：方正品牌，3GHZ 的 CPU，内存为 512MB，价格为 7 200 元。

□ 参考文献

[1] R. Duncan Luce & John W. Tukey. Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement. *Journal of Mathematical Psychology*, 1964, Vol. 1, pp. 1-27

[2] P. E. Green & V. R. Rao. Conjoint Measurement for Quantifying Judgement Data. *Journal of Marketing Research*, 1971, Vol. 8, pp. 355-363

[3] P. Cattin & D. R. Wittink. Commercial Use of Conjoint Analysis: A Survey. *Journal of Marketing*, 1982, Vol. 46, pp. 44-53

[4] Joseph F. Hair, Rolph E. Anderson, Ronald L. Tatham, William C. Black. *Multivariate Data Analysis*. Fifth Edition. Prentice-Hall, 1998

[5] 何晓群，陈少杰. 联合分析及其在公寓调查中的应用. 统计研究，2001(12)

[6] 周纪芗，茆诗松. 质量管理统计方法. 北京：中国统计出版社，1999

[7] 方开泰. 均匀设计与均匀设计表. 北京：科学出版社，1994

[8] 张文彤. SPSS 统计分析高级教程. 北京：高等教育出版社，2004

□ 思考与练习

1. 简述联合分析的思想。
2. 联合分析的设计应注意哪些问题？
3. 简述联合分析在市场研究中的应用。
4. 试对某种商品进行调查，利用联合分析方法得出分析结果并予以解释。

第 13 章

Chapter 13 多变量的图表示法

学习目标

1. 理解各种多变量图表示法的作图思想；
2. 了解各种多变量图表示法的作图方法；
3. 能够利用软件对多元资料作图；
4. 能够利用所作的多变量图形对数据进行探索性分析。

图形是对资料进行探索性研究的重要工具，人们在运用其他统计方法对所得资料进行分析之前，往往习惯于把各资料在一张图上画出来，以直观地反映资料的分布情况及各变量之间的相关关系。当变量较少时，可以采用直方图、条形图、饼图、散点图或是经验分布的密度图等方法。对于变量个数少于 3 的情况，这样做是简单而有效的。而当变量个数为 3 时，虽然仍可以作三维的散点图，但这样做已经不是很方便。当变量个数大于 3 时，就不能用通常的方法作图了。20 世纪 70 年代以来，统计学家研究发明了很多多维变量的图表示方法，借助图形来描述多元资料的统计特性，使图形直观、简洁的优点延伸到多变量的研究中。本章主要介绍散点图矩阵、脸谱图、雷达图等多变量的图表示法的基本思想及作图方法。

因为对资料的图表示法只是以一种直观的方式再现资料，不同的研究者习惯的资料显示方式可能会有很大不同，因此，不同于其他统计方法，大部分图表示法都没有非常严格的画图方法，研究者可以根据自己的习惯设定某些规则，以更方便地揭示资料之间的联系。因此，本章对各种图表示方法原则上只给出作图的思想及思路，而不对严格的数学公式做过多说明。

13.1 散点图矩阵

散点图矩阵是借助两变量散点图的作图方法，它可以看作一个大的图形方阵，其每一个非主对角元素的位置上是对应行的变量与对应列的变量的散点图，而主对角元素位置上是各变量名，这样可以清晰地看到所研究多个变量两两之间的相关关系。由此也可以看出，散点图矩阵方法还不是真正意义上的多变量作图方法，它研究的仍是两两变量之间的相关关系，而不能直接反映多个变量之间的关系，借助它来对资料分类也是比较困难的。然而，因其直观、简单、容易理解，散点图矩阵还是越来越受广大实际工作者的喜爱，很多统计软件也加入了作散点图矩阵的功能。下面举例说明如何用 SPSS 软件作散点图矩阵对资料进行分析。

例 13—1

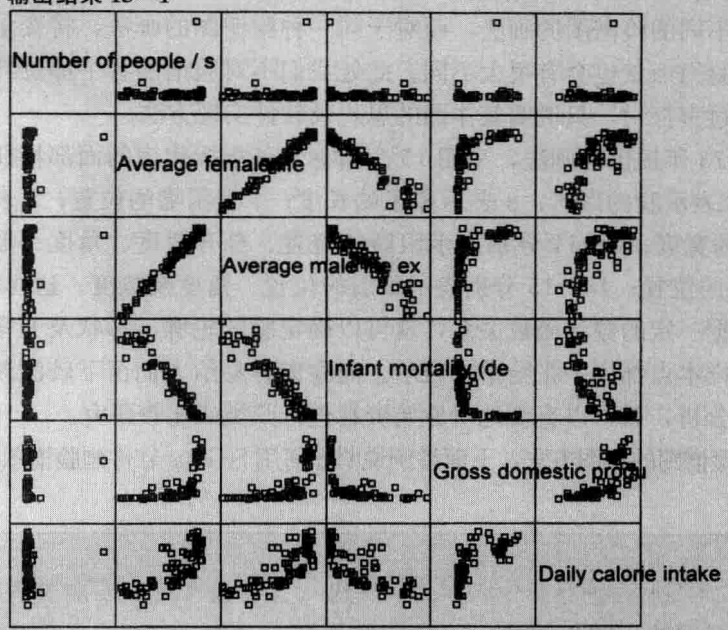
以 World95.sav 资料为例，该资料共有 26 个变量、109 条观测，是 1995 年世界 109 个国家和地区的基本发展情况的资料。选择该数据如下几个变量作图：density（每平方公里人口数），lifeexpf（女性预期寿命），lifeexpm（男性预期寿命），babymort（婴儿死亡率），gdp_cap（GDP 是总资产的倍数），calories（每日摄入热量）。

打开资料集 World95.sav，依次点选 Graphs→Scatter/Dot…进入 Scatterplot 对话框，选中 Matrix（矩阵）左侧的图标，点击 Define 按钮，进入 Scatterplot Matrix 对话框，依次选择上面 6 个变量，点击 OK 键运行，则生成如下图形，见输出结果 13—1。

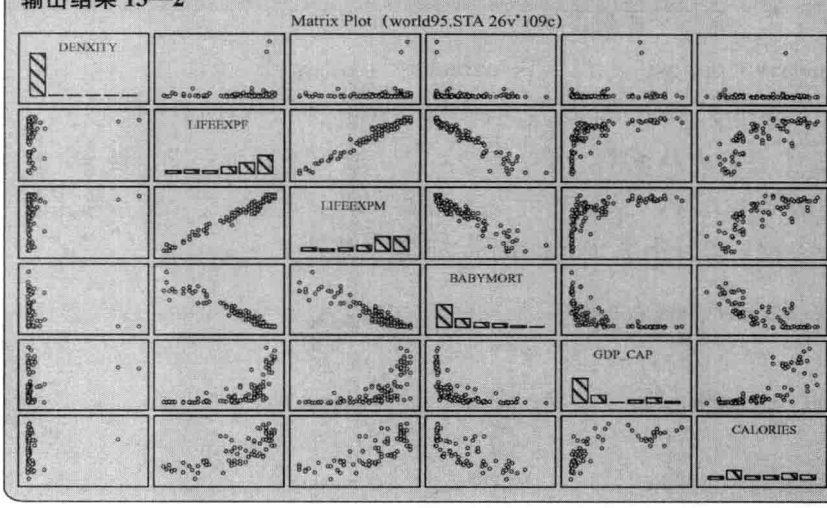
由散点图矩阵可以看到，每平方公里人口数与其他各变量的相关关系均不明显，男性的预期寿命、女性的预期寿命及婴儿死亡率三个变量之间有明显的线性相关关系，而 GDP 是总资产的倍数与上面三个变量存在着某种曲线相关关系。还可以看出其他变量之间的相关关系，在此不再赘述。另外，SPSS 软件还有一些选项可以帮助我们由散点图矩阵得到更多信息。资料集 World95.sav 中变量 religion 的含义是主要的宗教信仰，在 Scatterplot Matrix 对话框中将 religion 变量选作标记变量（选入 Set markers by），则在输出的散点图矩阵中，不同宗教信仰的国家以不同的颜色画出，借此可以做更详细的分析，此处不再详述。

另外，有些统计软件（如 Statistica 软件）提供的画散点图矩阵的模块可以同时在主对角线上画出相应变量的直方图，这样就在散点图矩阵上提供了更多的信息，即同时能通过该图了解各变量的分布情况。对于 Statistica 软件的用法，这里不做过多说明。对于上面的资料，用 Statistica 软件作散点图矩阵得到输出结果 13—2。

输出结果 13—1



输出结果 13—2



13.2　脸谱图

脸谱图是用脸谱来表达多变量的样品，由美国统计学家 H. 切尔诺夫（H. Chernoff）于 1970 年首先提出。该方法是将观测的 p 个变量（指标）分别用脸的某一部位的形状或大小来表示，一个样品（观测）可以画成一张脸谱。他首先将该方法用于聚类分析，引起了各国统计学家的极大兴趣，并对他的画法做出了改进。一些统计软件也收入了脸谱图分析法，国内也有很多研究者将该方法应用于多元统计分析中。

脸谱图分析法的基本思想是由 15～18 个指标决定脸部特征，若实际资料变量更多将被忽略（有新的画图方法取消了脸的对称性并引入更多脸部特征，从而最多

可以用 36 个变量来画脸谱），若实际资料变量较少则脸部有些特征将被自动固定。统计学家曾给出了几种不同的脸谱图的画法，而对于同一种脸谱图的画法，将变量次序重新排列，得到的脸谱形状也会有很大不同。此处我们不对脸谱的各个部位与原始变量的数学关系做过多探讨，只说明其作图的思想及软件实现方法。

按照切尔诺夫于 1973 年提出的画法，采用 15 个指标，各指标代表的面部特征为：1 表示脸的范围；2 表示脸的形状；3 表示鼻子的长度；4 表示嘴的位置；5 表示笑容曲线；6 表示嘴的宽度；7～11 分别表示眼睛的位置、分开程度、角度、形状和宽度；12 表示瞳孔的位置；13～15 分别表示眼眉的位置、角度及宽度。这样，按照各变量的取值，根据一定的数学函数关系，就可以确定脸的轮廓、形状及五官的部位、形状，每一个样本点都用一张脸谱来表示。而脸谱容易给人们留下较深刻的印象，通过对脸谱的分析，就可以直观地对原始资料进行归类或比较研究。

S-Plus 软件收入了脸谱图的作图方法，下面举例说明如何用 S-Plus 软件画脸谱图。

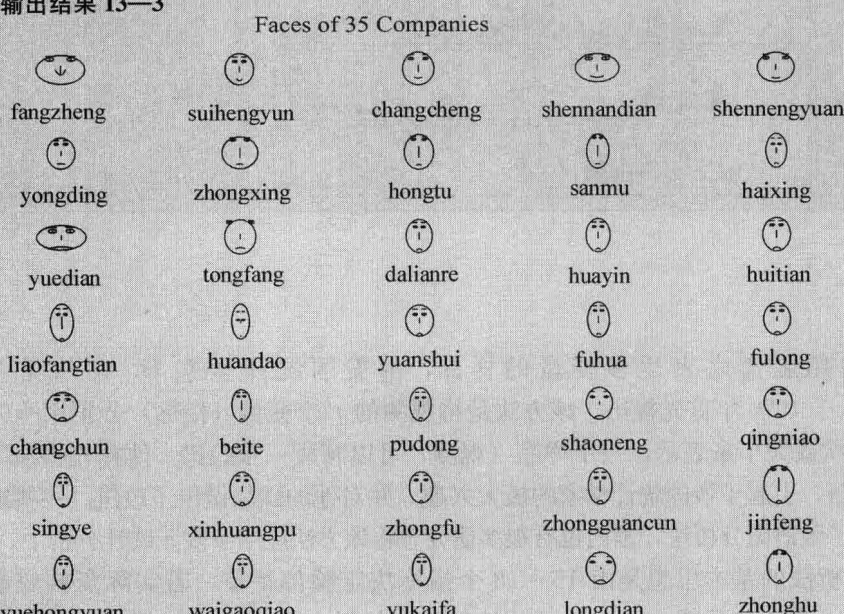
例 13—2

仍以我国 35 个上市公司的八大评价指标为例说明。S-Plus 画脸谱图的方法非常简单，只要调用 faces 函数就可以实现。将前面资料的数字部分输入 S-Plus，并将文件命名为 gongsi.sdd，在命令窗口调用下面的函数：

```
faces (data.matrix (gongsi), fill=T, which=1:8,
    head="Faces of 35 Companies", ncol=5, scale=T, byrow=T)
```

回车运行就可以生成 35 个公司的脸谱图，每一个公司用一张脸谱表示出来，但是，此时生成的脸谱图不好与公司名对应，可将 35 个公司名放入一个向量 ***a*** 中，然后在上面的命令中加入选项 labels=a，即可生成如下脸谱图，见输出结果 13—3。

输出结果 13—3



对 faces 函数的子选项作简要说明，因为完整的脸谱图共需 15 个变量，而此处只有 8 个变量，fill＝T 是指将由后 7 个变量决定的脸的部位画在相应的中央位置，which＝1:8 是指用资料集 gongsi 的前 8 列画脸谱图，head 指定图的标题，ncol 确定输出时每行输出脸谱图的个数，scale＝T 指在画脸谱图时将各变量都变换到（0，1）之间，byrow＝T 是指输出时脸谱图按行排列，这有助于我们将脸谱图与相应的公司名对应起来。

脸谱图给人的感觉形象直观，容易留下较深刻的印象，可以根据脸谱图来对各公司的运营能力进行比较。比如从脸的范围（净资产收益率）看来，方正科技、清华同方、粤电力、深南电、金丰投资等公司处于较高水平，而渝开发、粤宏远、寰岛实业等公司明显处于较低水平，类似可以对其他指标进行分析。利用脸谱图，还可以直观地对各个公司进行归类。由输出结果 13—3 来看，方正科技、深南电、深能源、中兴通讯、粤电力、清华同方、金丰等公司大致可以归为一类，穗恒运、长城计算机、永鼎光缆、宏图高科大致可以归为一类，富龙热力、韶能股份、惠天热电、大连热电、华银电力、长春经开、新黄浦、辽房天、三木集团、青鸟华光、海星科技、龙电股份等公司可以归为一类，剩余的公司大体可以归为一类。此处不再详述。

在利用脸谱图工具对观测进行比较分析时，值得注意的一点是脸谱的形状受各变量次序的影响很大。在本例中如果把 8 个指标的次序换一下，得到的脸谱图就会有很大不同。而且，根据脸谱图对各公司的归类有很强的主观性，因为不同的人所关注的脸的部位有很大不同，如有些人对脸的胖瘦比较在意，而有的人对五官的印象特别深，因此对同样的脸谱图，不同的人可能得到不同的结论。在实际分析中，该方法必须与聚类、相关等定量分析相结合，才能得到比较合理可信的结论。

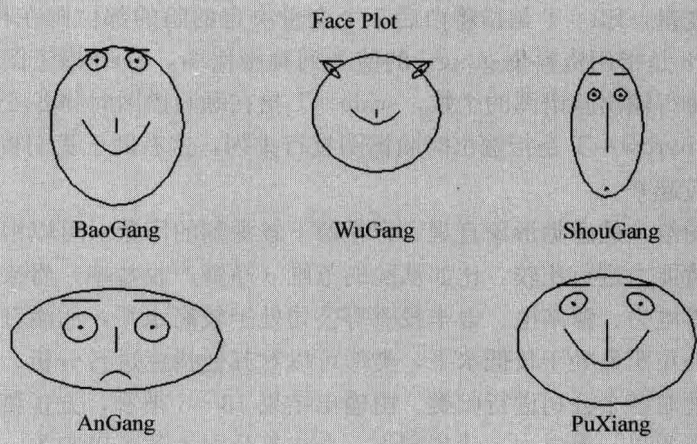
例 13—3

表 13—1 是反映五大钢铁公司经营状况的十大指标。为了比较国内钢铁公司与韩国浦项钢铁公司的差距，下面作出韩国浦项钢铁公司、宝钢、鞍钢、武钢、首钢五家钢铁公司的脸谱图（见输出结果 13—4）。

表 13—1

项目	宝钢	鞍钢	武钢	首钢	浦项
负债保障率	2.89	2.95	2.34	1.85	3.12
长期负债倍数	5.16	9.15	6.07	2.63	6.96
流动比率	1.31	1.83	1.16	2.22	2.1
资产利润率	21.71	17.34	24.77	11.89	25.34
收入利润率	23.17	11.33	19.55	7.6	22.28
成本费用利润率	30.23	12.76	24.81	8.05	28.52
净利润现金比率	1.79	0.9	1.7	1.09	1.3
三年资产平均增长率	1.48	7.28	63.3	11.76	13.18
三年销售平均增长率	20.07	29.19	52.88	18.77	24.16
三年平均资本增长率	11.04	10.5	48.95	7.63	17.51

输出结果 13—4



13.3 雷达图与星图

13.3.1 雷达图

雷达图是目前应用较为广泛的对多元资料进行作图的方法，利用雷达图可以很方便地研究各样本点之间的关系并对样品进行归类。设要分析的资料共有 p 个变量，雷达图的标准画法如下：先画一个圆，将圆 p 等分并由圆心连接各分点，将所得的 p 条线段作为坐标轴，根据各变量的取值对各坐标轴作适当刻度，这样，对每个观测的每个变量的取值，在相应坐标轴上都有一个刻度。对任一样本点，可以分别在 p 个轴上确定其坐标，在各坐标轴上点出其坐标并依次连接 p 个点，可以得到一个 p 边形，这样，每一个样本点用一个 p 边形表示出来，通过观察各个 p 边形的形状，就可以对各个样本点的相似性进行分析。当样本数目较小时，可以在一个圆中画出所有的样本点；当样本数目较大时，也可以每一个样本点画一个 p 边形进行分析。

Excel 软件提供了画雷达图的功能，它适用于观测数较少的情形，这时可以方便地把各观测画到一张图里，便于对各指标进行对比，但是，当观测数比较多时，画到一张雷达图里面就不太容易看出各观测之间的接近程度，用 Excel 当然也可以对每一个观测画一张雷达图，但此时转差率已经很低了。S-Plus 软件也收入了雷达图的画法，下面举例说明雷达图的画法。

例 13—4

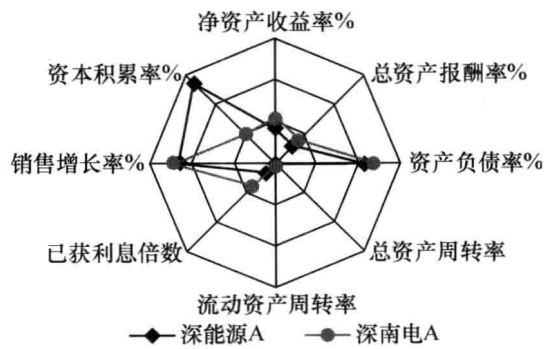
沿用我国 35 个上市公司的资料。常见的 Excel 就可画出很漂亮的雷达图。用 Excel 画雷达图，比如仅对深能源和深南电两公司画雷达图，方法如下。

在 Excel 窗口中，输入资料格式如下：

公司简称	净资产收益率%	总资产报酬率%	资产负债率%	总资产周转率	流动资产周转率	已获利息倍数	销售增长率%	资本积累率%
深能源 A	16.85	12.35	42.32	0.37	1.78	7.18	45.73	54.54
深南电 A	22	15.30	46.51	0.76	1.77	15.67	48.11	19.41

用鼠标选中该部分资料，依次点选插入→其他图表，进入图表向导对话框，在标准类型中选择雷达图，在子图表类型中选择第二项资料点雷达图，点下一步按钮，可以看到产生雷达图的示意图，系统产生默认是行，对本例资料，若不是行，则应改为行。点击下一步，进入图表选项对话框，在此可以对雷达图的有关设置进行重新设定，点击完成则生成如下雷达图，见输出结果 13—5。

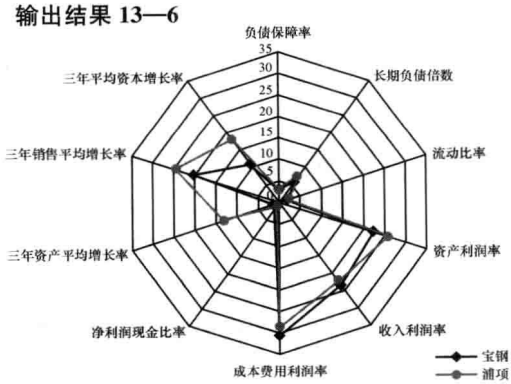
输出结果 13—5

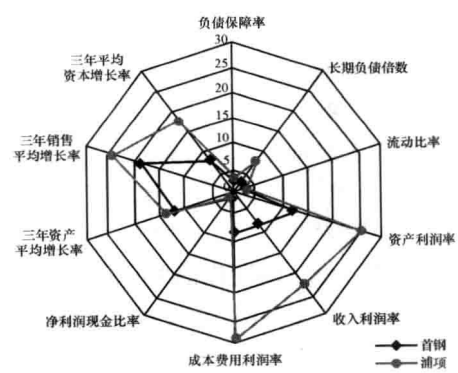


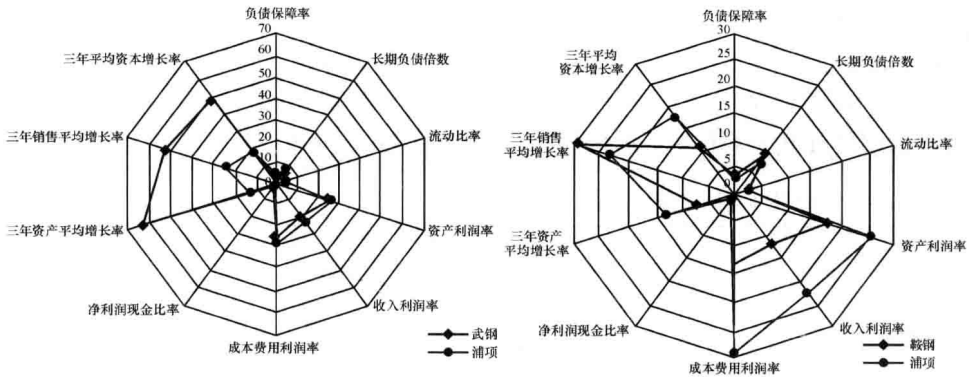
Excel 提供了很强的图形再编辑功能，对产生的雷达图可以进行各种编辑修改以使其界面显示更为友好，此处不再详细说明。根据此雷达图，可以对深能源与深南电的运营能力进行分析，深能源的资本积累率远高于深南电，深南电的已获利息倍数要高于深能源，两公司的其余指标大体相似。

根据反映五大钢铁公司经营状况的十大指标，我们可以作出韩国浦项钢铁公司与国内宝钢、鞍钢、武钢、首钢四家钢铁公司之间的雷达图，见输出结果 13—6。

输出结果 13—6







13.3.2 星图

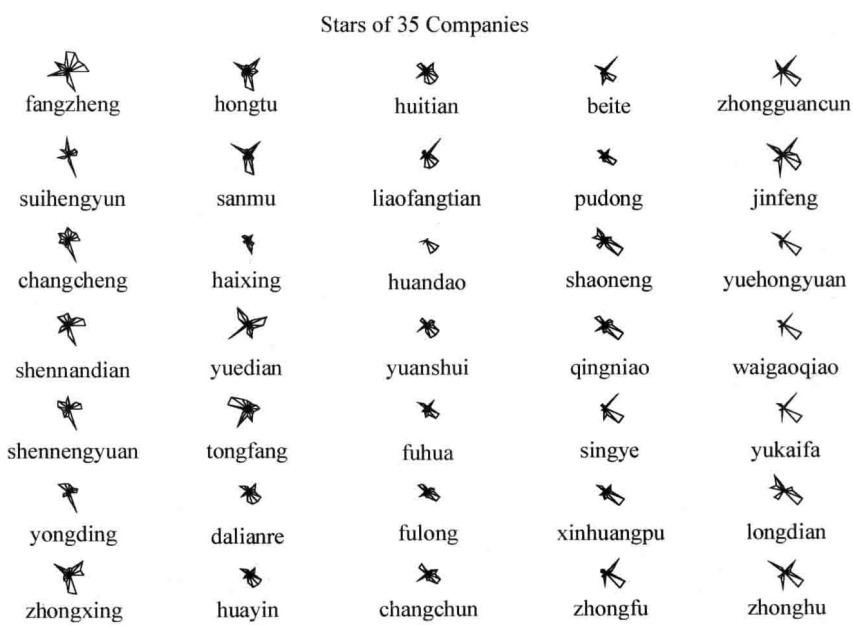
星图的形状与雷达图很相似，甚至有的文献把两者看成一回事。S-Plus 软件可以一次生成多个观测的星图，每一个观测生成一张星图。对上面的 35 个公司的资料，调用 stars 函数就可以方便地生成各个公司的星图。

如果数据文件已经建立，直接调用 stars 函数如下：

```
stars(data. matrix(gongsi),full=T,scale=T,radius=T,type="1",labels=a,
    head="Stars of 35 Companies",ncol=5)
```

则生成如下星图，见输出结果 13—7。

输出结果 13—7



对 stars 函数作简要说明。full=T，是指定每一个星图都包括一个整圆，若此项选为 F，则每一个星图仅包括上边半圆。scale=T，指将每一个指针都转换到范

围 [0，1]，即最大取值为 1，最小取值为 0，其他取值均转换为 0～1 之间的数。radius=T 指画出每一变量取值的半径，取 F 时将不画出。type="l"指的是对每一星图仅画出线（半径）而不画出各点，若要仅画出点或线与点都画出则应分别将 type 设为 "p" 和 "b"。labels 与 head 分别指定图的标题及各公司星图的标签。ncol 指定输出时每一行输出的星图个数。

S-Plus 所作星图各半径与原指标的对应关系为：从右边起，水平的半径为第一指针，逆时针旋转，星图的各半径分别对应第二、第三等各个指标，根据星图各条半径的长短，可以很容易地判断对应指标在各公司中的相对水平，以此来分析各公司的运营能力。同时，也可以利用星图来对各公司进行归类分析。与脸谱图相比，星图所受各指针排列次序的影响更小，受人的主观影响也较小。此处略去根据星图对各公司的比较研究。

13.4　星座图

所谓星座图，就是将所有样本点都点在一个半圆里面，就像天文学中表示星座的图像，根据样本点的位置可以直观地对各样本点之间的相关性进行分析。利用星座图可以方便地对样本点进行分类，在星座图上比较靠近的样本点比较相似，可以分为一类，相距较远的样本点的差异较大。

星座图的基本画图方法为：

（1）先将资料（X_{1i}，X_{2i}，…，X_{pi}）（$i=1, 2, \cdots, n$）进行变换，使其取值范围落到（0，π）之间，也就是构造函数 $f_j(X)$，使得

$$\begin{cases} B_{ji}=f_j(X_{ji}) \\ 0\leqslant f_j(X_{ji})\leqslant\pi \end{cases} \quad i=1,2,\cdots,n;j=1,2,\cdots,p \tag{13.1}$$

取 $f_j(X_{ji})=\dfrac{X_{ji}-\min\{X_{ji},\ i=1,\ 2,\ \cdots,\ n\}}{\max\{X_{ji},\ i=1,\ 2,\ \cdots,\ n\}-\min\{X_{ji},\ i=1,\ 2,\ \cdots,\ n\}}\pi$。

（2）对每一变量赋予一个权重 w_j，满足

$$\sum_{j=1}^{p} w_j = 1 \tag{13.2}$$

作图时，权数可以采用随机数的方法产生，也可以取 $w_j=\dfrac{1}{p}$。

（3）画一个半径为 1 的上半圆及底部的直径，以圆点 O 为圆心，w_1 为半径再画一个上半圆，将其弧度为 B_{11} 的地方记为 O_1，以 O_1 为圆心画上半圆，将其弧度为 B_{12} 的地方记为 O_2，依此类推，则 O_p 点即为第一个样本点的位置，同理可以画出所得资料所有的点。可以看出，第 k 组样品的星座 Z_k 为：

$$Z_k = \sum_{j=1}^{p} w_j e^{iB_{jk}} \tag{13.3}$$

则 Z_k 的路径为：

$$\sum_{j=1}^{p} w_j \cos B_{jk} \text{ 和} \sum_{j=1}^{p} w_j \sin B_{jk}, \quad k = 1,2,\cdots,n \tag{13.4}$$

（4）根据星座图上点的位置及路径判断各样本点之间的接近程度，进而可以对样本点进行归类分析。

在实际工作中，人们往往去掉各样本点的路径部分而仅保留其在星座图上的位置，并根据各点位置的接近程度分析样本点之间的接近程度。目前常用的统计软件均没有直接生成星座图的模块，但是，画星座图实际上非常简单。按照上面的方法，对数据进行规格化，对每一个变量赋予适当的权重，然后以式（13.4）各点的路径作为在星座图中的坐标画出各点的散点图，则画出的散点图实际上就是星座图。这里不再详细说明。上面 35 个上市公司的数据按此方法可以得到如下星座图（见图 13—1）。

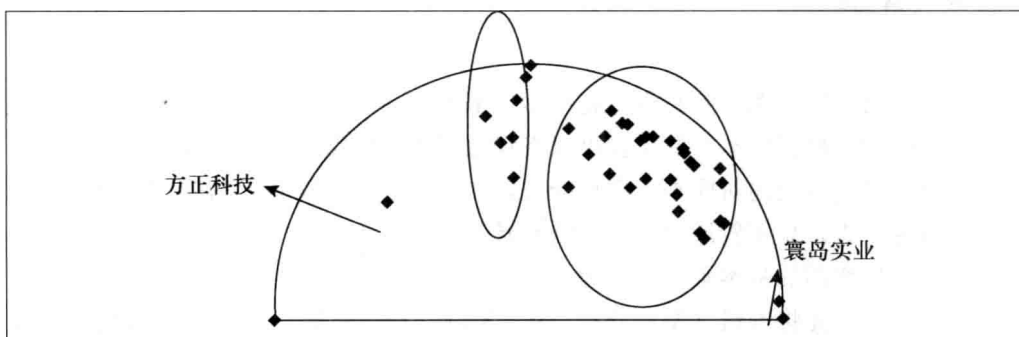


图 13—1 星座图

由图 13—1 中各公司在星座图中的接近程度，可以直观地对各公司进行分类。可以考虑把 35 个公司分成 4 类：方正科技、寰岛实业可单独成类，长城电脑、深能源 A、深南电 A、中兴通讯、清华同方、粤电力 A 可以归为一类，其他公司可以归为一类，此种归类与上面根据其他多变量图表示法得到的归类是有区别的。因此，实际工作中应用这些方法时，建议多种方法结合使用，以得到比较可信的结论。另外，此图还存在一个困难，就是不好将各个点与相应的公司对应起来，实际上，可以根据需要在生成星座图的同时画出各点的标签，此处出于图形清晰的考虑没有生成。此外，对于大部分软件来讲，当图形生成之后，只要将鼠标在相应的点上稍作停留，就会显示出该点对应的观测信息。

除本章介绍的几种方法外，多变量的图表示法还有塑像图、轮廓图、树形图等方法。这几种方法也是对每一个观测生成一张图，图形的不同部分表示观测不同指标的取值。有兴趣的读者可以参阅参考文献［1］。总体来说，多变量的图表示法使资料的呈现方式更直观、更形象，借助这些工具可以使研究者对资料有较深的印象，同时利用这些作图方法，可以帮助研究者对资料进行探索性分析，有助于进行更为专业的定量分析，形成合理结论。但是，多变量的图表示法只是给人一种大概

的印象，利用它来形成结论还是很不够，实践中必须结合其他统计分析方法并结合所分析的具体问题，综合定量分析与定性分析，才能得到较为合理可信的结论。

□ 参考文献

[1] 方开泰. 实用多元统计分析. 上海：华东师范大学出版社，1989

[2] 吴国富，安万福，刘景海. 实用数据分析方法. 北京：中国统计出版社，1992

[3] 王学仁，王松桂. 实用多元统计分析. 上海：上海科学技术出版社，1990

□ 思考与练习

1. 试述多变量图示法的思想方法和实际意义。
2. 试对某一多变量实际问题分别画散点图矩阵、脸谱图、雷达图、星座图等。

第 14 章
Chapter 14 多维标度法

学习目标

1. 理解多维标度法的模型；
2. 了解求解的古典法和非度量法；
3. 能够解释维数在空间图中的表现；
4. 掌握实现多维标度法的软件模块。

多维标度法（multidimensional scaling，MDS）是通过一系列技巧，使研究者识别构成受测者对样品进行评价基础的关键维数。比如，多维标度法常用于市场研究中，以识别构成顾客对产品、服务或者公司的评价基础的关键维数。其他的应用包括比较自然属性（比如食品口味或者不同的气味），对政治候选人或事件的了解，甚至评估不同群体的文化差异。多维标度法通过受测者所提供的对样品的相似性或者偏好的判断推导出内在的维数。一旦有数据，多维标度法可以确定：（1）评价样品时受测者使用什么维数；（2）在特定情况下可能使用多少维数；（3）每个维数的相对重要性；（4）如何获得样品关联的感性认识。

本章主要根据参考文献［1］来介绍多维标度法的思想原理及其应用。

14.1 多维标度法的基本理论和方法

多维标度法是在一个确定维数的空间中估计一组样品的坐标，它的数据是配对样品间的距离。有很多模型可以用来计算距离和使距离与实际数据相关联。多维标度模型有两因子和三因子的度量和非度量的多维标度模型。多维标度过程的数据包括样品间的一个或者多个对称或者不对称方阵。这样的数据也称为相似性数据。在

心理分析中，每个矩阵对应于一个主体，对每个主体拟合不同参数的模型称为个体差异模型。

为了说明多维标度法，先看一个经典的例子，参见参考文献 [1]。

例 14—1

表 14—1 列出了英国 12 个城市间公路的距离，由于公路弯弯曲曲，这些距离并不是城市间真正的距离。我们希望在地图上重新标出这 12 个城市，使它们之间的距离很接近表 14—1 中的距离。

表 14—1　　**英国 12 个城市间的公路距离**

	1	2	3	4	5	6	7	8	9	10	11	12
1												
2	244											
3	218	350										
4	284	77	369									
5	197	167	347	242								
6	312	444	94	463	441							
7	215	221	150	236	279	245						
8	469	583	251	598	598	169	380					
9	166	242	116	257	269	210	55	349				
10	212	53	298	72	170	392	168	531	190			
11	253	325	57	340	359	143	117	264	91	273		
12	270	168	284	164	277	378	143	514	173	111	256	

1=Aberystwyth，2=Brighton，3=Carlisle，4=Dover，5=Exeter，6=Glasgow，7=Hull，8=Inverness，9=Leeds，10=London，11=Newcastle，12=Norwich.

如果用 $\boldsymbol{D}=(d_{ij})$ 表示表 14—1 的矩阵，它名义上是距离阵，但并不一定是 n 个点的距离，即不是通常所理解的距离阵。于是首先需要将距离阵的概念加以拓展。

定义 14.1　一个 $n\times n$ 矩阵 $\boldsymbol{D}=(d_{ij})$，若满足 $\boldsymbol{D}'=\boldsymbol{D}$，$d_{ii}=0$，$d_{ij}\geqslant 0$（$i$，$j=1$，…，$n$；$i\neq j$），则称 $\boldsymbol{D}$ 为距离阵。

这样定义的距离并不一定满足通常距离的三角不等式。

对于距离阵 $\boldsymbol{D}=(d_{ij})$，多维标度法的目的是要寻找 k 和 R^k 中的 n 个点 $\boldsymbol{x}_1$，…，$\boldsymbol{x}_n$，用 $\hat{d}_{ij}$ 表示 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 的欧氏距离，$\hat{\boldsymbol{D}}=(\hat{d}_{ij})$，使得 $\hat{\boldsymbol{D}}$ 与 $\boldsymbol{D}$ 在某种意义下相近。在实际中，常取 $k=1$，2 或 3。

令

$$\boldsymbol{X}=(\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \cdots,\ \boldsymbol{x}_n) \tag{14.1}$$

为了叙述简单，我们称 $\boldsymbol{X}$ 为 $\boldsymbol{D}$ 的拟合构造点。当 $\hat{\boldsymbol{D}}=\boldsymbol{D}$ 时，$\boldsymbol{X}$ 称为 $\boldsymbol{D}$ 的构造点。

需要指出的是，多维标度法的解并不唯一。若 $\boldsymbol{X}$ 是解，令

$$\boldsymbol{y}_i=\boldsymbol{\Gamma}\boldsymbol{x}_i+\boldsymbol{a}$$

式中，$\boldsymbol{\Gamma}$ 为正交阵；$\boldsymbol{a}$ 为常数向量，则 $\boldsymbol{Y}=(\boldsymbol{y}_1,\ \boldsymbol{y}_2,\ \cdots,\ \boldsymbol{y}_n)$ 也是解，因为平移和正交变换不改变欧氏距离。

下面利用主成分分析的思想给出求古典解的方法，并讨论古典解的优良性。本章还给出非度量法的一些描述。

14.2 多维标度法的古典解

14.2.1 欧氏型距离阵

定义 14.2 一个距离阵 $\boldsymbol{D}=(d_{ij})$ 称作欧氏型的，若存在某个正整数 p 及 p 维空间的 n 个点 $\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \cdots,\ \boldsymbol{x}_n$，使得

$$d_{ij}^2=(\boldsymbol{x}_i-\boldsymbol{x}_j)'(\boldsymbol{x}_i-\boldsymbol{x}_j),\quad i,j=1,\cdots,n \tag{14.2}$$

如何判断一个距离是不是欧氏型的？如何求得欧氏型距离阵对应的 n 个点？这是下面首先要解决的问题。

令

$$\boldsymbol{A}=(a_{ij}),\ a_{ij}=-\frac{1}{2}d_{ij}^2 \tag{14.3}$$

$$\boldsymbol{B}=\boldsymbol{HAH},\ \boldsymbol{H}=\boldsymbol{I}_n-\frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}'_n \tag{14.4}$$

借助这些定义，下面的定理给出判断 $\boldsymbol{D}$ 是否为欧氏型的充分必要条件。

定理 14.1 设 $\boldsymbol{D}$ 是 $n\times n$ 距离阵，$\boldsymbol{B}$ 由式（14.4）所定义，则 $\boldsymbol{D}$ 是欧氏型的，当且仅当 $\boldsymbol{B}\geqslant\boldsymbol{0}$。

证明：设 $\boldsymbol{D}$ 是欧氏型的，则存在 $\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \cdots,\ \boldsymbol{x}_n\in R^p$，使得

$$d_{ij}^2=-2a_{ij}=(\boldsymbol{x}_i-\boldsymbol{x}_j)'(\boldsymbol{x}_i-\boldsymbol{x}_j) \tag{14.5}$$

由式（14.4）可得

$$\boldsymbol{B}=\boldsymbol{HAH}=\boldsymbol{A}-\frac{1}{n}\boldsymbol{AJ}-\frac{1}{n}\boldsymbol{JA}+\frac{1}{n^2}\boldsymbol{JAJ} \tag{14.6}$$

式中，$\boldsymbol{J}=\boldsymbol{1}_n\boldsymbol{1}'_n$。注意

$$\frac{1}{n}\boldsymbol{AJ}=\begin{pmatrix}\bar{a}_{1\cdot}\\ \vdots\\ \bar{a}_{n\cdot}\end{pmatrix}\boldsymbol{1}'_n,\ \frac{1}{n}\boldsymbol{JA}=\boldsymbol{1}_n(\bar{a}_{\cdot 1},\cdots,\bar{a}_{\cdot n}),\ \frac{1}{n^2}\boldsymbol{JAJ}=\bar{a}_{\cdot\cdot}\boldsymbol{1}_n\boldsymbol{1}'_n$$

其中

$$\bar{a}_{i\cdot}=\frac{1}{n}\sum_{j=1}^{n}a_{ij},\quad \bar{a}_{\cdot j}=\frac{1}{n}\sum_{i=1}^{n}a_{ij},\quad \bar{a}_{\cdot\cdot}=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij} \tag{14.7}$$

将它们代入式（14.6）中，得到

$$b_{ij}=a_{ij}-\bar{a}_{i\cdot}-\bar{a}_{\cdot j}+\bar{a}_{\cdot\cdot} \tag{14.8}$$

再由式（14.5）可求得 a_{ij}，$\bar{a}_{i\cdot}$，$\bar{a}_{\cdot j}$，$\bar{a}_{\cdot\cdot}$，将它们代入上式，得

$$b_{ij}=(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'(\boldsymbol{x}_j-\bar{\boldsymbol{x}}) \tag{14.9}$$

式中

$$\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$$

式（14.9）用矩阵表达为：

$$\boldsymbol{B}=(\boldsymbol{HX})(\boldsymbol{HX})'\geqslant\boldsymbol{0} \tag{14.10}$$

因为 $\boldsymbol{HX}$ 正是将 $\boldsymbol{X}$ 的数据中心化，即

$$\boldsymbol{HX}=(\boldsymbol{X}_1-\bar{\boldsymbol{X}},\cdots,\boldsymbol{X}_n-\bar{\boldsymbol{X}})' \tag{14.11}$$

反之，若设 $\boldsymbol{B}\geqslant\boldsymbol{0}$，记 $p=\mathrm{rank}(\boldsymbol{B})$，$\lambda_1,\lambda_2,\cdots,\lambda_p(\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p)$ 为 $\boldsymbol{B}$ 的正特征根，$\boldsymbol{x}_{(1)}$，$\boldsymbol{x}_{(2)}$，$\cdots$，$\boldsymbol{x}_{(p)}$ 为相应的特征向量，且

$$\boldsymbol{x}'_{(i)}\boldsymbol{x}_{(j)}=\delta_{ij}\lambda_i \tag{14.12}$$

这里

$$\delta_{ij}=\begin{cases}1, & i=j\\ 0, & i\neq j\end{cases}$$

令 $\boldsymbol{X}=(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)},\cdots,\boldsymbol{x}_{(p)})$，它是一个 $n\times p$ 阵，它的行用 $\boldsymbol{x}_1$，$\boldsymbol{x}_2$，$\cdots$，$\boldsymbol{x}_n$ 表示，我们欲指出，$\boldsymbol{X}$ 正好是 $\boldsymbol{D}$ 的构造点，从而证明定理的充分性。

令 $\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p)$，$\boldsymbol{\Gamma}=\boldsymbol{X}\boldsymbol{\Lambda}^{-\frac{1}{2}}$，故 $\boldsymbol{\Gamma}$ 的列为 $\boldsymbol{B}$ 的标准正交化的特征向量，于是

$$\boldsymbol{B}=\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}'=\boldsymbol{X}\boldsymbol{X}' \tag{14.13}$$

即　　$b_{ij}=\boldsymbol{x}'_i\boldsymbol{x}_j$

由此

$$\begin{aligned}(\boldsymbol{x}_i-\boldsymbol{x}_j)'(\boldsymbol{x}_i-\boldsymbol{x}_j)&=\boldsymbol{x}'_i\boldsymbol{x}_i-2\boldsymbol{x}'_i\boldsymbol{x}'_j+\boldsymbol{x}'_j\boldsymbol{x}_j=b_{ii}-2b_{ij}+b_{jj}\\&=a_{ii}-2a_{ij}+a_{jj}&&(\text{由式(14.8)})\\&=-2a_{ij}&&(\text{由 } a_{ii}=a_{jj}=0)\\&=d_{ij}^2\end{aligned}$$

这表明 $\boldsymbol{X}$ 正好是 $\boldsymbol{D}$ 的构造点，$\boldsymbol{D}$ 是欧氏型的。

在上面的证明过程中，下面列举的一些事实是颇为重要的。

（1）若 $\boldsymbol{D}$ 是欧氏型的，则相应的 $\boldsymbol{B}$ 有式（14.9），这表明 b_{ij} 是 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 中心化后的内积，下面简称 $\boldsymbol{B}$ 是 $\boldsymbol{X}$ 中心化的内积阵。

（2）定理的充分性证明给出了从 $\boldsymbol{D}$ 构造 $\boldsymbol{X}$ 的办法，即

$$\boldsymbol{D}\rightarrow\boldsymbol{A}\rightarrow\boldsymbol{B}\rightarrow\boldsymbol{X} \tag{14.14}$$

此外，我们还进一步指出：

（3）充分性中所定义的 $\boldsymbol{x}_1$，$\boldsymbol{x}_2$，…，$\boldsymbol{x}_n$ 的均值为 $\boldsymbol{0}$。

由 $\boldsymbol{H}$ 的定义 $\boldsymbol{H}\boldsymbol{1}_n=\boldsymbol{0}$，从而

$$\boldsymbol{0}=\boldsymbol{1}'\boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{1}=\boldsymbol{1}'\boldsymbol{B}\boldsymbol{1}=\boldsymbol{1}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{1}$$

必有 $\boldsymbol{X}'\boldsymbol{1}_n=\boldsymbol{0}$，即 $\bar{\boldsymbol{x}}=\boldsymbol{0}$。

（4）0 是 $\boldsymbol{B}$ 的特征根，相应的特征向量是 $\boldsymbol{1}_n$。这是因为

$$\boldsymbol{B}\boldsymbol{1}_n=\boldsymbol{H}\boldsymbol{A}\boldsymbol{H}\boldsymbol{1}_n=\boldsymbol{0}=0\boldsymbol{1}_n$$

14.2.2 多维标度法的古典解

当 $\boldsymbol{D}$ 是欧氏型的时，定理 14.1 已给出了构造点 $\boldsymbol{X}$ 的办法；当 $\boldsymbol{D}$ 不是欧氏型的时，不存在 $\boldsymbol{D}$ 的构造点，只能寻求 $\boldsymbol{D}$ 的拟合构造点，记作 $\hat{\boldsymbol{X}}$，以区分真正的构造点 $\boldsymbol{X}$。在实际中，即使 $\boldsymbol{D}$ 是欧氏型的，它的构造点也是 $n\times p$ 阵。当 p 较大时失去了实用的价值，这时宁可不用 $\boldsymbol{X}$，而去寻求低维的拟合构造点 $\hat{\boldsymbol{X}}$。

在定理 14.1 中，由 $\boldsymbol{D}$ 获得 $\boldsymbol{X}$ 的途径式（14.14）给我们一个启示，可仿造这个途径来给出（非欧氏型）距离阵的拟合构造点，基于这种思想得到的拟合构造点称为多维标度法的古典解。

求古典解的步骤如下：

（1）由距离阵 $\boldsymbol{D}=(d_{ij})$ 构造 $\boldsymbol{A}=(a_{ij})=(-\frac{1}{2}d_{ij}^2)$。

（2）令 $\boldsymbol{B}=(b_{ij})$，使

$$b_{ij}=a_{ij}-\bar{a}_{i\cdot}-\bar{a}_{\cdot j}+\bar{a}_{\cdot\cdot}$$

（3）求 $\boldsymbol{B}$ 的特征根 λ_1，λ_2，…，λ_n（$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_n$），若无负特征根，表明 $\boldsymbol{B}\geqslant\boldsymbol{0}$，从而 $\boldsymbol{D}$ 是欧氏型的；若有负特征根，$\boldsymbol{D}$ 一定不是欧氏型的。令

$$a_{1,k}=\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{n}|\lambda_i|} \tag{14.15}$$

$$a_{2,k}=\frac{\sum_{i=1}^{k}\lambda_i^2}{\sum_{i=1}^{n}\lambda_i^2} \tag{14.16}$$

这两个量相当于主成分分析中的累积贡献率，当然我们希望 k 不要取太大，而使 $a_{1,k}$ 和 $a_{2,k}$ 比较大。当 k 取定后，用 $\hat{\boldsymbol{x}}_{(1)}$，$\hat{\boldsymbol{x}}_{(2)}$，…，$\hat{\boldsymbol{x}}_{(k)}$ 表示 $\boldsymbol{B}$ 对应于 λ_1，λ_2，…，λ_k 的正交化的特征向量，使得

$$\hat{\boldsymbol{x}}_{(i)}'\hat{\boldsymbol{x}}_{(i)}=\lambda_i,\quad i=1,2,\cdots,k$$

通常还要求 $\lambda_k>0$。若 $\lambda_k<0$，要缩小 k 的值。

(4)令 $\hat{\boldsymbol{X}}=(\hat{\boldsymbol{x}}_{(1)},\hat{\boldsymbol{x}}_{(2)},\cdots,\hat{\boldsymbol{x}}_{(k)})$，则 $\hat{\boldsymbol{X}}$ 的行向量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n$ 即为欲求的古典解。

为了说明上述求解的步骤，下面看两个例子。

例 14—2

设有距离阵如下：

$$\boldsymbol{D}=\begin{bmatrix}
0 & 1 & \sqrt{3} & 2 & \sqrt{3} & 1 & 1\\
 & 0 & 1 & \sqrt{3} & 2 & \sqrt{3} & 1\\
 & & 0 & 1 & \sqrt{3} & 2 & 1\\
 & & & 0 & 1 & \sqrt{3} & 1\\
 & & & & 0 & 1 & 1\\
 & & & & & 0 & 1\\
 & & & & & & 0
\end{bmatrix}$$

由 $a_{ij}=-\frac{1}{2}d_{ij}^2$，求得 $\boldsymbol{A}$，$\bar{a}_{i\cdot}$，$\bar{a}_{\cdot j}$，$\bar{a}_{\cdot\cdot}$ 如下：

$$\left[\begin{array}{ccccccc|c}
0 & -\frac{1}{2} & -\frac{3}{2} & -2 & -\frac{3}{2} & -\frac{1}{2} & -\frac{1}{2} & -\frac{13}{14}\\
 & 0 & -\frac{1}{2} & -\frac{3}{2} & -2 & -\frac{3}{2} & -\frac{1}{2} & -\frac{13}{14}\\
 & & 0 & -\frac{1}{2} & -\frac{3}{2} & -2 & -\frac{1}{2} & -\frac{13}{14}\\
 & & & 0 & -\frac{1}{2} & -\frac{3}{2} & -\frac{1}{2} & -\frac{13}{14}\\
 & & & & 0 & -\frac{1}{2} & -\frac{1}{2} & -\frac{13}{14}\\
 & & & & & 0 & -\frac{1}{2} & -\frac{13}{14}\\
 & & & & & & 0 & -\frac{3}{7}\\
\hline
-\frac{13}{14} & -\frac{13}{14} & -\frac{13}{14} & -\frac{13}{14} & -\frac{13}{14} & -\frac{13}{14} & -\frac{3}{7} & -\frac{6}{7}
\end{array}\right]$$

再由式 (14.8) 得到

$$\boldsymbol{B}=\frac{1}{2}\begin{bmatrix}
2 & 1 & -1 & -2 & -1 & 1 & 0\\
 & 2 & 1 & -1 & -2 & -1 & 0\\
 & & 2 & 1 & -1 & -2 & 0\\
 & & & 2 & 1 & -1 & 0\\
 & & & & 2 & 1 & 0\\
 & & & & & 2 & 0\\
 & & & & & & 0
\end{bmatrix}$$

由于 $\boldsymbol{B}$ 的列有如下的线性关系：

$$\boldsymbol{b}_{(3)}=\boldsymbol{b}_{(2)}-\boldsymbol{b}_{(1)},\ \boldsymbol{b}_{(4)}=-\boldsymbol{b}_{(1)},\ \boldsymbol{b}_{(5)}=-\boldsymbol{b}_{(2)},\ \boldsymbol{b}_{(6)}=\boldsymbol{b}_{(1)}-\boldsymbol{b}_{(2)},\ \boldsymbol{b}_{(7)}=\boldsymbol{0}$$

故 $\boldsymbol{B}$ 的秩最多为 2，再由 $\boldsymbol{B}$ 的第一个二阶主子式非退化，故 $\mathrm{rank}(\boldsymbol{B})=2$，并求得

$$\lambda_1=\lambda_2=3,\quad \lambda_3=\cdots=\lambda_7=0$$

特征向量 $\boldsymbol{x}_{(1)}$ 和 $\boldsymbol{x}_{(2)}$ 可取对应于 $\lambda=3$ 的子空间中任一对正交化的向量，比如取

$$\boldsymbol{x}_{(1)}=(a,a,0,-a,-a,0,0)',\qquad a=\frac{\sqrt{3}}{2}$$

$$\boldsymbol{x}_{(2)}=(b,-b,-2b,-b,b,2b,0)',\qquad b=\frac{1}{2}$$

于是 7 个点的坐标分别为：

$$\left(\frac{\sqrt{3}}{2},\frac{1}{2}\right),\left(\frac{\sqrt{3}}{2},-\frac{1}{2}\right),(0,-1),\left(-\frac{\sqrt{3}}{2},-\frac{1}{2}\right),\left(-\frac{\sqrt{3}}{2},\frac{1}{2}\right),(0,1),(0,0)$$

因为 $\boldsymbol{B}\geqslant\boldsymbol{0}$，所以原距离阵 $\boldsymbol{D}$ 是欧氏型的，故这个古典解是 $\boldsymbol{D}$ 的古典解。

例 14—3

考虑例 14—1 中英国 12 个城市的距离阵，相应 $\boldsymbol{B}$ 的特征根如下：

$$\lambda_1=394\,473,\ \lambda_2=63\,634,\ \lambda_3=13\,544,\ \lambda_4=10\,245,\ \lambda_5=2\,465,\ \lambda_6=1\,450$$
$$\lambda_7=501,\ \lambda_8=0,\ \lambda_9=-17,\ \lambda_{10}=-214,\ \lambda_{11}=-1\,141,\ \lambda_{12}=-7\,063$$

最后 4 个特征根是负的，表明 $\boldsymbol{D}$ 不是欧氏型的。当 $k=2$ 时，

$$a_{1,2}=92.6\%,\qquad a_{2,2}=99.8\%$$

故取 $k=2$ 就可以了。前两个主成分相应的特征向量（满足式 (14.12)）为：

$$\boldsymbol{x}_{(1)}=(45,203,-138,212,189,-234,-8,-382,-32,153,-120,112)'$$
$$\boldsymbol{x}_{(2)}=(140,-18,31,-76,140,31,-50,-26,-5,-27,-34,-106)'$$

于是可将 $\boldsymbol{x}_{(1)}$，$\boldsymbol{x}_{(2)}$ 相应的 12 个坐标点画在平面图上，就可看到由古典解确定的 12 个城市的位置。本例图此处从略，有兴趣者可参见参考文献 [1]。

14.2.3 相似系数矩阵

在有些问题中，已知的是 n 个样品之间的相似系数矩阵 $\boldsymbol{C}$，而不是距离阵 $\boldsymbol{D}$。前面在聚类分析中曾讨论过相似系数的概念。本节称 $\boldsymbol{C}=(c_{ij})$ 为相似系数矩阵，若 $\boldsymbol{C}$ 满足 $\boldsymbol{C}'=\boldsymbol{C}$，且

$$c_{ij} \leqslant c_{ii}, \quad \forall i, j \tag{14.17}$$

这样定义的相似系数并不一定满足 $c_{ii}=1$，由于相似系数和距离之间有一定的联系，我们可从相似阵 $\boldsymbol{C}$ 来产生一个距离阵 $\boldsymbol{D}=(d_{ij})$，其中

$$d_{ij}=(c_{ii}+c_{jj}-2c_{ij})^{\frac{1}{2}} \tag{14.18}$$

由式（14.17）有 $c_{ii}+c_{jj}-2c_{ij}\geqslant 0$，故 d_{ij} 的定义有意义，显见 $d_{ii}=0$ 及 $d_{ij}=d_{ji}$，故 $\boldsymbol{D}$ 为距离阵。

当相似系数矩阵 $\boldsymbol{C}$ 为非负定阵时，有如下定理。

定理 14.2　若 $\boldsymbol{C}\geqslant\boldsymbol{0}$，则由式（14.18）定义的距离阵为欧氏型的。

证明：设 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 分别为式（14.3）和式（14.4）所定义，那么 $a_{ij}=-\frac{1}{2}d_{ij}^2$，以及

$$\begin{aligned}
-2b_{ij} &= d_{ij}^2-\frac{1}{n}\sum_{j=1}^{n}d_{ij}^2-\frac{1}{n}\sum_{i=1}^{n}d_{ij}^2-\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}d_{ij}^2\\
&= c_{ii}+c_{jj}-2c_{ij}-\frac{1}{n}\sum_{j=1}^{n}(c_{ii}+c_{jj}-2c_{ij})-\frac{1}{n}\sum_{i=1}^{n}(c_{ii}+c_{jj}-2c_{ij})\\
&\quad+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(c_{ii}+c_{jj}-2c_{ij})\\
&= -2c_{ij}+2\bar{c}_{i\cdot}+2\bar{c}_{\cdot j}-2\bar{c}_{\cdot\cdot}
\end{aligned}$$

式中，$\bar{c}_{i\cdot}$，$\bar{c}_{\cdot j}$，$\bar{c}_{\cdot\cdot}$ 与 $\bar{d}_{i\cdot}$，$\bar{d}_{\cdot j}$，$\bar{d}_{\cdot\cdot}$ 有类似的定义。因此

$$b_{ij}=c_{ij}-\bar{c}_{i\cdot}-\bar{c}_{\cdot j}+\bar{c}_{\cdot\cdot} \tag{14.19}$$

回顾式（14.8）的证明，有 $\boldsymbol{B}=\boldsymbol{HAH}\geqslant\boldsymbol{0}$，由定理 14.1 得 $\boldsymbol{D}$ 是欧氏型的。

在定理的证明中，只有最后才用到 $\boldsymbol{C}\geqslant\boldsymbol{0}$ 的假设。若 $\boldsymbol{C}$ 为相似系数矩阵，$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的定义如前，此时总有

$$\boldsymbol{B}=\boldsymbol{HAH} \tag{14.20}$$

这是一个很重要的事实，在求古典解时是需要的。

例 14—4

表 14—2 是一个相似阵 $\boldsymbol{C}$，求它的古典解。

由式（14.19）可方便地求得 $\boldsymbol{B}$ 以及 $\boldsymbol{B}$ 的特征根。

表 14—2　　**相似系数矩阵 $\boldsymbol{C}$**

	1	2	3	4	5	6	7	8	9	10
1	84									
2	62	89								
3	16	59	86							
4	6	23	33	89						
5	12	8	27	56	90					
6	12	14	33	34	30	86				
7	20	25	17	24	18	65	85			
8	37	25	16	13	10	22	65	88		
9	57	28	9	7	5	8	31	58	91	
10	52	18	9	7	5	18	15	39	79	94

$\boldsymbol{B}$ 的特征根为：

$$\lambda_1=187.4,\ \lambda_2=121.0,\ \lambda_3=95.4,\ \lambda_4=55.4,\ \lambda_5=46.6$$
$$\lambda_6=31.5,\ \lambda_7=9.6,\ \lambda_8=4.5,\ \lambda_9=0,\ \lambda_{10}=-4.1$$

由于 $\lambda_{10}=-4.1<0$，说明 $\boldsymbol{B}$ 不是非负定阵，从而 $\boldsymbol{D}$ 不是欧氏型的。若取 $k=2$，前两个特征向量（满足式（14.12））为：

$$\boldsymbol{x}_{(1)}=(-4.2,-0.3,3.7,5.6,5.4,3.8,0.9,-3.0,-6.2,-5.7)'$$
$$\boldsymbol{x}_{(2)}=(-3.2,-5.8,-4.3,-0.6,0,4.0,5.5,3.6,0.6,0.2)'$$

14.3 古典解的优良性

前面是从 n 阶距离阵 $\boldsymbol{D}$ 出发，求它的构造点或拟合构造点，本节试图从一些侧面来考察拟合构造点的古典解的优良性，为此首先需要给出主坐标的定义。

设 $\boldsymbol{X}$ 为 $n\times p$ 数据阵，令 $\boldsymbol{A}=\boldsymbol{X}'\boldsymbol{H}\boldsymbol{X}$，$\boldsymbol{H}=\boldsymbol{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}'_n$，$\boldsymbol{A}$ 的特征根记作 λ_1，λ_2，…，λ_p（$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$）。为简单起见，我们设 $\lambda_1>\lambda_2>\cdots>\lambda_p>0$，显见 λ_1，λ_2，…，λ_p 也是 $\boldsymbol{B}=\boldsymbol{H}\boldsymbol{X}\boldsymbol{X}'\boldsymbol{H}$ 的非零特征根。注意 $\boldsymbol{H}\boldsymbol{X}$ 的行是将 $\boldsymbol{X}$ 的行中心化，故 $\boldsymbol{B}=(b_{ij})$ 的元素可表示为：

$$b_{ij}=(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'(\boldsymbol{x}_j-\bar{\boldsymbol{x}}) \tag{14.21}$$

记 $\boldsymbol{v}_{(i)}$ 为 $\boldsymbol{B}$ 对应于 λ_i 的特征向量，且 $\boldsymbol{v}'_{(i)}\boldsymbol{v}_{(i)}=\lambda_i$（$i=1,2,\cdots,p$），令

$$\boldsymbol{V}_{(k)}=(\boldsymbol{v}_{(1)},\boldsymbol{v}_{(2)},\cdots,\boldsymbol{v}_{(k)})=(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n)' \tag{14.22}$$

则称 $\boldsymbol{v}_1$，$\boldsymbol{v}_2$，…，$\boldsymbol{v}_n$ 为 $\boldsymbol{X}$ 的 k 维主坐标。

显然，主坐标的概念是从构造点的古典解引申出来的。若将 $\boldsymbol{X}$ 的行看成 R^p 中的 n 个点，它们之间的欧氏距离阵记作 $\boldsymbol{D}$。由定理 14.1 的必要性证明，$\boldsymbol{D}$ 在 R^k 中拟合构造点的古典解正是 $\boldsymbol{X}$ 的 k 维主坐标。下一个定理进一步给出了 $\boldsymbol{X}$ 的 k 维主坐标和主成分之间的关系。

定理 14.3 $\boldsymbol{X}$ 的 k 维主坐标正好是将 $\boldsymbol{X}$ 中心化后 n 个样品的前 k 个主成分的值。

证明：由主成分分析章节中有关论述知，$\boldsymbol{X}$ 的主成分是求 $\boldsymbol{A}=\boldsymbol{X}'\boldsymbol{H}\boldsymbol{X}$ 的特征根 λ_1，λ_2，…，λ_p（$\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$）和相应的特征向量 $\boldsymbol{t}_{(1)}$，$\boldsymbol{t}_{(2)}$，…，$\boldsymbol{t}_{(p)}$，记

$$\boldsymbol{\Lambda}=\mathrm{diag}(\lambda_1,\lambda_2,\cdots,\lambda_p),\quad \boldsymbol{T}=(\boldsymbol{t}_{(1)},\boldsymbol{t}_{(2)},\cdots,\boldsymbol{t}_{(p)})$$

则

$$\boldsymbol{T}'\boldsymbol{T}=\boldsymbol{I}$$

$$\boldsymbol{X}'\boldsymbol{H}\boldsymbol{X}=\boldsymbol{T}\boldsymbol{\Lambda}\boldsymbol{T}' \tag{14.23}$$

若 $\boldsymbol{W}$ 为任一 $n\times p$ 阵，由矩阵的分解，$\boldsymbol{W}$ 可分解为：

$$\boldsymbol{W}=\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{Q}$$

式中，$\boldsymbol{U}$：$n\times p$，$\boldsymbol{U}'\boldsymbol{U}=\boldsymbol{I}_p$，$\boldsymbol{\Lambda}$ 为对角阵，$\boldsymbol{Q}$ 为 p 阶正交阵。而且 $\boldsymbol{\Lambda}^2$ 的对角元素为 $\boldsymbol{W}'\boldsymbol{W}$（或 $\boldsymbol{W}\boldsymbol{W}'$）的特征根，$\boldsymbol{U}$ 的列向量为 $\boldsymbol{W}\boldsymbol{W}'$ 的特征向量，$\boldsymbol{Q}$ 的行向量为 $\boldsymbol{W}'\boldsymbol{W}$ 的特征向量。

现将 $\boldsymbol{W}$ 取成 $\boldsymbol{H}\boldsymbol{X}$，并将 $\boldsymbol{\Lambda}$ 吸收到 $\boldsymbol{U}$ 中，且调整 $\{\boldsymbol{t}_{(i)}\}$ 的符号，使得

$$\boldsymbol{H}\boldsymbol{X}=\boldsymbol{V}\boldsymbol{T}' \tag{14.24}$$

其中 $\boldsymbol{V}=\boldsymbol{V}_p=(\boldsymbol{v}_{(1)},\ \boldsymbol{v}_{(2)},\ \cdots,\ \boldsymbol{v}_{(p)})$（参见式（14.22））。

由于 $\boldsymbol{H}\boldsymbol{X}$ 和 $\boldsymbol{X}$ 的样本离差阵均为 $\boldsymbol{X}'\boldsymbol{H}\boldsymbol{X}$，它们的主成分一样，由前面主成分分析中的叙述，$\boldsymbol{H}\boldsymbol{X}$ 的前 k 个主成分的值为：

$$\boldsymbol{H}\boldsymbol{X}\boldsymbol{T}_k=\boldsymbol{V}\boldsymbol{T}'\boldsymbol{T}_k=\boldsymbol{V}(\boldsymbol{I}_k\boldsymbol{0})'=\boldsymbol{V}_k$$

式中，$\boldsymbol{T}_k=(\boldsymbol{t}_{(1)},\ \boldsymbol{t}_{(2)},\ \cdots,\ \boldsymbol{t}_{(k)})$。注意 $\boldsymbol{H}\boldsymbol{X}$ 是 $\boldsymbol{X}$ 的中心化，上式是 $\boldsymbol{H}\boldsymbol{X}$ 的前 k 个主成分值，它正好是 $\boldsymbol{X}$ 的 k 维主坐标。

由这个定理的结论，我们可以从另一个角度来描述 $\boldsymbol{X}$ 的 k 维主坐标。若 $\boldsymbol{D}$ 是欧氏型的，$n\times p$ 阵 $\boldsymbol{X}$ 是它的构造点，$\hat{\boldsymbol{X}}$ 是 $n\times k$ 阵（$k<p$），是 $\boldsymbol{D}$ 的低维拟合构造点，$\hat{\boldsymbol{X}}$ 相应的距离阵为 $\hat{\boldsymbol{D}}$。定理 14.3 和以上的讨论指出，这个低维拟合构造点是 $\boldsymbol{H}\boldsymbol{X}\boldsymbol{T}_k$，由于 $\boldsymbol{H}$ 仅起中心化的作用，故拟合构造点等价于 $\boldsymbol{X}\boldsymbol{T}_k$，即 $\boldsymbol{X}$ 右乘一个列单位正交矩阵。

现在考虑一切形如 $\hat{\boldsymbol{X}}=\boldsymbol{X}\boldsymbol{\Gamma}_1$ 的拟合构造点，其中，$\boldsymbol{\Gamma}_1$：$p\times k$，$\boldsymbol{\Gamma}=(\boldsymbol{\Gamma}_1,\ \boldsymbol{\Gamma}_2)=(\boldsymbol{\gamma}_{(1)},\ \boldsymbol{\gamma}_{(2)},\ \cdots,\ \boldsymbol{\gamma}_{(p)})$ 为 p 阶正交阵，易见

$$d_{ij}^2=\sum_{t=1}^{p}(x_{it}-x_{jt})^2=\sum_{t=1}^{p}(\boldsymbol{x}_i'\boldsymbol{\gamma}_{(t)}-\boldsymbol{x}_j'\boldsymbol{\gamma}_{(t)})^2$$

$$\hat{d}_{ij}^2=\sum_{t=1}^{k}(\boldsymbol{x}_i'\boldsymbol{\gamma}_{(t)}-\boldsymbol{x}_j'\boldsymbol{\gamma}_{(t)})^2$$

后者为拟合构造点之间的距离。上两式表明 $\hat{d}_{ij}\leqslant d_{ij}$，因此可以用

$$\phi=\sum_{i=1}^{n}\sum_{j=1}^{n}(d_{ij}^2-\hat{d}_{ij}^2) \tag{14.25}$$

来度量 $\hat{\boldsymbol{X}}$ 拟合 $\boldsymbol{X}$（或 $\boldsymbol{D}$）的程度。下面的定理指出，在一切形如 $\hat{\boldsymbol{X}}=\boldsymbol{X}\boldsymbol{\Gamma}_1$ 的 k 维构造点中，$\boldsymbol{\Gamma}_1=\boldsymbol{T}_k$ 为最优，即相应的 ϕ 最小，而 $\hat{\boldsymbol{X}}=\boldsymbol{X}\boldsymbol{T}_k$ 正是 $\boldsymbol{X}$ 的 k 维主坐标，故给出了 k 维主坐标的优良性的一种描述。

定理 14.4　设 $\boldsymbol{D}$ 是欧氏型的距离阵，$\boldsymbol{X}(n\times p)$ 是它的构造点。给定 $k(1\leqslant k\leqslant p)$，则一切形如 $\hat{\boldsymbol{X}}=\boldsymbol{X}\boldsymbol{\Gamma}_1$（使 $\boldsymbol{\Gamma}=(\boldsymbol{\Gamma}_1,\ \boldsymbol{\Gamma}_2)$ 为 p 阶正交阵）的 k 维构造点中，$\boldsymbol{\Gamma}_1=\boldsymbol{T}_k$ 使 ϕ 达到最小。证明略。有兴趣者请参见参考文献 [1]。

14.4 非度量方法

古典解是基于主成分分析的思想，这时

$$d_{ij}=\hat{d}_{ij}+e_{ij}$$

$\hat{d}_{ij}$是拟合于d_{ij}的值，e_{ij}是误差。但有时，d_{ij}和$\hat{d}_{ij}$之间的拟合关系可以表示为：

$$d_{ij}=f(\hat{d}_{ij}+e_{ij}) \tag{14.26}$$

式中，f为一个未知的单调增加的函数。这时，我们用来构造$\hat{d}_{ij}$的唯一信息是利用$\{d_{ij}\}$的秩，将$\{d_{ij},\ i<j\}$由小到大排列为：

$$d_{i_1j_1}\leqslant d_{i_2j_2}\leqslant\cdots\leqslant d_{i_mj_m},\quad m=\frac{1}{2}n(n-1)$$

$(i,\ j)$所对应的d_{ij}在上面的排列中的名次（由小到大）称为$(i,\ j)$或d_{ij}的秩。我们欲寻找一个拟合构造点，使后者相互之间的距离也有如上的次序：

$$\hat{d}_{i_1j_1}\leqslant\hat{d}_{i_2j_2}\leqslant\cdots\leqslant\hat{d}_{i_mj_m}$$

并记为：

$$\hat{d}_{ij}\xrightarrow{\text{单调}}d_{ij}$$

这种模型大多出现在相似系数矩阵的场合，因为相似系数强调的是物品之间的相似，而不是它们的距离。

求这个模型的解有一些方法，其中以 Shepard-Kruskal 算法最为流行，它的步骤如下：

(1)已知一个相似系数矩阵$\boldsymbol{D}=(d_{ij})$(这里仍用$\boldsymbol{D}$来记相似系数矩阵)，并将其非对角线元素由小到大排列：

$$d_{i_1j_1}\leqslant d_{i_2j_2}\leqslant\cdots\leqslant d_{i_mj_m},\quad m=\frac{1}{2}n(n-1);\ i_l<j_l;l=1,2,\cdots,m$$

(2) 设$\hat{\boldsymbol{X}}(n\times k)$是$k$维拟合构造点，相应的距离阵$\hat{\boldsymbol{D}}=(\hat{d}_{ij})$，令

$$S^2(\hat{\boldsymbol{X}})=\frac{\min\sum\limits_{i<j}(d_{ij}^*-\hat{d}_{ij})^2}{\sum\limits_{i<j}d_{ij}^2} \tag{14.27}$$

极小是对一切$\{d_{ij}^*\}$($d_{ij}^*\xrightarrow{\text{单调}}d_{ij}$)进行的，使上式达到极小的$\{d_{ij}^*\}$称为$\hat{d}_{ij}$对$\{d_{ij}\}$的最小二乘单调回归。

如果$\hat{d}_{ij}\xrightarrow{\text{单调}}d_{ij}$，在式(14.27)中取$d_{ij}^*=\hat{d}_{ij}(i<j)$，这时，$S^2(\hat{\boldsymbol{X}})=0$，$\hat{\boldsymbol{X}}$是$\boldsymbol{D}$的构造点。

若将 $\boldsymbol{X}$ 的列作一正交平移变换 $\boldsymbol{y}_t=\boldsymbol{\Gamma x}_t+\boldsymbol{b}$，$\boldsymbol{\Gamma}$ 为正交阵，$\boldsymbol{b}$ 为常向量，则式（14.27）的分子不变。

（3）若 k 固定，且能存在一个 $\hat{\boldsymbol{X}}_0$，使得

$$S(\hat{\boldsymbol{X}}_0)=\min_{\hat{\boldsymbol{X}}:n\times k} S(\hat{\boldsymbol{X}})\equiv S_k$$

则称 $\hat{\boldsymbol{X}}_0$ 为 k 维最佳拟合构造点。

（4）由于 S_k（也称为压力指数）是 k 的单调下降序列，取 k，使 S_k 适当地小。例如 $S_k\leqslant 5\%$最好，$5\%<S_k\leqslant 10\%$次之，$S_k>10\%$较差。

求解可以用梯度法进行迭代（见参考文献［1］）。

14.5　多维标度法的上机实现

14.5.1　多维标度法的步骤及逻辑框图

多维标度法的实现主要有以下几个步骤：（1）确定研究的目的；（2）选择需要进行比较分析的样品和原始变量（或者距离矩阵）；（3）选择适当的求解方法，分析样品间的距离矩阵；（4）选择适当的维数，得到距离阵的古典解，将各个样品直观地呈现出来并对结果进行解释；（5）检验模型的拟合情况。

多维标度法实现的逻辑框图如图 14—1 所示。

图 14—1　多维标度法实现框图

14.5.2　多维标度法的计算机实现

多维标度法可以通过 SPSS 软件中的 Multidimensional Scaling 来实现。本章的

数据文件仍使用 World95. sav 数据，我们从其中选取亚洲国家和地区进行比较分析（Region=3）。

例 14—5

为了分析亚洲国家和地区的经济发展和文教卫生水平，我们可以使用多维标度法对这些国家和地区进行分析，将结果直观地呈现在图上。这里选取了如下变量：urban（城市人口的比例），lifeexpf（女性平均寿命），lifeexpm（男性平均寿命），literacy（会阅读的人的比例），babymort（婴儿死亡率），gdp _ cap（人均 GDP），death _ rt（死亡率），lit _ male（男性中会阅读的比例），lit _ fema（女性中会阅读的比例）。

进入 SPSS 软件后，使用菜单 Analyze→Scale→Multidimensional Scaling（ALSCAL），进入多维标度法的对话框。将上述 9 个变量选入变量框内，Individual Matrices for 是指用来分割样本的属性变量（通常是分类变量）。在 Distances 选项中，如果输入的是样品间的距离阵，就选择 Data are distances；如果输入的是原始变量，样品间的距离阵要通过原始变量来计算，就选择 Create distances from data，这里我们选择后者，点击 Measure 选择间隔尺度（Interval），样本间的欧氏距离，并且将变量标准化（Transform Values 的 Standardize 选项中选择 Z scores）。距离矩阵（Create Distance Matrix）选择样品间（Between cases）。点击 Model，选项中测量水平（Level of Measurement）选择间隔长度（Interval），标度模型（Scaling Model）选择欧氏距离、二维模型。在 Options 的 Display 选项中，可以选择需要的输出结果，这里我们全选。得到输出结果 14—1 至输出结果 14—3。

每个观测代表的样品是：

Afghanistan	N. Korea	Bangladesh	Cambodia	China Mainland	Hong Kong	India	Indonesia
1	2	3	4	5	6	7	8
Malaysia	Pakistan	Philippines	S. Korea	Singapore	Thailand	Vietnam	
9	10	11	12	13	14	15	

说明：原数据中 Japan，Taiwan 有缺失，处理时剔除。

输出结果 14—1　　样品的距离阵（标准化变量的样品距离阵）

Raw (unscaled) Data for Subject 1

	1	2	3	4	5
1	.000				
2	3.062	.000			
3	2.089	1.165	.000		
4	6.397	4.024	4.603	.000	
5	8.724	6.656	7.179	4.113	.000

续输出结果 14—1

6	4.194	1.619	2.276	2.450	5.349
7	5.566	3.378	3.849	1.010	4.497
8	7.008	4.531	5.172	1.122	3.262
9	7.836	5.666	6.189	2.011	3.546
10	3.636	.921	1.819	3.604	5.988
11	6.684	4.520	5.022	1.174	4.013
12	8.066	5.969	6.468	2.603	2.516
13	8.967	6.976	7.467	4.273	.868
14	7.213	4.954	5.467	1.201	4.084
15	6.501	4.331	4.781	.911	4.470
	6	7	8	9	10
6	.000				
7	1.829	.000			
8	2.974	1.704	.000		
9	4.085	2.443	1.587	.000	
10	1.320	3.069	4.017	5.192	.000
11	2.970	1.256	1.281	1.260	4.140
12	4.437	2.952	1.872	1.217	5.453
13	5.629	4.601	3.403	3.354	6.352
14	3.413	1.708	1.318	1.460	4.618
15	2.797	1.068	1.521	1.872	4.057
	11	12	13	14	15
11	.000				
12	1.957	.000			
13	3.941	2.202	.000		
14	1.015	2.113	4.050	.000	
15	.906	2.574	4.483	.803	.000

上面的结果是样品之间的距离阵，这里采用的是欧氏距离，距离阵为欧氏距离阵。

输出结果 14—2　　迭代过程和距离阵的古典解

```
Iteration history for the 2 dimensional solution (in squared distances)
Young's S-stress formula 1 is used.
Iteration        S-stress        Improvement
1    .035 46
2    .027 68        .007 77
3    .026 10        .001 58
4    .025 60        .000 51
Iterations stopped because
S-stress improvement is less than .001 000
Stress and squared correlation (RSQ) in distances
RSQ values are the proportion of variance of the scaled data (disparities)
in the partition (row, matrix, or entire data) which
is accounted for by their corresponding distances.
Stress values are Kruskal's stress formula 1.
For matrix
Stress = .033 35     RSQ = .995 99
Configuration derived in 2 dimensions
Stimulus Coordinates
Dimension
```

续输出结果 14—2

Stimulus Number	Stimulus Name	1	2
1	VAR1	2.750 4	−.359 0
2	VAR2	1.585 8	−.094 1
3	VAR3	1.872 9	−.106 2
4	VAR4	−.298 4	.408 5
5	VAR5	−1.576 1	−1.130 4
6	VAR6	.792 2	.041 8
7	VAR7	.049 3	.307 2
8	VAR8	−.654 3	.120 5
9	VAR9	−1.125 0	.460 4
10	VAR10	1.283 5	−.351 1
11	VAR11	−.516 7	.440 8
12	VAR12	−1.336 8	−.032 7
13	VAR13	−1.767 7	−.996 2
14	VAR14	−.714 7	.642 0
15	VAR15	−.344 5	.648 4

Optimally scaled data (disparities) for subject 1

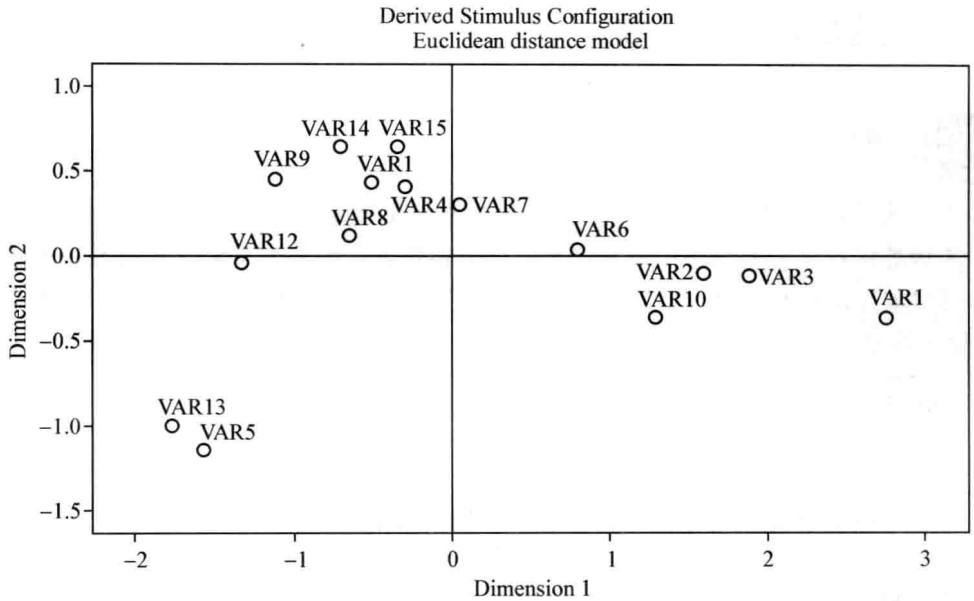
	1	2	3	4	5
1	.000				
2	1.417	.000			
3	.902	.414	.000		
4	3.180	1.925	2.231	.000	
5	4.410	3.317	3.593	1.972	.000
6	2.015	.654	1.001	1.093	2.626
7	2.740	1.584	1.833	.332	2.176
8	3.503	2.193	2.532	.391	1.523
9	3.941	2.793	3.070	.861	1.672
10	1.720	.284	.759	1.703	2.964
11	3.332	2.188	2.453	.419	1.920
12	4.062	2.954	3.218	1.174	1.128
13	4.539	3.486	3.746	2.057	.257
14	3.611	2.417	2.688	.433	1.957
15	3.235	2.087	2.326	.279	2.161

	6	7	8	9	10
6	.000				
7	.765	.000			
8	1.370	.699	.000		
9	1.958	1.089	.637	.000	
10	.495	1.420	1.921	2.543	.000
11	1.368	.462	.475	.464	1.987
12	2.143	1.358	.787	.441	2.681
13	2.774	2.230	1.597	1.571	3.156
14	1.602	.701	.494	.570	2.239
15	1.276	.362	.602	.787	1.943

	11	12	13	14	15
11	.000				
12	.833	.000			
13	1.882	.962	.000		
14	.335	.915	1.939	.000	
15	.277	1.159	2.168	.222	.000

输出结果 14—2 反映了迭代过程的一些结果，并且得到了各个样品的古典解（二维的坐标）。从结果可以看出，Young 压力指数为 0.025 60，K 压力指数小于 5%，RSQ 为 99.599%，所以认为该模型拟合的效果比较好。

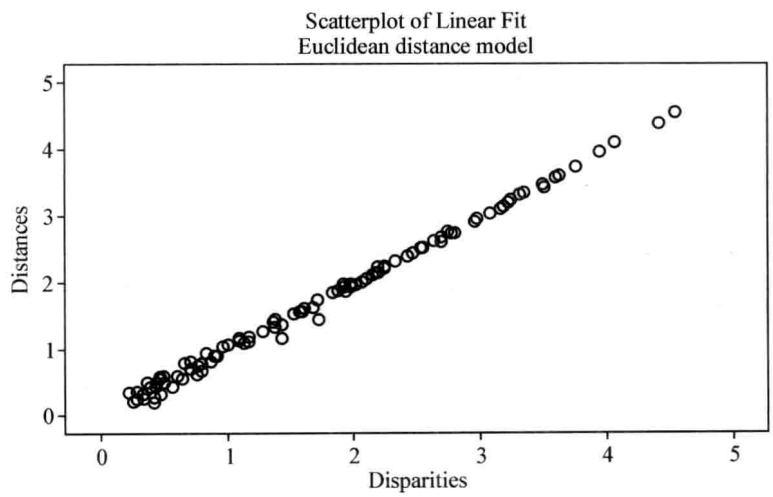
输出结果 14—3　　**二维平面图**



输出结果 14—3 就是在二维平面上直观地反映 15 个亚洲国家或地区在 9 个变量的指标体系中所处的位置。

输出结果 14—4 是欧氏距离线性拟合的散点图，由此图可评价模型拟合的优劣。

输出结果 14—4　　**欧氏距离线性拟合的散点图**



上面的线性拟合图是欧氏距离模型线性拟合的散点图。由散点图可以看出，欧氏距离（实际距离）与差异（disparities）点（拟合距离）在 $y=x$ 直线的附近，这说明模型拟合的效果是比较理想的。

14.6 社会经济案例研究

例 14—6

仍然采用 35 家上市公司的例子，为了能够清楚地显示公司之间的相似性，这里选取 9 家信息技术公司来做多维标度分析。这里仍然选择这 8 个变量来构造样品间的距离阵（使用欧氏距离），考虑到量纲的影响，我们将变量进行标准化。得到输出结果 14—5 和输出结果 14—6。

输出结果 14—5

For matrix

Stress =0.077 74　　RSQ =0.968 69

该结果说明压力指数小于 10%，拟合结果还是可以接受的。

输出结果 14—6

Raw (unscaled) Data for Subject 1

	1	2	3	4	5
1	.000				
2	2.719	.000			
3	3.857	2.755	.000		
4	4.311	3.291	3.969	.000	
5	3.764	2.032	1.418	3.595	.000
6	1.345	2.491	3.233	4.559	3.300
7	4.769	3.642	2.780	5.695	3.138
8	3.956	3.730	6.048	5.354	5.226
9	3.975	3.441	2.523	5.346	3.033

	6	7	8	9
6	.000			
7	3.565	.000		
8	4.732	6.974	.000	
9	2.748	1.128	6.717	.000

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

	Iteration	S-stress	Improvement
1	.096 66		
2	.090 81	.005 86	
3	.090 37	.000 43	

Iterations stopped because

S-stress improvement is less than .001 000

Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data (disparities)

续输出结果 14—6

in the partition (row, matrix, or entire data) which
is accounted for by their corresponding distances.
Stress values are Kruskal's stress formula 1.
For matrix
Stress = .077 74　　RSQ = .968 69
Configuration derived in 2 dimensions
Stimulus Coordinates
Dimension

Stimulus Number	Stimulus Name	1	2
1	VAR1	.729 3	−.764 7
2	VAR2	.355 6	.073 0
3	VAR3	−.845 3	.494 2
4	VAR4	.747 0	1.698 7
5	VAR5	−.363 9	.565 3
6	VAR6	−.025 5	−.765 1
7	VAR7	−1.669 8	−.331 8
8	VAR8	2.454 2	−.560 0
9	VAR9	−1.381 5	−.409 7

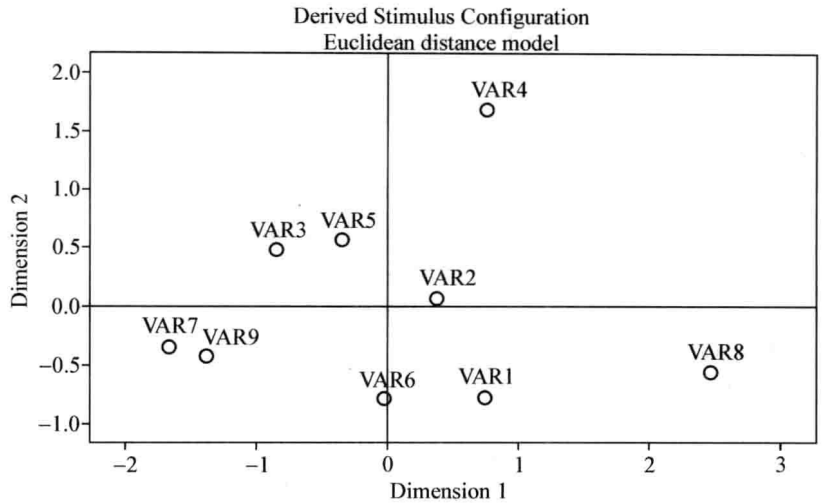
Optimally scaled data (disparities) for subject 1

	1	2	3	4	5
1	.000				
2	1.208	.000			
3	1.973	1.233	.000		
4	2.278	1.593	2.048	.000	
5	1.911	.747	.334	1.797	.000
6	.285	1.055	1.554	2.445	1.599
7	2.586	1.828	1.249	3.208	1.490
8	2.039	1.887	3.446	2.979	2.893
9	2.052	1.693	1.076	2.974	1.419

	6	7	8	9
6	.000			
7	1.777	.000		
8	2.561	4.068	.000	
9	1.228	.139	3.895	.000

该输出结果是每个样品的相对位置，也就是二维坐标值，可以根据这个值在二维平面上直观地显示每个样品（见输出结果 14—7）。

输出结果 14—7



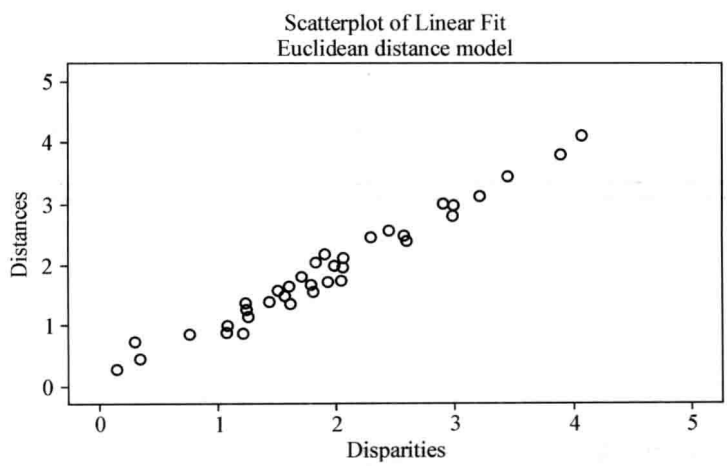
样品顺序对照表

中兴通讯	长城电脑	青鸟华光	清华同方	永鼎光缆	宏图高科	海星科技	方正科技	复华实业
1	2	3	4	5	6	7	8	9

该图直观地显示了每个样品所处的位置。方正科技（Case 8）和清华同方（Case 4）位于图形的两端。结合原始数据可以认为，方正科技的盈利能力是 9 家信息产业上市公司中最强的，而清华同方的负债情况较好，偿债能力和成长能力是 9 家公司中最强的。综合分析认为，第一维（横轴）主要反映公司的盈利能力，第二维主要反映公司的成长能力。我们可以从图中找出比较相似的公司，比如海星科技（Case 7）与复华实业（Case 9）是比较类似的，它们的盈利能力和成长能力都较差。

输出结果 14—8 是反映实际距离与拟合距离的散点图（这里采用欧氏距离），由图来看，拟合情况还是可以接受的，样本点基本落在 $y=x$ 直线附近。

输出结果 14—8



例 14—7

表 14—3 列出了我国 10 个城市间的距离（单位：公里），由于公路不是平直的，所以它们不是城市之间的最短距离，只可以看作这些城市之间的近似距离。我们希望利用这些距离数据画一张平面地图，标出这 10 个城市的位置，使之尽量接近表中所给出的距离数据，从而反映它们的真实地理位置。

表 14—3　　国内 10 个城市之间的距离数据

城市	天津	北京	锦州	沈阳	长春	哈尔滨	满洲里	齐齐哈尔	牡丹江	吉林
天津	0									
北京	137	0								
锦州	499	486	0							
沈阳	741	728	242	0						
长春	1 046	1 033	547	305	0					
哈尔滨	1 288	1 275	789	547	242	0				
满洲里	2 326	2 210	1 724	1 482	1 177	935	0			
齐齐哈尔	1 451	1 335	849	760	530	288	693	0		
牡丹江	1 746	1 630	1 144	902	597	355	1 290	643	0	
吉林	1 187	1 174	688	446	128	275	1 210	563	630	0

打开上表数据，在 SPSS 中进行以下操作：Analyze→Scale→Multidimensional Scaling，进入多维标度法的对话框。将 10 个变量（城市）选入变量框中。上面的操作中也用到了另外两个对话框，在 Model 子对话框中将数据的测量尺度改为比率尺度，后面的 Optional 子对话框中要求结果中显示空间匹配图。注意，空间匹配图是多维标度分析中非常有用的工具，如果在 SPSS 中不是默认的输出结果，则要手动选择它。

结果输出标题为“Alscal”，随后方框中显示的是在 SPSS 默认的情况下，两维空间的迭代记录。可以看出，在迭代 3 次后 S-stress 值的变化为0.000 08，小于默认的 0.001，达到收敛标准（见表 14—4）。

表 14—4

Iteration history for the 2 dimensional solution (in squared distances)

Young's S-stress formula 1 is used.

Iteration	S-stress	Improvement
1	.067 32	
2	.063 62	.003 70
3	.063 53	.000 08

Iterations stopped because

S-stress improvement is less than .001 000

表 14—5 中是统计量 Stress 和 RSQ 的具体解释和计算结果。RSQ 即决定系数，表示总变异中能够被相对空间距离解释的比例。Stress 是 K 压力指数。这里

Stress 值为 0.052 59，RSQ 的值为 0.985 05，已经非常接近 1。所以，关于中国 10 个城市距离的多维标度模型的拟合效果相当好。

表 14—5

Stress and squared correlation (RSQ) in distances

RSQ values are the proportion of variance of the scaled data (disparities) in the partition row, matrix, or entire data which is accounted for by their corresponding distances.

Stress values are Kruskal's stress formula 1.

For matrix

Stress = .052 59　　RSQ = .985 05

接下来表 14—6 显示的是我国 10 个城市在二维空间的一种坐标值。根据多维标度法解的概念和有关性质，它的解不是唯一的。

表 14—6

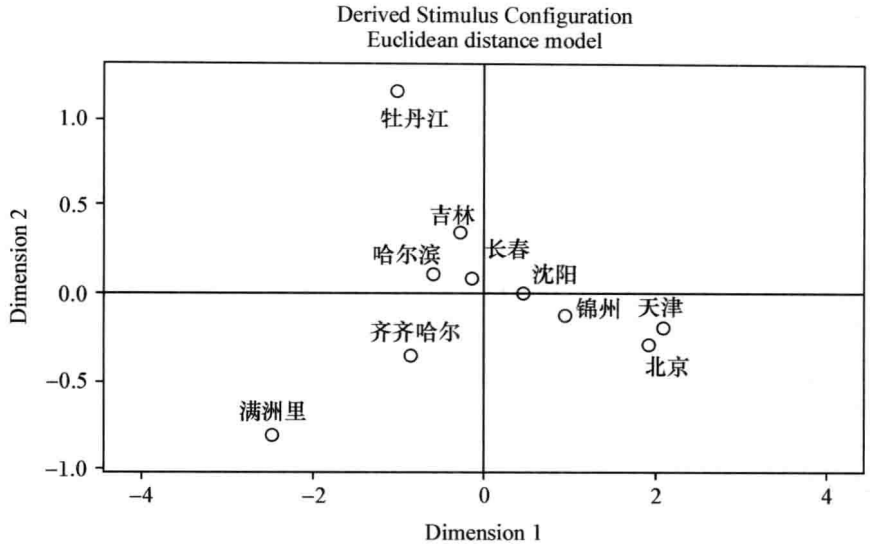
Configuration derived in 2 dimensions

Stimulus Coordinates

Stimulus Number	Stimulus Name	Dimension 1	Dimension 2
1	天津	1.921 9	—.282 5
2	北京	2.095 0	—.191 9
3	锦州	.931 1	—.116 0
4	沈阳	.468 4	.011 5
5	长春	—.141 3	.089 8
6	哈尔滨	—.616 2	.107 8
7	满洲里	—2.475 5	—.802 3
8	齐齐哈尔	—.863 4	—.344 7
9	牡丹江	—1.026 4	1.172 6
10	吉林	—.293 6	.358 7

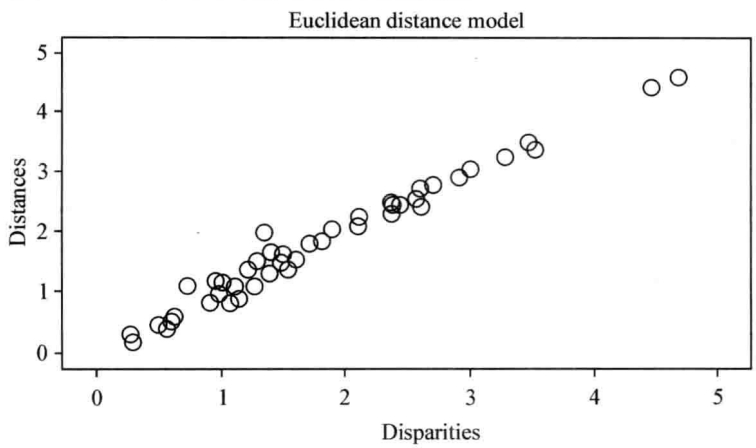
输出结果 14—9 是多维尺度分析中输出的我国 10 城市二维空间匹配图。它是系统将各个城市（变量）按照由实际距离计算出来的相应距离在空间中排列起来的。由于它的解不是唯一的，所以看上去与地图上的排列不完全相同。事实上大家可以发现，和实际地图相比，其差异就在于地图的坐标系进行了旋转，而各城市的相对位置均保持不变，基本上都是吻合的。这正是多维标度法在正交（旋转、平移）变换下有不变性的具体表现。

输出结果 14—9　10 城市二维空间匹配图



输出结果 14—10 是欧氏距离模型线性拟合散点图，提供的是原始数据的不一致程度和用线性模型计算出来的欧氏距离间的散点图。如果模型的拟合程度较高，则所有散点应当在一条直线上。从图中可见各点基本上处在一条直线上，没有明显的离群点，因此模型的拟合效果是比较好的。最后，如果在操作时选择 Options 子对话框中的 Model and options summary 复选框，则结果会输出非常详细的模型拟合参数汇总表，这有助于深入理解多维标度法在拟合时所需要考虑的各种问题。

输出结果 14—10　欧氏距离模型线性拟合散点图



在上面的例子中，城市间的直线距离是可以精确测量的，且为最准确的比率测量。但并非所有的距离都可以像这样被准确地测量，比如品牌间的差异，或者两个概念间的差异。在市场研究中大量的问题只能用问卷的方式以有序测量尺度收集，此时传统的古典多维标度模型不一定适用，而非度量的多维标度模型则更为合适。下面给出一个非度量模型的例子。

例 14—8

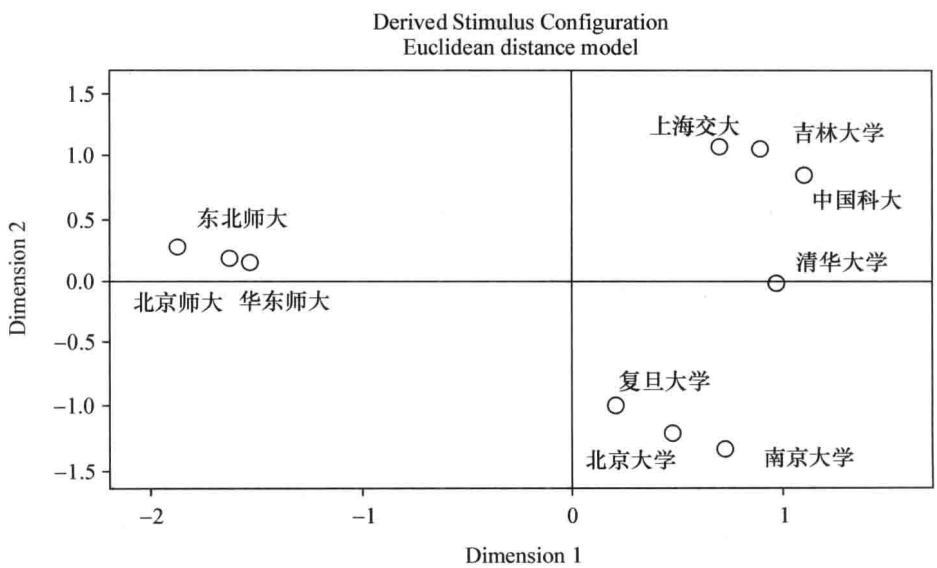
在一次调查中，收集了中国人民大学统计学院某年级 40 位大学生对中国 10 所大学差异性的评分。1 分表示差异较小，9 分表示差异较大，从 1 分到 9 分的差异程度逐渐增加。下面是某位学生的打分数据（见表 14—7）。

表 14—7　　某位大学生的测量数据

学校名称	北京大学	南京大学	吉林大学	中国科大	复旦大学	华东师大	清华大学	北京师大	上海交大	东北师大
北京大学	0									
南京大学	4	0								
吉林大学	6	5	0							
中国科大	5	6	4	0						
复旦大学	3	4	6	5	0					
华东师大	6	7	7	7	6	0				
清华大学	4	6	5	5	4	6	0			
北京师大	7	7	7	7	6	2	6	0		
上海交大	6	6	4	3	5	6	4	7	0	
东北师大	7	8	7	8	6	2	8	3	7	0

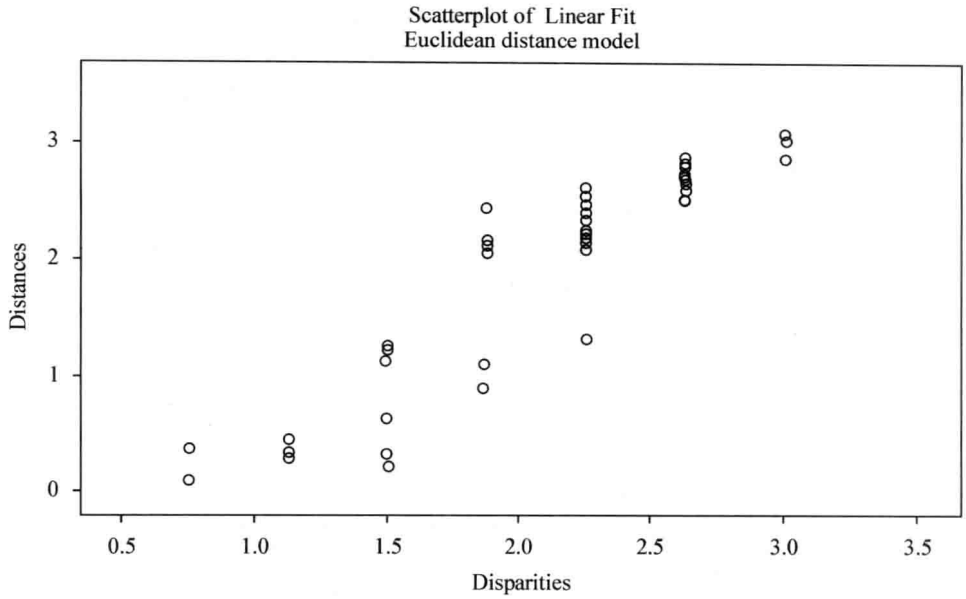
若将该数据直接进行古典多维标度模型的分析，则最终得到的定位见输出结果 14—11 和输出结果 14—12。

输出结果 14—11　使用古典模型对某位学生评价结果的空间定位图



可见，该同学非常明显地将师范大学、文理综合性大学和工科大学区分开。同时模型的拟合指标为 Stress=0.234 6，RSQ=0.840 02，模型的解释程度一般。

输出结果 14—12　使用古典模型的拟合效果图



由于该数据显然是以问卷打分的方式来判断距离的远近，作为有序尺度来加以分析更为合适。

在 Model 子对话框中将测量尺度改为 Ordinal，且选中下方的 Untie tied observation 复选框以区分相同的分值，重新进行分析，主要结果如表 14—8 所示。

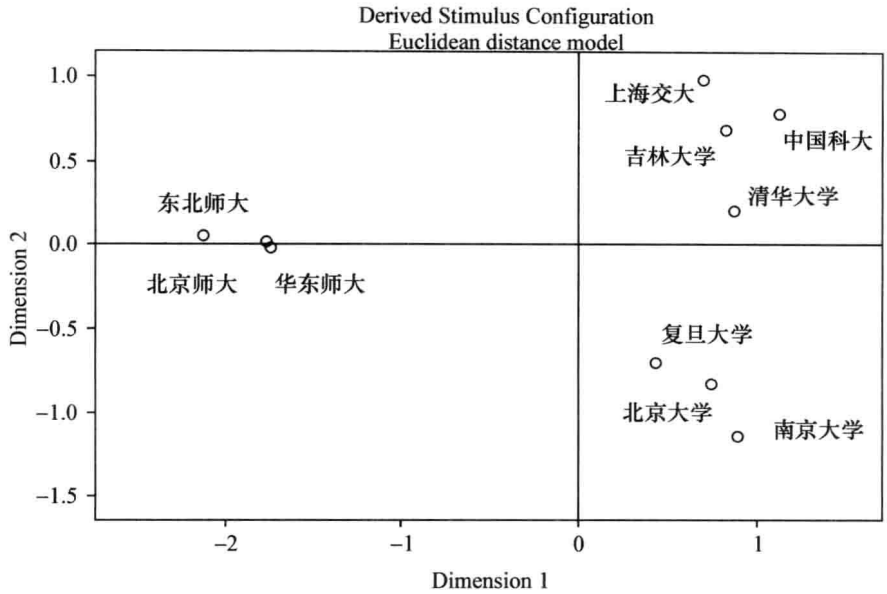
表 14—8

Stress and squared correlation (RSQ) in distances
RSQ values are the proportion of variance of the scaled data (disparities)
in the partition (row, matrix, or entire data) which
is accounted for by their corresponding distances.
Stress values are Kruskal's stress formula 1.
For matrix
Stress = .045 116　　RSQ = .991 41

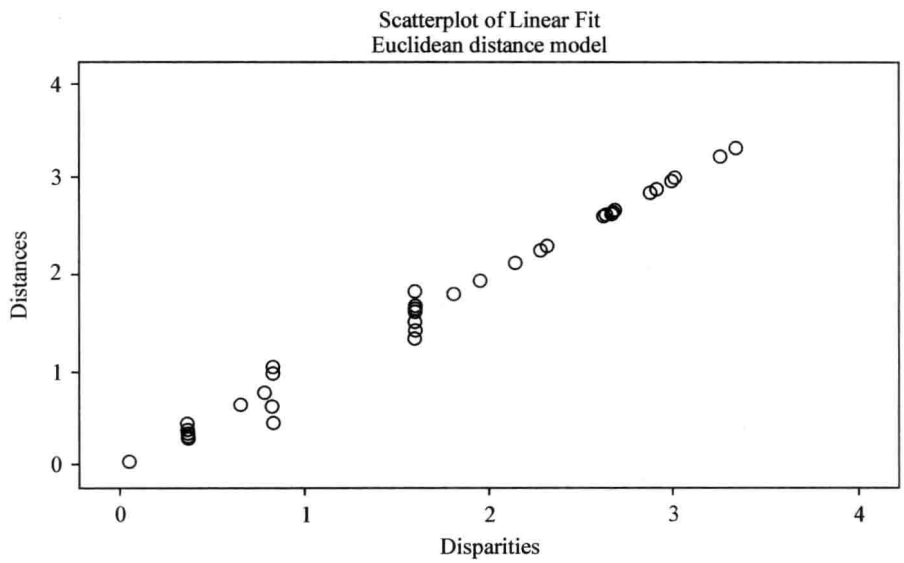
模型的拟合指标为 Stress=0.045 116，RSQ=0.991 41，显然效果要比古典模型好得多。但应注意结果中对这些指标的说明，这里的决定系数实际上指的是变换后数据的解释度，而变换中显然会损失一部分信息，因此该研究对原数据的解释效果是否更好还很难说，见输出结果 14—13。

可见和古典模型的结果比较，现在的结果很明显地缩小了师范大学间的差异，同时综合大学和理工科大学间的界面也开始变得模糊起来。清华、北大、复旦三所大学之间变得更为接近了，和原来的结果相比，这更合乎现实中人们对大学的印象。输出结果 14—14 给出的是变换后数据的拟合效果散点图，显然模型对变换后数据的解释度非常高。

输出结果 14—13　使用非古典模型对某位学生评价结果的空间定位图

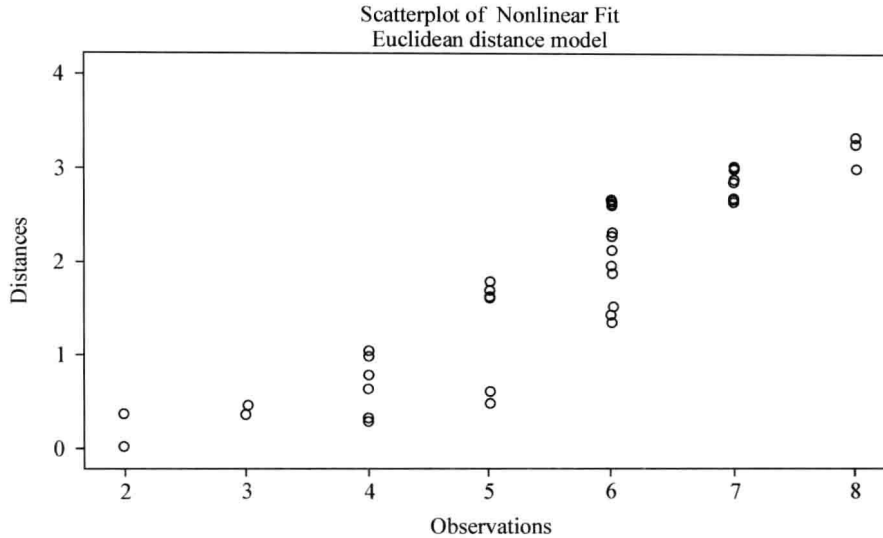


输出结果 14—14　使用非古典模型的拟合效果图

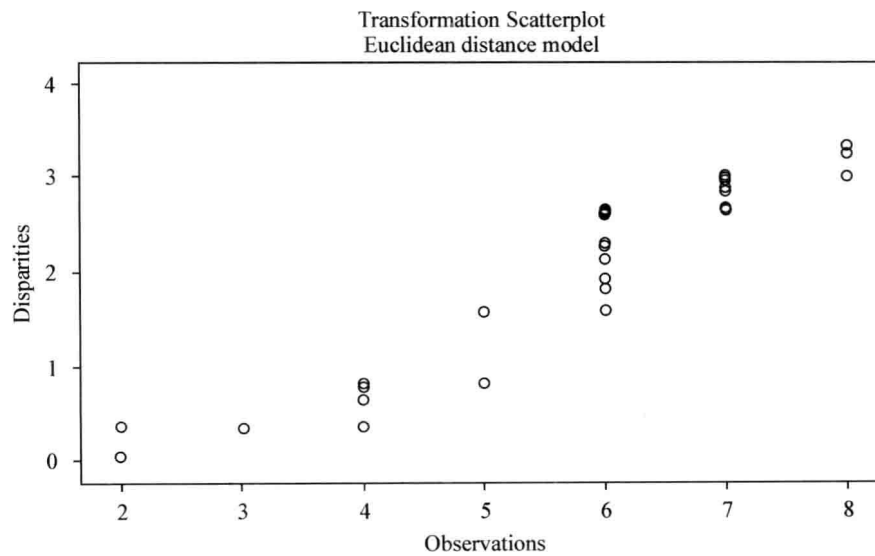


输出结果 14—15、输出结果 14—16 是模型对原始数据的解释程度，输出结果 14—15 给出的是经过连续变换后原始数据和最终的模型距离间的对应关系，输出结果 14—16 给出的是原始数据和变换数据间的关系。显然，和古典模型相比，当前模型对原始数据的解释程度更高。

输出结果 14—15



输出结果 14—16



□ 参考文献

[1] 方开泰. 实用多元统计分析. 上海：华东师范大学出版社，1989

[2] 张尧庭，方开泰. 多元统计分析引论. 北京：科学出版社，1982

[3] 王国梁，何晓群. 多变量经济数据统计分析. 西安：陕西科学技术出版社，1993

[4] Joseph F. Hair, Rolph E. Anderson, Ronald L. Tatham, William C.

Black. *Multivariate Data Analysis*. Fifth Edition. Prentice-Hall，1998

[5] Seber，G. A. F. *Multivariate Observations*. John Wiley & Sons，Inc.，1984

□ 思考与练习

1. 简述多维标度法的基本思想。
2. 简述实现多维标度法的步骤。
3. 给定距离阵

$$\boldsymbol{D}=\begin{bmatrix}0\\1&0\\2&1&0\\2&2&1&0\\2&2&2&1&0\\1&2&2&2&1&0\\1&1&1&1&1&1&0\end{bmatrix}$$

求它的拟合构造点，并说明它是不是欧氏型的。

4. 试解释样本间相似性的含义。

数理统计学	茆诗松　吕晓玲
统计学概论	贾俊平
统计学（第二版）	金勇进
统计学（第六版）	贾俊平　何晓群　金勇进
“十二五”普通高等教育本科国家级规划教材；教育部推荐教材；国家统计局优秀统计教材	
《统计学（第五版）》学习指导书	贾俊平
统计学——基于 SPSS	贾俊平
应用统计学——基于 SPREADSHEET 工具	耿修林
应用回归分析（第四版）	何晓群　刘文卿
普通高等教育“十一五”国家级规划教材	
统计分析与 SPSS 的应用（第四版）	薛　薇
抽样技术（第三版）	金勇进　杜子芳　蒋　妍
普通高等教育“十一五”国家级规划教材	
教育部普通高等教育精品教材	
应用时间序列分析（第三版）	王　燕
多元统计分析（第四版）	何晓群
“十二五”普通高等教育本科国家级规划教材	
应用随机过程（第三版）	张　波　商　豪
国民经济核算原理与中国实践（第三版）	高敏雪　李静萍　等
普通高等教育“十一五”国家级规划教材	
教育部普通高等教育精品教材	
《国民经济核算原理与中国实践（第三版）》学习指导书	高敏雪　等
宏观经济统计分析（第二版）	赵彦云
市场调查方法与技术（第三版）	简　明　金勇进　蒋　妍
经济社会统计（第三版）	李静萍　高敏雪
现代工业统计与质量管理	王　庚　等

21世纪统计学系列教材

ISBN 978-7-300-20848-0
9 787300 208480 >
定价：39.00 元

人大经管图书在线　www.rdjg.com.cn
了解图书出版信息　下载教学辅助资料

策划编辑　王伟娟　陈永凤
责任编辑　王　前　张佳佳
封面设计　三众工作室／耿中虎