When the group means are unequal both T^2 and F increase as N_1 (and N_2) increases, while D^2 does not. Therefore D^2 is a better description of the distance between groups when distances are compared.

The above formulas assume that the data are available in each case for all the variables compared. If some data values are equal to missing value codes or are out of range, observations for all variables may not be present in each case. Then N_1 and N_2 are replaced by the harmonic means of the frequencies of variables of the first and second groups respectively; this provides an approximate test of the equality of means. The formula is given in Appendix A.3.

TEST

HOTELLing. no/prev.

When HOTELLING is specified, Hotelling's T^2 and Mahalanobis D^2 are computed.

Example 3D.4 Restricting Analysis to Complete Cases

Usually an analysis of a single variable uses all the acceptable values for the variable whether or not the value of any other variable is acceptable. Conversely, the usual definition of Hotelling's T^2 requires that the data values be acceptable for all the variables for any case that is included in the computations.

Cases containing acceptable data values for all variables that are included in the analysis are called complete cases.

In P3D you can specify whether all the computations are to be based on COMPLETE cases only, or on all acceptable values. If COMPLETE is specified, the univariate statistics and the t statistics are also computed using only complete cases.

To illustrate the effect COMPLETE has on the results, we add the command COMPLETE to the TEST paragraph of Example 3D.3 as follows

```
-----------------------------------
/ TEST    VARIABLES ARE CHOLSTRL, ALBUMIN,
                        CALCIUM, URICACID.
          HOTELLING.
          COMPLETE.
-----------------------------------
```

The results for modified analysis are presented in Output 3D.4 and can be compared with the analysis in Output 3D.3. We do not display the results for ALBUMIN, CALCIUM and URICACID. The small differences between the two results are due to the fact that there are only eight cases containing unacceptable data. Note the different frequencies (sample sizes) and degrees of freedom used in the two analyses.

A large difference between the two analyses would indicate that the results may be biased due to the pattern of missing values or values out of range. If your analyses show a large difference, you may want to examine the data by using PAM (Section 12.2) to study the pattern of values excluded from the analysis.

TEST

COMPlete. no/prev.

When COMPLETE is specified, only complete cases are used in all the computations. Complete cases are cases in which the data are acceptable (not missing or out of range) for all variables specified in the USE statement of the VARIABLE paragraph (all variables if USE is not specified). COMPLETE or NO COMPLETE can be specified in only the first TEST paragraph of any problem. It cannot be altered until a new problem begins.

Example 3D.5 Correlation of Variables in Each Group

The CORRELATIONS between the variables in each group are printed when CORRELATION is specified in the TEST paragraph. If we submit the Control Language of Example 3D.1 with the added TEST paragraph

```
-----------------------------------
/ TEST    VARIABLES ARE CHOLSTRL, ALBUMIN,
                        CALCIUM, URICACID.
          CORRELATIONS.
-----------------------------------
```

we obtain the results shown in Output 3D.5.

Output 3D.5 Correlation matrices for each group

```
------------------------------------------------------------

TEST TITLE. . . . . . . . .WERNER BLOOD CHEMISTRY DATA
INDEXES OF VARIABLES TO BE ANALYZED . . . . . .   6  7  8  9
USE COMPLETE CASES ONLY . . . . . . . . . . . .     NO
PRINT GROUP CORRELATION MATRICES. . . . . . . .    YES
COMPUTE HOTELLINGS T SQUARE . . . . . . . . . .     NO
INDEX OF GROUPING VARIABLE. . . . . . . . . . .      5


GROUPS USED IN COMPUTATIONS . . . . . . . . . . . .  1  2


CORRELATION MATRIX FOR GROUP   1 NOPILL

                  CHOLSTRL   ALBUMIN    CALCIUM    URICACID
                         6         7          8           9

 CHOLSTRL    6     1.0000
 ALBUMIN     7     0.0296    1.0000
 CALCIUM     8     0.2874    0.4452     1.0000
 URICACID    9     0.2739    0.0858     0.2009      1.0000


CORRELATION MATRIX FOR GROUP   2 PILL

                  CHOLSTRL   ALBUMIN    CALCIUM    URICACID
                         6         7          8           9

 CHOLSTRL    6     1.0000
 ALBUMIN     7     0.1160    1.0000
 CALCIUM     8     0.2153    0.4258     1.0000
 URICACID    9     0.2473   -0.0485     0.1916      1.0000

--- analyses of variables 6 to 9 as in Output 3D.1 ---

------------------------------------------------------------
```

```
/ TEST        VARIABLES ARE CHOLSTRL, ALBUMIN,
                            CALCIUM, URICACID.
              GROUPS ARE 1, 4.
/ END
          ----------------------------------
```

These instructions are similar to Example 1D.1 except we now use AGE as the GROUPING variable; AGE is used to classify the cases into four groups. The results are presented in Output 3D.2.

Two TEST paragraphs are used. The first paragraph specifies that only the variable CHOLSTRL is to be analyzed and that three groups (with subscripts 1, 2 and 3) are to be compared. These groups are named '25ORLESS', '26 TO 35' and '36 TO 45'. There are three possible pairings of three groups, therefore three analyses of CHOLSTRL are computed, each using a different pair of groups.

The second TEST paragraph specifies that two groups (with subscripts 1 and 4) are to be compared using the data from the four blood chemistry measurements. This analysis follows the interpretation of the second TEST paragraph. We omit the results for CALCIUM and URICACID.

A title can be specified in each TEST paragraph to label the analysis.

```
TEST ------------------------------------------------
  VARiable = v list.        all variables except the
                            GROUPING variable/prev.
    Names or subscripts of the VARIABLES to be
    analyzed. When USE is stated in the VARIABLE
    paragraph, VARIABLES in the TEST paragraph must
    be included in the USE list.
  GROUP = g list.                   all groups/prev.
    The groups to be compared. Lists are the GROUP
    NAMES or group subscripts. A group subscript is
    the sequence number of the group in the list of
    CODES or CUTPOINTS specified in the GROUP
    paragraph, or, if not specified in a GROUP
    paragraph, the rank order of the group. If more
    than two GROUPS are specified, each possible
    pair of groups is compared.
  TITLE = 'c'.       ≤ 80 char.                blank
    A title for the analysis.
```

Output 3D.2 A subset of groups and variables are selected for analysis by P3D

```
TEST TITLE. . . . . . . . . .WERNER BLOOD CHEMISTRY DATA
INDEXES OF VARIABLES TO BE ANALYZED . . . . . . .   6
USE COMPLETE CASES ONLY . . . . . . . . . . . . .  NO
PRINT GROUP CORRELATION MATRICES. . . . . . . . .  NO
COMPUTE HOTELLINGS T SQUARE . . . . . . . . . . .  NO
INDEX OF GROUPING VARIABLE. . . . . . . . . . . .   2

GROUPS USED IN COMPUTATIONS . . . . . . . . . . .   1   2   3


DIFFERENCES ON SINGLE VARIABLES

************
* CHOLSTRL *  VARIABLE NUMBER    6      GROUP     1 25ORLESS  2 26 TO 35      1 25ORLESS(N=  50)         2 26 TO 35(N=  60)
************                            MEAN        222.1198    224.6331
         STATISTICS     P VALUE   DF    STD DEV      37.4441     35.7419         M                          X
                                        S.E.M.        5.2954      4.6143         M                          X    X
T (SEPARATE)    -0.36   0.721   102.6   SAMPLE SIZE       50          60         M  M  M                  X X   XXXX
T (POOLED)      -0.36   0.720   108     MAXIMUM     330.0000    317.0000      MM MMMMMMMM                  XXXXXXXXXX
                                        MINIMUM     155.0000    160.0000     MMM MMMMMMMMMMMM    M        XXXXXXXXXXXXXX XXX
F(FOR VARIANCES)                                                           MIN---------------------MAX   MIN-------------------MAX
  LEVENE          0.01   0.912    1, 108                                     AN M =     2 CASES          AN X =      2 CASES

DIFFERENCES ON SINGLE VARIABLES

************
* CHOLSTRL *  VARIABLE NUMBER    6      GROUP     1 25ORLESS  3 36 TO 45      1 25ORLESS(N=  50)         3 36 TO 45(N=  42)
************                            MEAN        222.1198    248.3332
         STATISTICS     P VALUE   DF    STD DEV      37.4441     44.8088         M
                                        S.E.M.        5.2954      6.9141         M
T (SEPARATE)    -3.01   0.003    80.1   SAMPLE SIZE       50          42         M  M  M                       X
T (POOLED)      -3.06   0.003    90     MAXIMUM     330.0000    335.0000      MM MMMMMMMM                  X X XXXXXX  X
                                        MINIMUM     155.0000    160.0000     MMM MMMMMMMMMMMM    M        XXX X X XXXXXXX XXXX
F(FOR VARIANCES)                                                           MIN---------------------MAX   MIN-------------------MAX
  LEVENE          1.87   0.174    1,  90                                     AN M =     2 CASES          AN X =      2 CASES

DIFFERENCES ON SINGLE VARIABLES

************
* CHOLSTRL *  VARIABLE NUMBER    6      GROUP     2 26 TO 35  3 36 TO 45      2 26 TO 35(N=  60)         3 36 TO 45(N=  42)
************                            MEAN        224.6331    248.3332
         STATISTICS     P VALUE   DF    STD DEV      35.7419     44.8088         M
                                        S.E.M.        4.6143      6.9141         M    M
T (SEPARATE)    -2.85   0.006    75.3   SAMPLE SIZE       60          42       M M   MMMM                      X
T (POOLED)      -2.97   0.004   100     MAXIMUM     317.0000    335.0000       MMMMMMMMMM                  X X XXXXXX  X
                                        MINIMUM     160.0000    160.0000     MMMMMMMMMMMMMM MMM           XXX X X XXXXXXX XXXX
F(FOR VARIANCES)                                                           MIN---------------------MAX   MIN-------------------MAX
  LEVENE          2.60   0.110    1, 100                                     AN M =     2 CASES          AN X =      2 CASES
```

(output continued)

Example 3D.7 The One-Sample (or Matched Pairs) t Test

The Werner data (Table 5.1) consist of 94 pairs of age-matched women. In each pair the first woman is not on the pill and the second woman is. We perform a paired t test by reading each pair of data records as a single case. We then use BMDP transformations to form four new variables that represent differences in the blood measurement variables. The Control Language rules for the TRANSFORM paragraph are described in Chapter 6. Note that we state that we ADD four variables in the VARIABLE paragraph. We request a one-sample t test for each new variable (those representing differences) by not specifying a GROUPING variable. The Control Language is as follows

```
-----------------------------------
/ PROBLEM   TITLE IS 'WERNER BLOOD CHEMISTRY DATA'.
/ INPUT     VARIABLES ARE 13.
            FORMAT IS '(A4, 5F4.0, 3F4.1/ 20X,
                       F4.0, 3F4.1)'.
/ VARIABLE  NAMES ARE ID, AGE, HEIGHT, WEIGHT,
                      BRTHPILL, CHOL1, ALB1, CAL1,
                      URIC1, CHOL2, ALB2, CAL2,
                      URIC2, CHOLDIFF, ALBDIFF,
                      CALDIFF, URICDIFF.
            MAXIMUMS ARE (6)400, (10)400.
            MINIMUMS ARE (6)150, (10)150.
            LABEL IS ID.
            ADD IS 4.

/ TRANSFORM CHOLDIFF = CHOL1 - CHOL2.
            ALBDIFF  = ALB1  - ALB2.
            CALDIFF  = CAL1  - CAL2.
            URICDIFF = URIC1 - URIC2.

/ TEST      VARIABLES ARE 14 TO 17.
            HOTELLING.

/ END
           -----------------------------------
```

Output 3D.7 shows the results for CHOLDIFF and ALBDIFF; the results for ALBDIFF (the difference in

Output 3D.7 Paired t test by P3D

```
-----------------------------------------------------------------------------------------

          MAHALANOBIS D SQUARE             0.1303
          HOTELLING T SQUARE              11.8567
          F VALUE                          2.8654       P VALUE    0.028
            DEGREES OF FREEDOM       4,    87.0


        WARNING - SINCE SPECIAL MISSING VALUE FORMULAS ARE USED,
                  THESE MULTIVARIATE STATISTICS ARE ONLY APPROXIMATE.


        DIFFERENCES ON SINGLE VARIABLES


        ************
        * CHOLDIFF *  VARIABLE NUMBER  14
        ************
                                       MEAN        -6.1848
        T STATISTIC  P VALUE  DF STD DEV       59.5390              H
                                       S.E.M.       6.2074           H  H
            -1.00      0.322  91 SAMPLE SIZE        92           HHHHH H H
                                       MAXIMUM    155.0000         HHHHHHHH H
                                       MINIMUM   -145.0000     HHHHHHHHHHHHHHHHH   H
                                                             MIN------------------MAX
                                                                AN H =    3 CASES


        ************
        * ALBDIFF  *  VARIABLE NUMBER  15
        ************
                                       MEAN         0.1804               H
        T STATISTIC  P VALUE  DF STD DEV        0.5315               HH
                                       S.E.M.       0.0554           H  HHH  H
             3.26     0.002   91 SAMPLE SIZE        92            HHHHHHH HH
                                       MAXIMUM      1.3000        HHHHHHHHHHHHHH H
                                       MINIMUM     -1.2000     HHHHHHHHHHHHHHHHHH H
                                                             MIN------------------MAX
                                                                AN H =    2 CASES

                  --- similar analyses for CALDIFF and URICDIFF ---
-----------------------------------------------------------------------------------------
```

(4

Output 3D.3 Hotelling's T^2 and Mahalanobis D^2 - using all acceptable values

```
TEST TITLE. . . . . . . . .WERNER BLOOD CHEMISTRY DATA
INDEXES OF VARIABLES TO BE ANALYZED . . . . . .   6   7   8   9
USE COMPLETE CASES ONLY . . . . . . . . . . . .     NO
PRINT GROUP CORRELATION MATRICES. . . . . . . .     NO
COMPUTE HOTELLINGS T SQUARE . . . . . . . . . .    YES
INDEX OF GROUPING VARIABLE. . . . . . . . . . .      5

GROUPS USED IN COMPUTATIONS . . . . . . . . . . . . .   1   2

DIFFERENCES AMONG GROUP MEANS USING ALL VARIABLES
FOR THE FOLLOWING GROUPS
************
* NOPILL   *
* PILL     *
************

 MAHALANOBIS D SQUARE              0.2819
 HOTELLING T SQUARE               13.0364
 F VALUE                           3.2057      P VALUE     0.014
   DEGREES OF FREEDOM        4,   180.0

WARNING - SINCE SPECIAL MISSING VALUE FORMULAS ARE USED,
          THESE MULTIVARIATE STATISTICS ARE ONLY APPROXIMATE.

DIFFERENCES ON SINGLE VARIABLES


************
* CHOLSTRL *  VARIABLE NUMBER   6       GROUP     1 NOPILL     2 PILL            1 NOPILL   (N=  94)                  2 PILL    (N=  92)
************                            MEAN       232.9678    239.4019
        STATISTICS      P VALUE   DF    STD DEV     43.4914     41.5620                  H
                                        S.E.M.       4.4858      4.3331            H     H                               X XXXX
T (SEPARATE)    -1.03  0.304  183.9     SAMPLE SIZE      94          92          H HHHHHHH                               X XXXX
T (POOLED)      -1.03  0.304  184       MAXIMUM    335.0000    390.0000         HHHHHHHHHH   H                         XXXXXXXXXX
                                        MINIMUM    155.0000    160.0000         HHHHHHHHHHHHH HH                      XXXXXXXXXXXXXXXX   X
F(FOR VARIANCES)                                                             MIN--------------------MAX          MIN----------------------MAX
   LEVENE        1.49  0.223    1, 184                                         AN H =      3 CASES                  AN X =       3 CASES

                         --- similar analyses for variables 7 to 9 ---
```

Output 3D.4 Hotelling's T^2 and Mahalanobis D^2 - using complete cases only

```
TEST TITLE. . . . . . . . .WERNER BLOOD CHEMISTRY DATA
INDEXES OF VARIABLES TO BE ANALYZED . . . . . .   6   7   8   9
USE COMPLETE CASES ONLY . . . . . . . . . . . .    YES
PRINT GROUP CORRELATION MATRICES. . . . . . . .     NO
COMPUTE HOTELLINGS T SQUARE . . . . . . . . . .    YES
INDEX OF GROUPING VARIABLE. . . . . . . . . . .      5

GROUPS USED IN COMPUTATIONS . . . . . . . . . . . . .   1   2

DIFFERENCES AMONG GROUP MEANS USING ALL VARIABLES
FOR THE FOLLOWING GROUPS
************
* NOPILL   *
* PILL     *
************

 MAHALANOBIS D SQUARE              0.2864
 HOTELLING T SQUARE               13.0284
 F VALUE                           3.2028      P VALUE     0.014
   DEGREES OF FREEDOM        4,   177.0

DIFFERENCES ON SINGLE VARIABLES


************
* CHOLSTRL *  VARIABLE NUMBER   6       GROUP     1 NOPILL     2 PILL            1 NOPILL   (N=  90)                  2 PILL    (N=  92)
************                            MEAN       232.0886    239.4019
        STATISTICS      P VALUE   DF    STD DEV     43.4700     41.5620                  H
                                        S.E.M.       4.5821      4.3331            H     H                               X XXXX
T (SEPARATE)    -1.16  0.248  179.2     SAMPLE SIZE      90          92          H HHH HHH                               X XXXX
T (POOLED)      -1.16  0.247  180       MAXIMUM    335.0000    390.0000         HHHHHHHHHH   H                         XXXXXXXXXX
                                        MINIMUM    155.0000    160.0000         HHHHHHHHHHHHH HH                      XXXXXXXXXXXXXXXX   X
F(FOR VARIANCES)                                                             MIN--------------------MAX          MIN----------------------MAX
   LEVENE        1.79  0.182    1, 180                                         AN H =      3 CASES                  AN X =       3 CASES

                         --- similar analyses for variables 7 to 9 ---
```

(5) Correlation matrix.

(6) Squared multiple correlation (SMC) of each variable with all other variables. The condition number is the ratio of the largest eigenvalue to the smallest eigenvalue and is of interest to see how nearly singular the correlation matrix might be. The condition number 0.4920D02 is read as 49.2.

(7), (8) The eigenvalues of the factors in (8) are all listed (under the heading "Variance Explained"). The preassigned criterion for the number of factors is the number of factors with eigenvalues gre than one (see third line of (1)). Therefore, in communalities are obtained for three factors (t with eigenvalues greater than one). The communa of a variable is its squared multiple correla with the factors extracted.

The cumulative proportion of total variance (8) is the sum of the variance explai (eigenvalues) up to and including the factor divi by the sum of all the eigenvalues. A success factor analysis explains a large proportion variance with a very few factors.

Output 4M.1 (continued)

```
     CORRELATION MATRIX     (5)
     ------------------

                   CONCENTR ANNOY   SMOKING1 SLEEPY   SMOKING2 TENSE    SMOKING3 ALERT    IRRITABL TIRED    CONTENT  SMOKING4
                      1        2       3        4        5        6        7        8        9       10       11       12

     CONCENTR  1    1.000
     ANNOY     2    0.562    1.000
     SMOKING1  3    0.086    0.144   1.000
     SLEEPY    4   *0.457    0.360   0.140    1.000
     SMOKING2  5    0.200    0.119   0.785    0.211    1.000
     TENSE     6    0.579    0.705   0.222    0.273    0.301    1.000
     SMOKING3  7    0.041    0.060   0.810    0.126    0.816    0.120    1.000
     ALERT     8    0.802    0.578   0.101    0.606    0.223    0.594    0.039    1.000
     IRRITABL  9    0.595    0.796   0.189    0.337    0.221    0.725    0.108    0.605    1.000
     TIRED    10    0.512    0.413   0.199    0.798    0.274    0.364    0.139    0.698    0.428    1.000
     CONTENT  11    0.492    0.739   0.239    0.240    0.235    0.711    0.100    0.605    0.697    0.394    1.000
     SMOKING4 12    0.228    0.122   0.775    0.277    0.813    0.214    C.845    0.201    0.156    0.271    0.171    1.000
```

```
    SQUARED MULTIPLE CORRELATIONS (SMC) OF
    EACH VARIABLE WITH ALL OTHER VARIABLES
    --------------------------------------

      1 CONCENTR    0.70351        (6)
      2 ANNOY       0.74250
      3 SMOKING1    0.73312
      4 SLEEPY      0.68377
      5 SMOKING2    0.78201
      6 TENSE       0.66472
      7 SMOKING3    0.82062
      8 ALERT       0.80208
      9 IRRITABL    0.71437
     10 TIRED       0.72627
     11 CONTENT     0.69130
     12 SMOKING4    0.80294

    CONDITION NUMBER = 0.492195D 02
    ----------------
```

```
    COMMUNALITIES OBTAINED FROM  3 FACTORS AFTER    1 ITERATIONS.
    ---------------------------------------------------------------

    THE COMMUNALITY OF A VARIABLE IS ITS SQUARED MULTIPLE
    CORRELATION (OR COVARIANCE) WITH THE FACTORS.

       1 CONCENTR    0.6607
       2 ANNOY       0.7956        (7)
       3 SMOKING1    0.8391
       4 SLEEPY      0.8474
       5 SMOKING2    0.8561
       6 TENSE       0.7804
       7 SMOKING3    0.8941
       8 ALERT       0.8258
       9 IRRITABL    0.7978
      10 TIRED       0.8453
      11 CONTENT     0.7715
      12 SMOKING4    0.8698
```

```
    FACTOR    VARIANCE EXPLAINED   CUMULATIVE PROPORTION OF TOTAL VARIANCE
    ------------------------------------------------------------------------
      1           5.425688              0.452141
      2           2.996636              0.701860          (8)
      3           1.360520              0.815237
      4           0.560300              0.861929
      5           0.363261              0.892200
      6           0.302254              0.917388
      7           0.240804              0.937455
      8           0.199752              0.954101
      9           0.158162              0.967281
     10           0.145653              0.979419
     11           0.136736              0.990814
     12           0.110235              1.000000

    THE VARIANCE EXPLAINED BY EACH FACTOR IS THE EIGENVALUE FOR THAT FACTOR.

    TOTAL VARIANCE IS DEFINED AS THE SUM OF THE DIAGONAL ELEMENTS OF THE
    CORRELATION (COVARIANCE) MATRIX.
```

Factor Analysis

(6)

(9) Unrotated factor loadings (pattern) for principal components. These loadings are the eigenvectors of the correlation matrix multiplied by the square roots of the corresponding eigenvalues. They are the correlations of the principal components with the original variables. The eigenvalues (VP) are printed at the bottom of each column.

(10) Orthogonal rotation is performed. Gamma is preassigned to 1 because varimax rotation is performed. At each iteration the simplicity criterion G (p. 488) is printed.

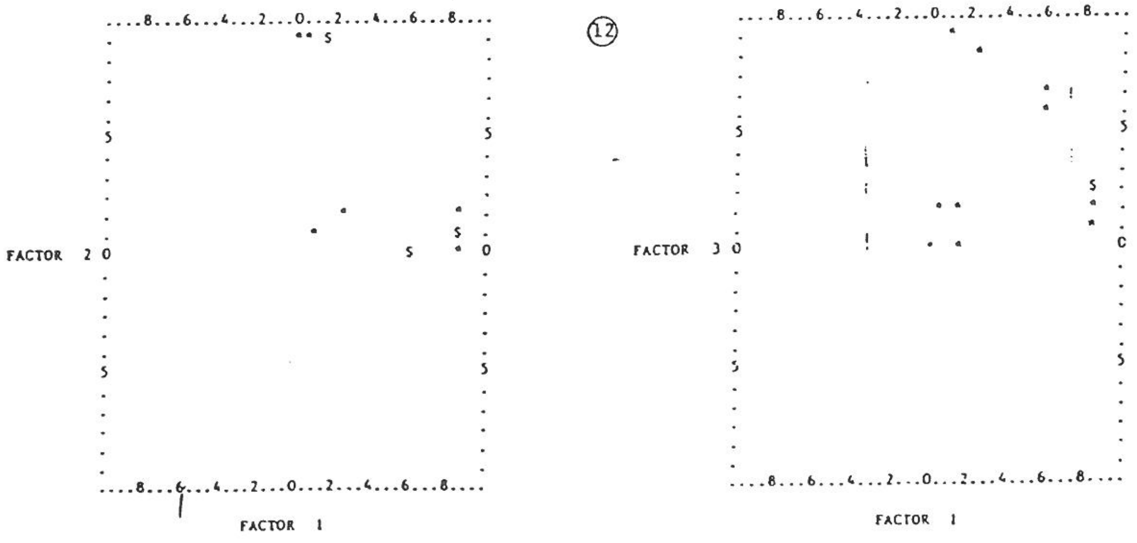
(11) Rotated factor loadings (pattern) -- coefficients of the factors after rotation. The sum of squares of the coefficients are printed below each column (VP). When the rotation is orthogonal, as in this example, VP is the variance explained by the factor and the rotated loadings are the correlations of the variables with the factors.

(12) Plots of the rotated factor loadings. The loadings for one factor are plotted against those of another factor.

```
UNROTATED FACTOR LOADINGS (PATTERN)
-----------------------------------
FOR PRINCIPAL COMPONENTS

                     FACTOR      FACTOR      FACTOR        (9)
                        1           2           3

CONCENTR     1       0.742      -0.309       0.117
ANNOY        2       0.755      -0.361      -0.309
SMOKING1     3       0.491       0.763      -0.124
SLEEPY       4       0.611      -0.117       0.679
SMOKING2     5       0.561       0.735      -0.030
TENSE        6       0.770      -0.232      -0.366
SMOKING3     7       0.417       0.847      -0.055
ALERT        8       0.808      -0.337       0.244
IRRITABL     9       0.783      -0.302      -0.306
TIRED       10       0.702      -0.138       0.577
CONTENT     11       0.748      -0.256      -0.382
SMOKING4    12       0.540       0.757       0.070
            VP       5.426       2.997       1.361

THE VP FOR EACH FACTOR IS THE SUM OF THE SQUARES OF THE
ELEMENTS OF THE COLUMN OF THE FACTOR LOADING MATRIX
CORRESPONDING TO THAT FACTOR.  THE VP IS THE VARIANCE
EXPLAINED BY THE FACTOR.

ORTHOGONAL ROTATION, GAMMA =        1.0000

ITERATION    SIMPLICITY
             CRITERION
    0        -1.900373      (10)
    1        -6.017688
    2        -6.019553
    3        -6.019557
```

```
ROTATED FACTOR LOADINGS (PATTERN)
---------------------------------

                     FACTOR      FACTOR      FACTOR        (11)
                        1           2           3

CONCENTR     1       0.601       0.034       0.546
ANNOY        2       0.867       0.021       0.209
SMOKING1     3       0.131       0.907       0.007
SLEEPY       4       0.117       0.116       0.906
SMOKING2     5       0.141       0.905       0.128
TENSE        6       0.859       0.147       0.144
SMOKING3     7       0.005       0.945       0.010
ALERT        8       0.590       0.030       0.691
IRRITABL     9       0.863       0.085       0.214
TIRED       10       0.249       0.143       0.873
CONTENT     11       0.862       0.117       0.125
SMOKING4    12       0.061       0.910       0.195

            VP       3.802       3.443       2.538

THE VP FOR EACH FACTOR IS THE SUM OF THE SQUARES OF THE
ELEMENTS OF THE COLUMN OF THE FACTOR PATTERN MATRIX
CORRESPONDING TO THAT FACTOR.  WHEN THE ROTATION IS
ORTHOGONAL, THE VP IS THE VARIANCE EXPLAINED BY THE FACTOR.
```

ROTATED FACTOR LOADINGS



OVERLAP IS INDICATED BY A DOLLAR SIGN. SCALE IS FROM -1 TO +1.

--- we omit the plot of factor 3 versus factor 2 ---

(output continued)

```
ITERATION FOR MAXIMUM LIKELIHOOD          21
--------------------------------

ITERATION   MAXIMUM CHANGE IN SQRT(UNIQUENESS)    LIKELIHOOD CRITERION TO BE MINIMIZED    STEP HALVINGS

                                                          0.969191
    1               0.120209                              0.900708                           2
    2               0.061833                              0.836862                           0
    3               0.012304                              0.834094                           0
    4               0.000609                              0.834089                           0
    5               0.000002

AN ASTERISK (IF ANY) AFTER THE ITERATION NUMBER INDICATES
THAT APPROXIMATE DERIVATIVES WERE USED.


CANONICAL CORRELATIONS
----------------------

              0.9790
              0.9668                22
              0.9082


COMMUNALITIES OBTAINED FROM  3 FACTORS AFTER    5 ITERATIONS.
-------------------------------------------------------------

THE COMMUNALITY OF A VARIABLE IS ITS SQUARED MULTIPLE
CORRELATION (OR COVARIANCE) WITH THE FACTORS.

   1 CONCENTR    0.5753
   2 ANNOY       0.7457
   3 SMOKING1    0.7630
   4 SLEEPY      0.7596
   5 SMOKING2    0.8011
   6 TENSE       0.7110
   7 SMOKING3    0.8784
   8 ALERT       0.7561
   9 IRRITABL    0.7551
  10 TIRED       0.8043
  11 CONTENT     0.6993
  12 SMOKING4    0.8352


FACTOR   VARIANCE EXPLAINED   CUMULATIVE PROPORTION OF TOTAL VARIANCE
---------------------------------------------------------------------
   1         4.793273               0.399439
   2         3.162091               0.662947
   3         1.128832               0.757016

TOTAL VARIANCE IS DEFINED AS THE SUM OF THE DIAGONAL ELEMENTS OF THE
CORRELAT :N (COVARIANCE) MATRIX.
```

```
UNROTATED FACTOR LOADINGS (PATTERN)
-----------------------------------

FOR MAXIMUM LIKELIHOOD CANONICAL FACTORS

                     FACTOR     FACTOR     FACTOR
                        1          2          3

CONCENTR    1        0.529      0.543      0.027
ANNOY       2        0.527      0.599     -0.329
SMOKING1    3        0.732     -0.466     -0.097
SLEEPY      4        0.518      0.389      0.583
SMOKING2    5        0.790     -0.419     -0.023
TENSE       6        0.579      0.490     -0.369
SMOKING3    7        0.722     -0.596     -0.040
ALERT       8        0.587      0.622      0.157
IRRITABL    9        0.574      0.562     -0.332
TIRED      10        0.590      0.448      0.506
CONTENT    11        0.552      0.507     -0.370
SMOKING4   12        0.789     -0.456      0.066

           VP        4.793      3.162      1.129

THE VP FOR EACH FACTOR IS THE SUM OF THE SQUARES OF THE
ELEMENTS OF THE COLUMN OF THE FACTOR LOADING MATRIX
CORRESPONDING TO THAT FACTOR.  THE VP IS THE VARIANCE
EXPLAINED BY THE FACTOR.

ORTHOGONAL ROTATION, GAMMA =     1.0000

ITERATION   SIMPLICITY
            CRITERION
    0       -0.611441
    1       -5.852844
    2       -5.864646
    3       -5.864750
    4       -5.864750


ROTATED FACTOR LOADINGS (PATTERN)
---------------------------------

                     FACTOR     FACTOR     FACTOR
                        1          2          3

CONCENTR    1        0.595      0.051      0.468
ANNOY       2        0.839      0.030      0.204
SMOKING1    3        0.128      0.864      0.023
SLEEPY      4        0.164      0.116      0.848
SMOKING2    5        0.144      0.874      0.127
TENSE       6        0.818      0.142      0.146
SMOKING3    7        0.007      0.937      0.011
ALERT       8        0.597      0.039      0.631
IRRITABL    9        0.840      0.090      0.204
TIRED      10        0.283      0.137      0.840
CONTENT    11        0.817      0.111      0.142
SMOKING4   12        0.068      0.893      0.183

           VP        3.604      3.264      2.216

THE VP FOR EACH FACTOR IS THE SUM OF THE SQUARES OF THE
ELEMENTS OF THE COLUMN OF THE FACTOR PATTERN MATRIX
CORRESPONDING TO THAT FACTOR.  WHEN THE ROTATION IS
ORTHOGONAL, THE VP IS THE VARIANCE EXPLAINED BY THE FACTOR.
```

--- the remainder of the results is analogous to 12 to 18 in Output 4M.1 ---

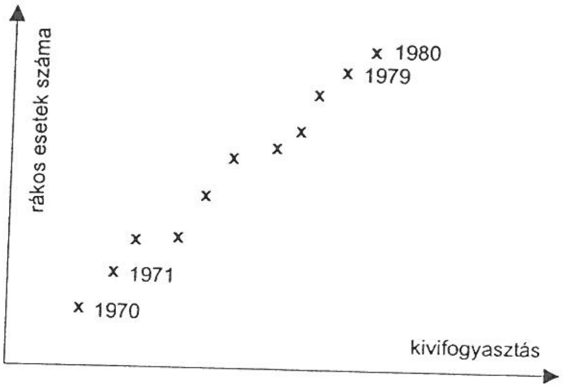
Output 1R.1 Multiple linear regression. Circled numbers correspond to those in the text

```
        --- the BMDP instructions read by P1R are printed and interpreted ---


    REGRESSION INTERCEPT. . . . . . . . . . . . . . .NON-ZERO
    GROUPING VARIABLE . . . . . . . . . . . . . . .
    WEIGHT VARIABLE . . . . . . . . . . . . . . . .
    PRINT COVARIANCE MATRIX . . . . . . . . . . . .      NO
    PRINT CORRELATION MATRIX. . . . . . . . . . . .      NO
    PRINT CORRELATION OF REGRESSION COEFFICIENTS. .      NO   ①
    PRINT RESIDUALS . . . . . . . . . . . . . . . .      NO
    PRINT NORMAL PROBABILITY PLOT . . . . . . . . .      NO
    PRINT DETRENDED NORMAL PROBABILITY PLOT . . . .      NO

    NUMBER OF CASES READ. . . . . . . . . . . . . .     188
       CASES WITH DATA MISSING OR BEYOND LIMITS . .       8
          REMAINING NUMBER OF CASES . . . . . . . .     180



  VARIABLE     ②        MEAN  STANDARD DEVIATION  COEFFICIENT      MINIMUM        MAXIMUM
                                                  OF VARIATION
    2 AGE          33.53819          9.89836         0.29514      19.00000       55.00000
    3 HEIGHT       64.46597          2.48213         0.03850      57.00000       71.00000
    4 WEIGHT      131.09384         20.49977         0.15637      94.00000      215.00000
    5 BRTHPILL      1.50551          0.50136         0.33302       1.00000        2.00000
    6 CHOLSTRL    235.83821         42.74364         0.18124     155.00000      390.00000
    7 ALBUMIN       4.12052          0.35871         0.08706       3.20000        5.00000
    8 CALCIUM       9.96773          0.47279         0.04743       8.80000       11.10000
    9 URICACID      4.75551          1.12111         0.23575       2.20000        9.90000



   REGRESSION TITLE. . . . . . . . . . . . . . . . .WERNER BLOOD CHEMISTRY DATA
   DEPENDENT VARIABLE. . . . . . . . . . . . . . .      6 CHOLSTRL
   TOLERANCE . . . . . . . . . . . . . . . . . . .  0.0100

  ALL DATA CONSIDERED AS A SINGLE GROUP

  MULTIPLE R         ③  0.4175            STD. ERROR OF EST.          39.1698
  MULTIPLE R-SQUARE      0.1743



  ANALYSIS OF VARIANCE
                         SUM OF SQUARES      DF      MEAN SQUARE      F RATIO     P(TAIL)
   ④      REGRESSION       57004.242        3       19001.414        12.385      0.0
          RESIDUAL        270032.000      176        1534.273


                                                       STD. REG
    VARIABLE          COEFFICIENT    STD. ERROR          COEFF          T    P(2 TAIL) TOLERANCE

  INTERCEPT      ⑤     151.42036
  AGE           2        1.38971        0.309           0.322      4.497      0.000      0.915924
  WEIGHT        4        0.00289        0.153           0.001      0.019      0.985      0.869294
  URICACID      9        7.87099        2.769           0.206      2.843      0.005      0.889443
```

magában. A felismert összefüggés látszólagos lehet, ha az analízis mögött nem állnak elméleti megfontolások.

Ok és okozat

Saville és Wood könyvéből [1] vettük az alábbi példát. A *4.1. ábra* az Egyesült Államokban megfigyelt rákos esetek számát mutatja a kivifogyasztás függvényében. Mivel 1970 és 1980 között mindkét mennyiség növekedett, ezek évente megfigyelt értékei *korreláltak*. Jóllehet ez matematikai bizonyosság, mégsem állíthatjuk, hogy a rákos esetek számának a növekedését az *okozta*, hogy az emberek több kivit ettek. A ténylegesen talált (és statisztikailag bizonyított) korrelációt csak akkor szabad *ok–okozati* kapcsolatnak tekinteni, ha erre *elméleti indok* van.



4.1. ábra. Kapcsolat az Egyesült Államokban megfigyelt rákos esetek száma és a kivifogyasztás között

Hasonló példákat lehet az élet legkülönbözőbb területén találni. Például határozottan pozitív korreláció van a Duna vízállása és a BME területén tartózkodó hallgatók száma között. Nyilván épeszű ember nem tételez fel ezek között ok-okozati kapcsolatot. A matematikai statisztika, vagy inkább az azt rosszul alkalmazó áltudomány iránt bizalmatlan emberek gyakran köszörülik szellemességüket az ilyen korrelációkon. Akkor mire vezethetők vissza ezek a látszólagos összefüggések? A válasz egyszerű. Az ilyen példákban általában lehet találni egy közvetítő mennyiséget, ami legtöbbször az idő. Mikor magas ugyanis a Duna vízszintje? Koratavasszal és késő ősszel. Éppen ezek az időszakok előzik meg a vizsgaidőszakokat, amikor a hallgatók a legszorgalmasabban járnak az egyetemre. Hasonlóan az idő a közvetítő a *4.1. ábrán* mutatott példában is.

Az extrapoláció veszélyei

Nem csak a lineáris regresszióban, hanem – általánosabban – a polinomillesztésben (vö. 4.2. alfejezet) is nagyon veszélyes az illesztésben kapott függvényt a vizsgált valószínűségi változók mérési tartományán túl extrapolálni. Súlyos tévedések forrása az ilyesmi. A probléma hangsúlyozottan főleg a polinomillesztésnél merül fel, ugyanis többnyire akkor fordulunk eh-

ANOVA táblázatok

Egyszempontos varianciaanalízis

A szóródás oka	Négyzetösszeg	Szabadsági	Empírikus
		fok	szórásnégyzet
Csoportok között	$Q_1 = \sum_{i=1}^{k} n_i(\bar{X}_{i.} - \bar{X}_{..})^2$	$k - 1$	$s_1^2 = \frac{Q_1}{k-1}$
Csoportokon belül	$Q_2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2$	$n - k$	$s_2^2 = \frac{Q_2}{n-k}$
Teljes	$Q = \sum_{i=1}^{k} \sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{..})^2$	$n - 1$	-

Kétszempontos varianciaanalízis (interakció nélkül)

A szóródás oka	Négyzetösszeg	Szabadsági	Empírikus
		fok	szórásnégyzet
a-hatások	$Q_1 = p \sum_{i=1}^{k}(\bar{X}_{i.} - \bar{X}_{..})^2$	$k - 1$	$s_1^2 = \frac{Q_1}{k-1}$
b-hatások	$Q_2 = k \sum_{j=1}^{p}(\bar{X}_{.j} - \bar{X}_{..})^2$	$p - 1$	$s_2^2 = \frac{Q_2}{p-1}$
Véletlen hiba	$Q_3 = \sum_{i=1}^{k} \sum_{j=1}^{p}(X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2$	$(k - 1)(p - 1)$	$s_3^2 = \frac{Q_3}{(k-1)(p-1)}$
Teljes	$Q = \sum_{i=1}^{k} \sum_{j=1}^{p}(X_{ij} - \bar{X}_{..})^2$	$kp - 1$	-

Kétszempontos varianciaanalízis (interakcióval)

A szóródás oka	Négyzetösszeg	Szabadsági	Empírikus
		fok	szórásnégyzet
a-hatások	$Q_1 = pn \sum_{i=1}^{k}(\bar{X}_{i..} - \bar{X}_{...})^2$	$k - 1$	$s_1^2 = \frac{Q_1}{k-1}$
b-hatások	$Q_2 = kn \sum_{j=1}^{p}(\bar{X}_{.j.} - \bar{X}_{...})^2$	$p - 1$	$s_2^2 = \frac{Q_2}{p-1}$
ab-interakció	$Q_3 = n \sum_{i=1}^{k} \sum_{j=1}^{p}(\bar{X}_{ij.} - \bar{X}_{i..} - \bar{X}_{.j.} + \bar{X}_{...})^2$	$(k - 1)(p - 1)$	$s_3^2 = \frac{Q_2}{(k-1)(p-1)}$
Véletlen hiba	$Q_4 = \sum_{i=1}^{k} \sum_{j=1}^{p} \sum_{l=1}^{n}(X_{ijl} - \bar{X}_{ij.})^2$	$kp(n - 1)$	$s_4^2 = \frac{Q_3}{kp(n-1)}$
Teljes	$Q = \sum_{i=1}^{k} \sum_{j=1}^{p} \sum_{l=1}^{n}(X_{ijl} - \bar{X}_{...})^2$	$kpn - 1$	-

Typeset by $\mathcal{A}\!\mathcal{M}\!\mathcal{S}$-TEX

not have an acceptable value. Two values are excluded from the first group in our example.

The midpoint for each interval is printed to the left of the histograms. Each interval includes its upper limit. For example, 210.0 and 225.0 are successive midpoints, so the value 217.5 would be classified into the interval with midpoint 210.0.

② For each group, P7D prints
- mean: $\bar{x}$
- standard deviation: s based on sample variance
- standard error of the mean (S.E.M.) based on sample variance
- robust estimate of standard deviation based on mean deviation from the mean
- maximum and minimum observed value (not out of range)
- sample size (frequency): N

③ For all groups combined, P7D prints the mean, standard deviation, standard error of the mean, maximum, minimum and frequency. The standard deviation is computed from the overall mean for the variable (not from the group means).

④ A one-way analysis of variance (ANOVA) that tests the equality of group means.

Let x_{ij} represent the jth observation in the ith group and $\bar{x}_i$ the mean and N_i the number of

Output 7D.1 Comparison of groups. Circled numbers correspond to those in the text.

```
                     --- the BMDP instructions read by P7D are printed and interpreted ---

               ************                                                              ************
  HISTOGRAM OF * CHOLSTRL * (VARIABLE    6). CASES DIVIDED INTO GROUPS BASED ON VALUES OF * AGE        * (VARIABLE   2)
               ************                                                              ************

           25ORLESS               26 TO 35                36 TO 45                OVER 45
           .......................+.......................+.......................+.......................+
   VAR  6
  EXCLUDED
   VALUES
           **
             TABULATIONS AND COMPUTATIONS WHICH FOLLOW EXCLUDE VALUES LISTED ABOVE
  MIDPOINTS
   435.000)
   420.000)
   405.000)
   390.000)
   375.000)                                                                         *
   360.000)
   345.000)
   330.000)*       ①                                  **                      *
   315.000)                *                          **
   300.000)*               **                         ***                     **
   285.000)**              **                         ***                     ******
   270.000)***             *                          *******
   255.000)****            **********                 M****                   ***
   240.000)********        ***********                *****                   M********
   225.000)M******         M******                    **                      **
   210.000)*******         *****                      *****                   ******
   195.000)*********       ***********                ****                    **
   180.000)****            *******                    **                      **
   165.000)***             ***                        **
   150.000)*
   135.000)
   120.000)
  GROUP MEANS ARE DENOTED BY M'S IF THEY COINCIDE WITH *'S, N'S OTHERWISE

  MEAN        222.120              224.633                248.333                 262.059
  STD.DEV.     37.444               35.742                 44.809                  43.267
  R.E.S.D.     38.044               37.427                 46.613                  42.448
  S. E. M.      5.295     ②          4.614                  6.914                   7.420
  MAXIMUM     330.000              317.000                335.000                 390.000
  MINIMUM     155.000              160.000                160.000                 190.000
  SAMPLE SIZE     50                    60                     42                      34

      ALL GROUPS COMBINED        ***************************** ANALYSIS OF VARIANCE TABLE *******************************
  (EXCEPT CASES WITH UNUSED VALUES  *
     FOR AGE       )             *   SOURCE             SUM OF SQUARES     DF     MEAN SQUARE   F VALUE    TAIL PROBABILITY
                                 *
  MEAN           236.150         *   BETWEEN GROUPS      46857.3398        3     15619.1133  ④  9.86        0.0000
  STD.DEV.        42.556  ③      *   WITHIN  CROUPS     288172.4290      182      1583.3650
  S. E. M.         3.120         *
  MAXIMUM        390.000         *   TOTAL              335029.7500      185
  MINIMUM        155.000         ****************************************************************************************
  SAMPLE SIZE        186         *   LEVENE'S TEST FOR EQUAL VARIANCES    3, 182              ⑤  0.98        0.4054
                                 ****************************************************************************************
                                 *   ONE-WAY ANALYSIS OF VARIANCE
                                 *   TEST STATISTICS FOR WITHIN-GROUP
                                 *   VARIANCES NOT ASSUMED TO BE EQUAL
                                 *     WELCH                          3,  90              ⑥  9.11        0.0000
                                 *     BROWN-FORSYTHE                 3, 151                 9.42        0.0000

  --- the analyses for variables 2 to 5 precede that for CHOLSTRL, those for variables 7 to 9 follow ---
```

pressure as a covariate or perform the latter analysis using the logarithm of blood pressure. Note that when there are only two grouping factors, a more detailed analysis can be obtained with P7D.

In the ANOVA table the main effect of each grouping factor is identified by the name of the grouping variable as specified in the VARIABLE paragraph. The first character of the grouping variable names are used to label interactions; therefore the two grouping variables are given names that begin with different letters. In BMDP instructions only the DESIGN paragraph is specific to P2V. The other BMDP instructions are explained in Chapter 5.

```
------------------------------------
/ PROBLEM   TITLE IS 'KUTNER SYSTOLIC BLOOD PRESSURE
                      DATA'.
/ INPUT     VARIABLES ARE 3.
            FORMAT IS '(3F3.0)'.
/ VARIABLE  NAMES ARE TREATMNT, DISEASE, SYSINCR.
/ DESIGN    DEPENDENT IS SYSINCR.
            GROUPING ARE TREATMNT, DISEASE.

/ GROUP     CODES(1) ARE     1,       2,       3,       4.
            NAMES(1) ARE DRUG1, DRUG2, DRUG3, DRUG4.
            CODES(2) ARE       1,         2,         3.
            NAMES(2) ARE DISEASE1, DISEASE2, DISEASE3.

/ END
------------------------------------
```

(See end of this P2V Section for organization of systems information, BMDP instructions and data)

If the FORM statement is used, the DESIGN paragraph is written

```
/ DESIGN   FORM IS '2G,Y'.
```

2G specifies that the first two variables are grouping factors, and Y specifies that the third variable is the dependent variable.

The results are presented in Output 2V.1. Circled numbers below correspond to those in the output.

(1) The DESIGN paragraph is interpreted by P2V.

(2) Number of cases read. Only cases containing acceptable values for all variables specified in the DESIGN paragraph are used in the analysis. An acceptable value is a value that is not missing or out of range. In addition, if CODES are specified for any GROUPING factors (variables), a case is included only if the value of the GROUPING factor is equal to a specified CODE.

(3) The frequency (COUNT) of observations in each cell is printed.

Output 2V.1 A two-way analysis of variance by P2V. Circled numbers correspond to those in the text

```
--- the BMDP instructions read by P2V are printed and interpreted ---


DESIGN SPECIFICATIONS

      GROUP  -   1   2    (1)
      DEPEND -   3



VARIABLE    MINIMUM  MAXIMUM  MISSING  CATEGORY  CATEGORY   GREATER  LESS THAN
NO. NAME     LIMIT    LIMIT    CODE      CODE      NAME       THAN    OR - TO

  1 TREATMNT
                                        1.00000  DRUG1
                                        2.00000  DRUG2
                                        3.00000  DRUG3
                                        4.00000  DRUG4

  2 DISEASE
                                        1.00000  DISEASE1
                                        2.00000  DISEASE2
                                        3.00000  DISEASE3



NUMBER OF CASES READ. . . . . . . . . . . . . .      58  (2)
```

```
GROUP STRUCTURE        (3)

TREATMNT  DISEASE     COUNT
DRUG1     DISEASE1      6.
DRUG1     DISEASE2      4.
DRUG1     DISEASE3      5.
DRUG2     DISEASE1      5.
DRUG2     DISEASE2      4.
DRUG2     DISEASE3      6.
DRUG3     DISEASE1      3.
DRUG3     DISEASE2      5.
DRUG3     DISEASE3      4.
DRUG4     DISEASE1      5.
DRUG4     DISEASE2      6.
DRUG4     DISEASE3      5.
```

(output continued)

13/6

Output 2V.1 (continued)

```
          CELL MEANS FOR  1-ST DEPENDENT VARIABLE         (4)

TREATMNT=  DRUG1      DRUG1      DRUG1      DRUG2      DRUG2      DRUG2      DRUG3      DRUG3      DRUG3      DRUG4
DISEASE =  DISEASE1   DISEASE2   DISEASE3   DISEASE1   DISEASE2   DISEASE3   DISEASE1   DISEASE2   DISEASE3   DISEASE1

SYSINCR     29.33333   28.25000   20.40000   28.00000   33.50000   18.16667   16.33333    4.40000    8.50000   13.60000

COUNT         6          4          5          5          4          6          3          5          4          5


                                 MARGINAL
TREATMNT=  DRUG4      DRUG4
DISEASE =  DISEASE2   DISEASE3

SYSINCR     12.83333   14.20000   18.87931

COUNT         6          5         58


STANDARD DEVIATIONS FOR  1-ST DEPENDENT VARIABLE

TREATMNT=  DRUG1      DRUG1      DRUG1      DRUG2      DRUG2      DRUG2      DRUG3      DRUG3      DRUG3      DRUG4
DISEASE =  DISEASE1   DISEASE2   DISEASE3   DISEASE1   DISEASE2   DISEASE3   DISEASE1   DISEASE2   DISEASE3   DISEASE1

SYSINCR     13.01794    5.85235   13.37161   10.97725    2.08167   12.52863   14.18920    6.91375    9.00000   10.54988


TREATMNT=  DRUG4      DRUG4
DISEASE =  DISEASE2   DISEASE3

SYSINCR     10.34247    8.92749


ANALYSIS OF VARIANCE FOR  1-ST          (5)
DEPENDENT VARIABLE - SYSINCR

      SOURCE                    SUM OF       DEGREES OF       MEAN           F        TAIL
                                SQUARES      FREEDOM          SQUARE                  PROB.

      MEAN                  20037.61301         1        20037.61301      181.41    0.0000
      TREATMNT               2997.47186         3          999.15729        9.05    0.0001
      DISEASE                 415.87305         2          207.93652        1.88    0.1637
      TD                      707.26626         6          117.87771        1.07    0.3958
1     ERROR                  5080.81667        46          110.45254
--------------------------------------------------------------------------------------------------------
```

(4) The mean, frequency and standard deviation of each cell for each dependent variable are printed.

(5) An ANOVA table is printed.

The sums of squares in the one-way ANOVA are well known. The sums of squares in the two-way, or higher, ANOVA depend upon the hypothesis of interest unless each cell contains the same number of observations. The hypotheses tested by P2V are the same for equal or unequal cell size problems, and are not affected by losing some of the cases. Although the hypotheses tested are independent, the sums of squares for unequal cell size problems are not in general orthogonal. Orthogonal sums of squares methods (or "sequential" methods) test hypotheses that are functions of cell sizes; P2V does not use a sequential method. For more detailed discussions, see Kutner (1974) and Speed and Hocking (1976). (The hypotheses tested by P2V for the main effects are labelled A and B by Kutner and H1 and H2 by Speed and Hocking.) They, as well as others, recommend these hypotheses for experimental data. Searle (1971, pp. 316-317) points out that sequential methods test hypotheses that depend on the cell sizes and cautions against their use. More general hypotheses can be tested in BMDP4V.

Hypotheses tested. In our example of a two-way ANOVA, let $E(Y_{ij}) = \mu_{ij}$ where Y_{ij} is an observation of the group (i, j). The test of equality of row means is the test that

$$\Sigma_j \mu_{ij} = \Sigma_j \mu_{kj} \qquad \text{for all } i, k.$$

The test of equality of column means is the test that

$$\Sigma_i \mu_{ij} = \Sigma_i \mu_{i\ell} \qquad \text{for all } j, \ell.$$

NB.! Az interakció tehát mást jelent a medicinában és mást a biometriában. Amit a biometriában interakciónak nevezünk, azt a medicinában potenciálásnak - esetleg blokkolásnak - hívjuk!

Az egyes Q_K-értékek mutatják az egyes gyógyszerek egyedi hatását, terminus technicus: *főhatás*. Az egyes Q-$_K$ összege és az össz-Q_K közti különbség mutatja az interakciót. Mind az egyes Q_K-értéket, mind az interakciót (Q_i) a Q_B-hez hasonlítjuk. Ha a Q_i/Q_B-ből megfelelő módon számított F-érték szignifikáns, akkor van interakció, van potenciálás. Ha nem szignifikáns, akkor nincs vagy legalábbis nem jelentős. Ha nincs, akkor az interakcióra jutó négyzetes eltéréseket és szabadságfokot bele szoktuk olvasztani a Q_B-be. Ezt azért tesszük, mert így a „hiba" (Q_B) szabadságfokát növelve megbízhatóbbá tehetjük az analízist.

16.3 táblázat

Kétszempontos varianciaanalízis
(a fehérjék minőségének és mennyiségének hatása a patkányok súlygyarapodására)

Nagy fehérjebevitel			Kis fehérjebevitel		
marha	disznó	gabona	marha	disznó	gabona
73	94	98	90	49	107
102	79	74	76	82	95
118	96	56	90	73	97
104	98	111	64	86	80
81	102	95	86	81	98
107	102	88	51	97	74
100	108	82	72	106	74
87	91	77	90	70	67
117	120	86	95	61	89
111	105	92	78	82	58
Σ 1000	995	859	792	787	839

$$N = 6 \cdot 10 = 60 \qquad \Sigma\Sigma x = 5272 \qquad x = 87{,}87 \qquad \Sigma\Sigma x^2 = 479\,435{,}7$$

$$\frac{(\Sigma\Sigma x)^2}{n} = \frac{27\,793\,984}{60} = 463\,233 \qquad Q_T = \quad 16\,202{,}7$$

$$\frac{1000^2+995^2+859^2+792^2+787^2+839^2}{10} = \begin{array}{r} 467\,846 \\ -463\,233 \\ \hline \end{array}$$

$$Q_K = \quad 4\,613$$

$$\frac{(1000+995+859)^2+(792+787+839)^2}{30} = \begin{array}{r} 466\,401{,}3 \\ -463\,233 \\ \hline \end{array}$$

$$Q_M = \quad 3\,168{,}3$$

$$\frac{(1000+792)^2+(995+787)^2+(859+839)^2}{20} = \begin{array}{r} 463\,499{,}5 \\ -463\,233 \\ \hline \end{array}$$

$$Q_P = \quad 266{,}5$$

$$Q_i = 4613-(266{,}5+3168{,}3) = \quad 1\,178{,}2$$

$$Q_B = 16\,200{,}7-4613 = \quad 11\,585{,}7$$

atosak
igyunk
llítunk
sra egy
ızottan
hogy a
nástól.
talános
nényei,
önböző
l 16.5.2

öl indu-
sa van.
közép-
értelme
gos ha-
z anyag

oportok
értelme.
elyhiány
chnikája
erpretá-

ett tekin-
iban volt
szerepelt,

mert en-
k-e (itt a
elengedő
ogy „van
kombiná-
a para-

llítetthez.
ajd a cso-
kombiná-
m tudjuk
an a 12.4-
akkor az
o, mint az
akkor a
tatisztiká-

Q_P a három különböző fehérje között talált különbségek összegét jelenti, Q_M pedig ugyanezt a fehérje mennyiségére. *N*

16.4 táblázat

Varianciaanalízis

	Sz. f.	*SSQ*	*MSQ*	*F*
A fehérje eredete	2	266,5	133,2	0,6
Adag	1	3168,3	3168,3	14,8
Interakció	2	1179,2	589,6	2,7
Kezelés	5	4613,0	922,6	4,3
Hiba	54	11585,7	214,6 = s^2	
Összesen	59	16202.7	—	

valamenn
volt a szo
adat ki
A Q_P-n
10+10
ugyane
szorzó
A
(*16.4*
juk
lön
ter
ös
es

Éppen emiatt *Finney*-nek [4] egy példáját alakitottam át minimálisan: a sorok sorrendjét cseréltem meg. Ennek következtében nem változott meg sem a Q_T, sem a készítményekre (betű) jutó, sem a nyulakra (oszlopok) jutó, de a sorrendre (sorok) jutó lényegesen csökkent, és emiatt az „error"-ra jutó ugyanannyit nőtt *(16.5 táblázat)*.

16.5 táblázat

Latin négyzet
(insulinkészitmények összehasonlitása, vércukor, mg%)

Napok	Nyulak				Számitások		
	I	II	III	IV	Σx	$\bar{x}$	$(\Sigma x)^2$
1.	B 47	A 90	C 79	D 50	266	66,50	70756
2.	D 46	B 61	A 87	C 66	260	65,00	67600
3.	A 62	C 74	D 58	B 59	253	63,25	64009
4.	C 76	D 63	B 63	A 69	271	67,75	73441
Σx	231	288	287	244	1050		275806
$\bar{x}$	57,75	72,00	71.75	61.00	$\bar{\bar{x}} = 65{,}625$		

Insulinkészítmények	A	B	C	D
Σx	308	230	295	217
x	77,0	57,5	73,8	54,2

$$\Sigma\Sigma x^2 = 71452$$

$$\text{Korrekciós faktor} = 1050^2/4 = 68906{,}26$$

$$Q_T = 2545{,}75$$

A Q_T és a 3 darab Q_K hozzájárulását úgy kell kiszámítani, mint az előzőekben. Az „error", mint a 12.4-ben volt, mert itt sem voltak paralelek. Itt nem látszik interakció. Ha gyanú lenne rá, akkor paralelekkel kellene megismételni a vizsgálatot. (Megjelent adatokat használok és az átszámítás S. I.-re rontaná az áttekinthetőséget.)

16.6. táblázat

Varianciaanalízis

	Sz. f.	*SSQ*	*MSQ*	$F_{[3;6]}$
Nyulak	3	646,25	215,4	4,44
Napok	3	45,25	15,1	–
Insulin	3	1563,25	521,1	10,74
Hiba	6	291,00	48,5	6,96 = *s*
Összesen	15	2545,75	–	–

16-1

Az $F_{[3;6]}$ kritikus értéke 4,76. Az insulinkészítmények ezt erősen meghaladták, a nyulak majdnem elérték, a napok erősen elmaradtak tőle *(16.6 táblázat)*. A napokra nem is számoltuk ki. Ugyanis, ha a beavatkozások – itt a napok – hatástalanok, akkor a csoportok közti eltérés akkora, mint amit az egyedek közti eltérés okoz, mint ezt a 12.4 pontban leírtuk. A véletlen ingadozások miatt azonban hatástalanság esetén sem kapunk „pont 1-et", hanem ingadozik körülötte. (Viszont ha annyiszor nagyobb, hogy ezt már a véletlen csak ritkán okoz, szignifikánsnak minősítjük.) Természetesen ugyanigy eltérhet a tört értéke lefelé is. Itt a kérdés csak az lehet: „növelte a variabilitást"? Itt nem fordulhat elő az a nagyon de nagyon ritka eset, hogy csök-

n emiatt is
árul.
. varianciá-
övényekkel
gi eredet a
.. az utóbbi
smertetve a
tivum vizs-
akkor ő az

place idejé-
azokat az
k az orvosi
olgozására.
elések.
:eket, hogy
vételt úgy,
:engedhető

lrendezést,
ntos terve-
lehetővé a
ezek miatt
designokat
án is.

ormája az
st adtunk
Student-
e. Az egy-
kívánton
továbbá az
getlenség.
hogy míg
zempont-

adatokat,
ek száma

bözö helyeken akarunk azonos kezeléseket [illegible] lönbözö stb. Azokat az eseteket, amikor szándékosan nem egyforma a létszam, később beszéljük meg, de már a kétmintás t-próbát ismertetve is tettünk erről említést.

Egy igen jó hatású, de csak parenteralisan adható készitményt igyekeztek oralisan is hatásossá tenni. Különbözö időpontokban mérték a szer vérszintjét. Az egyik ilyen kritikus időpontban mért adatokat mutatja a *16.1 táblázat*. (További részleteket nem közölhetünk, mert a „védö anyag” toxicusnak bizonyult, és ezért további próbálkozások folynak.) A könnyebb áttekinthetőség céljából utólag nagyság szerint rendeztük az adatokat. A létszám különbözőségének oka: el kellett dönteni, hogy a 3×1 p. os elegendő-e avagy 3×2-re van szükség, ezért vettek többet. Az injectiós adagolást már eléggé ismerték, egyformának is szánták a két csoport létszámát, de az im. csoportból verekedés miatt kettő elhullott. (Az előző vizsgálatok során nem tapasztaltak ilyen hatást az állatokon, emiatt tartották őket együtt.) Történetesen 32 megfelelő állatuk volt, és ezért tervezték 6 – 6 – 10 – 10 elosztásúnak.

16.1 táblázat

Egyszempontos varianciaanalizis
(Vérszintértékek különbözö adagolás mellett)

	im.	sc.	p. os (3×1)	p. os (3×2)	Összesen
Mért értékek	9	9	6	10	
	10	10	7	11	
	12	11	7	13	
	13	13	8	13	
		14	9	14	
		15	10	14	
			12	14	
			12	15	
			13	16	
			13	19	
n	4	6	10	10	30
Σx	44	72	97	139	352
$\bar{x}$	*11,0*	*12,0*	*9,7*	*13,9*	11,73
Σx^2	494	892	1005	1989	4380
$n\bar{x}^2$	484	864	940,9	1932,1	4221
Q_B	10	28	64,1	56,9	4130, 13
s^2	3,33	5,60	7,21	69,2	Q_K = 90,87
					249,87

A p. os készítményben egy uj anyag is szerepel, tehát gondolni kell a [illegible] hatásra. Később valóban be is bizonyosodott.

A számítás menete logikája azonos a 12.4-ben ismertetettel, de kissé bonyolultabb. Az össz-négyzetes eltérés (Q_T) és a csoporton belüli (Q_B) kiszámítása változatlan. A csoportok között (Q_K) azonban mindegyik csoportnál más és más a szorzószám, mert más és más a létszám *(16.2 táblázat)*. A 12.4 példájában a kezelések köztit számítva, a középértékek négyzeteit összeadtuk, és csak az összegüket szoroztuk 5-tel, a csoportlétszámmal. A fészkek különbségét vizsgálva is előbb összeadhattuk a középértékek négyzeteit, de itt már 4-gyel szoroztuk az összeget, mert ennyi volt az egyes fészkek létszáma. Itt azonban *külön-külön kell elvégezni a szorzást*, és csak azután szabad összegezni. Könnyebbség, hogy az $n\bar{x}^2$-értéket a csoporton belüli kiszámításához is ki kellett számítani. (A kerekítési hibák miatt előnyösebb a $\frac{(\Sigma x)^2}{n}$ értékkel számolni.)

16.2 táblázat

Varianciaanalízis

	Sz. f.	*SSQ*	*MSQ*
Csoporton belül	26	159,0	6,12
Csoportok között	3	90,9	30,30 = s^2
Összesen	29	249,9	–

$F_{[3,26]} = 4,95$

Az *SSQ* a „négyzetes eltérés összege” (sum of squares) és az MSQ a „négyzetes eltérések átlaga” (mean sqares) a nemzetközi irodalomban leggyakrabban használt rövidítések.

Tehát a két ha
csoporton beli
jelentősége.
Ebben a péld
hettük az egye
előfordulhat, h
nem szignifiká
különbözik. Ily
Minden szigi
egyetlen összehc
akarunk összeh
port van, akkor
Ezután a másoc
$\binom{n}{2}$; itt tehát
Ezek száma – t
esetén tehát 3 a f
. A $P = 5\%$ an
esetben kapunk
kapunk, vagyis a
valószínűsége, h
kozás közül egys
hogy kettő közü
három próbálko

- Bartlett's (1947) test for the significance of the k smallest eigenvalues is printed, where k can be 1, 2, etc. The uppermost line (chi-square = 56.31) tests whether the eigenvalues differ significantly from zero; this is a test that the correlations between the two sets of variables are zero. A significant chi-square indicates that the two sets or variables are not independent. The next line (chi-square = 23.66) tests whether all eigenvalues but the largest differ significantly from zero; this is a test of whether the first canonical variable is sufficient to describe the dependence between the two sets of variables. The number of canonical variables of practical value is less than or equal to the smallest number of eigenvalues for which Bartlett's test for the remaining eigenvalues is nonsignificant.

(9) Canonical variable loadings. These are the correlations of the canonical variables with the original variables. CNVRF1 is the name assigned by P6M to the 1st canonical variable in the first set; CNVRF2 to the 2nd, etc. CNVRS1 is the name assigned to the 1st canonical variable in the second set, etc. These correlations are analogous to unrotated factor loadings.

The canonical paragraph. The variables included in each set of variables must be specified in the CANONICAL paragraph. Each set should contain at least two variables; otherwise a regression program (Chapter 13) should be used.

The number of canonical variables to be obtained can be stated explicitly (NUMBER). If not stated the number is determined by the program as being all canonical variables whose correlations are greater than CONSTANT. (CONSTANT is preset to zero.)

In addition, you can specify the tolerance for matrix inversion and whether covariances and correlations are computed about the mean or about the origin.

CANONical

FIRST v list. required none
Names or subscripts of variables in the first set of variables. At least two variables must be specified.

SECOND v list. required none
Names or subscripts of variables in the second set of variables. At least two variables must be specified.

NUMBer = #. # of vars. in smaller set
Maximum number of canonical variables to be obtained.

CONSTant = #. 0.0/prev.
Canonical variables obtained must have a canonical correlation that exceeds CONSTANT.

TOLerance = #. between 0.0 & 1.0 0.0001/prev.
Tolerance for matrix inversion. Inversion is performed by stepwise pivoting. A variable is not pivoted if its squared multiple correlation with already pivoted variables exceeds 1 minus TOLERANCE, or if pivoting causes an already pivoted variable to have a squared multiple correlation with other pivoted variables that exceeds 1 minus TOLERANCE. Note that if a zero intercept model is used, then the R^2 is estimated under the assumption that all variables have zero means.

ZERO.
Covariances and correlations are computed about the origin and not about the mean. This is a rarely used option.

Example 6M.2 Printing the Coefficients of the Canonical Variables and the Canonical Variable Scores

In addition to the correlation matrix and the canonical variable loadings printed in 6M.1, P6M can print the covariance matrix, the canonical variables and the regression coefficients for the canonical variables. The number of cases for which the data

Output 6M.2 Scores and coefficients of the canonical variables

```
                    --- (2) to (8) in Output 6M.1 are printed ---

COEFFICIENTS FOR CANONICAL VARIABLES FOR FIRST SET OF VARIABLES             (10)

                   CNVRF1          CNVRF2          CNVRF3          CNVRF4
                      1               2               3               4

SMOKING1    3  0.378543D-01 -0.976451D 00 -0.965493D 00 -0.900841D 00
SMOKING2    5 -0.109322D 01 -0.646536D 00  0.134105D 01  0.182999D 00
SMOKING3    7  0.119115D 01 -0.173899D 00 -0.333693D-01  0.145194D 01
SMOKING4   12 -0.704060D 00  0.128569D 01 -0.660196D 00 -0.214955D 00


STANDARDIZED COEFFICIENTS FOR CANONICAL VARIABLES FOR FIRST SET OF VARIABLES
(THESE ARE THE COEFFICIENTS FOR THE STANDARDIZED VARIABLES - MEAN ZERO, STANDARD DEVIATION ONE.)

                CNVRF1   CNVRF2   CNVRF3   CNVRF4
                     1        2        3        4

SMOKING1    3    0.043   -1.104   -1.092   -1.019
SMOKING2    5   -1.160   -0.686    1.423    0.194
SMOKING3    7    1.383   -0.202   -0.039    1.686
SMOKING4   12   -0.898    1.641   -0.842   -0.274
```

(output continued)

Canonical correlations: 0.52229
0.37588
0.24040
0.13719

Number of canonical correlations	χ^2	df	ε Significance
0	56.31	32	0.00502
1	23.66	21	0.30975
2	8.05	12	0.70912
3	1.41	5	0.85637

Output 6M.2 (continued)

```
        COEFFICIENTS FOR CANONICAL VARIABLES FOR SECOND SET OF VARIABLES

                          CNVRS1           CNVRS2           CNVRS3           CNVRS4
                               1                2                3                4

        CONCENTR    1 -0.441692D 00   0.745510D 00 -0.470381D 00 -0.163811D 00
        ANNOY       2  0.801410D 00   0.461495D 00 -0.605503D 00 -0.739549D 00
        SLEEPY      4 -0.250790D 00   0.581216D 00 -0.685988D 00  0.615867D 00
        TENSE       6 -0.692552D 00  -0.380734D 00  0.421877D 00  0.448775D 00
        ALERT       8  0.140028D 00   0.204741D 00  0.150159D 01 -0.685341D 00
        IRRITABL    9  0.900002D-01  -0.795294D 00  0.425982D 00  0.113746D 01
        TIRED      10 -0.327905D 00  -0.616257D 00 -0.246355D 00  0.172116D 00
        CONTENT    11 -0.402041D 00  -0.595032D 00 -0.971468D 00 -0.795208D 00


       STANDARDIZED COEFFICIENTS FOR CANONICAL VARIABLES FOR SECOND SET OF VARIABLES
       (THESE ARE THE COEFFICIENTS FOR THE STANDARDIZED VARIABLES - MEAN ZERO, STANDARD DEVIATION ONE.)

                       CNVRS1   CNVRS2   CNVRS3   CNVRS4
                            1        2        3        4

       CONCENTR    1   -0.474    0.800   -0.505   -0.176
       ANNOY       2    0.781    0.450   -0.590   -0.721
       SLEEPY      4   -0.257    0.595   -0.702    0.630
       TENSE       6   -0.687   -0.378    0.418    0.445
       ALERT       8    0.143    0.208    1.529   -0.698
       IRRITABL    9    0.070   -0.622    0.333    0.890
       TIRED      10   -0.313   -0.588   -0.235    0.164
       CONTENT    11   -0.339   -0.501   -0.818   -0.670


      CANONICAL VARIABLES (CASE NUMBERS REFER TO DATA BEFORE DELETION OF CASES)

              CASE
   LABEL       NO.     WEIGHT      CNVRF1      CNVRF2      CNVRF3      CNVRF4      CNVRS1      CNVRS2      CNVRS3      CNVRS4

                 1     1.0000     -0.1954      1.8242      1.2321     -1.3620      0.1242      1.1330     -0.5973     -0.9229
                 2     1.0000     -0.6712     -0.8597     -0.7208      0.7465     -2.5299     -1.3010      0.2394     -0.1880
  (11)           3     1.0000     -1.1961     -0.9950      0.9382      0.4104     -2.5714      0.9062      1.3725      0.0831
                 4     1.0000     -1.9002      0.2907      0.2780      0.1954     -2.1278     -0.7060      1.2109      0.6072
                 5     1.0000     -0.1029     -0.3485     -0.4028      0.2274     -0.8327      0.8508     -0.0588     -1.4278

                          --- canonical variables for cases 6 to 105 ---

               106     1.0000     -0.1029     -0.3485     -0.4028      0.2274      0.5286     -0.8461      1.3204      0.2293
               107     1.0000      0.9525      1.2745     -0.7784      0.9452      1.0792      0.1966      0.0289      0.1056
               108     1.0000      0.9958      1.6503      1.1987      0.0899     -0.7903      1.9462      0.4103      0.2740
               109     1.0000     -0.2387      1.4484     -0.7450     -0.5067     -0.2426      1.8573     -0.3426     -0.1194
               110     1.0000     -0.4109     -2.8813      2.6100     -1.1308      1.0021     -1.0009      0.4685     -0.3382


     NUMERICAL CONSISTENCY CHECK

     THE FOLLOWING VARIANCES OF CANONICAL VARIABLES SHOULD ALL BE EQUAL TO ONE

     CANONICAL VARIABLE       VARIANCE     RELATIVE ERROR


              CNVRF1     0.100000D 01      0.187350D-14
              CNVRF2     0.100000D 01      0.301148D-14
              CNVRF3     0.100000D 01      0.213718D-14
  (12)        CNVRF4     0.100000D 01      0.337230D-14
              CNVRS1     0.100000D 01     -0.155431D-14
              CNVRS2     0.100000D 01      0.141553D-14
              CNVRS3     0.100000D 01      0.366374D-14
              CNVRS4     0.100000D 01      0.327516D-14


                            --- (9) in Output 6M.1 is printed ---
```

(7) Summary table. This contains a one line summary of each step including the F-to-enter (or remove) for the variable entered (or removed), the Wilks' lambda U statistic and the approximate F statistic.

(8) Classification of each case. For each case Mahalanobis D is computed to each group mean. The posterior probability for the distance of a case from a group is the ratio of $\exp(D^2)$ for the group over the sum of $\exp(D^2)$ for all groups. Prior probabilities, if assigned, affect these computations (see Appendix A.23, step 4). Outliers can be identified as cases with large D^2 from their group means. For large samples from a multivariate normal distribution, the D^2 from a case to its group mean is approximately distributed as a chi-square with degrees of freedom equal to the number of variables selected.

Each case incorrectly classified is noted in the output (cases 5, 9 and 12).

Output 7M.1 (continued)

```
                        --- results for steps 2 and 3 ---

 STEP NUMBER    4
 VARIABLE ENTERED    1 SEPALLEN

  VARIABLE          F TO   FORCE  TOLERANCE  *       VARIABLE          F TO   FORCE    TOLERANCE
                   REMOVE LEVEL              *                        ENTER  LEVEL
               DF=   2  144                  *                DF=   2  143
   1 SEPALLEN     4.721    1     0.347993    *
   2 SEPALWID    21.936    1     0.608860    *
   3 PETALLEN    35.590    1     0.365126    *
   4 PETALWID    24.904    1     0.649314    *


 U-STATISTIC OR WILKS' LAMBDA    0.0234386    DEGREES OF FREEDOM    4    2      147
 APPROXIMATE F-STATISTIC           199.145    DEGREES OF FREEDOM     8.00    288.00


 F - MATRIX          DEGREES OF FREEDOM =     4   144

            SETOSA    VERSICOL
 VERSICOL   550.19
 VIRGINIC  1098.27    105.31


 CLASSIFICATION FUNCTIONS

           GROUP =     SETOSA         VERSICOL       VIRGINIC
 VARIABLE
   1 SEPALLEN        23.54416       15.69820       12.44584      (6)
   2 SEPALWID        23.58786        7.07252        3.68529
   3 PETALLEN       -16.43063        5.21145       12.76655
   4 PETALWID       -17.39839        6.43422       21.07909

 CONSTANT           -86.30843      -72.85257     -104.36826
```

```
 CLASSIFICATION MATRIX

 GROUP      PERCENT    NUMBER OF CASES CLASSIFIED INTO GROUP -
            CORRECT
                        SETOSA    VERSICOL  VIRGINIC
 SETOSA     100.0       50          0          0
 VERSICOL    96.0        0         48          2
 VIRGINIC    98.0        0          1         49

 TOTAL       98.0       50         49         51
```

```
 JACKKNIFED CLASSIFICATION

 GROUP      PERCENT    NUMBER OF CASES CLASSIFIED INTO GROUP -
            CORRECT
                        SETOSA    VERSICOL  VIRGINIC
 SETOSA     100.0       50          0          0
 VERSICOL    96.0        0         48          2
 VIRGINIC    98.0        0          1         49

 TOTAL       98.0       50         49         51
```

```
 SUMMARY TABLE         (7)

  STEP           VARIABLE            F VALUE TO       NUMBER OF        U-STATISTIC   APPROXIMATE     DEGREES OF
 NUMBER     ENTERED     REMOVED   ENTER OR REMOVE  VARIABLES INCLUDED                F-STATISTIC      FREEDOM
   1     3 PETALLEN                  1180.1597            1               0.0586      1180.161     2.00   147.00
   2     2 SEPALWID                    43.0353            2               0.0369       307.104     4.00   292.00
   3     4 PETALWID                    34.5686            3               0.0250       257.503     6.00   290.00
   4     1 SEPALLEN                     4.7211            4               0.0234       199.145     8.00   288.00
```

(9) Eigenvalues of the matrix $\underline{W}^{-\frac{1}{2}}\underline{B}\underline{W}^{-\frac{1}{2}}$ are computed where $\underline{B}$ is the between-groups sums of cross products and $\underline{W}$ is the pooled (within-groups) sum of squares (see Appendix A.23 for a more precise definition). The eigenvalues, canonical correlations between the variables entered and dummy variables representing the groups, and the coefficients for the canonical variables are printed. The first canonical variable is the linear combination of variables entered that best discriminates among the groups (largest one-way ANOVA F statistic), the second canonical variable is the next best linear combination orthogonal to the first one, etc. The canonical variables are adjusted so that the (pooled) within-group variances are one and their overall mean is zero. The canonical variables are evaluated at the group means.

(10) The group means only are plotted in a scatter plot. The axes are the first two canonical variables. (This plot is not reproduced in Output 7M.1).

```
              INCORRECT             MAHALANOBIS D-SQUARE FROM AND
           CLASSIFICATIONS          POSTERIOR PROBABILITY FOR GROUP -

GROUP  SETOSA                      SETOSA        VERSICOL       VIRGINIC

  CASE
   1                             0.2 1.000     90.7 0.000    181.6 0.000
   6                             1.3 1.000     84.0 0.000    170.1 0.000   (8)
  10                             2.3 1.000    113.7 0.000    210.0 0.0
  18                             2.8 1.000     67.5 0.000    145.7 0.000
  26                             4.0 1.000    113.2 0.000    210.2 0.0

--- similar statistics for the remaining SETOSA cases ---

GROUP  VERSICOL                    SETOSA        VERSICOL       VIRGINIC

  CASE
   3                           105.3 0.000      2.2 0.996     13.1 0.004
   8                           131.7 0.000      8.4 0.960     14.8 0.040
   9                  VIRGINIC 130.9 0.000      8.7 0.253      6.5 0.747
  11                            99.2 0.000      1.3 0.998     13.8 0.002
  12                  VIRGINIC 149.0 0.000      8.4 0.143      4.9 0.857

--- similar statistics for the remaining VERSICOL cases ---

GROUP  VIRGINIC                    SETOSA        VERSICOL       VIRGINIC

  CASE
   2                           208.6 0.0       27.3 0.000      1.9 1.000
   4                           207.9 0.0       31.7 0.000      4.5 1.000
   5                  VERSICOL 133.1 0.000      5.3 0.729      7.2 0.271
   7                           173.2 0.000     26.6 0.000     11.0 1.000
  13                           159.0 0.000     12.8 0.003      1.2 0.997

--- similar statistics for the remaining VIRGINIC cases ---
```

```
EIGENVALUES        (9)                         VARIABLE     COEFFICIENTS FOR CANONICAL VARIABLES

               32.19192     0.28539             1 SEPALLEN      0.82938      0.02410
                                                2 SEPALWID      1.53447      2.16452
                                                3 PETALLEN     -2.20121     -0.93192
                                                4 PETALWID     -2.81046      2.83919

CUMULATIVE PROPORTION OF TOTAL DISPERSION      CONSTANT         2.10510     -6.66147

                0.99121     1.00000

                                               GROUP        CANONICAL VARIABLES EVALUATED AT GROUP MEANS
                                                  SETOSA        7.60760      0.21514
CANONICAL CORRELATIONS                            VERSICOL     -1.82505     -0.72790
                                                  VIRGINIC     -5.78255      0.51277
                0.98482     0.47120

                                                     --- (10) plot of group means ---

                                                                            (output continued)
```

```
CLUSTER  MEANS
--------------
              DOCDNT     PHARM      NURSES     HOSPBEDS   ANIMAL     STARCH     LIFEEXP

  1           3.8300     0.8600     3.0000     17.0867    14.6667    71.0000    53.6667      (5)
  2           3.3329     0.8314     3.4614     23.5600    25.4285    55.0000    54.2857
  3           1.2900     0.2300     3.5000     33.9200    21.0000    57.0000    35.0000

GRAND MEAN    3.2827     0.7845     3.3391     22.7363    22.0909    59.5454    52.3636


CLUSTER  STANDARD DEVIATIONS
----------------------------
              DOCDNT     PHARM      NURSES     HOSPBEDS   ANIMAL     STARCH     LIFEEXP

  1           0.9571     0.3226     1.2826     4.3406     1.2472     1.6330     1.2472      (6)
  2           2.6076     0.5289     1.5452     9.9481     4.4994     3.7417     2.9137
  3           0.0        0.0        0.0        0.0        0.0        0.0        0.0

MEAN SQUARES
  BETWEEN    2.4435     0.1700     0.2378   112.7889   122.2632   272.3643   166.2253
  WITHIN     6.2934     0.2838     2.7061    93.6598    18.2976    13.2500     8.0119

  D.F.-S     2,   8     2,   8     2,   8     2,   8     2,   8     2,   8     2,   8

  F-RATIO     0.388      0.599      0.088      1.204      6.682     20.556     20.747

  P-VALUE     0.765      0.633      0.965      0.369      0.014      0.000      0.000


CLUSTER PROFILES - VARIABLES ARE ORDERED BY F-RATIO SIZE     (7)
--------------------------------------------------------

                           *                                        *                                      *


LIFEEXP                    -1                                      -2-                             3
STARCH                         1-                               --2-                                      3
ANIMAL                 1-                                         --2--                                    3
HOSPBEDS              --1--                                   ------2------                                      3
PHARM                   ----1------                         --------2--------                      3
DOCDNT                    ---1----                        ----------2----------                     3
NURSES  ---------------1-----------------         --------------------2-------------------                   3

                           *                                        *                                      *


EACH COLUMN DESCRIBES A CLUSTER .
THE CLUSTER NUMBER IS PRINTED AT THE MEAN OF EACH VARIABLE
DASHES INDICATE ONE STANDARD DEVIATION ABOVE AND BELOW


POOLED WITHIN CLUSTER COVARIANCES
---------------------------------
              DOCDNT     PHARM      NURSES     HOSPBEDS   ANIMAL     STARCH     LIFEEXP      (8)
                   3          4          5          6          7          8          9


DOCDNT    3     6.29
PHARM     4     1.17       0.28
NURSES    5     2.90       0.38       2.71
HOSPBEDS  6    16.27       2.07      12.41      93.66
ANIMAL    7     9.80       1.89       3.51      22.75      18.30
STARCH    8    -4.21      -0.73      -0.95     -12.72      -7.63      13.25
LIFEEXP   9     5.59       0.67       3.78      24.04       8.98      -4.12       8.01


POOLED WITHIN CLUSTER CORRELATIONS
----------------------------------
              DOCDNT     PHARM      NURSES     HOSPBEDS   ANIMAL     STARCH     LIFEEXP
                   3          4          5          6          7          8          9


DOCDNT    3   1.0000
PHARM     4   0.8757     1.0000
NURSES    5   0.7039     0.4304     1.0000
HOSPBEDS  6   0.6702     0.4017     0.7793     1.0000
ANIMAL    7   0.9134     0.8276     0.4986     0.5495     1.0000
STARCH    8  -0.4610    -0.3784    -0.1578    -0.3611    -0.4897     1.0000
LIFEEXP   9   0.7868     0.4425     0.8109     0.8777     0.7414    -0.4004     1.0000
```

each cluster for each value of k. The remaining results are printed for the largest value of k.

① For each cluster two histograms display the distance from the cluster center to each case: a) for cases in the cluster, and, b) for cases not in the cluster. The digits in the display indicate the cluster assignment for each case. The scale for each pair of histograms is set to cover the maximum distance from that cluster center.

② The cases in cluster 1 are listed with their weight and distance from the center of cluster 1. When case labels are not used, the case number is printed. The average distance for cases in cluster 1 is also printed.

③ The program computes univariate statistics using the standardized data from the three countries in cluster 1: the center (mean), standard deviation and minimum and maximum values.

Output KM.1 K-means cluster analysis of health indicators. Circled numbers correspond to those in the text

```
                  --- the BMDP instructions are printed and interpreted ---


CLUSTER  1 OF  3 CONTAINS    3 CASES                ①
-------------------------------------
  STATISTICS ARE COMPUTED FROM THE STANDARDIZED DATA
                                                                                              ⓐ
                  1          1 1
DISTANCE +.........+.........+.........+.........+.........+.........+.........+.........+.........+.........+
                                                        3.5000                                             7.0000
FROM CENTER TO CASES IN THIS CLUSTER
                                                   2                                          ⓑ
                                          2      2 2     2       2       3                          2
DISTANCE +.........+.........+.........+.........+.........+.........+.........+.........+.........+.........+
                                                        3.5000                                             7.0000
FROM CENTER TO CASES IN OTHER CLUSTERS


C A S E     WEIGHT  DISTANCE   I   VARIABLE     MINIMUM     CENTER    MAXIMUM  ST.DEV.
--------------------------     I  -------------------------------------------------
SYRIA       1.0000    1.5064   I   3 DOCDNT      1.1331     1.7085    2.1546    0.5229
TURKEY      1.0000    0.6877   I   4 PHARM       1.1697     1.7648    2.6882    0.8108
U.A.R.      1.0000    1.6138   I   5 NURSES      0.9868     2.1145    3.2000    1.1072   ③
                               I   6 HOSPBEDS    1.2349     1.8143    2.3626    0.5645
          ②                    I   7 ANIMAL      2.1799     2.4593    2.6829    0.2561
                               I   8 STARCH      8.9675     9.2275    9.4874    0.2599
                               I   9 LIFEEXP     8.6573     8.9348    9.1567    0.2543

--------------------------
AVERAGE DISTANCE      1.2693




CLUSTER  2 OF  3 CONTAINS    7 CASES
-------------------------------------
  STATISTICS ARE COMPUTED FROM THE STANDARDIZED DATA
                                                                                                      2
                               2   2      2 2      2   2
DISTANCE +.........+.........+.........+.........+.........+.........+.........+.........+.........+.........+
                                                        2.5000                                             5.0000
FROM CENTER TO CASES IN THIS CLUSTER

                                                                1        1   1       3
DISTANCE +.........+.........+.........+.........+.........+.........+.........+.........+.........+.........+
                                                        2.5000                                             5.0000
FROM CENTER TO CASES IN OTHER CLUSTERS


C A S E     WEIGHT  DISTANCE   I   VARIABLE     MINIMUM     CENTER    MAXIMUM  ST.DEV.
--------------------------     I  -------------------------------------------------
IRAN        1.0000    1.7081   I   3 DOCDNT      0.4193     1.4867    4.1620    1.2564
IRAQ        1.0000    1.1302   I   4 PHARM       0.5335     1.7062    3.9400    1.1722
JORDAN      1.0000    1.3758   I   5 NURSES      1.6423     2.4397    4.3136    1.1764
LEBANON     1.0000    4.6609   I   6 HOSPBEDS    1.1818     2.5017    4.3227    1.1410
LIBYA       1.0000    2.3137   I   7 ANIMAL      3.5213     4.2639    5.8689    0.8149
MOROCCO     1.0000    2.1473   I   8 STARCH      6.3683     7.1480    7.7979    0.5252
TUNISIA     1.0000    1.8280   I   9 LIFEEXP     8.4908     9.0378    9.9892    0.5240
--------------------------
AVERAGE DISTANCE      2.1663
```

(output continued)

CLUSTER 3 OF 3 CONTAINS 1 CAFES

ALGERIA

23

Hierarchikus, agglomeratív

Output 1M.1 Cluster analysis of the Jarvik smoking questionnaire data. Circled numbers correspond to those in the text

--- the BMDP instructions read by P1M are printed and interpreted ---

```
①   PROCEDURE MEASURE . . . . . . . . . . . . . . . . .ABSCORR
    PROCEDURE AMALGAMATION RULE IS MINIMUM DISTANCE (SINGLE LINKAGE)

②   NUMBER OF CASES READ. . . . . . . . . . . . . .       110

        VARIABLE
③   NAME       NO.        MEAN        STANDARD
                                      DEVIATION

    CONCENTR    1        2.691        1.073
    ANNOY       2        2.118        0.974
    SMOKING1    3        3.364        1.131
    SLEEPY      4        2.609        1.024
    SMOKING2    5        3.582        1.061
    TENSE       6        2.445        0.992
    SMOKING3    7        3.427        1.161
    ALERT       8        2.809        1.018
    IRRITABL    9        2.218        0.783
    TIRED      10        3.091        0.953
    CONTENT    11        2.455        0.842
    SMOKING4   12        3.500        1.276
```

```
         VARIABLE     OTHER BOUNDARY    NUMBER OF ITEMS    DISTANCE OR SIMILARITY
     NAME      NO.    OF CLUSTER        IN CLUSTER         WHEN CLUSTER FORMED
     CONCENTR   1          12               12                   30.07
④   ALERT      8           1                2                   80.21
     SLEEPY     4          10                2                   79.82
     TIRED     10           1                4                   69.85
     ANNOY      2           6                4                   72.48
     IRRITABL   9           2                2                   79.61
     CONTENT   11           2                3                   73.92
     TENSE      6           1                8                   60.54
     SMOKING1   3          12                4                   80.98
     SMOKING2   5          12                3                   81.65
     SMOKING3   7          12                2                   84.53
     SMOKING4  12           1               12                   30.07
```

```
⑤  TREE PRINTED OVER ABSOLUTE CORRELATION MATRIX.
    CLUSTERING BY MINIMUM DISTANCE METHOD.
       VARIABLE
    NAME      NO.
              ---------------------------------/
CONCENTR(  1) 80/45 51/56 59 49 57/ 8 19  4 22/
               /     /           /           /
              /     /           /           /
ALERT   (  8)/60 69/57 60 60 59/10 22  3 20/
                  /           /           /
             ----/           /           /
SLEEPY  (  4) 79/35 33 24 27/13 21 12 27/
               /           /           /
              /           /           /
TIRED   ( 10)/41 42 39 36/19 27 13 27/
                        /           /
             ----------/           /
ANNOY   (  2) 79/73/70/14 11  6 12/
               /  /  /           /
              /  /  /           /
IRRITABL(  9)/69/72/18 22 10 15/
               /  /           /
              /  /           /
CONTENT ( 11)/71/23 23  9 17/
               /           /
              /           /
TENSE   (  6)/22 30 12 21/
                        /
             ----------/
SMOKING1(  3) 78 80 77/
                     /
              ------/
SMOKING2(  5) 81 81/
                  /
              ---/
SMOKING3(  7) 84/
               /
              /
SMOKING4( 12)/
```

```
THE VALUES IN THIS TREE HAVE BEEN SCALED 0 TO 100          ⑥
ACCORDING TO THE FOLLOWING TABLE

VALUE                                        VALUE
ABOVE       CORRELATION                      ABOVE       CORRELATION
  0           0.000                           50           0.500
  5           0.050                           55           0.550
 10           0.100                           60           0.600
 15           0.150                           65           0.650
 20           0.200                           70           0.700
 25           0.250                           75           0.750
 30           0.300                           80           0.800
 35           0.350                           85           0.850
 40           0.400                           90           0.900
 45           0.450                           95           0.950
```

--- ⑦ an explanation of the clustering process ---

--- ⑧ the shaded correlation matrix appears here (see Output 1M.3) ---

246

Output 1M.2 Using the average linkage rule to join clusters

```
PROCEDURE MEASURE . . . . . . . . . . . . . . . .ANC
PROCEDURE AMALGAMATION RULE IS AVERAGE DISTANCE (AVERAGE LINKAGE)



NUMBER OF CASES READ. . . . . . . . . . . . . .        110



--- means and standard deviations for each variable ---



     VARIABLE        OTHER BOUNDARY    NUMBER OF ITEMS    DISTANCE OR SIMILARITY
NAME        NO.       OF CLUSTER         IN CLUSTER       WHEN CLUSTER FORMED
CONCENTR     1           12                 12.                 61.80
ALERT        8            1                  2                  90.10
ANNOY        2            6                  4                  89.18
IRRITABL     9            2                  2                  89.80
CONTENT     11            2                  3                  87.90
TENSE        6            1                  6                  84.00
SLEEPY       4           10                  2                  89.91
TIRED       10            1                  8                  76.49
SMOKING1     3           12                  4                  92.03
SMOKING2     5           12                  3                  92.41
SMOKING3     7           12                  2                  92.26
SMOKING4    12            1                 12                  61.80



TREE PRINTED OVER CORRELATION MATRIX (SCALED 0-100).
CLUSTERING BY AVERAGE DISTANCE METHOD.
  VARIABLE
NAME     NO.
             ----------------------------------/
CONCENTR(  1) 90/78 79 74 78/72 75/54 59 52 61/
               /           /     /           /
              /           /     /           /
ALERT   (  8)/78 80 80 79/80 84/55 61 51 60/
                        /     /           /
             ----------/     /           /
ANNOY   (  2) 89/86/85/67 70/57 55 53 56/
               /  /  /     /           /
              /  /  /     /           /
IRRITABL(  9)/84/86/66 71/59 61 55 57/
               /  /     /           /
              /  /     /           /
CONTENT ( 11)/85/62 69/61 61 54 58/
               /     /           /
              /     /           /
TENSE   (  6)/63 68/61 65 56 60/
                  /           /
             ----/           /
SLEEPY  (  4) 89/56 60 56 63/
               /           /
              /           /
TIRED   ( 10)/59 63 56 63/
                        /
             ----------/
SMOKING1(  3) 89 90 88/
                     /
             -------/
SMOKING2(  5) 90 90/
                  /
             ----/
SMOKING3(  7) 92/
               /
              /
SMOKING4( 12)/
```

THE VALUES IN THIS TREE HAVE BEEN SCALED 0 TO 100
ACCORDING TO THE FOLLOWING TABLE

VALUE ABOVE	CORRELATION	VALUE ABOVE	CORRELATION
0	-1.000	50	0.000
5	-0.900	55	0.100
10	-0.800	60	0.200
15	-0.700	65	0.300
20	-0.600	70	0.400
25	-0.500	75	0.500
30	-0.400	80	0.600
35	-0.300	85	0.700
40	-0.200	90	0.800
45	-0.100	95	0.900

--- explanation of the clustering ---

choice of instrumental variable, since it is unlikely to be correlated with measurement errors for x or with the disturbance term in the regression.

Table 7.2.1 Capital–labour substitution data (Kmenta, 1971, p. 313)

Country	y	x	z
United States	0.7680	3.5459	3.4241
Canada	0.4330	3.2367	3.1748
New Zealand	0.4575	3.2865	3.1686
Australia	0.5002	3.3202	3.2989
Denmark	0.3462	3.1585	3.1742
Norway	0.3068	3.1529	3.0492
United Kingdom	0.3787	3.2101	3.1175
Colombia	−0.1188	2.6066	2.5681
Brazil	−0.1379	2.4872	2.5682
Mexico	−0.2001	2.4280	2.6364
Argentina	−0.3845	2.3182	2.5703

The values of z are given in Table 7.2.1. The instrumental variable estimates from (7.2.7) and (7.2.8) (with estimated standard errors) lead to the equation

$$y_i = \underset{(0.10)}{-2.30} + \underset{(0.03)}{0.84}\, x_i.$$

It will be noted that IV estimates and OLS estimates are very similar. Thus, in this example, the measurement errors do not seem to be severe.

7.2.2 Two-stage least squares (2SLS) estimation

The instrumental variable matrix $\mathbf{Z}(n \times k)$ in Section 7.2.1 is assumed to have the same dimension as the "independent" variable matrix $\mathbf{X}(n \times q)$, i.e. $k = q$. However, if $k > q$ then an extension of IV estimation may be given using the method of *two-stage least squares* (2SLS). This method is defined as follows.

First, regress $\mathbf{X}$ on $\mathbf{Z}$ using the usual OLS multivariate regression estimates to get a fitted value of $\mathbf{X}$,

$$\hat{\mathbf{X}} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}. \qquad (7.2.9)$$

Note that $\hat{\mathbf{X}}(n \times q)$ is a linear combination of the columns of $\mathbf{Z}$.

Second, substitute $\hat{\mathbf{X}}$ for $\mathbf{X}$ in the original equation (7.2.1) and use OLS

to estimate

Then $\boldsymbol{\beta}^{**}$ i

Note that

so that $\boldsymbol{\beta}^{*}$

instrumenta

estimator $\boldsymbol{\beta}$

seen that $\boldsymbol{\beta}$

plim $n^{-1}\mathbf{Z}'\mathbf{Z}$

q. (See Exer

We have

practice. Int

correlated w

However, in

between $\mathbf{Z}$ a

In the con

problem of c

that the mea

errors, so tha

section that t

estimation dc

7.3 Simul

7.3.1 Structu

Imagine an ec

endogenous va

by a number

variables are

An econome

simultaneous

dogenous vari

of some of the

$\mathbf{Y}(n \times p)$ and

observations or

studies) or obs

(cross-sectional

26

Table 7.4.1 Data for food consumption and prices model

Q_t	P_t	D_t	F_t	A_t
98.485	100.323	87.4	98.0	1
99.187	104.264	97.6	99.1	2
102.163	103.435	96.7	99.1	3
101.504	104.506	98.2	98.1	4
104.240	98.001	99.8	110.8	5
103.243	99.456	100.5	108.2	6
103.993	101.066	103.2	105.6	7
99.900	104.763	107.8	109.8	8
100.350	96.446	96.6	108.7	9
102.820	91.228	88.9	100.6	10
95.435	93.085	75.1	81.0	11
92.424	98.801	76.9	68.6	12
94.535	102.908	84.6	70.9	13
98.757	98.756	90.6	81.4	14
105.797	95.119	103.1	102.3	15
100.225	98.451	105.1	105.0	16
103.522	86.498	96.4	110.5	17
99.929	104.016	104.4	92.5	18
105.223	105.769	110.7	89.3	19
106.232	113.490	127.1	93.0	20

Table 7.4.2 Estimators (with standard errors) for food consumption and prices model

	True coefficient	OLS	2SLS	LIML	3SLS	FIML
Demand equation						
Constant	96.5	99.90	94.63 (7.9)	93.62 (8.0)	Same as 2SLS	Same as LIML
P	−0.25	−0.32	−0.24 (0.10)	−0.23 (0.10)		
D	0.30	0.33	0.31 (0.05)	0.31 (0.05)		
Supply equation						
Constant	62.5	58.28	49.53 (12.01)	Same as 2SLS	52.11 (11.89)	51.94 (12.75)
D	0.15	0.16	0.24 (0.10)		0.23 (0.10)	0.24 (0.11)
F	0.20	0.25	0.26 (0.05)		0.23 (0.04)	0.22 (0.05)
A	0.36	0.25	0.25 (0.10)		0.36 (0.07)	0.37 (0.08)

Values for D_t an
the data is summ
The estimated
Table 7.4.2. Note
variable in each
Standard errors
it can be seen h
other procedures.
equation are ider
Exercises 7.4.1 a
described in the n

7.5 System E

7.5.1 Seemingly

Before we turn to
case when only *on*
a slightly different
model as

where $\mathbf{X}_j(n \times q_j)$ d
equation. Note tha
hand side of each e
the assumptions of
applied to each eq
parameters. Howe
between equations
system as a whole.
Write the model

where $\mathbf{Y}^{\text{V}} = (\mathbf{y}'_{(1)}, \ldots$
one another and se

If Σ is known,
eneralized least

(7.5.3)

'hen the data is

(7.5.4)

(7.5.5)

* and the OLS

variate regres-

:nt, but by the
ι and hence is

timate (7.5.3)
to estimate Σ
S estimate.
ls to estimate

(7.5.6)

mate of Δ,

(7.5.7)

d regularity
imated by

Example 7.5.1 (Kmenta, 1971, p. 527) Consider data on the investment performance of two firms, General Electric Company and Westinghouse Electric Company, over the period 1935–1954. Each firm's investment (I) is related to the value of its capital stock (C), and the value of its shares (F). The assumed relationship is

$$I_t = \alpha C_t + \beta F_t + \gamma + u_t, \qquad t = 1935, \ldots, 1954.$$

The results for General Electric are as follows (standard errors in parentheses):

(a) Using ordinary least squares,

$$I_t = \underset{(0.026)}{0.152C} + \underset{(0.016)}{0.027F_t} - \underset{(31.37)}{9.956}.$$

(b) Using Zellner's two-stage method,

$$I_t = \underset{(0.025)}{0.139C_t} + \underset{(0.015)}{0.038F_t} - \underset{(29.32)}{27.72}.$$

The results for Westinghouse were as follows:

(a) Using ordinary least squares,

$$I_t = \underset{(0.056)}{0.092C_t} + \underset{(0.016)}{0.053F_t} - \underset{(8.02)}{0.509}.$$

(b) Using Zellner's two-stage method,

$$I_t = \underset{(0.053)}{0.058C_t} + \underset{(0.015)}{0.064F_t} - \underset{(7.55)}{1.25}.$$

It can be seen that in the case of each of the six coefficients, Zellner's estimate has a lower estimated standard error than does the ordinary least squares estimate.

7.5.2 Three-stage least squares (3SLS)

The method of three-stage least squares involves an application of Zellner's estimator to the general system of structural equations.

As in Section 7.4.1, write each of the structural equations as a regression-like equation,

$$\mathbf{y}_{(j)} = \mathbf{Z}_j \boldsymbol{\delta}_{(j)} + \mathbf{u}_{(j)}, \qquad j = 1, \ldots, p - r.$$

Here $\mathbf{Z}_j = (\mathbf{Y}_{*,j}, \mathbf{X}_{0,j})$ denotes those endogenous and exogenous variables (other than $\mathbf{y}_{(j)}$) appearing in the jth equation and $\boldsymbol{\delta}_{(j)} = (-\boldsymbol{\beta}'_{*,(j)}, -\boldsymbol{\gamma}'_{0,(j)})'$ represents the corresponding structural coefficients. Also, r denotes the number of exact identities in the system which we omit from consideration.