Canonical Correlation Analysis

(정준상관분석)

H. Park

HUFS

Correlation Coefficient

- 두 변수 간의 상관관계를 나타내는 척도

X	Y
x_1	y_1
x_2	y_2
…	…
x_n	y_n

What if

x1	x2	x3	y1	y2
x_11	x_12	x_13	y_11	y_12
x_21	x_22	x_23	y_21	y_22
…	…	…	…	…
x_n1	x_n2	x_n3	y_n1	y_n2

(   )   (   )

(Q): “__________” 과
“__________” 상관관계는?

$\rho_{x1,y1}$, $\rho_{x1,y2}$

$\rho_{x2,y1}$, $\rho_{x2,y2}$

$\rho_{x3,y1}$, $\rho_{x3,y2}$

정준상관계수
(canonical correlation coefficient)

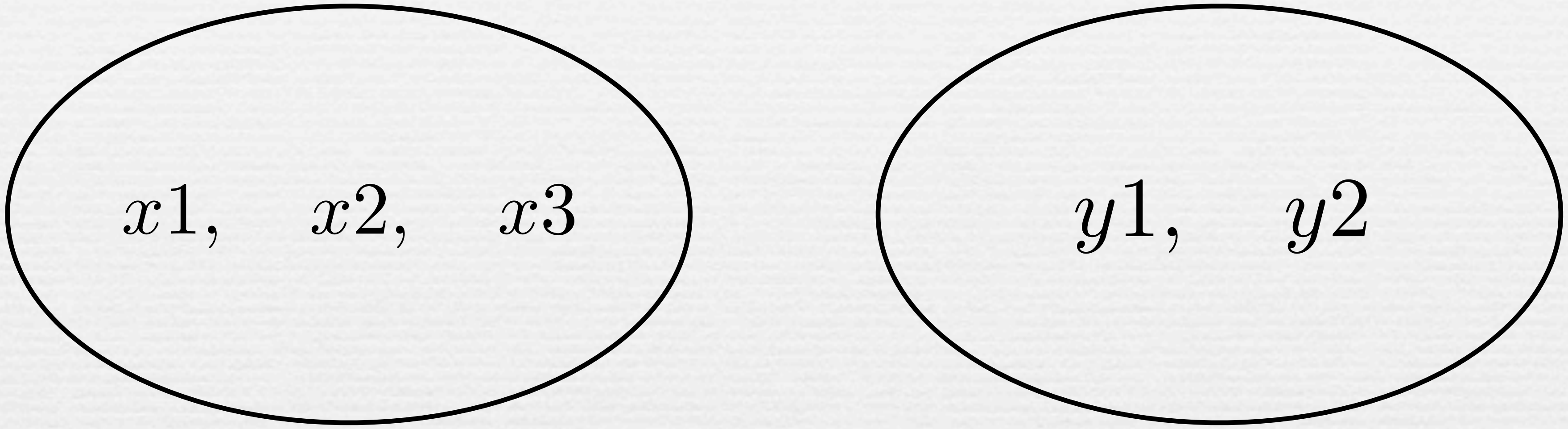
“____”

- X 그룹의 선형조합 :
- Y 그룹의 선형조합:
- 상관계수, ______________를 최대로 하는 선형조합

  =____________________ (__________________)

- 1차 정준상관변수와 직교하되, 2번째로 상관계수를 최대로 하는 선형조합 = ____________________

정준상관계수
(canonical correlation coefficient)



$x1, \quad x2, \quad x3$

$y1, \quad y2$

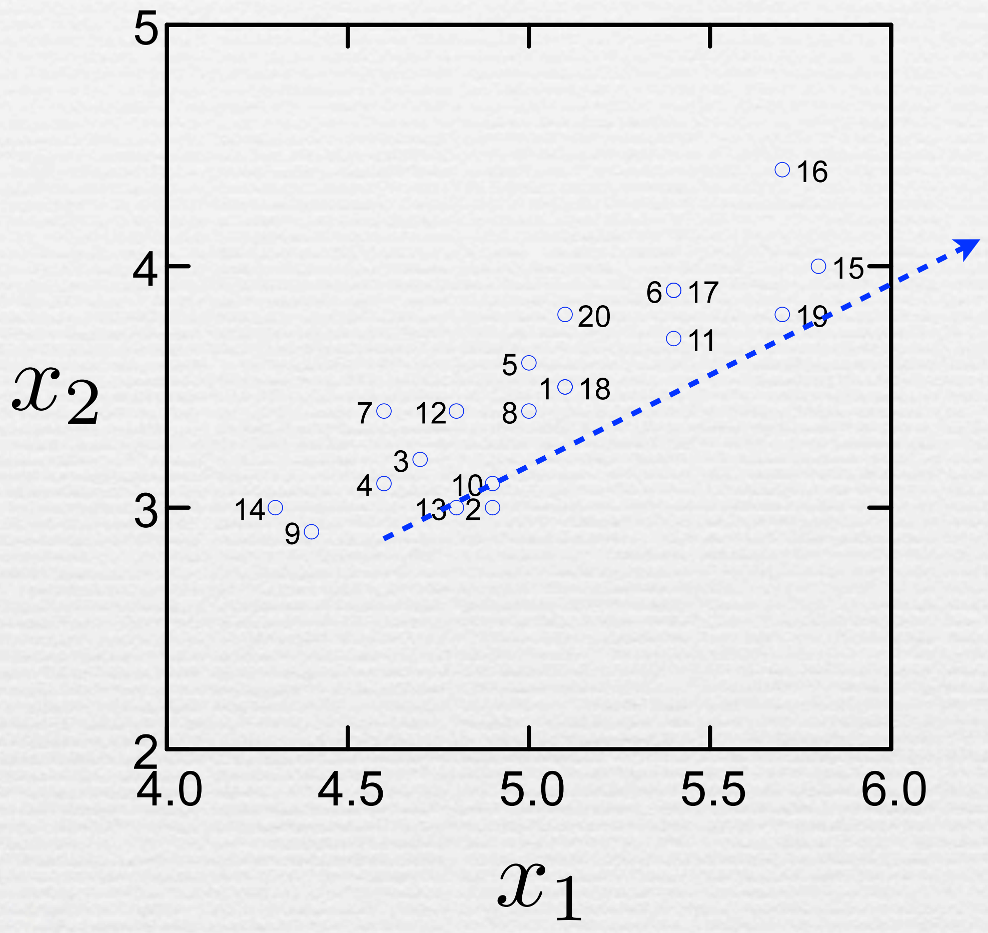
$$\max_{i} \ \mathrm{Corr}(V_i, W_i) = $$ ____________________

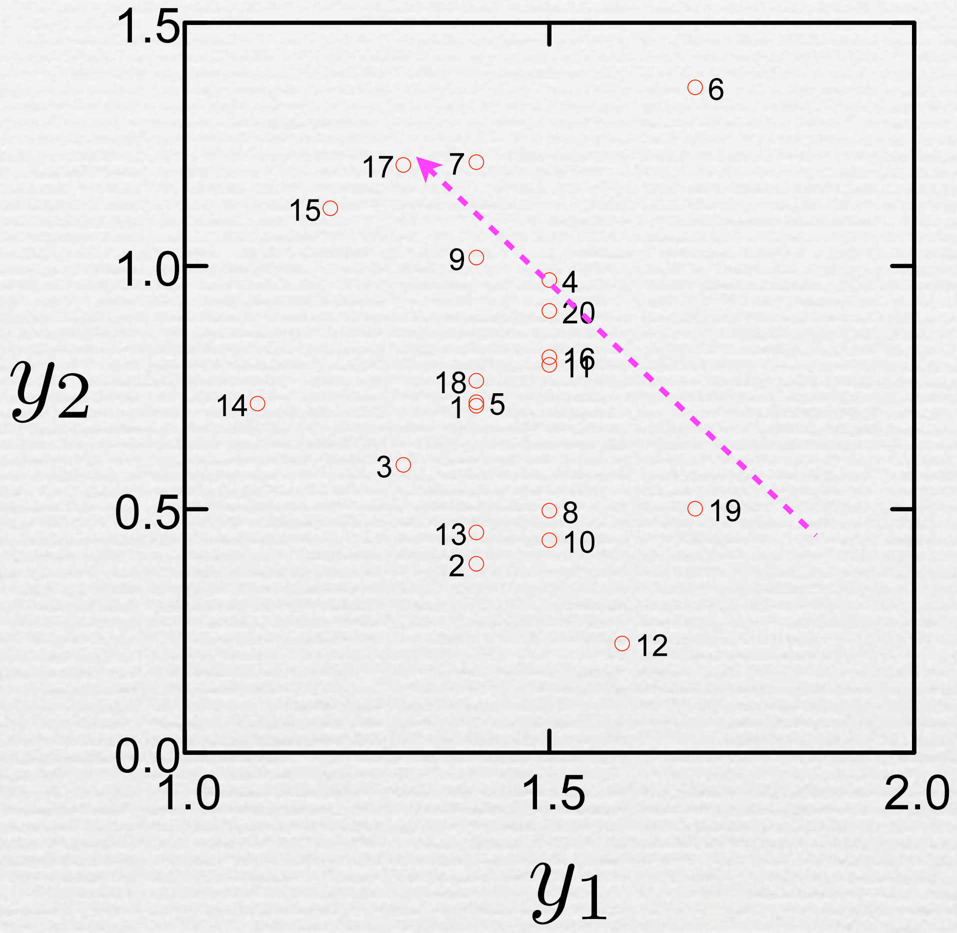
____

(Q) 몇 차 까지 가능? →

$$(V_i, W_i) = $$ ______________________________

정준상관분석





$$V_1 = a_{11}x_1 + a_{12}x_2$$

$$V_2 = a_{21}x_1 + a_{22}x_2$$

$$W_1 = b_{11}y_1 + b_{12}y_2$$

$$W_2 = b_{21}y_1 + b_{22}y_2$$

$$\max_i \ \mathrm{Corr}(V_i, W_i)$$

정준상관변수의 해석

$$V_1 = a_1x_1 + a_2x_2 + a_3x_3 \qquad W_1 = b_1y_1 + b_2y_2$$

>> 선형결합으로 표현되기 때문에, ______________________________

>> ______________________________ 의 필요성

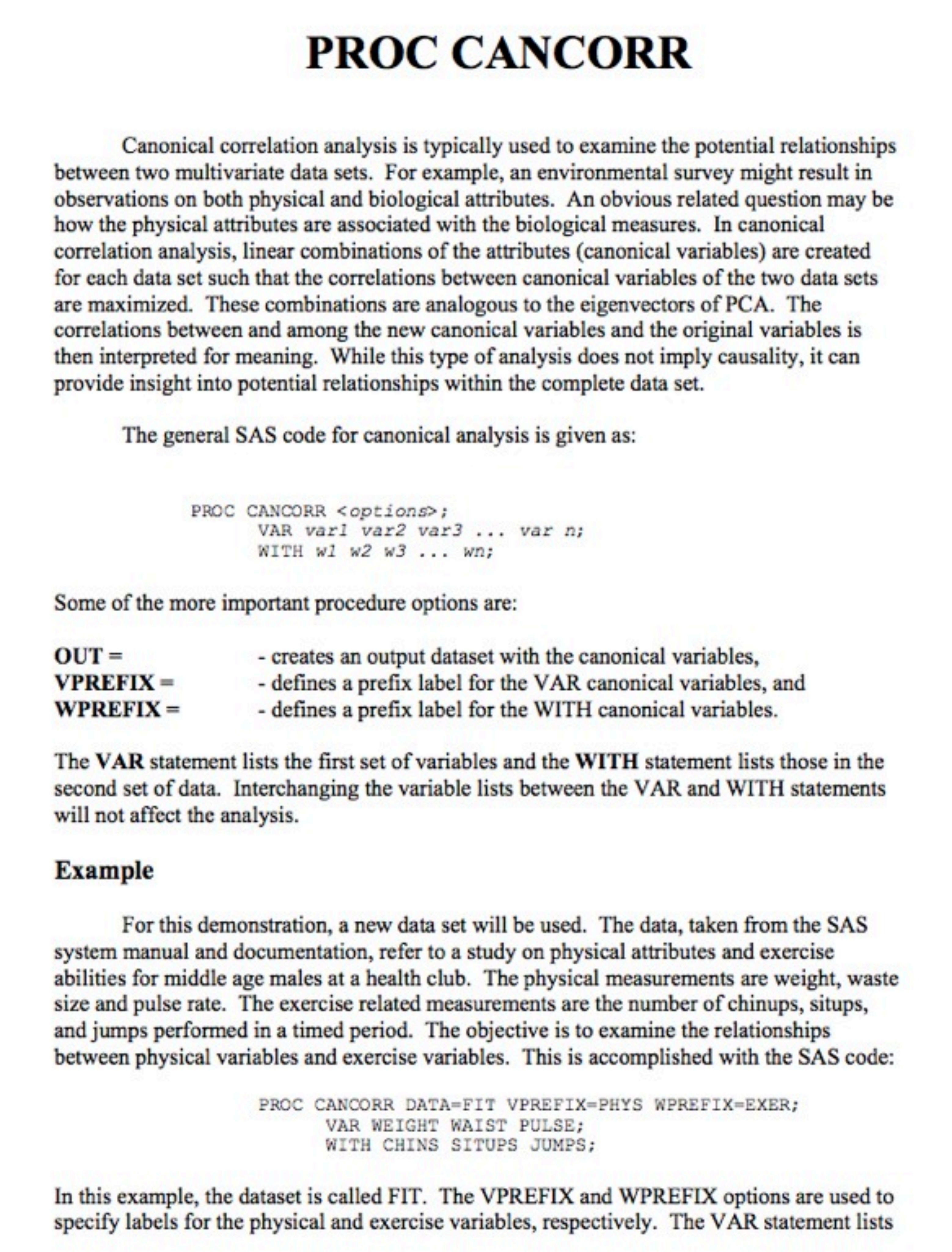
Redundancy Analysis (중복분석):

“how much of variance in one set of variables can be explained by the other set of variables”

$$R^2(V_1 \text{ on } x_1, x_2, x_3) \quad R^2(W_1 \text{ on } y_1, y_2) \quad R^2(V_1 \text{ on } W_1)$$

Proc CanCorr (SAS)

```
proc cancorr data=a ;
     vprefix =<...Vi ...이름>
    vname =’<...Vi ...설명>’
    wprefix =<...Wi ...이름>
    wname =’<...Wi ...설명>’ ;
var____________;
with ___________;
run;
```



PROC CANCORR

Canonical correlation analysis is typically used to examine the potential relationships between two multivariate data sets. For example, an environmental survey might result in observations on both physical and biological attributes. An obvious related question may be how the physical attributes are associated with the biological measures. In canonical correlation analysis, linear combinations of the attributes (canonical variables) are created for each data set such that the correlations between canonical variables of the two data sets are maximized. These combinations are analogous to the eigenvectors of PCA. The correlations between and among the new canonical variables and the original variables is then interpreted for meaning. While this type of analysis does not imply causality, it can provide insight into potential relationships within the complete data set.

The general SAS code for canonical analysis is given as:

```
PROC CANCORR <options>;
      VAR var1 var2 var3 ... var n;
      WITH w1 w2 w3 ... wn;
```

Some of the more important procedure options are:

OUT =	- creates an output dataset with the canonical variables,
VPREFIX =	- defines a prefix label for the VAR canonical variables, and
WPREFIX =	- defines a prefix label for the WITH canonical variables.

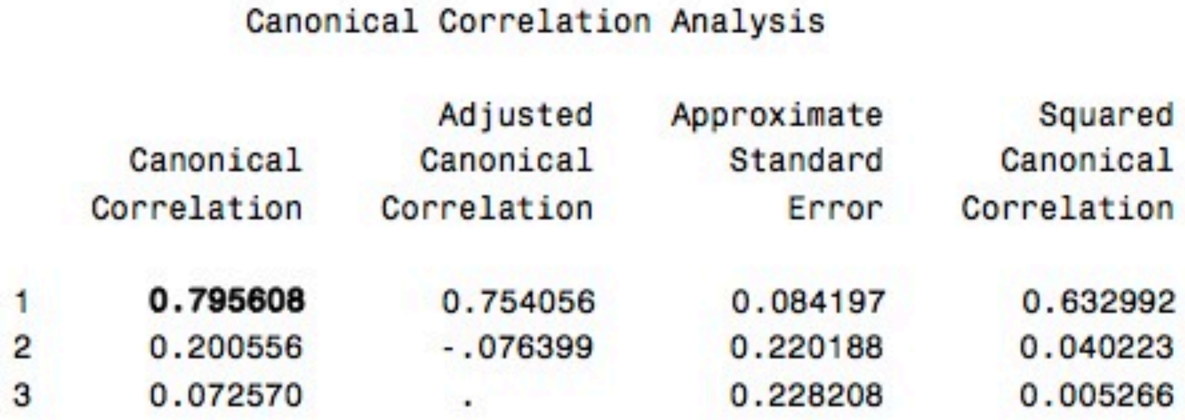
The **VAR** statement lists the first set of variables and the **WITH** statement lists those in the second set of data. Interchanging the variable lists between the VAR and WITH statements will not affect the analysis.

Example

For this demonstration, a new data set will be used. The data, taken from the SAS system manual and documentation, refer to a study on physical attributes and exercise abilities for middle age males at a health club. The physical measurements are weight, waste size and pulse rate. The exercise related measurements are the number of chinups, situps, and jumps performed in a timed period. The objective is to examine the relationships between physical variables and exercise variables. This is accomplished with the SAS code:

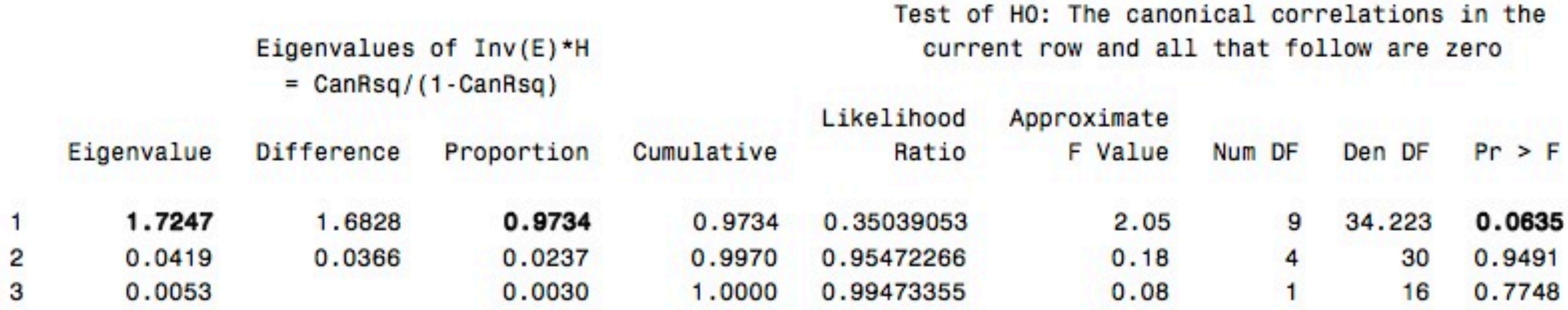
```
PROC CANCORR DATA=FIT VPREFIX=PHYS WPREFIX=EXER;
      VAR WEIGHT WAIST PULSE;
      WITH CHINS SITUPS JUMPS;
```

In this example, the dataset is called FIT. The VPREFIX and WPREFIX options are used to specify labels for the physical and exercise variables, respectively. The VAR statement lists the physical variables. Likewise, the WITH statement lists the exercise variables. The output from resulting from the code is:.



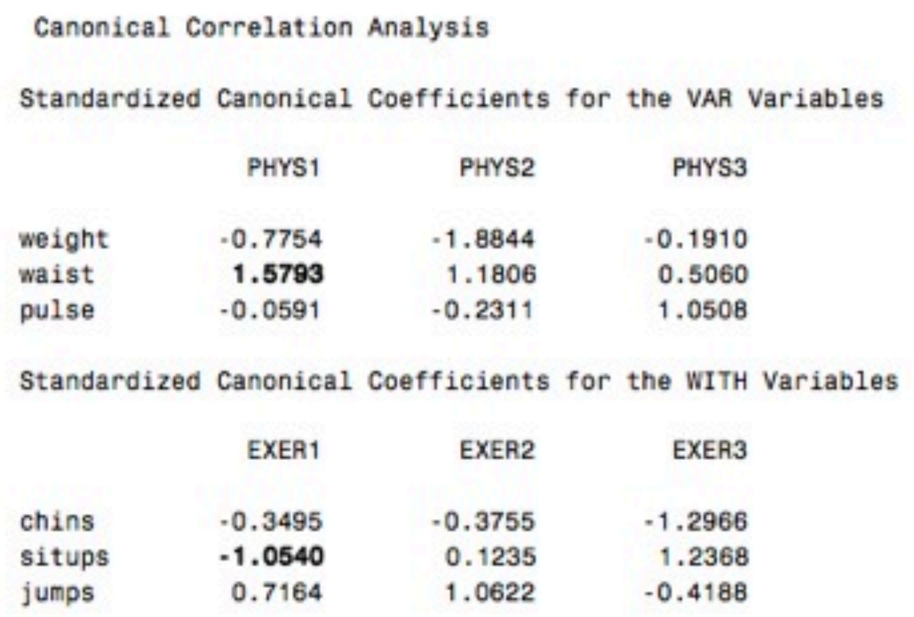
```
                         Canonical Correlation Analysis

                                 Adjusted    Approximate        Squared
                  Canonical     Canonical       Standard      Canonical
                Correlation   Correlation          Error    Correlation

          1        0.795608      0.754056       0.084197       0.632992
          2        0.200556      -.076399       0.220188       0.040223
          3        0.072570          .          0.228208       0.005266
```

This initial section prints the correlation between the "new" canonical variables. PROC CANCORR has created new canonical variable from each of the VAR and WITH variable lists and then has calculated the resulting correlations. For example, the first canonical variable for physical attributes was created from the VAR list and labeled PHYS1. This new variable is just a linear combination of weight, waste and pulse. Another canonical variable, EXER1, has been made from the WITH list. SAS computes the correlation between the new canonical variables PHYS1 and EXER1. This "canonical correlation" has a value of 0.7956. Two more canonical variables are created from each list and their correlation computed (0.2006). These values give an overall feel for the degree of association between the physical and exercise variables. At 0.79, the relationship is fairly strong.



```
                                                              Test of H0: The canonical correlations in the
                  Eigenvalues of Inv(E)*H                        current row and all that follow are zero
                    = CanRsq/(1-CanRsq)

                                                            Likelihood  Approximate
     Eigenvalue   Difference   Proportion   Cumulative          Ratio      F Value    Num DF    Den DF    Pr > F

1        1.7247       1.6828       0.9734       0.9734     0.35039053         2.05         9    34.223    0.0635
2        0.0419       0.0366       0.0237       0.9970     0.95472266         0.18         4        30    0.9491
3        0.0053                    0.0030       1.0000     0.99473355         0.08         1        16    0.7748
```

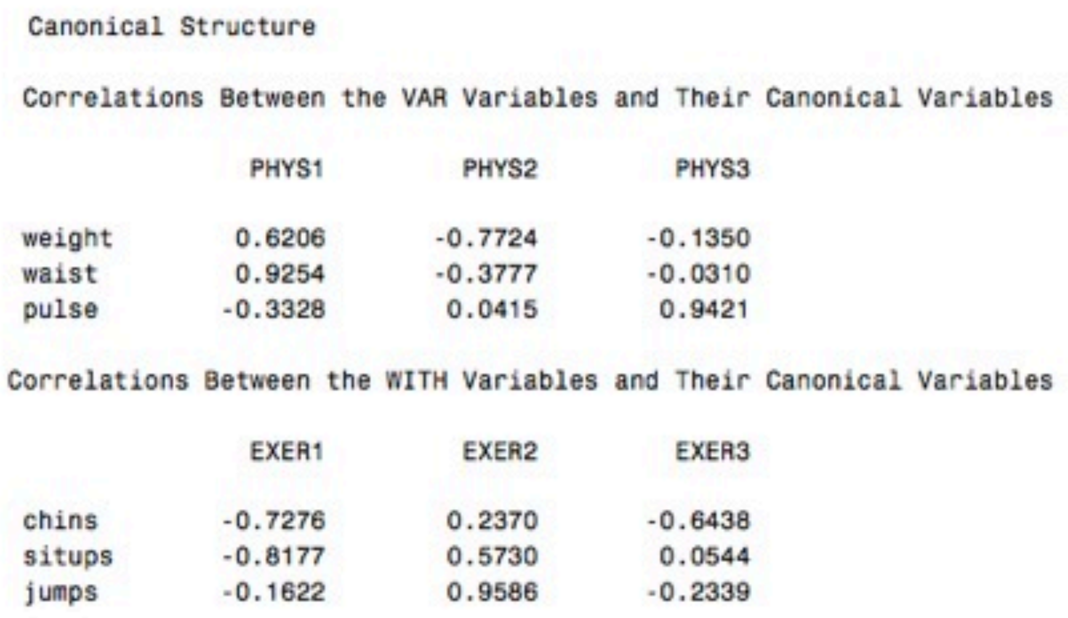
The next section of the printout provides a PCA type decomposition of the canonical variables. From this we see that the first canonical variables account for most of the data variability (97.34%) and the approximate F test shows only the first component to be significant (p = 0.0635). Thus, only the first canonical variable from each of the VAR and WITH lists has any potential meaning. The next obvious question is: "What makes up each canonical variable?". SAS provides this information in two forms: RAW Canonical Coefficients and Standardized Canonical Coefficients. Because the various variables may have different units and scales, it is best to interpret the standardized coefficients. These appear below. The RAW coefficients have been omitted from the output.



```
 Canonical Correlation Analysis

Standardized Canonical Coefficients for the VAR Variables

                  PHYS1          PHYS2          PHYS3

weight          -0.7754        -1.8844        -0.1910
waist            1.5793         1.1806         0.5060
pulse           -0.0591        -0.2311         1.0508

Standardized Canonical Coefficients for the WITH Variables

                  EXER1          EXER2          EXER3

chins           -0.3495        -0.3755        -1.2966
situps          -1.0540         0.1235         1.2368
jumps            0.7164         1.0622        -0.4188
```

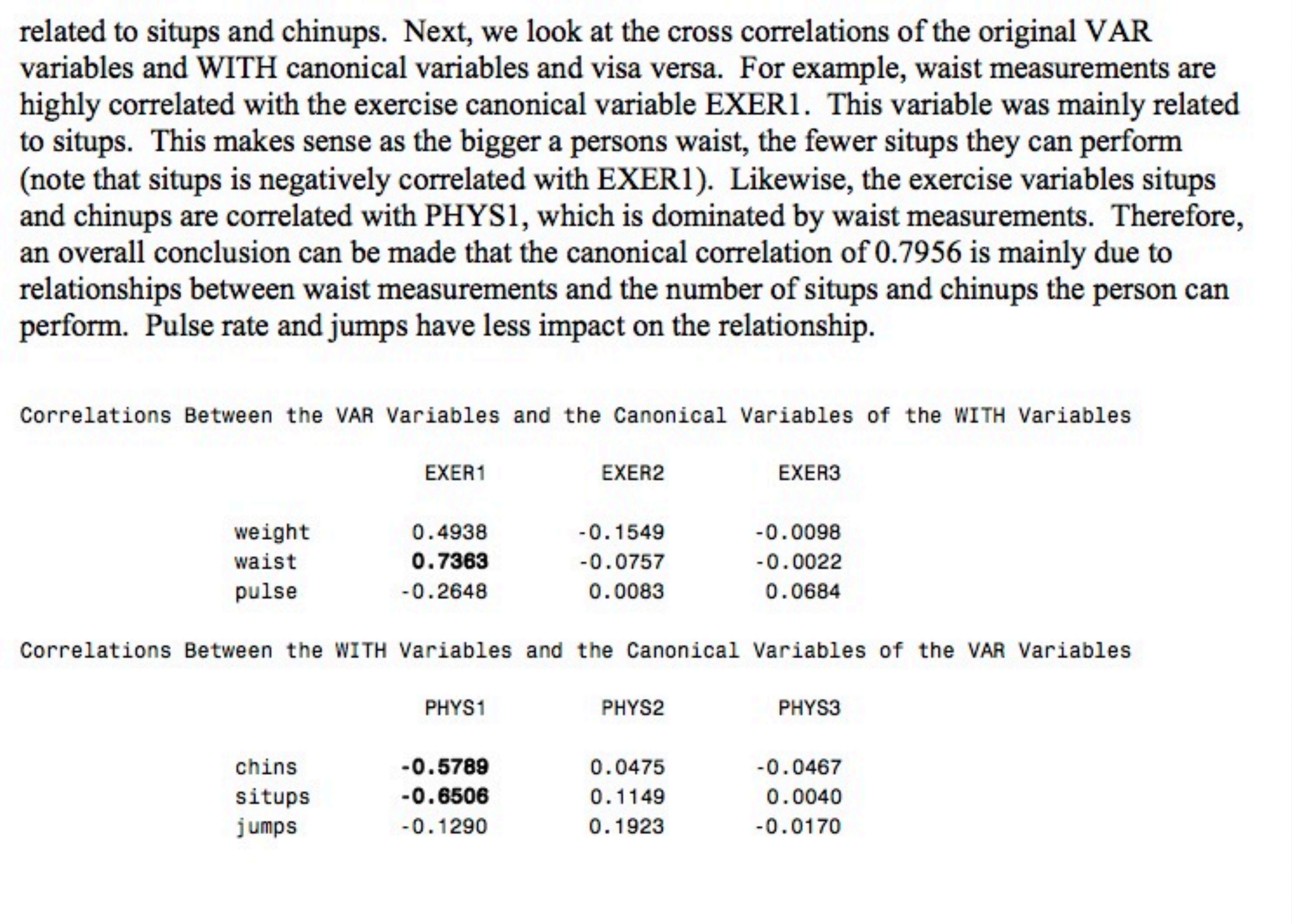
The coefficients are reported in two sections. The first is for the VAR variables and the second for the WITH variables. The coefficients are interpreted in a similar manner to PCA eigenvector loadings. Hence, we see in the canonical variable PHYS1 that the measurement waist is important. In the exercise canonical variable, situps is an important factor.

Another means of assessing the association of the original variables with the canonical variables is through correlation. The remaining section of output provides all possible combination of such correlations. The first set of correlations examines the relationships between the attributes and their respective canonical variables.



```
 Canonical Structure

 Correlations Between the VAR Variables and Their Canonical Variables

                  PHYS1          PHYS2          PHYS3

 weight          0.6206        -0.7724        -0.1350
 waist           0.9254        -0.3777        -0.0310
 pulse          -0.3328         0.0415         0.9421

Correlations Between the WITH Variables and Their Canonical Variables

                  EXER1          EXER2          EXER3

 chins          -0.7276         0.2370        -0.6438
 situps         -0.8177         0.5730         0.0544
 jumps          -0.1622         0.9586        -0.2339
```

From this we see that PHYS1 is highly correlated with waist measurements and that EXER1 is



related to situps and chinups. Next, we look at the cross correlations of the original VAR variables and WITH canonical variables and visa versa. For example, waist measurements are highly correlated with the exercise canonical variable EXER1. This variable was mainly related to situps. This makes sense as the bigger a persons waist, the fewer situps they can perform (note that situps is negatively correlated with EXER1). Likewise, the exercise variables situps and chinups are correlated with PHYS1, which is dominated by waist measurements. Therefore, an overall conclusion can be made that the canonical correlation of 0.7956 is mainly due to relationships between waist measurements and the number of situps and chinups the person can perform. Pulse rate and jumps have less impact on the relationship.

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

                           EXER1          EXER2          EXER3

                 weight    0.4938        -0.1549        -0.0098
                 waist     0.7363        -0.0757        -0.0022
                 pulse    -0.2648         0.0083         0.0684

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

                           PHYS1          PHYS2          PHYS3

                 chins    -0.5789         0.0475        -0.0467
                 situps   -0.6506         0.1149         0.0040
                 jumps    -0.1290         0.1923        -0.0170
```