Multivariate

With MATLAB

Professor Dr. Taha Hussein Ali

2020

Part I

Department of Statistics and Informatics, College of Administration and Economics Salahaddin University, Erbil, Kurdistan Region, Iraq

Multivariate With MATLAB

Prof. Dr. Taha Hussein Ali Alzubaydi

Department of Statistics and Informatics, College of Administration and Economics, Salahaddin University, Erbil, Kurdistan Region, Iraq, 2020

Introduction

This book has been prepared for the beginners to help them understand basic to advanced functionality of MATLAB. After completing this chapter 1 (Which included an explanation of the Matlab language) you will find yourself at a moderate level of expertise in using MATLAB from where you can take yourself to next levels.

On other side, I have long been fascinated by the interplay of variables in multivariate data and by the challenge of unraveling the effect of each variable. My continuing objective has been to present the power and utility of multivariate analysis in a highly readable format.

Practitioners and researchers in all applied disciplines often measure several variables on each subject or experimental unit. In some cases, it may be productive to isolate each variable in a system and study it separately. Typically, however, the variables are not only correlated with each other, but each variable is influenced by the other variables as it affects a test statistic or descriptive statistic. Thus, in many instances, the variables are intertwined in such a way that when analyzed individually they yield little information about the system. Using multivariate analysis, the variables can be examined simultaneously in order to access the key features of the process that produced them. The multivariate approach enables us to (1) explore the joint performance of the variables and (2) determine the effect of each variable in the presence of the others.

Multivariate analysis provides both descriptive and inferential procedures—we can search for patterns in the data or test hypotheses about patterns of a priori interest. With multivariate descriptive techniques, we can peer beneath the tangled web of variables on the surface and extract the essence of the system. Multivariate inferential procedures include hypothesis tests that (1) process any number of variables without inflating the Type I error rate and (2) allow for whatever intercorrelations the variables possess. A wide variety of multivariate descriptive and inferential procedures is readily accessible in statistical software packages.

CONTENTS

Chapter one

Introduction to MATLAB

$$S_2 = \begin{pmatrix} 9.1361 & 7.5494 & 4.8639 & 4.1512 \\ 7.5494 & 18.604 & 10.225 & 5.4456 \\ 4.8639 & 10.225 & 30.039 & 13.494 \\ 4.1512 & 5.4456 & 13.494 & 27.996 \end{pmatrix}$$

Table 2.3: Four Psychological Test Scores on 32 Males and 32 Females

Males				Females			
y_1	y_2	y_3	y_4	y_1	y_2	y_3	y_4
15	17	24	14	13	14	12	21
17	15	32	26	14	12	14	26
15	14	29	23	12	19	21	21
13	12	10	16	12	13	10	16
20	17	26	28	11	20	16	16
15	21	26	21	12	9	14	18
15	13	26	22	10	13	18	24
13	5	22	22	10	8	13	23
14	7	30	17	12	20	19	23
17	15	30	27	11	10	11	27
17	17	26	20	12	18	25	25
17	20	28	24	14	18	13	26
15	15	29	24	14	10	25	28
18	19	32	28	13	16	8	14
18	18	31	27	14	8	13	25
15	14	26	21	13	16	23	28
18	17	33	26	16	21	26	26
10	14	19	17	14	17	14	14
18	21	30	29	16	16	15	23
18	21	34	26	13	16	23	24
13	17	30	24	2	6	16	21
16	16	16	16	14	16	22	26
11	15	25	23	17	17	22	28
16	13	26	16	16	13	16	14
16	13	23	21	15	14	20	26
18	18	34	24	12	10	12	9
16	15	28	27	14	17	24	23
15	16	29	24	13	15	18	20
18	19	32	23	11	16	18	28
18	16	33	23	7	7	19	18
17	20	21	21	12	15	7	28
19	19	30	28	6	5	6	13

The sample covariance matrices do not appear to indicate a disparity in the population covariance matrices. (A significance test to check this assumption is carried out in Example 4.3.2, and the hypothesis $H_0 : \Sigma_1 = \Sigma_2$ is not rejected.) The pooled covariance matrix is

$$S_{pl} = \frac{1}{32+32-2}[(32-1)S_1 + (32-1)S_2]$$

$$= \begin{pmatrix} 7.164 & 6.047 & 5.693 & 4.701 \\ 6.047 & 15.89 & 8.492 & 5.856 \\ 5.693 & 8.492 & 29.36 & 13.98 \\ 4.701 & 5.856 & 13.98 & 22.32 \end{pmatrix}$$

By (2.9), we obtain

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(\bar{y}_1 - \bar{y}_2)' S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2) = 97.601$$

From interpolation in Table A.7, we obtain $T^2_{0.01,4,62} = 15.363$, and we therefore reject $H_0 : \mu_1 = \mu_2$. See Example 2.5 for a discussion of which variables contribute most to separation of the two groups.

Example 2.4.2[The program name mt4.m]	Applications using MATLAB

```
clc
clear all
data=[15 17 24 14 13 14 12 21;17 15 32 26 14 12 14 26; 15 14 29 23 12 19 21
21;13 12 10 16 12 13 10 16;20 17 26 28 11 20 16 16;15 21 26 21 12 9 14 18;
15 13 26 22 10 13 18 24;13 5 22 22 10 8 13 23;14 7 30 17 12 20 19 23;17 15 30 27
11 10 11 27;17 17 26 20 12 18 25 25;17 20 28 24 14 18 13 26;15 15 29 24 14 10
25 28; 18 19 32 28 13 16 8 14;18 18 31 27 14 8 13 25;15 14 26 21 13 16 23 28;
18 17 33 26 16 21 26 26;10 14 19 17 14 17 14 14;18 21 30 29 16 16 15 23;
18 21 34 26 13 16 23 24;13 17 30 24 2 6 16 21;16 16 16 16 14 16 22 26;
11 15 25 23 17 17 22 28;16 13 26 16 16 13 16 14;16 13 23 21 15 14 20 26;
18 18 34 24 12 10 12 9;16 15 28 27 14 17 24 23;15 16 29 24 13 15 18 20;
18 19 32 23 11 16 18 28;18 16 33 23 7 7 19 18;17 20 21 21 12 15 7 28;
19 19 30 28 6 5 6 13];y11=data(:,1);y12=data(:,2);y13=data(:,3);
y14=data(:,4);y21=data(:,5);y22=data(:,6);y23=data(:,7);
y24=data(:,8);y24=data(:,8);m1=mean(data(:,1:4))',m2=mean(data(:,5:8))'
S1=cov(data(:,1:4))',S2=cov(data(:,5:8))',n1=length(y11);n2=length(y21);
Spl=(1/(n1+n2-2))*((n1-1)*S1+(n2-1)*S2)
Ts=((n1*n2)/(n1+n2))*(m1-m2)'*inv(Spl)*(m1-m2)
alfa=0.01;p=4;v=n1+n2-p-1;
tabTs=(finv(1-alfa,p,v)*(n1+n2-2)*p)/v
```

Ans.

```
m1 =
        15.969
        15.906
        27.188
         22.75
m2 =
        12.344
        13.906
        16.656
        21.938
S1 =
        5.1925        4.5454        6.5222          5.25
        4.5454        13.184        6.7601        6.2661
        6.5222        6.7601        28.673        14.468
          5.25        6.2661        14.468        16.645
S2 =
        9.1361        7.5494        4.8639        4.1512
        7.5494        18.604        10.225        5.4456
        4.8639        10.225        30.039        13.494
        4.1512        5.4456        13.494        27.996
Spl =
        7.1643         6.0474          5.693         4.7006
        6.0474         15.894         8.4924         5.8558
         5.693         8.4924         29.356         13.981
        4.7006         5.8558         13.981         22.321
Ts =
        97.601
tabTs =
        15.363
```

2.4.3: Likelihood Ratio Tests

The maximum likelihood approach to estimation was introduced ; the likelihood function is the joint density of $y_1, y_2, \cdots, y_n$. The values of the parameters that maximize the likelihood function are the maximum likelihood estimators.

The *likelihood ratio* method of test construction uses the ratio of the maximum value of the likelihood function assuming H_0 is true to the maximum under H_1, which is essentially unrestricted. Likelihood ratio tests usually have good power and sometimes have optimum power over a wide class of alternatives.

When applied to multivariate normal samples and $H_0 : \mu_1 = \mu_2$, the likelihood ratio approach leads directly to Hotelling's T^2-test in (2.9). Similarly, in the one sample case, the T^2-statistic in (2.5) is the likelihood ratio test. Thus the T^2-test, which we introduced rather informally, is the best test according to certain criteria.

2.5: TESTS ON INDIVIDUAL VARIABLES CONDITIONAL ON REJECTION OF H_0 BY THE T^2-TEST

If the hypothesis $H_0 : \mu_1 = \mu_2$ is rejected, the implication is that $\mu_{1j} \neq \mu_{2j}$ for at least one $j = 1,2,\cdots,p$. But there is no guarantee that $H_0 : \mu_{1j} = \mu_{2j}$ will be rejected for some *j* by a univariate test. However, if we consider a linear combination of the variables, $z = \mathrm{a}'\mathrm{y}$, then there is at least one coefficient vector **a** for which

$$t(\mathrm{a}) = \frac{\bar{z}_1 - \bar{z}_2}{\sqrt{(1/n_1 + 1/n_2)s_z^2}} \tag{2.12}$$

will reject the corresponding hypothesis $H_0 : \mu_{z_1} = \mu_{z_2}$ or $H_0 : \mathrm{a}'\mu_1 = \mathrm{a}'\mu_2$. $\bar{z}_1 = \mathrm{a}'\bar{\mathrm{y}}_1$ and $\bar{z}_2 = \mathrm{a}'\bar{\mathrm{y}}_2$, the variance estimator s_z^2 is the pooled estimator $\mathrm{a}'\mathrm{S}_{\mathrm{pl}}\mathrm{a}$. Thus (2.12) can be written as

$$t(\mathrm{a}) = \frac{\mathrm{a}'\bar{\mathrm{y}}_1 - \mathrm{a}'\bar{\mathrm{y}}_2}{\sqrt{[(n_1 + n_2)/n_1 n_2]\mathrm{a}'\mathrm{S}_{\mathrm{pl}}\mathrm{a}}} \tag{2.13}$$

Since $t(\mathrm{a})$ can be negative, we work with $t^2(\mathrm{a})$. The linear function $z = \mathrm{a}'\mathrm{y}$ is a projection of **y** onto a line through the origin. We seek the line (direction) on which the difference $\bar{\mathrm{y}}_1 - \bar{\mathrm{y}}_2$ is maximized when projected. The projected difference $\mathrm{a}'(\bar{\mathrm{y}}_1 - \bar{\mathrm{y}}_2)$ [standardized by $\mathrm{a}'\mathrm{S}_{\mathrm{pl}}\mathrm{a}$ as in (2.13)] will be less in any other direction than that parallel to the line joining $\bar{\mathrm{y}}_1$ and $\bar{\mathrm{y}}_2$. The value of **a** that projects onto this line, or, equivalently, maximizes $t^2(\mathrm{a})$ in (2.13), is (any multiple of)

$$\mathrm{a} = \mathrm{S}_{\mathrm{pl}}^{-1}(\bar{\mathrm{y}}_1 - \bar{\mathrm{y}}_2) \tag{2.14}$$

Since a in (2.14) projects $\bar{\mathrm{y}}_1 - \bar{\mathrm{y}}_2$ onto a line parallel to the line joining $\bar{\mathrm{y}}_1$ and $\bar{\mathrm{y}}_2$, we would expect that $t^2(\mathrm{a}) = T^2$, and this is indeed the case (see Problem 2.3).

When $\mathrm{a} = \mathrm{S}_{\mathrm{pl}}^{-1}(\bar{\mathrm{y}}_1 - \bar{\mathrm{y}}_2)$, then $z = \mathrm{a}'\mathrm{y}$ is called the discriminant function. Sometimes the vector **a** itself in (2.14) is loosely referred to as the discriminant function.

If $H_0: \mu_1 = \mu_2$ is rejected by T^2 in (2.9), the discriminant function $\mathrm{a}'\mathrm{y}$ will lead to rejection of $H_0: \mathrm{a}'\mu_1 = \mathrm{a}'\mu_2$ using (2.13), with $\mathrm{a} = \mathrm{S}_{\mathrm{pl}}^{-1}(\bar{\mathrm{y}}_1 - \bar{\mathrm{y}}_2)$. We can then examine each a_j in **a** for an indication of the contribution of the corresponding y_j to rejection of H_0. This follow-up examination of each a_j should be done only if $H_0: \mu_1 = \mu_2$ is rejected by T^2. The discriminant function will appear again in Section 2.6.2 and in Chapters 5 and 6.

We list these and other procedures that could be used to check each variable following rejection of H_0 by a two-sample T^2-test:

1. Univariate t-tests, one for each variable,

$$t_j = \frac{\bar{y}_{1j} - \bar{y}_{2j}}{\sqrt{[(n_1 + n_2)/n_1 n_2]s_{jj}}}, \quad j = 1,2,\cdots,p \tag{2.15}$$

   where s_{jj} is the jth diagonal element of S_{pl}. Reject $H_0: \mu_{1j} = \mu_{2j}$ if $|t_j| > t_{\alpha/2, n_1+n_2-2}$. For confidence intervals on $\mu_{1j} - \mu_{2j}$, see Rencher (1998, Section 3.6).

2. To adjust the α-level resulting from performing the p tests in (2.15), we could use a Bonferroni critical value $t_{\alpha/2p, n_1+n_2-2}$ for (2.15) (Bonferroni 1936). A critical value $t_{\alpha/2p}$ is much greater than the corresponding $t_{\alpha/2}$, and the resulting overall α-level is conservative. Bonferroni critical values $t_{\alpha/2p,v}$ are given in Table A.8, from Bailey (1977).

3. Another critical value that could be used with (2.15) is T_{α,p,n_1+n_2-2}, where T_α is the square root of T_α^2 from Table A.7; that is, $T_{\alpha,p,n_1+n_2-2} = \sqrt{T^2_{\alpha,p,n_1+n_2-2}}$. This allows for all p variables to be tested as well as all possible linear combinations, as in (2.13), even linear combinations chosen after seeing the data. Consequently, the use of T_α is even more conservative than using $t_{\alpha/2p}$; that is, $T_{\alpha,p,n_1+n_2-2} > t_{\alpha/2p,n_1+n_2-2}$.

4. Partial *F*- or *t*-tests [test of each variable adjusted for the other variables; see (2.32) in Section 2.8].
5. Standardized discriminant function coefficients (see Section 5.5)
6. Correlations between the variables and the discriminant function (see Section 5.7.3).
7. Stepwise discriminant analysis (see Section 5.9).

The first three methods are univariate approaches that do not use covariances or correlations among the variables in the computation of the test statistic. The last four methods are multivariate in the sense that the correlation structure is explicitly taken into account in the computation.

Method 6, involving the correlation between each variable and the discriminant function, is recommended in many texts and software packages. However, Rencher (1988) has shown that these correlations are proportional to individual *t*- or *F*-tests (see Section 5.7.3). Thus this method is equivalent to method 1 and is a univariate rather than a multivariate approach. Method 7 is often used to identify a subset of important variables or even to rank the variables according to order of entry. But Rencher and Larson (1980) have shown that stepwise methods have a high risk of selecting spurious variables, unless the sample size is very large.

We now consider the univariate procedures 1, 2, and 3. The probability of rejecting one or more of the p univariate tests when H_0 is true is called the overall α or *experimentwise* error rate. If we do univariate tests only, with no T^2-test, then the tests based on $t_{\alpha/2p}$ and T_α in procedures 2 and 3 are conservative (overall α too low), and tests based on $t_{\alpha/2}$ in procedure 1 are liberal (overall α too high). However, when these tests are carried out only after rejection by the T^2-test (such tests are sometimes called protected tests), the experimentwise error rates change. Obviously the tests will reject less often (under H_0) if they are carried out only if T^2 rejects. Thus the tests using $t_{\alpha/2p}$ and T_α become even more conservative, and the test using $t_{\alpha/2}$ becomes more acceptable.

Hummel and Sligo (1971) studied the experimentwise error rate for univariate *t*-tests following rejection of H_0 by the T^2-test (protected tests). Using α = 0.05, they found that using $t_{\alpha/2}$ for a critical value yields an overall α acceptably close to the nominal 0.05. In fact, it is slightly conservative, making this the preferred univariate test (within the limits of their study). They also compared this procedure with that of performing univariate tests without a prior T^2-test (unprotected tests). For this case, the overall α is too high, as expected. Table 2.2 gives an excerpt of Hummel and Sligo's results. The sample size is for each of the two samples; the r^2 in common is for every pair of variables.

Hummel and Sligo therefore recommended performing the multivariate T^2-test followed by univariate *t*-tests. This procedure appears to have the desired overall α level and will clearly have better power than tests using T_α or $t_{\alpha/2p}$ as a critical value. Table 2.2 also highlights the importance of using univariate *t*-tests only if the multivariate T^2-test is significant. The inflated α's resulting if *t*-tests are used without regard to the outcome of the T^2-test are clearly evident. Thus among the three univariate procedures (procedures 1, 2, and 3), the first appears to be preferred.

Among the multivariate approaches (procedures 4, 5, and 7), we prefer the fifth procedure, which compares the (absolute value of) coefficients in the discriminant function to find the effect of each variable in separating the two groups of observations. These coefficients will often tell a different story from the univariate tests, because the univariate tests do not take into account the correlations among the variables or the effect of each variable on T^2 in the presence of the other variables. A variable will typically have a different effect in the presence of other variables than it has by itself. In the discriminant function $z = \mathrm{a}'\mathrm{y} = a_1y_1 + a_2y_2 + \cdots + a_py_p$, where $\mathrm{a} = \mathrm{S}_{\mathrm{pl}}^{-1}(\bar{\mathrm{y}}_1 - \bar{\mathrm{y}}_2)$, the coefficients $a_1, a_2, \cdots, a_p$ indicate the relative importance of the variables in a multivariate context, something the univariate *t*-tests cannot do. If the variables are not commensurate (similar in scale and variance), the coefficients should be standardized, as in Section 5.5; this allows for more valid comparisons among the variables. Rencher and Scott (1990) provided a decomposition of the information in the standardized discriminant

function coefficients. For a detailed analysis of the effect of each variable in the presence of the other variables, see Rencher (1993; 1998, Sections 3.3.5 and 3.5.3).

Table 2.4: Comparison of Experimentwise Error Rates (Nominal α= .05)

Sample Size	Number of Variables	Common r^2 .10	.30	.50	.70
		Univariate Tests Only[a]			
10	3	.145	.112	.114	.077
10	6	.267	.190	.178	.111
10	9	.348	.247	.209	.129
30	3	.115	.119	.117	.085
30	6	.225	.200	.176	.115
30	9	.296	.263	.223	.140
50	3	.138	.124	.102	.083
50	6	.230	.190	.160	.115
50	9	.324	.258	.208	.146
		Multivariate Test Followed by Univariate Tests[b]			
10	3	.044	.029	.035	.022
10	6	.046	.029	.030	.017
10	9	.050	.026	.025	.018
30	3	.037	.044	.029	.025
30	6	.037	.037	.032	.021
30	9	.042	.042	.030	.021
50	3	.038	.041	.033	.028
50	6	.037	.039	.028	.027
50	9	.036	.038	.026	.020

[a] Ignoring multivariate tests.
[b] Carried out only if multivariate test rejects.

Example 2.5: For the psychological data in Table 2.3, we obtained $\bar{y}_1$, $\bar{y}_2$, and S_{pl} in Example 2.4.2. The discriminant function coefficient vector is obtained from (2.14) as

$$a = S_{pl}^{-1}(\bar{y}_1 - \bar{y}_2) = \begin{pmatrix} 0.5104 \\ -0.2033 \\ 0.4660 \\ -0.3097 \end{pmatrix}$$

Thus the linear combination that best separates the two groups is

$$\mathrm{a}'\mathrm{y} = 0.5104y_1 - 0.2033y_2 + 0.4660y_3 - 0.3097y_4$$

in which y_1 and y_3 appear to contribute most to separation of the two groups. (After standardization, the relative contribution of the variables changes somewhat; see the answer to Problem 5.7 in Appendix B.)

Example 2.5[The program name mt5.m]	Applications using MATLAB

```
clc
clear all
data=[15 17 24 14 13 14 12 21;17 15 32 26 14 12 14 26; 15 14 29 23
12 19 21 21;13 12 10 16 12 13 10 16;20 17 26 28 11 20 16 16;15 21 26
21 12 9 14 18; 15 13 26 22 10 13 18 24;13 5 22 22 10 8 13 23;14 7 30
17 12 20 19 23;17 15 30 27 11 10 11 27; 17 17 26 20 12 18 25 25;17
20 28 24 14 18 13 26;15 15 29 24 14 10 25 28;18 19 32 28 13 16 8
14;18 18 31 27 14 8 13 25;15 14 26 21 13 16 23 28;18 17 33 26 16 21
26 26;10 14 19 17 14 17 14 14;18 21 30 29 16 16 15 23; 18 21 34 26
13 16 23 24;13 17 30 24 2 6 16 21;16 16 16 16 14 16 22 26;11 15 25
23 17 17 22 28;16 13 26 16 16 13 16 14;16 13 23 21 15 14 20 26;18 18
34 24 12 10 12 9;16 15 28 27 14 17 24 23;15 16 29 24 13 15 18 20;
18 19 32 23 11 16 18 28;18 16 33 23 7 7 19 18;17 20 21 21 12 15 7 28;
19 19 30 28 6 5 6 13];y11=data(:,1);y12=data(:,2);y13=data(:,3);
y14=data(:,4);y21=data(:,5);y22=data(:,6);y23=data(:,7);
y24=data(:,8);y24=data(:,8);
m1=mean(data(:,1:4))';
m2=mean(data(:,5:8))';
S1=cov(data(:,1:4))';S2=cov(data(:,5:8))';
n1=length(y11);n2=length(y21);
Spl=(1/(n1+n2-2))*((n1-1)*S1+(n2-1)*S2);
a=inv(Spl)*(m1-m2),
```

Ans.

```
a =
    0.51042
   -0.20329
    0.46604
   -0.30967
```

2.6: COMPUTATION OF T^2

If one has a program available with matrix manipulation capability, it is a simple matter to compute T^2 using (2.9). However, this approach is somewhat cumbersome for those not accustomed to the use of such a programming language, and many would prefer a more automated procedure. But very few general-purpose statistical programs provide for direct calculation of the two-sample T^2-statistic, perhaps because it is so easy to obtain from other procedures. We will discuss two types of widely available procedures that can be used to compute T^2.

2.6.1: Obtaining T^2 from a MANOVA Program

Multivariate analysis of variance (MANOVA) is discussed in Chapter 3, and the reader may wish to return to the present section after becoming familiar with that material. One-way MANOVA involves a comparison of mean vectors from several samples. Typically, the number of samples is three or more, but the procedure will also accommodate two samples. The two-sample T^2 test is thus a special case of MANOVA.

Four common test statistics are defined in Section 3.1: Wilks' Λ, the Lawley–Hotelling $\mathrm{U}^{(s)}$, Pillai's $\mathrm{V}^{(s)}$, and Roy's largest root θ. Without concerning ourselves here with how these are defined or calculated, we show how to use each to obtain the two-sample T^2:

$$T^2 = (n_1 + n_2 - 2)\frac{1-\Lambda}{\Lambda} \tag{2.16}$$

$$T^2 = (n_1 + n_2 - 2)\mathrm{U}^{(s)} \tag{2.17}$$

$$T^2 = (n_1 + n_2 - 2)\frac{\mathrm{V}^{(s)}}{1-\mathrm{V}^{(s)}} \tag{2.18}$$

$$T^2 = (n_1 + n_2 - 2)\frac{\theta}{1-\theta} \tag{2.19}$$

(For the special case of two groups, $\mathrm{V}^{(s)} = \theta$.) These relationships are demonstrated in Section 3.1.7. If the MANOVA program gives eigenvectors of $\mathrm{E}^{-1}\mathrm{H}$ (**E** and **H** are defined in Section 3.1.2), the eigenvector corresponding to the largest eigenvalue will be equal to (**a** constant multiple of) the discriminant function $\mathrm{S}_{\mathrm{pl}}^{-1}(\bar{\mathrm{y}}_1 - \bar{\mathrm{y}}_2)$.

2.6.2: Obtaining T^2 from Multiple Regression

In this section, the *y*'s become independent variables in a regression model. For each observation vector y_{1i} and y_{2i} in a two-sample T^2, define a "dummy" group variable as

$$w_i = \frac{n_2}{n_1+n_2} \text{ for each of } \mathrm{y}_{11}, \mathrm{y}_{12}, \cdots, \mathrm{y}_{1n_1} \text{ in sample 1}$$

$$= -\frac{n_1}{n_1+n_2} \text{ for each of } \mathrm{y}_{21}, \mathrm{y}_{22}, \cdots, \mathrm{y}_{2n_2} \text{ in sample 2}$$

Then $\overline{w} = 0$ for all n_1+n_2 observations. The prediction equation for the regression of *w* on the *y*'s can be written as

$$\hat{w}_i = b_0 + b_1 y_{i1} + b_2 y_{i2} + \cdots + b_p y_{ip}$$

where *i* ranges over all n_1+n_2 observations and the least squares estimate b_0 is [see (7.15)]

$$b_0 = \overline{w} - b_1\overline{y}_1 - b_2\overline{y}_2 - \cdots - b_p\overline{y}_p$$

Substituting this into the regression equation, we obtain

$$\hat{w}_i = \overline{w} + b_1(y_{i1} - \overline{y}_1) + b_2(y_{i2} - \overline{y}_2) + \cdots + b_p(y_{ip} - \overline{y}_p)$$

$$= b_1(y_{i1} - \overline{y}_1) + b_2(y_{i2} - \overline{y}_2) + \cdots + b_p(y_{ip} - \overline{y}_p) \text{ (since } \overline{w} = 0)$$

Let $\mathrm{b}' = (b_1, b_2, \cdots, b_p)$ be the vector of regression coefficients and R^2 be the squared multiple correlation. Then we have the following relationships:

$$T^2 = (n_1+n_2-2)\frac{R^2}{1-R^2} \qquad (2.20)$$

$$\mathrm{a} = \mathrm{S}_{\mathrm{pl}}^{-1}(\overline{\mathrm{y}}_1 - \overline{\mathrm{y}}_2) = \frac{n_1+n_2}{n_1 n_2}(n_1+n_2-2+T^2)\mathrm{b} \qquad (2.21)$$

Thus with ordinary multiple regression, one can easily obtain T^2 and the discriminant function $\mathrm{S}_{\mathrm{pl}}^{-1}(\overline{\mathrm{y}}_1 - \overline{\mathrm{y}}_2)$. We simply define w_i as above for each of the n_1+n_2 observations, regress the *w*'s on the *y*'s, and use the resulting R^2 in (2.20). For **b**, delete the intercept from the regression coefficients for use in (2.21). Actually, since only the relative values of the elements of $\mathrm{a} = \mathrm{S}_{\mathrm{pl}}^{-1}(\overline{\mathrm{y}}_1 - \overline{\mathrm{y}}_2)$ are of interest, it is not necessary to

convert from **b** to a in (2.21). We can use **b** directly or standardize the values $b_1, b_2, \cdots, b_p$ as in Section 5.5.

Example 2.6.2: We illustrate the regression approach to computation of T^2 using the psychological data in Table 2.3. We set $w = n_2/(n_1+n_2)$ = 32/64 = 1/2 for each observation in group 1 (males) and equal to $-n_1/(n_1+n_2) = -1/2$ in the second group (females). When *w* is regressed on the 64 *y*'s, we obtain

$$\begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} -0.75114 \\ 0.05117 \\ -0.02038 \\ 0.046721 \\ -0.031044 \end{pmatrix}, \quad R^2 = 0.61153$$

By (2.20), $T^2 = (n_1+n_2-2)\dfrac{R^2}{1-R^2} = (32+32-2)\dfrac{0.61153}{1-0.61153} = 97.601$

as was obtained before in Example 2.4.2. Note that $\mathrm{b}' = (b_1, b_2, b_3, b_4)$ = (.05117,−.02038, .046721,−.031044), with the intercept deleted, is proportional to the discriminant function coefficient vector a from Example 2.5, as we would expect from (2.21).

Example 2.5[The program name mt6.m]	Applications using MATLAB

```
clc
clear all
data=[15 17 24 14 13 14 12 21;17 15 32 26 14 12 14 26; 15 14 29 23 12 19 21
21;13 12 10 16 12 13 10 16;20 17 26 28 11 20 16 16;15 21 26 21 12 9 14 18; 15 13
26 22 10 13 18 24;13 5 22 22 10 8 13 23;14 7 30 17 12 20 19 23;17 15 30 27 11 10
11 27; 17 17 26 20 12 18 25 25;17 20 28 24 14 18 13 26;15 15 29 24 14 10 25 28;
18 19 32 28 13 16 8 14;18 18 31 27 14 8 13 25;15 14 26 21 13 16 23 28; 18 17 33
26 16 21 26 26;10 14 19 17 14 17 14 14;18 21 30 29 16 16 15 23; 18 21 34 26 13
16 23 24;13 17 30 24 2 6 16 21;16 16 16 16 14 16 22 26;11 15 25 23 17 17 22
28;16 13 26 16 16 13 16 14;16 13 23 21 15 14 20 26; 18 18 34 24 12 10 12 9;16 15
28 27 14 17 24 23;15 16 29 24 13 15 18 20; 18 19 32 23 11 16 18 28;18 16 33 23 7
7 19 18;17 20 21 21 12 15 7 28; 19 19 30 28 6 5 6 13]; y11=data(:,1); y12=
data(:,2); y13=data(:,3); y14=data(:,4);y21=data(:,5);y22=data(:,6); y23=data(:,7);
y24=data(:,8);n1=length(y11);n2=length(y21); w1=n2/(n1+n2);w2=-n1/(n1+n2);
data1=[y11 y12 y13 y14];data2=[y21 y22 y23 y24]; w=[ones(size(y11))*w1;
ones(size(y21))*w2]; x=[data1;data2]; n=n1+n2;x=[ones(n,1) x];b=x\w, W=x*b;
e=w-W; Rs=corr(w,W)^2 ,Ts=(n1+n2-2)*Rs/(1-Rs)
```

Ans.

```
b =
      -0.75114
       0.05117
      -0.02038
       0.046721
     -0.031044
Rs =
       0.61153
Ts =
        97.601
```

2.7: PAIRED OBSERVATIONS TEST

As usual, we begin with the univariate case to set the stage for the multivariate presentation.

2.7.1: Univariate Case

Suppose two samples are not independent because there exists a natural pairing between the ith observation y_i in the first sample and the ith observation x_i in the second sample for all *i*, as, for example, when a treatment is applied twice to the same individual or when subjects are matched according to some criterion, such as IQ or family background. With such pairing, the samples are often referred to as *paired observations* or *matched pairs*. The two samples thus obtained are correlated, and the two-sample test statistic in (2.9) is not appropriate because the samples must be independent in order for (2.9) to have a *t*-distribution. [The two-sample test in (2.9) is somewhat robust to heterogeneity of variances and to lack of normality but not to dependence.] We reduce the two samples to one by working with the differences between the paired observations, as in the following layout for two treatments applied to the same subject:

Pair Number	Treatment 1	Treatment 2	Difference $d_i = y_i - x_i$
1	y_1	x_1	d_1
2	y_2	x_2	d_2
⋮	⋮	⋮	⋮
n	y_n	x_n	d_n

To obtain a *t*-test, it is not sufficient to assume individual normality for each of *y* and *x*. To allow for the covariance between *y* and *x*, we need the additional assumption that *y* and *x* have a bivariate normal distribution with

$$\mu = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix} \quad , \quad \Sigma = \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix}$$

where $d_i = y_i - x_i$ is $N\left(\mu_y - \mu_x, \sigma_d^2\right)$, where $\sigma_d^2 = \sigma_y^2 - 2\sigma_{yx} + \sigma_x^2$. From $d_1, d_2, \cdots, d_n$ we calculate

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \text{ and } s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2$$

To test $H_0 : \mu_y = \mu_x$, that is, $H_0 : \mu_d = 0$, we use the one-sample statistic

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \tag{2.22}$$

which is distributed as t_{n-1} if H_0 is true. We reject H_0 in favor of $H_1 : \mu_d \neq 0$ if $|t| > t_{\alpha/2, n-1}$. It is not necessary to assume $\sigma_y^2 = \sigma_x^2$ because there are no restrictions on Σ.

This test has only $n - 1$ degrees of freedom compared with $2(n - 1)$ for the two independent-sample *t*-test (2.8). In general, the pairing reduces the within-sample variation s_d and thereby increases the power.

If we mistakenly treated the two samples as independent and used (2.8) with $n_1 = n_2 = n$, we would have

$$t = \frac{\bar{y} - \bar{x}}{s_{\text{pl}}\sqrt{2/n}} = \frac{\bar{y} - \bar{x}}{\sqrt{2s_{\text{pl}}^2/n}}$$

However,

$$E\left(\frac{2s_{\text{pl}}^2}{n}\right) = 2E\left[\frac{(n-1)s_y^2 + (n-1)s_x^2}{(n+n-2)n}\right] = \frac{\sigma_y^2 + \sigma_x^2}{n}$$

whereas $Var\left(\bar{y} - \bar{x}\right) = \left(\sigma_y^2 + \sigma_x^2 - 2\sigma_{yx}\right)/n$. Thus if the test statistic for independent samples (2.8) is used for paired data, it does not have a *t*-distribution and, in fact, underestimates the true average *t*-value

(assuming H_0 is false), since $\sigma_y^2 + \sigma_x^2 > \sigma_y^2 + \sigma_x^2 - 2\sigma_{yx}$ if $\sigma_{yx} > 0$, which would be typical in this situation. One could therefore use

$$t = \frac{\bar{y} - \bar{x}}{\sqrt{\left(s_y^2 + s_x^2 - 2s_{yx}\right)/n}} \tag{2.23}$$

but $t = \sqrt{n}\bar{d}/s_d$ in (2.22) is equal to it and somewhat simpler to use.

2.7.2: Multivariate Case

Here we assume the same natural pairing of sampling units as in the univariate case, but we measure *p* variables on each sampling unit. Thus y_i from the first sample is paired with x_i from the second sample, *i* = 1, 2, . . . , *n*. In terms of two treatments applied to each sampling unit, this situation is as follows:

Pair Number	Treatment 1	Treatment 2	Difference $\mathbf{d}_i = \mathbf{y}_i - \mathbf{x}_i$
1	$\mathbf{y}_1$	$\mathbf{x}_1$	$\mathbf{d}_1$
2	$\mathbf{y}_2$	$\mathbf{x}_2$	$\mathbf{d}_2$
⋮	⋮	⋮	⋮
n	$\mathbf{y}_n$	$\mathbf{x}_n$	$\mathbf{d}_n$

In Section 2.7.1, we made the assumption that *y* and *x* have a bivariate normal distribution, in which *y* and *x* are correlated. Here we assume **y** and **x** are correlated and have a multivariate normal distribution:

$$\begin{pmatrix} \mathrm{y} \\ \mathrm{x} \end{pmatrix} \quad \text{is} \quad N_{2p}\left[\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx} \end{pmatrix}\right]$$

To test $H_0 : \mu_d = 0$, which is equivalent to $H_0 : \mu_y = \mu_x$ since $\mu_d = E(\mathrm{y} - \mathrm{x}) = \mu_y - \mu_x$, we calculate

$$\bar{\mathrm{d}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{d}_i \text{ and } \mathrm{S}_d = \frac{1}{n-1}\sum_{i=1}^{n} (\mathrm{d}_i - \bar{\mathrm{d}})(\mathrm{d}_i - \bar{\mathrm{d}})'$$

We then have

$$T^2 = \bar{\mathrm{d}}'\left(\frac{\mathrm{S}_d}{n}\right)^{-1}\bar{\mathrm{d}} = n\,\bar{\mathrm{d}}'\mathrm{S}_d^{-1}\bar{\mathrm{d}} \tag{2.24}$$

Under H_0, this paired comparison T^2-statistic is distributed as $T^2_{p,n-1}$. We reject H_0 if $T^2 > T^2_{\alpha,p,n-1}$. Note that S_d estimates $Cov(\mathrm{y} - \mathrm{x}) =$