Objective of Multivariate Analysis of variance (MANOVA) is to determine whether the differences on criterion or dependent variables among groups are “reliable” or statistically significant. Variable(s) that create group membership are called control or independent variable(s).

• Application

There are two broad class of situation where MANOVA may be used.

• Experiment: In this situation, control or independent variable(s) within and / or between experimental unit is changed and measures of difference(s) on criterion variables is used to demonstrate the cause (experimental treatment) and the effect. For example, higher number of brand advertising exposures increases the likelihood of higher preference towards brand and purchase intentions.

• Groups with “natural” belonging: In this situation, groups with natural belongings are used to demonstrate reliable differences exists between groups. For example, attitude towards consumer’s willingness to search for bargains and consumer price responsiveness differ significantly across females and males.

• Data Requirement

• Criterion variables must have interval or ratio scale measurement properties.

• Control variable(s) may be either nominal or ordinal. If control variable is interval scale and it is desired to understand the impact of each level of control variable on the criterion variable then, control variable must be converted to ordinal scale.

In following notes, we will present situation involving

• Two Group Comparison,
• Matched Comparison,
• Multiple Group Comparison, and
• Multiple Factorial Comparison.

The main idea behind MANOVA is to compare two measures of variability. One measure of variability is a result of differences among values of dependent variables within each group and the second measure of variability is a result of differences between group means. Unlike ANOVA, these comparisons are made simultaneously over several variables.
Two Group Comparison

Two groups, one treatment group and another control group are compared on multiple dependent variables. That is, test units (individuals, stores etc.) assigned randomly to two groups and we measure two or more dependent variables.

- **Experimental Situation**

Consider an experiment where $R$ is used to denote random assignment of treatment and $X$ is used to denote treatment application to test units. Moreover we take measurement on $n_1$ and $n_2$ test units for treatment and control condition respectively. That is, we observe two or more responses per test unit such as (if we have two response variables) $y_{1-11}, y_{2-11}, y_{1-12}, y_{2-12} \cdots y_{1-1n_1}, y_{2-1n_1}$ and $y_{1-21}, y_{2-21}, y_{1-22}, y_{2-22}, \cdots y_{1-2n_2}, y_{2-2n_2}$ dependent variable values for treatment (denoted by 1 after dash “-” sign in subscript) and control (denoted by 2 after dash sign) groups. Here number before dash sign are used to denote variable number.

- **Hypothesis Tested**

\[
H_0 : \left( \bar{y}_{1-1}, \bar{y}_{2-1} \right) = \left( \bar{y}_{1-2}, \bar{y}_{2-2} \right)
\]

and

\[
H_A : \left( \bar{y}_{1-1}, \bar{y}_{2-1} \right) \neq \left( \bar{y}_{1-2}, \bar{y}_{2-2} \right).
\]

Note that $\bar{y}_{2-1}$ is the mean for second dependent variable and the first group. Here the first number in subscript denote variable number and the second number reflects group number. For univariate case, we relied on the standard deviation which depended on sums of squares. In multivariate case, we also have to think about covariation between two dependent variables.

- **Illustrative Example**

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>

Suppose we ask nine respondents two questions, attitude towards family and attitude towards church on ten point scale. There are five respondents in group 1 and four in the second group. Note that means for the first group for the first variable is 6 and the second variable is 8. For the second group means are are 3 and 2 for the first and second variable respectively. Intuitively, these measures indicate that two groups are different. We see below statistical reasoning behind this conclusion.
We can write combined matrix and decompose such matrix from each group mean. That is,

\[
\begin{pmatrix}
6 & 7 \\
5 & 9 \\
8 & 6 \\
4 & 9 \\
7 & 9 \\
2 & 3 \\
5 & 1 \\
3 & 1 \\
2 & 3
\end{pmatrix}
= \begin{pmatrix}
6 & 8 \\
6 & 8 \\
6 & 8 \\
6 & 8 \\
6 & 8 \\
3 & 2 \\
3 & 2 \\
3 & 2 \\
3 & 2
\end{pmatrix}
+ \begin{pmatrix}
0 & -1 \\
-1 & 1 \\
2 & -2 \\
-2 & 1 \\
1 & 1 \\
-1 & 1 \\
2 & -1 \\
0 & -1 \\
-1 & 1
\end{pmatrix}
\]

\[
\text{Responses} = \text{Treatment Means} + \text{Error}
\]

Note that by decomposing responses into two or more components, we can better understand source of variation and attribute that source variation to that “factor”. In order to better understand various components of variation, we could square each entry. Following squaring, we could add first five rows of each matrix. These summed entries will be called sums of squares. For each group thus, I will get two sums of squares. In addition, I could also compute product of two numbers in each row and add them (again separating group), I will get sum of cross products. Matrices thus, formed are called sums of squares and cross products (SSCP). For my example, I could construct

\[
\begin{pmatrix}
190 & 234 \\
-32 & 28 \\
42 & 20 \\
20 & 20
\end{pmatrix}
= \begin{pmatrix}
180 & 240 \\
240 & 320 \\
36 & 24 \\
24 & 16
\end{pmatrix}
+ \begin{pmatrix}
10 & -6 \\
-6 & 8 \\
6 & -4 \\
-4 & 4
\end{pmatrix}
\]

Note that entry 190 is based on sums of squares of first variable for the first group or treatment group. That is, \(6^2 + 7^2 + \cdots + 7^2\) and entry 234 is based on sums of cross products of two measured variables for the first group. That is, \(6 \times 7 + 5 \times 9 + \cdots + 7 \times 9\). We could also decompose the mean for variables from the overall sample mean for each variable. This should help us differentiate variation due to treatment groups and overall variation in the sample. The overall mean for the first variable is \(4\frac{2}{7}\) and second variable it is \(5\frac{1}{7}\). Thus, our decomposition
would be

$$
\begin{pmatrix}
6 & 7 \\
5 & 9 \\
8 & 6 \\
4 & 9 \\
7 & 9 \\
\end{pmatrix}
= 
\begin{pmatrix}
4 & \frac{2}{3} & \frac{5}{3} \\
\frac{4}{3} & \frac{5}{3} \\
\frac{4}{3} & \frac{1}{3} \\
\frac{4}{3} & \frac{5}{3} \\
\frac{4}{3} & \frac{5}{3} \\
\end{pmatrix}
+ 
\begin{pmatrix}
1 & \frac{1}{3} & \frac{2}{3} \\
\frac{1}{3} & \frac{2}{3} \\
\frac{1}{3} & \frac{2}{3} \\
\frac{1}{3} & \frac{2}{3} \\
\frac{1}{3} & \frac{2}{3} \\
\end{pmatrix}
+ 
\begin{pmatrix}
0 & -1 \\
-1 & 1 \\
2 & -2 \\
-2 & 1 \\
1 & 1 \\
\end{pmatrix}
$$

Response = Overall Mean + Treatment Mean + Error.

Based on this decomposition, we can also write various sums of squares. In such cases, our decomposition is based on overall sample and that can be written as,

$$
\begin{pmatrix}
232 & 254 \\
254 & 348 \\
\end{pmatrix}
= 
\begin{pmatrix}
196 & 224 \\
224 & 256 \\
\end{pmatrix}
+ 
\begin{pmatrix}
20 & 40 \\
40 & 80 \\
\end{pmatrix}
+ 
\begin{pmatrix}
16 & -10 \\
-10 & 12 \\
\end{pmatrix}.
$$

You may think of these matrices as pooled or between group variation matrices. Moreover, the total sums of squares matrix (one that is left of equal sign) and squared values of overall mean (first matrix on the right hand of equal sign) are less informative and often ignored in analysis. Now let us go back to error variation matrix for the first group. That is,

$$
SS_1 = \begin{pmatrix} 10 & -6 \\ -6 & 8 \end{pmatrix}.
$$

Since we have 5 observations, \(n_1 - 1 = 4\). Thus, if we divide \(SS_1\) by 4, the resulting matrix is called within group variance-covariance matrix. Thus,

$$
S_1 = \begin{pmatrix} 2.5 & -1.5 \\ -1.5 & 2.0 \end{pmatrix}
$$

Similarly, we can construct \(S_2\) by first writing \(SS_2\) matrix. That is,

$$
SS_2 = \begin{pmatrix} 6 & -4 \\ -4 & 4 \end{pmatrix}
$$

and then dividing by 3, we get

$$
S_2 = \begin{pmatrix} 2 & -4/3 \\ -4/3 & 4/3 \end{pmatrix}
$$
Recall that in univariate case, we constructed between group variation measure \((s^2)\), in multivariate case we construct between group variance-covariance matrix and it is equal to

\[
S_{\text{pooled}} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \quad \text{or} \quad \frac{SS_1 + SS_2}{n_1 + n_2 - 2}
\]

For our example, we can write

\[
S_{\text{pooled}} = \frac{1}{n_1 + n_2 - 2} \left[ \begin{pmatrix} 10 & -6 \\ -6 & 8 \end{pmatrix} + \begin{pmatrix} 6 & -4 \\ -4 & 4 \end{pmatrix} \right]
\]

\[
= \frac{1}{5 + 4 - 2} \begin{pmatrix} 16 & -10 \\ -10 & 12 \end{pmatrix}
\]

\[
= \frac{1}{7} \begin{pmatrix} 16 & -10 \\ -10 & 12 \end{pmatrix}
\]

- **Test Statistic**

\[
T^2 = \left( \bar{y}_{1-1} - \bar{y}_{1-2} \quad \bar{y}_{2-1} - \bar{y}_{2-2} \right) \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \left( \begin{pmatrix} 10 & -6 \\ -6 & 8 \end{pmatrix} \right)^{-1} \left( \bar{y}_{1-1} - \bar{y}_{1-2} \quad \bar{y}_{2-1} - \bar{y}_{2-2} \right)
\]

Note that the first and third parenthesis are meant to compute square of differences and middle parenthesis is same as dividing by \(s^2\). Since we squared everything on the right hand side, this test is called \(T^2\) and often referred to it as Hotelling’s \(T\) statistic or Hotelling - Lawley Trace statistic and it is always positive.

- **Decision Rule**

Note that

\[
F = \frac{n_1 + n_2 - p - 1}{p(n_1 + n_2 - 2)} T^2
\]

is distributed as \(F\)-distribution with \(p\) and \((n_1 + n_2 - p - 1)\) degrees of freedom and \(p\) is number of dependent variables. If \(F > F_{p,n_1+n_2-p-1}(\alpha)\), then reject the null hypothesis. That is, look at tabled \(F\) value with \(p\) and \(n_1 + n_2 - p - 1\) degrees of freedom with probability level of \(\alpha\). Alternatively, if the probability of \(F\) statistic is less than or equal to \(\alpha\), then I would reject the null hypothesis and conclude that statistically means across two groups are not equal.

Note that for our illustrative example,

\[
T^2 = \begin{pmatrix} 6 & 3 \\ 8 & -2 \end{pmatrix} \left( \frac{1}{5} + \frac{1}{4} \right) \frac{1}{7} \begin{pmatrix} 16 & -10 \\ -10 & 12 \end{pmatrix}^{-1} \begin{pmatrix} 6 & 3 \\ 8 & -2 \end{pmatrix}
\]
Note that middle parenthesis becomes
\[
\begin{pmatrix}
\frac{36}{35} & -\frac{9}{14} \\
-\frac{9}{14} & \frac{27}{35}
\end{pmatrix}
\]
which when inverted becomes
\[
\begin{pmatrix}
\frac{140}{69} & \frac{350}{207} \\
\frac{350}{207} & \frac{560}{207}
\end{pmatrix}
\]
This results in $T^2 = 176.52$ and computed $F$-value becomes $\frac{3}{7} \times 176.52$ or 75.65. Critical $F$-value for 2 and 6 degrees of freedom is 5.14 and thus we reject null hypothesis that means for both groups are equal.

- **Assumptions**

1. Each test unit is independent from others.
2. Variance-covariance matrices of both groups are about equal.
3. Each observation has equal influence in constructing group means and variance-covariance matrices.
4. Variations within group is a similar.
5. $y_{1-11}, y_{2-11}, y_{1-12}, y_{2-12}, \ldots, y_{1-1n_1}, y_{2-1n_1}$ are multivariate normally distributed, and also $y_{1-21}, y_{2-21}, y_{1-22}, y_{2-22}, \ldots, y_{1-2n_2}, y_{2-2n_2}$ are multivariate normally distributed. Alternatively, error values across treatments are multivariate normally distributed.

- **Using SAS to Conduct above analysis**

Above analysis also can be done using SAS. Following inputs were used to conduct such an analysis.

```plaintext
options nodate nocenter ps=60 ls=70;
data ex1;
input group y1 y2;
cards;
1 6 7
1 5 9
1 8 6
1 4 9
1 7 9
2 2 3
2 5 1
2 3 1
2 2 3
;
```

Multivariate Research Methods
proc glm;
class group;
model y1 y2 = group;
manova h=group / printe printh;
means group;
run;

SAS output produced following:

The SAS System
General Linear Models Procedure
Class Level Information

Class Levels Values
GROUP 2 1 2

Number of observations in data set = 9

Dependent Variable: Y1

Source      DF  Sum of Squares     F Value  Pr > F
Model        1  20.000000000     8.75     0.0212
Error        7  16.000000000
Corrected Total 8  36.000000000

R-Square  C.V.  Y1 Mean
0.555556  32.39695  4.66666667

Dependent Variable: Y2

Source      DF  Sum of Squares     F Value  Pr > F
Model        1  80.000000000    46.67     0.0002
Error        7  12.000000000
Corrected Total 8  92.000000000

R-Square  C.V.  Y2 Mean
0.869565  24.54951  5.33333333

E = Error SS&CP Matrix

<table>
<thead>
<tr>
<th></th>
<th>Y1</th>
<th>Y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>16</td>
<td>-10</td>
</tr>
<tr>
<td>Y2</td>
<td>-10</td>
<td>12</td>
</tr>
</tbody>
</table>

General Linear Models Procedure
Multivariate Analysis of Variance

H = Type III SS&CP Matrix for GROUP

<table>
<thead>
<tr>
<th></th>
<th>Y1</th>
<th>Y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>Y2</td>
<td>40</td>
<td>80</td>
</tr>
</tbody>
</table>

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no Overall GROUP Effect
H = Type III SS&CP Matrix for GROUP  E = Error SS&CP Matrix
S=1  M=0  N=2

Statistic  Value  F  Num DF  Den DF  Pr > F
Multivariate Analysis of Variance

<table>
<thead>
<tr>
<th></th>
<th>Wilks' Lambda</th>
<th>Pillai's Trace</th>
<th>Hotelling-Lawley Trace</th>
<th>Roy's Greatest Root</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.03814262</td>
<td>0.96185738</td>
<td>25.2173913</td>
<td>25.2173913</td>
</tr>
<tr>
<td></td>
<td>75.6522</td>
<td>75.6522</td>
<td>75.6522</td>
<td>75.6522</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

General Linear Models Procedure

<table>
<thead>
<tr>
<th>GROUP</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>6.00000000</td>
<td>1.58113883</td>
<td>8.00000000</td>
<td>1.41421356</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>3.00000000</td>
<td>1.41421356</td>
<td>2.00000000</td>
<td>1.15470054</td>
</tr>
</tbody>
</table>

- **Comments about SAS output**

  1. Matrix \( E \) is error covariance matrix and it is same as \( S_{\text{Pooled}} \) or between group error covariance matrix.

  2. Matrix \( H \) is hypothesis covariance matrix or between group covariance matrix.

  3. In this instance, various statistical tests results in same \( F \)-statistic. These tests will be discussed in section dealing with multiple group comparisons.

- **Evaluating MANOVA Assumptions**

  - Multivariate Normality: There number of tests available to test assumptions about multivariate normality. I have SAS macro `MULTNORM` that computes\(^1\) and assesses null hypothesis that a sample is from normal distribution. `multnorm` prints univariate tests of skewness, kurtosis and omnibus (jointly looking at both skewness and kurtosis) and as well as Mardia’s (1970), Small’s (1980) and Srivastava’s (1984)\(^2\) measures of multivariate skewness and kurtosis. In addition, there is graphic approach as well (plotting chi-square Q-Q).

  - Equality of variance-covariance matrices: Box (1949)’s\(^3\) test often is used to compare equality of variance-covariance matrices across various treatment groups. Suppose \( n \) be

---

\(^1\)This macro is placed in file `G:\Courses\26-606\multnorm.sas`.


Multivariate Analysis of Variance

total sample size. That is, \( n = n_1 + n_2 \cdots n_k \). There are \( k \) subgroups with variance-covariance matrix denoted by \( S_1, S_2 \cdots S_k \) and \( S_{\text{pooled}} \) be pooled variance-covariance matrix. Then Box’s M test is defined as

\[
M = (n - k) \log |S_{\text{pooled}}| - \sum_{j=1}^{k} (n_j - 1) \log |S_j|,
\]

where \( \log |S_j| \) is logarithm of the determinant of the variance-covariance matrix for group \( j \). Notice that if we are dealing with univariate case, then \( S_{\text{pooled}} \) will become pooled between group variance. Moreover, as pooled variance-covariance become equal to individual ones, \( M \) will approach close to zero. Box showed that \( M(1 - C) \) is approximately distributed as chi-square with \( p(p + 1)(k - 1)/2 \) degrees of freedom where \( C \) is equal to

\[
C = \frac{2p^2 + 3p - 1}{6(p + 1)(k - 1)} \left( \sum_{j=1}^{k} \frac{1}{n_j - 1} - \frac{1}{n - k} \right)
\]

and \( p \) is number of dependent variables.

For our illustrative example,

\[
S_1 = \begin{pmatrix}
2.5 & -1.5 \\
-1.5 & 2.0
\end{pmatrix}
\]

which means determinant of \( S_1 \) is \( |S_1| = 2.75 \) and \( \log |S_1| \) then is 1.0116. Similarly,

\[
S_2 = \begin{pmatrix}
2 & -4/3 \\
-4/3 & 4/3
\end{pmatrix}
\]

which means \( |S_2| \) is 0.8889 and \( \log |S_2| \) is -.1178. Finally,

\[
S_{\text{pooled}} = \frac{1}{7} \begin{pmatrix}
16 & -10 \\
-10 & 12
\end{pmatrix}
\]

and that means \( |S_{\text{pooled}}| \) is 1.8776 and \( \log |S_{\text{pooled}}| \) is 0.63. Thus, \( M = (9 - 2) \times (.63) - [(5 - 1)(1.0116) + (4 - 1)(-.1178)] \) which means \( M = 0.717 \). Note that \( C = (8 + 6 - 1)/(6 \times 3 \times 1) \times (1/4 + 1/3 - 1/7) = (13/18) \times (0.25 + 0.333 - 0.1429) \) which is equal to 0.318. Thus, \( M(1 - C) \) is 0.717(1 - .318) or 0.489 with 3 degrees of freedom. This would indicate that we could not reject null hypothesis since critical value of chi-square is 7.815 at \( \alpha = .05 \) and 3 degrees of freedom.

- **A Complete Analysis**

Johnson and Wichern (1992)\(^4\) provide costs of running trucks that use gasoline and diesel. Three components of cost were fuel, repair and capital all measured in cents per mile for a firm

involved in dairy products transportation. There were 36 observations with gasoline powered
trucks while 23 used diesel. I have used this dataset to indicate nature of analysis involved in
this instance. SAS input is given below first.

```
options ps =60 ls=80 nocenter nodate;
data trucks;
input fuelc repair capital fueltype $6.;
datalines;
  16.44 12.43 11.23 gasoline
  7.19  2.70  3.92 gasoline
  9.92  1.35  9.75 gasoline
  4.24  5.78  7.78 gasoline
  11.20  5.05 10.67 gasoline
  14.25  5.78  9.88 gasoline
  13.50 10.98 10.60 gasoline
  13.32 14.27  9.45 gasoline
  29.11 15.09  3.28 gasoline
  12.68  7.61 10.23 gasoline
  7.51  5.80  8.13 gasoline
  9.90  3.63  9.13 gasoline
 10.25  5.07 10.17 gasoline
  11.11  6.15  7.61 gasoline
  12.17 14.26 14.39 gasoline
  10.24  2.59  6.09 gasoline
  10.18  6.05 12.14 gasoline
   8.88  2.70 12.23 gasoline
  12.34  7.73 11.68 gasoline
  8.51 14.02 12.01 gasoline
  26.16 17.44 16.89 gasoline
  12.95  8.24  7.18 gasoline
  16.93 13.37 17.59 gasoline
  14.70 10.78 14.58 gasoline
  10.32  5.16 17.00 gasoline
   8.98  4.49  4.26 gasoline
   9.70 11.59  6.83 gasoline
  12.72  5.63  5.59 gasoline
  9.49  2.16  6.23 gasoline
   8.22  7.95  6.72 gasoline
  13.70 11.22  4.91 gasoline
   8.21  9.85  8.17 gasoline
  15.86 11.42 13.06 gasoline
   9.18  9.18  9.49 gasoline
  12.49  4.67 11.94 gasoline
  17.32  6.86  4.44 gasoline
  12.80 12.26  9.11  diesel
   7.42  5.13 17.15  diesel
  10.28  3.32 11.23  diesel
  10.16 14.72  5.99  diesel
  12.79  4.17 29.28  diesel
   9.60 12.72 11.00  diesel
   6.47  8.89 19.00  diesel
  11.35  9.95 14.53  diesel
  19.15  2.94 13.68  diesel
  12.70  5.06 20.84  diesel
  19.77 17.86 35.18  diesel
  11.61 11.75 17.00  diesel
  12.09 13.25 20.66  diesel
  19.83 10.14 17.45  diesel
   8.29  6.22 16.38  diesel
  15.90 12.90 19.09  diesel
```
Multivariate Analysis of Variance

11.94  5.69  14.77  diesel
9.54  16.77  22.66  diesel
10.43  17.65  10.66  diesel
10.87  21.52  28.47  diesel
7.13  13.22  19.44  diesel
11.88  12.18  21.20  diesel
12.03  9.22  23.09  diesel

;;;;
proc glm;
class fueltype;
model fuelc repair capital = fueltype;
manova h = fueltype / printh printe;
means fueltype;
run;
data gtrucks;
  set trucks;
  if fueltype="gasol";
  %include "c:\sas6_12\multnorm.sas";
  %multnorm(data=gtrucks,var=fuelc repair capital);
run;
data dtrucks;
  set trucks;
  if fueltype="diese";
  %multnorm(data=dtrucks,var=fuelc repair capital);
run;
proc discrim pool=test data=trucks;
class fueltype;
var fuelc repair capital;
run;

Few comments about SAS input.

- Options of MANOVA, that is `printh` and `printe` provide mechanism for printing hypothesis and error sums of squares and cross products matrices.

- Statement `means` is used to generate means for classification variable.

- There are two separate datasets used to determine multivariate normality. Furthermore, `proc discrim` is used test equality of covariance matrices. We will ignoring rest of output from discriminant analysis.

SAS output from above run produced following.

```
General Linear Models Procedure
Class Level Information

Class Levels Values
FUELTYPE  2  diese gasol

Number of observations in data set = 59

Dependent Variable: FUELC

  Source   DF  Sum of Squares  Square  F Value  Pr > F
Model     1  62.65567880  62.65567880  3.96   0.0513
Error    57  901.43859577 15.81471221
```

Multivariate Research Methods  Course: COST*6060
### Multivariate Analysis of Variance

Corrected Total 58 964.09427458

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>98.52879810</td>
<td>98.52879810</td>
<td>4.75</td>
<td>0.0335</td>
</tr>
<tr>
<td>Error</td>
<td>57</td>
<td>1182.77106630</td>
<td>20.75036958</td>
<td>20.75036958</td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>58</td>
<td>1281.29986441</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dependent Variable: REPAIR

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>98.52879810</td>
<td>98.52879810</td>
<td>4.75</td>
<td>0.0335</td>
</tr>
<tr>
<td>Error</td>
<td>57</td>
<td>1182.77106630</td>
<td>20.75036958</td>
<td>20.75036958</td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>58</td>
<td>1281.29986441</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### R-Square C.V. Root MSE REPAIR Mean

<table>
<thead>
<tr>
<th>Variable</th>
<th>R-Square</th>
<th>C.V.</th>
<th>Root MSE</th>
<th>REPAIR Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUELC</td>
<td>0.064989</td>
<td>34.89953</td>
<td>3.9767716</td>
<td>11.394915</td>
</tr>
<tr>
<td>REPAIR</td>
<td>0.076898</td>
<td>49.80914</td>
<td>4.5552574</td>
<td>9.1454237</td>
</tr>
</tbody>
</table>

### Dependent Variable: CAPITAL

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Squares</th>
<th>Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>1032.5347352</td>
<td>1032.5347352</td>
<td>38.84</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>57</td>
<td>1515.1134885</td>
<td>26.5809384</td>
<td>26.5809384</td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>58</td>
<td>2547.6482237</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### R-Square C.V. Root MSE CAPITAL Mean

<table>
<thead>
<tr>
<th>Variable</th>
<th>R-Square</th>
<th>C.V.</th>
<th>Root MSE</th>
<th>CAPITAL Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUELC</td>
<td>0.076898</td>
<td>49.80914</td>
<td>4.5552574</td>
<td>12.934068</td>
</tr>
<tr>
<td>REPAIR</td>
<td>0.076898</td>
<td>49.80914</td>
<td>4.5552574</td>
<td>9.1454237</td>
</tr>
</tbody>
</table>

### E = Error SS&CP Matrix

<table>
<thead>
<tr>
<th>Variable</th>
<th>FUELC</th>
<th>REPAIR</th>
<th>CAPITAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUELC</td>
<td>901.43859577</td>
<td>449.54134239</td>
<td>153.6974965</td>
</tr>
<tr>
<td>REPAIR</td>
<td>449.54134239</td>
<td>1182.7710663</td>
<td>336.1439837</td>
</tr>
<tr>
<td>CAPITAL</td>
<td>153.6974965</td>
<td>336.1439837</td>
<td>1515.1134885</td>
</tr>
</tbody>
</table>

### H = Type III SS&CP Matrix for FUELTYPE

<table>
<thead>
<tr>
<th>Variable</th>
<th>FUELC</th>
<th>REPAIR</th>
<th>CAPITAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUELC</td>
<td>62.65567803</td>
<td>-78.57091527</td>
<td>-254.3504762</td>
</tr>
<tr>
<td>REPAIR</td>
<td>-78.57091527</td>
<td>98.528798102</td>
<td>318.95831461</td>
</tr>
<tr>
<td>CAPITAL</td>
<td>-254.3504762</td>
<td>318.95831461</td>
<td>1032.5347352</td>
</tr>
</tbody>
</table>

### Manova Test Criteria and Exact F Statistics for the Hypothesis of no Overall FUELTYPE Effect

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>F</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks’ Lambda</td>
<td>0.52820432</td>
<td>16.3755</td>
<td>3</td>
<td>55</td>
<td>0.0001</td>
</tr>
<tr>
<td>Pillai’s Trace</td>
<td>0.47179568</td>
<td>16.3755</td>
<td>3</td>
<td>55</td>
<td>0.0001</td>
</tr>
<tr>
<td>Hotelling-Lawley Trace</td>
<td>0.89320679</td>
<td>16.3755</td>
<td>3</td>
<td>55</td>
<td>0.0001</td>
</tr>
<tr>
<td>Roy’s Greatest Root</td>
<td>0.89320679</td>
<td>16.3755</td>
<td>3</td>
<td>55</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

### Level of FUELTYPE

<table>
<thead>
<tr>
<th>FUELTYPE</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>diesel</td>
<td>23</td>
<td>10.1056522</td>
<td>2.08861595</td>
<td>10.7621739</td>
<td>5.08441107</td>
</tr>
<tr>
<td>gasol</td>
<td>36</td>
<td>12.2186111</td>
<td>4.79722429</td>
<td>8.1125000</td>
<td>4.18856905</td>
</tr>
</tbody>
</table>

### Level of CAPITAL

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>---------------</td>
<td>--------</td>
<td>--------</td>
</tr>
</tbody>
</table>
Following observations can be made from above.

- All costs differ by the type of fuel used by trucks. The cost of fuel for gasoline powered trucks is higher than diesel trucks by almost 2 cents, and statistically not significant at prob. 0.05. The repair cost of gasoline trucks is lower by more than 2.5 cents on an average and it is statistically significant at prob. of 0.05. The cost of capital for diesel trucks is almost twice as that of gasoline powered trucks.

- Overall, costs are significantly different for gasoline and diesel truck and biggest contributor to cost differences is cost of capital.

- Gasoline trucks have higher variability than diesel trucks.

**Testing Normality for Gasoline Trucks**

### Chi-square Q-Q plot for Gasoline Trucks

<table>
<thead>
<tr>
<th>Variable</th>
<th>Skewness (g1)</th>
<th>Sqrt(b1)</th>
<th>Normalized b1</th>
<th>Prob (b1=0)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUEL C</td>
<td>1.866</td>
<td>1.787</td>
<td>3.868</td>
<td>0.00011</td>
<td>36</td>
</tr>
<tr>
<td>REPAIR</td>
<td>0.364</td>
<td>0.348</td>
<td>0.964</td>
<td>0.33521</td>
<td>36</td>
</tr>
<tr>
<td>CAPITAL</td>
<td>0.357</td>
<td>0.342</td>
<td>0.946</td>
<td>0.34422</td>
<td>36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Kurtosis (g2)</th>
<th>b2</th>
<th>Normalized b2</th>
<th>Prob (b2=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUEL C</td>
<td>4.880</td>
<td>7.066</td>
<td>3.164</td>
<td>0.00156</td>
</tr>
<tr>
<td>REPAIR</td>
<td>-0.773</td>
<td>2.168</td>
<td>-1.243</td>
<td>0.21380</td>
</tr>
</tbody>
</table>
Multivariate Analysis of Variance

<table>
<thead>
<tr>
<th>Variable</th>
<th>Skewness (g1)</th>
<th>Sqrt(b1)</th>
<th>Normalized b1</th>
<th>Prob (b1=0)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUEL</td>
<td>0.702</td>
<td>0.655</td>
<td>1.484</td>
<td>0.13783</td>
<td>23</td>
</tr>
<tr>
<td>REPAIR</td>
<td>0.184</td>
<td>0.172</td>
<td>0.406</td>
<td>0.68503</td>
<td>23</td>
</tr>
<tr>
<td>CAPITAL</td>
<td>0.594</td>
<td>0.555</td>
<td>1.271</td>
<td>0.20367</td>
<td>23</td>
</tr>
</tbody>
</table>

Chi-square Q-Q plot for Diesel Trucks

<table>
<thead>
<tr>
<th>Variable</th>
<th>Kurtosis (g2)</th>
<th>b2</th>
<th>Normalized b2</th>
<th>Prob (b2=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FUEL</td>
<td>1.418</td>
<td>3.878</td>
<td>1.448</td>
<td>0.14759</td>
</tr>
<tr>
<td>REPAIR</td>
<td>-0.622</td>
<td>2.256</td>
<td>-0.640</td>
<td>0.52197</td>
</tr>
<tr>
<td>CAPITAL</td>
<td>0.633</td>
<td>3.254</td>
<td>0.887</td>
<td>0.39589</td>
</tr>
</tbody>
</table>

Normality tests for diesel trucks

Chi-square Q-Q plot for Diesel Trucks
Testing equality of covariance matrices.

Discriminant Analysis

59 Observations 58 DF Total
3 Variables 57 DF Within Classes
2 Classes 1 DF Between Classes

Class Level Information

Prior
FUELTYPE Frequency Weight Proportion Probability
diese 23 23.0000 0.389831 0.500000
gasol 36 36.0000 0.610169 0.500000

Discriminant Analysis Within Covariance Matrix Information

Covariance Natural Log of the Determinant
FUELTYPE Matrix Rank of the Covariance Matrix

diese 3 8.48879
gasol 3 8.06241
Pooled 3 8.79777

Discriminant Analysis
Test of Homogeneity of Within Covariance Matrices

Test Chi-Square Value = 30.544284
with 6 DF Prob > Chi-Sq = 0.0001
Two observations are made on the same test unit, once before treatment and once after the treatment. In multivariate case, multiple dependent variables are measured.

- **Experimental Situation**

  \[ O_1 \times X \times O_2 \times n \]  

  Note that we take measurement on \( n \) test units at two different times. That is, we observe two or more responses per test unit such as (if we have two response variables) \( y_{1-11}, y_{2-11}, y_{1-12}, y_{2-12} \ldots y_{1-1n}, y_{2-1n} \) and \( y_{1-21}, y_{2-21}, y_{1-22}, y_{2-22} \ldots y_{1-2n}, y_{2-2n} \) dependent variable values for pre- and post-treatment. Note that subscript before dash sign ("-") indicates variable number and subscript after dash sign ("-"), first number indicates occasion (pre = 1 and post = 2) and the second number indicates test unit number (subject, store etc.).

- **Hypothesis Tested**

  \[
  H_0 : \begin{pmatrix}
  \bar{y}_{1-d} \\
  \bar{y}_{2-d}
  \end{pmatrix} = \begin{pmatrix}
  0 \\
  0
  \end{pmatrix} \quad \text{and} \quad
  H_A : \begin{pmatrix}
  \bar{y}_{1-d} \\
  \bar{y}_{2-d}
  \end{pmatrix} \neq \begin{pmatrix}
  0 \\
  0
  \end{pmatrix},
  \]

  where \( \bar{y}_{1-d} \) and \( \bar{y}_{2-d} \) are mean differences for variable 1 and variable 2 respectively. Our interest lies in finding whether differences are statistically significant and on an average different from zero. That is,

  \[
  \begin{align*}
  y_{1-d1} &= y_{1-11} - y_{1-21} & y_{2-d1} &= y_{2-11} - y_{2-21} \\
  y_{1-d2} &= y_{1-12} - y_{1-22} & y_{2-d2} &= y_{2-12} - y_{2-22} \\
  \cdots & \cdots & \cdots & \cdots \\
  y_{1-dn} &= y_{1-1n} - y_{1-2n} & y_{2-dn} &= y_{2-1n} - y_{2-2n}
  \end{align*}
  \]

- **Illustrative Example:**

  A recent issue of *Consumer Reports*\(^5\) contained comparisons for the major chains for taste and nutrition. As add-on to this report, *Consumer Reports* also reported same information for food sold in Canada. I have chosen 17 food items for comparison purpose below. Each product is "closely" matched in both countries. I excluded number of items that were in the report, because I could not find comparable product either because the chains are different or item itself did not match. With global nature of fast food business, one would expect that food served in different countries might be very much similar, at least nutritionally. To test this

idea, I focussed my comparison on three attributes; calories, sodium content and fat content. Details about data is given below along with SAS set-up to analyze this comparison.

```sas
options nocenter nodate ps=60 ls=80;
/* Fast food data from Consumer Reports, Dec. 1997 */
/* Data for Items that are sold in Canada and US */
data fastf;
input obsno c_prc c_wtgm c_tfat c_sfat c_fatcal c_cal c_sod u_prc u_wtoz u_tfat u_sfat u_fatcal u_cal u_sod outlet $20.;
/*
c_prc - Price of item in Canada in Canadian $
c_wtgm - Weight of item in Canada in grams or ounces in US
c_tfat - Total fat for item sold in Canada
c_sfat - Saturated fat for item sold in Canada
c_fatcal - % of Calories from fat for item sold in Canada
c_cal - Calories
c_sod - Sodium in mg
*/
dif_cal = c_cal - u_cal ;
dif_sod = c_sod - u_sod ;
dif_tf = c_tfat - u_tfat;
inter = 1;
datalines;
1 3.70 246 7 2.0 16 350 924 3.20 9 6 1.0 16 348 978 Subway - Roasted Chicken
2 3.80 190 8 1.5 23 310 790 2.75 7 8 1.5 23 310 790 Wendy's - Grilled Chicken
3 3.00 . 15 3.0 42 319 737 2.90 9 26 5.0 44 530 1060 Burger King - BK Broiler
4 3.10 239 4 1.0 15 23 310 790 2.75 8 5 1.0 15 303 939 Subway - Roast Beef
5 3.60 . 28 11. 42 319 942 2.95 9 8 11. 45 555 1561 Arby's - Giant Roast Beef
6 2.00 90 17 3.5 61 250 550 1.00 3 14 3.0 60 210 460 Wendy's - Chicken Nuggets
7 2.80 115 14 3.0 54 235 521 1.90 4 17 3.5 53 290 510 McDonald's - Chicken McNuggets
8 2.60 . 11 4.0 21 111 390 2.50 4 22 7.0 57 350 940 Burger King - Chicken Tender
9 2.85 210 20 7.0 43 420 810 2.10 8 20 7.0 43 420 920 Wendy's - Single Burger with Everything
10 1.50 170 24 8.0 51 420 570 1.05 6 24 8.0 51 420 530 Burger King - Whopper Jr.
11 2.80 210 30 10. 50 541 1047 2.20 8 28 10. 57 530 880 McDonald's - Big Mac
12 3.30 282 30 12. 47 580 1460 2.75 10 30 12. 47 580 1460 Wendy's - Big Bacon Classic Burger
13 3.30 250 34 12. 52 591 1250 2.65 9 34 12. 50 610 1250 McDonald's - Arch Deluxe with Bacon
14 3.00 230 39 18. 55 641 1220 2.55 8 39 18. 55 640 1240 Burger King - Double Cheese Burger with Bacon
15 2.10 284 46 16. 57 730 1350 1.85 10 46 16. 57 730 1350 Burger King - Whopper with Cheese
16 1.40 160 23 3.5 44 470 150 1.05 6 23 3.5 44 470 150 Wendy's - Biggie French Fries
17 1.60 174 22 12. 42 520 290 1.33 5 22 4.0 44 490 290 McDonald's - Large French Fries
;;;
proc corr cov;
var dif_cal dif_sod dif_tf;
run;
proc glm ;
model dif_cal dif_sod dif_tf = inter / noint;
manova h=inter ;
run;
```

Note several things about this comparison.

- Three difference variables are created, each compares difference between the Canada and US. This is done outside proc glm.

- Variable inter and subsequent option to suppress use of intercept is meant to test the null hypothesis that overall means across three variables; calories (dif_cal), sodium

```sas
proc corr cov;
var dif_cal dif_sod dif_tf;
run;
proc glm ;
model dif_cal dif_sod dif_tf = inter / noint;
manova h=inter ;
run;
```
(dif_sod), and total fat (dif_tf) is zero. Variable inter takes fixed value of 1 for all observation. This is alternative way to estimate intercept in your model.

- **Proc corr** is used here to obtain descriptive information about these variables as well as correlations among them. It is not surprising that there is very high correlations among differenced variables.

Note that **proc glm** contains independent variable that is fixed to 1 across all observations. This one way to test

- **Test Statistic**

\[ T^2 = n \left( \begin{array}{c} \bar{y}_{1-d} \\ \bar{y}_{2-d} \\ \bar{y}_{3-d} \end{array} \right) S_d^{-1} \left( \begin{array}{c} \bar{y}_{1-d} \\ \bar{y}_{2-d} \\ \bar{y}_{3-d} \end{array} \right) \]

\( S_d^{-1} \) is the inverse of variance-covariance of differences.

- **Decision Rule**

If there are \( p \) variables to be compared then \( T^2 \left[ \frac{n-p}{(n-1)p} \right] > F_{p,n-p}(\alpha) \), then we would reject the null hypothesis. Note that \( F_{p,n-p}(\alpha) \) is used to denote type I error of \( (1 - \alpha) \) and \( p \) and \( n - p \) degrees of freedom.

- **Assumptions**

1. Each test unit is independent from others.
2. Each test unit has equal influence on estimated statistics.
3. Variations across test units is a similar.
4. \( y_1-d_1, y_1-d_2, \cdots y_1-d_n \) are multivariate normally distributed.

- **Back to illustrative example:**

Note that SAS produced following output from **proc corr**.

Correlation Analysis

3 'VAR' Variables: DIF_CAL DIF_SOD DIF_TF

Simple Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Sum</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
</table>

Multivariate Research Methods

Course: COST*6060
Note that from above analysis, food sold by these chains in the Canada appears to contain more calories, more sodium and more fat than that sold in the US. We can use statistical test to confirm whether these findings are by chance alone. To accomplish this, inverse of covariance matrix for differences ($S_d$) needs to be computed. That will be

$$S_d^{-1} = \begin{pmatrix}
47.9613 & 7.1157 & -1001.5506 \\
7.1157 & 3.0439 & -222.5103 \\
-1001.5506 & -222.5103 & 24300.6240 \\
\end{pmatrix} \times 10^{-4}.$$

This would result in $T^2$ value of

$$T^2 = n \begin{pmatrix}
23.76 & -43.88 & -1.06 \\
\end{pmatrix}
\begin{pmatrix}
\end{pmatrix}
^{-1}
\begin{pmatrix}
23.76 & -43.88 & -1.06 \\
\end{pmatrix}
= 6.6813.$$

This finally would result in F-value of $\frac{n-p}{(n-1)p}T^2$ is equal to $\frac{17-3}{16\times3} \times 6.6813 = 1.95$ We can confirm all this analysis using proc glm.

**General Linear Models Procedure**

Number of observations in data set = 17

Dependent Variable: DIF_CAL

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>9600.9411765</td>
<td>9600.9411765</td>
<td>2.39</td>
<td>0.1419</td>
</tr>
<tr>
<td>Error</td>
<td>16</td>
<td>64349.0588235</td>
<td>4021.8161765</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uncorrected Total</td>
<td>17</td>
<td>73950.0000000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Multivariate Analysis of Variance

R-Square  C.V.  Root MSE  DIF_CAL Mean
0.129830  -266.8570  63.417791  -23.764706

NOTE: No intercept term is used: R-square is not corrected for the mean.

Source  DF  Type I SS  Mean Square  F Value  Pr > F
INTER 1  9600.9411765  9600.9411765  2.39  0.1419

Dependent Variable: DIF_SOD

Sum of Mean
Source  DF  Squares  Square  F Value  Pr > F
Model 1  32736.235294  32736.235294  1.23  0.2846
Error 16  427227.764706  26701.735294
Uncorrected Total 17  459964.000000

R-Square  C.V.  Root MSE  DIF_SOD Mean
0.071171  -372.3744  163.40666  -43.882353

NOTE: No intercept term is used: R-square is not corrected for the mean.

Source  DF  Type I SS  Mean Square  F Value  Pr > F
INTER 1  32736.235294  32736.235294  1.23  0.2846

Dependent Variable: DIF_TF

Sum of Mean
Source  DF  Squares  Square  F Value  Pr > F
Model 1  19.05882353  19.05882353  1.22  0.2866
Error 16  250.94117647  15.68382353
Uncorrected Total 17  270.00000000

R-Square  C.V.  Root MSE  DIF_TF Mean
0.070588  -374.0265  3.9602807  -1.0588235

NOTE: No intercept term is used: R-square is not corrected for the mean.

Source  DF  Type I SS  Mean Square  F Value  Pr > F
INTER 1  19.05882353  19.05882353  1.22  0.2866

Manova Test Criteria and Exact F Statistics for
the Hypothesis of no Overall INTER Effect
H = Type III SS&CP Matrix for INTER  E = Error SS&CP Matrix
S=1  M=0.5  N=6

Statistic  Value  F  Num DF  Den DF  Pr > F
Wilks' Lambda  0.70537271  1.9492  3  14  0.1681
Pillai's Trace  0.29462729  1.9492  3  14  0.1681
Hotelling-Lawley Trace  0.41769023  1.9492  3  14  0.1681
Roy's Greatest Root  0.41769023  1.9492  3  14  0.1681

Since $F$-statistic is 1.95 and probability of this is 0.17, we can not reject the null hypothesis that product sold in the Canada and US contain different amount of calories, sodium and fat. In other words, while there are differences on these attributes for product sold by fast food chains, there might be more variation in observed differences or our sample size is too small to detect these differences.
Suppose that more than two treatments are to be compared with the control group on multiple dependent variables. Test units (individuals, stores etc.) assigned randomly to their respective groups. As an example, suppose there are two treatments and one control group and we have measured two dependent variables per test unit.

- **Experimental Situation**

  Treatment 1 \( R \ X_1 \ O_1 \ n_1 \) test units.

  Treatment 2 \( R \ X_1 \ O_2 \ n_2 \) test units.

  Control Group \( R \ O_3 \ n_3 \) test units.

  Note that we take measurement on \( n_1, n_2 \cdots n_k \) test units. That is, we obtain

\[
\begin{array}{cccccc}
& y_{11} & y_{21} & y_{12} & y_{22} & y_{13} & y_{23} \\
\text{Treatment 1} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\text{Treatment 2} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\text{Control} & \cdot & \cdot & y_{1n} & y_{2n} & \cdot & \cdot \\
\end{array}
\]

\[
\begin{align*}
\text{mean} & \quad \bar{y}_{11} & \quad \bar{y}_{21} & \quad \bar{y}_{12} & \quad \bar{y}_{22} & \quad \bar{y}_{13} & \quad \bar{y}_{23} \\
\end{align*}
\]

That is, we took measurement on \( n = n_1 + n_2 + n_3 \) test units. The last row in above table are computed averages. In univariate case where we were concerned about within group variation (for example, \( ss_1, ss_2, \cdots ss_k \) and \( k \) is number of groups to be compared), here we must also be concerned about cross variable variation or covariation. In other words, we will be concerned about following sums of squares (SS) and cross product (SCP) matrix.

\[
H = \begin{pmatrix}
\bar{y}_{1} \\
\bar{y}_{2}
\end{pmatrix}
\begin{pmatrix}
SS_{11} & SCP_{12} \\
SCP_{21} & SS_{22}
\end{pmatrix}.
\]

- **Illustrative Example:**

  Suppose we have three groups and two dependent variables. Measurements were

\[
\begin{align*}
\text{Group 1} &= \begin{pmatrix}
6 & 7 \\
5 & 9 \\
8 & 6 \\
4 & 9 \\
7 & 9
\end{pmatrix} \\
\text{Group 2} &= \begin{pmatrix}
3 & 3 \\
1 & 6
\end{pmatrix} \\
\text{Group 3} &= \begin{pmatrix}
2 & 3 \\
5 & 1
\end{pmatrix}
\end{align*}
\]
Note that means for group 1 are 6 and 8, group 2 are 2 and 4 and group 3 are 3 and 2 for variable 1 and 2 respectively and overall sample means are 4 and 5 for respective variables.

Note that in general, an observed variable value \((y_{j-i}, \text{that is variable } j, \text{ group } l \text{ and observation number within a group } i)\) can be decomposed as

\[
y_{j-i} = \bar{y}_{j-.} + (\bar{y}_{j-l} - \bar{y}_{j-.}) + (\bar{y}_{j-i} - \bar{y}_{j-l}).
\]

Let us see, how this identity might apply to the first variable.

\[
\begin{bmatrix}
6 & 3 & 2 \\
5 & 1 & 5 \\
8 & 2 & 3 \\
7 & 2 & \\
\end{bmatrix} = \begin{bmatrix}
4 & 4 & 4 \\
4 & 4 & 4 \\
4 & 4 & 4 \\
4 & 4 & \\
\end{bmatrix} + \begin{bmatrix}
2 & -2 & -1 \\
2 & -2 & -1 \\
2 & -2 & -1 \\
2 & -1 \\
\end{bmatrix} + \begin{bmatrix}
0 & 1 & -1 \\
-1 & -1 & 2 \\
-2 & -1 \\
\end{bmatrix}
\]

Suppose we squared each element of above matrix and added them, then we would have sums of squares for each component. That is,

\[
SS_{\text{response}} = SS_{\text{mean}} + SS_{\text{Treatment}} + SS_{\text{residual}}
\]

\[
246 = 192 + 36 + 18.
\]

We need to repeat this for the second variable as well. Various components would be

\[
\begin{bmatrix}
7 & 3 & 3 \\
9 & 6 & 1 \\
6 & 3 & 1 \\
9 & 3 & \\
\end{bmatrix} = \begin{bmatrix}
5 & 5 & 5 \\
5 & 5 & 5 \\
5 & 5 & 5 \\
5 & 5 & \\
\end{bmatrix} + \begin{bmatrix}
3 & -1 & -3 \\
3 & -1 & -3 \\
3 & -1 & -3 \\
3 & -3 \\
\end{bmatrix} + \begin{bmatrix}
-1 & -1 & 1 \\
1 & 2 & -1 \\
-2 & -1 & -1 \\
1 & 1 \\
\end{bmatrix}
\]

Moreover,

\[
SS_{\text{response}} = SS_{\text{mean}} + SS_{\text{Treatment}} + SS_{\text{residual}}
\]

\[
402 = 300 + 84 + 18.
\]

For sums of squares, we squared each item and then added squared values. On the other hand, to compute sums of cross products, we would multiply each test units scores for variable 1 and variable 2, and then sum them over all test units. Note that sums of cross product are symmetric, that is, whether we multiply first variable 1 or variable 2 should not change cross product. In addition, when you have two variables, we would have one sum of cross product summation. With three dependent variables, we would have three cross product terms. For example, \(SCP_{\text{response}}\) will be equal to

\[
\begin{bmatrix}
6 & 3 & 2 \\
5 & 1 & 5 \\
8 & 2 & 3 \\
4 & 2 & \\
7 & \\
\end{bmatrix} \times \begin{bmatrix}
7 & 3 & 3 \\
9 & 6 & 1 \\
6 & 3 & 1 \\
9 & 3 & \\
\end{bmatrix} = 6 \times 7 + 5 \times 9 + \cdots + 2 \times 3 = 275
\]
Thus, sums of cross product equation would be
\[
SCP_{\text{response}} = SCP_{\text{mean}} + SCP_{\text{Treatment}} + SCP_{\text{residual}}
\]
275 = 240 + 48 + −13.

We summarize our example as follows.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Matrix of sum of squares and cross products</th>
<th>Degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment</td>
<td>H = \begin{pmatrix} 36 &amp; 48 \ 48 &amp; 84 \end{pmatrix}</td>
<td>k − 1 = 2</td>
</tr>
<tr>
<td>Residual</td>
<td>E = \begin{pmatrix} 18 &amp; -13 \ -13 &amp; 18 \end{pmatrix} (\sum_{j=1}^{k} n_j) − k = 9</td>
<td></td>
</tr>
</tbody>
</table>

Note that matrix of sum of squares and cross products due to treatment is called H or B referring to hypothesis or between group sum of squares. Matrix of sum of squares due to residuals is called E or W referring to error (residual) or within group sum of squares. We will be using H and E notation to be consistent with SAS output.

- **Hypothesis Tested**

  \[H_0 : \begin{pmatrix} \bar{y}_{1-1} \\ \bar{y}_{2-1} \end{pmatrix} = \begin{pmatrix} \bar{y}_{1-2} \\ \bar{y}_{2-2} \end{pmatrix} = \begin{pmatrix} \bar{y}_{1-3} \\ \bar{y}_{2-3} \end{pmatrix}\]

  \[H_A : \text{at least two pairs of means are different from each other.}\]

- **Test Statistics:**

  To test null hypothesis, we could use Wilks’ Lambda (\(\Lambda\)) and it is equal to

  \[
  \Lambda = \frac{|E|}{|H + E|},
  \]

  where \(|E|\) represents determinant of residual or error sum of squares and cross products. It should come as no surprise that smaller the value of \(\Lambda\), more likely we to reject null hypothesis. This would happen, if residual sum of squares are small or treatment sum of squares are relatively large.

- **Decision Rules:**

  Distribution of Wilks’ Lambda depends upon number of dependent variables and number of groups compared. Following table summarizes relevant F-distribution.
Multivariate Analysis of Variance

<table>
<thead>
<tr>
<th>No. of variables (p)</th>
<th>No. of groups (k)</th>
<th>Compute</th>
<th>F–distribution degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>p = 1</td>
<td>k ≥ 2</td>
<td>(n−k)</td>
<td>Fk−1,n−k</td>
</tr>
<tr>
<td>p = 2</td>
<td>k ≥ 2</td>
<td>(n−k−1)</td>
<td>F2(k−1),2(n−k−1)</td>
</tr>
<tr>
<td>p ≥ 1</td>
<td>k = 2</td>
<td>(n−p−1)</td>
<td>Fp,n−p−1</td>
</tr>
<tr>
<td>p ≥ 1</td>
<td>k = 3</td>
<td>(n−p−2)</td>
<td>F2p,2(n−p−2)</td>
</tr>
</tbody>
</table>

In this table n is total sample size or \( \sum_{j=1}^{k} n_j \). For a large n there is alternative to Wilks’ test. Bartlett has shown that for a large \( n - \left( n - 1 - \frac{n+k}{2} \right) \log \Lambda \) has approximately a chi-square distribution with \( p(k-1) \) degrees of freedom. Consequently, for a large n, we reject null hypothesis at significance level \( \alpha \) if

\[- \left( n - 1 - \frac{n+k}{2} \right) \log \Lambda > \chi^2_{p(k-1)}(\alpha)\]

where \( \chi^2_{p(k-1)}(\alpha) \) is the upper (100\( \alpha \)) percentile of a chi-square distribution with \( p(k-1) \) degrees of freedom.

For our illustrative example, \(|E|\) is 155 and

\[|H + E| = \begin{pmatrix} 36 & 48 \\ 48 & 84 \end{pmatrix} + \begin{pmatrix} 18 & -13 \\ -13 & 18 \end{pmatrix} = \begin{pmatrix} 54 & 35 \\ 35 & 102 \end{pmatrix} = 4283.\]

This means that \( \Lambda \) is 155/4283 = 0.0362. Since we have two variables (\( p = 2 \)) and three groups (\( k = 3 \))

\[\left( \frac{n - k - 1}{k - 1} \right) \left( \frac{1 - \sqrt{\Lambda}}{\sqrt{\Lambda}} \right) = \left( \frac{1 - .19}{.19} \right) \frac{8}{2} = 17.05.\]

Since \( F_{4,16}(0.05) \) is 3.01, we would reject the null hypothesis and conclude that treatment differences exist at \( \alpha = 0.05 \).

- **Assumptions**

  1. Each test unit is independent from others.

  2. Variance-covariance matrices of all groups are about equal.

  3. Each test unit has equal influence in constructing group means and variance-covariance matrices.
4. Variations within group is a similar.

5. $y_{1-11}, y_{2-11}, y_{1-12}, y_{2-12}, \ldots y_{1-1n_1}, y_{2-1n_1}$ are multivariate normally distributed, and also $y_{1-21}, y_{2-21}, y_{1-22}, y_{2-22}, \ldots y_{1-kn_2}, y_{2-kn_2}$ are multivariate normally distributed. Alternatively, error values across treatments are multivariate normally distributed.

- **Alternative Tests:**

  There are three other tests that are often reported in the literature. All other three tests also use matrix $\mathbf{H}$ and / or $\mathbf{E}$. SAS prints out these as

  1. Pillai Trace,
  2. Hotelling-Lawley Trace and
  3. Roy’s greatest root.

  Without getting into details of these\(^6\), here is suggestion as to which might be appropriate in a particular situation.

  1. When there are two groups and number of dependent variables are two, all tests will give same $F$-statistic.
  2. Roy’s test works best when means across groups move in the same direction (that is, variables have high correlations). In other instances, this test is not as powerful as others.
  3. Pillai’s Trace performs very well when there is a departure from normality and covariances across groups are not equal.
  4. When there is only departure from normality, Pillai’s, Hotelling-Lawley and Wilks’ Lambda, will perform very well.
  5. Many individuals not familiar with the multivariate analysis often use Wilks’ Lambda, because this was the first test proposed in the literature and as well has intuitive appeal.

- **Attitude towards Science and Environment**

  This analysis is based on dataset from the Inter-University Consortium for Political and Social Research (ICPSR). In present analysis I will be using data from International Social Survey in 1993 about environment. I focussed here on data from five countries, United Kingdom (UK), United States (US), New Zealand (NZ), Ireland (IR) and Canada (CN), all predominantly English speaking countries. I selected randomly 20% of respondents from each country to save

\(^6\)Suppose eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ are $\lambda_1 > \lambda_2 > \cdots > \lambda_s$ where $s$ is min$(k, p)$, then Pillai’s trace is $\sum_{i=1}^{s} \frac{\lambda_i}{1+\lambda_i}$, Hotelling-Lawley trace is $\sum_{i=1}^{s} \lambda_i$ and Roy’s greatest root is $\frac{\lambda_1}{1+\lambda_1}$. 

Multivariate Research Methods  
Course: COST*6060
time required for computation. Finally, I have chosen four questions for analysis purpose. All these questions had Likert-scale responses (1=Strongly agree, to 5=Strongly disagree). The specific questions were:

- We believe too often in science, and not enough in feeling and faith (V9, negative wording).
- Overall, modern science does more harm than good (V10, negative wording).
- Any change humans cause in nature – no matter how scientific – is likely to make things worse (V11, negative wording).
- Modern science will solve our environmental problems with little change to our way of life (V12, positive wording).

I expected that the North American respondent would have similar response pattern than respondents from the UK and those from down under.

SAS Input for this particular analysis is as follows.

```sas
options ls=80 ps=60 nodate nocenter;
data intev;
infile "D:\26-606\issp93.dat" dlm='"';
input V3 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18 V19
V21 V22 V23 V200 V201 V202 V212 V329 V351 V352;
if v9 ge 6 then v9 = .;
if v10 ge 6 then v10 = .;
if v11 ge 6 then v11 =.;
if v12 ge 6 then v12 = .;
if v13 ge 6 then v13 = .;
if v14 ge 6 then v14 = .;
/* Randomly select 20% of observations */
seed =2421998;
for x = ranuni(seed);
if x le .2 ;
if v3 = 4 then country ="UK";
if v3 = 6 then country ="US";
if v3 = 10 then country ="IR";
if v3 = 18 then country ="NZ";
if v3 = 19 then country ="CN";
proc glm;
class country;
model v9 v10 v11 v12 = country;
manova h=country / printh printe;
output out=predatt
   residual=v9res v10res v11res v12res ;
means country;
run;
%include "c:sas6_12\multnorm.sas";
%multnorm(data=predatt, var=v9res v10res v11res v12res);
run;
```

Several comments about input setup.

- Note that using fixed random number seed, we will ensure that same set of observation will be used, if we had to repeat our analysis.
• I have used to residuals to determine multivariate normality.

• With 1166 observations in residual dataset, `MULTNORM` macro is very slow. It will take considerable time to compute various test statistics.

SAS output from analysis is as follows.

The SAS System
General Linear Models Procedure
Class Level Information
Class Levels Values
COUNTRY 5 IR CN NZ UK US

Number of observations in data set = 1308
NOTE: Observations with missing values will not be included in this analysis.
Thus, only 1166 observations can be used in this analysis.

General Linear Models Procedure

Dependent Variable: V9 - Too much faith in Science

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>14.51371280</td>
<td>3.62842820</td>
<td>3.53</td>
<td>0.0071</td>
</tr>
<tr>
<td>Error</td>
<td>1161</td>
<td>1192.34477776</td>
<td>1.02699809</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>1165</td>
<td>1206.85949057</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square: 0.012026
C.V.: 39.34849
Root MSE: 1.0134091
Mean: 2.5754717

Dependent Variable: V10 - Science harmful

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>39.70698852</td>
<td>9.92674713</td>
<td>9.16</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>1161</td>
<td>1258.05287426</td>
<td>1.08359421</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>1165</td>
<td>1297.75986278</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square: 0.030597
C.V.: 30.61951
Root MSE: 1.0409583
Mean: 3.3996569

Dependent Variable: V11 - Any change leads to worsening...

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>10.92392931</td>
<td>2.73098233</td>
<td>2.35</td>
<td>0.0522</td>
</tr>
<tr>
<td>Error</td>
<td>1161</td>
<td>1347.32735714</td>
<td>1.16048868</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>1165</td>
<td>1358.25128645</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Multivariate Analysis of Variance

R-Square | C.V. | Root MSE | V11 Mean
---------|------|----------|---------
0.008043 | 34.90094 | 1.0772598 | 3.0866209

Source | DF | Type I SS | Mean Square | F Value | Pr > F
COUNTRY 4 | 10.92392931 | 2.73098233 | 2.35 | 0.0522

Dependent Variable: V11 - Modern science will solve...

Sum of Mean
Source | DF | Squares | Square | F Value | Pr > F
Model | 4 | 23.58348290 | 5.89587073 | 6.18 | 0.0001
Error | 1161 | 1108.49884986 | 0.95477937 |
Corrected Total | 1165 | 1132.08233276 |

R-Square | C.V. | Root MSE | V12 Mean
---------|------|----------|---------
0.020832 | 27.29591 | 0.9771281 | 3.5797599

Source | DF | Type I SS | Mean Square | F Value | Pr > F
COUNTRY 4 | 23.58348290 | 5.89587073 | 6.18 | 0.0001

E = Error SS&CP Matrix

<table>
<thead>
<tr>
<th>V9</th>
<th>V10</th>
<th>V11</th>
<th>V12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1192.3447778</td>
<td>375.95102391</td>
<td>410.19684968</td>
<td>-30.45995576</td>
</tr>
<tr>
<td>375.95102391</td>
<td>1258.0528743</td>
<td>579.04398181</td>
<td>44.56493917</td>
</tr>
<tr>
<td>410.19684968</td>
<td>579.04398181</td>
<td>1347.3273571</td>
<td>-44.05838412</td>
</tr>
<tr>
<td>-30.45995576</td>
<td>44.56493917</td>
<td>-44.05838412</td>
<td>1108.4988499</td>
</tr>
</tbody>
</table>

Multivariate Analysis of Variance

H = Type III SS&CP Matrix for COUNTRY

<table>
<thead>
<tr>
<th>V9</th>
<th>V10</th>
<th>V11</th>
<th>V12</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.513712805</td>
<td>12.879164767</td>
<td>5.6805088063</td>
<td>10.441087838</td>
</tr>
<tr>
<td>12.879164767</td>
<td>39.706988517</td>
<td>19.590666556</td>
<td>20.266964775</td>
</tr>
<tr>
<td>5.6805088063</td>
<td>19.590666556</td>
<td>10.923929314</td>
<td>13.502637983</td>
</tr>
<tr>
<td>10.441087838</td>
<td>20.266964775</td>
<td>13.502637983</td>
<td>23.583482903</td>
</tr>
</tbody>
</table>

Manova Test Criteria and F Approximations for the Hypothesis of no Overall COUNTRY Effect

H = Type III SS&CP Matrix for COUNTRY  E = Error SS&CP Matrix

S=4  M=-0.5  N=578

Statistic | Value | F | Num DF | Den DF | Pr > F
Wilks' Lambda | 0.93901162 | 4.6024 | 16 | 3538.386 | 0.0001
Pillai's Trace | 0.06192428 | 4.5640 | 16 | 4644 | 0.0001
Hotelling-Lawley Trace | 0.06192428 | 4.6229 | 16 | 4626 | 0.0001
Roy's Greatest Root | 0.04357317 | 12.6471 | 4 | 1161 | 0.0001

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

General Linear Models Procedure

Level of V9-------- V10--------
COUNTRY N Mean SD Mean SD
IR 194 2.39690722 1.03422807 3.10824742 1.12145077
CN 253 2.71146245 1.04641633 3.61264822 1.03133356
Multivariate Analysis of Variance

<table>
<thead>
<tr>
<th>Country</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>NZ</td>
<td>212</td>
<td>2.613</td>
<td>0.972</td>
<td>3.490</td>
<td>0.990</td>
</tr>
<tr>
<td>UK</td>
<td>216</td>
<td>2.653</td>
<td>0.922</td>
<td>3.208</td>
<td>0.939</td>
</tr>
<tr>
<td>US</td>
<td>291</td>
<td>2.491</td>
<td>1.018</td>
<td>3.484</td>
<td>1.099</td>
</tr>
</tbody>
</table>

Mean Responses by Sampled Country

<table>
<thead>
<tr>
<th>Question</th>
<th>CN</th>
<th>NZ</th>
<th>US</th>
<th>UK</th>
<th>IR</th>
</tr>
</thead>
<tbody>
<tr>
<td>V9. Faith in science</td>
<td>2.71</td>
<td>2.61</td>
<td>2.49</td>
<td>2.65</td>
<td>2.40</td>
</tr>
<tr>
<td>V10. Science harmful</td>
<td>3.61</td>
<td>3.49</td>
<td>3.48</td>
<td>3.21</td>
<td>3.11</td>
</tr>
<tr>
<td>V11. Humans cause harm</td>
<td>3.22</td>
<td>3.14</td>
<td>3.10</td>
<td>2.95</td>
<td>2.99</td>
</tr>
<tr>
<td>V12. Modern science solve</td>
<td>3.82</td>
<td>3.64</td>
<td>3.47</td>
<td>3.43</td>
<td>3.53</td>
</tr>
</tbody>
</table>

Comments about our analysis.

- Canadian and Irish respondents are generally on opposite end of measurement scale.
- Canadians and New Zealanders have similar opinions on these items.
- One would have expected that the US respondents to have relatively more extreme views about environment and science. This notion is not supported in this dataset.

Testing normality of residuals
### Univariate Normality Tests

<table>
<thead>
<tr>
<th>Variable</th>
<th>Skewness (g1)</th>
<th>Sqrt(b1)</th>
<th>Normalized b1</th>
<th>Prob (b1=0)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>V9RES</td>
<td>0.357</td>
<td>0.356</td>
<td>4.854</td>
<td>0.00000</td>
<td>1166</td>
</tr>
<tr>
<td>V10RES</td>
<td>-0.514</td>
<td>-0.513</td>
<td>-6.803</td>
<td>0.00000</td>
<td>1166</td>
</tr>
<tr>
<td>V11RES</td>
<td>-0.250</td>
<td>-0.250</td>
<td>-3.449</td>
<td>0.00056</td>
<td>1166</td>
</tr>
<tr>
<td>V12RES</td>
<td>-0.485</td>
<td>-0.484</td>
<td>-6.455</td>
<td>0.00000</td>
<td>1166</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Kurtosis (g2)</th>
<th>b2</th>
<th>Normalized b2</th>
<th>Prob (b2=3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V9RES</td>
<td>-0.578</td>
<td>2.420</td>
<td>-5.737</td>
</tr>
<tr>
<td>V10RES</td>
<td>-0.509</td>
<td>2.488</td>
<td>-4.788</td>
</tr>
<tr>
<td>V11RES</td>
<td>-0.963</td>
<td>2.036</td>
<td>-14.480</td>
</tr>
<tr>
<td>V12RES</td>
<td>-0.449</td>
<td>2.548</td>
<td>-4.029</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Omnibus Chi-sq</th>
<th>Prob (Normal)</th>
<th>Shapiro-Wilk</th>
<th>Prob (Normal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>V9RES</td>
<td>56.48</td>
<td>0.00000</td>
<td>0.9215</td>
</tr>
<tr>
<td>V10RES</td>
<td>69.21</td>
<td>0.00000</td>
<td>0.9239</td>
</tr>
<tr>
<td>V11RES</td>
<td>221.56</td>
<td>0.00000</td>
<td>0.9048</td>
</tr>
<tr>
<td>V12RES</td>
<td>57.90</td>
<td>0.00000</td>
<td>0.9169</td>
</tr>
</tbody>
</table>

### Multivariate Normality Tests

<table>
<thead>
<tr>
<th>Test</th>
<th>Measure</th>
<th>Test Statistic</th>
<th>Prob(Normal)</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mardia’s Skewness</td>
<td>0.937</td>
<td>182.684</td>
<td>0.00000</td>
<td>1166</td>
</tr>
<tr>
<td>Mardia’s Kurtosis</td>
<td>25.010</td>
<td>2.489</td>
<td>0.01280</td>
<td>1166</td>
</tr>
<tr>
<td>Small’s Skewness</td>
<td>122.607</td>
<td>122.607</td>
<td>0.00000</td>
<td>1166</td>
</tr>
<tr>
<td>Small’s Kurtosis</td>
<td>274.478</td>
<td>274.478</td>
<td>0.00000</td>
<td>1166</td>
</tr>
<tr>
<td>Small’s Omnibus</td>
<td>397.085</td>
<td>397.085</td>
<td>0.00000</td>
<td>1166</td>
</tr>
<tr>
<td>Srivastava’s Skewness</td>
<td>0.085</td>
<td>65.735</td>
<td>0.00000</td>
<td>1166</td>
</tr>
<tr>
<td>Srivastava’s Kurtosis</td>
<td>3.164</td>
<td>2.292</td>
<td>0.02193</td>
<td>1166</td>
</tr>
</tbody>
</table>

All above tests indicate that our residuals are not normal. Chi-square Q-Q plot indicate that there might be 5% to 10% respondents more extreme that might be resulting in lack of normality. One would then identify these observations and determine causes of these extreme responses.
Multiple Factorial Designs

Consider a situation where interest is in determining impact of more than one experimental factor. One example might be price level (−5%, 0% and +5%) and whether brand was promoted via feature advertising (two levels). We might be interested in measuring impact of price and advertising on own brand sales ($y_1$), competitive brand sales ($y_2$) as well as total category sales ($y_3$). Our interest would be in comparing changes to test units for both factors as well as interaction between two factors. Note that interaction have both statistical as well as substantive meaning.

**Experimental Situation**

Consider a design with two factors. First factor has $L$-levels and the second factor has $K$-levels. In order to understand interactions, we may need at as many as $L \times K$ treatment groups. If $L$ is 3 and $K$ is 2, then we would need 6 treatment groups. These groups could be shown as follows:

<table>
<thead>
<tr>
<th>Group</th>
<th>Test Units</th>
<th>Factor A</th>
<th>Factor B</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$R \times O_1$</td>
<td>$n_1$</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>$R \times O_2$</td>
<td>$n_2$</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>$R \times O_3$</td>
<td>$n_3$</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>$R \times O_4$</td>
<td>$n_4$</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>$R \times O_5$</td>
<td>$n_5$</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>$R \times O_6$</td>
<td>$n_6$</td>
<td>3</td>
</tr>
</tbody>
</table>

Note that we take measurement on $n_1$, $n_2$, $\ldots$ $n_6$ test units for six different treatments respectively. That is, we obtain $y_{1-11}$, $y_{1-12}$, $y_{1-13}$, $y_{2-11}$, $y_{2-12}$, $y_{2-13}$, $y_{3-11}$, $y_{3-12}$, $y_{3-13}$, $y_{4-11}$, $y_{4-12}$, $y_{4-13}$, $y_{5-11}$, $y_{5-12}$, $y_{5-13}$, $y_{6-11}$, $y_{6-12}$, $y_{6-13}$ and so on dependent variable values.

**Illustrative Example:**

Stevens (1996)\(^7\) provides an example with two factors with following observations.

Note that overall mean for variable 1 and variable 2 is 11.55 and 11.45 respectively. We can also breakdown observations by both factors and compute means for each cell as well as overall means across cell. By examining means and possibly their standard deviations, we can get better ‘feel’ for summary information. I used SAS to create summary measures here, although some of these could also be computed with help of calculator.

Cell Means and Sample Sizes

<table>
<thead>
<tr>
<th></th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
<th>Factor A</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>6.50</td>
<td>9.00</td>
<td>13.67</td>
<td>16.33</td>
<td>12.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.75</td>
<td></td>
<td>3</td>
<td>12.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>18.33</td>
<td></td>
<td></td>
<td>12.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>15.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>9.00</td>
<td>8.67</td>
<td>11.67</td>
<td>11.67</td>
<td>10.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11.67</td>
<td>11.67</td>
<td>10.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.60</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.80</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>8.38</td>
<td>8.75</td>
<td>12.67</td>
<td>14.00</td>
<td>11.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>11.29</td>
<td>10.29</td>
<td>11.45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>14.13</td>
<td>13.25</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
<td>8</td>
<td></td>
</tr>
</tbody>
</table>

For this example, we need to construct \( H \) and \( E \) matrices for factor A, factor B and factor A*B. To accomplish that let us look at decomposition. Suppose we denote observed value as \( y_{j-lki} \) where \( j \) indicates variable number, \( l \) indicates level associated with factor A, \( k \) indicates level associated with factor B and \( i \) individual response such that \( i = 1, 2, \ldots, n \) where \( n \) is sample size or \( n = \sum_{l=1}^{L} \sum_{k=1}^{K} n_{lki} \). Each observation can be decomposed as

\[
\begin{pmatrix}
  y_{j-lki} \\
  \text{response}
\end{pmatrix}
= \begin{pmatrix}
  \bar{y}_{j-\cdots} \\
  \text{Grand} \\
  \text{Mean}
\end{pmatrix} + \begin{pmatrix}
  (\bar{y}_{j-l\cdots} - \bar{y}_{j-\cdots}) \\
  \text{Factor A} \\
  \text{Effect}
\end{pmatrix} + \begin{pmatrix}
  (\bar{y}_{j-\cdots-k} - \bar{y}_{j-\cdots}) \\
  \text{Factor B} \\
  \text{Effect}
\end{pmatrix} + \begin{pmatrix}
  (\bar{y}_{j-\cdots-lk} - \bar{y}_{j-\cdots-l\cdots-k} + \bar{y}_{j-\cdots}) \\
  \text{Factor A} \times \text{B} \\
  \text{Effect}
\end{pmatrix} + \begin{pmatrix}
  (y_{j-lki} - \bar{y}_{j-l\cdots}) \\
  \text{Residual}
\end{pmatrix}.
\]

Although SAS or SPSS, use above identity to construct various \( H \) matrices, generally they are not printed. This is partly due to the fact that our interest is about the particular hypothesis. Note that for this example, we could construct following table.

- **Hypothesis Tested:**

  Experiment of such nature have multiple hypotheses. We want to find out whether factor A, factor B and factor A and factor B taken jointly together influence means of dependent variables.

  1. **Test for Factor A**

     \[
     H_0 : \begin{pmatrix}
     \bar{y}_{1-\cdots} \\
     \bar{y}_{2-\cdots}
     \end{pmatrix} = \begin{pmatrix}
     \bar{y}_{1-2\cdots} \\
     \bar{y}_{2-2\cdots}
     \end{pmatrix} \quad \text{and}
     \]

     \[
     H_A : \begin{pmatrix}
     \bar{y}_{1-\cdots} \\
     \bar{y}_{2-\cdots}
     \end{pmatrix} \neq \begin{pmatrix}
     \bar{y}_{1-2\cdots} \\
     \bar{y}_{2-2\cdots}
     \end{pmatrix}
     \]

  2. **Test for Factor B**

     \[
     H_0 : \begin{pmatrix}
     \bar{y}_{1-\cdots} \\
     \bar{y}_{2-\cdots}
     \end{pmatrix} = \begin{pmatrix}
     \bar{y}_{1-2\cdots} \\
     \bar{y}_{2-2\cdots}
     \end{pmatrix} = \begin{pmatrix}
     \bar{y}_{1-3\cdots} \\
     \bar{y}_{2-3\cdots}
     \end{pmatrix},
     \]

Multivariate Research Methods
and $H_A$ : At least two sets of means are different from other.

3. **Test for Factor A×B** Our null hypothesis would test whether individual cell means are equal to factor means. On the other hand, alternative hypothesis would test whether individual cell means are not equal to factor means.

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Matrix of sum of squares and cross products</th>
<th>Degrees of freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor A</td>
<td>$H_A = \begin{pmatrix} 29.39 &amp; 31.93 \ 31.93 &amp; 34.69 \end{pmatrix}$</td>
<td>$L - 1 = 1$</td>
</tr>
<tr>
<td>Factor B</td>
<td>$H_B = \begin{pmatrix} 129.75 &amp; 113.50 \ 113.50 &amp; 122.18 \end{pmatrix}$</td>
<td>$K - 1 = 3$</td>
</tr>
<tr>
<td>Factor A×B</td>
<td>$H_{A\times B} = \begin{pmatrix} 83.67 &amp; 42.46 \ 42.46 &amp; 39.42 \end{pmatrix}$</td>
<td>$(L - 1) \times (K - 1) = 3$</td>
</tr>
<tr>
<td>Residual</td>
<td>$E = \begin{pmatrix} 148.37 &amp; 31.93 \ 31.93 &amp; 146.88 \end{pmatrix}$</td>
<td>$(\sum_{j=1}^k n_j) - (LK - 1) = 21$</td>
</tr>
<tr>
<td>Mean</td>
<td>$M = \begin{pmatrix} 3869.83 &amp; 3835.17 \ 3835.17 &amp; 3800.83 \end{pmatrix}$</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>$T = \begin{pmatrix} 4261 &amp; 4055 \ 4055 &amp; 4144 \end{pmatrix}$</td>
<td>$(\sum_{j=1}^k n_j)$</td>
</tr>
<tr>
<td>Mean Corrected</td>
<td>$T - M = \begin{pmatrix} 391.17 &amp; 219.83 \ 219.83 &amp; 343.17 \end{pmatrix}$</td>
<td>$(\sum_{j=1}^k n_j) - 1$</td>
</tr>
</tbody>
</table>

- **Test Statistic:**

We could use any of four tests summarized in previous pages, namely, Wilks’ Lambda, Pillai’s Trace, Hotelling-Lawley Trace and Roy’s greatest root. Here are estimated values of Wilks’ Lambda and associated $F$-statistic for dataset used for illustration.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Value of $F$-Statistic</th>
<th>Numerator</th>
<th>Error</th>
<th>Prob. $\geq F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.737</td>
<td>3.57</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>B</td>
<td>0.387</td>
<td>4.05</td>
<td>6</td>
<td>40</td>
</tr>
<tr>
<td>A×B</td>
<td>0.551</td>
<td>2.31</td>
<td>6</td>
<td>40</td>
</tr>
</tbody>
</table>

We should conclude from above summary that the null hypothesis for factor A and B may be rejected at $\alpha$ of 0.05. The null hypothesis about factor A×B may not be rejected. Given
that Factor B has four levels, interpretation of interaction in this instance is complicated. A closer examinations of means reveal that A₁ and B₁, A₂ and B₁; and A₁ and B₄ seem to be quite different groups from all others. Note also that some of these groups, we have very small sample and hence we may not have sufficient confidence in our findings.

- **Assumptions**

  1. Each test unit is independent from others.
  2. Each test unit has equal influence in constructing group means and variance-covariance matrices.
  3. Variations within group is a similar.
  4. Variance-covariance matrices within and between groups are about equal.
  5. Observations in each group are multivariate normally distributed.

**SAS Input for Example from Stevens**

```sas
options ps=65 ls=70 nocenter nodate;
data stev310;
input a b y1 y2 ;
cards;
  1 1 6 10
  1 2 13 16
  1 3 9 11
  1 4 21 19
  1 1 7 8
  1 2 11 15
  1 3 8 8
  1 4 18 15
  1 2 17 18
  1 3 14 9
  1 4 16 13
  1 3 13 11
  2 1 11 8
  2 2 10 12
  2 3 14 12
  2 4 11 10
  2 1 7 6
  2 2 11 13
  2 3 10 8
  2 4 9 8
  2 1 10 5
  2 2 14 10
  2 3 11 13
  2 4 8 15
  2 1 6 12
  2 4 17 12
  2 1 9 7
  2 4 13 14
  2 1 11 14
proc glm ;
class a b;
```

Multivariate Research Methods

Course: COST*6060
Comments about input.

- Option of generating all sums of squares using (ss1, ss2, ss3 and ss4) is meant to demonstrate that sums of squares differ, when number respondents per cell are different.
- Use two different manova statements is meant to indicate alternative sums of squares are used computation of errors and hypothesis testing purpose.

SAS output from above run

```
General Linear Models Procedure
Class Level Information
Class Levels Values
A 2 1 2
B 4 1 2 3 4

Number of observations in data set = 29

Dependent Variable: Y1

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>7</td>
<td>242.80574713</td>
<td>4.91</td>
<td>0.0021</td>
</tr>
<tr>
<td>Error</td>
<td>21</td>
<td>148.36666667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>28</td>
<td>391.17241379</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

R-Square C.V. Y1 Mean
0.620713 23.00974 11.5517241

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>29.39300203</td>
<td>4.16</td>
<td>0.0542</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>129.74723019</td>
<td>6.12</td>
<td>0.0037</td>
</tr>
<tr>
<td>A*B</td>
<td>3</td>
<td>83.66551491</td>
<td>3.95</td>
<td>0.0223</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type II SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>17.47972319</td>
<td>2.47</td>
<td>0.1307</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>129.74723019</td>
<td>6.12</td>
<td>0.0037</td>
</tr>
<tr>
<td>A*B</td>
<td>3</td>
<td>83.66551491</td>
<td>3.95</td>
<td>0.0223</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>12.64807256</td>
<td>1.79</td>
<td>0.1952</td>
</tr>
<tr>
<td>B</td>
<td>3</td>
<td>179.54366369</td>
<td>8.47</td>
<td>0.0007</td>
</tr>
<tr>
<td>A*B</td>
<td>3</td>
<td>83.66551491</td>
<td>3.95</td>
<td>0.0223</td>
</tr>
</tbody>
</table>
```
### Source DF Type IV SS F Value Pr > F

| A  | 1 | 12.64807256 | 1.79  | 0.1952 |
| B  | 3 | 179.54356369 | 8.47  | 0.0007 |
| A*B | 3 | 83.66514911 | 3.95  | 0.0223 |

Dependent Variable: Y2

### Source DF Sum of Squares F Value Pr > F

| Model | 7 | 196.28908046 | 4.01  | 0.0062 |
| Error | 21 | 146.88333333 |
| Corrected Total | 28 | 343.17241379 |

<table>
<thead>
<tr>
<th>R-Square</th>
<th>C.V.</th>
<th>Y2 Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.571984</td>
<td>23.10131</td>
<td>11.4482759</td>
</tr>
</tbody>
</table>

### Source DF Type I SS F Value Pr > F

| A  | 1 | 34.68711968 | 4.96  | 0.0370 |
| B  | 3 | 122.17770604 | 5.82  | 0.0047 |
| A*B | 3 | 39.42425474 | 1.88  | 0.1641 |

### Source DF Type II SS F Value Pr > F

| A  | 1 | 24.12098335 | 3.45  | 0.0774 |
| B  | 3 | 122.17770604 | 5.82  | 0.0047 |
| A*B | 3 | 39.42425474 | 1.88  | 0.1641 |

### Source DF Type III SS F Value Pr > F

| A  | 1 | 23.67902494 | 3.39  | 0.0800 |
| B  | 3 | 123.80365854 | 5.90  | 0.0044 |
| A*B | 3 | 39.42425474 | 1.88  | 0.1641 |

### Source DF Type IV SS F Value Pr > F

| A  | 1 | 23.67902494 | 3.39  | 0.0800 |
| B  | 3 | 123.80365854 | 5.90  | 0.0044 |
| A*B | 3 | 39.42425474 | 1.88  | 0.1641 |

**E = Error SS&CP Matrix**

<table>
<thead>
<tr>
<th></th>
<th>Y1</th>
<th>Y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>148.36666667</td>
<td>31.93333333</td>
</tr>
<tr>
<td>Y2</td>
<td>31.93333333</td>
<td>146.88333333</td>
</tr>
</tbody>
</table>

### Multivariate Analysis of Variance

### H = Type I SS&CP Matrix for A

<table>
<thead>
<tr>
<th></th>
<th>Y1</th>
<th>Y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>29.393002028</td>
<td>31.930527383</td>
</tr>
<tr>
<td>Y2</td>
<td>31.930527383</td>
<td>34.687119675</td>
</tr>
</tbody>
</table>
Manova Test Criteria and Exact F Statistics for the Hypothesis of no Overall A Effect

\( H = \text{Type I SS&CP Matrix for A} \quad E = \text{Error SS&CP Matrix} \)

\[
\begin{array}{l|lll|l}
\text{Statistic} & \text{Value} & \text{F} & \text{Num DF} & \text{Den DF} & \text{Pr > F} \\
\hline
\text{Wilks' Lambda} & 0.73696924 & 3.57412 & 2 & 20 & 0.0471 \\
\text{Pillai's Trace} & 0.26303076 & 3.57412 & 2 & 20 & 0.0471 \\
\text{Hotelling-Lawley Trace} & 0.35741158 & 3.57412 & 2 & 20 & 0.0471 \\
\text{Roy's Greatest Root} & 0.35741158 & 3.57412 & 2 & 20 & 0.0471 \\
\end{array}
\]

\( H = \text{Type I SS&CP Matrix for B} \)

\[
\begin{array}{l|ll}
\text{Y1} & \text{Y2} \\
\hline
\text{Y1} & 129.74723019 & 113.50207237 \\
\text{Y2} & 113.50207237 & 122.17770604 \\
\end{array}
\]

Multivariate Analysis of Variance

Manova Test Criteria and F Approximations for the Hypothesis of no Overall B Effect

\( H = \text{Type I SS&CP Matrix for B} \quad E = \text{Error SS&CP Matrix} \)

\[
\begin{array}{l|lll|l}
\text{Statistic} & \text{Value} & \text{F} & \text{Num DF} & \text{Den DF} & \text{Pr > F} \\
\hline
\text{Wilks' Lambda} & 0.38698893 & 4.04999 & 6 & 40 & 0.0029 \\
\text{Pillai's Trace} & 0.66833149 & 3.51313 & 6 & 42 & 0.0066 \\
\text{Hotelling-Lawley Trace} & 1.44110229 & 4.56349 & 6 & 38 & 0.0014 \\
\text{Roy's Greatest Root} & 1.33393766 & 9.33756 & 3 & 21 & 0.0004 \\
\end{array}
\]

\text{NOTE: F Statistic for Roy's Greatest Root is an upper bound.} \quad \text{NOTE: F Statistic for Wilks' Lambda is exact.}

\( H = \text{Type I SS&CP Matrix for A*B} \)

\[
\begin{array}{l|ll}
\text{Y1} & \text{Y2} \\
\hline
\text{Y1} & 83.665514905 & 42.461653117 \\
\text{Y2} & 42.461653117 & 39.424254743 \\
\end{array}
\]

Manova Test Criteria and F Approximations for the Hypothesis of no Overall A*B Effect

\( H = \text{Type I SS&CP Matrix for A*B} \quad E = \text{Error SS&CP Matrix} \)

\[
\begin{array}{l|lll|l}
\text{Statistic} & \text{Value} & \text{F} & \text{Num DF} & \text{Den DF} & \text{Pr > F} \\
\hline
\text{Wilks' Lambda} & 0.55108091 & 2.31384 & 6 & 40 & 0.0519 \\
\text{Pillai's Trace} & 0.68859196 & 2.26289 & 6 & 42 & 0.0555 \\
\text{Hotelling-Lawley Trace} & 0.74262456 & 2.35164 & 6 & 38 & 0.0498 \\
\text{Roy's Greatest Root} & 0.62798679 & 4.39591 & 3 & 21 & 0.0150 \\
\end{array}
\]

\text{NOTE: F Statistic for Roy's Greatest Root is an upper bound.} \quad \text{NOTE: F Statistic for Wilks' Lambda is exact.}

Multivariate Analysis of Variance
**Multivariate Analysis of Variance**

**E = Error SS&CP Matrix**

<table>
<thead>
<tr>
<th></th>
<th>Y1</th>
<th>Y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>148.36666667</td>
<td>31.93333333</td>
</tr>
<tr>
<td>Y2</td>
<td>31.93333333</td>
<td>146.88333333</td>
</tr>
</tbody>
</table>

**Multivariate Analysis of Variance**

**H = Type III SS&CP Matrix for A**

<table>
<thead>
<tr>
<th></th>
<th>Y1</th>
<th>Y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>12.648072562</td>
<td>17.305895692</td>
</tr>
<tr>
<td>Y2</td>
<td>17.305895692</td>
<td>23.679024943</td>
</tr>
</tbody>
</table>

**Manova Test Criteria and Exact F Statistics for the Hypothesis of no Overall A Effect**

**H = Type III SS&CP Matrix for A**  
**E = Error SS&CP Matrix**

*S=1  M=0  N=9*

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>F</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks' Lambda</td>
<td>0.02963475</td>
<td>2.0535</td>
<td>2</td>
<td>20</td>
<td>0.1545</td>
</tr>
<tr>
<td>Pillai's Trace</td>
<td>0.17036525</td>
<td>2.0535</td>
<td>2</td>
<td>20</td>
<td>0.1545</td>
</tr>
<tr>
<td>Hotelling-Lawley Trace</td>
<td>0.2053497</td>
<td>2.0535</td>
<td>2</td>
<td>20</td>
<td>0.1545</td>
</tr>
<tr>
<td>Roy's Greatest Root</td>
<td>0.2053497</td>
<td>2.0535</td>
<td>2</td>
<td>20</td>
<td>0.1545</td>
</tr>
</tbody>
</table>

**Multivariate Analysis of Variance**

**H = Type III SS&CP Matrix for B**

<table>
<thead>
<tr>
<th></th>
<th>Y1</th>
<th>Y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>179.54356369</td>
<td>133.9995935</td>
</tr>
<tr>
<td>Y2</td>
<td>133.9995935</td>
<td>123.80365854</td>
</tr>
</tbody>
</table>

**Manova Test Criteria and F Approximations for the Hypothesis of no Overall B Effect**

**H = Type III SS&CP Matrix for B**  
**E = Error SS&CP Matrix**

*S=2  M=0  N=9*

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>F</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks' Lambda</td>
<td>0.33927437</td>
<td>4.77879</td>
<td>6</td>
<td>40</td>
<td>0.0009</td>
</tr>
<tr>
<td>Pillai's Trace</td>
<td>0.73050266</td>
<td>4.02799</td>
<td>6</td>
<td>42</td>
<td>0.0028</td>
</tr>
<tr>
<td>Hotelling-Lawley Trace</td>
<td>1.74180149</td>
<td>5.5157</td>
<td>6</td>
<td>38</td>
<td>0.0003</td>
</tr>
<tr>
<td>Roy's Greatest Root</td>
<td>1.61440771</td>
<td>11.3009</td>
<td>3</td>
<td>21</td>
<td>0.0001</td>
</tr>
</tbody>
</table>

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

**H = Type III SS&CP Matrix for A*B**

<table>
<thead>
<tr>
<th></th>
<th>Y1</th>
<th>Y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y1</td>
<td>83.665514905</td>
<td>42.461653117</td>
</tr>
<tr>
<td>Y2</td>
<td>42.461653117</td>
<td>39.424254743</td>
</tr>
</tbody>
</table>

---

**Multivariate Research Methods**  
**Course: COST*6060**
### Multivariate Analysis of Variance

Manova Test Criteria and F Approximations for the Hypothesis of no Overall A*B Effect

\[ H = \text{Type III SS&CP Matrix for A*B} \quad E = \text{Error SS&CP Matrix} \]

\[ S=2 \quad M=0 \quad N=9 \]

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>F</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks’ Lambda</td>
<td>0.55108091</td>
<td>2.31384</td>
<td>6</td>
<td>40</td>
<td>0.0519</td>
</tr>
<tr>
<td>Pillai’s Trace</td>
<td>0.48859196</td>
<td>2.26289</td>
<td>6</td>
<td>42</td>
<td>0.0555</td>
</tr>
<tr>
<td>Hotelling-Lawley Trace</td>
<td>0.74262456</td>
<td>2.35164</td>
<td>6</td>
<td>38</td>
<td>0.0498</td>
</tr>
<tr>
<td>Roy’s Greatest Root</td>
<td>0.62798679</td>
<td>4.39591</td>
<td>3</td>
<td>21</td>
<td>0.0150</td>
</tr>
</tbody>
</table>

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

**NOTE:** F Statistic for Wilks’ Lambda is exact.

### Multivariate Regression Analysis

One natural extension of MANOVA model is multiple dependent variable regression model. A model that has received considerable attention in the marketing as well as economics is share allocation model. We will examine one such model below. Consider consumer expenditure on durables, nondurables and services. Suppose we postulated indirect utility function for an average consumer. With three goods (durables, nondurables and services), suppose we denote price of \( i \)th good to be \( p_i \) and \( \mu \) to be total expenditure, then we may write following function as an indirect utility for a representative consumer.

\[
V(P, \mu) = \log \mu - \sum_{i=1}^{n} \alpha_i \log p_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij} \log p_i \log p_j
\]  

(1)

with following restrictions \( \beta_{ij} = \beta_{ji} \) for all \( i \) and \( j \), \( \sum_{i=1}^{n} \alpha_i = 1 \) and \( \sum_{i=1}^{n} \beta_{ij} = 0 \) for all \( j \). To determine share of expenditure to the \( i \)th good, we need to use Roy’s identity which in general is

\[
S_i = -\frac{\partial V/\partial \log p_i}{\partial V/\partial \log \mu}
\]

One may write following share equation for above indirect utility function,

\[
S_i = \alpha_i + \sum_{i=1}^{n} \beta_{ij} \log p_j \quad \text{for all } i.
\]

---

8Direct utility function contains quantity chosen, while indirect utility function is minimized with respect to prices subject to budget constraint for both utility functions. To account for minimization we have changed sign of logarithm of prices to be negative.


---

Multivariate Research Methods

Course: COST*6060
Since shares must sum to 1, and there are \( n \) such equations one of the equation in above system is redundant and must be deleted from analysis. Suppose we started with the translog indirect utility function such as

\[
V(P, \mu) = -\sum_{i=1}^{n} \alpha_i \log(p_i/\mu) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \beta_{ij} \log(p_i/\mu) \log(p_j/\mu) \tag{2}
\]

with similar set of parameter restrictions as before or \( \beta_{ij} = \beta_{ji} \) for all \( i \) and \( j \) and \( \sum_{i=1}^{n} \alpha_i = 1 \).

The corresponding share equation is given by

\[
S_i = \frac{\alpha_i + \sum_{j=1}^{n} \beta_{ij} \log(p_j) - \log(\mu) \sum_{j=1}^{n} \beta_{ij}}{1 + \sum_{k=1}^{n} \sum_{i=1}^{n} \beta_{ki} \log(p_i/\mu)} \tag{3}
\]

There are atleast three important distinctions between MANOVA and above share model. First note that our model development upto this point does not depend on statistical reasoning. It is entirely based on our assumptions about consumption behaviour. Second there is no statement about error component. In other words, the economic theory does not specify that error terms are normally distributed or some other distribution such as Dirichlet distribution. Finally, above described share model is not linear in parameters and procedure such as generalized linear model (for example, \textsc{PROC GLM}) will not be adequate to estimate parameters.

There is an alternative modeling framework proposed in the marketing which does not depend upon consumer behaviour propositions. Suppose we argue that brand share depends upon the ratio of brand’s own attraction to the sum of all brands competing in the market. Let us denote \( S_i \) and \( A_i \) to be brand share and attraction respectively of brand \( i \). Assuming that there are \( n \) competitive brands in the market, we may write

\[
S_i = \frac{A_i}{\sum_{j=1}^{n} A_j}. \tag{4}
\]

To complete modelling specification we need to indicate characteristics of attraction or link between marketing variables and attraction\(^{11}\). Suppose that price is only variable that we want to link to brand shares. Then we may write \( A_i = \exp[\alpha_i + \sum_{j=1}^{n} \beta_{ij} \log(p_j)] \). On the surface such model appear to be plausible. Note that since shares for all brands within a product category must sum to 1. This results in \((n-1)\) equations to be independent and only \( n^2 - n \) parameters could be estimated. In other words, of \( n \) attractions that are to be estimated, parameters for only \( n - 1 \) could be estimated. This results in often imposing some arbitrary constraints on parameters. For example, some have argued that one brand can be treated as a base brand

\(^{10}\)Imposing following restrictions on the utility function \( \sum_{i=1}^{n} \beta_{ij} = 0 \) for all \( j \) would result in utility function described in equation (1).

\(^{11}\)Note that all attractions must be greater than or equal to zero and if attraction for a brand is zero, then brand share for that brand must be zero as well. Finally, we would expect that shares for all brand must sum to 1.
and comparison could be made to with respect to the base brand. Others have argued that one should impose constraint such that parameters in each column sum to a constant or zero\(^{12}\).

- **Parameter Estimation**

To estimate parameters of equation (3) is challenging because parameters are embedded in not linear equations. Software package such as SAS allows one to estimate complicated model given by the indirect utility function given in equation (2) and share equations given in (3). Following SAS program provides one approach to estimate share equations for resulting from consumer expenditure for three commodity groups (non-durables, durables and services). Data used in this instance came from the Bureau of Labour Statistics in the United States for year 1982Q1 to 1998Q4.

```sas
options nocenter nodate ps = 70 ls =80 nonumber;
filename uscons dde 'excel|US_CONS.xls!r2c3:r69c8';
data consum;
  infile uscons dlm='09'x notab dsd missover;
  input serv nondur dur servpr nondurpr durpr;
  obs = _n_;
  totcons = serv + nondur + dur;
  sersh = serv/totcons;
  nondursh = nondur/totcons;
  dursh = dur / totcons;
  servpr = servpr /100;
  nondurpr = nondurpr /100;
  durpr = durpr/100;
proc model data=consum;
  num1 = bpss*log(servpr) + bpsn*log(nondurpr) + bpsd*log(durpr);
  num2 = bpsn*log(servpr) + bpnn*log(nondurpr) + bpnd*log(durpr);
  num3 = bpsd*log(servpr) + bpnd*log(nondurpr) + bpdd*log(durpr);
  sum1 = bpss + bpsn + bpsd;
  sum2 = bpsn + bpnn + bpnd;
  sum3 = bpss + bpnn + bpdd;
  sersh = (as + num1 + sum1*log(totcons))/(as + an + ad + num1 + num2 + num3 + (sum1+sum2+sum3)*log(totcons));
  dursh = (ad + num3 + sum3*log(totcons))/(as + an + ad + num1 + num2 + num3 + (sum1+sum2+sum3)*log(totcons));
  nondursh = (an + num2 + sum2*log(totcons))/(as + an + ad + num1 + num2 + num3 + (sum1+sum2+sum3)*log(totcons));
  restrict as + an + ad = 1;
  restrict sum1 + sum2 + sum3 = 0;
  fit sersh dursh / fiml out=predsh outpredicted;
run;
data new1(drop=sersh dursh);
  set predsh;
  psersh = sersh;
  pnondush = nondursh;
  pdursh = 1 - sersh - nondursh;
  obs = _n_;
proc sort data = new1; by obs; run;
proc sort data=consum; by obs; run;
data new;
  merge new1 consum;
```

\(^{12}\)There are simpler forms of above model. One specification involves imposing constraint such that \(\beta_{ij} = 0\) when \(i \neq j\) with only \(n\) parameters estimated for each marketing variable. Since parameters in such case vary by brand, model refered as differential effects model. Finally, imposing constraints that \(\beta_{ij} = 0\) when \(i \neq j\) and all \(\beta_{ii} = \beta\) for all \(i\), the model reduces to the constant effects competitive interactions model while the general model is refered as cross effects.
Multivariate Analysis of Variance

by obs;
proc reg data=new ;
model sersh = psersh /dw;
model dursh = pdursh/dw;
model nondursh = pnondush/dw;
run;

SAS Output resulting from above code

MODEL Procedure
FIML Estimation

Nonlinear FIML Summary of Residual Errors

<table>
<thead>
<tr>
<th>Equation</th>
<th>DF</th>
<th>SSE</th>
<th>MSE</th>
<th>Root MSE</th>
<th>R-Square</th>
<th>Adj R-Sq</th>
</tr>
</thead>
<tbody>
<tr>
<td>SERSH</td>
<td>3.5</td>
<td>0.00160</td>
<td>0.00002488</td>
<td>0.0049881</td>
<td>0.3455</td>
<td>0.3201</td>
</tr>
<tr>
<td>NONDURSH</td>
<td>3.5</td>
<td>0.0002195</td>
<td>3.33893E-6</td>
<td>0.0018273</td>
<td>0.9835</td>
<td>0.9828</td>
</tr>
</tbody>
</table>

Nonlinear FIML Parameter Estimates

| Parameter | Estimate | Std Err | Ratio | Prob>|T| | Label |
|-----------|----------|---------|-------|--------|-------|-------|
| BPSS      | -0.015910 | 0.02379 | -0.67 | 0.5061 |
| BPSN      | -0.017917 | 0.02043 | -0.88 | 0.3838 |
| BPSD      | 0.073489  | 0.03798 | 1.93  | 0.0574 |
| BPND      | 0.00489379| 0.02462 | 0.20  | 0.8431 |
| BPDD      | -0.039239 | 0.02043 | -2.85 | 0.0059 |
| AS        | 0.231003  | 0.18273 | 1.26  | 0.2107 |
| AN        | 0.750268  | 0.06365 | 11.79 | 0.0001 |
| AD        | 0.018729  | 0.039638| 0.50  | 0.6083 |
| Restrict1 | -55.983817| 60.45666| -0.93 | 0.3586 SUM1 + SUM2 + SUM3 = 0 |
| Restrict0 | 0.061511  | 14.53666| 0.00  | 0.9967 AS + AN + AD = 1 |

Number of Observations

<table>
<thead>
<tr>
<th>Used</th>
<th>Missing</th>
</tr>
</thead>
<tbody>
<tr>
<td>68</td>
<td>0</td>
</tr>
</tbody>
</table>

Model: MODEL1
Dependent Variable: SERSH

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>0.00085</td>
<td>0.00085</td>
<td>34.842</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>66</td>
<td>0.00160</td>
<td>0.00002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>67</td>
<td>0.00245</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Root MSE   | 0.00493 | R-square | 0.3455 |
| Dep Mean   | 0.66045 | Adj R-sq | 0.3356 |
| C.V.       | 0.87962 |          |        |
Multivariate Analysis of Variance

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > |T| |
|----------|----|-------------------|----------------|------------------------|---------|---|
| INTERCEP | 1  | 0.005058          | 0.09409266     | 0.054                  | 0.9573  |
| PSERSH   | 1  | 0.990971          | 0.16788428     | 5.903                  | 0.0001  |

Durbin-Watson D 0.226
(For Number of Obs.) 68
1st Order Autocorrelation 0.865

Model: MODEL2
Dependent Variable: NONDURSH

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>0.01281</td>
<td>0.01281</td>
<td>3925.376</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>66</td>
<td>0.0002153421</td>
<td>3.262759E-6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>67</td>
<td>0.01302</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.00181
Dep Mean 0.31992
C.V. 0.56461

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > |T| |
|----------|----|-------------------|----------------|------------------------|---------|---|
| INTERCEP | 1  | -0.000390         | 0.00511718     | -0.076                 | 0.9395  |
| PNONDUSH | 1  | 1.001216          | 0.01598038     | 62.653                 | 0.0001  |

Durbin-Watson D 0.457
(For Number of Obs.) 68
1st Order Autocorrelation 0.738

Model: MODEL1
Dependent Variable: SERSH

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>0.00085</td>
<td>0.00085</td>
<td>34.842</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>66</td>
<td>0.00160</td>
<td>0.00002</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>67</td>
<td>0.00245</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.00493
Dep Mean 0.56045
C.V. 0.87982

Parameter Estimates
Multivariate Analysis of Variance

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > |T| |
|----------|----|-------------------|----------------|------------------------|--------|---|
| INTERCEP | 1  | 0.005058          | 0.09409265     | 0.054                 | 0.9573 |
| PSERSH   | 1  | 0.990971          | 0.16788428     | 5.903                 | 0.0001 |

Durbin-Watson D 0.226  
(For Number of Obs.) 68  
1st Order Autocorrelation 0.865

Model: MODEL2  
Dependent Variable: DURSH

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>0.00756</td>
<td>0.00756</td>
<td>235.247</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>66</td>
<td>0.00212</td>
<td>0.00003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>67</td>
<td>0.00969</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.00567  
Dep Mean 0.11963  
C.V. 4.73968

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | T for H0: Parameter=0 | Prob > |T| |
|----------|----|-------------------|----------------|------------------------|--------|---|
| INTERCEP | 1  | -0.002910         | 0.00801901     | -0.363                 | 0.7178 |
| PDURSH   | 1  | 1.024354          | 0.06678644     | 15.338                 | 0.0001 |

Durbin-Watson D 0.216  
(For Number of Obs.) 68  
1st Order Autocorrelation 0.860

Model: MODEL3  
Dependent Variable: NONDURSH

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Prob&gt;F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>0.01281</td>
<td>0.01281</td>
<td>3925.376</td>
<td>0.0001</td>
</tr>
<tr>
<td>Error</td>
<td>66</td>
<td>0.0002153421</td>
<td>3.2627591E-6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>67</td>
<td>0.01302</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.00181  
Dep Mean 0.31992  
C.V. 0.56461

Parameter Estimates
To estimate parameters resulting from equation (4) is relatively straightforward, although interpreting parameters is still complicated. In general to estimate parameters in equation (4), we may take logarithms of both sides and write

$$\log(S_i) = \log(A_i) - \log\left(\sum_{j=1}^{n} A_j\right), \quad (5)$$

and there will be $n$ such equations in $n$ brand market. Similar equation may be written for a base brand or brand $n$. That is,

$$\log(S_n) = \log(A_n) - \log\left(\sum_{j=1}^{n} A_j\right). \quad (6)$$

If we subtract both sides of equation (5) and (6), we may write,

$$\log(S_i) - \log(S_n) = \log(A_i) - \log(A_n)$$

and there will be $n - 1$ such equations. Moreover, above model can be estimated using any regression package. To illustrate above model and its estimation, I compiled observations from Advertising Age’s web page on revenue and advertising spending for the top 10 burger restaurant chains in the United States (see table below).

### Top 10 fast-food burger restaurants in the US

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>McDonald’s</td>
<td>43.10</td>
<td>627.3</td>
<td>43.30</td>
<td>571.9</td>
<td>42.20</td>
<td>577.8</td>
<td>41.10</td>
<td>597.6</td>
</tr>
<tr>
<td>Burger King</td>
<td>21.90</td>
<td>403.7</td>
<td>22.60</td>
<td>407.5</td>
<td>21.40</td>
<td>423.2</td>
<td>20.90</td>
<td>360.5</td>
</tr>
<tr>
<td>Wendy’s</td>
<td>12.20</td>
<td>217.8</td>
<td>11.60</td>
<td>188.4</td>
<td>11.90</td>
<td>171.9</td>
<td>11.50</td>
<td>155.0</td>
</tr>
<tr>
<td>Hardee’s</td>
<td>5.60</td>
<td>48.1</td>
<td>6.50</td>
<td>50.4</td>
<td>8.80</td>
<td>80.5</td>
<td>10.10</td>
<td>82.2</td>
</tr>
<tr>
<td>Jack in the Box</td>
<td>4.00</td>
<td>63.5</td>
<td>3.60</td>
<td>51.2</td>
<td>3.30</td>
<td>50.5</td>
<td>3.20</td>
<td>44.1</td>
</tr>
<tr>
<td>Sonic Drive-In</td>
<td>3.70</td>
<td>32.2</td>
<td>3.30</td>
<td>28.1</td>
<td>2.90</td>
<td>23.0</td>
<td>2.60</td>
<td>15.7</td>
</tr>
<tr>
<td>Carl’s Jr</td>
<td>2.00</td>
<td>37.4</td>
<td>1.70</td>
<td>34.1</td>
<td>1.70</td>
<td>27.8</td>
<td>1.60</td>
<td>25.0</td>
</tr>
<tr>
<td>Whataburger</td>
<td>1.10</td>
<td>8.2</td>
<td>1.10</td>
<td>6.7</td>
<td>1.20</td>
<td>5.4</td>
<td>1.10</td>
<td>4.6</td>
</tr>
<tr>
<td>White Castle</td>
<td>1.00</td>
<td>10.6</td>
<td>1.00</td>
<td>10.5</td>
<td>0.90</td>
<td>10.3</td>
<td>0.90</td>
<td>10.0</td>
</tr>
<tr>
<td>Steak n Shake</td>
<td>1.00</td>
<td>5.7</td>
<td>0.90</td>
<td>5.7</td>
<td>0.80</td>
<td>4.1</td>
<td>0.70</td>
<td>3.3</td>
</tr>
<tr>
<td><strong>Total top 10</strong></td>
<td><strong>95.60</strong></td>
<td><strong>1454.5</strong></td>
<td><strong>95.60</strong></td>
<td><strong>1354.5</strong></td>
<td><strong>95.10</strong></td>
<td><strong>1374.5</strong></td>
<td><strong>93.70</strong></td>
<td><strong>1298.0</strong></td>
</tr>
<tr>
<td><strong>Remaining</strong></td>
<td><strong>4.40</strong></td>
<td><strong>19.5</strong></td>
<td><strong>4.40</strong></td>
<td><strong>23.0</strong></td>
<td><strong>4.90</strong></td>
<td><strong>6.4</strong></td>
<td><strong>6.30</strong></td>
<td><strong>2.1</strong></td>
</tr>
<tr>
<td><strong>Total market in dollars</strong></td>
<td><strong>44.05</strong></td>
<td><strong>1474.0</strong></td>
<td><strong>41.83</strong></td>
<td><strong>1377.5</strong></td>
<td><strong>40.6</strong></td>
<td><strong>1380.9</strong></td>
<td><strong>39.1</strong></td>
<td><strong>1300.1</strong></td>
</tr>
</tbody>
</table>

Advertising expenses are in millions of dollars and total market is in billion dollars. All of above observations were obtained from Advertising Age’s web page, [http://adage.com/dataplace/](http://adage.com/dataplace/)
Note at the outset that overall market has remained relatively stable over these four years with modest increase in total revenue to the industry. Consequently, one could argue that higher spending on advertising by organization leads to decrease in share to competitor(s), holding all else constant. Using regression package within Excel, following estimates were obtained. Following estimate suggest that media advertising has strong influence on market share. The last two columns in following table also indicate that constructed model has good forecasting accuracy, one that has room for improvement.

<table>
<thead>
<tr>
<th></th>
<th>Coeff.</th>
<th>Std. error</th>
<th>t-Stat</th>
<th>P-value</th>
<th>2000 Shares</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Actual</td>
<td>Predicted</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adv</td>
<td>0.6000</td>
<td>0.07</td>
<td>8.84</td>
<td>0.000</td>
<td>43.1</td>
</tr>
<tr>
<td>McDonald’s</td>
<td>1.0004</td>
<td>0.33</td>
<td>3.00</td>
<td>0.005</td>
<td>43.5</td>
</tr>
<tr>
<td>Burger King</td>
<td>0.5692</td>
<td>0.31</td>
<td>1.86</td>
<td>0.073</td>
<td>21.1</td>
</tr>
<tr>
<td>Wendy’s</td>
<td>0.4302</td>
<td>0.25</td>
<td>1.70</td>
<td>0.100</td>
<td>12.7</td>
</tr>
<tr>
<td>Hardee’s</td>
<td>0.6161</td>
<td>0.18</td>
<td>3.36</td>
<td>0.002</td>
<td>5.3</td>
</tr>
<tr>
<td>Jack in the Box</td>
<td>-0.0290</td>
<td>0.17</td>
<td>-0.17</td>
<td>0.866</td>
<td>4.4</td>
</tr>
<tr>
<td>Sonic Drive-Ins</td>
<td>0.3103</td>
<td>0.12</td>
<td>2.57</td>
<td>0.015</td>
<td>4.0</td>
</tr>
<tr>
<td>Carl’s Jr</td>
<td>-0.4139</td>
<td>0.14</td>
<td>-3.03</td>
<td>0.005</td>
<td>2.1</td>
</tr>
<tr>
<td>Whataburger</td>
<td>0.1184</td>
<td>0.05</td>
<td>2.46</td>
<td>0.020</td>
<td>1.4</td>
</tr>
<tr>
<td>White Castle</td>
<td>-0.3705</td>
<td>0.07</td>
<td>-5.24</td>
<td>0.000</td>
<td>1.1</td>
</tr>
<tr>
<td>Steak n Shake</td>
<td>0.0000</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>1.1</td>
</tr>
</tbody>
</table>

**Testing Multivariate Normality**

The purpose of this material is to provide procedures that can be used to evaluate the multivariate normality. If tests reveal problems, then it is advisable to turn to the alternative approaches to analysis, including transformation. For illustration purpose, bivariate normal distribution with means for both variables to be zero and variance for $y_1$ and $y_2$ to be 2 and 1 respectively. Finally both variables were uncorrelated. A distribution plot such variables is shown below.

If these two variables were highly correlated, (say corr($y_1, y_2$) = 0.8), then distribution plot would be as follows (see distribution below).

To detect departures from univariate normality, we examined third and fourth moments about the mean. To detect departures from multivariate normality, we examine the vector of means.
Multivariate Analysis of Variance

and their covariances. If observations are multivariate normally distributed, then the vector of means and their covariances are sufficient to describe those observations. That is, the multivariate coefficient of skewness \( b_{1,p} \), where \( p \) is number of variables for which we are interested in testing multivariate normality is zero. Furthermore, the multivariate coefficient of kurtosis \( b_{2,p} \) is equal to \( p \times (p + 2) \). We could use these indices as our null hypotheses to test departures from multivariate normality for observed variables.

Below we will look at the second, third and fourth moments for a sample below.

**The Sample Variance-Covariance Matrix**

Suppose that \( Y_{ij} \) are value of observed variables with subscript \( i \) indicating variable number \((i = 1, \ldots, p)\) and \( j \) indicating observation number \((j = 1, \ldots, n \text{ and } p < n)\). Then, the vector of means can be written as \( Y_i \), and it is computed as

\[
Y_i = \frac{1}{n} \sum_{j=1}^{n} Y_{ij}.
\]

The sample variance-covariance (\( S \)) matrix is computed as

\[
S_{il} = \frac{1}{n - 1} \sum_{i=1}^{p} \sum_{l=1}^{p} \sum_{j=1}^{n} (Y_{ij} - Y_{il})(Y_{lj} - Y_{ll}).
\]

Note that \( S \) is \( p \times p \) matrix with diagonal elements are variance associated with \( p \)th variable.

**Multivariate Skewness**

Mardia (1970)\(^\text{13}\) proposed multivariate tests of skewness. His sample measure of skewness, \( b_{1,p} \) is

\[
b_{1,p} = \frac{1}{n^2} \sum_{i=1}^{p} \sum_{l=1}^{p} \sum_{j=1}^{n} \left( (Y_{ij} - Y_{il})S_{il}^{-1}(Y_{lj} - Y_{ll}) \right)^3.
\]

If a set of variables are multivariate normally distributed, then we would expect that \( b_{1,p} \) to be equal to zero. To test the departure from normality, note that variable \( n kb_{1,p} \) is chi-square

distributed with \( \frac{p(p + 1)(p + 2)}{6} \) degrees of freedom, where \( k = \frac{(p + 1)(n + 1)(n + 3)}{n[(n + 1)(p + 1) - 6]} \).

Small (1980) proposed a measure of multivariate skewness based on normalized values of univariate skewness and estimate of the correlation matrix. Suppose the normalized measure of skewness for \( i \)th variable is denoted by \( Z_{\sqrt{b_{1i}}} \), then, Small’s measure of skewness (\( Q_1 \)) is equal to

\[
Q_1 = \sum_{l=1}^{p} \sum_{i=1}^{p} Z_{\sqrt{b_{1i}}} r_{li}^3 Z_{\sqrt{b_{1i}}}
\]

where \( r_{li} \) is the estimated correlation between variable \( l \) and \( i \). This measure (\( Q_1 \)) is chi-square distribution with \( p \) degrees of freedom.

Srivastava (1984) suggested measures of multivariate skewness and kurtosis that are computed from the skewness and kurtosis coefficients of principal components (PC) extracted from the covariance matrix. Let \( b'_{1i} \) is the square of the skewness coefficient for the \( i \)th PC extracted from \( S \). Srivastava’s measure of skewness is defined as

\[
(b_{1p})^2 = \frac{1}{p} \sum_{i=1}^{p} b'_{1i}
\]

and \( \frac{np}{6} (b_{1p})^2 \) is distributed as chi-square with \( p \) degrees of freedom.

**Multivariate Kurtosis**

Mardia’s measure of multivariate kurtosis (\( b_{2,p} \)) for a sample is

\[
b_{2,p} = \frac{1}{n} \sum_{i=1}^{p} \sum_{l=1}^{p} \sum_{j=1}^{n} (Y_{ij} - Y_{i.}) S_{il}^{-1} (Y_{lj} - Y_{l.})^2.
\]

Note that \( b_{2,p} \) is normally distributed with the mean of \( p(p+2) \) and variance of \( \frac{8p(p+2)}{n} \).

Small’s measure of multivariate kurtosis (\( Q_2 \)) is defined by

\[
Q_2 = \sum_{l=1}^{p} \sum_{i=1}^{p} Z_{b_{2l}} r_{li}^4 Z_{b_{2i}}
\]

where \( Z_{b_{2i}} \) is normalized measure of kurtosis for variable \( i \). Measure \( Q_2 \), multivariate measure of kurtosis is also chi-square distributed with \( p \) degrees of freedom\(^{16}\).


\(^{16}\)Small also proposed an omnibus measure \( Q_3 \) to determine departure from multivariate normality and \( Q_3 = Q_1 + Q_2 \). This omnibus measure is distributed as chi-square with \( 2p \) degrees of freedom.
Srivastava’s measure of multivariate kurtosis is computed from the kurtosis coefficients of principal components (PC) extracted from the covariance matrix. Let $b_{2i}$ be the kurtosis coefficient for the $i$th PC extracted from $S$. Srivastava’s measure of kurtosis is defined as

$$b_{2p} = \frac{1}{p} \sum_{i=1}^{p} b_{2i}$$

and $\sqrt{\frac{np}{24}}(b_{2p} - 3)$ is distributed normally with the mean of zero and the standard deviation of 1.