Multivariate Analysis of Variance and Covariance Hypothesis Tests

For multivariate analysis of variance (MANOVA) and multivariate analysis of covariance (MANCOVA), SAS (and other packages) display a number of tests which are:

- Wilks' lambda
- Pillai's trace
- Hotelling-Lawley trace
- Roy's maximum root

In the case of MANOVA with one class variable, the null and alternative hypotheses are:

\[ H_0: \bar{\mu}_1 = \bar{\mu}_2 = \bar{\mu}_3 = \ldots \bar{\mu}_k \]

\[ H_0: \text{not all centroids are equal for all levels of the class variable} \]

In the case of two class variables, there would be three null hypotheses: i) there is no interaction between the two class variables; and, then, if there is no interaction: ii) the centroids are the same for the levels of the first class variable; and iii) the centroids are the same for the second class variable. In the case of MANCOVA, the same hypotheses are tested (for one class variable or for two class variables), but the tests are based on centroids that are adjusted for the covariate(s).

In all cases, there will be a B (between matrix) and a within (E) matrix for these hypothesis tests. SAS labels the within as E for error matrix, and the B as H for hypothesis. You must ask for the MANOVA tests from SAS, and specify what is to be used as the E and the H matrices in your test. For example, in the case of MANOVA with a single class variable, method:

```
PROC GLM;
    CLASS method;
    MODEL logn logp logk = method;
    LSMEANS method / stderr pdiff;
    MANOVA H=method/printe printh;
    output out=glmout predicted=predict1 predict2 predict3
        residual=resid1 resid2 resid3;
    TITLE 'MULTIVARIATE ANALYSIS OF VARIANCE';
run;
```
The H is method, and the E is the default and therefore, is not stated. In the case of MANOVA with two class variables, there are three hypothesis tests. Using an example with the class variables BECZONE and RSKGRP, and a covariate, AGE, the MANCOVA to look for differences in centroids for the y-variables, Indcay Inwst, is obtained using:

```* assumes that beczone is fixed and rskgrp is random; PROC GLM;   CLASS beczone rskgrp;   MODEL lndcay lnwst=beczone rskgrp beczone*rskgrp age;   LSMEANS beczone*rskgrp;   MANOVA H=beczone E=beczone*rskgrp/printe printh;   MANOVA H=rskgrp /printe printh;   MANOVA H=beczone*rskgrp/printe printh;   MANOVA H=age/printe printh;   TITLE 'MULTIVARIATE ANALYSIS OF COVARIANCE';   Output out=pout r=resid1 resid2 p=pred1 pred2; RUN;```

Why are there four tests? The four tests are all the same in the case of one class variable with two levels. Otherwise, different tests are more robust, in that the p-values are more accurate, to failures to meet the assumptions of analysis of variance, MANOVA, and MANCOVA.

- In terms of the univariate F tests (ANOVA or ANCOVA; one y variable at a time), the F-test is reasonably robust to small departures from normality.
- All four multivariate tests are robust to failures to meet the homogeneity of covariance assumption, IF the number of observations in each level of the class variable are equal.
- Pillai’s trace is more robust to failure to meet this assumption, in the case of unequal numbers of observations in each level of the class variable.
- Wilk’s lambda, Hotelling’s T test, and Roy’s greatest root are more powerful (i.e., can detect smaller differences) than Pillai’s when the first treatment level is most important (i.e., the first treatment level is a control).

**DETAILS** of each test from the SAS documentation follow
**Multivariate Analysis of Variance**

If you fit several dependent variables to the same effects, you may want to make tests jointly involving parameters of several dependent variables. Suppose you have \( p \) dependent variables, \( k \) parameters for each dependent variable, and \( n \) observations. The models can be collected into one equation:

\[
Y = X\beta + \varepsilon
\]

where \( Y \) is \( n \times p \), \( X \) is \( n \times k \), \( \beta \) is \( k \times p \), and \( \varepsilon \) is \( n \times p \). Each of the \( p \) models can be estimated and tested separately. However, you may also want to consider the joint distribution and test the \( p \) models simultaneously.

For multivariate tests, you need to make some assumptions about the errors. With \( p \) dependent variables, there are \( n \times p \) errors that are independent across observations but not across dependent variables. Assume

\[
\text{vec}(\varepsilon) \sim N(0, I_p \otimes \Sigma)
\]

where \( \text{vec}(\varepsilon) \) strings \( \varepsilon \) out by rows, \( \otimes \) denotes Kronecker product multiplication, and \( \Sigma \) is \( p \times p \). \( \Sigma \) can be estimated by

\[
S = [(e'e)/(n - r)] = [((Y - Xb)'(Y - Xb))/(n - r)]
\]

where \( b = (X'X)^{-1}X'Y \), \( r \) is the rank of the \( X \) matrix, and \( e \) is the matrix of residuals.

If \( S \) is scaled to unit diagonals, the values in \( S \) are called partial correlations of the \( Y \)s adjusting for the \( X \)s. This matrix can be displayed by PROC GLM if PRINTE is specified as a MANOVA option.

The multivariate general linear hypothesis is written

\[
LM = 0
\]

You can form hypotheses for linear combinations across columns, as well as across rows of \( \beta \).

The MANOVA statement of the GLM procedure tests special cases where \( L \) corresponds to Type I, Type II, Type III, or Type IV tests, and \( M \) is the \( p \times p \) identity matrix. These tests are joint tests that the given type of hypothesis holds for all dependent variables in the model, and they are often sufficient to test all hypotheses of interest.

Finally, when these special cases are not appropriate, you can specify your own \( L \) and \( M \) matrices by using the CONTRAST statement before the MANOVA
Another alternative is to use a REPEATED statement, which automatically generates a variety of \( M \) matrices useful in repeated measures analysis of variance. See the "REPEATED Statement" section and the "Repeated Measures Analysis of Variance" section for more information.

One useful way to think of a MANOVA analysis with an \( M \) matrix other than the identity is as an analysis of a set of transformed variables defined by the columns of the \( M \) matrix. You should note, however, that PROC GLM always displays the \( M \) matrix in such a way that the transformed variables are defined by the rows, not the columns, of the displayed \( M \) matrix.

All multivariate tests carried out by the GLM procedure first construct the matrices \( H \) and \( E \) corresponding to the numerator and denominator, respectively, of a univariate \( F \)-test.

\[
H = M'(Lb)'(L(X'X)−L)'−1(Lb)M \\
E = M'(Y'Y − b'(X'X)b)M
\]

The diagonal elements of \( H \) and \( E \) correspond to the hypothesis and error SS for univariate tests. When the \( M \) matrix is the identity matrix (the default), these tests are for the original dependent variables on the left-hand side of the MODEL statement. When an \( M \) matrix other than the identity is specified, the tests are for transformed variables defined by the columns of the \( M \) matrix. These tests can be studied by requesting the SUMMARY option, which produces univariate analyses for each original or transformed variable.

Four multivariate test statistics, all functions of the eigenvalues of \( E'H \) (or \( (E+H)H' \)), are constructed:

- Wilks' lambda = \( \det(E)/\det(H+E) \)
- Pillai's trace = \( \text{trace}(H(H+E)'\))
- Hotelling-Lawley trace = \( \text{trace}(E'H) \)
- Roy's maximum root = \( \lambda \), largest eigenvalue of \( E'H \)

By default, all four are reported with \( p \)-values based on \( F \) approximations, as discussed in the "Multivariate Tests" section in Chapter 2, "Introduction to Regression Procedures." Alternatively, if you specify MSTAT=EXACT on the associated MANOVA or REPEATED statement, \( p \)-values for three of the four tests are computed exactly (Wilks' Lambda, the Hotelling-Lawley Trace, and Roy's Greatest Root), and the \( p \)-values for the fourth (Pillai's trace) are based on an \( F \)-approximation that is more accurate than the default. See the "Multivariate Tests" section in Chapter 2, "Introduction to Regression Procedures," for more details on the exact calculations.
**Multivariate Tests**

Multivariate hypotheses involve several dependent variables in the form

\[ H_0 : \mathbf{L} \beta \mathbf{M} = \mathbf{d} \]

where \( \mathbf{L} \) is a linear function on the regressor side, \( \beta \) is a matrix of parameters, \( \mathbf{M} \) is a linear function on the dependent side, and \( \mathbf{d} \) is a matrix of constants. The special case (handled by PROC REG) in which the constants are the same for each dependent variable is written

\[ (\mathbf{L} \beta - \mathbf{c}) \mathbf{M} = \mathbf{0} \]

where \( \mathbf{c} \) is a column vector of constants and \( \mathbf{j} \) is a row vector of 1s. The special case in which the constants are 0 is

\[ \mathbf{L} \beta \mathbf{M} = \mathbf{0} \]

These multivariate tests are covered in detail in Morrison (1976); Timm (1975); Mardia, Kent, and Bibby (1979); Bock (1975); and other works cited in Chapter 5, "Introduction to Multivariate Procedures."

To test this hypothesis, construct two matrices, \( \mathbf{H} \) and \( \mathbf{E} \), that correspond to the numerator and denominator of a univariate \( \mathbf{F} \) test:

\[
\mathbf{H} = \mathbf{M}'(\mathbf{L} \beta - \mathbf{c} \mathbf{j})'(\mathbf{L}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{L} \beta - \mathbf{c} \mathbf{j})\mathbf{M} \\
\mathbf{E} = \mathbf{M}'(\mathbf{Y}'\mathbf{W} \mathbf{Y} - \mathbf{B}'(\mathbf{X}'\mathbf{W}\mathbf{X})\mathbf{B}) \mathbf{M} 
\]

Four test statistics, based on the eigenvalues of \( \mathbf{E}^{-1} \mathbf{H} \) or \( (\mathbf{E} + \mathbf{H})^{-1} \mathbf{H} \), are formed. Let \( \lambda_i \) be the ordered eigenvalues of \( \mathbf{E}^{-1} \mathbf{H} \) (if the inverse exists), and let \( \xi_i \) be the ordered eigenvalues of \( (\mathbf{E} + \mathbf{H})^{-1} \mathbf{H} \). It happens that \( \xi_i = \lambda_i/(1 + \lambda_i) \) and \( \lambda_i = \xi_i/(1 - \xi_i) \), and it turns out that \( \lambda_i = \sqrt{\xi_i} \) is the \( i \)th canonical correlation.

Let \( p \) be the rank of \( (\mathbf{H} + \mathbf{E}) \), which is less than or equal to the number of columns of \( \mathbf{M} \). Let \( q \) be the rank of \( \mathbf{L}(\mathbf{X}'\mathbf{W}\mathbf{X})' \mathbf{L}' \). Let \( v \) be the error degrees of freedom and \( s = \min(p,q) \). Let \( m = (|p-q|-1)/2 \), and let \( n=(v-p-1)/2 \). Then the following statistics test the multivariate hypothesis in various ways, and their p-values can be approximated by \( \mathbf{F} \) distributions. Note that in the special case that the rank of \( \mathbf{H} \) is 1, all of these \( \mathbf{F} \) statistics will be the same and the corresponding p-values will in fact be exact, since in this case the hypothesis is really univariate.

*Wilks' Lambda*

If
\[
A = \frac{\det(\mathbf{E})}{\det(\mathbf{H} + \mathbf{E})} = \prod_{i=1}^{n} \frac{1}{1 + \lambda_i} = \prod_{i=1}^{n} (1 - q_i)
\]

then

\[
F = \frac{1 - \lambda^{1/t}}{\lambda^{1/t}} \cdot \frac{rt - 2u}{pq}
\]

is approximately \(F\), where

\[
r = v - \frac{p - q + 1}{2}
\]

\[
u = \frac{pq - 2}{2}
\]

\[
t = \begin{cases} 
\sqrt{\frac{(s^2 - 4)}{s^2 + q^2 - 5}} & \text{if } p^2 + q^2 - 5 > 0 \\
1 & \text{otherwise}
\end{cases}
\]

The degrees of freedom are \(pq\) and \(rt-2u\). The distribution is exact if \(\min(p, q) \leq 2\). (Refer to Rao 1973, p. 556.)

**Pillai’s Trace**

If

\[
V = \text{trace} (\mathbf{H}(\mathbf{H}-\mathbf{E})^{-1}) = \sum_{i=1}^{n} \frac{1}{1 + \lambda_i} = \sum_{i=1}^{n} \xi_i
\]

then

\[
F = [(2n+s+1)/(2m+s+1)] \cdot [(V)/(s-V)]
\]

is approximately \(F\) with \(s(2m+s+1)\) and \(s(2n+s+1)\) degrees of freedom.

**Hotelling-Lawley Trace**

If

\[
U = \text{trace} (\mathbf{E}^{-1}\mathbf{H}) = \sum_{i=1}^{n} \frac{\lambda_i}{1 + \frac{\xi_i}{q_i}} = \sum_{i=1}^{n} \frac{\xi_i}{1 - q_i}
\]

then for \(n>0\)

\[
F = (U/c)((4 + (pq+2)/(b-1))/(pq))
\]
is approximately $F$ with $pq$ and $4 + (pq+2)/(b-1)$ degrees of freedom, where 
$b = (p+2n)(q+2n)/(2(2n+1)(n-1))$ and $c = (2+(pq+2)/(b-1))/(2n)$; while for $\pi \leq 0$

$$F = [(2(sn+1)U)/(s^2 (2m+s+1))]$$

is approximately $F$ with $s(2m+s+1)$ and $2(sn+1)$ degrees of freedom.

**Roy's Maximum Root**

If

$$\Theta = \lambda_1$$

then

$$F = \Theta r c - r + q$$

where $r = \max(p,q)$ is an upper bound on $F$ that yields a lower bound on the significance level. Degrees of freedom are $r$ for the numerator and $v-r+q$ for the denominator.

Tables of critical values for these statistics are found in Pillai (1960).

**Exact Multivariate Tests**

Beginning with release 9.0 of SAS/STAT software, if you specify the MSTAT=EXACT option on the appropriate statement, $p$-values for three of the four tests are computed exactly (Wilks' Lambda, the Hotelling-Lawley Trace, and Roy's Greatest Root), and the $p$-values for the fourth (Pillai's trace) are based on an $F$-approximation that is more accurate (but occasionally slightly more liberal) than the default. The exact $p$-values for Roy's Greatest Root give an especially dramatic improvement, since in this case the $F$-approximation only provides a lower bound for the $p$-value. If you use the $F$-based $p$-value for this test in the usual way, declaring a test significant if $p < 0.05$, then your decisions may be very liberal. For example, instead of the nominal 5% Type I error rate, such a procedure can easily have an actual Type I error rate in excess of 30%. By contrast, basing such a procedure on the exact $p$-values will result in the appropriate 5% Type I error rate, under the usual regression assumptions.

The exact $p$-values are based on the following sources:

- **Wilks' Lambda**: Lee (1972), Davis (1979)
- **Pillai's Trace**: Muller (1998)
- **Hotelling-Lawley Trace**: Davis (1970), Davis (1980)
Note that although the MSTAT=EXACT $p$-value for Pillai’s Trace is still approximate, it has "substantially greater accuracy" than the default approximation (Muller 1998).

Since most of the MSTAT=EXACT $p$-values are not based on the $F$-distribution, the columns in the multivariate tests table corresponding to this approximation---in particular, the $F$ value and the numerator and denominator degrees of freedom---are no longer displayed, and the column containing the $p$-values is labeled "P Value" instead of "Pr > F". Thus, for example, suppose you use the following PROC ANOVA code to perform a multivariate analysis of an archaeological data set: