

Multivariate Linear Models

Stanley Sawyer — Washington University

September 8, 2007 rev August 6, 2008

1. Introduction. Suppose that we have n observations, each of which has d components. For example, we may have measurements of $d = 5$ air pollutants (CO, NO, etc.) on $n = 42$ widely-separated days, d test scores for n different students, best results for d Olympic events for teams from n different countries, or d different physical measurements for n individuals (human or animal) that we are trying to classify. These observations take the form of a $n \times d$ matrix

$$Y = \begin{pmatrix} Y_{11} & Y_{12} & \dots & Y_{1d} \\ Y_{21} & Y_{22} & \dots & Y_{2d} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{nd} \end{pmatrix} \quad (1.1)$$

The i^{th} row in the display corresponds to the i^{th} observation, such as the i^{th} day, the i^{th} student, the i^{th} country, or the i^{th} individual. Each column corresponds to a particular pollutant, test, event, or measurement (height, weight, sitting height, head breadth, etc.).

As in univariate ($d = 1$) observations, we also assume that we have p covariates for each observation (day or student or country or individual). For air pollution, these might be wind strength and solar intensity ($p = 2$), age, sex, and income for students ($p = 3$), or species or country of origin for physical measurements. These are connected in the regression model

$$Y_{ij} = \mu_j + \sum_{a=1}^p X_{ia}\beta_{aj} + e_{ij} \quad (1.2)$$

for the j^{th} component of the i^{th} individual, where $1 \leq a \leq p$ refers to covariates. As in the univariate case ($d = 1$), there are $r = p + 1$ parameters for each component including the intercept terms μ_j . If we leave out the intercept terms μ_j , so that we have a *no-intercept* regression, then $r = p$.

A key assumption in (1.2) is that the measured covariate terms X_{ia} are the same for all components of the observations Y_{ij} . For example, wind strength and solar intensity have the same numerical values for all pollutants,

although the response to wind and solar intensity (measured by μ_j and β_{aj}) may be different for different pollutants. Similarly, the same student has the same values of age, sex, and income for all tests.

In the example of athletic events, this assumption means that each individual belongs to the same country and has the same physical measurements for all athletic events. (According to the news, it is not unheard of for athletes to accept dual citizenship in another country to join an Olympic team, but we assume that this does not happen here.) In contrast, the *parameters* μ_j and β_{aj} can depend on the individual components j .

The form of (1.2) means that the sum on the right-hand side of (1.2) has the form of a matrix product rather than being more complicated, which means that the resulting statistical analysis is much simpler than it would be otherwise. In particular, the fact that the X_{ia} are the same for all j also means that (1.2) has the form of d parallel univariate regressions for the d components with the same design matrix X .

The errors e_{ij} in (1.2) are assumed to be a jointly normal random vector with mean zero in R^{nd} , where $1 \leq i \leq n$ for observations and $1 \leq j \leq d$ for components. The *rows* of e_{ij} are assumed to be independent, since they correspond to different observations.

However, the *columns* of e_{ij} are allowed to be correlated. In practice, the values of Y_{ij} for a particular i are often positively correlated over j . For example, if one pollutant is high after correcting for wind and solar intensity, then the other pollutants may be high as well. If a student does well on one test after correcting for age, sex, and income, then he or she is likely to do well on the other tests as well. If one physical measurement on an individual is large after correctly for country of origin, then other physical measurements may be large as well.

In more detail, we assume that the errors e_{ij} in (1.2) are mean-zero jointly normal random variables and satisfy

$$\begin{aligned} \text{Cov}(e_{ij}, e_{kl}) &= 0, & i \neq k \\ \text{Cov}(e_{ij}, e_{i\ell}) &= \Sigma_{j\ell} \end{aligned} \tag{1.3}$$

for all i, j, k, ℓ . The assumption of the same $d \times d$ covariance matrix Σ for all i replaces the assumption of a constant variance σ^2 for a univariate regression. To keep things simple, we assume that Σ is positive definite (or invertible). An equivalent way of writing (1.3) is

$$\text{Cov}(e_{ij}, e_{kl}) = (I_n)_{ik} \Sigma_{j\ell} \tag{1.4}$$

where I_n is the $n \times n$ identity matrix.

2. The Regression Model (1.2) in Terms of Matrices: As in the univariate case, we set $r = p + 1$ and shift indices in (1.2) so that the first component $\beta_1 = \mu$ if we include intercept terms, but have $r = p$ for a no-intercept regression. In either case, we can write the regression

$$Y_{ij} = \sum_{a=1}^r X_{ia}\beta_{aj} + e_{ij}$$

in matrix notation as

$$Y = X\beta + e \tag{2.1}$$

In (2.1), Y is $n \times d$, X is $n \times r$, and

$$\beta = \begin{pmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1d} \\ \beta_{21} & \beta_{22} & \dots & \beta_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \dots & \beta_{rd} \end{pmatrix}$$

is an $r \times d$ matrix. If we allow intercepts, the first column of X is identically one and the first row of β are the intercepts μ_j . In general, the a^{th} row of β corresponds to the a^{th} covariate (or intercept). The j^{th} column of β are the regression coefficients for the j^{th} component of Y_{ij} .

For example, suppose that we measure $d = 5$ air pollutants on $n = 42$ different days. Each pollutant has $p = 2$ parameters for response to wind strength and solar intensity. Adding an intercept term means $r = p + 1 = 3$ coefficients and the parameter matrix β is 3×5 . In a particular numerical example, the estimated values of the parameters β were

$$\hat{\beta} = \begin{pmatrix} 4.718 & 4.106 & 10.115 & 8.276 & 2.358 \\ -0.138 & -0.192 & -0.211 & -0.787 & 0.071 \\ 0.012 & -0.006 & 0.021 & 0.095 & 0.003 \end{pmatrix} \tag{2.2}$$

Each column in (2.2) is the estimated parameter values β for a particular component of Y . The first row $\{\beta_{1j}\}$ contains all of the intercepts of the $d = 5$ univariate regressions on wind strength and solar intensity. The second row $\{\beta_{2j}\}$ are the coefficients for wind, which might scatter some pollutants but not others, and the third row $\{\beta_{3j}\}$ are the coefficients for solar intensity.

3. Kronecker Products of Matrices. In a univariate regression ($d = 1$), the observations Y and parameters β in $Y = X\beta + e$ are column vectors. For a multivariate regression ($d > 1$), Y is a $n \times d$ matrix and β is an $r \times d$

matrix. Sometimes it will be more convenient to treat the observations Y as an nd -dimensional vector or β as an rd -dimensional vector, where $nd = 210$ and $rd = 15$ in this case. If $d = 1$, then $\text{Cov}(Y)$ and $\text{Cov}(e)$ are $n \times n$ matrices, but if $d > 1$ they are not obviously defined as matrices, but would be 210×210 if they were defined.

We will use the subscript L when we view Y , β , and e as column vectors. Thus Y and e are $n \times d$ matrices, but Y_L and e_L will be $nd \times 1$ column vectors. Similarly, β_L will be a $rd \times 1$ column vector. To be explicit, we assume that the matrix entries are stored in the column vector by rows. This means that the I^{th} entry of the column vector Y_L , for example, is

$$(Y_L)_I = Y_{ij} \quad \text{for } I = (i - 1)d + j \tag{3.1}$$

Note that the relation $I = (i - 1)d + j$ gives a one-one correspondence between pairs (i, j) with $1 \leq j \leq d$ and $1 \leq i \leq n$ and indices I with $1 \leq I \leq nd$. (**Exercise:** Prove this.) This is called the *lexicographic* ordering of (i, j) , since it is the same as alphabetical ordering if i, j were replaced by letters. In particular, if $n = 2$ and $d = 3$, then the $N = nd = 6$ indices ij are ordered 11, 12, 13, 21, 22, 23.

If the basic regression equation $Y = X\beta + e$ in (2.1) is written in terms of vectors, it should take the form

$$Y_L = X_{(L)}\beta_L + e_L \tag{3.2}$$

where $X_{(L)}$ is an $nd \times rd$ matrix that depends on the $n \times d$ matrix X . The notions of *Kronecker product* or *tensor product* of vectors or matrices are a useful way to describe these larger matrices.

In general, if $A = \{A_{ij}\}$ is an $m_1 \times n_1$ matrix and $B = \{B_{pq}\}$ is an $m_2 \times n_2$ matrix, the *tensor product* or *Kronecker product* (matrix) of A and B is the matrix $C = A \otimes B$ with components

$$C_{ip,jq} = A_{ij}B_{pq} \quad (C = A \otimes B) \tag{3.3}$$

The matrix C is $m_1m_1 \times n_2n_2$, which can be much larger than either A or B . A subtlety here is that (3.3) does not define a matrix in the usual sense since the pairs ip and jq are not obviously linearly ordered. To complete the definition, we define

$$C_{IJ} = A_{ij}B_{pq} \quad (I = (i - 1)m_2 + p, \quad J = (j - 1)n_2 + q) \tag{3.4}$$

using the same lexicographic ordering as in (3.1). Then C_{IJ} is an $m_1m_1 \times n_2n_2$ matrix.

For example, the basic regression equation (2.1) can be written

$$\begin{aligned}
 Y_{ij} &= \sum_{a=1}^r X_{ia} \beta_{aj} + e_{ij} \\
 &= \sum_{a=1}^r \sum_{b=1}^d (X_{ia} \delta_{jb}) \beta_{ab} + e_{ij}
 \end{aligned}$$

In terms of the ordering of indices in (3.1) and (3.3), this is

$$Y_L = (X \otimes I_d) \beta_L + e_L \tag{3.5}$$

and (3.2) holds with $X_{(L)} = X \otimes I_d$. In general, the matrix relation

$$W_{ij} = \sum_{a=1}^k A_{ia} B_{aj} = \sum_{a=1}^k \sum_{b=1}^n (A_{ia} \delta_{jb}) B_{ab}$$

where W is $m \times n$, A is $n \times k$, and B is $k \times n$ implies

$$W_L = (A \otimes I_n) B_L \tag{3.6}$$

This is like the matrix equation $W = AB$, but now W_L and B_L are vectors. Similarly, the relations

$$\text{Cov}(e_{ij}, e_{kl}) = (I_n)_{ik} \Sigma_{j\ell}$$

in (1.4) are equivalent to

$$\text{Cov}(e_L) = I_n \otimes \Sigma \tag{3.7}$$

(Exercise: Why is $\text{Cov}(e_L) = I_n \otimes \Sigma$ in (3.7) and not $\Sigma \otimes I_n$ or something different? Explain clearly.)

With lexicographic ordering of the indices, the entries of

$$C_{IJ} = C_{ip,jq} = A_{ij} B_{pq}$$

in (3.4) for $1 \leq p \leq m_2$ and $1 \leq q \leq n_2$ are adjacent in C_{IJ} if i and j are fixed. This means that the matrix $C = A \otimes B$ can be written in block partitioned form as

$$C = \begin{pmatrix}
 a_{11}B & a_{12}B & \dots & a_{1n_1}B \\
 a_{21}B & a_{22}B & \dots & a_{2n_2}B \\
 \vdots & \vdots & \ddots & \vdots \\
 a_{m_1 1}B & a_{n_1 2}B & \dots & a_{m_1 n_1}B
 \end{pmatrix}$$

In particular by (3.7)

$$\text{Cov}(e_L) = \begin{pmatrix} \Sigma & 0 & \dots & 0 \\ 0 & \Sigma & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Sigma \end{pmatrix} \tag{3.8}$$

is an $nd \times nd$ block diagonal matrix with n blocks of the $d \times d$ matrix Σ down the diagonal.

4. The MLE of the $r \times d$ matrix β . We first give explicit component-wise derivations of the matrix MLE $\hat{\beta}$ and its covariance matrix $\text{Cov}(\hat{\beta})$ and follow it by shorter derivations that use Kronecker products.

In terms of components, the errors e_{ij} in (2.1) are jointly normal, are independent for different i , and have covariance matrix Σ in j for fixed i . This means that the likelihood function of the first observation Y_1 in the regression $Y = X\beta + e$ in (2.1) (or equivalently of the first row of the $n \times d$ matrix Y) is the multivariate normal density

$$L(Y_1, \beta) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp(-S_1/2) \quad \text{where} \tag{4.1}$$

$$S_1 = \sum_{a=1}^d \sum_{b=1}^d (Y_{1a} - (X\beta)_{1a}) \Sigma_{ab}^{-1} (Y_{1b} - (X\beta)_{1b})$$

Since the rows of e_{ij} are independent, the likelihood function of all n observations Y in (2.1) is the product

$$L(Y, \beta) = \frac{1}{\sqrt{(2\pi)^{nd} \det(\Sigma)^n}} \exp(-S/2) \quad \text{where} \tag{4.2}$$

$$S = \sum_{i=1}^n \sum_{a=1}^d \sum_{b=1}^d (Y_{ia} - (X\beta)_{ia}) \Sigma_{ab}^{-1} (Y_{ib} - (X\beta)_{ib})$$

Finding the matrix MLE $\hat{\beta}$ is equivalent to minimizing the triple sum S in (4.2) as a function of β . Since $(X\beta)_{ia} = \sum_{j=1}^r X_{ij}\beta_{ja}$ in (4.2), setting $(\partial/\partial\beta_{kc})S = 0$ leads to the set of equations

$$2 \sum_{i=1}^n \sum_{b=1}^d X_{ik} \Sigma_{cb}^{-1} (Y_{ib} - (X\beta)_{ib}) = 2 \sum_{b=1}^d \Sigma_{cb}^{-1} ((X'Y)_{kb} - (X'X\beta)_{kb}) = 0$$

for all k and c . This is $\Sigma^{-1}(X'X\beta - X'Y)' = 0$ in matrix form. Premultiplying by Σ leads to the matrix “normal equations”

$$X'X\beta = X'Y \quad \text{or} \quad (X'X \otimes I_d)\beta_L = (X' \otimes I_d)Y_L$$

by applying (3.6). If the $r \times r$ design matrix $X'X$ is invertible, then the matrix-valued MLE of β is

$$\hat{\beta} = (X'X)^{-1}X'Y \quad \text{or} \quad \hat{\beta}_L = ((X'X)^{-1}X' \otimes I_d)Y_L \quad (4.3)$$

The first formula in (4.3) is exactly the same formula as in the univariate case ($d = 1$), except that now $\hat{\beta}$ is $r \times d$. The columns of $\hat{\beta}$ for individual components of Y_{ij} are formed by applying the same $r \times n$ matrix $(X'X)^{-1}X'$ to each of the columns of Y .

In terms of components, (4.3) and $Y = X\beta + e$ imply

$$\hat{\beta}_{aj} = \beta_{aj} + \sum_{\ell=1}^n Q_{a\ell}e_{\ell j}, \quad Q = (X'X)^{-1}X'$$

Then by (1.4)

$$\begin{aligned} \text{Cov}(\hat{\beta}_{aj}, \hat{\beta}_{bk}) &= \sum_{\ell=1}^n \sum_{m=1}^n Q_{a\ell}Q_{bm} \text{Cov}(e_{\ell j}, e_{mk}) \\ &= \sum_{\ell=1}^n Q_{a\ell}Q_{b\ell} \Sigma_{jk} = (QQ')_{ab} \Sigma_{jk} \\ &= ((X'X)^{-1})_{ab} \Sigma_{jk} \end{aligned} \quad (4.4)$$

since $QQ' = (X'X)^{-1}X'X(X'X)^{-1} = (X'X)^{-1}$. Thus

$$\text{Cov}(\hat{\beta}) = (X'X)^{-1} \otimes \Sigma \quad (4.5)$$

We can also derive the relations (4.3) and (4.4) using tensor products. First, we need

Lemma 4.1. Suppose that $C = A \otimes B$ and $F = D \otimes E$ and assume that the matrices AD and BE can be defined. Then

(i)
$$CF = (A \otimes B)(D \otimes E) = AD \otimes BE \quad (4.6)$$

(ii) $I_m \otimes I_n = I_{mn}$ for all integers $m, n \geq 1$

(iii) The transpose $C' = A' \otimes B'$

(iv) If A and B are invertible, then so is C and

$$C^{-1} = A^{-1} \otimes B^{-1} \tag{4.7}$$

Proof. (i) Write $C_{ia,jb} = A_{ij}B_{ab}$ and $F_{jb,kc} = D_{jk}E_{bc}$. Then

$$\begin{aligned} (CF)_{ia,kc} &= \sum_{jb} C_{ia,jb} F_{jb,kc} \\ &= \sum_j \sum_b A_{ij} B_{ab} D_{jk} E_{bc} = (AD)_{ik} (BE)_{ac} \end{aligned}$$

which implies $CF = AD \otimes BE$. Parts (ii) and (iii) are similar. Part (iv) follows from parts (i) and (ii).

To begin a second proof of (4.3) and (4.4), recall that the MLE $\hat{\beta}$ of the univariate regression equation with possible correlated errors

$$Y = X\beta + e, \quad e \approx N(0, V)$$

where V is a positive definite $n \times n$ matrix is given by

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}Y \tag{4.8}$$

The fastest way to verify (4.8) is to note that

$$Y_1 = V^{-1/2}Y = (V^{-1/2}X)\beta + V^{-1/2}e = X_1\beta + e_1$$

Since $\text{Cov}(e_1) = V^{-1/2} \text{Cov}(e)V^{-1/2} = V^{-1/2}VV^{-1/2} = I_n$, the usual formula for $\text{Cov}(e) = \sigma^2 I_n$ implies $\hat{\beta} = (X_1'X_1)^{-1}X_1'Y_1$. Substituting $X_1 = V^{-1/2}X$ and $Y_1 = V^{-1/2}Y$ then implies (4.8).

Write the multivariate regression $Y = X\beta + e$ in vector form as

$$Y_L = (X \otimes I_d)\beta_L + e_L, \quad \text{Cov}(e_L) = V = I_n \otimes \Sigma \tag{4.9}$$

where V is $nd \times nd$. Then by (4.8)

$$\begin{aligned} \hat{\beta}_L &= ((X \otimes I_d)'V^{-1}(X \otimes I_d))^{-1}(X \otimes I_d)'V^{-1}Y_L \\ &= ((X \otimes I_d)'(I_n \otimes \Sigma)^{-1}(X \otimes I_d))^{-1}(X \otimes I_d)'(I_n \otimes \Sigma)^{-1}Y_L \\ &= ((X'X)^{-1} \otimes \Sigma)(X' \otimes \Sigma^{-1})Y_L \\ &= ((X'X)^{-1}X' \otimes I_d)Y_L \end{aligned}$$

by multiple applications of Lemma 4.1. This is (4.3). Since $\text{Cov}(AX) = A \text{Cov}(X)A'$ for any random variable X and matrix A and $\text{Cov}(Y_L) = \text{Cov}(e_L)$,

$$\begin{aligned} \text{Cov}(\widehat{\beta}) &= ((X'X)^{-1}X' \otimes I_d) \text{Cov}(Y_L)((X'X)^{-1}X' \otimes I_d)' \\ &= ((X'X)^{-1}X' \otimes I_d)(I_n \otimes \Sigma)(X(X'X)^{-1} \otimes I_d) \\ &= (X'X)^{-1} \otimes \Sigma \end{aligned} \tag{4.10}$$

since $(X'X)^{-1}X'X(X'X)^{-1} = (X'X)^{-1}$. This completes the second proof of (4.3) and (4.4).

5. The MLE of the $d \times d$ matrix Σ . In principle, this should be easier than finding $\widehat{\beta}$ since no Kronecker products should be involved. The next result shows that the maximum likelihood estimator of the matrix Σ is a natural generalization of the corresponding one-dimensional result.

Theorem 5.1. The maximum likelihood estimators for (β, Σ) for the likelihood (4.2) is given by $\widehat{\beta} = (X'X)^{-1}X'Y$ (by (4.3)) and

$$\widehat{\Sigma}_{ab} = \frac{1}{n} \sum_{i=1}^n (Y_{ia} - (X\widehat{\beta})_{ia})(Y_{ib} - (X\widehat{\beta})_{ib}) \tag{5.1}$$

That is, the maximum likelihood estimator $\widehat{\Sigma}$ is the sample covariance matrix of the residuals of the multivariate regression $Y = X\beta + e$ in Section 1 with $n - 1$ replaced by n .

Proof. The likelihood (4.2) is

$$L(Y, \beta, \Sigma) = \frac{1}{\sqrt{(2\pi)^{nd} \det(\Sigma)^n}} \exp(-S_\Sigma/2) \quad \text{where} \tag{5.2}$$

$$S_\Sigma = \sum_{i=1}^n \sum_{a=1}^d \sum_{b=1}^d (Y_{ia} - (X\beta)_{ia}) \Sigma_{ab}^{-1} (Y_{ib} - (X\beta)_{ib})$$

Note

$$S_\Sigma = \sum_{a=1}^d \sum_{b=1}^d Q_{ab} \Sigma_{ab}^{-1} = \text{tr}(Q\Sigma^{-1})$$

where

$$Q_{ab} = \sum_{i=1}^n (Y_{ia} - (X\beta)_{ia})(Y_{ib} - (X\beta)_{ib}) \tag{5.3}$$

Thus the likelihood in (5.2) can also be written

$$L(Y, \beta, \Sigma) = \frac{1}{\sqrt{(2\pi)^{nd} \det(\Sigma)^n}} \exp\left(-\frac{1}{2} \text{tr}(Q\Sigma^{-1})\right) \tag{5.4}$$

Since the maximum likelihood estimators are defined by the maximum over both β and Σ , and since the maximum over β in (5.4) does not depend on Σ (see (4.3)), Theorem 5.1 follows from

Lemma 5.1. Let Q be a $d \times d$ positive definite matrix and $n > 0$ an arbitrary number. Then the maximum of

$$\frac{1}{\sqrt{\det(\Sigma)^n}} \exp\left(-\frac{1}{2} \text{tr}(Q\Sigma^{-1})\right) \tag{5.5}$$

for $d \times d$ positive definite matrices Σ is attained at $\Sigma = (1/n)Q$.

Proof of Lemma 5.1. By the spectral theorem for symmetric matrices, we can write $Q = E^2$ where E is positive definite. By (5.5), it is sufficient to maximize

$$\phi(\Sigma) = -n \log \det(\Sigma) - \text{tr}(Q\Sigma^{-1}) \tag{5.6}$$

Set $A = E\Sigma^{-1}E$. Then $A^{-1} = E^{-1}\Sigma E^{-1}$ and $EA^{-1}E = \Sigma$. Then

$$\begin{aligned} \phi(\Sigma) &= \phi(EA^{-1}E) \\ &= -n \log \det(EA^{-1}E) - \text{tr}(QE^{-1}AE^{-1}) \\ &= -n \log \det(E)^2 + n \log \det(A) - \text{tr}(E^{-1}E^2E^{-1}A) \\ &= -n \log \det(Q) + n \log \det(A) - \text{tr}(A) \end{aligned}$$

where Q is fixed. By the spectral theorem again, $A = U'DU$ where D is diagonal and U is orthogonal, and $\det(A) = \det(U'DU) = \det(D)$ and $\text{tr}(A) = \text{tr}(U'DU) = \text{tr}(D)$ by properties of the determinant and trace of matrices. Thus if $D = \text{diag}(v_1, v_2, \dots, v_d)$

$$\phi(\Sigma) = -n \log \det(Q) + \sum_{i=1}^d (n \log(v_i) - v_i)$$

The expression on the right above is maximized when $v_i = n$ for all i . This implies $D = nI_d$, hence $A = U'DU = U'(nI_d)U = nI_d$, and hence

$$\Sigma = EA^{-1}E = (1/n)E^2 = (1/n)Q$$

This completes the proof of Lemma 5.1 and hence of Theorem 5.1.

6. Hypothesis Testing: A natural generalization of univariate tests for whether or not coefficients in the regression $Y = X\beta + e$ in (2.1) are nonzero is

$$H_0(a) : \beta_{aj} = 0, \quad 1 \leq j \leq d \tag{6.1}$$

This hypothesis says that the a^{th} row of the $r \times d$ matrix β is identically zero, which is equivalent to saying that the a^{th} covariate column in X_{ia} does not affect any of the components of $Y = X\beta + e$. That is, the data vectors $\{Y_i \in R^d : 1 \leq i \leq n\}$ do not depend on the a^{th} covariate.

A natural generalization of (6.1) is

$$H_0 : h'\beta = 0, \quad h \text{ is } r \times 1 \tag{6.2}$$

where h is a $r \times 1$ column vector. This is equivalent to

$$(h'\beta)_j = \sum_{b=1}^r h_b \beta_{bj} = 0, \quad 1 \leq j \leq d \tag{6.3}$$

This says that the same linear relationship (6.3) holds for the coefficients β_{aj} in the d componentwise univariate regressions ($1 \leq j \leq d$) that are implicit in the multivariate regression $Y = X\beta + e$.

If $d = 1$, the usual way to test $h'\beta = 0$ (or $\beta_a = 0$) is to use the identity

$$\text{Var}(h'\hat{\beta}) = h' \text{Cov}(\hat{\beta})h = \sigma^2 h'(X'X)^{-1}h \quad (d = 1)$$

If $H_0 : h'\beta = 0$ is correct, then the test statistic

$$T = \frac{h'\hat{\beta}}{\sqrt{(\text{MSE}) h'(X'X)^{-1}h}} \quad \text{where} \tag{6.4}$$

$$\text{MSE} = \frac{1}{n-r} \sum_{i=1}^n (Y_i - (X\hat{\beta})_i)^2 \tag{6.5}$$

has a Student's t distribution with $n - r$ degrees of freedom.

If $d > 1$, then $h'\beta$ is a $1 \times d$ row vector, and a plausible generalization is to compare the $d \times d$ matrix

$$\begin{aligned} H_h &= (h'\hat{\beta})'(h'\hat{\beta}) / (h'(X'X)^{-1}h) \\ &= (\hat{\beta}'h)(\hat{\beta}'h) / (h'(X'X)^{-1}h) \end{aligned} \tag{6.6}$$

with the $d \times d$ residual error matrix

$$E = (Y - X\hat{\beta})'(Y - X\hat{\beta}) \tag{6.7}$$

with entries

$$E_{ab} = \sum_{i=1}^n (Y_{ia} - (X\hat{\beta})_{ia})(Y_{ib} - (X\hat{\beta})_{ib}) \tag{6.8}$$

If $d = 1$, H_h/E is the same as $t^2/(n - r)$ for t in (6.4), and has distribution $F_{1,n-r}/(n - r)$ if $h'\beta = 0$.

If $d > 1$, the fact that H_h/E is a ratio of matrices for H_h and E in (6.6) and (6.7) is made even more awkward by the fact that the three matrices

$$E^{-1}H_h \quad H_hE^{-1} \quad E^{-1/2}H_hE^{-1/2} \tag{6.9}$$

are in general different. However, the *eigenvalues* of the three matrices (6.9) are exactly the same. This follows because all three matrices have the same characteristic polynomial (for example for $E^{-1}H_h$)

$$\begin{aligned} f(\lambda) &= \det((E^{-1}H_h) - \lambda I) = \det(E^{-1}(H_h - \lambda E)) \\ &= \det(H_h - \lambda E) / \det(E) \end{aligned} \tag{6.10}$$

For the “rank one” case (6.2), it turns out that (i) the three matrices in (6.9) have exactly one nonzero eigenvalue λ_1 for H_h in (6.6), (ii) the eigenvalue λ_1 can be easily found, (iii) the eigenvalue λ_1 has an F distribution given $H_0 : h'\beta = 0$, and hence (iv) a practical test of $H_0 : h'\beta = 0$ can be based on λ_1 .

To show (i) and (ii), first note that $\beta'h$ and $\hat{\beta}'h$ in (6.6) are $d \times 1$ column vectors and that (6.6) can be written

$$H_h = vv'/C, \quad v = \hat{\beta}'h, \quad C = h'(X'X)^{-1}h \tag{6.11}$$

which is a $d \times d$ matrix of rank one. Similarly

$$E^{-1/2}H_hE^{-1/2} = (E^{-1/2}v)(E^{-1/2}v)'/C \tag{6.12}$$

is also a $d \times d$ rank-one matrix.

In general, if $A = cxx'$ is a symmetric $d \times d$ matrix for a $d \times 1$ column vector x and a constant c , then A has at most one nonzero eigenvalue λ_1 , which is given by $\lambda_1 = cx'x$. To show this, first assume $Ay = \lambda y$ for some eigenvalue $\lambda \neq 0$ and eigenvector $y \neq 0$. Then $Ay = cx(x'y) = \lambda y$. Since $\lambda \neq 0$, $y = cx(x'y)/\lambda = ax$ for $a = c(x'y)/\lambda$. Thus $Ay = \lambda y = A(ax) = cax(x'x) = \lambda ax$, which implies $\lambda = cx'x$ since $y = ax \neq 0$. Conversely, if $x'y = 0$, then $Ay = cx(x'y) = 0$. It follows that the $d \times d$ matrix A has one eigenvalue $\lambda_1 = cx'x$ and an additional $d - 1$ zero eigenvalues.

This implies that the matrices in (6.9) and (6.12) have the sole nonzero eigenvalue

$$\begin{aligned} \lambda_1 &= (E^{-1/2}v)'(E^{-1/2}v)/C = v'E^{-1}v/C \\ &= (\hat{\beta}'h)'E^{-1}(\hat{\beta}'h)/(h'(X'X)^{-1}h) \end{aligned} \tag{6.13}$$

For the special case of $H_0(a) : \beta_{aj} = 0$ in (6.1), this is

$$\lambda_1(a) = \widehat{\beta}_a E^{-1} \widehat{\beta}'_a / ((X'X)^{-1})_{aa} \tag{6.14}$$

where $\widehat{\beta}_a$ is the a^{th} row of $\widehat{\beta}$ and E is the $d \times d$ covariance sum matrix in (6.8). Note that the matrix X appears directly in the statistic λ_1 in (6.13) only as the scalar constant $h'(X'X)^{-1}h$, exactly as in the univariate case.

We will derive the exact distribution of λ_1 given $H_0 : h'\beta = 0$ in Sections 8 and 10 below. This will allow us to find P-values for “rank one” multivariate hypotheses using standard probability distributions. Before proceeding, let’s show how a simple multivariate two-sample problem also leads to a statistic that is very similar to (6.14).

7. A Multivariate Two-Sample t -Test: Suppose that we have two independent d -dimensional vector-valued samples

$$\begin{aligned} (Z_1)_1, (Z_1)_2, \dots, (Z_1)_{n_1} & \quad \text{where} \quad (Z_1)_i \approx N(\mu_1, \Sigma) & (7.1) \\ (Z_2)_1, (Z_2)_2, \dots, (Z_2)_{n_2} & \quad \text{where} \quad (Z_2)_j \approx N(\mu_2, \Sigma) \end{aligned}$$

with the same covariance matrix Σ and that we want to test $H_0 : \mu_1 = \mu_2$.

Examples of (7.1) would be two sets of d -dimensional pollution profiles for two different cities, d tests for two sets of students, Olympic results for two sets of athletes from two different countries, or d physical measurements on two sets of human skulls.

Note that this is exactly the same setup as in the classical two-sample t -test. The only difference is that the observations Z_{ij} in (7.1) are vector-valued with the same unknown $d \times d$ covariance matrix Σ , as opposed to being univariate normal with the same unknown variance σ^2 .

We could analyze the data in (7.1) by carrying out d different two-sample t -tests on the d components of Z_{ij} . However, this can definitely lead to misleading results if the random vectors Z_{ij} have a significant vector difference that is not aligned with one of the coordinates axes. An appropriate test of (7.1) would take this possibility into account.

If $d = 1$, the standard classical test of $H_0 : \mu_1 = \mu_2$ is based on the statistic

$$\begin{aligned} T &= \sqrt{\frac{n_1 n_2}{n_1 + n_2}} (\bar{Z}_1 - \bar{Z}_2) / \sqrt{s^2} \quad \text{where} & (7.2) \\ s^2 &= \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)^2 \end{aligned}$$

Here s^2 is the *pooled variance* estimator of σ^2 . If $\mu_1 = \mu_2$, then T has a Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom. If $\mu_1 \neq \mu_2$, then T has a *noncentral* Student's t distribution with $n_1 + n_2 - 2$ degrees of freedom.

A generalization of T for $d > 1$ due to Hotelling (1931) is

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\bar{Z}_1 - \bar{Z}_2)' S^{-1} (\bar{Z}_1 - \bar{Z}_2) \quad \text{where} \quad (7.3)$$

$$S = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^2 \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_i)(Z_{ij} - \bar{Z}_i)'$$

Here S is called the *pooled sample covariance estimator* of the matrix Σ . The statistic T^2 in (7.3) is called the *Hotelling T^2 -statistic* for the two-sample multivariate problem (7.1).

The data in (7.1) can be put in the form of a multivariate regression $Y = X\beta + e$ by viewing the data $(Z_1)_i, (Z_2)_j$ in (7.1) as the row vectors of a $n \times d$ matrix Y with entries

$$\begin{aligned} Y_{ij} &= (Z_1)_{ij}, & 1 \leq i \leq n_1, & \quad 1 \leq j \leq d \\ Y_{ij} &= (Z_2)_{i-n_1,j}, & n_1 + 1 \leq i \leq n, & \quad 1 \leq j \leq d \end{aligned}$$

where $n = n_1 + n_2$. Then the model (7.1) is equivalent to

$$\begin{aligned} Y_{ij} &= (\mu_1)_j + e_{ij}, & 1 \leq i \leq n_1, & \quad 1 \leq j \leq d \\ Y_{ij} &= (\mu_2)_j + e_{ij}, & n_1 + 1 \leq i \leq n, & \quad 1 \leq j \leq d \end{aligned}$$

where the rows e_i of the $n \times d$ matrix e are independent random normal vectors with distribution $N(0, \Sigma)$. This can be written in matrix form as

$$Y = X\beta + e \quad \text{for} \quad X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \dots & \dots \\ 0 & 1 \\ 0 & 1 \\ \dots & \dots \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad (7.4)$$

where μ_1 and μ_2 are now viewed as row vectors. Here X is an $n \times 2$ matrix with n_1 rows equal to $(1 \ 0)$ followed by n_2 rows equal to $(0 \ 1)$. Notice that this is a no-intercept regression. With only slightly more effort, we could also have transformed the problem into a regression in which the first column corresponds to an intercept.

If $h = (1 \ -1)'$, then $h'\beta = \mu_1 - \mu_2$ in (7.4) and $H_0 : \mu_1 = \mu_2$ is equivalent to $H_0 : h'\beta = 0$. We now apply (6.2) through (6.13) in Section 6. For X and β in (7.4),

$$X'X = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}, \quad \hat{\beta} = (X'X)^{-1}X'Y = \begin{pmatrix} \bar{Z}_1 \\ \bar{Z}_2 \end{pmatrix} \quad (7.5)$$

where $\bar{Z}_a = (1/n_a) \sum_{i=1}^{n_a} Z_{ai}$ are the two sample means in (7.1) viewed as row vectors. In particular, $\hat{\beta}_a = \bar{Z}_a$ for $a = 1, 2$ for the two rows of the $2 \times d$ matrix $\hat{\beta}$. Similarly

$$\begin{aligned} \hat{\beta}'h &= \begin{pmatrix} \bar{Z}_1 \\ \bar{Z}_2 \end{pmatrix}' \begin{pmatrix} 1 \\ -1 \end{pmatrix} = (\bar{Z}_1 - \bar{Z}_2)' \quad \text{and} \\ h'(X'X)^{-1}h &= (1 \ -1) \begin{pmatrix} 1/n_1 & 0 \\ 0 & 1/n_2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{1}{n_1} + \frac{1}{n_2} \end{aligned}$$

The eigenvalue λ_1 in (6.13) can now be written

$$\begin{aligned} \lambda_1 &= (\hat{\beta}'h)'E^{-1}(\hat{\beta}'h)/(h'(X'X)^{-1}h) \\ &= \frac{n_1n_2}{n_1 + n_2}(\bar{Z}_1 - \bar{Z}_2)'E^{-1}(\bar{Z}_1 - \bar{Z}_2) \end{aligned} \quad (7.6)$$

where $E = (Y - X\hat{\beta})'(Y - X\hat{\beta})$ is the residual error matrix in (6.7)–(6.8) and \bar{Z}_a are now viewed as column vectors. Since the matrix X in (7.4) is $(n_1 + n_2) \times 2$, λ_1 in (7.6) corresponds to $n = n_1 + n_2$ and $r = 2$ in Section 6. The residual error matrix E in (6.7) depends on the matrix of fitted values

$$(X\hat{\beta})_{ij} = \begin{cases} (\bar{Z}_1)_j & 1 \leq i \leq n_1, \quad 1 \leq j \leq d \\ (\bar{Z}_2)_j & n_1 + 1 \leq i \leq n, \quad 1 \leq j \leq d \end{cases}$$

so that

$$E = \sum_{i=1}^{n_1} (Z_{1i} - \bar{Z}_1)(Z_{1i} - \bar{Z}_1)' + \sum_{i=1}^{n_2} (Z_{2i} - \bar{Z}_2)(Z_{2i} - \bar{Z}_2)' \quad (7.7)$$

Thus the pooled covariance matrix S in the two-sample Hotelling T^2 statistic in (7.3) is $S = E/(n_1 + n_2 - 2)$ for E in (7.6), and the eigenvalue λ_1 in (7.6) can be written

$$\begin{aligned} \lambda_1 &= \frac{n_1n_2}{n_1 + n_2}(\bar{Z}_1 - \bar{Z}_2)'E^{-1}(\bar{Z}_1 - \bar{Z}_2) \\ &= \frac{1}{n_1 + n_2 - 2}T^2 \end{aligned} \quad (7.8)$$

for the two-sample Hotelling T^2 statistic in (7.3).

8. The Distribution of λ_1 for “rank one” tests $H_0 : h'\beta = 0$:

The test procedure of Section 6 compares the $d \times d$ rank-one matrix

$$H_h = (\widehat{\beta}'h)(\widehat{\beta}'h)' / (h'(X'X)^{-1}h) \tag{8.1}$$

with the $d \times d$ residual error matrix

$$E = (Y - X\widehat{\beta})'(Y - X\widehat{\beta})$$

We showed in Section 6 that the matrix H_h in (8.1) is rank one, and that the three matrices $E^{-1}H_h$, H_hE^{-1} , and $E^{-1/2}H_hE^{-1/2}$ in (6.9) have the single nonzero eigenvalue

$$\lambda_1 = (\widehat{\beta}'h)'E^{-1}\widehat{\beta}'h / (h'(X'X)^{-1}h) \tag{8.2}$$

Since $\widehat{\beta}'h$ is a $d \times 1$ column vector and E and $X'X$ are positive semidefinite matrices that are generally also positive definite, λ_1 in (8.2) is a nonnegative number that is generally positive.

We next derive a representation of the distribution of the test statistic λ_1 in (8.2) given $H_0 : h'\beta = 0$. By (7.6), this will also give us the distribution of the two-sample Hotelling T^2 statistic (7.3).

First, if L is any $q \times r$ matrix, then $L\widehat{\beta}$ is $q \times d$ and

$$(L\widehat{\beta})_L = (L \otimes I_d)\widehat{\beta}_L$$

by (3.6). Hence

$$\begin{aligned} \text{Cov}((L\widehat{\beta})_L) &= (L \otimes I_d) \text{Cov}(\widehat{\beta}_L)(L \otimes I_d)' \\ &= (L \otimes I_d)((X'X)^{-1} \otimes \Sigma)(L' \otimes I_d) \\ &= (L(X'X)^{-1}L') \otimes \Sigma \end{aligned} \tag{8.3}$$

by Lemma 4.1. If $L = h'$ is $1 \times r$, then $L(X'X)^{-1}L' = h'(X'X)^{-1}h$ is a number and

$$\text{Cov}((h'\widehat{\beta})_L) = \text{Cov}(\widehat{\beta}'h) = (h'(X'X)^{-1}h)\Sigma$$

Since $\widehat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'e$, $\widehat{\beta}$ has a joint normal distribution. Hence the random vector $\widehat{\beta}'h$ has the distribution

$$\widehat{\beta}'h \approx N(\beta'h, (h'(X'X)^{-1}h)\Sigma) \tag{8.4}$$

(Exercise: Derive (8.4) by writing $h'\widehat{\beta}$ in terms of its components and using (1.4) or (4.4) directly.)

The relation (8.4) implies that the random vector

$$Z_0 = (\widehat{\beta}'h - \beta'h) / \sqrt{h'(X'X)^{-1}h} \approx N(0, \Sigma) \tag{8.5}$$

If $h'\beta = 0$, it follows that the principal eigenvalue λ_1 in (8.2) can be written

$$\lambda_1 = Z_0'E^{-1}Z_0 \tag{8.6}$$

where $Z_0 \approx N(0, \Sigma)$ is given by (8.5) and $E = (Y - X\widehat{\beta})'(Y - X\widehat{\beta})$ is the residual error matrix. I now claim that if $h'\beta = 0$, then λ_1 can be written

$$\lambda_1 = Z_0' \left(\sum_{i=1}^{n-r} Z_i Z_i' \right)^{-1} Z_0 \quad \text{where} \tag{8.7}$$

Z_0, Z_1, \dots, Z_{n-r} are independent $N(0, \Sigma)$

It will also turn out that the distribution (8.7) does not depend on Σ .

The distribution in (8.7) is called the *Hotelling T^2* distribution (abbreviated $T^2(d, n - r)$) in honor of Hotelling (1931), and is the d -dimensional generalization of a Student's t distribution.

To prove (8.7), first note that the fitted value matrix

$$X\widehat{\beta} = X((X'X)^{-1}X'Y) = X(X'X)^{-1}X'(X\beta + e) = X\beta + Ke \tag{8.8}$$

where $K = X(X'X)^{-1}X'$ is an $n \times n$ orthogonal projection matrix. (That is, $K = K^2 = K'$.) The residual value matrix is

$$Y - X\widehat{\beta} = (X\beta + e) - (X\beta + Ke) = (I_n - K)e$$

Since $K = K' = K^2$, it follows from the spectral theorem that

$$K = U'DU, \quad D = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} \tag{8.9}$$

where U is an $n \times n$ orthogonal matrix and k is the number of nonzero eigenvalues of K . Note that k is the same as the dimension of the range space of K . Since $\text{tr}(K) = \text{tr}(U'DU) = \text{tr}(DUU') = \text{tr}(D) = k$ and $\text{tr}(K) = \text{tr}(X(X'X)^{-1}X') = \text{tr}((X'X)^{-1}X'X) = \text{tr}(I_r) = r$, it follows that $k = r$.

Define $Z = Ue$ for the $n \times n$ matrix U in (8.9). Thus

$$Z_{ij} = \sum_{a=1}^n U_{ia}e_{aj}, \quad 1 \leq i \leq n, \quad 1 \leq j \leq d$$

This means that the same $n \times n$ rotation matrix U is applied to each column of e . Viewing the $n \times d$ matrix $Z = Ue$ as an $nd \times 1$ column vector as in Section 3, $Z_L = (U \otimes I_d)e_L$ by (3.6) and

$$\begin{aligned} \text{Cov}(Z_L) &= (U \otimes I_d) \text{Cov}(e_L)(U \otimes I_d)' \\ &= (U \otimes I_d) (I_n \otimes \Sigma) (U' \otimes I_d) = (UU') \otimes \Sigma = I_n \otimes \Sigma \end{aligned}$$

by (3.7) and (4.5). This means that Z_L has the same distribution as e_L . In particular, the n rows of Z are independent random vectors with distribution $Z_i \approx N(0, \Sigma)$.

By (8.8) and (8.9), the fitted values

$$X\hat{\beta} = X\beta + Ke = X\beta + U'DUe = X\beta + (U'D)Z \quad (8.10)$$

Similarly, the residual matrix $Y - X\hat{\beta}$ is

$$Y - X\hat{\beta} = (I_n - K)e = U'(I_n - D)Ue = U'(I_n - D)Z$$

In particular

$$\begin{aligned} E &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= Z'(I_n - D)'UU'(I_n - D)Z = Z'(I_n - D)Z \end{aligned}$$

We also have

$$\begin{aligned} \hat{\beta} &= (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'e \\ &= \beta + (X'X)^{-1}X'Ke = \beta + ((X'X)^{-1}X'U')DZ \end{aligned}$$

since $X'K = X'X(X'X)^{-1}X' = X'$ and $K = U'DU$ by (8.9). Thus if we write E and $\hat{\beta}$ in terms of their components

$$\begin{aligned} E_{ab} &= \sum_{i=1}^n \sum_{j=1}^n Z_{ia}(I_n - D)_{ij}Z_{jb} = \sum_{i=r+1}^n Z_{ia}Z_{ib} \quad (8.11) \\ \hat{\beta}_{aj} &= \beta_{aj} + \sum_{i=1}^n \sum_{k=1}^n L_{ai}D_{ik}Z_{kj} = \beta_{aj} + \sum_{i=1}^r A_{ai}Z_{ij} \end{aligned}$$

where $A = (X'X)^{-1}X'U'$. This means that $\hat{\beta}$ and $\hat{\beta}'h$ depend on only the first r rows of Z , while E depends only on the last $n - r$ rows of Z . In particular, $\hat{\beta}$ and E are independent. Finally by (8.11)

$$E = \sum_{i=r+1}^n Z_i Z_i' \quad (8.12)$$

where Z_i is the i^{th} row of Z viewed as a column vector. Since Z_0 in (8.5) is a linear function of $\hat{\beta}$, it follows from (8.11) that Z_0, Z_{r+1}, \dots, Z_n are independent random vectors. The relations (8.5), (8.6), and (8.12) complete the proof of (8.7).

To put the multivariate distribution (8.7) in more perspective, we need some more definitions.

9. Wishart and Hotelling T^2 Distributions: A $d \times d$ random matrix W is said to have a *Wishart distribution* with parameters Σ, d , and m (abbreviated $W \approx W(\Sigma, d, m)$) if W has the same distribution as the random $d \times d$ matrix

$$\sum_{i=1}^m Z_i Z_i' \quad \text{where } Z_1, \dots, Z_m \text{ are independent } N(0, \Sigma) \quad (9.1)$$

In particular, the Wishart distribution is a distribution of random nonnegative-definite $d \times d$ matrices, rather than of a single univariate random variable. The random matrix (9.1) can be shown to be positive definite and invertible (with probability one) if and only if $m \geq d$.

We can sum up the first principal result of Section 8 in a theorem:

Theorem 9.1. Consider the multivariate regression

$$Y = X\beta + e, \quad e_L \approx N(0, I_n \otimes \Sigma) \quad (9.2)$$

where Y is $n \times d$, X is an $n \times r$ matrix of rank r , and β is $r \times d$, and A_L for a matrix A means the column vector of the matrix entries of A written in lexicographic order. Let $\hat{\beta} = (X'X)^{-1}X'Y$ be the MLE of β (Section 4). Then

- (i) $\hat{\beta}_L \approx N(\beta_L, (X'X)^{-1} \otimes \Sigma)$
- (ii) $E = (Y - X\hat{\beta})'(Y - X\hat{\beta}) \approx W(\Sigma, d, n - r + 1)$
- (iii) $\hat{\beta}$ and E are independent.

Proof. Part (i): See (4.5) or (4.10). Parts (ii,iii): See Section 8.

It follows from (7.4) that the residual error matrix in the multivariate two-sample problem (7.1) also satisfies $E \approx W(\Sigma, d, n_1 + n_2 - 2)$.

The Wishart distribution is a multivariate generalization of the chi-square distribution, but also depends on the matrix Σ . For simplicity, let $W(d, m) = W(I_d, d, m)$ denote the Wishart distribution with $\Sigma = I_d$. Then

Lemma 9.1. In terms of distributions, for any $r \times d$ matrix A ,

- (i) $W(\Sigma, d, m) \approx \Sigma^{1/2}W(d, m)\Sigma^{1/2}$
- (ii) $AW(\Sigma, d, m)A' \approx W(A\Sigma A', r, m)$

Proof. If $W = \sum_{i=1}^m Z_i Z_i'$ where Z_i are independent $N(0, \Sigma)$, then

$$AWA' = A \sum_{i=1}^m Z_i Z_i' A' = \sum_{i=1}^m (AZ_i)(AZ_i)'$$

Since $\text{Cov}(AZ_i) = A \text{Cov}(Z_i) A' = A \Sigma A'$ and A is $r \times d$, it follows that AWA' is Wishart $W(A \Sigma A', r, m)$. It follows from the same argument that if $W = \sum_{i=1}^m N_i N_i'$ for independent $N_i \approx N(0, I_d)$ and $A = \Sigma^{1/2}$, then $AWA \approx W(\Sigma, d, m)$.

A random variable T is said to have a *Hotelling's T^2 distribution* with parameters (d, m) (abbreviated $T \approx T^2(d, m)$) if T has the distribution

$$T \approx Y' S^{-1} Y, \quad S = \frac{1}{m} \sum_{i=1}^m Z_i Z_i' \tag{9.3}$$

where Y, Z_1, \dots, Z_m are $m+1$ independent $N(0, \Sigma)$ for some positive definite matrix Σ , or equivalently if

$$T \approx Y' \left(\frac{1}{m} W(\Sigma, d, m) \right)^{-1} Y, \quad Y \approx N(0, \Sigma) \tag{9.4}$$

where Y is independent of $W(\Sigma, d, m)$.

The distribution $T \approx T^2(d, m)$ does not depend on Σ . For, by (9.4) and Lemma 9.1,

$$S \approx \frac{1}{m} W(\Sigma, d, m) \approx \Sigma^{1/2} \left(\frac{1}{m} W(d, m) \right) \Sigma^{1/2}$$

and hence

$$\begin{aligned} T &\approx Y' S^{-1} Y \approx (\Sigma^{1/2} N_0)' (\Sigma^{-1/2} S_N^{-1} \Sigma^{-1/2}) \Sigma^{1/2} N_0 \\ &\approx N_0' S_N^{-1} N_0, \quad S_N = \frac{1}{m} \sum_{i=1}^m N_i N_i' \end{aligned}$$

where N_0, N_1, \dots, N_m are independent $N(0, I_d)$. It follows that the distribution of $T^2(d, m)$ does not depend on Σ , so that we can assume $\Sigma = I_d$ in the definitions (9.3) and (9.4).

We can state the second principal result of Section 8 as a second theorem:

Theorem 9.2. For the multivariate regression (9.2), consider the test statistic λ_1 in (8.2) for the hypothesis $H_0 : h' \beta = 0$ for an arbitrary $r \times 1$ column vector h . Then

$$\lambda_1 \approx \frac{T^2(d, n-r)}{n-r} \tag{9.5}$$

has a Hotelling T^2 distribution divided by $n - r$. In particular, the null distribution for λ_1 for the test $H_0 : h'\beta = 0$ is a scaled Hotelling T^2 distribution.

Proof. See (8.7) in Section 8 and the definition (9.3).

We prove in the next section that Hotelling $T^2(d, n)$ distributions are F -distributions with

$$T \approx T^2(d, n) \approx \frac{dn}{n - d + 1} F_{d, n-d+1} \tag{9.6}$$

for $n \geq d$. In particular $T^2(d, n) \approx d^2 F_{d,1}$ if $n = d$. If $n < d$, the matrices (9.1) are not invertible (with probability one) and (9.3) and (9.4) cannot be defined. If $n \geq d$, the eigenvalue λ_1 in (9.5) satisfies

$$\lambda_1 \approx \frac{T^2(d, n - r)}{n - r} \approx \frac{d}{n - r - d + 1} F_{d, n-r-d+1} \approx \frac{V_1}{V_2} \tag{9.7}$$

where V_1 and V_2 are independent chi-square random variables with d and $n - r - d + 1$ degrees of freedom, respectively.

Examples of (9.6) for Tests $H_0 : h'\beta = 0$: By (8.1)–(8.2) and (8.7), the sole nonzero eigenvalue λ_1 of the three random matrices $E^{-1}H_h$, $H_h E^{-1}$, and $E^{-1/2}H_h E^{-1/2}$ in (6.9) has the distribution (9.7) if $h'\beta = 0$, $h \neq 0$, and $n - r \geq d$.

Similarly, the two-sample Hotelling T^2 statistic in (7.3) has the distribution

$$\begin{aligned} T^2 &= \frac{n_1 n_2}{n_1 + n_2} (\bar{Z}_1 - \bar{Z}_2)' S^{-1} (\bar{Z}_1 - \bar{Z}_2) \\ &\approx (n_1 + n_2 - 2) \lambda_1 \\ &\approx T^2(d, n_1 + n_2 - 2) \approx \frac{d(n_1 + n_2 - 2)}{n_1 + n_2 - d - 1} F_{d, n_1+n_2-d-1} \end{aligned}$$

by (7.8), since $n = n_1 + n_2$ and $r = 2$ for λ_1 in (7.6) or (7.8), and (9.7).

Exercise 9.1: Suppose that $h'\beta \neq 0$ in (8.1)–(8.2). Show that

$$\lambda_1 = Z_0 \left(\sum_{i=1}^{n-r} Z_i Z_i' \right)^{-1} Z_0' \tag{9.8}$$

where Z_0, Z_1, \dots, Z_{n-r} are normally-distributed independent random vectors, Z_0 is $N(\gamma, I_d)$ for some $\gamma \neq 0$, and Z_1, \dots, Z_r are $N(0, I_d)$. Find γ in terms of h, β , and Σ .

10. The Distribution of $T^2(d, n)$: The purpose of this section is to prove that Hotelling distributions are F -distributions. Specifically,

Theorem 10.1. If T has the Hotelling $T^2(d, n)$ distribution defined in (9.3) and $n \geq d$, then

$$T = T^2(d, n) \approx \frac{dn}{n - d + 1} F_{d, n-d+1} \tag{10.1}$$

Generally, if

$$T \approx Z'_0 \left(\frac{1}{n} W \right)^{-1} Z_0 \quad \text{for} \quad W = \sum_{i=1}^n Z_i Z'_i \tag{10.2}$$

where Z_0, Z_1, \dots, Z_n are independent $N(0, I_d)$, then T has the F -distribution (10.1).

Proof. The main step is to show that the conditional distribution

$$\{ Z'_0 W^{-1} Z_0 \mid Z_0 = z_0 \} \approx (z'_0 z_0) / V, \quad V \approx \chi^2(n - d + 1) \tag{10.3}$$

That is, the conditional distribution of $Z'_0 W^{-1} Z_0$ in (10.2) given $Z_0 = z_0$ depends on z_0 only through $z'_0 z_0$ and is given by (10.3).

I claim that (10.3) is sufficient to prove (10.1). For, (10.3) implies that the conditional distribution

$$\left\{ \frac{Z'_0 W^{-1} Z_0}{Z'_0 Z_0} \mid Z_0 = z_0 \right\} \approx 1/V, \quad V \approx \chi^2(n - d + 1) \tag{10.4}$$

in particular does not depend on z_0 . By Lemma 10.1 below, this implies that the unconditioned random variable $Q = (Z'_0 W^{-1} Z_0) / (Z'_0 Z_0)$ is independent of Z_0 and has the same distribution (10.4). In turn, this implies that

$$\begin{aligned} T &= n(Z'_0 Z_0)Q \approx n \frac{V_1}{V_2} \approx \frac{nd(V_1/d)}{(n - d + 1)(V_2/(n - d + 1))} \\ &\approx \frac{nd}{n - d + 1} F(d, n - d + 1) \end{aligned}$$

where $V_1 = Z'_0 Z_0 \approx \chi^2(d)$, $V_2 = V \approx \chi^2(n - d + 1)$, and V_1 and V_2 are independent. Given (10.3), this completes the proof of (10.1).

To prove (10.3), first note that the independence of Z_0 and W implies that for any $n \times n$ orthogonal matrix Q

$$\begin{aligned} z'_0 W^{-1} z_0 &= z'_0 \left(\sum_{i=1}^n Z_i Z'_i \right)^{-1} z_0 \\ &\approx z'_0 \left(\sum_{i=1}^n (QZ_i)(QZ_i)' \right)^{-1} z_0 \approx z'_0 \left(Q \sum_{i=1}^n Z_i Z'_i Q' \right)^{-1} z_0 \\ &\approx (Q' z_0)' W^{-1} (Q' z_0) \end{aligned}$$

where, by (10.3), Q could depend on z_0 . In particular, we can choose Q so that $Q'z_0 = (\sqrt{z_0'z_0})e_1$ where e_1 is the first coordinate vector in R^d . In that case

$$z_0'W^{-1}z_0 \approx (z_0'z_0)(W^{-1})_{11} \tag{10.5}$$

where the last expression above means the (1, 1) entry of the random matrix W^{-1} . Next, write W in the partitioned form

$$W = \sum_{i=1}^n Z_i Z_i' = \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \tag{10.6}$$

where W_{11} is 1×1 , W_{12} is a $1 \times r$ row vector for $r = d - 1$, $W_{21} = W_{12}'$, and W_{22} is $r \times r$. Then by Lemma 10.2 below

$$(W^{-1})_{11} = (W_{11} - W_{12}W_{22}^{-1}W_{21})^{-1}$$

Since W_{11} and $W_{12}W_{22}^{-1}W_{21}$ are 1×1 (that is, are numbers), (10.5) implies

$$z_0'W^{-1}z_0 \approx \frac{z_0'z_0}{W_{11} - W_{12}W_{22}^{-1}W_{21}} \tag{10.7}$$

and it is sufficient for (10.3) to prove

$$W_{11} - W_{12}W_{22}^{-1}W_{21} \approx \chi^2(n - d + 1) \tag{10.8}$$

for the $d \times d$ matrix W . We express W_{ij} in (10.6) in terms of the independent standard normal random variables Z_{ia} in (10.2), using the fact that the Z_i are independent $N(0, I_d)$. Specifically, let $Y_i = Z_{i1}$ ($1 \leq i \leq n$) be the first column of Z and consider the $n \times r$ random matrix $X_{ia} = Z_{i,a+1}$ for $1 \leq a \leq r = d - 1$ composed of the remaining columns. Then by (10.6)

$$\begin{aligned} W_{11} &= \sum_{i=1}^n Z_{i1}Z_{i1} = \sum_{i=1}^n Y_i^2 = Y'Y \\ (W_{12})_a &= \sum_{i=1}^n Z_{i1}Z_{i,a+1} = \sum_{i=1}^n Y_i X_{ia} = (Y'X)_{ia} \\ (W_{22})_{ab} &= \sum_{i=1}^n Z_{i,a+1}Z_{i,b+1} = (X'X)_{ab} \end{aligned}$$

Since $W_{21} = W_{12}' = X'Y$,

$$W_{11} - W_{12}W_{22}^{-1}W_{21} = Y'Y - Y'X(X'X)^{-1}X'Y = Y(I_n - K)Y \tag{10.9}$$

Here $K = X(X'X)^{-1}X'$ is a (random) $n \times n$ orthogonal projection matrix with $\text{rank}(K) = r = d - 1$ that is independent of Y . In fact, K is the same matrix as in (8.8) except that now X is random.

Since X is random, we cannot conclude from (10.9) directly that the expression in (10.9) has a chi-square distribution as in (8.9) and (8.11), but we can conclude as before that the conditional distribution

$$\{ W_{11} - W_{12}W_{22}^{-1}W_{21} \mid X = x \} \approx \chi^2(n - r) \approx \chi^2(n - d + 1) \quad (10.10)$$

for an arbitrary $n \times r$ constant matrix x using the fact that the matrix entries Z_{ia} are independent. Then Lemma 10.1 implies that (10.10) holds without any conditioning. This implies (10.8) and hence (10.3), which completes the proof of (10.1).

We now give the statements and proofs of three lemmas.

Lemma 10.1. Assume that Q and X are two arbitrary random variables with a joint density $f(q, x)$. (Either or both of Q and X may be vector valued.) Suppose that the conditional distribution of Q given $X = x$ does not depend on x , which we can write as

$$f_{Q|X}(q \mid x) = f_{Q|X}(q) \quad (10.11)$$

Then

- (i) Q and X are independent and
- (ii) $f_{Q|X}(q) = f_Q(q)$ is the same as the marginal distribution of Q .

Proof. The joint density $f(q, x)$ for any two random variables Q and X can be written

$$f(q, x) = f_X(x)f_{Q|X}(q \mid x) \quad (10.12)$$

where $f_X(x)$ is the marginal density of X and $f_{Q|X}(q \mid x)$ is the conditional density of Q given $X = x$. Similarly, the marginal distribution of Q is

$$f_Q(q) = \int_x f(q, x) dx$$

It then follows from (10.11) and (10.12) that

$$f_Q(q) = \int_x f(q, x) dx = f_{Q|X}(q)$$

Thus the conditional density (10.11) is the same as the marginal density $f_Q(q)$. Moreover, by (10.11) and (10.12)

$$f(q, x) = f_X(x)f_Q(q)$$

This implies that Q and X are independent, which completes the proof of Lemma 10.1.

Lemma 10.2. Let A be a symmetric $d \times d$ matrix that we can write in partitioned form as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

where A_{11} , A_{22} , and A are invertible square matrices. Define

$$A^{-1} = B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$$

in the same partitioned form. Then

$$(A^{-1})_{11} = B_{11} = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} \tag{10.13}$$

Proof. This is just the generalization of Cramer’s Rule for 2×2 real matrices to 2×2 partitioned matrices, but is still somewhat nasty to prove. First, define

$$D = \begin{pmatrix} I & C \\ 0 & I \end{pmatrix} = \begin{pmatrix} I_{d_1} & C \\ 0 & I_{d_2} \end{pmatrix}$$

where A_{11} is $d_1 \times d_1$, C is a $d_1 \times d_2$ matrix, etc. Then

$$\begin{aligned} DAD' &= \begin{pmatrix} I & C \\ 0 & I \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ C' & I \end{pmatrix} \\ &= \begin{pmatrix} A_{11} + CA_{21} & A_{12} + CA_{22} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ C' & I \end{pmatrix} \\ &= \begin{pmatrix} A_{11} + CA_{21} + A_{21}C' + CA_{22}C' & A_{12} + CA_{22} \\ A_{21} + A_{22}C' & A_{22} \end{pmatrix} \end{aligned}$$

Let $C = -A_{12}A_{22}^{-1}$, so that $A_{12} + CA_{22} = 0$. Since $A_{21} = A'_{12}$, $A_{21} + A_{22}C' = 0$ as well and

$$DAD' = \begin{pmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} & 0 \\ 0 & A_{22} \end{pmatrix}$$

Note D is invertible with $D^{-1} = \begin{pmatrix} I & -C \\ 0 & I \end{pmatrix}$ and define B_{11} by (10.13). Then

$$A = D^{-1} \begin{pmatrix} B_{11}^{-1} & 0 \\ 0 & A_{22} \end{pmatrix} (D')^{-1}$$

and hence

$$\begin{aligned}
 A^{-1} &= D' \begin{pmatrix} B_{11} & 0 \\ 0 & A_{22}^{-1} \end{pmatrix} D \\
 &= \begin{pmatrix} I & 0 \\ C' & I \end{pmatrix} \begin{pmatrix} B_{11} & 0 \\ 0 & A_{22}^{-1} \end{pmatrix} \begin{pmatrix} I & C \\ 0 & I \end{pmatrix} \\
 &= \begin{pmatrix} B_{11} & 0 \\ C'B_{11} & A_{22}^{-1} \end{pmatrix} \begin{pmatrix} I & C \\ 0 & I \end{pmatrix} \\
 &= \begin{pmatrix} B_{11} & B_{11}C \\ C'B_{11} & C'B_{11}C + A_{22}^{-1} \end{pmatrix}
 \end{aligned}$$

This completes the proof of Lemma 10.2.

We include a final lemma, which can be used to obtain a shorter proof of Lemma 10.2 if A is positive definite.

Lemma 10.3. Assume $X \approx N(\mu, \Sigma)$ is a normally-distributed random vector that we write in partitioned form

$$X = \begin{pmatrix} Y \\ Z \end{pmatrix} \quad \mu = \begin{pmatrix} a \\ b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

If Σ_{22} is invertible, then the conditional distribution

$$\{Y \mid Z = z\} \approx N(a + \Sigma_{12}\Sigma_{22}^{-1}(z - b), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}) \tag{10.13}$$

Proof. Write $Y = Y - CZ + CZ$ for a matrix C . Then

$$\begin{aligned}
 \text{Cov}(Y - CZ, CZ) &= \text{Cov}(Y, CZ) - \text{Cov}(CZ, CZ) \\
 &= \text{Cov}(Y, Z)C' - C \text{Cov}(Z, Z)C' \\
 &= (\Sigma_{12} - C\Sigma_{22})C'
 \end{aligned}$$

Set $C = \Sigma_{12}\Sigma_{22}^{-1}$. Then $\text{Cov}(Y - CZ, CZ) = 0$, which implies that $Y - CZ$ and CZ are independent. In turn, this implies

$$\begin{aligned}
 \{Y \mid Z = z\} &\approx \{Y - CZ + CZ \mid Z = z\} \\
 &\approx (Y - CZ) + Cz
 \end{aligned} \tag{10.14}$$

Thus the conditional distribution $\{Y \mid Z = z\}$ is normal with

$$E(Y \mid Z = z) = E(Y - CZ) + Cz = a - Cb + Cz = a + C(z - b)$$

and by (10.14)

$$\begin{aligned}
 \text{Cov}(Y \mid Z = z) &= \text{Cov}(Y - CZ) = \text{Cov}(Y - CZ, Y - CZ) \\
 &= \text{Cov}(Y) - C \text{Cov}(Z, Y) - \text{Cov}(Y, Z)C' + C \text{Cov}(Z, Z)C' \\
 &= \Sigma_{11} - C\Sigma_{21} - \Sigma_{12}C' + C\Sigma_{22}C' \\
 &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}
 \end{aligned}$$

This completes the proof of Lemma 10.3.

Exercise 10.1: Assume that the matrix A in Lemma 10.2 is positive definite. Use (10.13) to provide a shorter proof of Lemma 10.2. (*Hint:* Consider the form of the conditional densities of Y and X .)

11. A Higher-Rank Version of $H_0 : h'\beta = 0$: A natural generalization of tests of the form $H_0 : h'\beta = 0$ for the regression $Y = X\beta + e$ is

$$H_0 : L\beta = 0 \tag{11.1}$$

where L is a $q \times r$ matrix with $\text{rank}(L) = q$. Since $L\beta$ is $q \times d$, equation (11.1) is shorthand for q different relations of the form $h'\beta = 0$ for $r \times 1$ column vectors h . If $q = 1$, then L is $1 \times r$, so that $L = h'$ for an $r \times 1$ column vector h .

An example of (11.1) would be three independent vector-valued samples

$$\begin{aligned} (Z_1)_1, (Z_1)_2, \dots, (Z_1)_{n_1} & \quad \text{where } (Z_1)_i \approx N(\mu_1, \Sigma) \\ (Z_2)_1, (Z_2)_2, \dots, (Z_2)_{n_2} & \quad \text{where } (Z_2)_j \approx N(\mu_2, \Sigma) \\ (Z_3)_1, (Z_3)_2, \dots, (Z_3)_{n_3} & \quad \text{where } (Z_3)_k \approx N(\mu_3, \Sigma) \end{aligned} \tag{11.2}$$

with $H_0 : \mu_1 = \mu_2 = \mu_3$. The one-way layout (11.2) can be put in the form $Y = X\beta + e$ as in (7.4) where now X is $n \times 3$, $\beta = (\mu_1 \ \mu_2 \ \mu_3)'$, and $n = n_1 + n_2 + n_3$. In this case, $H_0 : \mu_1 = \mu_2 = \mu_3$ is equivalent to

$$L\beta = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} \mu_1 - \mu_2 \\ \mu_1 - \mu_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{11.3}$$

which is $H_0 : L\beta = 0$ for a 2×3 matrix L .

In the univariate case ($d = 1$), one can show that if $H_0 : L\beta = 0$ holds and MSE is defined by (6.5), then

$$F = (L\hat{\beta})'(L(X'X)^{-1}L')^{-1}(L\hat{\beta})/MSE \tag{11.4}$$

has a F -distribution with $(q, n - r)$ degrees of freedom.

Exercise 11.1: Show that, if $d = 1$ and L is $q \times r$, the matrix dimensions in (11.4) work out so that (11.4) exists as a number.

Exercise 11.2: Prove or disprove: If $d = 1$ and the one-way layout (11.2) is written as $Y = X\beta + e$ for $\beta = (\mu_1 \ \mu_2 \ \mu_3)'$ analogously to (7.4) for L in (11.3), then F in (11.4) is the same as the classical one-way ANOVA test statistic.

Multivariate ANOVA and Regression Tests: A multivariate ($d > 1$) version of the test $H_0 : L\beta = 0$ for rank $q > 1$ can be based on comparing the $d \times d$ matrix

$$H_L = (L\hat{\beta})'(L(X'X)^{-1}L')^{-1}(L\hat{\beta}) \tag{11.5}$$

with the $d \times d$ residual error matrix

$$E = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

as before. Recall

$$\text{Cov}((L\hat{\beta})_L) = (L(X'X)^{-1}L') \otimes \Sigma$$

by (8.3). If $d = 1$, then H_L and E are numbers and H_L/E has an F distribution given $L\beta = 0$. As in the rank-one case ($q = 1$), the multivariate ($d > 1$) analog is more complicated, since H_L and E are $d \times d$ matrices and the three matrices

$$E^{-1}H_L \quad H_LE^{-1} \quad E^{-1/2}H_LE^{-1/2} \tag{11.6}$$

are generally different. However, as in (6.9)–(6.10), the *eigenvalues* of the three matrices (11.6) are the same. Since E^{-1} is invertible, the number of nonzero eigenvalues is the same as the rank of H_L , which can be shown to be the same as $q = \text{rank}(L)$ if $\beta \neq 0$.

If $q = \text{rank}(L) = 1$, the three matrices (11.6) have a unique nonzero eigenvalue λ_1 , which has the F -distribution (9.7) if $h'\beta = 0$.

If $q > 1$, the matrices (11.6) are generally not of rank one and have more than one nonzero eigenvalue. Since the third matrix in (11.6) is positive semidefinite, we can assume $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$. Tests of $L\beta = 0$ that do not depend on which matrix is chosen in (11.6) can be based on expressions that depend on different functions of the eigenvalues λ_i .

The four most-common tests of $H_0 : L\beta = 0$ for $q > 1$ and the corresponding functions of λ_i are:

1. Wilk's Lambda: $\Lambda = \det(E) / \det(H_L + E) = \prod_{i=1}^d \frac{1}{\lambda_i + 1}$
2. Pillai's Trace: $S_1 = \text{tr}(H_L(H_L + E)^{-1}) = \sum_{i=1}^d \frac{\lambda_i}{\lambda_i + 1}$
3. Hotelling-Lawley Trace: $S_2 = \text{tr}(H_LE^{-1}) = \sum_{i=1}^d \lambda_i$
4. Roy's Greatest Root: $S_3 = \lambda_1$

The last test is named after the Indian statistician S. N. Roy, so that Roy is not a first name. Wilk's Lambda is essentially the likelihood ratio test statistic for $H_0 : L\beta = 0$.

If $q = \text{rank}(L) = 1$, then only one eigenvalue $\lambda_1 > 0$, and that eigenvalue has the F -distribution (9.8) if $h'\beta = 0$. In that case, the four tests above are equivalent and have identical P-values.

If $q = \text{rank}(L) > 1$, the four tests use different approximations of their test statistics in terms of F distributions and give different P -values. In this case, the four tests can be viewed as tests of $H_0 : L\beta = 0$ against different alternatives.

The standard test for Roy's Greatest Root is a little different than the others in that the approximation only gives a *lower bound* for the true P -value. That is, one concludes $P \geq 0.01$ (for example) and not that P is approximately 0.01, as is the case for the other three tests. In fact, it often happens that the P-value for Roy's Greatest Root is significantly smaller than the others, which could then be significantly misleading.

See the SAS documentation for references and more details, and in particular for references for approximations of the four P-values.

References.

1. Anderson, T. W. (2003) An introduction to multivariate statistical analysis, 3rd edn. John Wiley and Sons, New York.