Multivariate Linear Models

Stanley Sawyer — Washington University

September 8, 2007 rev November 8, 2010

1. Introduction. Suppose that we have n observations, each of which has d components, which we can represent as the $n \times d$ matrix

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

For example, we may have (i) measurements of $d = 5$ air pollutants (CO, NO, etc.) on $n = 42$ widely-separated days, (ii) d test scores for n different students, (iii) best results for d Olympic events for teams from n different countries, or (iv) d different physical measurements for n individuals (human or animal) that we are trying to classify. In each case, the i^{th} row corresponds to the i^{th} multivariate observation, and the j^{th} column corresponds to the j^{th} variable measured.

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

$$Y_{ij} = \sum_{a=1}^{p} X_{ia}\beta_{aj} + e_{ij} \tag{1.2}$$

for the j^{th} component of the i^{th} individual, where $1 \le a \le p$ refers to covariates and $p = r + 1$ if there is an intercept and $p = r$ otherwise. In most cases, the first column in X corresponds to an intercept, so that $X_{i1} = 1$ for $1 \le i \le n$ and $\beta_{1j} = \mu_j$ for $1 \le j \le d$.

A key assumption in the multivariate model (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 differ. Similarly, the same student has the same age, sex, and income for all tests. 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. Also, 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 jointly normal with mean zero in R^{nd}, where $1 \le i \le n$ for observations and $1 \le j \le d$ 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 more likely to do well on the other tests as well.

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

$$\begin{aligned} \mathrm{Cov}(e_{ij}, e_{k\ell}) &= 0, \qquad i \ne k \\ \mathrm{Cov}(e_{ij}, e_{i\ell}) &= \Sigma_{j\ell} \end{aligned} \tag{1.3}$$

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

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

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

To avoid pathologies, we will assume in the following that the $p \times p$ design matrix $X'X$ is invertible and that $n \ge p + d$.

2. The Regression Model (1.2) in Terms of Matrices: As in the univariate case, we can write the regression (1.2)

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

in matrix notation as

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

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

$$\beta = \begin{pmatrix} \beta_{11} & \beta_{12} & \dots & \beta_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p1} & \beta_{p2} & \dots & \beta_{pd} \end{pmatrix}$$

is an $p \times d$ matrix. If X_{i1} is identically one, 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 $r = 2$ parameters for response to wind strength and solar intensity. Adding an intercept term means $p = r + 1 = 3$ coefficients and the parameter matrix β is 3×5. In a particular numerical example, the estimated values of the parameters β were

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

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 $p \times d$ matrix. Sometimes it will be more convenient to treat the observations Y as an nd-dimensional vector or β as an pd-dimensional vector, where $nd = 210$ and $pd = 15$ if $n = 42$, $d = 5$, and $p = 3$. If $d = 1$, then $\mathrm{Cov}(Y)$ and $\mathrm{Cov}(e)$ 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 instead of matrices. 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 $pd \times 1$ 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

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

Note that the relation $I = (i-1)d+j$ gives a one-one correspondence between pairs (i, j) with $1 \le j \le d$ and $1 \le i \le n$ and indices I with $1 \le I \le nd$. (***Exercise:*** Prove this.)

The ordering in (3.1) is called *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.

In the representation $I = (i-1)d + j$, the index j is sometimes called the *fast index* and i the *slow index*, since j always changes when I changes to $I + 1$ but i only changes when I moves on to the next row, or after j has completed a full cycle of values $1 \le j \le d$.

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 $nd \times pd$ matrix that depends somehow on the $n \times p$ matrix X. The notions of *Kronecker product* or *tensor product* of vectors or matrices are a useful way to describe these larger matrices.

Definition. Let $A = \{\, A_{ij} \,\}$ be an $m_1 \times n_1$ matrix and $B = \{\, B_{ab} \,\}$ an $m_2 \times n_2$ matrix. Then, the *tensor product* or *Kronecker product* matrix of A and B is the $m_1 m_2 \times n_1 n_2$ matrix $C = A \otimes B$ with components

$$C_{ia,jb} = A_{ij} B_{ab} \qquad (C = A \otimes B) \tag{3.3}$$

for $1 \le i \le m_1$, $1 \le j \le m_2$, $1 \le a \le m_2$, $1 \le b \le n_2$, with the notation $ia = (i-1) * m_2 + a$ and $jb = (j-1) * n_2 + b$. More exactly, C is the $m_1 m_2 \times n_1 n_2$ matrix

$$C_{IJ} = A_{ij} B_{ab} \quad \text{for} \quad I = (i-1)m_2 + a, \;\; J = (j-1)n_2 + b \tag{3.4}$$

Note that i, a in (3.3) are the slow indices (row indices) of A and B (respectively) while j, b in (3.3) are the fast indices (or column indices).

As an example, the covariance matrix (1.4) of the error terms in (3.2) can be written

$$\mathrm{Cov}(e_L) = I_n \otimes \Sigma \tag{3.5}$$

The next lemma shows how to represent the "super-matrix" X_L in (3.2) in terms of tensor products.

Lemma 3.1. Let A be an $m \times n$ matrix, B a $n \times d$ matrix, and $W = AB$ the matrix product

$$W_{ik} = \sum_{j=1}^{n} A_{ij} B_{jk} \tag{3.6}$$

Then for W_L, B_L defined as in (3.1)

$$W_L = \big(A \otimes I_d\big)\, B_L \tag{3.7}$$

where d is the second dimension for the $n \times d$ matrix B.

Proof. By (3.6)

$$\begin{aligned} W_{ik} \;&=\; \sum_{j=1}^{n} A_{ij} B_{jk} \;=\; \sum_{j=1}^{n}\sum_{\ell=1}^{d} A_{ij}\delta_{k\ell}\, B_{j\ell} \\ &=\; \sum_{j=1}^{n}\sum_{\ell=1}^{d} \big(A \otimes I_d\big)_{ik,j\ell}\, B_{j\ell} \end{aligned}$$

by (3.3), which implies (3.7).

By Lemma 3.1, the basic regression equation (2.1)

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

can be written

$$Y_L \;=\; (X \otimes I_d)\beta_L \;+\; e_L \tag{3.8}$$

so that $X_L = X \otimes I_d$ in (3.2).

With lexicographic ordering of the indices, the entries of

$$C_{IJ} = C_{ia,jb} = A_{ij} B_{ab}$$

for fixed $I = (i-1)m_2 + a$ and increasing $J = (j-1)n_2 + b$ trace out the a^{th} row of B repeatedly for each value of j, with each row of B values multiplied by A_{ij}.

This means that the matrix $C = A \otimes B$ can be written in block partitioned form as

$$C \;=\; \begin{pmatrix} a_{11}B & a_{12}B & \dots & a_{1n_1}B \\ a_{21}B & a_{22}B & \dots & a_{2n_2}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m_1 1}B & a_{n_1 2}B & \dots & a_{m_1 n_1}B \end{pmatrix} \tag{3.9}$$

In particular by (3.5)

$$\operatorname{Cov}(e_L) \;=\; \begin{pmatrix} \Sigma & 0 & \dots & 0 \\ 0 & \Sigma & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Sigma \end{pmatrix} \tag{3.10}$$

We conclude this section by proving a number of basic properties of tensor products:

Lemma 3.2. Suppose that the matrix A is $m_1 \times n_1$, B is $m_2 \times n_2$, D is a $n_1 \times k_1$, and E is $n_2 \times k_2$, so that the matrices AD and BE are defined. Then

(i) Let $C = A \otimes B$ and $F = D \otimes E$. Then

$$CF \;=\; (A \otimes B)\,(D \otimes E) \;=\; AD \otimes BE \tag{3.11}$$

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

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

(iv) Assume that A and B are invertible square matrices. Then $C = A \otimes B$ is also invertible and

$$\big(A \otimes B\big)^{-1} = A^{-1} \otimes B^{-1} \tag{3.12}$$

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 $CF = AD \otimes BE$.

(ii) By definition, $(I_n \otimes I_m)_{ia,jb} = (I_n)_{ij} (I_m)_{ab} = \delta_{ij} \delta_{ab}$, which equals one if $i = j$ and $a = b$ (or, equivalently, $ia = jb$), and is otherwise zero. This implies $I_n \otimes I_m = I_{mn}$.

(iii) By definition $(A \otimes B)_{ia,jb} = A_{ij} B_{ab}$. Hence

$$(A \otimes B)'_{ia,jb} = (A \otimes B)_{jb,ia} = A_{ji} B_{ba} = (A')_{ij} (B')_{ab} = \big(A' \otimes B'\big)_{ia,jb}$$

and $(A \otimes B)' = A' \otimes B'$.

(iv) By parts (i) and (ii),

$$(A \otimes B)\big(A^{-1} \otimes B^{-1}\big) \;=\; AA^{-1} \otimes BB^{-1} \;=\; I_n \otimes I_m \;=\; I_{mn}$$

Thus $A^{-1} \otimes B^{-1}$ is a right inverse of $A \otimes B$ and hence is the unique inverse matrix.

The next lemma is an application of Lemmas 3.1 and 3.2 that will be useful for the analysis of multivariate ANOVAs and regressions.

Lemma 3.3. Let $e = e_{ij}$ be an $n \times d$ random matrix whose rows are independent $N(0, \Sigma)$ for the same $d \times d$ covariance matrix Σ. (That is, $\mathrm{Cov}(e_L) = I_n \otimes \Sigma$ as in (3.5).) Let

$$Z_{ij} \;=\; \sum_{k=1}^{n} R_{ik} e_{kj} \tag{3.13}$$

where R is an $n \times n$ orthogonal matrix. Then Z_{ij} has the same distribution as e_{ij}. That is, the rows of Z_{ij} are independent $N(0, \Sigma)$.

Proof. Thus $Z = Re$ by (3.13), so that $Z_L = (R \otimes I_d) e_L$ by Lemma 3.1. Since e_L is a joint normal vector and $R \otimes I_d$ is a $nd \times nd$ matrix, Z_L is also joint normal, and it is sufficient to prove $\mathrm{Cov}(Z_L) = \mathrm{Cov}(e_L) = I_n \otimes \Sigma$. By (3.5), (3.13), and Lemma 3.2,

$$\begin{aligned} \mathrm{Cov}(Z_L) \;&=\; \mathrm{Cov}\big((R \otimes I_d)\, e\big) \;=\; (R \otimes I_d)\,(I_n \otimes \Sigma)\,(R \otimes I_d)' \\ &=\; R I_n R' \otimes I_d \Sigma I_d \;=\; RR' \otimes \Sigma \;=\; I_n \otimes \Sigma \end{aligned}$$

and Z has the same distribution as e.

4. The MLE of the $\boldsymbol{p \times d}$ matrix $\boldsymbol{\beta}$. The purpose of this section is to find the maximum likelihood estimator $\widehat{\beta}$ and its covariance matrix $\mathrm{Cov}(\widehat{\beta}_L)$. We first derive $\mathrm{Cov}(\widehat{\beta}_L)$ using its individual components and then show a shorter proof using tensor 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. Thus the likelihood function of the i^{th} observation Y_i in the regression $Y = X\beta + e$ in (2.1) (or equivalently of the i^{th} row of the $n \times d$ matrix Y) is the multivariate normal density

$$L(\beta, \Sigma, Y_i) = \frac{1}{\sqrt{(2\pi)^d \det\big(\Sigma\big)}} \exp(-S_i/2) \qquad \text{where} \tag{4.1}$$

$$S_i \;=\; \sum_{a=1}^{d} \sum_{b=1}^{d} \big(Y_{ia} - (X\beta)_{ia}\big)\, \Sigma^{-1}_{ab}\, \big(Y_{ib} - (X\beta)_{ib}\big)$$

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

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

$$S \;=\; \sum_{i=1}^{n}\sum_{a=1}^{d}\sum_{b=1}^{d}\big(Y_{ia}-(X\beta)_{ia}\big)\,\Sigma_{ab}^{-1}\,\big(Y_{ib}-(X\beta)_{ib}\big) \tag{4.3}$$

Finding the matrix MLE $\widehat{\beta}$ is equivalent to minimizing the triple sum S in (4.3) as a function of β. Since $(X\beta)_{ia} = \sum_{j=1}^{r} X_{ij}\beta_{ja}$ in (4.3), setting $(\partial/\partial\beta_{kc})S = 0$ leads to the set of equations

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

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”

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

by Lemma 3.1. If the $p\times p$ design matrix $X'X$ is invertible, then the matrix-valued MLE of β is

$$\widehat{\beta} = (X'X)^{-1}X'Y \qquad \text{or} \qquad \widehat{\beta}_L = \big((X'X)^{-1}X'\otimes I_d\big)Y_L \tag{4.4}$$

The first formula in (4.4) is exactly the same formula as in the univariate case ($d = 1$), except that now $\widehat{\beta}$ is a $p \times d$ matrix. The columns of $\widehat{\beta}$ for individual components of Y_{ij} are formed by applying the same $p\times n$ matrix $(X'X)^{-1}X'$ to each of the columns of Y.

In terms of components, (4.4) implies

$$\widehat{\beta}_{aj} = \beta_{aj} + \sum_{i=1}^{n} M_{ai}e_{ij}, \quad \text{where} \quad M = (X'X)^{-1}X' \tag{4.5}$$

Then since $\mathrm{Cov}(e_{ia},e_{kb}) = \delta_{ik}\Sigma_{ab}$ by (1.4)

$$\mathrm{Cov}(\widehat{\beta}_{aj},\widehat{\beta}_{bk}) \;=\; \mathrm{Cov}\left(\sum_{i=1}^{n} M_{ai}e_{ij},\ \sum_{\ell=1}^{n} M_{b\ell}e_{\ell k}\right)$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\sum_{\ell=1}^{n} M_{ai}M_{b\ell}\,\mathrm{Cov}(e_{ij},e_{\ell k}) \\
&= \sum_{i=1}^{n}\sum_{\ell=1}^{n} M_{ai}M_{b\ell}\delta_{i\ell}\Sigma_{jk} \\
&= \sum_{i=1}^{n} M_{ai}M_{bi}\,\Sigma_{jk} \;=\; (MM')_{ab}\,\Sigma_{jk} \\
&= ((X'X)^{-1})_{ab}\,\Sigma_{jk} \qquad (4.6)
\end{aligned}
$$

since $MM' = (X'X)^{-1}X'\,X(X'X)^{-1} = (X'X)^{-1}$. Thus

$$\mathrm{Cov}(\widehat{\beta}) \;=\; (X'X)^{-1}\otimes\Sigma \qquad (4.7)$$

We can derive (4.7) more easily using tensor products. By (4.5)

$$\widehat{\beta}_L \;=\; \beta_L + (Me)_L \;=\; \beta_L + \big(M\otimes I_d\big)\,e_L$$

by Lemma 3.1, and hence (using the relation $\mathrm{Cov}(AX) = A\,\mathrm{Cov}(X)A'$)

$$
\begin{aligned}
\mathrm{Cov}(\widehat{\beta}_L) &= \mathrm{Cov}\big((M\otimes I_d)\,e_L\big) \;=\; \big(M\otimes I_d\big)\,\mathrm{Cov}(e_L)\big(M\otimes I_d\big)' \\
&= \big(M\otimes I_d\big)\big(I_n\otimes\Sigma\big)\big(M'\otimes I_d\big) \\
&= \big(MI_nM'\big)\otimes\big(I_d\Sigma I_d\big) \;=\; MM'\otimes\Sigma \\
&= (X'X)^{-1}\otimes\Sigma
\end{aligned}
$$

by Lemma 3.2, since $MM' = (X'X)^{-1}$ as in (4.6).

5. The MLE of the $d\times d$ matrix Σ. The next result shows that the maximum likelihood estimator of the matrix Σ is essentially the sample covariance matrix of the multivariate residuals, which is a natural generalization of the corresponding one-dimensional result.

Theorem 5.1. Assume $n \ge p+d$. Then, the maximum likelihood estimator of Σ for the likelihood (4.2) is

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

or equivalently

$$\widehat{\Sigma} \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(Y_{i}-(X\widehat{\beta})_{i}\big)\big(Y_{i}-(X\widehat{\beta})_{i}\big)' \qquad (5.2)$$

That is, the maximum likelihood estimator $\widehat{\Sigma}$ of the $d\times d$ covariance matrix Σ 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. Let Q be the $d \times d$ matrix with entries

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

which is the right-hand side of (5.1) multiplied by n. We show in Section 8 below that the matrix Q in (5.3) is positive definite with probability one if $n \ge p + d$, which we assume. Then the object is to show $\widehat{\Sigma} = (1/n)Q$.

By (4.2)–(4.3), the likelihood $L(\widehat{\beta}, \Sigma, Y)$ is

$$L(\widehat{\beta}, \Sigma, Y) \;=\; \frac{1}{\sqrt{(2\pi)^{nd}\det\big(\Sigma\big)^{n}}}\,\exp(-S_\Sigma/2) \tag{5.4}$$

where by (5.3)

$$\begin{aligned}
S_\Sigma \;&=\; \sum_{i=1}^{n}\sum_{a=1}^{d}\sum_{b=1}^{d} \big(Y_{ia} - (X\widehat{\beta})_{ia}\big)\,\Sigma^{-1}_{ab}\,\big(Y_{ib} - (X\widehat{\beta})_{ib}\big) \\
&=\; \sum_{a=1}^{d}\sum_{b=1}^{d}\left(\sum_{i=1}^{n} \big(Y_{ia} - (X\widehat{\beta})_{ia}\big)\,\Sigma^{-1}_{ab}\,\big(Y_{ib} - (X\widehat{\beta})_{ib}\big)\right) \\
&=\; \sum_{a=1}^{d}\sum_{b=1}^{d} Q_{ab}\Sigma^{-1}_{ab} \;=\; \mathrm{tr}(Q\Sigma^{-1})
\end{aligned}$$

Thus the likelihood can be written

$$L(\widehat{\beta}, \Sigma, Y) \;=\; \frac{1}{\sqrt{(2\pi)^{nd}\det\big(\Sigma\big)^{n}}}\,\exp\big(-\frac{1}{2}\,\mathrm{tr}(Q\Sigma^{-1})\big) \tag{5.5}$$

After taking logarithms and multiplying by 2, the maximum of (5.5) over $d \times d$ positive definite matrices Σ can be found by maximizing

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

Define $A = Q^{1/2}\Sigma^{-1}Q^{1/2}$. Then $\Sigma^{-1} = Q^{-1/2}AQ^{-1/2}$ and

$$\begin{aligned}
\phi(\Sigma) \;&=\; n\log\det(Q^{-1/2}AQ^{-1/2}) \;-\; \mathrm{tr}(QQ^{-1/2}AQ^{-1/2}) \\
&=\; n\log\det(A) - n\log\det(Q^{1/2})^2 \;-\; \mathrm{tr}(AQ^{-1/2}QQ^{-1/2}) \\
&=\; -n\log\det(Q) + n\log\det(A) - \mathrm{tr}(A)
\end{aligned}$$

where Q in (5.3) is fixed. By the spectral theorem, $A = Q^{1/2}\Sigma^{-1}Q^{1/2} = RDR'$ where D is diagonal and R is orthogonal. Then $\det(A) = \det(RDR') = \det(D)$ and $\operatorname{tr}(A) = \operatorname{tr}(RDR') = \operatorname{tr}(D)$. Thus if $D = \operatorname{diag}(v_1, v_2, \ldots, v_d)$ are the eigenvalues of A,

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

The expression on the right above is maximized over either A or Σ when $v_i = n$ for all i. This implies $D = nI_d$ and $A = RDR' = RnI_dR' = nI_d$. Thus $\phi(\Sigma)$ and $L(\widehat{\beta}, \Sigma, Y)$ are maximized at

$$\widehat{\Sigma} \;=\; Q^{1/2}A^{-1}Q^{1/2} \;=\; (1/n)Q$$

for Q in (5.3). This completes the proof 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

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

or

$$H_0(a) : \widetilde{\beta}_a = 0$$

where $\widetilde{\beta}_a$ is the a^{th} row of the $p \times d$ matrix β. This is equivalent to saying that the a^{th} covariate column in X_{ia} does not affect any of the components of $Y = X\beta + e$, or the data matrix $\{\, Y_i \in R^d : 1 \le i \le n \,\}$ does not depend on the a^{th} covariate.

A natural generalization of (6.1) is

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

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

$$(h'\beta)_j = \sum_{a=1}^{p} h_a\beta_{aj} = 0, \qquad 1 \le j \le d \tag{6.3}$$

or $\sum_{a=1}^{p} h_a\widetilde{\beta}_a = 0$. Equivalently, 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 $h'\beta = \sum_{a=1}^{p} h_a \beta_a = 0$ (or $\beta_a = 0$ for a single value of a) is to use the identity

$$\mathrm{Var}(h'\widehat{\beta}) = h' \operatorname{Cov}(\widehat{\beta}) h = \sigma^2 h'(X'X)^{-1} h \qquad (d = 1)$$

If $d = 1$ and $H_0 : h'\beta = 0$, the one-dimensional test statistic

$$T = \frac{h'\widehat{\beta}}{\sqrt{(\mathrm{MSE})\, h'(X'X)^{-1}h}} \qquad \text{where} \tag{6.4}$$

$$\mathrm{MSE} = \frac{1}{n-p} \sum_{i=1}^{n} (Y_i - (X\widehat{\beta})_i)^2 \tag{6.5}$$

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

If $d > 1$, then $h'\beta = \sum_{a=1}^{p} h_a \widetilde{\beta}_a$ is a $1 \times d$ row vector, and a plausible generalization is to compare the $d \times d$ matrix

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

with the $d \times d$ residual error matrix

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

with entries

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

The matrix E in (6.8) is sometimes called an SSCP matrix, for "Sum of Squares and Cross Products", to distinguish it from the "Sum of Squares of Errors" (or SSE) for univariate regressions.

If $d = 1$, then $H_h/E = t^2/(n-p)$ for t in (6.4), which has distribution $F_{1,n-p}/(n-p)$ if $h'\beta = 0$.

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

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

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

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

Note that the matrix H_h in (6.6) has $\mathrm{rank}(H_h) = 1$, since H_h is the outer product $H_h = ww'$ for $w = (\widehat{\beta}'h)/\sqrt{h'(X'X)^{-1}h}$. In addition

Lemma 6.1. For $w = (\widehat{\beta}'h)/\sqrt{h'(X'X)^{-1}h}$ as above,

(i) The three matrices in (6.9) have the same unique nonzero eigenvalue

$$\lambda_1 \;=\; w'E^{-1}w \;=\; \frac{(\widehat{\beta}'h)'E^{-1}\widehat{\beta}'h}{h'(X'X)^{-1}h} \tag{6.11}$$

(ii) The matrices $A_1 = H_h E^{-1}$, $A_2 = E^{-1}H_h$, and $A_3 = E^{-1/2}H_h E^{-1/2}$ have eigenvectors $w_1 = w$, $w_2 = E^{-1}w$, and $w_3 = E^{-1/2}w$, respectively, for λ_1, and

(iii) If $H_0 : h'\beta = 0$ and $n \ge p + d$, the eigenvalue λ_1 has the distribution

$$\lambda_1 \;=\; w'E^{-1}w \;\approx\; \frac{d}{n-p-d+1}\,F_{d,n-p-d-1} \tag{6.12}$$

Thus the hypothesis $H_0 : h'\beta = 0$ has a test based on a F distribution for a test statistic that is essentially λ_1. We defer the proof of part (iii) to Sections 8 and 10.

Proof. First, I claim that a $d \times d$ matrix A has $\mathrm{rank}(A) = 1$ if and only if $A = uv'$ is the outer product of two non-zero vectors $u, v \in R^d$. (*Exercise*: Prove this.)

If $A = uv'$ for $u, v \ne 0$ and $Ax = \lambda x$ for $x, \lambda \ne 0$, then $Ax = u(v'x) = \lambda x$. Since $\lambda x \ne 0$, we must have $v'x \ne 0$ and $x = cu$ for some $c \ne 0$, which implies $uc(v'u) = \lambda cu$ and $\lambda = v'u$. The choice $c = 1$ (and hence $x = u$) corresponds to the normalization $v'x = \lambda$.

Assume $H_h = ww'$ as above. Then $A_1 = H_h E^{-1} = ww'E^{-1} = w(E^{-1}w)'$. Thus $w_1 = w$ is an eigenvector for eigenvalue $\lambda = w'E^{-1}w$. Similarly, $A_2 = E^{-1}H_h = (E^{-1}w)w'$ has eigenvector $w_2 = E^{-1}w$ and the same eigenvalue. The argument for A_3 is similar.

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

$$\lambda_1(a) \;=\; \widehat{\beta}_a E^{-1}\widehat{\beta}_a'/\big((X'X)^{-1}\big)_{aa} \tag{6.13}$$

where $\widehat{\beta}_a$ is the a^{th} row of $\widehat{\beta}$ and E is the SSCP matrix in (6.8). Note that the matrix X appears in the statistics λ_1 in (6.11) and (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. Before proceeding, let's show how a simple multivariate two-sample problem leads to the same test statistic (6.13).

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

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

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

$$T = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\overline{Z}_1 - \overline{Z}_2\right) / \sqrt{s^2} \qquad \text{where} \tag{7.2}$$

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

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

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

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

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 rewriting the data $(Z_1)_i, (Z_2)_j$ in (7.1) as a $(n_1 + n_2) \times d$ matrix Y with entries

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

for $n = n_1 + n_2$. The model (7.1) is then equivalent to

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

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

$$Y = X\beta + e \qquad \text{for} \qquad X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \multicolumn{2}{c}{\dots} \\ 0 & 1 \\ 0 & 1 \\ \multicolumn{2}{c}{\dots} \end{pmatrix} \qquad \text{and} \quad \beta = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \tag{7.4}$$

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, β, and Y in (7.4),

$$X'X = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix}, \qquad \widehat{\beta} = (X'X)^{-1} X'Y = \begin{pmatrix} \overline{Z}_1 \\ \overline{Z}_2 \end{pmatrix} \tag{7.5}$$

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

$$\widehat{\beta}' h \;=\; \begin{pmatrix} \overline{Z}_1 \\ \overline{Z}_2 \end{pmatrix}' \begin{pmatrix} 1 \\ -1 \end{pmatrix} \;=\; (\overline{Z}_1 - \overline{Z}_2)' \qquad \text{and}$$

$$h'(X'X)^{-1}h \;=\; (1 \quad -1) \begin{pmatrix} 1/n_1 & 0 \\ 0 & 1/n_2 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \;=\; \frac{1}{n_1} + \frac{1}{n_2}$$

The eigenvalue λ_1 in (6.11) is now

$$\begin{aligned} \lambda_1 \;&=\; (\widehat{\beta}'h)'E^{-1}(\widehat{\beta}'h)/\big(h'(X'X)^{-1}h\big) \\ &=\; \frac{n_1 n_2}{n_1 + n_2}(\overline{Z}_1 - \overline{Z}_2)'E^{-1}(\overline{Z}_1 - \overline{Z}_2) \end{aligned} \tag{7.6}$$

where $E = (Y - X\widehat{\beta})'(Y - X\widehat{\beta})$ is the residual error matrix in (6.7)–(6.8) for $n = n_1 + n_2$, with $\overline{Z}_a$ now viewed as column vectors. By (7.4), the matrix of fitted values is

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

so that the residual error matrix (6.8) is

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

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

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

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

8. The Distribution of $\boldsymbol{\lambda_1}$ for "rank one" tests $\boldsymbol{H_0 : h'\beta = 0}$:
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

$$E = (Y - X\widehat{\beta})'(Y - X\widehat{\beta}) \tag{8.2}$$

We show below that the matrix E is invertible with probability one if $n \geq p + d$. By Lemma 6.1 in Section 6, the three matrices $H_h E^{-1}$, $E^{-1} H_h$, and $E^{-1/2} H_h E^{-1/2}$ have the single nonzero eigenvalue

$$\lambda_1 = (\widehat{\beta}' h)' E^{-1} \widehat{\beta}' h \big/ \big(h'(X'X)^{-1} h\big) \tag{8.3}$$

We next derive a representation of the distribution of the test statistic λ_1 in (8.3) 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).

Theorem 8.1. We can write E in (8.2) as

$$E \;=\; \sum_{i=1}^{n-p} Z_i Z_i' \tag{8.4}$$

where $Z_1, \ldots, Z_{n-p}$ are independent $N(0, \Sigma)$. If $\beta' h = 0$, the eigenvalue λ_1 in (8.3) can be written

$$\lambda_1 \;=\; Z_0' \left(\sum_{i=1}^{n-p} Z_i Z_i' \right)^{-1} Z_0 \tag{8.5}$$

where $Z_0, Z_1, \ldots, Z_{n-p}$ are independent $N(0, \Sigma)$.

Remarks. (1) It follows from (8.4) that the $d \times d$ matrix E is invertible with probability one if and only if $n \geq p + d$. (*Exercise*: Prove this.)

(2) By Lemma 9.2 in Section 9 below, the distribution in (8.5) does not depend on Σ.

Proof of Theorem 8.1. Since $(\widehat{\beta}' h)_j = \sum_{a=1}^{p} h_a \widehat{\beta}_{aj}$, it follows from (4.6) that

$$\mathrm{Cov}(\widehat{\beta}' h)_{jk} = \sum_{a=1}^{p} \sum_{b=1}^{p} h_a h_b \,\mathrm{Cov}(\widehat{\beta}_{aj}, \widehat{\beta}_{bk}) = \sum_{a=1}^{p} \sum_{b=1}^{p} h_a h_b \left(X'X)^{-1} \right)_{ab} \Sigma_{jk}$$

and

$$\mathrm{Cov}(\widehat{\beta}' h) = \big(h'(X'X)^{-1} h\big)\, \Sigma$$

Thus the column vector $\widehat{\beta}' h$ has the multivariate normal distribution

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

and hence

$$Z_0 = (\widehat{\beta}'h - \beta'h)/\sqrt{h'(X'X)^{-1}h} \quad\approx\quad N(0,\Sigma) \tag{8.6}$$

If $h'\beta = 0$, it follows that the eigenvalue λ_1 in (8.3) can be written

$$\lambda_1 = Z_0' E^{-1} Z_0 \tag{8.7}$$

where $E = (Y - X\widehat{\beta})'(Y - X\widehat{\beta})$ is the residual error matrix.

The next step is to find the distribution of the residual error matrix E. The fitted value matrix satisfies

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

where $K = X(X'X)^{-1}X'$ is $n \times n$ and $K = K^2 = K'$. It follows from the spectral theorem that

$$K = R'DR, \qquad D = \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix} \tag{8.9}$$

where R is an $n \times n$ orthogonal matrix and k is the number of nonzero eigenvalues of K. Since $\mathrm{tr}(K) = \mathrm{tr}(R'DR) = \mathrm{tr}(DRR') = \mathrm{tr}(D) = k$ and $\mathrm{tr}(K) = \mathrm{tr}(X(X'X)^{-1}X') = \mathrm{tr}((X'X)^{-1}X'X) = \mathrm{tr}(I_p) = p$, it follows that $k = p$.

Let $Z = Re$ for the $n \times n$ matrix R in (8.9), so that

$$Z_{ij} \;=\; \sum_{a=1}^{n} R_{ia} e_{aj}, \qquad 1 \le i \le n,\ 1 \le j \le d \tag{8.10}$$

Thus the same $n \times n$ matrix R is applied to each column of e. It follows from Lemma 3.3 in Section 3 that the matrix Z has the same distribution as e. 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 matrix is

$$X\widehat{\beta} \;=\; X\beta + Ke \;=\; X\beta + R'DRe \;=\; X\beta + (R'D)Z$$

and

$$\begin{aligned} \widehat{\beta} \;&=\; (X'X)^{-1}X'X\,\widehat{\beta} \\ &=\; \beta + (X'X)^{-1}X'(R'D)Z \;=\; \beta + A(DZ) \end{aligned} \tag{8.11}$$

where $A = (X'X)^{-1}X'R'$. By (8.8), the residual values matrix $Y - X\widehat{\beta}$ is

$$\begin{aligned} Y - X\widehat{\beta} &= (X\beta + e) - (X\beta + Ke) \\ &= (I_n - K)e \;=\; R'(I_n - D)Re \;=\; R'(I_n - D)Z \end{aligned}$$

and

$$\begin{aligned} E &= (Y - X\widehat{\beta})'(Y - X\widehat{\beta}) \\ &= Z'(I_n - D)'RR'(I_n - D)Z \;=\; Z'(I_n - D)Z \end{aligned} \tag{8.12}$$

If we write $\widehat{\beta}$ and E in terms of their components,

$$\begin{aligned} \widehat{\beta}_{aj} &= \beta_{aj} + \sum_{i=1}^{n} A_{ai} \sum_{k=1}^{n} D_{ik} Z_{kj} = \beta_{aj} + \sum_{i=1}^{p} A_{ai} Z_{ij} \quad \text{and} \\ E_{ab} &= \sum_{i=1}^{n} Z_{ia} \sum_{k=1}^{n} (I_n - D)_{ik} Z_{kb} = \sum_{i=p+1}^{n} Z_{ia} Z_{ib} \end{aligned} \tag{8.13}$$

This means that $\widehat{\beta}$ and $\widehat{\beta}'h$ depend on only the first p rows of Z, while E depends only on the last $n - p$ rows of Z. In particular, (i) $\widehat{\beta}$ and E are independent and (ii) $E = \sum_{i=p+1}^{n} Z_i Z_i'$ where Z_i is the i^{th} row of Z viewed as a column vector.

Since Z_0 in (8.6) is a linear function of $\widehat{\beta}$, it follows from (8.13) that $Z_0, Z_{p+1}, \ldots, Z_n$ are independent random vectors. The relations (8.6), (8.7), and (8.13) complete the proof of Theorem 8.1.

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 $W \approx W(d, m, \Sigma)$) if W has the same distribution as the random $d \times d$ matrix

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

In particular, the Wishart distribution is a distribution of random positive semidefinite $d \times d$ 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 many of the results in Sections 2–8 in the following theorem.

Theorem 9.1. Consider the multivariate regression

$$Y = X\beta + e, \qquad e_L \;\approx\; N(0, I_n \otimes \Sigma) \tag{9.2}$$

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

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

Proof. Part (i): See (4.7) or (4.12). 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 $E \approx W(d, n_1 + n_2 - 2, \Sigma)$.

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

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

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

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 $\mathrm{Cov}(AZ_i) = A\,\mathrm{Cov}(Z_i)A' = A\Sigma A'$ and A is $p \times d$, it follows that AWA' is Wishart $W(r, m, A\Sigma A')$. 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 $AWA \approx W(d, m, \Sigma)$.

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

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

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

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

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

Lemma 9.2. The distribution $T \approx T^2(d,m)$ in (9.3) does not depend on Σ.

Proof. By (9.4) and Lemma 9.1,

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

and hence

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

where N_0, $N_1, \ldots, N_m$ 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).

The second part of Theorem 8.1 can be stated in a second theorem.

Theorem 9.2. For the multivariate regression (9.2), consider the test statistic λ_1 in (8.3) for the hypothesis $H_0 : h'\beta = 0$ for an arbitrary $p \times 1$ column vector h. Then

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

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

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

$$T \;\approx\; T^2(d,m) \;\approx\; \frac{dm}{m-d+1}\, F_{d,\,m-d+1} \tag{9.6}$$

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

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

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

Examples of (9.6) for Tests $\boldsymbol{H_0 : h'\beta = 0}$: By (8.1)–(8.3) and Theorem 8.1, the sole nonzero eigenvalue λ_1 of the three random matrices $E^{-1}H_h$, H_hE^{-1}, and $E^{-1/2}H_hE^{-1/2}$ in (6.9) has the distribution (9.7) if $h'\beta = 0$, $h \neq 0$, and $n - p \geq d$.

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

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

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)–(8.3). Show that

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

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

10. The Distribution of $\boldsymbol{T^2(d,m)}$: The purpose of this section is to prove that Hotelling distributions are F-distributions. Recall that a random variable T is said to have a Hotelling $T^2(d,m)$ distribution if it has the same distribution as

$$T \;\approx\; Z_0' \left(\frac{1}{m}\sum_{i=1}^{m} Z_i Z_i'\right)^{-1} Z_0 \tag{10.1}$$

where $Z_0, Z_1, \ldots, Z_m$ are $m+1$ independent d-dimensional standard normal vectors (that is, $Z_i \approx N(0, I_d)$) and $m \geq d$. Then

Theorem 10.1. If $T \approx T^2(d,m)$ as in (10.1) for $m \geq d$, then T has the F distribution

$$T \;=\; T^2(d,m) \;\approx\; \frac{dm}{m-d+1}\, F_{d,\,m-d+1} \tag{10.2}$$

Corollary 10.1. Given $H_0 : \beta' h = 0$ and $n \geq p + d$, the quantity λ_1 in Theorem 8.1 has the F distribution

$$\lambda_1 \;\approx\; \frac{T^2(d, n-p)}{n-p} \;\approx\; \frac{d}{n-p-d+1}\, F_{d,\,n-p-d+1}$$

Remark. Note that (10.2) is equivalent to saying

$$T \;=\; T^2(d,m) \;\approx\; m\,\frac{V_1}{V_2} \tag{10.3}$$

where $V_1 \approx \chi^2_d$, $V_2 \approx \chi^2_{m-d+1}$, and V_1 and V_2 are independent.

We begin with the statements and proofs of two 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

$$f_{Q|X}(q \mid x) \;=\; f_{Q|X}(q) \tag{10.4}$$

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

Proof of Lemma 10.1. The joint density $f(q,x)$ for any two random variables Q and X can be written

$$f(q,x) = f_X(x) f_{Q|X}(q \mid x) \tag{10.5}$$

where $f_X(x)$ is the marginal density of X and $f_{Q|X}(q \mid x)$ is the conditional density of Q given $X=x$. By (10.4), the marginal distribution of Q is

$$\begin{aligned} f_Q(q) \;&=\; \int_X f(q,x)\,dx \;=\; \int_X f_X(x) f_{Q|X}(q \mid x)\,dx \\ &=\; \int_X f_X(x)\,dx f_{Q|X}(q) \;=\; f_{Q|X}(q) \end{aligned}$$

so that the marginal density $f_Q(q)$ is the same as the conditional density (10.4). Thus $f_{Q|X}(q \mid x)\,dx = f_{Q|X}(q) \;=\; f_Q(q)$ and by (10.5)

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

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

Lemma 10.2. Let A be an invertible $d \times d$ matrix that we write (along with its inverse) in partitioned form as

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

where A_{11} and B_{11} are $d_1 \times d_1$ matrices for $d = d_1 + d_2$, $d_1, d_2 > 0$. It follows that A_{12}, B_{12} are $d_1 \times d_2$ matrices, A_{21}, B_{21} are $d_2 \times d_1$, and A_{22}, B_{22} are $d_2 \times d_2$. Assume that A_{22} is invertible. Then

$$B_{11} \;=\; \big(A_{11} - A_{12}A_{22}^{-1}A_{21}\big)^{-1} \tag{10.6}$$

Proof of Lemma 10.2. This is a generalization of Cramér's rule for 2×2 real matrices to 2×2 partitioned matrices. By definition

$$\begin{aligned} AB \;&=\; \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} \;=\; \begin{pmatrix} I_{d_1} & 0 \\ 0 & I_{d_2} \end{pmatrix} \\ &=\; \begin{pmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{pmatrix} \end{aligned}$$

Thus $A_{11}B_{11} + A_{12}B_{21} = I_{d_1}$ and $A_{21}B_{11} + A_{22}B_{21} = 0$. The second relation implies $A_{22}B_{21} = -A_{21}B_{11}$ and hence $B_{21} = -A_{22}^{-1}A_{21}B_{11}$. The first relation then becomes

$$\begin{aligned} A_{11}B_{11} + A_{12}B_{21} \;&=\; A_{11}B_{11} - A_{12}A_{22}^{-1}A_{21}B_{11} \\ =\; \big(A_{11} - A_{12}A_{22}^{-1}A_{21}\big)\, B_{11} \;&=\; I_{d_1} \end{aligned}$$

It follows that $A_{11} - A_{12}A_{22}^{-1}A_{21}$ is invertible and (10.6) holds.

Lemma 10.2 has an interesting corollary, for which we give an alternative proof:

Corollary 10.1. Assume $X \approx N(\mu, \Sigma)$ is a normally-distributed random vector that we write in partitioned form

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

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

$$\{\, Y \mid Z = z \,\} \;\approx\; N\big(a + \Sigma_{12}\Sigma_{22}^{-1}(z - b),\; \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\big) \tag{10.7}$$

Proof of Corollary 10.1. Write $Y = Y - CZ + CZ$ for a matrix C. Then

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

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

$$\begin{aligned} \{\, Y \mid Z = z \,\} \;&\approx\; \{\, Y - CZ + CZ \mid Z = z \,\} \\ &\approx\; (Y - CZ) + Cz \end{aligned} \tag{10.8}$$

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 (10.8)

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

This completes the proof of the corollary.

We are now ready to begin the proof of Theorem 10.1.

Proof of Theorem 10.1. Let $W = \sum_{i=1}^{m} Z_i' Z_i$ where $Z_0, Z_1, \ldots, Z_m$ are independent $N(0, I_d)$. The main step is to show that the conditional distribution of $Z_0' W^{-1} Z_0$ given $Z_0 = z_0$ is

$$\{\, Z_0' W^{-1} Z_0 \mid Z_0 = z_0 \,\} \;\approx\; (z_0' z_0)/V_2 \tag{10.9}$$

where $V_2 \approx \chi^2(m - d + 1)$.

Proof that (10.9) implies Theorem 10.1. Since Z_0 and W are independent, the relation (10.9) implies that

$$\left\{ \frac{Z_0' W^{-1} Z_0}{Z_0' Z_0} \;\middle|\; Z_0 = z_0 \right\} \;\approx\; 1/V_2, \qquad V_2 \approx \chi^2(n - d + 1) \tag{10.10}$$

Note that the right-hand side of (10.10) does not depend on z_0. This implies by Lemma 10.1 that

(i) The unconditioned $Q = (Z_0' W^{-1} Z_0)/(Z_0' Z_0)$ has the same distribution (10.10) and

(ii) $Q = (Z_0' W^{-1} Z_0)/(Z_0' Z_0)$ is independent of Z_0.

Since $T = m(Z_0' Z_0)\,Q$, this implies

$$T \;=\; m(Z_0' Z_0)\,Q \;\approx\; mV_1 \frac{1}{V_2} \;\approx\; m\frac{V_1}{V_2}$$

where $V_1 = Z_0'Z_0 \approx \chi_d^2$ is independent of $Q = 1/V_2$. This implies (10.3) and hence (10.2), which completes the proof of Theorem 10.1 given (10.9). It only remains to prove (10.9).

Proof of (10.9). Since Z_0 and W are independent, the distribution

$$\{\ Z_0'W^{-1}Z_0 \mid Z_0 = z_0\ \} \approx z_0'W^{-1}z_0$$

Let B be a $d \times d$ orthogonal matrix. Since $Z_1, \ldots, Z_m$ are independent $N(0, I_d)$, it follows that $BZ_1, \ldots, BZ_m$ are also independent $N(0, I_d)$ and

$$\begin{aligned} z_0'W^{-1}z_0 &= z_0'\left(\sum_{i=1}^m Z_iZ_i'\right)^{-1} z_0 \;\approx\; z_0'\left(\sum_{i=1}^m (BZ_i)(BZ_i)'\right)^{-1} z_0 \\ &= z_0'\left(B\sum_{i=1}^m Z_iZ_i'\,B'\right)^{-1} z_0 \;=\; z_0'B\left(\sum_{i=1}^m Z_iZ_i'\right)^{-1} B'z_0 \\ &= (B'z_0)'W^{-1}(B'z_0) \end{aligned}$$

Since B can depend on z_0, we can choose B so that $B'z_0 = (\sqrt{z_0'z_0})\ e_1$ where e_1 is the first coordinate vector in R^d. Then

$$z_0'W^{-1}z_0 \;\approx\; (B'z_0)'W^{-1}(B'z_0) \;=\; (z_0'z_0)\big(W^{-1}\big)_{11} \tag{10.11}$$

where the last expression above means the $(1,1)$ entry of the $d \times d$ random matrix W^{-1}.

Write W in the partitioned form

$$W \;=\; \sum_{i=1}^m Z_iZ_i' \;=\; \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \tag{10.12}$$

where W_{11} is 1×1, W_{12} is a $1 \times r$ for $r = d-1$, $W_{21} = W_{12}'$ is $r \times 1$, and W_{22} is $r \times r$. Lemma 10.2 above then implies

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

Since W_{11} and $W_{12}W_{22}^{-1}W_{21}$ are 1×1 (that is, are numbers), (10.11) implies

$$z_0'W^{-1}z_0 \;\approx\; \frac{z_0'z_0}{W_{11} - W_{12}W_{22}^{-1}W_{21}} \tag{10.13}$$

To prove (10.9), it is now sufficient to prove

$$W_{11} - W_{12}W_{22}^{-1}W_{21} \;\approx\; \chi^2_{m-d+1} \tag{10.14}$$

Since $W = \sum_{i=1}^m Z_i Z_i'$ where $Z_1, \ldots, Z_m$ are independent $N(0, I_d)$, we can write $W_{ab} = \sum_{i=1}^m Z_{ia} Z_{ib}$ where Z_{ia} are univariate independent standard normal random variables for $1 \le i \le m$ and $1 \le a \le d$.

For definiteness, let $Y_i = Z_{i1}$ $(1 \le i \le m)$ be the first column of Z and let X be the $n \times r$ random matrix $X_{ia} = Z_{i,a+1}$ for $1 \le a \le r = d-1$ defined by the remaining columns. Then for $1 \le a \le r$ and $1 \le b \le r$

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

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

$$W_{11} - W_{12} W_{22}^{-1} W_{21} \;=\; Y'Y - Y'X(X'X)^{-1}X'Y \;=\; Y'(I_m - K)Y \tag{10.15}$$

where $K = X(X'X)^{-1}X'$ is independent of Y.

Conditional on $X = x \in R^r$, K is an $m \times m$ orthogonal projection matrix with $\mathrm{rank}(K) = r = d - 1$. Similarly $\mathrm{rank}(I_m - K) = m - r = m - d + 1$. Since Y_i are independent $N(0,1)$ for $1 \le i \le m$, $Y'Y \approx \chi_m^2$ and $Y'(I_m - K)Y \approx \chi_{m-r}^2 = \chi_{m-d+1}^2$ conditional on $X = x$. Since the latter distribution does not depend on x, it follows from a second application of Lemma 10.1 that the unconditional distribution of $Y'(I_m - K)Y$ in (10.15) is also χ_{m-d+1}^2.

This implies (10.14), which by (10.13) implies (10.9) and completes the proof of Theorem 10.1.

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

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

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

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

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

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

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

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

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

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

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

Exercise 11.1: Show that, if $d = 1$ and A is $q \times p$, 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 A 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 $H_0 : A\beta = 0$ for rank $q > 1$ can be based on comparing the $d \times d$ matrix

$$H_A = (A\widehat{\beta})'(A(X'X)^{-1}A')^{-1}(A\widehat{\beta}) \tag{11.5}$$

with the $d \times d$ residual error matrix

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

as before. Since A is $q \times p$ and $\widehat{\beta}$ is $p \times d$,

$$\begin{aligned} \mathrm{Cov}\big((A\widehat{\beta})_L\big) &= \mathrm{Cov}\Big((A \otimes I_p)\widehat{\beta}_L\Big) = (A \otimes I_p)\,\mathrm{Cov}(\widehat{\beta}_L)(A' \otimes I_p) \\ &= (A \otimes I_p)((X'X)^{-1} \otimes \Sigma)(A' \otimes I_p) = (A(X'X)^{-1}A') \otimes \Sigma \end{aligned}$$

as in (3.9). If $d = 1$, then H_A and E are numbers and H_A/E has an F distribution given $A\beta = 0$. As in the rank-one case ($q = 1$), the multivariate ($d > 1$) analog is more complicated, since H_A and E are $d \times d$ matrices and the three matrices

$$E^{-1}H_A \qquad H_A E^{-1} \qquad E^{-1/2}H_A E^{-1/2} \tag{11.6}$$

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_A, which can be shown to be the same as $q = \mathrm{rank}(A)$ if $\beta \neq 0$.

If $q = \mathrm{rank}(A) = 1$, 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 $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$. Tests of $A\beta = 0$ 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 $H_0 : A\beta = 0$ 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: $S_1 = \mathrm{tr}\big(H_L(H_L+E)^{-1}\big) = \sum_{i=1}^d \frac{\lambda_i}{\lambda_i+1}$
3. Hotelling-Lawley Trace: $S_2 = \mathrm{tr}(H_L E^{-1}) = \sum_{i=1}^d \lambda_i$
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 $q = \mathrm{rank}(L) = 1$, 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 $q = \mathrm{rank}(L) > 1$, 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.