



Universidad Autónoma Chapingo
División de Ciencias Forestales

Métodos Estadísticos Multivariados en Dasonomía //

Tesis que presenta:

Arturo Antonio Alvarado Segura

como requisito parcial
para obtener el grado de:

**Maestro en Ciencias
en Ciencias Forestales**



DIRECCION ACADEMICA
DEPTO. DE SERVICIOS ESCOLARES
OFICINA DE EXAMENES PROFESIONALES

C.A.

859031

Chapingo, Edo. de México

1998



Métodos Estadísticos Multivariados en Dasonomía

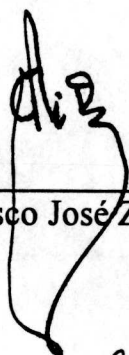
Por:

Arturo Antonio Alvarado Segura

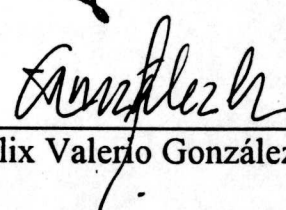
Esta tesis se realizó bajo la dirección del Dr. Francisco José Zamudio Sánchez, misma que ha sido aprobada por el siguiente Comité Asesor y aceptada como requisito parcial para obtener el grado de:

Maestro en Ciencias en Ciencias Forestales

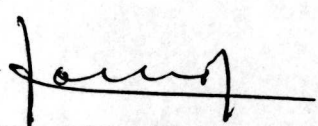
Director:


Dr. Francisco José Zamudio Sánchez

Asesor:


Dr. Félix Valerio González Cossío

Asesor:


Dr. José Luis Romo Lozano

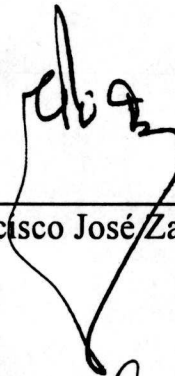
32519

Chapingo, Edo. de Méx., México. Enero de 1998.

Métodos Estadísticos Multivariados en Dasonomía

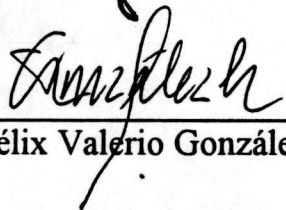
El jurado que revisó y aprobó el examen de grado de **Arturo Antonio Alvarado Segura**, autor de la presente tesis de Maestría en Ciencias, estuvo constituido por:

Presidente:



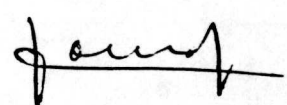
Dr. Francisco José Zamudio Sánchez

Asesor:



Dr. Félix Valerio González Cossío

Asesor:



Dr. José Luis Romo Lozano

Chapingo, Edo. de Méx., México. Enero de 1998.

Le can dzo'ojke', chen bin k-kalante' ba'ax k-yakumaj.

Chen bin k-yakumte' ba'ax k-na'atic.

Chen bin k-na'ate' ba'ax ka'ansmaj.

Al final, sólo conservaremos aquello que amamos.

Sólo amaremos aquello que entendamos.

Sólo entenderemos aquello que hemos enseñado.

**Baba Dioum
Calan-sihnalil Senegal
(Conservacionista de Senegal)**

*Dedico cariñosamente
este trabajo a Blanca*

AGRADECIMIENTOS

Deseo hacer patente mi agradecimiento al **Consejo Nacional de Ciencia y Tecnología (CONACyT)** y a la **Universidad Autónoma Chapingo (UACH)**, instituciones que financiaron mis estudios de maestría. Un reconocimiento especial a mis profesores del Postgrado de la **División de Ciencias Forestales (DiCiFo)** de la UACH, que con esmero han construido un programa académico, en mi opinión, atractivo y de calidad.

Estoy muy agradecido con los miembros de mi **Comité Asesor**, por la invaluable ayuda que me han proporcionado, ya que a través de sus sugerencias, puntos de vista y seguimiento, se fue conformando este trabajo: al Dr. Zamudio, por la dirección y la revisión pormenorizada y oportuna que hizo del mismo, por sus enseñanzas y palabras de estímulo, por todas las facilidades y apoyos decisivos que me ha otorgado y, ciertamente, por su inmensurable tolerancia hacia mí; al Dr. González, por la revisión fructífera que realizó; al Dr. Romo, por sus comentarios para mejorar la estructura y la redacción de partes del documento, así como por su manifiesta disposición en ayudarme.

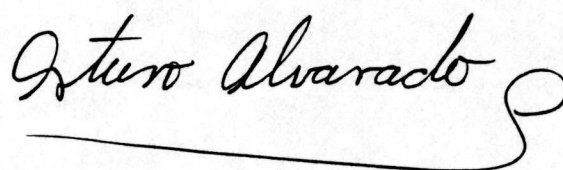
Deseo mencionar a la Dra. Amparo Borja, por su amable dedicación en traducir algunos textos de folletos en francés que tratan sobre métodos multivariados que tienen aplicaciones en la agronomía. Mis expresiones de agradecimiento también para el Ing. Carlos F. Romahn, el Mtro. Gabriel Rodríguez y el Mtro. Carlos Cíntora, por la ayuda múltiple que he recibido de ellos durante mucho tiempo y por distinguirme con su amistad.

Para el desarrollo de los ejemplos en este documento, así como otros que no se reportan, fue indispensable contar con la cortesía de las siguientes personas que me autorizaron el uso de sus datos: el Profr. Daniel Suárez y los ingenieros Tomás Martínez, Miguel Ávalos y Nancy May. Los créditos de los datos usados se citan a pie de cuadro.

Durante esta etapa de mi formación fue importante haber contado con la maravillosa amistad y el apoyo de las siguientes personas: Lic. Maggie Cano, Biól. Clarita Alfaro, Ing. Ramiro López, Mat. Tomás Vera, Profr. Eduardo Zavala, Quím. Carlos Huerta y Profra. Esperanza Victoria; mi profundo agradecimiento a ellos.

Ocupan un lugar en mi mente, mis amigos y compañeros siguientes: Francisco Vargas, Carlos Caba, Martina Esparza, J. Alberto Chablé, Margarita Fuentes, Géner Díaz, Antonio García D., Bernardo Caamal, Ángel Leyva, Felipe Puc, Rebeca González, Felipe Puerto, Arisbe Mendoza, Gustavo Rojo, Alfonso Alvarado, Adrián López, Rosa Martínez, Joel Díaz, Norma Quirós, Vidal Cob y Juan Basulto; gracias a todos por el privilegio de sus ideas y los momentos compartidos. A quienes sin intención he obviado, pido una disculpa.

Muchísimas gracias por su apoyo. Sinceramente,

A handwritten signature in dark ink, reading "Gustavo Alvarado". The signature is written in a cursive style with a long, sweeping underline that extends to the right.

CONTENIDO

AGRADECIMIENTOS	vi
CONTENIDO	viii
ÍNDICE DE CUADROS	xiv
SÍMBOLOS USADOS	xvi
RESUMEN	xviii
ABSTRACT	xix
I INTRODUCCIÓN	1
II REVISIÓN DE LITERATURA	4
2.1. La Dasonomía y sus Divisiones	4
2.2. La Estadística y sus Divisiones	7
2.3. Relación de la Estadística y la Dasonomía	10
2.4. Elementos de Álgebra Matricial	11
2.4.1. Combinación lineal	11
2.4.2. Inversa de una matriz	12
2.4.3. Productos Kronecker	13
2.4.4. Valores característicos y vectores característicos	14
2.4.5. Teorema de la descomposición espectral	15
2.4.6. Teorema de la descomposición del valor singular	19
III MATERIALES Y MÉTODOS	20
3.1. Presentación de los Métodos Multivariados	20
3.2. Datos Sobre Varias Dimensiones	21

IV LAS TÉCNICAS MULTIVARIADAS Y SUS APLICACIONES	25
4.1. Generalidades	26
4.1.1. El contexto multivariado	26
4.1.2. La matriz de datos	27
4.1.3. Muestra aleatoria multivariada	30
4.1.4. El vector de medias y la matriz de covarianzas	30
4.1.5. Medidas de dispersión multivariada	34
4.1.6. Combinaciones lineales	34
4.1.6.1. Transformaciones de escala	36
4.1.6.2. Transformación de Mahalanobis	36
4.1.6.3. Transformación componentes principales	36
4.1.7. La distribución multinormal	37
4.1.7.1. Algunas propiedades	38
4.1.7.2. La distribución normal de una matriz	38
4.1.8. La distribución Wishart	39
4.1.8.1. Propiedades de las matrices Wishart	39
4.1.8.2. Matrices Wishart particionadas	40
4.1.9. La distribución T^2 de Hotelling	42
4.1.10. El estadístico T^2 de Hotelling para dos muestras	44
4.1.11. Estadísticos con base en la distribución Wishart	45
4.2. Pruebas de Hipótesis	49
4.2.1. Introducción	49
4.2.2. Pruebas de hipótesis sobre el vector μ y la matriz Σ	50
4.2.2.1. La hipótesis $H_0: \mu = \mu_0$ cuando Σ es conocida	50
4.2.2.2. La hipótesis $H_0: \mu = \mu_0$ cuando Σ es desconocida	51
4.2.2.3a. La hipótesis $H_0: \Sigma = \Sigma_0$ con μ desconocida (PRV)	52
4.2.2.3b. La hipótesis $H_0: \Sigma = \Sigma_0$ con μ desconocida (PUI)	53
4.2.3. Hipótesis con base en muestras individuales	55

4.2.3.1. Hipótesis sobre μ con base en una sola muestra	55
4.2.3.2. Hipótesis sobre Σ con base en una sola muestra	58
4.2.4. Hipótesis sobre muestras múltiples	65
4.2.4.1. La hipótesis $H_{0b}: \mu_1 = \mu_2 = \dots = \mu_k$ dado que $\Sigma_1 = \Sigma_2 = \dots = \Sigma_k$	65
4.2.4.2. La prueba de homogeneidad de covarianzas	71
4.2.4.3. La prueba de homogeneidad completa	74
4.2.5. Intervalos de confianza simultáneos	74
4.2.5.1. El caso de la T^2 de Hotelling para una muestra	74
4.2.5.2. El caso de la T^2 de Hotelling para dos muestras	77
4.2.6. Anexo de programas en SAS	78
4.3. Regresión Lineal Multivariada	90
4.3.1. Introducción	90
4.3.2. Regresión múltiple univariada	90
4.3.2.1. El modelo	90
4.3.2.2. La estimación del vector de parámetros y las sumas de cuadrados	92
4.3.2.3. Propiedades muestrales de los estimadores mínimos cuadrados ordinarios	93
4.3.2.4. Intervalos de confianza	95
4.3.3. Regresión lineal multivariada	99
4.3.3.1. El modelo de regresión lineal multivariada	99
4.3.3.2. Estimadores de máxima verosimilitud de $\mathbf{B}$ y Σ y las sumas de cuadrados y productos cruzados	101
4.3.3.3. Propiedades muestrales de los estimadores y la distribución de $\hat{\mathbf{B}}$ y $\hat{\Sigma}$	103
4.3.4. Pruebas de razón de verosimilitud de los parámetros de regresión	106
4.3.4.1. Hipótesis de que las respuestas no dependen de las últimas $q - r$ variables de $\mathbf{X}$	106
4.3.5. Predicciones de la regresión multivariada	108

4.3.6. Hipótesis lineal general	111
4.3.6.1. La prueba de razón de verosimilitud (PRV)	111
4.3.6.2. La prueba unión intersección (PUI)	114
4.3.6.3. Intervalos de confianza simultáneos (ICS)	115
4.3.7. Matrices diseño de rango incompleto	115
4.3.8. Correlación múltiple	117
4.3.8.1. El efecto de la media	117
4.3.8.2. Coeficiente de correlación múltiple	118
4.3.8.3. Coeficiente de correlación parcial	119
4.3.8.4. Medidas de correlación entre vectores	120
4.3.9. Anexo de datos y programas	122
4.4 Análisis de Varianza Multivariado	129
4.4.1. Introducción	129
4.4.2. Diseño completamente al azar	132
4.4.2.1. Formulación univariada	132
4.4.2.1.1. Sumas de cuadrados y cuadrados medios	134
4.4.2.1.2. Prueba de hipótesis sobre los tratamientos	136
4.4.2.1.3. Esperanzas de las sumas de cuadrados	137
4.4.2.2. Formulación multivariada estándar	137
4.4.2.2.1. Sumas de cuadrados y productos cruzados (SCPC)	140
4.4.2.2.2. Prueba de hipótesis sobre los tratamientos	142
4.4.3. El enfoque unión intersección	146
4.4.4. Diseño de bloques completos al azar	148
4.4.4.1. Formulación univariada estándar	148
4.4.4.2. Formulación univariada con submuestreo	151
4.4.4.2.1. El modelo lineal	151
4.4.4.2.2. Sumas de cuadrados y esperanzas de los cuadrados medios	153
4.4.4.2.3. Pruebas de hipótesis	154
4.4.4.3. Formulación multivariada estándar	155
4.4.4.3.1. El modelo lineal	156

4.4.4.3.2. Sumas de cuadrados y productos cruzados	157
4.4.4.4. Formulación multivariada y con submuestreo	157
4.4.4.4.1. El modelo lineal	158
4.4.4.4.2. Planteamiento de las pruebas de hipótesis	159
4.4.4.4.3. Sumas de cuadrados y productos cruzados	160
4.4.4.4.4. Prueba para interacciones	161
4.4.4.4.5. Prueba para efectos principales	162
4.4.5. Anexo de programas	166
4.5. Método de Componentes Principales	168
4.5.1. Introducción	168
4.5.2. Componentes principales poblacionales	172
4.5.2.1. Propiedades de las componentes principales poblacionales	173
4.5.3. Componentes principales muestrales	176
4.5.3.1. Propiedades de las componentes principales muestrales	178
4.5.4. Correlación entre las variables originales y las componentes principales	180
4.5.5. Estructura de correlación muestral	183
4.5.6. Graficación de las componentes principales	186
4.5.7. Pruebas de hipótesis sobre las componentes principales	188
4.5.7.1. La hipótesis que $(\lambda_1 + \dots + \lambda_k) / (\lambda_1 + \dots + \lambda_p) = \psi$	189
4.5.7.2. La hipótesis sobre variación isotrópica	191
4.5.7.3. Pruebas de hipótesis que involucran a la matriz de correlaciones	195
4.5.7.3.1. La hipótesis de que todos los eigenvalores de la matriz poblacional P de correlaciones son iguales	195
4.5.7.3.2. La hipótesis que los últimos $(p - k)$ valores característicos de P son iguales	196
4.5.8. El uso de las componentes principales en la regresión múltiple	196
4.5.9. Anexo de programas	201
4.6. Método de Correlaciones Canónicas	203
4.6.1. Introducción	203
4.6.2. Correlaciones canónicas poblacionales	206

4.6.3. Correlaciones canónicas muestrales	209
4.6.4. La hipótesis $\Sigma_{12} = \mathbf{O}$	211
4.6.5. Puntaje y predicción	212
4.6.6. Datos cualitativos y el uso de variables ficticias (“de engaño”)	214
4.6.7. Anexo de Programas	217
V CONCLUSIONES	221
VI LITERATURA CITADA	223

ÍNDICE DE CUADROS

Cuadro 3.2.1	Mediciones de longitud y ancho del sépalo y del pétalo de tres especies de <i>Iris</i> spp.	22
Cuadro 3.2.2	Mediciones de variables sobre los sitios de anidamiento de pájaros de helecho y sitios ubicados aleatoriamente en la misma área	23
Cuadro 3.2.3	Número de individuos/ha de la regeneración natural en una selva mediana subperennifolia en Escárcega, Campeche, México	23
Cuadro 3.2.4	Mediciones del caparazón (mm) de 48 tortugas pintas	24
Cuadro 3.2.5	Pesos (centigramos) de depósitos de corcho de 28 árboles en los cuatro puntos cardinales	24
Cuadro 4.2.1	Mediciones de 4 dimensiones en 31 coníferas	31
Cuadro 4.3.1	Diámetro normal, altura y volumen de una muestra de 35 árboles	122
Cuadro 4.3.2	Peso verde de la leña (kg) de ramas de <i>Gliricidia sepium</i>	123
Cuadro 4.3.3	Peso verde del forraje (kg) de ramas de <i>Gliricidia sepium</i>	123
Cuadro 4.4.1	Peso (mg) de 105 pupas de <i>Hypsipyla grandella</i>	133
Cuadro 4.4.2	Efectos de fertilizantes de N y P y sus combinaciones, sobre tres variables-respuestas de plántulas de <i>Gmelina arborea</i>	138
Cuadro 4.4.3	Análisis de varianza multivariado del diseño completamente al azar	143
Cuadro 4.4.4	Observaciones de un diseño en bloques completos al azar estándar	149
Cuadro 4.4.5	Sumas de cuadrados y esperanzas de los cuadrados medios del diseño en bloques completos al azar estándar	150
Cuadro 4.4.6	Observaciones de un bloques al azar con submuestreo	151
Cuadro 4.4.7	Sumas de cuadrados y esperanzas de los cuadrados medios del diseño en bloques completos al azar con submuestreo	153
Cuadro 4.4.8	Observaciones de un bloques completos al azar; el caso	

multivariado estándar	156
Cuadro 4.4.9 Matrices de las sumas de cuadrados y productos cruzados del diseño en bloques completos al azar multivariado estándar	157
Cuadro 4.4.10 Observaciones de un bloques completos al azar; el caso multivariado con submuestreo	158
Cuadro 4.4.11 Datos de tres variables-respuesta en un experimento de regulado- res del crecimiento con plántulas de <i>Acacia mangium</i> de catorce semanas	163
Cuadro 4.5.1 Calificaciones de 55 estudiantes en 6 materias	169
Cuadro 4.5.2 Las seis componentes principales generadas a partir de las calificaciones del Cuadro 4.5.1	181
Cuadro 4.5.3 Variables dasométricas de 20 eucaliptos y nutrimentos de las muestras de suelo correspondientes	199
Cuadro 4.5.4 Componentes principales generadas a partir de los nueve nutrimentos de los 20 sitios del Cuadro 4.5.3	199
Cuadro 4.6.1 Mediciones de 4 dimensiones en 31 coníferas	205

SÍMBOLOS USADOS

$a = b$: a es igual a b .

$a \neq b$: a no es igual que b ; a es diferente de b .

$a > b$: a es mayor que b .

$a \geq b$: a es mayor que ó igual a b .

$p < q$: p es menor que q .

$g \leq h$: g es menor que ó igual a h .

$i = 1, 2, \dots, n$: i toma los valores de 1, 2, así sucesivamente, hasta que vale n .

$\sum_{i=1}^m x_i = x_1 + x_2 + \dots + x_m$: la sumatoria de las x_i , desde $i = 1$ hasta $i = m$.

$\prod_{j=1}^k x_j = (x_1)(x_2) \dots (x_k)$: el producto de las x_j , desde $j = 1$ hasta $j = k$.

UE: unidad experimental

f.v.: fuente de variación

g.l.: grados de libertad

$\text{tr}(\cdot)$: la traza de $(\cdot)$. Por ej., $\text{tr}(\mathbf{A})$ es la traza de la matriz $\mathbf{A}$

$r(\cdot)$: el rango de $(\cdot)$. Por ej., $r(\mathbf{A})$ es el rango de la matriz $\mathbf{A}$

H_0 : así se denotará a la hipótesis nula

H_1 : normalmente así se denota a la hipótesis alternativa

EP: estadístico de prueba

$\sim$: se distribuye como

$X \sim N(\mu, \sigma^2)$: la variable aleatoria X tiene distribución normal con media μ y varianza σ^2 .

$\mathbf{x} = [X_1, X_2]' \sim N_2(\mu, \Sigma)$: el vector aleatorio $\mathbf{x}$ tiene una distribución normal bivariada con vector de medias μ y matriz de covarianzas Σ

$\mathbf{x} \sim N_p(\mu, \Sigma)$: el vector aleatorio $\mathbf{x}$ tiene distribución normal p -variada con vector de medias μ y matriz de covarianzas Σ

$X \sim \chi_p^2$: la variable aleatoria X tiene una distribución ji-cuadrada con p g.l.

$Y \sim F_{m,n}$: la variable aleatoria Y tiene una distribución F de Snedecor con m g.l. en el numerador y n g.l. en el denominador.

$EP \stackrel{H_0}{\sim} \chi_k^2$: EP se distribuye, bajo el supuesto de que H_0 es verdadera, como una ji-cuadrada con k g.l.

$Y_i = X_i^2 Z_i \forall i$: la igualdad se cumple para todos los valores que tome i .

RESUMEN

Arturo Antonio Alvarado Segura

MÉTODOS ESTADÍSTICOS MULTIVARIADOS EN DASONOMÍA

1998

(Bajo la dirección de Francisco José Zamudio Sánchez, Ph.D.)

Se discuten algunas técnicas estadísticas multivariadas ejemplificándolas con datos del bosque y/o sus recursos asociados. Los métodos multivariados están relacionados con el análisis de observaciones que contienen mediciones sobre más de una variable. Estos métodos permiten una comprensión global del fenómeno bajo estudio ya que indagan las relaciones que existen tanto entre variables como entre individuos. Frecuentemente los estudios que se hacen en la dasonomía involucran la necesidad de coleccionar varias variables —como altura, diámetro, ancho de copa, peso, magnitudes anatómicas de la madera, magnitudes corporales, etc— sobre un grupo de individuos —como árboles, especies faunísticas, especies herbáceas o florísticas, etc. Por lo tanto, la presentación metodológica de algunas técnicas multivariadas que tienen aplicaciones en la dasonomía puede ser de utilidad para los investigadores cuyo interés principal no sea la estadística, pero deseen aplicarla adecuadamente para reportar, con aceptación generalizada y formal, los resultados de sus estudios.

Palabras clave: análisis de varianza multivariado, componentes principales, correlaciones canónicas, dasonomía, matriz de covarianza, pruebas de hipótesis, regresión lineal multivariada.

ABSTRACT

Arturo Antonio Alvarado Segura

MULTIVARIATE STATISTICAL METHODS IN FORESTRY

1998

(Under the direction of Francisco José Zamudio Sánchez, Ph.D.)

Some multivariate statistical techniques are discussed using examples in the context of forest and/or their associated resources. Multivariate methods are involved in analyzing observations on more than one variable. These methods allow a whole comprehension of the phenomena being studied since they are concerned with relationships as among variables as between different individuals. Usually surveys that are done in forestry involve several variables —such as height, diameter, crown width, weight, anatomical measurements of wood, magnitudes of a body, and so on— on a set of individuals —such as trees, species of wildlife, species of grasses and flowers, etc. Thus, methodological presentation of some multivariate techniques which have applications in forestry can be useful for researchers whose main interest is not statistics but wish a formal and general acceptance of their works through an appropriate use of statistics.

Key words: canonical correlations, covariance matrix, forestry, hypothesis testing, multivariate analysis of variance, multivariate linear regression, principal components.

I INTRODUCCIÓN

La dasonomía constituye una estructura de conocimientos relacionada con el bosque y sus recursos asociados, entendiéndose por bosque, a una comunidad biológica de plantas y animales en una interacción compleja con el ambiente físico (Young *et al.*, 1991). En las investigaciones forestales es común que se consideren varias características (representadas por las variables) ya sea como predictoras de otras o como variables-respuesta de diferentes condiciones experimentales. Por otro lado, el análisis multivariado es la parte de la metodología estadística que tiene por objeto estudiar mediciones simultáneas sobre varias variables de un grupo de individuos (o unidades); esta disciplina estadística tiene aplicaciones en los campos de la medicina, la economía y la administración, la agronomía, la biología, la educación, la psicología, los estudios ambientales, la meteorología, la geología, entre otros (Johnson & Wichern, 1982; Harris, 1985; Mardia *et al.*, 1989).

En este escrito se pretende entrelazar algunas de las técnicas multivariadas generadas en la estadística y, posiblemente, aplicadas en otras ramas del conocimiento, con los estudios forestales que involucren dos o más variables. El *objetivo* es incorporar a nuestro entorno académico algunas de estas metodologías que tratan desde una perspectiva más global y amplia —en comparación con la difundida contraparte univariada— los fenómenos bajo estudio.

Aunque todavía es predominante el análisis de datos desde la perspectiva univariada (sobre una variable), pueden mencionarse los trabajos de Caballero (1967), Cortez (1994), López (1995), Sánchez (1993), entre otros, que hicieron uso de alguna técnica multivariada en sus análisis estadísticos. Como ejemplos de algunos trabajos publicados

en el ámbito internacional que emplearon métodos multivariados, se encuentran Zimmerman *et al.* (1994), Olff *et al.* (1994), Clark & Haynes (1992) y Repasky & Schluter (1994).

Para el investigador en recursos naturales, que desee reportar los resultados de sus investigaciones de una manera ordenada, formal y universalmente aceptada a través de la estadística, es de gran utilidad un conocimiento general sobre las ventajas, las desventajas y los supuestos sobre los que están contruidos los métodos estadísticos que emplee, para hacerlo apropiadamente. La estadística permite presentar los resultados de las investigaciones como tendencias provisionales y aproximadas, proveyendo una estimación de los niveles de confiabilidad en las conclusiones que se alcanzan; la estadística multivariada permite además indagar sobre las interrelaciones que ocurren en las variables sobre los individuos. Sin embargo, aún cuando la estadística es un instrumento valioso en el análisis de información experimental u observacional, algunas veces se usa inadecuadamente: en unos casos, para presentar como buenas, investigaciones pobremente concebidas y/o realizadas; en otros, aunque no de manera intencional, se descuidan los supuestos requeridos para la validez de los métodos empleados (Ostle, 1973, p. 19). En esa tónica Infante & Zárate (1997) comentan que “es mejor no usar estadística que usarla mal”. De ahí la importancia de que el investigador comprenda con claridad las técnicas que lo auxilien en sus análisis. En este documento se tiene la intención de presentar algunas de las técnicas multivariadas que tienen aplicaciones en la dasonomía de forma que, tanto sus estructuras conceptuales básicas como su uso en casos particulares, resulten fácilmente comprensibles para los investigadores cuyo interés se enfoque a los recursos naturales, como es el caso de los recursos relacionados con el bosque.

Con respecto a las aplicaciones de la estadística, es importante recordar que ante una *mala planeación* de la investigación o una *toma deficiente de los datos*, no existe método estadístico —por más depurado o fino que sea— que pueda superar los resultados incorrectos que se obtengan; por lo que, cuando así se requiera, la consulta a un

estadístico deberá ser desde las etapas iniciales de la investigación (Cochran & Cox, 1983).

Con base en las consideraciones anteriores, se estructuraron las técnicas multivariadas abordadas en el Capítulo IV. Por cada método, se hace una pequeña introducción, se escriben sus resultados importantes tomados de la literatura y se desarrollan algunos ejemplos con datos relacionados con el bosque, para esbozar alguna situación en que la técnica pueda ser de utilidad; adicionalmente, se escriben los programas en SAS (1987) —Statistical Analysis System, Sistema de Análisis Estadístico— con los que se hicieron los cálculos y que, en un momento dado, pueden ser útiles para tratar situaciones parecidas a las de los ejemplos.

Es conveniente dejar explícito que este trabajo, de ninguna manera puede considerarse un sustituto —ni siquiera mediano— de los libros de métodos multivariados, sino que su propósito sólo es servir de puente inicial entre algunas de las técnicas multivariadas más conocidas y el análisis de datos provenientes de investigaciones forestales, en particular, y de los recursos naturales, en términos más amplios.

II REVISIÓN DE LITERATURA

En este Capítulo, se escriben los conceptos de *dasonomía* y *estadística* así como sus ramas o disciplinas más importantes, para comentar sobre la forma en que interactúan estos cuerpos de conocimiento. Al final, se escribe una sección de álgebra matricial en la que se resumen algunas de sus herramientas tomadas de Searle (1982) y Mardia *et al.* (1989) que se usan en repetidas ocasiones en lo que resta del trabajo.

2.1 LA DASONOMÍA Y SUS DIVISIONES

Dasonomía es una palabra formada por los vocablos griegos *dasos*, bosque y *nomos*, tratado, estudio o tratado del bosque. Siguiendo a Young *et al.* (1991), se puede definir a la dasonomía como el arte y el cuerpo organizado de conocimientos relacionados con la administración (y planificación a largo plazo) de los recursos naturales que se dan sobre, y en asociación con, los terrenos de aptitud forestal, para el beneficio humano. En su concepción más amplia, la dasonomía debe incluir no sólo el estudio de las especies maderables, sino todo aquello que interactúe con el bosque y genere condiciones que alteren —favorezcan o perjudiquen— las acciones culturales que ahí se realicen, así como también las externalidades que se producen, como la reducción de los problemas causados por la erosión eólica en los terrenos agrícolas aledaños; en este sentido, la dasonomía incluye el estudio y el manejo de la fauna silvestre, la preservación de las cuencas hidrográficas, las especies no maderables, la influencia de las masas forestales en la calidad de vida de los ciudadanos, el control de la erosión hídrica, el manejo de los incendios en el bosque, el manejo de las plagas y enfermedades, la sucesión vegetal, etc.

Se entenderá por bosque, a una comunidad biológica de plantas y animales en una interacción compleja con el ambiente físico; las asociaciones vegetales están formadas por árboles y/o arbustos, predominantemente, y en ocasiones involucran a las formas crasas y el estrato herbáceo (Santillán, 1986 y Young *et al.*, 1991).

A la dasonomía se le nombra también muchas veces como “ciencia forestal”; sin embargo, esta acepción es difícil de justificar desde la concepción moderna de que toda ciencia debe involucrar, entre otras características, la descripción y explicación de los fenómenos que estudia y que con base en ciertos principios amplios aunque no totalizadores, puede predecir fenómenos parecidos (Hawking, 1996; Fregoso, 1996) los cuales no necesariamente deben tener aplicación inmediata; por la alta variabilidad de los aspectos biológicos que estudia la dasonomía, es casi imposible hacer predicciones exactas del comportamiento de una variable, y menos de un grupo de ellas. Pero puede afirmarse, de forma menos irreal, que la dasonomía es auxiliada por varias ciencias para cumplir sus objetivos, entre las que se pueden mencionar a la biología, la química, las matemáticas y la estadística.

Por otra parte, la dasonomía puede dividirse para su estudio en tres grandes ramas: la biología forestal, el manejo forestal y los productos forestales (Young *et al.*, 1991), que se resumen a continuación.

Biología Forestal. Estudia los factores genéticos y del ambiente que, ya sea de forma individual o a través de sus interacciones, determinan la forma y funcionamiento de los sistemas vitales, el crecimiento de los árboles y de los sitios forestales. Considera también la influencia de los árboles y las propiedades del sitio sobre otras comunidades y sistemas biológicos. El objetivo de la biología forestal es el estudio de los principios generales que rigen las prácticas silvícolas (Hocker Jr., 1984). En Young *et al.* (1991), se incluyen dentro la biología forestal los aspectos de: fisiología de crecimiento del árbol, suelos forestales, genética forestal, ecología y ecosistemas forestales y entomología y patología forestales.

Manejo forestal. Dentro de esta rama de la dasonomía se incluye una serie de aspectos biológicos y socioeconómicos, que constituyen diferentes disciplinas como: la silvicultura, la dasometría, la administración forestal, la recreación forestal, el manejo de incendios forestales y las interacciones del bosque con la fauna silvestre (Young *et al.*, 1991). A continuación se intenta una demarcación de los primeros tres conceptos.

Silvicultura. (Por sus raíces, cultivo del bosque). Es una especialidad de la dasonomía que estudia los aspectos biológicos del manejo forestal y está sujeta a restricciones económicas y ambientales; incluye los estudios de regeneración, composición de especies, y crecimiento de los bosques. El objetivo de la práctica silvicultural puede enfocarse a destacar la producción de madera, los productos no maderables, el hábitat para la fauna silvestre, el flujo de las corrientes o las características estéticas del bosque (belleza escénica). Por tanto, de forma general, puede decirse que la silvicultura es la manipulación de las masas forestales con la finalidad de alcanzar un conjunto de objetivos establecidos (Young *et al.*, 1991).

Dasometría. (Proviene de las palabras griegas, *dasos*, bosque y *metrón*, medida, significa medición del bosque). Es la parte de la dasonomía relacionada con la medición, cálculo o estimación del volumen, edad e incremento de las masas forestales (Santillán, 1986). Se divide a su vez: en dendrometría, que estudia las medidas del bosque en un tiempo determinado (de forma transversal); y en epidometría que estudia el crecimiento de los árboles y las masas forestales, es decir, mediciones del bosque en forma longitudinal. En la dasometría, los métodos cuantitativos, como los estadísticos, contribuyen de una forma contundente para el análisis de la información.

Administración forestal. El bosque genera productos de alta demanda; la sociedad comienza a administrar el bosque (un recurso) cuando hay productos forestales insuficientes (escasez) que satisfagan su demanda o cuando es demasiado costoso obtener esos productos por algún otro medio (bienes sustitutos). Anteriormente, el concepto de

administración forestal sólo incluía la manipulación biológica de los árboles y sus rodales para lograr ciertos objetivos establecidos; este aspecto proporciona muchas alternativas para la administración forestal, pero ahora sólo se enfatiza en los cursos de silvicultura y ecología. Administrar el bosque, involucra la manipulación de variables internas, como las condiciones climáticas, el suelo y las especies presentes en un determinado rodal, pero también incluye variables externas que hay que considerar, como la legislación, los impuestos y el mercado de los productos forestales. Actualmente, se puede enunciar que la administración forestal consiste en la aplicación de técnicas analíticas para decidir qué alternativas de producción son las que contribuyen mejor a los objetivos de una sociedad determinada. (Young *et al.*, 1991).

Productos forestales. Incluye los aspectos de anatomía y tecnología de la madera, comercialización y demanda y oferta de los productos provenientes del bosque o de sus derivados en el mercado local y nacional (Young *et al.*, 1991).

2.2 LA ESTADÍSTICA Y SUS DIVISIONES

Con la finalidad de demarcar el concepto **Estadística** y su apoyo en la generación y validación de conocimientos en las diferentes áreas aplicadas, puede aceptarse que “es la ciencia involucrada con la planeación, análisis e interpretación de los resultados de una investigación” y que en general trata con: (Ostle, 1973; p. 19)

- 1) Colección y ordenamiento de datos
- 2) Diseño y análisis de experimentos y diagnósticos
- 3) Medición de la variación, tanto de datos experimentales como de datos observacionales
- 4) Estimación de parámetros poblacionales y suministro de varias medidas de exactitud¹ y precisión² de esas estimaciones

¹ La “exactitud” se refiere a qué tan bien, el valor estimado de un parámetro, se acerca a su valor verdadero. (Freese, 1969; p. 6)

- 5) Pruebas de hipótesis respecto a poblaciones
- 6) Estudio de la relación entre dos o más variables (o entre dos o más individuos)

Una característica distintiva de la estadística es el uso del *razonamiento inductivo* en el proceso de obtener los resultados, ya que se requieren hacer conclusiones sobre un grupo grande de datos (la población) disponiendo sólo con una parte representativa de ellos (la muestra). Los resultados finales, de la estimación de un parámetro por ejemplo, se pueden presentar en un rango de valores, determinando el grado de error que se puede cometer al tomarlo como base de una decisión.

De una forma concisa, se puede definir a la estadística como “la ciencia, pura y aplicada, que crea, desarrolla y aplica técnicas de modo que pueda evaluarse la incertidumbre de inferencias inductivas” (Steel & Torrie, 1988).

Algunas divisiones de la estadística se escriben a continuación.

La estadística puede dividirse en dos grandes áreas: la **estadística descriptiva** y la **estadística inferencial**. La primera se refiere a un grupo de técnicas para la recolección, organización, descripción y presentación de grupos de datos y cálculos de cantidades “representativas” de dicho grupo, con el objeto de que los aspectos relevantes del mismo sean rápida y fácilmente aprehensibles. La estadística inferencial se refiere a las técnicas que permiten cuantificar, mediante los principios probabilísticos, la incertidumbre que es inherente de la información obtenida del análisis de datos y que, en última instancia, nos ayuda a tomar decisiones con base en los resultados. La idea más generalizada de la estadística en la mayoría de la gente, se relaciona con la estadística descriptiva; sin embargo, de acuerdo a la concepción actual de esta ciencia, el recuento y la numeración de grandes cantidades de información, aunque constituye uno de los aspectos que abarca la estadística, no es el más importante (Infante & Zárate, 1997).

² La “precisión” se refiere al agrupamiento de los valores muestrales alrededor de su propia media. (Freese, 1969; p. 6)

Por el número de variables consideradas en el análisis, la estadística se divide en **univariada** y **multivariada**. La *estadística univariada* está relacionada con el análisis de un fenómeno tomando en consideración una sola variable. Por ejemplo, se puede estar interesado en conocer la altura media de los árboles de una área sujeta a contaminación y compararla con la altura media de los árboles en ambiente silvestre con la finalidad de cuantificar los efectos de un ambiente contaminado sobre la vegetación; atención, sólo con la información que provea la altura de los árboles (i.e., mediciones sobre una variable). Por otra parte, la *estadística multivariada* está involucrada con el análisis de información de observaciones que se han registrado sobre varias dimensiones de los individuos (Mardia *et al.*, 1989); en vez de estudiar el efecto de la contaminación en los árboles sólo con las alturas, esta evaluación se puede hacer con base en la información recabada de altura, diámetro del fuste a cierta altura, longitud de raíces, proyección de la copa y coloración de las hojas, por ejemplo. La idea es que usando más de una variable, se cuenta con más elementos para hacer una conclusión del fenómeno que se estudia.

Como ejemplos de datos que contienen mediciones de dos o más variables, medidas cada una, sobre un conjunto de individuos, se pueden mencionar: (véanse los Cuadros 3.2.1, 3.2.2, 3.2.3, 3.2.4 y 4.5.1)

- Mediciones florales de diferentes especies de *Iris*
- Datos de pájaros sobre dos tipos de sitios en el bosque
- Número de árboles de la regeneración natural en una selva mediana subperennifolia
- Mediciones de la coraza de un grupo de tortugas
- Calificaciones obtenidas en diferentes materias por un grupo de estudiantes

En estos cinco ejemplos, las variables son: las mediciones florales; la altura del dosel, la distancia del nido al contorno y el perímetro de la masa de vegetación; los estratos; el largo y el ancho de los caparazones; las calificaciones. Asimismo, los individuos son: las flores, los sitios, las etapas sucesionales; las tortugas y los estudiantes, respectivamente.

Como sinónimos de variable se puede usar: atributo, característica, descripción, medición o respuesta; así también, algunos sinónimos de “individuo” son: observación, objeto, parcela, lectura y unidad (*ibid.*).

2.3 RELACIÓN DE LA ESTADÍSTICA Y LA DASONOMÍA

Las siguientes definiciones sobre aplicaciones particulares de la estadística se presentan para entender la interacción de la dasonomía y la estadística.

Bioestadística. Las herramientas de la estadística se emplean en muchos campos —los negocios, la educación, la psicología, la agricultura y la economía por mencionar sólo unos cuantos. Cuando los datos que se están analizando se obtienen de las ciencias biológicas y de la medicina, se usa el término bioestadística para distinguir a esta aplicación particular de las herramientas y conceptos estadísticos (Daniel, 1983).

Estadística Ecológica. Incluye todos los métodos estadísticos que son de valor en el estudio de las relaciones entre animales y plantas y su ambiente, pero más particularmente aquellos métodos estadísticos desarrollados especialmente para su uso en esta área. Puede abarcar tanto el estudio de especies individuales como la ecología de poblaciones así como experimentos en ecología. (Manly, 1994, p.307)

Biometría Forestal. Se comenta de la biometría en general, que la evolución de la práctica estadística ha sido guiada por las necesidades de los estudiosos de fenómenos biológicos para resolver problemas y contestarse preguntas que resultan de sus aplicaciones cotidianas en la disciplina en que trabajan. Particularmente esto es cierto en dasonomía, donde se requiere de información cuantitativa confiable para el manejo de los bosques. La biometría forestal incluye los procedimientos estadísticos usados para cuantificar el volumen y forma de los árboles, la calidad de sitio (curvas altura-edad), la

densidad del rodal, y las relaciones crecimiento-cosecha. (Burkhardt & Gregoire, 1994; p. 378)

Considerando las tres definiciones anteriores, y tomando en cuenta que los estudios en dasonomía —que se dedica al estudio del bosque— tienen una importante componente biológica y que pueden integrar al mismo tiempo los aspectos de conservación (ecológicas) y de aprovechamiento, se puede estar de acuerdo en que, el punto donde se intersectan la dasonomía y la estadística es, de manera general, la bioestadística —puesto que los bosques y los recursos asociados son entes biológicos— y muy particularmente lo hacen en la biometría forestal. La cuestión ecológica, se considera implícitamente como parte de la biología y del concepto de manejo forestal que incluye aspectos físicos, ecológicos, socio-económicos, entre otros.

2.4 ELEMENTOS DE ÁLGEBRA MATRICIAL

Para poder resumir algunas herramientas matriciales que se usarán frecuentemente en este escrito, se asume que se conocen, de forma general, qué es una matriz y las siguientes operaciones: suma, producto, transpuestas, determinantes e inversas. Se tiene como referencias el libro de Searle (1982) y el apéndice de álgebra matricial que presentan Mardia *et al.* (1989).

2.4.1 Combinación Lineal

Sean $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$ un conjunto de p vectores del mismo orden, y $a_1, a_2, \dots, a_p$, p valores escalares. Entonces, la combinación lineal del conjunto de p vectores se puede definir como:

$$a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \dots + a_p\mathbf{x}_p = \sum_{j=1}^p a_j\mathbf{x}_j = \mathbf{X}\mathbf{a} \quad (2.4-1)$$

donde $\mathbf{X} = [\mathbf{x}_{(1)} \quad \mathbf{x}_{(2)} \quad \cdots \quad \mathbf{x}_{(p)}]$ y $\mathbf{a} = [a_1 \quad a_2 \quad \cdots \quad a_p]'$. Observe que $\mathbf{Xa}$ es un vector columna que presenta las combinaciones lineales de las columnas de $\mathbf{X}$. De forma análoga, dada una matriz $\mathbf{X}$ y un vector $\mathbf{b}$, si $\mathbf{b}'\mathbf{X}$ existe, será un vector hilera de las combinaciones lineales de las hileras de $\mathbf{X}$.

La combinación lineal $\mathbf{Xa}$ es una combinación lineal estandarizada si se cumple que $\mathbf{a}'\mathbf{a} = 1$.

2.4.2 Inversa de una Matriz

La inversa de la matriz cuadrada $\mathbf{A}$ es la matriz única $\mathbf{A}^{-1}$ que satisface:

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}. \quad (2.4-2)$$

La inversa existe si, y sólo si, $\mathbf{A}$ es una matriz no-singular; en otras palabras, $\mathbf{A}^{-1}$ existe si, y sólo si, el determinante de $\mathbf{A}$ es diferente de cero ($|\mathbf{A}| \neq 0$).

El (i, j) -ésimo elemento de $\mathbf{A}^{-1}$ se denota por a^{ij} , quedando como:

$$\mathbf{A}^{-1} = \begin{bmatrix} a^{11} & \cdots & a^{1n} \\ \vdots & \ddots & \vdots \\ a^{n1} & \cdots & a^{nn} \end{bmatrix}.$$

Para $\mathbf{A}$ particionada, se escribe:

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix}.$$

Propiedades

$$(I) \quad \mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} \text{adj}(\mathbf{A}),$$

donde

$$|\mathbf{A}| = \sum_{j=1}^n a_{ij}(-1)^{i+j} |\mathbf{M}_{ij}| \quad \text{para toda hilera } i,$$

$$= \sum_{i=1}^n a_{ij} (-1)^{i+j} |M_{ij}| \quad \text{para toda columna } j,$$

es el determinante de la matriz A con $|M_{ij}|$ igual al valor del determinante que se obtiene después de eliminar la i -ésima hilera y la j -ésima columna de A , y se denomina como el menor de a_{ij} . El cofactor de a_{ij} es $(-1)^{i+j}$ veces el menor de a_{ij} ; y $\text{adj}(A)$ es la transpuesta de la matriz de cofactores, conocida como la matriz adjunta de A .

$$(II) (cA)^{-1} = c^{-1}A^{-1}$$

$$(III) (AB)^{-1} = B^{-1}A^{-1}$$

$$(IV) \text{ La solución única de } Ax = b \text{ es } x = A^{-1}b$$

$$(V) \text{ Si existen todas las inversas necesarias, entonces para } A_{p \times p}, B_{p \times n}, C_{n \times n} \text{ y } D_{n \times p},$$

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

$$(VI) \text{ Si existen todas las inversas necesarias, entonces para una matriz } A \text{ particionada, los elementos de } A^{-1} \text{ son}$$

$$A^{11} = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}, \quad A^{22} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1},$$

$$A^{12} = -A^{11}A_{12}A_{22}^{-1}, \quad A^{21} = -A_{22}^{-1}A_{21}A^{11}$$

En la definición de A^{-1} se asume que A es una matriz cuadrada con determinante diferente de cero (i.e., se asume que A es una matriz no-singular). Es claro que la inversa de A^{-1} es A .

2.4.3 Productos Kronecker

Definición 2.4.3.1. Sea $A = (a_{ij})$ una matriz de dimensión $(m \times n)$ y $B = (b_{kl})$ una matriz de orden $(p \times q)$. Entonces el producto Kronecker de A y B se define como:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}_{mp \times nq}$$

Definición 2.4.3.2. Si $\mathbf{X}$ es una matriz de dimensión $(n \times p)$ denotar con $\mathbf{X}^V$ al vector de orden np que se obtiene de “vectorizar” a $\mathbf{X}$; es decir, de colocar las columnas de $\mathbf{X}$ una debajo de otra de la primera a la última, de forma que:

$$\mathbf{X}^V = \begin{bmatrix} \mathbf{x}_{(1)} \\ \mathbf{x}_{(2)} \\ \vdots \\ \mathbf{x}_{(p)} \end{bmatrix},$$

donde, $\mathbf{x}_{(j)}$ es la j -ésima columna de $\mathbf{X}$, $j = 1, \dots, p$. De estas definiciones se dan las siguientes propiedades:

- (I) $\alpha(\mathbf{A} \otimes \mathbf{B}) = (\alpha\mathbf{A}) \otimes \mathbf{B} = \mathbf{A} \otimes (\alpha\mathbf{B})$ para todo escalar α , y de ahí puede escribirse sin ambigüedad como $\alpha\mathbf{A} \otimes \mathbf{B}$.
- (II) $\mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{C}) = (\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{C}$. Por lo que puede escribirse como $\mathbf{A} \otimes \mathbf{B} \otimes \mathbf{C}$.
- (III) $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$.
- (IV) $(\mathbf{A} \otimes \mathbf{B})(\mathbf{F} \otimes \mathbf{G}) = (\mathbf{A}\mathbf{F}) \otimes (\mathbf{B}\mathbf{G})$. Aquí son necesarios los paréntesis.
- (V) $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$ para matrices $\mathbf{A}$ y $\mathbf{B}$ no-singulares.
- (VI) $(\mathbf{A} + \mathbf{B}) \otimes \mathbf{C} = \mathbf{A} \otimes \mathbf{C} + \mathbf{B} \otimes \mathbf{C}$.
- (VII) $\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C}$.
- (VIII) $(\mathbf{A}\mathbf{X}\mathbf{B})^V = (\mathbf{B}' \otimes \mathbf{A})\mathbf{X}^V$.

2.4.4 Valores Característicos y Vectores Característicos

Si $\mathbf{A}_{p \times p}$ es cualquier matriz cuadrada, entonces:

$$q(\lambda) = |\mathbf{A} - \lambda\mathbf{I}| \quad (2.4-3)$$

es un polinomio de orden p en λ . Las p raíces de $q(\lambda)$, $\lambda_1, \dots, \lambda_p$, de las cuales, algunas pueden ser cero o números complejos, son conocidos como los *valores característicos* de la matriz A . Estas raíces son conocidas también como *raíces latentes*, *eigenvalores*, *valores propios*, *raíces características*, *autovalores* o *raíces- λ* . En este trabajo se usa la acepción de valores característicos. Algunas de las λ_i serán iguales si $q(\lambda)$ tiene raíces múltiples.

A cada valor característico λ_i corresponde un vector u_i diferente de cero (para esto es necesario que $|A - \lambda I| = 0$), que satisface

$$Au_i = \lambda_i u_i \quad \text{para } i = 1, \dots, p; \quad (2.4-4)$$

a los vectores $u_1, u_2, \dots, u_p$, se les conoce como *vectores característicos*, *vectores latentes*, *eigenvectores*, *vectores propios* o *autovectores*. Un eigenvector u será un vector característico estandarizado si

$$u' u = 1. \quad (2.4-5)$$

Propiedad 2.4.4.1. Todos los valores característicos de una matriz simétrica $A_{p \times p}$ son reales.

2.4.5 Teorema de la Descomposición Espectral

(Mardia *et al.*, 1989; p. 469). Se puede demostrar que $AA' = \sum_j a_{(j)} a'_{(j)}$, donde $a_{(j)}$ es la j -ésima columna de A . Si Γ es una matriz ortogonal, i.e., $I = \Gamma\Gamma'$, entonces se tiene que:

$$I = \sum_i \gamma_{(i)} \gamma'_{(i)}. \quad (2.4-6)$$

Ahora bien, si los $\gamma_{(i)}$ son vectores característicos de $A = A'$, premultiplicando por A en ambos lados de (2.4-6), se tiene que:

$$\begin{aligned} \mathbf{A} &= \mathbf{A} \sum_i \gamma_{(i)} \gamma'_{(i)} = \sum_i \mathbf{A} \gamma_{(i)} \gamma'_{(i)} = \sum_i \lambda_i \gamma_{(i)} \gamma'_{(i)} \\ &= \mathbf{\Gamma} \mathbf{\Lambda} \mathbf{\Gamma}' \end{aligned} \quad (2.4-7)$$

Esta es la descomposición espectral de $\mathbf{A}$, donde $\mathbf{\Lambda}$ es una matriz que en su diagonal principal contiene los valores característicos de $\mathbf{A}$, y $\mathbf{\Gamma}$ es la matriz ortogonal que contiene en sus columnas los vectores característicos asociados de $\mathbf{A}$. (La prueba de este teorema está en Mardia *et al.*, 1989, p. 469)

Corolario 2.4.5.1. Si $\mathbf{A}$ es una matriz simétrica no-singular, entonces para cualquier entero n ,

$$\mathbf{\Lambda}^n = \text{diag}(\lambda_i^n) \quad \text{y} \quad \mathbf{A} = \mathbf{\Gamma} \mathbf{\Lambda}^n \mathbf{\Gamma}'. \quad (2.4-8)$$

Si todos los valores característicos de $\mathbf{A}$ son positivos entonces se pueden definir las potencias racionales:

$$\mathbf{A}^{r/s} = \mathbf{\Gamma} \mathbf{\Lambda}^{r/s} \mathbf{\Gamma}', \quad \text{donde} \quad \mathbf{\Lambda}^{r/s} = \text{diag}(\lambda_i^{r/s}), \quad (2.4-9)$$

donde $s > 0$ y r son números enteros. Si alguno de los valores característicos de $\mathbf{A}$ son cero, entonces (2.4-8) y (2.4-9) se cumplen si se tiene la restricción de que los exponentes sean no-negativos (La prueba está en Mardia *et al.*, 1989, p.470)

Dos casos especiales de (2.4-9) son:

$$\mathbf{A}^{1/2} = \mathbf{\Gamma} \mathbf{\Lambda}^{1/2} \mathbf{\Gamma}', \quad \text{donde} \quad \mathbf{\Lambda}^{1/2} = \text{diag}(\lambda_i^{1/2}), \quad (2.4-10)$$

cuando $\lambda_i \geq 0, \forall i$ y

$$\mathbf{A}^{-1/2} = \mathbf{\Gamma} \mathbf{\Lambda}^{-1/2} \mathbf{\Gamma}', \quad \text{donde} \quad \mathbf{\Lambda}^{-1/2} = \text{diag}(\lambda_i^{-1/2}), \quad (2.4-11)$$

cuando $\lambda_i > 0, \forall i$. La descomposición (2.4-10) es conocida como la *descomposición raíz cuadrada simétrica* de $\mathbf{A}$.

Corolario 2.4.5.2. El rango de $\mathbf{A}$, denotado por $r(\mathbf{A})$, es igual al número de valores característicos diferentes de cero. (ver la prueba en Mardia *et al.*, 1989, p. 471)

Observaciones (1) El Teorema de la descomposición espectral muestra que una matriz simétrica A es unívocamente determinada por sus valores y vectores característicos, o más específicamente por sus diferentes valores característicos y sus correspondientes eigenespacios.

(2) Como $A^{1/2}$ tiene los mismos vectores característicos que A y tiene valores característicos dados como funciones de los valores característicos de A , entonces la matriz raíz cuadrada simétrica es unívocamente definida.

(3) Si los λ_i son todos distintos y están escritos en orden decreciente, entonces Γ es unívocamente determinado.

(4) Si $\lambda_{k+1} = \dots = \lambda_p = 0$ entonces (2.4-7) puede escribirse más compactamente como:

$$A = \Gamma_1 \Lambda_1 \Gamma_1' = \sum_{i=1}^k \lambda_i \gamma_{(i)} \gamma_{(i)}',$$

donde $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ y $\Gamma_1 = (\gamma_{(1)}, \dots, \gamma_{(k)})$.

(5) Una matriz simétrica A tiene rango 1 si y sólo si

$$A = \mathbf{x} \mathbf{x}'$$

para alguna $\mathbf{x}$. Entonces el único valor característico diferentes-de-cero de A está dado por:

$$\text{tr}(A) = \text{tr} \mathbf{x} \mathbf{x}' = \mathbf{x}' \mathbf{x}$$

y el correspondiente eigen-espacio es generado por $\mathbf{x}$. La traza de la matriz A se denota como $\text{tr}(A)$ y se define como la suma de los elementos de su diagonal principal.

(6) Como $J = \mathbf{1} \mathbf{1}'$ tiene rango 1 con valor característico p y su correspondiente vector característico $\mathbf{1}$, se tiene que la matriz de equicorrelación $E = (1 - \rho)I + \rho J$ tiene valores característicos $\lambda_1 = 1 + (p - 1)\rho$ y $\lambda_2 = \dots = \lambda_p = 1 - \rho$ y los mismos vectores característicos de J . Para los vectores característicos $\gamma_{(1)}, \dots, \gamma_{(p)}$ se puede seleccionar cualquier conjunto de vectores estandarizados ortogonales a $\mathbf{1}$ y algún otro.

(7) Si A es simétrica e idempotente (es decir, $A = A'$ y $A^2 = A$), entonces $\lambda_i = 0$ ó 1 para toda i , ya que $A^2 = A \Rightarrow \Lambda = \Lambda^2$.

(8) Si $A = A'$ y $A^2 = A$ entonces $r(A) = \text{tr}(A)$.

(9) Como ejemplo, considerar

$$\mathbf{A} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \quad (2.4-12)$$

Usando (2.4-3), los valores característicos de $\mathbf{A}$ son las soluciones de:

$$\begin{vmatrix} 1-\lambda & \rho \\ \rho & 1-\lambda \end{vmatrix} = 0,$$

conocidos como $\lambda_1 = 1 + \rho$ y $\lambda_2 = 1 - \rho$. Entonces:

$$\Lambda = \text{diag}(1 + \rho, 1 - \rho) = \begin{pmatrix} 1 + \rho & 0 \\ 0 & 1 - \rho \end{pmatrix}. \quad (2.4-13)$$

Para $\rho \neq 0$, el vector característico asociado a $\lambda_1 = 1 + \rho$ de (2.4-4) es

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (1 + \rho) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix},$$

que conduce a $x_1 = x_2$, por lo tanto el primer vector característico estandarizado es

$$\gamma_{(1)} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}.$$

Análogamente, el vector característico correspondiente a $\lambda_2 = 1 - \rho$ es

$$\gamma_{(2)} = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}.$$

De ahí que,

$$\Gamma = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}. \quad (2.4-14)$$

(10) La fórmula (2.4-10) sugiere un método para calcular la raíz cuadrada simétrica de una matriz. Por ejemplo, para la matriz en (2.4-12) con $\rho^2 < 1$, se observa que usando Λ y Γ de (2.4-10) y la relación $\Gamma \Lambda \Gamma' = \Lambda$ (derivada de 2.4-7) se tiene

$$\mathbf{A}^{1/2} = \Gamma \Lambda^{1/2} \Gamma = \begin{pmatrix} a & b \\ b & a \end{pmatrix},$$

donde $2a = (1 + \rho)^{1/2} + (1 - \rho)^{1/2}$, $2b = (1 + \rho)^{1/2} - (1 - \rho)^{1/2}$.

(11) Para matrices generales $A_{n \times p}$, se puede usar el teorema de la descomposición espectral para derivar el siguiente resultado.

2.4.6 Teorema de la Descomposición del Valor Singular

Si A es una matriz de dimensión $(n \times p)$ de rango r , entonces A puede escribirse como:

$$A = ULV' \quad (2.4-15)$$

donde $U_{n \times r}$ y $V_{p \times r}$ son matrices ortonormales por columnas ($U'U = V'V = I_r$) y L es una matriz diagonal con elementos positivos.

Las columnas de la matriz U son vectores característicos de AA' y las columnas de la matriz V son vectores característicos de $A'A$. De la Propiedad 2.4.4.1 se tiene que los valores característicos de AA' y $A'A$ son iguales.

III MATERIALES Y MÉTODOS

3.1 PRESENTACIÓN DE LOS MÉTODOS MULTIVARIADOS

En cada uno de los seis subcapítulos del Capítulo 4 se hace una introducción al tema; se presentan, con base en la investigación bibliográfica, los conceptos y resultados importantes; dentro de la estructura de cada tema, se desarrollan algunos ejemplos usando datos dasonómicos reportados en la literatura, con la finalidad de apreciar aplicaciones específicas del método en cuestión. También se incluyen los programas de SAS usados para resolver los ejemplos con la idea de que puedan ser útiles para el análisis de información análoga al de los casos que se describen.

En muchas ocasiones los programas se escribieron, con fines didácticos, usando el subprograma IML (Interactive Matrix Language) de SAS. Es probable que resulte más cómodo usar las funciones preconstruidas de SAS (como MANOVA y PRINCOMP) pero es sugerible usar los programas en IML cuando se desee entender un poco más sobre la forma en que opera el método o la técnica que se esté abordando.

En la presentación del material se tuvo como referencia principal a Mardia *et al.* (1989); otras fuentes de soporte son Morrison (1981) y Johnson & Wichern (1982). Se aclara que, frecuentemente, cuando se mencione alguna propiedad, se citará el teorema o bien la página de la referencia, en que se puede localizar para facilitar su cotejamiento. Como se ha mencionado, este trabajo hace énfasis en la cuestión aplicada y los ejemplos

que se presentan tratan de reforzar dicho enfoque; sin embargo, en algunas ocasiones, se han escrito demostraciones que ya sea que son fáciles de desarrollar y entender o que por su relativa importancia, se tomó de la literatura para que pueden ayudar en darle alguna continuidad a la temática. Para el propósito de este trabajo, es suficiente con prestar atención a los resultados directos y a las aplicaciones que de ellos se pueden hacer más que, sin dar a entender que es menos importante, entrar en detalle con la formalidad algebraica.

Muchas de las fórmulas se escribieron usando la notación matricial, pero sólo es necesario conocer los conceptos básicos de esta notación, para comprender con claridad la exposición. En el subcapítulo de Generalidades, se destina una sección corta para tratar los conceptos matriciales que se usaron en repetidas ocasiones en la tesis. Sin embargo, si se desea revisar con más profundidad algún concepto matricial, se puede recurrir a Searle (1982) y a Mardia *et al.* (1989, pp. 452-485) o Johnson & Wichern (1982, pp. 37-90) que presentan un apéndice de álgebra de matrices.

3.2 DATOS SOBRE VARIAS DIMENSIONES

A continuación se presentan algunos conjuntos de datos multivariados que se usan a lo largo del trabajo, al igual que otros, para desarrollar los ejemplos numéricos en los capítulos subsiguientes. En el Cuadro 3.2.1, se cuentan con las mediciones de tres especies de *Iris* en las siguientes variables:

- x_1 : longitud de sépalos,
- x_2 : ancho de sépalos,
- x_3 : longitud de pétalos,
- x_4 : ancho de pétalos.

Cuadro 3.2.1 Mediciones de longitud y ancho del sépalo y del pétalo de tres especies de *Iris* spp.

<i>Iris setosa</i>				<i>Iris versicolour</i>				<i>Iris virginica</i>			
x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4	x_1	x_2	x_3	x_4
5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4	6.3	3.3	6.0	2.5
4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5	5.8	2.7	5.1	1.9
4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5	7.1	3.0	5.9	2.1
4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3	6.3	2.9	5.6	1.8
5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5	6.5	3.0	5.8	2.2
5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3	7.6	3.0	6.6	2.1
4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6	4.9	2.5	4.5	1.7
5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0	7.3	2.9	6.3	1.8
4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3	6.7	2.5	5.8	1.8
4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4	7.2	3.6	6.1	2.5
5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0	6.5	3.2	5.1	2.0
4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5	6.4	2.7	5.3	1.9
4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0	6.8	3.0	5.5	2.1
4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4	5.7	2.5	5.0	2.0
5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3	5.8	2.8	5.1	2.4
5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4	6.4	3.2	5.3	2.3
5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5	6.5	3.0	5.5	1.8
5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0	7.7	3.8	6.7	2.2
5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5	7.7	2.6	6.9	2.3
5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1	6.0	2.2	5.0	1.5
5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8	6.9	3.2	5.7	2.3
5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3	5.6	2.8	4.9	2.0
4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5	7.7	2.8	6.7	2.0
5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2	6.3	2.7	4.9	1.8
4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3	6.7	3.3	5.7	2.1
5.0	3.0	1.6	0.2	6.6	3.0	4.4	1.4	7.2	3.2	6.0	1.8
5.0	3.4	1.6	0.4	6.8	2.8	4.8	1.4	6.2	2.8	4.8	1.8
5.2	3.5	1.5	0.2	6.7	3.0	5.0	1.7	6.1	3.0	4.9	1.8
5.2	3.4	1.4	0.2	6.0	2.9	4.5	1.5	6.4	2.8	5.6	2.1
4.7	3.2	1.6	0.2	5.7	2.6	3.5	1.0	7.2	3.0	5.8	1.6
4.8	3.1	1.6	0.2	5.5	2.4	3.8	1.1	7.4	2.8	6.1	1.9
5.4	3.4	1.5	0.4	5.5	2.4	3.7	1.0	7.9	3.8	6.4	2.0
5.2	4.1	1.5	0.1	5.8	2.7	3.9	1.2	6.4	2.8	5.6	2.2
5.5	4.2	1.4	0.2	6.0	2.7	5.1	1.6	6.3	2.8	5.1	1.5
4.9	3.1	1.5	0.2	5.4	3.0	4.5	1.5	6.1	2.6	5.6	1.4
5.0	3.2	1.2	0.2	6.0	3.4	4.5	1.6	7.7	3.0	6.1	2.3
5.5	3.5	1.3	0.2	6.7	3.1	4.7	1.5	6.3	3.4	5.6	2.4
4.9	3.6	1.4	0.1	6.3	2.3	4.4	1.3	6.4	3.1	5.5	1.8
4.4	3.0	1.3	0.2	5.6	3.0	4.1	1.3	6.0	3.0	4.8	1.8
5.1	3.4	1.5	0.2	5.5	2.5	4.0	1.3	6.9	3.1	5.4	2.1
5.0	3.5	1.3	0.3	5.5	2.6	4.4	1.2	6.7	3.1	5.6	2.4
4.5	2.3	1.3	0.3	6.1	3.0	4.6	1.4	6.9	3.1	5.1	2.3
4.4	3.2	1.3	0.2	5.8	2.6	4.0	1.2	5.8	2.7	5.1	1.9
5.0	3.5	1.6	0.6	5.0	2.3	3.3	1.0	6.8	3.2	5.9	2.3
5.1	3.8	1.9	0.4	5.6	2.7	4.2	1.3	6.7	3.3	5.7	2.5
4.8	3.0	1.4	0.3	5.7	3.0	4.2	1.2	6.7	3.0	5.2	2.3
5.1	3.8	1.6	0.2	5.7	2.9	4.2	1.3	6.3	2.5	5.0	1.9
4.6	3.2	1.4	0.2	6.2	2.9	4.3	1.3	6.5	3.0	5.2	2.0
5.3	3.7	1.5	0.2	5.1	2.5	3.0	1.1	6.2	3.4	5.4	2.3
5.0	3.3	1.4	0.2	5.7	2.8	4.1	1.3	5.9	3.0	5.1	1.8

Fuente: Fisher (1936), reportados por Mardia *et al.* (1989) y Johnson & Wichern (1982)

Cuadro 3.2.2 Mediciones de variables sobre los sitios de anidamiento de pájaros de helecho y sitios ubicados aleatoriamente en la misma área

En los sitios de anidamiento			En los sitios disponibles		
x_1	x_2	x_3	x_1	x_2	x_3
1.20	14.0	8.90	0.47	13.5	3.17
0.58	25.0	4.34	0.62	8.0	3.23
0.74	14.0	2.30	0.75	19.0	2.44
0.70	12.0	5.16	0.52	5.0	1.56
1.36	14.5	2.92	0.73	8.0	2.28
0.78	17.0	3.30	0.62	16.0	3.16
0.45	15.0	3.17	0.60	17.0	2.78
0.78	15.0	4.81	0.26	4.5	3.07
0.63	16.0	2.40	0.46	15.0	3.84
0.75	12.0	3.74	0.28	12.0	3.33
0.55	12.0	4.86	0.53	11.0	2.80
0.45	20.0	2.88	0.42	17.0	2.92
1.56	16.0	4.90	0.47	20.0	4.40
0.85	23.0	4.65	0.50	13.0	3.86
0.58	12.0	4.02	0.54	16.0	3.48
0.75	13.0	4.54	0.56	18.0	2.36
0.55	18.0	3.22	0.32	7.0	3.08
0.56	18.0	3.08	0.62	16.0	5.07
0.57	19.5	4.43	0.39	15.0	2.02
0.41	16.0	3.48	0.56	8.0	1.81
0.65	18.0	4.50	0.27	9.0	2.05
0.78	14.0	2.96	0.42	11.0	1.74
0.64	18.0	5.25	0.70	13.0	2.85
0.71	18.0	3.07	0.26	14.0	3.64
			0.34	9.5	2.40

Fuente: Manly, B.F.J. (1994).

(x_1 : altura del dosel (metros), x_2 : distancia del eje exterior del nido a la superficie exterior más cercana del grupo de vegetación en que se sitúa el nido (centímetros), x_3 : el perímetro del grupo de vegetación (metros))

Cuadro 3.2.3 Número de individuos/ha de la regeneración natural en una selva mediana subperennifolia en Escárcega, Campeche, México.

Etapas sucesional	Estratos de alturas (en cm)			
	0-30	31-100	100-300	>300
I	29,050	7,475	9,450	11,350
II	16,025	5,575	5,025	5,250
III	43,250	8,600	4,225	2,225
IV	83,175	9,275	2,075	2,000

Fuente: Centeno (1989) reportado por López (1995, p. 7)

Cuadro 3.2.4 Mediciones del caparazón (mm) de 48 tortugas pintas (24 hembras y 24 machos)

Hembras			Machos		
Largo (x_1)	Ancho (x_2)	Alto (x_3)	Largo (x_1)	Ancho (x_2)	Alto (x_3)
98	81	38	93	74	37
103	84	38	94	78	35
103	86	42	96	80	35
105	86	42	101	84	39
109	88	44	102	85	38
123	92	50	103	81	37
123	95	46	104	83	39
133	99	51	106	83	39
133	102	51	107	82	38
133	102	51	112	89	40
134	100	48	113	88	40
136	102	49	114	86	40
138	98	51	116	90	43
138	99	51	117	90	41
141	105	53	117	91	41
147	108	57	119	93	41
149	107	55	120	89	40
153	107	56	120	93	44
155	115	63	121	95	42
155	117	60	125	93	45
158	115	62	127	96	45
159	118	63	128	95	45
162	124	61	131	95	46
177	132	67	135	106	47

Fuente: Datos tomados de Johnson & Wichern (1982, p. 281)

Cuadro 3.2.5 Pesos (en centigramos) de depósitos de corcho de 28 árboles en los cuatro puntos cardinales

N	E	S	W	N	E	S	W
72	66	76	77	91	79	100	75
60	53	66	63	56	68	47	50
56	57	64	58	79	65	70	61
41	29	36	38	81	80	68	58
32	32	35	36	78	55	67	60
30	35	34	26	46	38	37	38
39	39	31	27	39	35	34	37
42	43	31	25	32	30	30	32
37	40	31	25	60	50	67	54
33	29	27	36	35	37	48	39
32	30	34	28	39	36	39	31
63	45	74	63	50	34	37	40
54	46	60	52	43	37	39	50
47	51	52	43	48	54	57	43

Fuente: Rao (1948; p. 62)

IV LAS TÉCNICAS MULTIVARIADAS Y SUS APLICACIONES

4.1 GENERALIDADES

4.1.1 El Contexto Multivariado

La estadística multivariada está involucrada con el análisis de datos colectados, simultáneamente, sobre varias dimensiones del mismo individuo. Como en el caso univariado, se asume que se colecta una muestra aleatoria de observaciones con varias componentes provenientes de diferentes individuos u otras unidades muestrales independientes. Sin embargo, la fuente común de cada observación individual generará dependencia o correlación entre dimensiones y esta es una característica distintiva entre las técnicas y datos multivariados y los prototipos univariados (Morrison, 1981).

Si sobre cada uno de n individuos de una muestra se consideran las mediciones sobre p diferentes variables, los datos registrados podrán arreglarse de forma rectangular en una **matriz de datos** de manera que las hileras contengan a los individuos y las columnas a las mediciones de las respectivas variables. Con los datos de *Iris spp.* de el Cuadro 3.3.1 se pueden conformar tres matrices de datos, una para cada especie; las tres matrices de datos resultantes serán de dimensión 50×4 , ya que para cada especie, se tienen las mediciones de cuatro variables sobre 50 individuos. Las variables de datos multivariados no tienen necesariamente que ser del mismo tipo, es decir, pueden incluirse variables continuas (como la altura o el diámetro), discretas (como la edad del arbolado o el número de ramas o de raíces de una plántula) y/o dicotómicas (p. ej., macho o hembra, vivo o muerto, germinó o no germinó) [Mardia *et al.*, 1989, p. 2].

4.1.2 La Matriz de Datos

En general, si se toman p variables, $X_1, X_2, \dots, X_p$, sobre cada uno de n individuos, $o_1, o_2, \dots, o_n$, provenientes de una muestra aleatoria, se obtendrá un conjunto de datos con np elementos. Estos datos se pueden arreglar en una *matriz de datos* $\mathbf{X}$ de dimensión $(n \times p)$, en la que cada hilera corresponde a un individuo y cada columna, a una variable (*op. cit.*).

Normalmente se usará la notación $\mathbf{X}_{n \times p}$ [ó $\mathbf{X}(n \times p)$] para indicar la dimensión de la matriz de datos, o sencillamente se escribirá $\mathbf{X}$, cuando en el contexto sea obvio su orden o dimensión; este convenio en la notación del orden, será válido también para cualquier otra matriz en los siguientes capítulos.

La matriz de datos $\mathbf{X}_{n \times p}$ que contiene observaciones de p variables tomadas sobre n individuos, puede escribirse de manera ampliada como:

$$\mathbf{X}_{n \times p} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} \quad (4.1-1)$$

$$i = 1, 2, \dots, n, \quad j = 1, 2, \dots, p$$

en donde x_{ij} denota la observación de la j -ésima variable sobre el i -ésimo individuo.

La matriz de datos $\mathbf{X}$ también se puede expresar como sigue. Sea $\mathbf{X} = \{x_{ij}\}$; usando la notación de Mardia *et al.* (1989, pp. 8 y 453) donde $\mathbf{x}_i$ representa la i -ésima hilera de $\mathbf{X}$ escrita como columna y $\mathbf{x}_{(j)}$, la j -ésima columna de $\mathbf{X}$, se tiene que:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_i \\ \vdots \\ \mathbf{x}'_n \end{bmatrix} = [\mathbf{x}_{(1)} \quad \mathbf{x}_{(2)} \quad \cdots \quad \mathbf{x}_{(j)} \quad \cdots \quad \mathbf{x}_{(p)}], \quad (4.1-2)$$

donde

$$\mathbf{x}'_i = [x_{i1} \quad x_{i2} \quad \cdots \quad x_{ip}] \quad \text{y} \quad \mathbf{x}_{(j)} = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}.$$

EJEMPLO 4.1.1 Con los datos de el Cuadro 3.2.4 se pueden conformar dos matrices de datos, una para las tortugas hembras y otra para las tortugas machos. A continuación, se escriben parcialmente estas matrices de datos.

Para las hembras:

$$\mathbf{X} = \begin{bmatrix} 98 & 81 & 38 \\ 103 & 84 & 38 \\ \vdots & \vdots & \vdots \\ 162 & 124 & 61 \\ 177 & 132 & 67 \end{bmatrix}$$

Para los machos:

$$\mathbf{X} = \begin{bmatrix} 93 & 74 & 37 \\ 94 & 78 & 35 \\ \vdots & \vdots & \vdots \\ 131 & 95 & 46 \\ 135 & 106 & 47 \end{bmatrix}.$$

En las hileras de estas matrices de datos están representados los individuos y en sus columnas, las variables, que en este caso corresponden al largo, ancho y altura del caparazón de las tortugas. ♦

EJEMPLO 4.1.2 Con los datos de el Cuadro 3.2.5, de los depósitos de corcho de 28 árboles en los cuatro puntos cardinales, se puede representar la matriz de datos usando la notación de la expresión (4.1-2) como sigue:

$$X = \begin{bmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_{28} \end{bmatrix} = [\mathbf{x}_{(1)} \quad \mathbf{x}_{(2)} \quad \mathbf{x}_{(3)} \quad \mathbf{x}_{(4)}]$$

donde,

$$\mathbf{x}_1 = \begin{bmatrix} 72 \\ 66 \\ 76 \\ 77 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 60 \\ 53 \\ 66 \\ 63 \end{bmatrix}, \dots, \mathbf{x}_9 = \begin{bmatrix} 37 \\ 40 \\ 31 \\ 25 \end{bmatrix}, \dots, \mathbf{x}_{28} = \begin{bmatrix} 48 \\ 54 \\ 57 \\ 43 \end{bmatrix} \quad \text{y} \quad \mathbf{x}_{(1)} = \begin{bmatrix} 72 \\ 60 \\ \vdots \\ 48 \end{bmatrix}, \dots, \mathbf{x}_{(4)} = \begin{bmatrix} 77 \\ 63 \\ \vdots \\ 43 \end{bmatrix}.$$

Se observa que $\mathbf{x}_1$ es la primera hilera de la matriz de datos X escrita en forma de columna y corresponde a las mediciones de los depósitos del primer árbol en las cuatro direcciones (N,E,S,W); así enseguida, hasta que $\mathbf{x}_{28}$ corresponde a las observaciones de la hilera 28 de la matriz de datos (escrita como columna) y contiene las mediciones de los depósitos del árbol 28 en las cuatro direcciones. Por otro lado, $\mathbf{x}_{(1)}$ contiene los datos de la primera columna de la matriz de datos que corresponden a las mediciones del depósito de corcho de los 28 árboles en la dirección norte; así sucesivamente, hasta que, la última columna de X , $\mathbf{x}_{(4)}$, contiene las mediciones del depósito de corcho de los 28 árboles en la dirección oeste. ♦

Tomando en cuenta la matriz de datos, los estudios multivariados pueden clasificarse desde dos puntos de vista alternativos: los métodos que hacen comparaciones entre las columnas de la matriz de datos (estudian las relaciones entre las variables) y aquellos que comparan las hileras de la matriz de datos (indagan las relaciones entre los individuos). Los primeros son conocidos como *técnicas-R*, por la importancia que juega la matriz de correlaciones R en este enfoque; los métodos que se dirigen a las relaciones entre individuos se les conoce como *técnicas-Q*. Dentro las técnicas-R, se incluyen el análisis de componentes principales, análisis de correlaciones canónicas, análisis de factores; las técnicas-Q, incluyen el análisis discriminante y el escalamiento multidimensional. En este

trabajo sólo se van a revisar algunas de las técnicas-R; el cálculo de la matriz de correlaciones $\mathbf{R}$ de una matriz de datos se presenta en la expresión (4.1-10) [*ibid.* p. 8].

4.1.3 Muestra Aleatoria Multivariada

Una muestra aleatoria multivariada se define como un conjunto de observaciones en n individuos seleccionados al azar de una población, a los cuales se les registra p variables, $i=1,2,\dots,n$, $j=1,2,\dots,p$; en la matriz de datos, $\mathbf{x}_i'$ es la hilera que contiene los valores de las p variables registradas en el individuo i , y $\mathbf{x}_{(j)}$ es la columna que contiene los registros de la j -ésima variable sobre los n individuos de la muestra. Usualmente, las hileras $\mathbf{x}_i'$ forman una muestra aleatoria mientras que con las columnas $\mathbf{x}_{(j)}$ no sucede así; una vez seleccionado al azar al i -ésimo individuo se registran sobre él las p variables, lo cual es enfatizado en la notación, por el uso de paréntesis en los subíndices de los vectores columna de la matriz de datos (*ibid.* p. 9).

En una muestra aleatoria univariada se tienen los registros de una sola variable sobre un conjunto de n individuos seleccionados al azar de una población de tamaño N , $n \leq N$; en cambio, en el caso multivariado, sobre los n individuos seleccionados al azar, se consideran simultáneamente las mediciones de p variables o características de interés.

4.1.4 El Vector de Medias y la Matriz de Covarianzas Muestrales

La media muestral de la j -ésima variable —haciéndolo de una por una, como en el caso univariado— se define escalarmente por:

$$\bar{x}_{.j} = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad (4.1-3)$$

Las $\bar{x}_{.j}$ se pueden arreglar en el *vector muestral de medias* de dimensión $(p \times 1)$, que se obtiene matricialmente por:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i = \frac{1}{n} \mathbf{X}' \mathbf{1} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_{\cdot p} \end{bmatrix}_{p \times 1} \quad (4.1-4)$$

donde $\mathbf{1} = (1, 1, \dots, 1)'_{n \times 1}$ —es un vector columna de puros unos de tamaño n .

EJEMPLO 4.1.3 Usando los datos del Cuadro 3.2.5, se puede calcular el vector de medias de la siguiente manera usando la expresión (4.1-4):

$$\bar{\mathbf{x}} = \frac{1}{28} \left(\begin{bmatrix} 72 \\ 66 \\ 76 \\ 77 \end{bmatrix} + \begin{bmatrix} 60 \\ 53 \\ 66 \\ 63 \end{bmatrix} + \dots + \begin{bmatrix} 48 \\ 54 \\ 57 \\ 43 \end{bmatrix} \right) = \begin{bmatrix} 72 & 66 & 76 & 77 \\ 60 & 53 & 66 & 63 \\ \vdots & \vdots & \vdots & \vdots \\ 48 & 54 & 57 & 43 \end{bmatrix}' \begin{bmatrix} \frac{1}{28} \\ \frac{1}{28} \\ \vdots \\ \frac{1}{28} \end{bmatrix} = \begin{bmatrix} 50.54 \\ 46.18 \\ 49.68 \\ 45.18 \end{bmatrix}$$

Es claro que $\bar{x}_1 = 50.54$ es el depósito de corcho promedio en la dirección norte; $\bar{x}_2 = 46.18$ es el depósito de corcho promedio en la dirección este; y así sucesivamente.

♦

La varianza muestral de la j -ésima variable, definiéndola por analogía del contexto univariado es:

$$s_{jj} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_{\cdot j})^2 = s_j^2 \quad j = 1, 2, \dots, p, \quad (4.1-5)$$

y la covarianza muestral entre la j -ésima y la k -ésima variables es:

$$s_{jk} = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_{\cdot j})(x_{ik} - \bar{x}_{\cdot k}) \quad (4.1-6)$$

Los valores de s_{jj} y s_{jk} , de las expresiones (4.1-5) y (4.1-6), se pueden disponer en la *matriz de varianzas y covarianzas muestral*, denotada por $\mathbf{S}$, que se obtiene usando notación matricial con (4.1-7). A $\mathbf{S}$ usualmente se le nombra sencillamente como *matriz de covarianzas*.

$$\mathbf{S} = \frac{1}{n} \mathbf{X}' \mathbf{H} \mathbf{X} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} \quad (4.1-7)$$

donde $\mathbf{H} = \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}'$ denota la *matriz centrante*. La matriz $\mathbf{S}$ contiene en su diagonal principal a las $s_{jj} = s_j^2$ y fuera de la diagonal principal están las s_{jk} , cumpliéndose que $\mathbf{S} = \mathbf{S}'$.

Como en la estadística univariada, resulta conveniente definir una matriz de covarianzas que tenga como divisor $n - 1$, en lugar de n . Así:

$$\mathbf{S}_u = \frac{1}{n-1} \mathbf{X}' \mathbf{H} \mathbf{X} \quad (4.1-8)$$

es un estimador insesgado de la matriz de covarianzas poblacional Σ , si los datos forman una muestra aleatoria multivariada con los segundos momentos finitos.

EJEMPLO 4.1.4 Con los datos de depósito de corcho, usados en los ejemplos anteriores se ejemplifica el cálculo de la matriz de covarianzas.

$$\begin{aligned} \mathbf{S} = \frac{1}{n} \mathbf{X}' \mathbf{H} \mathbf{X} &= \frac{1}{28} \left\{ \begin{matrix} \begin{bmatrix} 72 & 60 & \cdots & 48 \\ 66 & 53 & \cdots & 54 \\ 76 & 66 & \cdots & 57 \\ 77 & 63 & \cdots & 43 \end{bmatrix} & \begin{bmatrix} \frac{27}{28} & -\frac{1}{28} & \cdots & -\frac{1}{28} \\ -\frac{1}{28} & \frac{27}{28} & \cdots & -\frac{1}{28} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{28} & -\frac{1}{28} & \cdots & \frac{27}{28} \end{bmatrix} & \begin{bmatrix} 72 & 66 & 76 & 77 \\ 60 & 53 & 66 & 63 \\ \vdots & \vdots & \vdots & \vdots \\ 48 & 54 & 57 & 43 \end{bmatrix} \end{matrix} \right\} \\ &\quad \begin{matrix} 4 \times 28 & 28 \times 28 & 28 \times 4 \end{matrix} \\ &= \frac{1}{28} \begin{bmatrix} 7840.96 & 6041.32 & 7787.82 & 6109.32 \\ 6041.32 & 5938.11 & 6184.61 & 4627.11 \\ 7787.82 & 6184.61 & 9450.11 & 7007.61 \\ 6109.32 & 4627.11 & 7007.61 & 6102.11 \end{bmatrix} = \begin{bmatrix} 280.03 & 215.76 & 278.14 & 218.19 \\ 215.76 & 212.08 & 220.88 & 165.25 \\ 278.14 & 220.88 & 337.50 & 250.27 \\ 218.19 & 165.25 & 250.27 & 217.93 \end{bmatrix}, \\ &\quad \begin{matrix} 4 \times 4 \end{matrix} \end{aligned}$$

y el estimador insesgado de Σ es:

$$S_u = \frac{1}{27} \begin{bmatrix} 7840.96 & 6041.32 & 7787.82 & 6109.32 \\ 6041.32 & 5938.11 & 6184.61 & 4627.11 \\ 7787.82 & 6184.61 & 9450.11 & 7007.61 \\ 6109.32 & 4627.11 & 7007.61 & 6102.11 \end{bmatrix} = \begin{bmatrix} 290.41 & 223.75 & 288.44 & 226.27 \\ 223.75 & 219.93 & 229.06 & 171.37 \\ 288.44 & 229.06 & 350.00 & 259.54 \\ 226.27 & 171.37 & 259.54 & 226.00 \end{bmatrix}.$$

♦

El coeficiente de correlación muestral entre la j -ésima y la k -ésima variable se define escalarmente como:

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj} s_{kk}}} = \frac{s_{jk}}{s_j s_k}, \quad (4.1-9)$$

y la *matriz muestral de correlaciones* que contiene a los r_{jk} se obtiene con:

$$\mathbf{R} = \mathbf{D}^{-\frac{1}{2}} \mathbf{S} \mathbf{D}^{-\frac{1}{2}} \quad (4.1-10)$$

donde $\mathbf{D}^{\frac{1}{2}} = \sqrt{\text{diag}(\mathbf{S})} = \text{diag}\{\sqrt{s_{jj}}\} = \text{diag}\{s_j\}$. Es decir,

$$\mathbf{D}^{-\frac{1}{2}} = \begin{bmatrix} s_{11}^{-\frac{1}{2}} & 0 & \cdots & 0 \\ 0 & s_{22}^{-\frac{1}{2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_{pp}^{-\frac{1}{2}} \end{bmatrix}.$$

Al aplicar (4.1-10) queda:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}_{p \times p};$$

los unos de la diagonal principal de $\mathbf{R}$, se desprenden directamente de la definición del coeficiente de correlación muestral r_{jk} en (4.1-9).

EJEMPLO 4.1.5 Continuando con los datos de el Cuadro 3.2.5 y haciendo uso de la matriz $\mathbf{S}$ del Ejemplo 4.1.4, se procede a calcular su correspondiente matriz de correlaciones $\mathbf{R}$. Así:

$$R = \begin{bmatrix} 0.0598 & 0 & 0 & 0 \\ 0 & 0.0687 & 0 & 0 \\ 0 & 0 & 0.0544 & 0 \\ 0 & 0 & 0 & 0.0677 \end{bmatrix} S = \begin{bmatrix} 0.0598 & 0 & 0 & 0 \\ 0 & 0.0687 & 0 & 0 \\ 0 & 0 & 0.0544 & 0 \\ 0 & 0 & 0 & 0.0677 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0.8854 & 0.9047 & 0.8832 \\ 0.8854 & 1 & 0.8256 & 0.7687 \\ 0.9047 & 0.8256 & 1 & 0.9228 \\ 0.8832 & 0.7687 & 0.9228 & 1 \end{bmatrix}$$

4.1.5 Medidas de Dispersión Multivariada

La matriz S es una posible generalización de la noción univariada de varianza, que mide la dispersión con respecto a la media. Sin embargo, algunas veces es conveniente contar con un solo número que mida la dispersión multivariada. Dos medidas usuales son:

- (1) la varianza generalizada que es el determinante de S , $|S|$; y
- (2) la variación total, definida como la traza de S , $\text{tr } S$.

La varianza generalizada juega un importante papel en la estimación de máxima verosimilitud y la variación total es un concepto útil en la presentación de las componentes principales (*ibid.* p.13).

4.1.6 Combinaciones Lineales (*ibid.* p. 13)

Una de las herramientas más importantes del análisis multivariado, son las combinaciones lineales de las variables. Unas cuantas combinaciones lineales bien seleccionadas pueden proveer mayor información que un gran número de las variables originales, frecuentemente porque pueden reducir la dimensionalidad de los datos. Las transformaciones lineales pueden también simplificar la estructura de la matriz de covarianzas, haciendo que la interpretación de los datos sea más directa.

Considerar una combinación lineal

$$y_r = a_1 x_{r1} + \dots + a_p x_{rp}, \quad r = 1, \dots, n, \quad (4.1-11)$$

donde $a_1, \dots, a_p$ son valores dados. De (4.1-4), la media de y_r , denotada por $\bar{y}$, está dada por:

$$\bar{y} = \frac{1}{n} \mathbf{a}' \sum_{r=1}^n \mathbf{x}_r = \mathbf{a}' \bar{\mathbf{x}}, \quad (4.1-12)$$

y se tiene que su varianza es:

$$s_y^2 = \frac{1}{n} \sum_{r=1}^n (y_r - \bar{y})^2 = \frac{1}{n} \sum_{r=1}^n \mathbf{a}' (\mathbf{x}_r - \bar{\mathbf{x}}) (\mathbf{x}_r - \bar{\mathbf{x}})' \mathbf{a} = \mathbf{a}' \mathbf{S} \mathbf{a}. \quad (4.1-13)$$

En general es de interés la transformación lineal q -dimensional:

$$\mathbf{y}_r = \mathbf{A} \mathbf{x}_r + \mathbf{b}, \quad r = 1, \dots, n, \quad (4.1-14)$$

que puede escribirse como $\mathbf{Y} = \mathbf{X} \mathbf{A}' + \mathbf{1} \mathbf{b}'$, donde $\mathbf{A}$ es una matriz de dimensión $(q \times p)$ y $\mathbf{b}$ es un vector de tamaño q .

El vector de medias y la matriz de covarianzas de los nuevos individuos $\mathbf{y}_r$ están dados por:

$$\bar{\mathbf{y}} = \mathbf{A} \bar{\mathbf{x}} + \mathbf{b}, \quad (4.1-15)$$

$$\mathbf{S}_y = \frac{1}{n} \sum_{r=1}^n (\mathbf{y}_r - \bar{\mathbf{y}}) (\mathbf{y}_r - \bar{\mathbf{y}})' = \mathbf{A} \mathbf{S} \mathbf{A}'. \quad (4.1-16)$$

Cuando $\mathbf{A}$ es una matriz no-singular (i.e., $|\mathbf{A}| \neq 0$) entonces:

$$\mathbf{S} = \mathbf{A}^{-1} \mathbf{S}_y (\mathbf{A}')^{-1} \quad (4.1-17)$$

Estos diferentes ejemplos de transformaciones lineales, están centrados con media cero.

4.1.6.1 Transformaciones de escala

Sea $y_r = D^{-1/2}(x_r - \bar{x})$, $r = 1, \dots, n$, donde $D^{\frac{1}{2}} = \text{diag}\{\sqrt{s_{jj}}\} = \text{diag}\{s_j\}$ [Observe que $D^{\frac{1}{2}}$ es igual a la de la expresión (4.1-10)]. Con esta transformación lineal se hace que cada variable tenga varianza uno (se dice que se estandariza la escala de las variables) eliminando así la selección arbitraria de escalas. Por ejemplo, si $x_{(1)}$ representa el diámetro normal¹, $y_{(1)}$ resultará la misma, independientemente que $x_{(1)}$ se mida en centímetros, en pulgadas, o en cualquier otra unidad de longitud. En este caso, $S_y = R$.

4.1.6.2 Transformación de Mahalanobis

Si $S > 0$ entonces S^{-1} tiene una única matriz raíz cuadrada simétrica y positiva definida $S^{-1/2}$ ($\exists \cdot S^{-1/2}S^{-1/2} = S^{-1}$). La transformación de Mahalanobis se define como:

$$z_r = S^{-1/2}(x - \bar{x}), \quad r = 1, \dots, n. \quad (4.1-18)$$

Entonces, la varianza de la transformación de Mahalanobis es $S_z = I$, lo que significa que esta transformación elimina la correlación entre variables y estandariza la varianza de cada variable.

4.1.6.2 Transformación componentes principales

Por el teorema de la descomposición espectral, la matriz de covarianzas S puede escribirse como:

$$S = GLG', \quad (4.1-19)$$

donde L es una matriz diagonal que contiene los valores característicos de S ordenados descendientemente ($l_1 \geq l_2 \geq \dots \geq l_p \geq 0$) y G es una matriz ortogonal ($GG' = G'G = I$) que contiene los vectores característicos de S asociados a los l_i , $i = 1, \dots, p$ de L . La transformación componentes principales se define como:

¹ Por convención, se entenderá por **diámetro normal** (*dn*), al diámetro de un árbol medido a la altura de 1.30 m de la base del mismo. En varios trabajos se usa el término **diámetro a la altura del pecho** (*dap* o *dbh*) como sinónimo.

$$\mathbf{w}_r = \mathbf{G}'(\mathbf{x}_r - \bar{\mathbf{x}}), \quad r = 1, 2, \dots, p. \quad (4.1-20)$$

Como la varianza de $\mathbf{w}$, $\mathbf{S}_w = \mathbf{G}'\mathbf{S}\mathbf{G} = \mathbf{L}$, es una matriz diagonal, las componentes principales contenidas en las columnas de $\mathbf{W}$, representan combinaciones lineales no correlacionadas de las variables originales. Aunque se necesitan p componentes principales para reproducir la variabilidad total de los datos originales, se espera que la mayor parte de ésta, esté contenida en un menor número de ellas, digamos $k < p$. Si sucede que existe casi la misma información en las k componentes principales que en las p variables originales, entonces éstas podrían ser reemplazadas por aquéllas; y el conjunto de datos originales (n mediciones sobre p variables) se reduce a uno que contiene n observaciones sobre k componentes principales. Cuando esto ocurre, los estadísticos dicen que la dimensionalidad del problema se reduce de p a k .

4.1.7 La Distribución Multinormal (*ibid.* p. 13)

La distribución de probabilidad multivariada más importante es la normal multivariada y su *función de densidad de probabilidad* (f.d.p) se puede escribir como extensión de su contraparte univariada. Si la f.d.p. de $N(\mu, \sigma^2)$, la normal univariada con media μ y varianza $\sigma^2 > 0$, se escribe como:

$$f(x) = \{2\pi\sigma^2\}^{-1/2} \exp\left\{-\frac{1}{2}(x - \mu)(\sigma^2)^{-1}(x - \mu)\right\},$$

entonces una extensión razonable con p variables es:

$$f(\mathbf{x}) = \{2\pi\Sigma\}^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)'\Sigma^{-1}(\mathbf{x} - \mu)\right\}, \quad (4.1-21)$$

donde $\Sigma > 0$.

Definición 4.1.7.1. Se dice que el vector aleatorio $\mathbf{x}$ tiene distribución normal p -variada (o normal p -dimensional o normal multivariada) con vector de medias μ y matriz de covarianzas Σ si su f.d.p. está dado por (4.1-21) y se denota por $\mathbf{x} \sim N_p(\mu, \Sigma)$. $\square$

Teorema 4.1.7.1. Suponer que $\mathbf{x}$ tiene la f.d.p. dada en (4.1-21) y sea

$$\mathbf{y} = \Sigma^{-1/2}(\mathbf{x} - \mu) \quad (4.1-22)$$

donde $\Sigma^{-1/2}$ es la raíz cuadrada simétrica positiva-definida de Σ^{-1} . Entonces $y_1, \dots, y_p$ son variables independientes con distribución normal estándar, $N(0,1)$. Esto significa que $E(\mathbf{y}) = \mathbf{0}$ y $Var(\mathbf{y}) = \mathbf{I}$. [La prueba está en Mardia *et al.*, 1989, p. 37] $\square$

Corolario 1. Suponer que $\mathbf{x}$ tiene la f.d.p. dada en (4.1-21) entonces:

$$E(\mathbf{x}) = \mu, \quad Var(\mathbf{x}) = \Sigma. \quad (4.1-23)$$

$\square$

4.1.7.1 Algunas propiedades

Si $\mathbf{x} \sim N_p(\mu, \Sigma)$ se tienen los siguientes resultados.

Teorema 4.1.7.2. Si $U = (\mathbf{x} - \mu)' \Sigma^{-1}(\mathbf{x} - \mu)$ entonces U tiene una distribución ji-cuadrada con p grados de libertad, denotándose como

$$U \sim \chi_p^2. \quad (4.1-24)$$

$\square$

Teorema 4.1.7.3. Todas las combinaciones lineales no-triviales de los elementos $\mathbf{x}$ tienen distribución normal univariada. $\square$

4.1.7.2 La distribución normal de una matriz

Sea $\mathbf{X}_{n \times p}$ una matriz cuyas n hileras, $\mathbf{x}'_1, \dots, \mathbf{x}'_n$, se distribuyen independientemente como $N_p(\mu, \Sigma)$. Entonces $\mathbf{X}$ tiene la distribución normal matricial y representa una muestra aleatoria que proviene de una población $N_p(\mu, \Sigma)$. La f.d.p. de $\mathbf{X}$ es:

$$\begin{aligned} f(\mathbf{X}) &= |2\pi\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \mu)' \Sigma^{-1}(\mathbf{x}_i - \mu)\right\} \\ &= |2\pi\Sigma|^{-n/2} \exp\left\{-\frac{1}{2} \text{tr}[\Sigma^{-1}(\mathbf{X} - \mathbf{1}\mu')'(\mathbf{X} - \mathbf{1}\mu')]\right\}. \end{aligned}$$

4.1.8 La Distribución Wishart (*ibid.* p. 66)

Definición 4.1.8.1. Si la matriz $\mathbf{M}_{p \times p}$ puede escribirse como $\mathbf{M} = \mathbf{X}'\mathbf{X}$ donde $\mathbf{X}_{m \times p}$ es una matriz de datos proveniente de $N_p(\mathbf{0}, \Sigma)$, entonces se dice que $\mathbf{M}$ tiene una distribución Wishart con matriz de escala Σ y parámetro de grados de libertad m . Esto se denota como $\mathbf{M} \sim W_p(\Sigma, m)$. Cuando $\Sigma = \mathbf{I}_p$, se dice que la distribución está en forma estándar. $\square$

La distribución Wishart es una generalización matricial de la distribución ji-cuadrada univariada. Cuando $p = 1$ variable, la distribución $W_1(\sigma^2, m)$ está dada por $\mathbf{x}'\mathbf{x}$ donde los elementos de $\mathbf{x}_{m \times 1}$ se distribuyen idéntica e independientemente como $N_1(0, \sigma^2)$; esto es, la distribución $W_1(\sigma^2, m)$ es la misma que la distribución $\sigma^2 \chi_m^2$. La matriz de escala Σ en la distribución Wishart juega el mismo papel que σ^2 en la distribución $\sigma^2 \chi_m^2$.

4.1.8.1 Propiedades de las matrices Wishart

Teorema 4.1.8.1. Si $\mathbf{M} \sim W_p(\Sigma, m)$ y $\mathbf{B}$ es una matriz de dimensión $(p \times q)$, entonces $\mathbf{B}'\mathbf{M}\mathbf{B} \sim W_q(\mathbf{B}'\Sigma\mathbf{B}, m)$. $\square$

Corolario 4.1.8.1.1. La submatrices diagonales de $\mathbf{M}$ por sí mismas tienen distribución Wishart. $\square$

Corolario 4.1.8.1.2. $\Sigma^{-1/2}\mathbf{M}\Sigma^{-1/2} \sim W_p(\mathbf{I}, m)$. $\square$

Corolario 4.1.8.1.3. Si $\mathbf{M} \sim W_p(\mathbf{I}, m)$ y $\mathbf{B}_{p \times q}$ satisface $\mathbf{B}'\mathbf{B} = \mathbf{I}_q$, entonces $\mathbf{B}'\mathbf{M}\mathbf{B} \sim W_q(\mathbf{I}, m)$. $\square$

Del Corolario 4.1.8.1.1, cada elemento de la diagonal de $\mathbf{M}$ tiene una distribución $\sigma_i^2 \chi_m^2$.

Teorema 4.1.8.2. Si $\mathbf{M} \sim W_p(\Sigma, m)$ y $\mathbf{a}$ es cualquier vector fijo de dimensión p tal que $\mathbf{a}'\Sigma\mathbf{a} \neq 0$, entonces

$$\frac{\mathbf{a}'\mathbf{M}\mathbf{a}}{\mathbf{a}'\Sigma\mathbf{a}} \sim \chi_m^2. \quad \square$$

Corolario 4.1.8.2.1. $m_{ii} \sim \sigma_i^2 \chi_m^2. \quad \square$

Teorema 4.1.8.3. Si $\mathbf{M}_1 \sim W_p(\Sigma, m_1)$ y $\mathbf{M}_2 \sim W_p(\Sigma, m_2)$ y si $\mathbf{M}_1$ y $\mathbf{M}_2$ son independientes, entonces $\mathbf{M}_1 + \mathbf{M}_2 \sim W_p(\Sigma, m_1 + m_2). \quad \square$

Teorema 4.1.8.4. Si $\mathbf{X}(n \times p)$ es una matriz de datos proveniente de una distribución $N_p(\mathbf{0}, \Sigma)$ y si $\mathbf{C}(n \times n)$ es una matriz simétrica, entonces:

- (a) $\mathbf{X}'\mathbf{C}\mathbf{X}$ tienen la misma distribución que una suma ponderada de matrices $W_p(\Sigma, 1)$ independientes, donde las ponderaciones son valores característicos de $\mathbf{C}$;
- (b) $\mathbf{X}'\mathbf{C}\mathbf{X}$ tiene una distribución Wishart si y sólo si $\mathbf{C}$ es una matriz idempotente, en cuyo caso $\mathbf{X}'\mathbf{C}\mathbf{X} \sim W_p(\Sigma, r)$ donde $r = \text{tr } \mathbf{C} = \text{rango}(\mathbf{C})$;
- (c) Si $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{H}\mathbf{X}$ es la matriz muestral de covarianzas, entonces $n\mathbf{S} \sim W_p(\Sigma, n-1).$

$\square$

4.1.8.2 Matrices Wishart particionadas

Si $\mathbf{M} \sim W_p(\Sigma, m)$, frecuentemente es útil hacer la partición:

$$\mathbf{M} = \begin{bmatrix} \mathbf{M}_{11}(a \times a) & \mathbf{M}_{12}(a \times b) \\ \mathbf{M}_{21}(b \times a) & \mathbf{M}_{22}(b \times b) \end{bmatrix}_{p \times p} \quad (4.1-25)$$

Se puede querer dividir las p variables originales en dos subgrupos, que consisten de a y b variables, respectivamente, donde $a + b = p$.

Las submatrices $\mathbf{M}_{11}$ y $\mathbf{M}_{22}$ tienen distribuciones Wishart, aunque estas distribuciones, en general, no son independientes. Sin embargo, $\mathbf{M}_{11}$ es independiente de:

$$\mathbf{M}_{22.1} = \mathbf{M}_{22} - \mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{M}_{12}. \quad (4.1-26)$$

Aquí $\mathbf{M}_{22.1}$ es m veces el análogo muestral de $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.

Cuando $p = 2$,

$$\mathbf{M} = m \begin{bmatrix} s_1^2 & rs_1s_2 \\ rs_1s_2 & s_2^2 \end{bmatrix}. \quad (4.1-27)$$

En este contexto, $\mathbf{M}_{22.1}$ se reduce a:

$$\mathbf{M}_{22.1} = ms_2^2(1 - r^2). \quad (4.1-28)$$

Teorema 4.1.8.5. Si $\mathbf{M} \sim W_p(\Sigma, m)$, $m > a$ entonces:

- (a) $\mathbf{M}_{22.1}$ tiene la distribución $W_b(\Sigma_{22.1}, m - a)$ y es independiente de $(\mathbf{M}_{11}, \mathbf{M}_{12})$, y
- (b) si $\Sigma_{12} = \mathbf{O}$, entonces $\mathbf{M}_{22} - \mathbf{M}_{22.1} = \mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{M}_{12}$ tiene la distribución $W_b(\Sigma_{22}, a)$, y $\mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{M}_{12}$, $\mathbf{M}_{11}$ y $\mathbf{M}_{22.1}$ son conjuntamente independientes. $\square$

Haciendo que $\mathbf{M}^{22} = \mathbf{M}_{22.1}^{-1} = (\mathbf{M}_{22} - \mathbf{M}_{21}\mathbf{M}_{11}^{-1}\mathbf{M}_{12})^{-1}$, se pueden escribir los dos incisos del Teorema anterior como:

Corolario 4.1.8.5.1

- (a) $(\mathbf{M}^{22})^{-1} \sim W_b((\Sigma^{22})^{-1}, m - a)$ y es independiente de $(\mathbf{M}_{11}, \mathbf{M}_{12})$. [Aquí $\Sigma^{22} = \Sigma_{22.1}^{-1}$]
- (b) Si $\Sigma_{12} = \mathbf{O}$ entonces $\mathbf{M}_{22} - (\mathbf{M}^{22})^{-1} \sim W_b(\Sigma_{22}, a)$ y $\mathbf{M}_{22} - (\mathbf{M}^{22})^{-1}$, $\mathbf{M}_{11}$ y $(\mathbf{M}^{22})^{-1}$ son conjuntamente independientes. $\square$

Teorema 4.1.8.6. Si $\mathbf{M} \sim W_p(\Sigma, m)$, $m > p$, entonces:

- (a) la relación $\mathbf{a}'\Sigma^{-1}\mathbf{a}/\mathbf{a}'\mathbf{M}^{-1}\mathbf{a}$ se distribuye como una χ_{m-p+1}^2 para cualquier vector fijo $\mathbf{a}$ de dimensión p ; en particular, $\sigma^{ii}/m^{ii} \sim \chi_{m-p+1}^2$ para $i = 1, \dots, p$.

(b) m^{ii} es independiente de todos los elementos de $\mathbf{M}$ excepto de m_{ii} .

□

Teorema 4.1.8.7. Si $\mathbf{M} \sim W_p(\Sigma, m)$, $m \geq p$, entonces $|\mathbf{M}|$ es $|\Sigma|$ veces (el producto de) p variables aleatorias ji-cuadradas independientes con $m, m-1, \dots, m-p+1$ grados de libertad. □

Corolario 4.1.8.7.1. Si $\mathbf{M} \sim W_p(\Sigma, m)$, $\Sigma > 0$ y $m > p$, entonces $\mathbf{M} > 0$ con probabilidad uno.

□

4.1.9 La Distribución T^2 de Hotelling (*ibid.* p. 73)

Definición 4.1.9.1. Si α puede escribirse como $m\mathbf{d}'\mathbf{M}^{-1}\mathbf{d}$ donde $\mathbf{d}$ y $\mathbf{M}$ se distribuyen independientemente como $N_p(\mathbf{0}, \mathbf{I})$ y $W_p(\mathbf{I}, m)$, respectivamente, entonces se dice que α tiene la distribución T^2 de Hotelling con parámetros p y m . La notación es $\alpha \sim T^2(p, m)$.

□

Teorema 4.1.9.1. Si $\mathbf{x}$ y $\mathbf{M}$ se distribuyen independientemente como $N_p(\mu, \Sigma)$ y $W_p(\Sigma, m)$, respectivamente, entonces:

$$m(\mathbf{x} - \mu)' \mathbf{M}^{-1}(\mathbf{x} - \mu) \sim T^2(p, m). \quad (4.1-29)$$

□

Corolario 4.1.9.1.1. Si $\bar{\mathbf{x}}$ y $\mathbf{S}$ son el vector de medias y la matriz de covarianzas de una muestra de tamaño n proveniente de una población $N_p(\mu, \Sigma)$, y $\mathbf{S}_u = (n/n-1)\mathbf{S}$, entonces:

$$(n-1)(\bar{\mathbf{x}} - \mu)' \mathbf{S}^{-1}(\bar{\mathbf{x}} - \mu) = n(\bar{\mathbf{x}} - \mu)' \mathbf{S}_u^{-1}(\bar{\mathbf{x}} - \mu) \sim T^2(p, n-1) \quad (4.1-30)$$

Corolario 4.1.9.1.2. El estadístico T^2 es invariante bajo cualquier transformación lineal no-singular $\mathbf{x} \rightarrow \mathbf{Ax} + \mathbf{b}$

El estadístico t univariado también tiene la anterior propiedad de invarianza. Sin embargo el cuadrado de la variable t_m univariada tiene la distribución $T^2(1, m)$. Es decir, la distribución $F_{1, m}$ y la distribución $T^2(1, m)$ son iguales.

Teorema 4.1.9.2.

$$T^2(p, m) = \{mp/(m - p + 1)\} F_{p, m-p+1}. \quad (4.1-31)$$

Corolario 4.1.9.2.1. Si $\bar{x}$ y S son la media y la covarianza de una muestra de tamaño n proveniente de $N_p(\mu, \Sigma)$, entonces:

$$\{(n - p)/p\}(\bar{x} - \mu)' S^{-1}(\bar{x} - \mu) \sim F_{p, n-p}. \quad (4.1-32)$$

Corolario 4.1.9.2.2. $|M|/|M + dd'| \sim B(\frac{1}{2}(m - p + 1), \frac{1}{2}p)$ donde $B(\cdot, \cdot)$ es una variable beta.

Teorema 4.1.9.3. Si d y M se distribuyen independientemente como $N_p(0, I)$ y $W_p(I, m)$, respectivamente, entonces:

$$d'd \left(1 + \frac{1}{d'M^{-1}d} \right) \sim \chi_{m+1}^2$$

y se distribuye independientemente de $d'M^{-1}d$.

El Teorema 4.1.9.3 es una extensión del siguiente resultado univariado: Si $d \sim N(0, 1)$ y $u^2 \sim \chi^2$, entonces $d^2 + u^2$ es independiente de d^2/u^2 . Antes de escribir otra generalización del mismo resultado se requiere el siguiente lema.

Lema. Sea W una matriz aleatoria simétrica de orden $(p \times p)$ y sea x un vector aleatorio de dimensión p . Si x es independiente de $(g_1' W g_1, \dots, g_p' W g_p)$ para toda matriz ortogonal no-aleatoria $G = (g_1, \dots, g_p)'$, entonces x es independiente de W .

Teorema 4.1.9.4. Si $\mathbf{d}$ y $\mathbf{M}$ se distribuyen independientemente como $N_p(\mathbf{0}, \mathbf{I})$ y $W_p(\mathbf{I}, m)$, respectivamente, entonces $\mathbf{d}'\mathbf{M}^{-1}\mathbf{d}$ es independiente de $\mathbf{M} + \mathbf{d}\mathbf{d}'$.

4.1.10 El Estadístico T^2 de Hotelling para Dos-Muestras (*ibid.* p. 76)

La llamada distancia de Mahalanobics entre dos poblaciones con medias μ_1 y μ_2 con matriz de covarianzas Σ está dada por Δ , donde:

$$\Delta^2 = (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2). \quad (4.1-33)$$

Por supuesto que sólo raras poblaciones cuentan con parámetros conocidos, por lo que es usual hacer estimaciones de sus correspondientes valores muestrales. Suponer que se tienen dos muestras de tamaño n_1 y n_2 de forma que $n_1 + n_2 = n$; entonces la distancia de Mahalanobics (muestral), D , puede escribirse como:

$$D^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_u^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2), \quad (4.1-34)$$

donde $\mathbf{S}_u = (n_1 \mathbf{S}_1 + n_2 \mathbf{S}_2) / (n - 2)$ es un estimador insesgado de Σ . (Observe que $\bar{\mathbf{x}}_i$ y $\mathbf{S}_i$, $i = 1, 2$, son el vector de medias y la matriz de covarianzas de la i -ésima muestra.) En el siguiente teorema se escribe la distribución estadística de D^2 bajo un conjunto particular de supuestos.

Teorema 4.1.10.1. Si $\mathbf{X}_1$ y $\mathbf{X}_2$ son matrices de datos independientes y las n_i hileras de $\mathbf{X}_i$ son idéntica e independientemente distribuidas (i.i.d.) como $N_p(\mu_i, \Sigma_i)$, $i = 1, 2$, entonces cuando $\mu_1 = \mu_2$ y $\Sigma_1 = \Sigma_2$, $(n_1 n_2 / n) D^2$ es una variable $T^2(p, n - 2)$.

La cantidad

$$(n_1 n_2 / n) D^2 = (n_1 n_2 / n) (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}_u^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \quad (4.1-35)$$

se conoce como estadístico T^2 de Hotelling para dos muestras. Usando la relación entre los estadísticos T^2 y F , se puede deducir también que, bajo los supuestos dados,

$$\frac{n_1 n_2 (n-p-1)}{n(n-2)p} D^2 \sim F_{p, n-p-1}. \quad (4.1-36)$$

4.1.11 Estadísticos con Base en la Distribución Wishart (*ibíd.* p. 80)

Muchas pruebas en el análisis univariado se basan sobre estadísticos que tienen distribuciones ji-cuadradas independientes. Una hipótesis particular podría implicar que $a \sim \sigma^2 \chi_a^2$ y $b \sim \sigma^2 \chi_b^2$, donde a y b son estadísticos basados sobre los datos. Si a y b son independientes entonces a/b es α/β veces una variable $F_{\alpha, \beta}$, y $a/(a+b)$ tiene una distribución beta con parámetros $\frac{1}{2}\alpha$ y $\frac{1}{2}\beta$. Ninguna de estas funciones involucra el parámetro σ que, en general, es desconocido.

En el escenario multivariado muchos estadísticos se basan en distribuciones Wishart independientes. (Recordar que la distribución Wishart multivariada es una generalización de la ji-cuadrada univariada.) Sea $\mathbf{A} \sim W_p(\Sigma, m)$ independiente de $\mathbf{B} \sim W_p(\Sigma, n)$ donde $m \geq p$. Dado que $m \geq p$, existe la inversa $\mathbf{A}^{-1}$, y son de interés los valores característicos diferentes de cero de $\mathbf{A}^{-1}\mathbf{B}$. Los valores característicos de $\mathbf{A}^{-1}\mathbf{B} = \mathbf{A}^{-1/2}\mathbf{A}^{-1/2}\mathbf{B}$ son los mismos que los de $\mathbf{A}^{-1/2}\mathbf{B}\mathbf{A}^{-1/2}$, y como esta última es positiva semidefinida (p.s.d), todos los raíces características diferentes de cero son positivos. Así también, con probabilidad uno, el número de valores característicos diferentes de cero es igual a $\min(n, p)$. Debido a que la matriz de escala Σ no afecta la distribución de las raíces características, sin pérdida de generalidad se puede suponer $\Sigma = \mathbf{I}$.

La distribución conjunta de la raíces de $\mathbf{A}^{-1}\mathbf{B}$ se denota con $\Psi(p, m, n)$. El siguiente teorema establece una importante relación entre las distribuciones Ψ para diferentes valores de los parámetros.

Teorema 4.1.11.1. Para $m \geq p$ y $n, p \geq 1$, la distribución $\Psi(p, m, n)$ es idéntica a la distribución $\Psi(n, m+n-p, p)$.

Teorema 4.1.11.2. Si $A \sim W_p(\Sigma, m)$ y $B \sim W_p(\Sigma, n)$ son independientes con $m \geq p$ y $n \geq p$, entonces:

$$\phi = \frac{|A^{-1}B|}{|A|} \quad (4.1-37)$$

es proporcional al producto de p variables F independientes, de las cuales la i -ésima tiene $(n-i+1)$ g.l. en el numerador y $(m-i+1)$ g.l. en el denominador.

Definición 4.1.11.1. Cuando $A \sim W_p(I, m)$ y $B \sim W_p(I, n)$ son independientes, con $m \geq p$, se dice que:

$$\Lambda = \frac{|A|}{|A+B|} = |I + A^{-1}B|^{-1} \sim \Lambda(p, m, n) \quad (4.1-38)$$

tiene una distribución lambda de Wilks con parámetros p, m , y n .

La familia de distribuciones Λ ocurre frecuentemente en el contexto de las pruebas de razón de verosimilitud. El parámetro m usualmente representa los g.l. del “error” y n los g.l. de la “hipótesis” por los que $m+n$ representa los g.l. del “total”. Al igual que el estadístico T^2 , la distribución lambda de Wilks es invariante bajo cambios de los parámetros de escala de A y B . La distribución de Λ está dada en el siguiente teorema.

Teorema 4.1.11.3. Se tiene:

$$\Lambda(p, m, n) \sim \prod_{i=1}^n u_i, \quad (4.1-39)$$

donde $u_1, \dots, u_n$ son n variables independientes y $u_i \sim B\left(\frac{1}{2}(m+i-p), \frac{1}{2}p\right)$, $i = 1, \dots, p$.

Teorema 4.1.11.4. Las distribuciones $\Lambda(p, m, n)$ y $\Lambda(n, m+n-p, p)$ son iguales.

Los que siguen son casos especiales de estos resultados:

- (a) Los estadísticos $\Lambda(p, m, 1)$ y $\Lambda(1, m+1-p, p)$ son equivalentes y cada uno corresponde a un estadístico $B(\frac{1}{2}(m-p+1), \frac{1}{2}p)$ simple.
- (b) $\Lambda(p, m, 2)$ y $\Lambda(2, m+2-p, p)$ son equivalentes y corresponden al producto de un estadístico $B(\frac{1}{2}(m-p+1), \frac{1}{2}p)$ con un estadístico $B(\frac{1}{2}(m-p+2), \frac{1}{2}p)$ independiente.

A partir de las relaciones entre las variables β y F , pueden expresarse también funciones de los estadísticos $\Lambda(p, m, 1)$ y $\Lambda(p, m, 2)$ en términos de la distribución F como sigue:

- (a) Si $\Lambda \sim \Lambda(p, m, 1)$, entonces

$$\frac{1-\Lambda}{\Lambda} \sim \frac{p}{m-p+1} F_{p, m-p+1}; \quad (4.1-40)$$

- (b) Si $\Lambda \sim \Lambda(1, m, n)$, entonces

$$\frac{1-\Lambda}{\Lambda} \sim \frac{n}{m} F_{n, m}; \quad (4.1-41)$$

- (c) Si $\Lambda \sim \Lambda(p, m, 2)$, entonces

$$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \sim \frac{p}{m-p+1} F_{2p, 2(m-p+1)}; \quad (4.1-42)$$

- (d) Si $\Lambda \sim \Lambda(2, m, n)$, entonces

$$\frac{1-\sqrt{\Lambda}}{\sqrt{\Lambda}} \sim \frac{n}{m-1} F_{2n, 2(m-1)}. \quad (4.1-43)$$

Para otros valores de n y p , y se cumple que m es grande, se usa la aproximación de Bartlett:

$$-\left\{m - \frac{1}{2}(p-n+1)\right\} \ln \Lambda(p, m, n) \sim \chi_{np}^2 \quad (4.1-44)$$

asintóticamente mientras $m \rightarrow \infty$.

Definición 4.1.11.2. Suponer que $\mathbf{A} \sim W_p(\mathbf{I}, m)$ independientemente de $\mathbf{B} \sim W_p(\mathbf{I}, n)$, donde $m \geq p$. Entonces el valor característico más grande de $(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B}$, representado por θ , es conocido como el estadístico de la raíz más grande y su distribución se denota con $\theta(p, m, n)$.

La $\theta(p, m, n)$ puede definirse también como la raíz más grande de la siguiente ecuación:

$$|\mathbf{B} - \theta(\mathbf{A} + \mathbf{B})| = 0.$$

Si λ es un valor característico de $\mathbf{A}^{-1}\mathbf{B}$, entonces $\lambda/(1 + \lambda)$ es un valor característico de $(\mathbf{A} + \mathbf{B})^{-1}\mathbf{B}$; θ está dado por:

$$\theta = \frac{\lambda_1}{1 + \lambda_1}, \quad (4.1-45)$$

donde λ_1 denota el valor característico más grande de $\mathbf{A}^{-1}\mathbf{B}$. Los valores de θ están en el intervalo $0 < \theta < 1$. De la expresión:

$$\Lambda(p, m, n) = |\mathbf{I} + \mathbf{A}^{-1}\mathbf{B}|^{-1} = \prod_{i=1}^p (1 + \lambda_i)^{-1} = \prod_{i=1}^k (1 + \lambda_i)^{-1}. \quad (4.1-46)$$

y del Teorema 4.1.11.1 se obtienen las siguientes propiedades:

$$(1) \theta(p, m, n) \text{ y } \theta(n, m + n - p, p) \text{ tienen la misma distribución;} \quad (4.1-47)$$

$$(2) \frac{\theta(1, m, n)}{1 - \theta(1, m, n)} = \frac{1 - \Lambda(1, m, n)}{\Lambda(1, m, n)} \sim \frac{n}{m} F_{n, m}; \quad (4.1-48)$$

$$(3) \frac{\theta(p, m, 1)}{1 - \theta(p, m, 1)} = \frac{1 - \Lambda(p, m, 1)}{\Lambda(p, m, 1)} \sim \frac{p}{m - p + 1} F_{p, m - p + 1}; \quad (4.1-49)$$

Para $p \geq 2$ los valores críticos de θ pueden verse de los cuadros; en Mardia *et al.* (1989; pp. 494-496) están los valores críticos para $p = 2$ y para valores generales de p pueden consultarse los tablas de Pearson & Hartley (*ibid.* p.85). En el caso de los parámetros del estadístico θ , p es la dimensión, m son los g.l. del “error” y n son los g.l. de la “hipótesis”.

4.2 PRUEBAS DE HIPÓTESIS

4.2.1 Introducción

Dos de las tareas importantes en la estadística son: (a) la estimación de parámetros poblacionales y (b) las pruebas de hipótesis junto con la construcción de intervalos de confianza. En este capítulo se presenta la metodología de algunas pruebas de hipótesis multivariadas; se plantea la hipótesis, se escribe el estadístico de prueba y la regla de decisión para rechazar la hipótesis nula. Algunas de las pruebas de hipótesis univariadas tienen su extensión directa a hipótesis multivariadas, pero existen otras que son exclusivas del enfoque multivariado (Mardia *et al.*, 1989).

De acuerdo a Winer *et al.* (1991), una *hipótesis estadística* es un enunciado postulado sobre una población, el cual se busca refutar o validar con base en la información extraída de la muestra. La *prueba estadística* es un conjunto de reglas que conducen a tomar una decisión acerca de la hipótesis. Otro concepto importante es el *nivel de significación* de las pruebas de hipótesis, que se define como la probabilidad que se da de margen para rechazar la hipótesis nula cuando ésta sea verdadera. El nivel de significancia se denota, usualmente, con la letra griega minúscula alfa, α .

Mardia *et al.* (1989) señalan que en las pruebas de hipótesis multivariadas, se pueden mencionar dos problemas fundamentales: el gran número de hipótesis diferentes que existen y la dificultad de escoger entre diferentes pruebas estadísticas plausibles. Los mismos autores presentan los estadísticos de prueba de ciertas hipótesis obtenidos desde dos criterios diferentes: la prueba de razón de verosimilitud (PRV) y la prueba unión intersección (PUI). En algunas ocasiones, ambos enfoques conducen al mismo estadístico

de prueba, pero en general no ocurre así. Sobre estas pruebas se desarrollan varios ejemplos numéricos en este capítulo y se anexan los programas en SAS que se usaron para hacer los cálculos.

4.2.2 Pruebas sobre el Vector μ y la Matriz Σ

Para plantear los estadísticos de prueba (*EP*) de las siguientes tres hipótesis, usando el enfoque de PRV, suponer que la muestra aleatoria de $(X_1, X_2, \dots, X_n)' \sim N_p(\mu, \Sigma)$, y sean $\mu_0 = [\mu_{01}, \mu_{02}, \dots, \mu_{0p}]'$ un vector de medias dado y Σ_0 , una matriz de covarianzas dada.

4.2.2.1 La hipótesis $H_0: \mu = \mu_0$ cuando Σ es conocida

El estadístico de prueba (*EP*) dado por el criterio de RV de esta prueba es:

$$n(\bar{x} - \mu_0)' \Sigma^{-1} (\bar{x} - \mu_0) \sim \chi_p^2 \quad (4.2-1)$$

donde, $\bar{x} = \frac{1}{n} \mathbf{X}'\mathbf{1}$ como se dio en (4.1-4), es el estimador de máxima verosimilitud de μ y χ_p^2 es el valor de una distribución ji-cuadrada con p grados de libertad. La regla de decisión a un nivel de significancia α es: rechazar H_0 si el valor calculado del *EP* es mayor que $\chi_{p; \alpha}^2$ donde p es el número de variables aleatorias; no rechazar H_0 , de otro modo. $\square$

EJEMPLO 4.2.1 Considerar el ancho de sépalo de las tres especies de *Iris spp.* del Cuadro 3.2.1 y la hipótesis $H_0: \mu = \mu_0$, con Σ conocida, donde:

$$\mu_0 = \begin{bmatrix} 3.0 \\ 3.0 \\ 3.0 \end{bmatrix} \text{ y } \Sigma = \begin{bmatrix} 0.1 & 0.0 & 0.0 \\ 0.0 & 0.1 & 0.0 \\ 0.0 & 0.0 & 0.1 \end{bmatrix}.$$

Con esta hipótesis se está postulando que la media poblacional del ancho de sépalo de las tres especies es 3.0 y se trata de ver si los datos del ancho de sépalo de las especies de *Iris*, no contradicen la hipótesis nula. Usando la expresión (4.2-1), se tiene que:

$$EP = 50(0.428, -0.23, -0.026) \begin{bmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{bmatrix} \begin{pmatrix} 0.428 \\ -0.23 \\ -0.026 \end{pmatrix} = 118.38.$$

Como $EP = 118.38 > \chi^2_{3, 0.01} = 11.34$, se rechaza la hipótesis nula con un nivel de significancia de 0.01¹. Es decir, se concluye que el vector de medias poblacional es diferente de μ_0 . Los cálculos de esta prueba se hicieron con el Programa 4.2.6.1. ♦

4.2.2.2 La hipótesis $H_0: \mu = \mu_0$ cuando Σ es desconocida

En esta hipótesis, el EP de la PRV es:

$$(n-1)(\bar{x} - \mu_0)' S^{-1}(\bar{x} - \mu_0) \quad (4.2-2)$$

que bajo la hipótesis nula se distribuye como una T^2 de Hotelling con $(p, n-1)$ grados de libertad. Además se tiene la siguiente proporcionalidad:

$$\frac{(n-1) - p + 1}{(n-1)p} T^2_{(p, n-1)} \stackrel{H_0}{\sim} F_{p, n-p}, \quad (4.2-3)$$

por lo que, el estadístico de prueba se puede replantear como:

$$EP = \frac{n-p}{p} (\bar{x} - \mu_0)' S^{-1}(\bar{x} - \mu_0) \stackrel{H_0}{\sim} F_{p, n-p} \quad (4.2-4)$$

Usando la expresión (4.2-4), la regla de decisión es: rechazar H_0 a un nivel de significancia α , si $EP \geq F_{p, n-p; \alpha}$; de otro modo, no rechazar H_0 . □

EJEMPLO 4.2.2 Se va a plantear una hipótesis similar a la del Ejemplo 4.2.1, para los anchos de sépalo de las especies de *Iris spp*, excepto que la matriz de covarianzas no se

¹ Cuando con un nivel de significancia $\alpha=0.01$ se rechaza H_0 , usualmente se dice que el rechazo de la hipótesis nula fue "altamente significativo".

conoce. En estas condiciones podemos usar el EP de la expresión (4.2-4) de tal forma que:

$$EP = \frac{50-3}{3} (0.428, -0.23, -0.026) \begin{bmatrix} 7.252 & 1.264 & 0.055 \\ 1.264 & 10.661 & 0.888 \\ 0.055 & 0.888 & 9.886 \end{bmatrix} \begin{pmatrix} 0.428 \\ -0.23 \\ -0.026 \end{pmatrix} = 25.9995.$$

Como $EP = 25.99 > F_{3,47;0.01} = 4.23$ se debe rechazar la hipótesis nula con un nivel de significancia $\alpha = 0.01$. Para hacer los cálculos de este ejemplo se hizo uso del Programa 4.2.6.2. ♦

4.2.2.3a La hipótesis $H_0: \Sigma = \Sigma_0$, con μ desconocida

El EP de esta prueba por la RV es:

$$EP = -2 \ln \lambda = np(a - \ln g - 1) \quad (4.2-5)$$

donde a y g , son la media aritmética y la media geométrica, respectivamente, de los valores característicos de $\Sigma_0^{-1}\mathbf{S}$; en este caso, sean $d_1, d_2, \dots, d_p$ los valores característicos de $\Sigma_0^{-1}\mathbf{S}$, entonces:

$$a = \frac{\text{tr}(\Sigma_0^{-1}\mathbf{S})}{p} = \frac{\sum_{i=1}^p d_i}{p} \quad \text{y} \quad g = |\Sigma_0^{-1}\mathbf{S}|^{1/p} = \left(\prod_{i=1}^p d_i \right)^{1/p}.$$

Comentario. Del resultado que si $\mathbf{A}_{m \times n}$ y $\mathbf{B}_{n \times m}$ son cualesquiera dos matrices no nulas, los productos matriciales $\mathbf{AB}$ y $\mathbf{BA}$ tienen los mismos valores característicos diferentes de cero, se sigue que los valores característicos diferentes de cero de la matriz $\Sigma_0^{-1}\mathbf{S} = \Sigma_0^{-1}\mathbf{S}^{1/2}\mathbf{S}^{1/2}$ son iguales que los de la matriz simétrica $\mathbf{S}^{1/2}\Sigma_0^{-1}\mathbf{S}^{1/2}$, los cuales, por ese hecho, se pueden pedir en SAS —ya que por Propiedad 2.4.4.1, todos los valores característicos de una matriz simétrica son números reales.

Bajo la hipótesis nula, el EP en (4.2-5) se distribuye asintóticamente como una χ_m^2 con $m = p(p+1)/2$. La regla de decisión para una prueba aproximada de tamaño α es: rechazar H_0 si $EP \geq \chi_{m, \alpha}^2$; no rechazar de otro modo. □

EJEMPLO 4.2.3 Considerando el largo de sépalo de las tres especies de *Iris spp* del Cuadro 3.2.1, se puede plantear la hipótesis 4.2.2.3.a donde:

$$\Sigma_0 = \begin{bmatrix} 0.15 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0 & 0.35 \end{bmatrix}.$$

Para calcular el *EP* en (4.2-5), se tienen que obtener:

$$S = \begin{bmatrix} 0.122 & -0.014 & 0.029 \\ -0.014 & 0.261 & -0.046 \\ 0.029 & -0.046 & 0.396 \end{bmatrix} \text{ y } \Sigma_0^{-1}S = \begin{bmatrix} 0.812 & -0.096 & 0.196 \\ -0.058 & 1.044 & -0.182 \\ 0.084 & -0.130 & 1.132 \end{bmatrix}.$$

Los valores característicos de $\Sigma_0^{-1}S$ son $d_1 = 1.2936955$, $d_2 = 0.9298699$ y $d_3 = 0.7647706$, por lo que $\alpha = 0.996112$ y $g = 0.9864377$; usando (4.2-5) tenemos que el $EP = 1.465$, valor que se compara con una $\chi^2_{6, 0.05} = 12.59$. La conclusión a un nivel de significancia $\alpha = 0.05$ es que no se puede rechazar la hipótesis nula; es decir, la matriz de covarianza poblacional es estadísticamente igual a Σ_0 . Los cálculos de este ejemplo se hicieron con el Programa 4.2.6.3. ♦

Los *EP* presentados para las tres pruebas anteriores, cuyas distribuciones se conocen bajo la hipótesis nula, fueron obtenidos usando el criterio de la PRV (Mardia, *et al.* 1989; pp. 124-127). A continuación se mencionan los *EP* de las mismas pruebas bajo la PUI.

Los *EP* bajo la PUI para las hipótesis 4.2.2.1 y 4.2.2.2, coinciden con los obtenidos usando el criterio de la PRV (*ibíd.*, pp. 129-130). Aquí sólo se presentará el *EP* bajo la PUI de la hipótesis 4.2.2.3.

4.2.2.3b La hipótesis $H_0: \Sigma = \Sigma_0$, con μ desconocida

Bajo el enfoque de unión intersección, se rechaza H_0 si:

$$\lambda_1(\Sigma_0^{-1}S) > c_2 \text{ o } \lambda_p(\Sigma_0^{-1}S) < c_1 \quad (4.2-6)$$

donde c_1 y c_2 se escogen de tal modo que el tamaño de la prueba (nivel de significancia) sea α ; $\lambda_1(\Sigma_0^{-1}S)$ y $\lambda_p(\Sigma_0^{-1}S)$ son el valor característico más grande y más pequeño, respectivamente, de la matriz $\Sigma_0^{-1}S$. Sin embargo, es muy complicada la distribución conjunta de los valores característicos de $\Sigma_0^{-1}S$ y no se tienen sus valores tabulados, por lo que estamos imposibilitados de conocer los valores de c_1 y c_2 . Se resalta el hecho de que el *EP* bajo la PUI no es el mismo que el correspondiente a la PRV, a pesar de que ambos dependen sólo de los valores característicos de la matriz $\Sigma_0^{-1}S$; se reitera así de que, en general, los *EP* obtenidos para una misma hipótesis por cada uno de los dos criterios de pruebas de hipótesis (PRV y PUI), no son los mismos (*ibid.*, pp. 130-131). $\square$

Para presentar las pruebas de hipótesis que siguen, se deben considerar los siguientes resultados, tomados de Mardia *et al.* (1989, p. 107), sobre los estimadores de máxima verosimilitud de μ , tanto con Σ conocida como con Σ desconocida.

El estimador por máxima verosimilitud del vector μ bajo $H_0: R\mu = r$, con Σ conocida, está dado por:

$$\hat{\mu} = \bar{x} - \Sigma R'(R\Sigma R')^{-1}(R\bar{x} - r) \quad (4.2-7)$$

donde $\bar{x} = \frac{1}{n} X1$.

Con Σ desconocida, el estimador de máxima verosimilitud de μ es:

$$\hat{\mu} = \bar{x} - SR'(RSR')^{-1}(R\bar{x} - r) \quad (4.2-8)$$

4.2.3 Hipótesis con Base en Muestras Individuales

4.2.3.1 Hipótesis sobre μ con base en una sola muestra (*ibíd.*, pp. 131-138)

Bajo el criterio de la PRV, se debe asumir que $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]'$ es una muestra aleatoria proveniente de una población normal p -variada, N_p , con media μ y matriz de covarianzas Σ . Del siguiente resultado general se puede derivar la PRV para pruebas particulares sobre μ :

Resultado. Si H_0 y H_1 son hipótesis que conducen a los estimadores de máxima verosimilitud $\hat{\mu}$ y $\bar{\mathbf{x}}$, respectivamente, y si no existen restricciones sobre Σ , entonces los estimadores de máxima verosimilitud de Σ son $\mathbf{S} + \mathbf{d}\mathbf{d}'$ y $\mathbf{S}$, respectivamente, donde $\mathbf{d} = \bar{\mathbf{x}} - \hat{\mu}$. La PRV para probar H_0 contra H_1 está dado por:

$$-2 \ln \lambda = n \mathbf{d}' \Sigma^{-1} \mathbf{d} \text{ si } \Sigma \text{ es conocida} \quad (4.2-9)$$

y

$$-2 \ln \lambda = n \ln(1 + \mathbf{d}' \mathbf{S}^{-1} \mathbf{d}) \text{ si } \Sigma \text{ es desconocida} \quad (4.2-10)$$

Para el enfoque de la PUI, no existe un resultado general que involucre las pruebas que se van a escribir a continuación; por consiguiente, los *EP* se obtienen para cada caso particular; aquí se escribirán directamente las derivaciones que presentan Mardia *et al.* (1989).

4.2.3.1a La hipótesis $H_0: \mathbf{R}\mu = \mathbf{r}$ cuando Σ es conocida (hipótesis de restricciones lineales)

En esta hipótesis, la matriz $\mathbf{R}$ y el vector $\mathbf{r}$ son proporcionados. El estimador de máxima verosimilitud de μ es el de la expresión (4.2-7) y la PRV está dado por:

$$-2 \ln \lambda = n(\mathbf{R}\bar{\mathbf{x}} - \mathbf{r})'(\mathbf{R}\Sigma\mathbf{R}')^{-1}(\mathbf{R}\bar{\mathbf{x}} - \mathbf{r}) \stackrel{H_0}{\sim} \chi_q^2 \quad (4.2-11)$$

donde $q < p$ es el número de elementos de $\mathbf{r}$. Con el EP en (4.2-11), la regla de decisión para una prueba de tamaño α es: rechazar la hipótesis nula si $EP \geq \chi_{q, \alpha}^2$; no rechazar H_0 , de otro modo. $\square$

Un caso especial de la prueba 4.2.3.1a se da cuando se hace la partición $\mu = (\mu'_1, \mu'_2)'$, y se plantea:

$$H_0: \mu_1 = \mathbf{0} \quad (4.2-12)$$

Para expresarlo en la forma de la prueba 4.2.3.1a, se hace $\mathbf{R} = (\mathbf{I}, \mathbf{0})$ y $\mathbf{r} = \mathbf{0}$; y el estadístico de prueba de (4.2.11) se reduce a:

$$-2 \ln \lambda = n\bar{\mathbf{x}}'_1 \Sigma_{11}^{-1} \bar{\mathbf{x}}_1 \quad (4.2-13)$$

donde $\bar{\mathbf{x}}_1 = (\mathbf{I}, \mathbf{0})\bar{\mathbf{x}} - \mathbf{0}$ y $\Sigma_{11}^{-1} = [(\mathbf{I}, \mathbf{0})\Sigma(\mathbf{I}, \mathbf{0})']^{-1}$. La regla de decisión para una prueba de tamaño α es: rechazar la hipótesis nula escrita en (2-12) si el lado derecho de (4.2-13) es mayor que $\chi_{q, \alpha}^2$. $\square$

Otro caso particular de la prueba 4.2.3.1a, es el planteamiento de la hipótesis nula:

$$H_0: \mu = k\mu_0 \quad (4.2-14)$$

donde k es una constante y $\mu_0 = [\mu_{01}, \mu_{02}, \dots, \mu_{0p}]'$ es un vector dado. Para expresar esta prueba en términos de $H_0: \mathbf{R}\mu = \mathbf{r}$, se toma $\mathbf{r} = \mathbf{0}$ y $\mathbf{R}_{(p-1) \times p}$ con $\text{rango}(\mathbf{R}) = p - 1$ y todas las hileras de $\mathbf{R}$ ortogonales a μ_0 . Bajo $H_0: \mathbf{R}\mu = \mathbf{0}$, el estimador de máxima verosimilitud de k es $\hat{k} = \mu'_0 \Sigma^{-1} \bar{\mathbf{x}} / \mu'_0 \Sigma^{-1} \mu_0$, que puede servir para expresar el estadístico de la PRV sin usar explícitamente la matriz $\mathbf{R}$. Así:

$$EP = -2 \ln \lambda = n\bar{\mathbf{x}}' \Sigma^{-1} \left\{ \Sigma - (\mu'_0 \Sigma^{-1} \mu_0)^{-1} \mu_0 \mu'_0 \right\} \Sigma^{-1} \bar{\mathbf{x}} \quad (4.2-15)$$

se distribuye bajo H_0 como una χ^2_{p-1} . La regla de decisión con un nivel de significancia α es: rechazar $H_0: R\mu = 0$ si $EP > \chi^2_{p-1; \alpha}$; de otro modo, no rechazar H_0 . $\square$

Mardia *et al.* (1989) reportan que para probar la hipótesis 4.2.3.1a, coinciden los EP obtenidos por la PRV y por la PUI.

4.2.3.1b La hipótesis $H_0: R\mu = r$ cuando Σ es desconocida (hipótesis de restricciones lineales)

Prueba razón de verosimilitud. De (4.2-8) y (4.2-10) se tiene que $-2 \ln \lambda = n \ln(1 + d'S^{-1}d)$, donde $d = SR'(RSR')^{-1}(R\bar{x} - r)$. Para probar la hipótesis 4.2.3.1b se usa el EP :

$$T^2 = (n-1)d'S^{-1}d = (n-1)(R\bar{x} - r)'(RSR')^{-1}(R\bar{x} - r), \quad (4.2-16)$$

que bajo H_0 se distribuye como una $T^2_{(q, n-1)}$. Si se plantea la forma particular $H_0: \mu_1 = 0$, con Σ desconocida, el EP se reduce a:

$$T^2 = (n-1)\bar{x}'_1 S^{-1}_1 \bar{x}_1.$$

La regla de decisión para una prueba de tamaño α es: rechazar H_0 si el estadístico T^2 que se calcula es mayor que $T^2(q, n-1; \alpha)$

EJEMPLO 4.2.4 (*ibid.*) Con los datos del Cuadro 3.2.5 se puede plantear la siguiente hipótesis nula: *los pesos del depósito de corcho de los 28 árboles no varían en las cuatro direcciones*. Si $\mathbf{x}$ es el vector aleatorio que contiene a las cuatro variables, $\mathbf{x} = (N, E, S, W)'$, entonces una forma de representar la hipótesis nula es $H_0: R\mu = 0$, donde:

$$R = \begin{bmatrix} -1 & 1 & -1 & 1 \\ 0 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 \end{bmatrix}.$$

De esta forma, puede verse la hipótesis nula como:

$$H_0: \begin{bmatrix} (E+W)-(N+S) \\ W-E \\ S-N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

En este ejemplo, $q = 3$ es el número de elementos del vector $\mathbf{r} = \mathbf{0}$ y $n = 28$ árboles. Para probar H_0 en 4.2.3.1b se usa la expresión (4.2-16), obteniendo que el estadístico de prueba es $T^2 = 20.742$; donde el vector de medias y la matriz de covarianzas, calculados en los Ejemplos 4.1.3 y 4.1.4, respectivamente, son:

$$\bar{\mathbf{x}} = \begin{bmatrix} 50.54 \\ 46.18 \\ 49.68 \\ 45.18 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 280.03 & 215.76 & 278.14 & 218.19 \\ 215.76 & 212.08 & 220.88 & 165.25 \\ 278.14 & 220.88 & 337.50 & 250.27 \\ 218.19 & 165.25 & 250.27 & 217.93 \end{bmatrix}.$$

Usando la equivalencia $T^2(p, m) = (mp/m - p + 1)F_{p, m-p+1}$ (*ibid.*, p. 74), se tiene que el correspondiente estadístico F calculado es: $F_{cal} = 6.402$, que se compara con el punto 1% superior de una distribución F de Snedecor con 3 y 25 g.l., $F_{3,25;0.01} = 4.68$. Como $F_{cal} > F_{3,25;0.01}$, se rechazar H_0 con un nivel de significancia $\alpha = 0.01$ y se concluye que la variación de los pesos del depósito de corcho en las cuatro direcciones es altamente significativa. Los cálculos de esta prueba se hicieron con el Programa 4.2.6.4. ♦

Prueba unión intersección. En esta hipótesis, la PUI coincide también con la PRV.

4.2.3.2 Hipótesis sobre Σ con base a una sola muestra (*ibid.*, p. 133)

Resultado. Sea $\mathbf{x}_1, \dots, \mathbf{x}_n$ una muestra aleatoria de una población $N_p(\mu, \Sigma)$. Si con H_0 y H_1 , los estimadores de máxima verosimilitud de Σ son $\hat{\Sigma}$ y $\mathbf{S}$, y $\bar{\mathbf{x}}$ es el estimador de máxima verosimilitud de μ bajo H_0 y H_1 , entonces la PRV para probar H_0 contra H_1 está dada por:

$$-2 \ln \lambda = np(a - \ln g - 1), \quad (4.2-17)$$

donde a y g son las medias aritmética y geométrica, respectivamente, de los valores característicos de la matriz $\hat{\Sigma}^{-1}\mathbf{S}$.

El resultado es similar cuando μ es conocida. Tampoco se conocen resultados generales bajo la PUI.

4.2.3.2a La hipótesis $H_0: \Sigma = k\Sigma_0$, con μ desconocida

Prueba de razón de verosimilitud. El estimador de máxima verosimilitud de k es $\hat{k} = \text{tr}\Sigma_0^{-1}\mathbf{S}/p$ y la PRV está dado por la expresión (4.2-17), donde a y g están referidos a los valores característicos de $\Sigma_0^{-1}\mathbf{S}/\hat{k}$. Para este caso particular se pueden hacer las siguientes simplificaciones, $a = a_0/\hat{k}$ y $g = g_0/\hat{k}$, donde a_0 y g_0 son las medias aritmética y geométrica, respectivamente, de los valores característicos¹ de $\Sigma_0^{-1}\mathbf{S}$; también se tiene que $\hat{k} = a_0$. Por lo tanto $a = 1$ y $g = g_0/a_0$. De (4.2-17), se tiene que el EP es:

$$-2 \ln \lambda = np \ln \left(\frac{a_0}{g_0} \right), \quad (4.2-18)$$

el cual tiene una distribución asintótica χ_m^2 con $m = \frac{1}{2}p(p+1) - 1 = \frac{1}{2}(p-1)(p+2)$ grados de libertad. La regla de decisión para una prueba de tamaño α es: rechazar H_0 si el EP en (4.2-18) es mayor que $\chi_{m, \alpha}^2$; no rechazar H_0 de otro modo. $\square$

EJEMPLO 4.2.5 Para ejemplificar la prueba de hipótesis 4.2.3.2a, se puede plantear la hipótesis $H_0: \Sigma = k\mathbf{I}$ conocida como la “prueba de esfericidad”. La PRV está dada en la

¹Ver comentario de la Sección 4.2.2.3a para la obtención de los valores característicos (eigenvalores) de $\Sigma_0^{-1}\mathbf{S}$

expresión (4.2-18), donde a_0 y g_0 , en este caso, son la media aritmética y la media geométrica, respectivamente, de los valores característicos de S de tal forma que:

$$a_0 = \frac{1}{p} \sum s_{ii} \text{ y } g_0^p = |S|.$$

Para los datos del ancho de sépalo de las tres especies de *Iris spp*, $n = 50$, $p = 3$, $a_0 = 0.1131$, $|S| = 0.0013462$, $g_0 = 0.1104$. De la expresión (4.2-18), $-2 \ln \lambda = 3.5752$ que se debe comparar con el punto $100\alpha \%$ superior de una distribución ji-cuadrada con $\frac{1}{2}(p-1)(p+2) = 5$ g.l., $\chi^2_{5; \alpha}$. Como $EP = 3.5752 < \chi^2_{5; 0.05} = 11.07$, entonces no se rechaza la hipótesis nula, concluyéndose que la matriz de covarianzas poblacional es proporcional a la matriz idéntica. Los cálculos de este ejemplo se pueden hacer con el Programa 4.2.6.5. ♦

4.2.3.2b La hipótesis $H_0: \Sigma_{12} = O$, con μ desconocida (*ibid.*, pp. 135-136)

Prueba de razón de verosimilitud. Se particionan las variables en dos grupos con dimensiones p_1 y p_2 , respectivamente, de tal forma que $p_1 + p_2 = p$. Considérese la partición de $\hat{\mu} = (\hat{\mu}'_1, \hat{\mu}'_2)' = \bar{x}$, y

$$\hat{\Sigma} = \begin{bmatrix} S_{11} & O \\ O & S_{22} \end{bmatrix}.$$

Como el estimador de máxima verosimilitud de μ bajo H_0 y H_1 es $\bar{x}$, del resultado general en (4.2-17) se tiene que la PRV de la hipótesis 4.2.3.2b depende de los valores característicos de:

$$\hat{\Sigma}^{-1}S = \begin{bmatrix} S_{11} & O \\ O & S_{22} \end{bmatrix}^{-1} \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} = \begin{bmatrix} I & S_{11}^{-1}S_{12} \\ S_{22}^{-1}S_{21} & I \end{bmatrix},$$

como $\hat{\Sigma}^{-1}S$ es de dimensión $(p \times p)$ y en su diagonal principal hay puros unos, $\text{tr}(\hat{\Sigma}^{-1}S) = p$; entonces la media aritmética de sus valores característicos es $a = 1$.

Se tiene también que:

$$g^p = |\hat{\Sigma}^{-1}S| = \frac{|S|}{|S_{11}||S_{22}|} = \frac{|S_{22} - S_{21}S_{11}^{-1}S_{12}|}{|S_{22}|}, \quad (4.2-19)$$

Por lo que,

$$\begin{aligned} -2 \ln \lambda &= -n \ln \left(\frac{|S_{22} - S_{21}S_{11}^{-1}S_{12}|}{|S_{22}|} \right) \\ &= -n \ln |S_{22}^{-1}S_{22} - S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}| \\ &= -n \ln |\mathbf{I} - S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}| \\ &= -n \ln \prod_{i=1}^k (1 - \theta_i) \end{aligned} \quad (4.2-20)$$

donde las θ_i denotan los valores característicos diferentes de cero de $S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}$ y $k = \min(p_1, p_2)$. Se puede escribir $-2 \ln \lambda$ de (4.2-20) en términos de la matriz de correlaciones como $|\mathbf{I} - \mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}|$; de hecho, $S_{22}^{-1}S_{21}S_{11}^{-1}S_{12} = \mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$.

Para tener la distribución de (4.2-20) se hace lo siguiente. Sean $\mathbf{M}_{11} = n\mathbf{S}_{11}$, $\mathbf{M}_{22} = n\mathbf{S}_{22}$, y $\mathbf{M}_{22.1} = n(\mathbf{S}_{22} - \mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12})$ y úsense los resultados de que bajo H_0 , $\mathbf{M}_{22.1}$ y $\mathbf{M}_{22} - \mathbf{M}_{22.1}$ son matrices Wishart independientemente distribuidas, $\mathbf{M}_{22.1} \sim W_{p_2}(\Sigma_{22}, n-1-p_1)$ y $\mathbf{M}_{22} - \mathbf{M}_{22.1} \sim W_{p_2}(\Sigma_{22}, p_1)$; cuando $n-1 \geq p_1 + p_2$ se tiene que:

$$\lambda^{2/n} = \frac{|\mathbf{M}_{22.1}|}{|\mathbf{M}_{22.1} + (\mathbf{M}_{22} - \mathbf{M}_{22.1})|} \sim \Lambda(p_2, n-1-p_1, p_1) \quad (4.2-21)$$

De (4.2-20) puede observarse que $\lambda^{2/n} = |\mathbf{I} - \mathbf{S}_{22}^{-1}\mathbf{S}_{21}\mathbf{S}_{11}^{-1}\mathbf{S}_{12}| = \prod_{i=1}^k (1 - \theta_i)$ es el mismo que el estadístico de prueba en (4.2-21). La regla de decisión para una prueba de tamaño α es: rechazar H_0 si $\lambda^{2/n} < \Lambda(p_2, n-1-p_1, p_1; \alpha)$, el punto $100\alpha\%$ inferior de una distribución $\Lambda(p_2, n-1-p_1, p_1)$; de otro modo, no rechazar H_0 . Para los casos en que, ya sea p_1 o p_2 , valgan 1 ó 2, se puede usar la distribución exacta del estadístico Λ a través de las expresiones (4.1-40), (4.1-41), (4.1-42) o (4.1-43). Para cualesquier otros valores de p_1 y p_2 , se puede usar la siguiente aproximación de Bartlett:

$$EP = -\left(n - \frac{1}{2}(p_1 + p_2 + 3)\right) \ln |\mathbf{I} - \mathbf{S}_{22}^{-1} \mathbf{S}_{21} \mathbf{S}_{11}^{-1} \mathbf{S}_{12}|, \quad (4.2-22)$$

que para valores grandes de n se distribuye asintóticamente como una χ_m^2 con $m = p_1 p_2$. Usando (4.2-22) como EP , la regla de decisión para una prueba de tamaño α es: rechazar H_0 si $EP > \chi_{m, \alpha}^2$; de otro modo, no rechazar la hipótesis nula. $\square$

Prueba unión intersección. Si $\mathbf{x}_1$ y $\mathbf{x}_2$ son subvectores de $\mathbf{x}$, entonces la independencia entre $\mathbf{x}_1$ y $\mathbf{x}_2$ implica que los escalares $\mathbf{a}'\mathbf{x}_1$ y $\mathbf{b}'\mathbf{x}_2$ son independientes también, para vectores $\mathbf{a}$ y $\mathbf{b}$ cualesquiera. Una forma de probar si los dos escalares son independientes, es observar su coeficiente de correlación muestral, r . Observe que:

$$r^2 = \frac{[\text{cov}(\mathbf{a}'\mathbf{x}_1, \mathbf{b}'\mathbf{x}_2)]^2}{\text{var}(\mathbf{a}'\mathbf{x}_1) \text{var}(\mathbf{b}'\mathbf{x}_2)} = \frac{(\mathbf{a}'\mathbf{S}_{12}\mathbf{b})^2}{\mathbf{a}'\mathbf{S}_{11}\mathbf{a} \mathbf{b}'\mathbf{S}_{22}\mathbf{b}}.$$

La PUI basada en esta descomposición usa como estadístico al valor característico más grande de $\mathbf{S}_{11}^{-1} \mathbf{S}_{12} \mathbf{S}_{22}^{-1} \mathbf{S}_{21}$ que es igual al correspondiente a $\mathbf{S}_{22}^{-1} \mathbf{S}_{21} \mathbf{S}_{11}^{-1} \mathbf{S}_{12}$. El valor característico más grande de $\mathbf{S}_{22}^{-1} \mathbf{S}_{21} \mathbf{S}_{11}^{-1} \mathbf{S}_{12}$, que denotamos con θ_1 , tiene la distribución de la raíz característica más grande $\theta(p_2, n-1-p_1, p_1)$. La regla de decisión para una prueba de tamaño α es: rechazar H_0 si $\theta_1 \geq \theta(p_2, n-1-p_1, p_1; \alpha)$. $\square$

Comentario. Para entender con claridad la distribución de θ_1 se procede así. Se hace $\mathbf{S}_{22}^{-1} \mathbf{S}_{21} \mathbf{S}_{11}^{-1} \mathbf{S}_{12} = [\mathbf{M}_{22.1} + (\mathbf{M}_{22} - \mathbf{M}_{22.1})]^{-1} (\mathbf{M}_{22} - \mathbf{M}_{22.1})$ donde $\mathbf{M}_{22} = n\mathbf{S}_{22}$ y como $\mathbf{M}_{22.1} = n(\mathbf{S}_{22} - \mathbf{S}_{21} \mathbf{S}_{11}^{-1} \mathbf{S}_{12})$, $(\mathbf{M}_{22} - \mathbf{M}_{22.1}) = n\mathbf{S}_{21} \mathbf{S}_{11}^{-1} \mathbf{S}_{12}$. Si $\mathbf{A} = \mathbf{M}_{22.1}$ y $\mathbf{B} = (\mathbf{M}_{22} - \mathbf{M}_{22.1})$, la distribución de θ_1 se desprende directamente de la Definición 4.1.11.2. Recuerde también que si λ_1 denota al eigenvalor más grande de $\mathbf{A}^{-1}\mathbf{B}$, igual que el de la matriz simétrica $\mathbf{B}^{1/2} \mathbf{A}^{-1} \mathbf{B}^{1/2}$, entonces $\theta_1 = \lambda_1 / (1 + \lambda_1)$ es el eigenvalor más grande de $(\mathbf{A} + \mathbf{B})^{-1} \mathbf{B}$.

Concluyendo, la PUI de la hipótesis 4.2.3.2b se basa en θ_1 mientras que su respectiva PRV se basa en $\prod (1 - \theta_i)$, $i = 1, \dots, k$.

EJEMPLO 4.2.6 Con los datos del Cuadro 4.2.1 suponer que X_1 = diámetro normal, X_2 = altura a la base de la copa (altura de fuste limpio), X_3 = radio de la base de la copa y X_4 = altura total. Se desea probar la hipótesis de la Sección 4.2.3.2b bajo ambos criterios, la PRV y la PUI. Suponga que la matriz de covarianzas poblacional Σ se particiona en $\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ con todas las submatrices cuadradas. Para probar $\Sigma_{12} = \mathbf{O}$ se requieren los valores característicos de:

$$S_{22}^{-1}S_{21}S_{11}^{-1}S_{12} = \begin{bmatrix} 0.2128161 & 0.0118967 \\ 3.5319338 & 0.6322949 \end{bmatrix}$$

siendo éstos, $\theta_1 = 0.7158$ y $\theta_2 = 0.1293$. En este ejemplo $n = 31$ árboles y $p_1 = p_2 = 2$.

Usando la PRV se tiene que calcular el estadístico de prueba en (4.2-21),

$$\lambda^{2/n} = (1 - \theta_1)(1 - \theta_2) = (1 - 0.7158)(1 - 0.1293) = 0.2474$$

que para una prueba de tamaño α , se debe comparar con $\lambda(2, 28, 2)$. De (4.1-45) se sabe que si $X \sim \lambda(2, 28, 2)$ entonces $\{[(28-1)/2] [(1 - X^{1/2})/X^{1/2}]\} \sim F_{4, 54}$. Así $F_{cal} = 13.64$ se compara con $F_{4, 54; 0.01} = 3.688$ y se decide que el rechazo de H_0 es "altamente significativo".

Con la PUI, tenemos que el valor característico más grande de $S_{22}^{-1}S_{21}S_{11}^{-1}S_{12}$ es $\theta_1 = 0.7158$ que se compara con $\theta(2, 25, 2; 0.01) = 0.3915$. Por lo tanto, también se concluye rechazar la hipótesis nula con un nivel de significancia $\alpha = 0.01$. Los cálculos de este ejemplo se hicieron con el Programa 4.2.6.10.

4.2.3.2c La hipótesis $H_0: R = I$, con μ desconocida (Mardia *et al.*, 1989; p. 137)

Esta hipótesis es para probar si todas las variables están o no correlacionadas con alguna otra y es equivalente a plantear si Σ es una matriz diagonal. Bajo la hipótesis nula, la media y la varianza de cada variable se obtienen de forma individual, así que

Cuadro 4.2.1 Mediciones de 4 dimensiones en 31 coníferas

<i>sp</i>	Diám. en <i>cm</i>	Altura total en <i>m</i>	altura a la ba- se de la copa (hbc) en <i>m</i>	Radio de la base de la copa en <i>cm</i>
A	86.1	32.7	9	279.9
A	97.3	33.8	9.6	682.5
A	47.8	27.7	2.8	261.6
A	40.9	19.5	9.1	139.5
A	80.3	35.1	12.4	279.9
A	49.5	24.9	7.3	351.8
B	75.2	33.7	11.7	340.9
B	63.5	30.8	7.2	239.8
B	35.6	12.0	1.6	193.3
B	27.4	10.7	1.1	166.6
B	64.8	33.7	13.2	239.0
B	76.5	38.6	13.7	299.5
B	68.3	31.1	11.2	400.3
C	53.3	22.1	3.1	181.6
C	46.2	16.2	4.1	117.8
C	30.5	11.2	0.0	211.3
C	67.6	26.8	12.3	147.9
C	69.9	4.4	2.4	244.1
C	51.3	21.1	8.2	332.5
D	38.6	20.8	6.1	336.1
D	71.6	34.1	9.8	151.4
D	56.1	28.7	11.0	263.6
D	55.9	29.1	6.4	308.1
D	34.0	13.0	0.4	136.2
D	28.2	13.4	0.6	168.4
D	29.5	14.4	0.5	207.8
E	79.0	32.1	6.2	358.7
E	36.1	21.3	3.1	258.6
E	45.7	34.7	5.1	110.0
E	70.4	34.8	10.0	287.3
E	75.4	36.8	9.3	251.9

Fuente: Tomado de Biging & Gill (1997, p. 27)

$\hat{\mu} = \bar{x}$ y $\hat{\Sigma} = \text{diag}(s_{11}, s_{22}, \dots, s_{pp})$. A partir de (4.2-17), se tiene que la PRV se plantea en términos de la matriz de correlaciones $\mathbf{R}$ con el EP en (4.2-23):

$$EP = -2 \ln \lambda = -n \ln |\mathbf{R}|, \quad (4.2-23)$$

que bajo H_0 se distribuye asintóticamente como una χ_m^2 donde $m = \frac{1}{2}p(p+1) - p = \frac{1}{2}p(p-1)$ grados de libertad. En 1949, Box demostró que si en el EP anterior, la n es sustituida por $n^* = n - 1 - \frac{1}{6}(2p+5) = n - \frac{1}{6}(2p+11)$ será mejor la aproximación de la distribución χ^2 , quedando el siguiente EP modificado:

$$EPM = -n^* \ln|\mathbf{R}| \sim \chi_{m^*}^2 \quad (4.2-24)$$

donde $m^* = \frac{1}{2}p(p-1)$. La regla de decisión para una prueba de tamaño α es: rechazar H_0 si $EPM > \chi_{m^*, \alpha}^2$; de otro modo, no rechazar H_0 .

EJEMPLO 4.2.7 Con los datos del Cuadro 4.2.1, se desea probar $H_0: \mathbf{R} = \mathbf{I}$ vs. $H_1: \mathbf{R} \neq \mathbf{I}$ (con μ desconocida). Usando la expresión (4.2-24) se tiene que $EPM = 65.98$ se debe comparar con $\chi_{6, 0.01}^2 = 16.81$. Por lo tanto, con un nivel de significancia $\alpha = 0.01$ se rechaza $H_0: \mathbf{R} = \mathbf{I}$; es decir, se rechaza que las variables no están correlacionadas. Los cálculos de este ejemplo se hicieron con el Programa 4.2.6.11.

4.2.4 Hipótesis sobre Muestras Múltiples (*ibid.*, pp. 138-141)

4.2.4.1 La hipótesis $H_{0b}: \mu_1 = \mu_2 = \dots = \mu_k$ dado que $\Sigma_1 = \Sigma_2 = \dots = \Sigma_k$

Esta hipótesis es útil en el análisis de varianza multivariado de sólo un criterio de clasificación (véase la Sección 4.4.2) para probar la igualdad de los vectores de medias poblacionales μ_j , $j = 1, \dots, k$, sobre k muestras independientes, bajo el supuesto de que sus matrices de covarianzas asociadas son iguales.

4.2.4.1.1 Prueba de razón de verosimilitud. Bajo la hipótesis nula H_{0b} los estimadores de máxima verosimilitud son $\bar{\mathbf{x}}$ y $\mathbf{S}$, ya que puede considerarse que todas las observaciones constituyen una sola muestra aleatoria. Bajo la hipótesis alternativa, el estimador de máxima verosimilitud del vector de medias poblacional de la j -ésima

muestra, μ_j , es $\bar{x}_j$ y el estimador de máxima verosimilitud de la matriz común de covarianzas es $n^{-1}\mathbf{W}$; en esta última expresión, $\mathbf{W} = \sum n_j \mathbf{S}_j$ es la matriz de las sumas de cuadrados y productos cruzados (SCPC) “dentro de muestras” y $n = \sum n_j$, donde n_j es el número de unidades (o individuos) que conforman la j -ésima muestra. Por la forma como se calcula, a $n^{-1}\mathbf{W}$ se le puede nombrar como la matriz de covarianzas “combinada”. Por otra parte, $\mathbf{T} = n\mathbf{S}$ es la matriz de las SCPC “total” que se deriva considerando todas las matrices de datos como si provinieran de una sola muestra. Con todo lo anterior, la PRV de esta hipótesis está dada por:

$$\lambda_b = \left(\frac{|\mathbf{W}|}{|n\mathbf{S}|} \right)^{n/2} = |\mathbf{T}^{-1}\mathbf{W}|^{n/2}. \quad (4.2-25)$$

{Comentario para el cálculo de $\mathbf{S}_j$ y $\mathbf{S}$. Suponer que en cada una de las k muestras independientes se toman p variables y que los valores de estas variables se arreglan en las matrices de datos $\mathbf{X}_1, \dots, \mathbf{X}_j, \dots, \mathbf{X}_k$, de dimensiones $(n_1 \times p), \dots, (n_j \times p), \dots, (n_k \times p)$, respectivamente. De acuerdo a (4.1-7),

$$\mathbf{S}_j = n_j^{-1} \mathbf{X}_j' \mathbf{H}_j \mathbf{X}_j, \text{ con } \mathbf{H}_j = \mathbf{I}_{n_j} - n_j^{-1} \mathbf{1}\mathbf{1}'.$$

En cambio, para el cálculo de $\mathbf{S}$ se usa directamente la fórmula (4.1-7), donde $\mathbf{X}$ está conformada por el arreglo vertical de las k matrices de datos, igual al número de muestras (grupos), i.e. $\mathbf{X} = (\mathbf{X}_1', \dots, \mathbf{X}_k')'$; a $\mathbf{S}$ se le puede denominar como la matriz de covarianzas “conjunta”.

También se tiene que la matriz que contiene las SCPC “entre grupos” es:

$$\mathbf{K} = \mathbf{T} - \mathbf{W} = \sum_{j=1}^k n_j (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})' \quad (4.2-26)$$

A partir de la expresión (4.2-25), se puede escribir:

$$\lambda_b^{2/n} = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{K}|} = |\mathbf{I} + \mathbf{W}^{-1}\mathbf{K}|^{-1}. \quad (4.2-27)$$

La matriz $\mathbf{W}^{-1}\mathbf{K}$ es una generalización de la relación $((n-k) SC(tratamientos))/((k-1) SC(Error))$ en el caso univariado, y su valor tiende a cero si la hipótesis nula es verdadera.

Mardia *et al.* (1989, p.139) demuestran que, bajo la hipótesis nula H_{0b} , si $n \geq p + k$ entonces:

$$|\mathbf{I} + \mathbf{W}^{-1}\mathbf{K}|^{-1} \sim \Lambda(p, n-k, k-1) \quad (4.2-28)$$

La regla de decisión a un nivel de significancia α es: rechazar H_{0b} si $|\mathbf{I} + \mathbf{W}^{-1}\mathbf{K}|^{-1} \leq \Lambda(p, n-k, k-1; \alpha)$; de otro modo, H_{0b} no se rechaza. Recuerde que $\Lambda(p, n-k, k-1; \alpha)$ es el punto $100\alpha\%$ inferior de una distribución lambda de Wilks con parámetros p , $n-k$ y $k-1$.

EJEMPLO 4.2.8 En el Cuadro 3.2.2 se cuenta con los datos de tres variables tomadas en dos tipos de sitios de pájaros helecho: sitios de anidamiento y sitios escogidos aleatoriamente en la misma área (sitios disponibles). Se desea comparar si son similares los vectores de medias poblacionales en las dos condiciones. En este ejemplo, $p = 3$ variables, $k = 2$ muestras, $n = n_1 + n_2 = 24 + 24 = 48$.

Usando la expresión (4.2-28), se calcula el estadístico de prueba:

$$\Lambda = |\mathbf{I} + \mathbf{W}^{-1}\mathbf{K}|^{-1} = 0.6291764$$

A través de la expresión (4.1-40), se tiene que $EP = 8.64$ debe compararse con $F_{3, 44; 0.01}$. Como $EP = 8.64 > F_{3, 44; 0.01} = 4.26$ se rechaza la hipótesis nula con un nivel de significancia $\alpha = 0.01$ y se concluye que los vectores de medias poblacionales de los sitios de anidamiento y los sitios escogidos aleatoriamente en la misma área no son iguales.

Para obtener el estadístico de prueba de la expresión (4.2-28), se calcularon vectores de medias muestrales, matrices de covarianzas muestrales y matrices de sumas de cuadrados y productos cruzados (SCPC). En este ejemplo, los vectores de medias de los sitios de anidamiento, los sitios aleatorios (disponibles), y el vector de medias “conjunto” son:

$$\bar{x}_1 = \begin{bmatrix} 0.733 \\ 16.250 \\ 4.037 \end{bmatrix}, \quad \bar{x}_2 = \begin{bmatrix} 0.495 \\ 12.750 \\ 2.956 \end{bmatrix} \quad \text{y} \quad \bar{x} = \begin{bmatrix} 0.614 \\ 14.500 \\ 3.496 \end{bmatrix},$$

respectivamente. Para calcular el vector de medias “conjunto”, se arreglaron las observaciones de las tres variables de los sitios disponibles abajo de las observaciones correspondientes a los sitios de anidamiento como si constituyeran una sola muestra, y se usó la expresión (4.1-4). Las matrices de covarianzas de los sitios de anidamiento, los sitios disponibles y la matriz de covarianzas “conjunta” son:

$$S_1 = \begin{bmatrix} 0.074 & -0.155 & 0.130 \\ -0.155 & 11.125 & -0.394 \\ 0.130 & -0.394 & 1.798 \end{bmatrix}, \quad S_2 = \begin{bmatrix} 0.021 & 0.170 & -0.004 \\ 0.170 & 18.708 & 1.483 \\ -0.004 & 1.483 & 0.700 \end{bmatrix},$$

$$S = \begin{bmatrix} 0.062 & 0.216 & 0.127 \\ 0.216 & 17.980 & 1.490 \\ 0.127 & 1.490 & 1.541 \end{bmatrix},$$

respectivamente. El estimador de máxima verosimilitud de la matriz de covarianzas común a las poblaciones es:

$$n^{-1}W = \begin{bmatrix} 0.047 & 0.008 & 0.063 \\ 0.008 & 14.917 & 0.544 \\ 0.063 & 0.544 & 1.249 \end{bmatrix}.$$

Recuerde que $n^{-1}W = n^{-1}(n_1S_1 + n_2S_2)$ es la matriz de covarianzas “combinada”.

Finalmente, las matrices de las SCPC “dentro de grupos” y “entre grupos” son:

$$\mathbf{W} = \begin{bmatrix} 2.274 & 0.363 & 3.012 \\ 0.363 & 716.000 & 26.120 \\ 3.012 & 26.120 & 59.956 \end{bmatrix} \text{ y } \mathbf{K} = \begin{bmatrix} 0.679 & 9.993 & 3.086 \\ 9.993 & 147.00 & 45.395 \\ 3.086 & 45.395 & 14.018 \end{bmatrix}$$

Todos los cálculos anteriores se hicieron con el Programa 4.2.6.6. ♦

4.2.4.1.2 Prueba unión intersección. El análogo univariado de la hipótesis 4.2.4.1 se prueba usando el estadístico:

$$\frac{\sum_{j=1}^k n_j (\bar{x}_j - \bar{x})^2}{\sum_{j=1}^k n_j s_j^2} \sim F_{k-1, n-k} \quad (4.2-29)$$

donde $\bar{x}$ es la media general. La fórmula correspondiente para una combinación lineal $\mathbf{Xa}$ es $\sum n_j (\mathbf{a}'(\bar{x}_j - \bar{x}))^2 / \sum n_j \mathbf{a}' \mathbf{S}_j \mathbf{a}$ cuyo valor máximo es el valor característico más grande de $\mathbf{W}^{-1} \mathbf{K}$, que es el que se usa como estadístico de prueba de esta hipótesis bajo el enfoque unión intersección. Si λ_1 es el eigenvalor más grande de $\mathbf{W}^{-1} \mathbf{K}$, entonces $\lambda_1 / (1 + \lambda_1) \sim \theta(p, n-k, k-1)$ y el criterio para rechazar la hipótesis 4.2.4.1 usando un nivel de significancia α es si el valor calculado de $\lambda_1 / (1 + \lambda_1)$ es mayor que $\theta(p, n-k, k-1; \alpha)$. Para valores de $p=2$ se pueden encontrar los valores de $\theta(p, n-k, k-1; \alpha)$ en la Tabla C.4 (*ibid.*, pp. 494-496) y para valores de $p=1$ son útiles las expresiones (4.1-48) y (4.1-49). Para $p > 2$ se pueden usar las Tablas de Pearson y Hartley citadas por Mardia *et al.* (1989). □

4.2.4.1.3 Prueba T^2 de Hotelling para dos muestras. Cuando $k=2$ muestras se trata de probar si $H_0: \mu_1 = \mu_2$; haciendo $\delta = \mu_1 - \mu_2$, podemos equivalentemente plantear la hipótesis:

$$H_{0b}: \delta = 0 \quad \text{versus} \quad H_{1b}: \delta \neq 0 \quad (4.2-30)$$

En este caso, la matriz de la SCPC “entre grupos” se reduce a

$$\mathbf{K} = n_1(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}})(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}})' + n_2(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}})(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}})'$$

Si hacemos $\mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$, se tiene que $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}} = (n_2/n)\mathbf{d}$ y $\bar{\mathbf{x}}_2 - \bar{\mathbf{x}} = -(n_1/n)\mathbf{d}$. Por lo tanto,

$$\mathbf{K} = \frac{n_1 n_2}{n} \mathbf{d} \mathbf{d}' \quad (4.2-31)$$

y

$$\begin{aligned} |\mathbf{I} + \mathbf{W}^{-1} \mathbf{K}| &= |\mathbf{I} + (n_1 n_2 / n) \mathbf{W}^{-1} \mathbf{d} \mathbf{d}'| \\ &= 1 + (n_1 n_2 / n) \mathbf{d}' \mathbf{W}^{-1} \mathbf{d} \end{aligned} \quad (4.2-32)$$

que es proporcional al estadístico T^2 de Hotelling para dos muestras. La regla de decisión para un nivel de significancia α es: rechazar $H_{0b}: \delta = 0$ si el estadístico T^2 calculado es mayor que $T^2(p, n-2; \alpha)$. $\square$

Comentario: El segundo término de (4.2-32), $(n_1 n_2 / n) \mathbf{d}' \mathbf{W}^{-1} \mathbf{d}$, es proporcional al estadístico T^2 de Hotelling para dos muestras; de hecho $T^2 = (n-2)(n_1 n_2 / n) \mathbf{d}' \mathbf{W}^{-1} \mathbf{d}$ sobre el cual se puede basar la prueba, ya que bajo $H_0: \mu_1 = \mu_2$, $T^2 \sim T^2(p, n-2)$. Para mayores detalles puede revisarse la Sección 4.1.10.

EJEMPLO 4.2.9 Como en el Ejemplo 4.2.8 se compararon los vectores de medias de dos poblaciones, se pudo haber usado alternativamente el estadístico T^2 . En este ejemplo $T^2 = 27.145$; usando para $p = 3$ y $m = 46 = n - 2$, la siguiente proporcionalidad:

$$T^2(p, m) = \{mp / (m - p + 1)\} F_{p, m-p+1},$$

tenemos que $F_{cal} = \frac{27.11145(44)}{(46)(3)} = 8.64423$, el cual coincide con el valor del EP del

Ejemplo 4.2.8, y se debe comparar con $F_{3,44;0.01} = 4.26$ llegando a la misma conclusión del ejemplo antes citado al rechazar $H_{0b}: \delta = 0$. Los cálculos se hicieron con el Programa 4.2.6.6. $\blacklozenge$

Por otro lado, el rango de $W^{-1}K$ es uno y su valor característico más grande es:

$$(n_1 n_2 / n) d' W^{-1} d \quad (4.2-33)$$

por lo que, para el caso particular de dos muestras, los EP obtenidos bajo la PRV y bajo la PUI, coinciden también. $\square$

4.2.4.2 La prueba de homogeneidad de covarianzas

Se desea inferir sobre la hipótesis $H_{0a}: \Sigma_1 = \dots = \Sigma_k$. Para probar esta hipótesis nula contra la hipótesis alternativa $H_{1a}: \Sigma_i \neq \Sigma_j$ para alguna $i \neq j$, $i, j = 1, \dots, k$, se tiene el EP siguiente:

$$-2 \ln \lambda_a = n \ln |S| - \sum n_i \ln |S_i| = \sum n_i \ln |S_i^{-1} S| \quad (4.2-34)$$

que se distribuye asintóticamente como una ji-cuadrada con $\frac{1}{2} p(p+1)(k-1)$ grados de libertad. Aquí $S = n^{-1}W$ es el estimador de máxima verosimilitud de la matriz de covarianzas común Σ bajo la hipótesis nula y S_i es el estimador de máxima verosimilitud de Σ_i bajo la hipótesis alternativa. Cuando n_i es pequeño, (4.2-34) da una ponderación alta a la contribución de S , consideración que condujo a Box a proponer en 1949, el estadístico M :

$$M = \gamma \sum (n_j - 1) \ln |S_{uj}^{-1} S_u|, \quad (4.2-35)$$

donde,

$$\gamma = 1 - \frac{2p^2 + 3p - 1}{6(p+1)(k-1)} \left(\sum \frac{1}{n_j - 1} - \frac{1}{n - k} \right),$$

y

$$S_u = \frac{n}{n-k} S, \quad S_{uj} = \frac{n_j}{n_j - 1} S_j$$

El estadístico M de Box tiene una distribución asintótica χ_m^2 con $m = \frac{1}{2}p(p+1)(k-1)$. La regla de decisión para una prueba de tamaño α es rechazar H_{0a} si $M > \chi_{m, \alpha}^2$. Se recomienda $n_j \geq 20$ y $k, p \leq 5$. $\square$

Para el caso univariado, cuando $p = 1$, de (4.2-34) se obtiene el estadístico de la prueba de homogeneidad de Bartlett:

$$n \ln s^2 - \sum n_i \ln s_i^2, \quad (4.2-36)$$

donde s_i^2 es la varianza de la muestra i y s^2 es la estimación de la varianza combinada. La regla de decisión a un nivel de significancia α es rechazar la hipótesis nula si el EP en (4.2-36) es mayor que $\chi_{k-1, \alpha}^2$. $\square$

EJEMPLO 4.2.10 En el Ejemplo 4.2.8 se estimaron las matrices de covarianzas para los sitios de anidamiento y los sitios disponibles, S_1 y S_2 , respectivamente, así como la matriz $S = n^{-1}W$ donde W es la matriz de las sumas de cuadrados y productos cruzados “dentro grupos”. A partir de estas matrices se pueden obtener los estimadores insesgados $S_{u1} = (24/23)S_1$, $S_{u2} = (24/23)S_2$ y $S_u = (48/46)S$, cuyos determinantes son:

$$|S_{u1}| = 1.428, |S_{u2}| = 0.230, |S_u| = 0.810.$$

Con estos datos se puede probar la hipótesis de que las matrices de covarianzas poblacionales (de los sitios de anidamiento y los sitios disponibles) son similares. En este ejemplo, $k = 2$ grupos, $p = 3$ variables y de la expresión (4.2-35), $\gamma = 0.9529$ y $M = 20.7774$ que se compara contra una $\chi_{6, 0.01}^2 = 16.81$. Como $M > \chi_{6, 0.01}^2$ se debe rechazar H_0 y se concluye con un nivel de significancia $\alpha = 0.01$ que las matrices de covarianzas poblacionales de los dos grupos no son iguales. Los cálculos para este ejemplo fueron obtenidos con el Programa 4.2.6.7. $\blacklozenge$

EJEMPLO 4.2.11 (*ibid.*, p. 141) En el Cuadro 3.2.4, se reportan las mediciones del largo, ancho y altura de la coraza de 48 tortugas, que están en dos grupos: 24 hembras y 24 machos. Con estos datos, se puede plantear la siguiente hipótesis nula: *la matriz de covarianzas poblacional de las hembras es igual a la matriz de covarianzas poblacional de los machos*. Igual que en el ejemplo anterior, $k = 2$, $p = 3$, $n_1 = 24$, $n_2 = 24$ y $n = 48$; haciendo cálculos similares a los del Ejemplo 4.2.10, se tiene que $\gamma = 0.9529$ y $M = 23.998$. Como $M = 23.998 > \chi^2_{6;0.01} = 16.81$, entonces se rechaza la hipótesis nula con un nivel de significancia $\alpha = 0.01$ y se concluye que las matrices de covarianzas poblacionales de machos y hembras no son iguales. Es lo mismo a decir que no se puede asumir que la variación de estas tres dimensiones en los machos y en las hembras son similares.

En este ejemplo, los vectores de medias y las matrices de covarianzas para hembras y machos que se calcularon con el Programa 4.2.6.8 son:

$$\bar{\mathbf{x}}_1 = \begin{bmatrix} 136.04 \\ 102.58 \\ 52.04 \end{bmatrix}, \quad \mathbf{S}_1 = \begin{bmatrix} 432.71 & 259.68 & 159.04 \\ 259.68 & 164.58 & 97.60 \\ 159.04 & 97.60 & 62.04 \end{bmatrix}$$

$$\bar{\mathbf{x}}_2 = \begin{bmatrix} 113.38 \\ 88.29 \\ 40.71 \end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix} 132.98 & 75.85 & 35.82 \\ 75.85 & 47.96 & 20.75 \\ 35.82 & 20.75 & 10.79 \end{bmatrix},$$

respectivamente. Y el estimador de máxima verosimilitud de la matriz de covarianzas común (matriz de covarianzas combinada) es:

$$\mathbf{S}_{comb} = (n_1 \mathbf{S}_1 + n_2 \mathbf{S}_2) / n = \begin{bmatrix} 282.85 & 167.77 & 97.43 \\ 167.77 & 106.27 & 59.18 \\ 97.43 & 59.18 & 36.41 \end{bmatrix}.$$

Así también, los determinantes de las matrices de covarianzas de hembras, de machos y combinada son:

$$|\mathbf{S}_{u1}| = 13105.511, \quad |\mathbf{S}_{u2}| = 795.072, \quad |\mathbf{S}_u| = 5580.835,$$

respectivamente. ♦

4.2.4.3 La prueba de homogeneidad completa

En esta sección se prueba la hipótesis nula $H_{0c}: \mu_1 = \dots = \mu_k$ y $\Sigma_1 = \dots = \Sigma_k$. Con la hipótesis H_{0b} de la Sección 4.2.4.1 se prueba la igualdad de los vectores de medias pero suponiendo que las matrices de covarianzas asociadas son iguales; con la hipótesis H_{0a} de la Sección 4.2.4.2, se prueba la igualdad de las matrices de covarianzas (es decir, se prueba lo que en la Sección 4.2.4.1 se supone válido). Ahora bien, en la prueba de homogeneidad completa se prueban, simultáneamente, ambas hipótesis.

Prueba de razón de verosimilitud. El *EP* por este enfoque es:

$$-2 \ln \lambda_c = -2 \ln \lambda_a - 2 \ln \lambda_b = n \ln |n^{-1} \mathbf{W}| - \sum n_j \ln |\mathbf{S}_j|, \quad (4.2-37)$$

que tiene una distribución asintótica χ_m^2 con $m = \frac{1}{2} p(k-1)(p+3)$. La regla de decisión para una prueba de tamaño α es: rechazar H_{0c} si $-2 \ln \lambda_c > \chi_{m, \alpha}^2$; de otro modo, no rechazar H_{0c} . $\square$

4.2.5 Intervalos de Confianza Simultáneos (*ibid.*, pp. 144-147)

4.2.5.1 El caso de la T^2 de Hotelling para una muestra

Sea,

$$t_a^2 = (n-1) \frac{\mathbf{a}'(\bar{\mathbf{x}} - \boldsymbol{\mu})(\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{a}}{\mathbf{a}' \mathbf{S} \mathbf{a}}. \quad (4.2-38)$$

Si se maximiza t_a^2 con respecto a $\mathbf{a}$, se obtiene $T^2 = (n-1)(\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu})$; de ahí, se tiene que:

$$t_a^2 \leq T^2 \quad \forall \text{ vector } \mathbf{a} \text{ de tamaño } p \quad (4.2-39)$$

donde T^2 es el estadístico T^2 de Hotelling para una muestra.

Se tiene también que T^2 es proporcional a una distribución F de Snedecor, por lo que, el límite superior al $100\alpha\%$ de T^2 es:

$$\Pr(T^2 \geq T_\alpha^2) = \alpha$$

donde

$$T_\alpha^2 = \{(n-1)p/(n-p)\} F_{p, n-p; \alpha}$$

siendo $F_{p, n-p; \alpha}$, el punto superior al $100\alpha\%$ de esta distribución. Si $T^2 \leq T_\alpha^2$, implica que:

$$t_a^2 \leq T_\alpha^2 \quad \forall \text{ vector } a \text{ } p\text{-dimensional} \quad (4.2-40)$$

A partir de lo anterior, se tiene que el intervalo de confianza al $100(1-\alpha)\%$ para $a'\mu$ es:

$$(a'\bar{x} - b, a'\bar{x} + b), \quad (4.2-41)$$

donde,

$$b = \left[\left(\frac{p}{n-p} a' S a \right) F_{p, n-p; \alpha} \right]^{1/2}.$$

El intervalo (4.2-41), se deriva de la expresión:

$$\Pr\{a'\mu \in (a'\bar{x} - b, a'\bar{x} + b) \quad \forall \text{ } a_{p \times 1}\} = 1 - \alpha, \quad (4.2-42)$$

la cual significa que todos los intervalos de confianza de la combinación lineal $a'\mu$ contruidos para diferentes valores de a , tienen un $100(1-\alpha)\%$ de confianza de ser verdaderos; por esto se les denomina intervalos de confianza simultáneos de $a'\mu$ para toda a . $\square$

Mardia *et al.* (1989), agregan que:

- (a) Para un valor fijo del vector $\mathbf{a}$, se pueden construir intervalos de confianza al $100(1-\alpha)\%$ para $\mathbf{a}'\mu$ a partir de $t_{\mathbf{a}} = \mathbf{a}'(\bar{\mathbf{x}} - \mu) / [\mathbf{a}'\mathbf{S}\mathbf{a}/(n-1)]^{1/2}$ que se distribuye como una t de Student con $n-1$ g.l.; esto se obtiene poniendo en la expresión (4.2-41), $b = [\{\mathbf{a}'\mathbf{S}\mathbf{a}/(n-1)\} F_{1, n-1; \alpha}]^{1/2}$ y haciendo uso de la equivalencia $F_{1, n-1} = (t_{n-1})^2$.
- (b) Se pueden usar los intervalos de confianza simultáneos para hipótesis más específicas sobre μ . Por ejemplo, la hipótesis $H_0: \mu = \mathbf{0}$, no se rechaza al nivel de significancia α si y sólo si todos los intervalos de confianza simultáneos al $100(1-\alpha)\%$ contienen a cero. Si existe al menos un vector $\mathbf{a}$, que haga que el intervalo de confianza para $\mathbf{a}'\mu$ no incluya al cero, entonces se rechaza $H_0: \mu = \mathbf{0}$. El método permite detectar qué combinación lineal en particular conduce al rechazo de la hipótesis nula.

EJEMPLO 4.2.12 (*ibid.*, p.145) Con los depósitos de corcho de árboles en 4 direcciones, se pueden generar las siguientes tres combinaciones lineales ortogonales:

$$y_1 = E + W - N - S, \quad y_2 = W - E, \quad y_3 = S - N.$$

Para obtener las observaciones de las variables transformadas (y_1, y_2 y y_3) con los datos del Cuadro 3.3.4 se hace $\mathbf{Y} = \mathbf{X}\mathbf{R}'$ donde $\mathbf{R}$ es la que se usó en el Ejemplo 4.2.4. Se asume que la matriz de datos $\mathbf{Y}$ proviene de una población $N_3(\mathbf{v}, \mathbf{R}\Sigma\mathbf{R}')$, donde $\mathbf{v} = \mathbf{R}\mu$. En el Ejemplo 4.2.4 y usando estos mismos datos, se rechazó la hipótesis $H_0: \mathbf{v} = \mathbf{0}$ por lo que convendría indagar qué media pudo haber propiciado el rechazo de la hipótesis nula. El vector de medias y la matriz de covarianzas de los datos transformados son:

$$\bar{\mathbf{y}} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \bar{y}_3 \end{bmatrix} = \begin{bmatrix} -8.857 \\ -1.000 \\ -0.857 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} 124.122 & -25.964 & -20.270 \\ -25.964 & 99.50 & 26.964 \\ -20.270 & 26.964 & 61.265 \end{bmatrix}.$$

En este ejemplo se tiene que $n = 28$ árboles, $p = 3$ variables transformadas, $F_{3,25;0.01} = 4.68$. Si $\mathbf{a} = (1,0,0)'$ entonces $\mathbf{a}'\mathbf{S}\mathbf{a} = s_{11} = 124.122$; así que para construir un intervalo de confianza para ν_1 a partir de la expresión (4.2-41), sólo falta calcular:

$$b = \left[\left(\frac{p}{n-p} \mathbf{a}'\mathbf{S}\mathbf{a} \right) F_{p,n-p;\alpha} \right]^{1/2} = \left[\left(\frac{3}{25} (124.122) \right) 4.68 \right]^{1/2} = 8.345.$$

Así, de la expresión (4.2-41) el intervalo de confianza al 99% para la media poblacional de la combinación lineal y_1 , ν_1 , está dado por:

$$(\mathbf{a}'\bar{\mathbf{x}} - b, \mathbf{a}'\bar{\mathbf{x}} + b) = (-8.857 - 8.345, 8.857 + 8.345) = (-17.20, -0.51)$$

De manera similar, se construyeron los intervalos de confianza simultáneos para ν_2 y ν_3 , resultando:

$$-8.47 < \nu_2 < 6.47 \quad \text{y} \quad -6.72 < \nu_3 < 5.00.$$

Obsérvese que el único intervalo que no contiene al cero es el que se construyó para ν_1 ; entonces puede concluirse que es esta media que propició particularmente el rechazo de la hipótesis nula en el Ejemplo 4.2.4. Los cálculos requeridos en este ejemplo se hicieron con el Programa 4.2.6.9. ♦

4.2.5.2. El caso de la T^2 de Hotelling para dos muestras

Se consideran dos muestras de tamaño n_1 y n_2 , respectivamente, con sus correspondientes vectores de medias muestrales, $\bar{\mathbf{x}}_1$ y $\bar{\mathbf{x}}_2$; así como en la Sección (4.2.4.1.3), sea $\boldsymbol{\delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ y $\mathbf{d} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$. Los intervalos de confianza simultáneos al $100(1 - \alpha)\%$ para la combinación lineal $\mathbf{a}'\boldsymbol{\delta}$ están dados por:

$$\mathbf{a}'\mathbf{d} \pm \left\{ \mathbf{a}'(n_1\mathbf{S}_1 + n_2\mathbf{S}_2)\mathbf{a} \frac{p(n_1 + n_2)}{n_1 n_2 m} F_{p,m;\alpha} \right\}^{1/2}, \quad (4.2-43)$$

donde $m = n_1 + n_2 - p - 1$. □

4.2.6 Anexo de programas en SAS

Se escriben los programas en SAS usando el procedimiento IML (Interactive Matrix Language), para resolver los diferentes ejemplos de este capítulo. De manera general debe recordarse que cada instrucción en una línea de programa SAS, termina con punto y coma (;). Así también, nótese que la transpuesta de una matriz, en IML se indica con una tilde inclinada de izquierda a derecha y de arriba hacia abajo ($'$); es decir, si A es una matriz $(m \times n)$, para representar su transpuesta de orden $(n \times m)$ en un programa SAS\IML, se escribe A' —SAS no acepta A' como, usualmente, se denota la transpuesta de la matriz A en los libros de álgebra de matrices. Finalmente, los comentarios que están entre $/* \dots */$ no forman parte del programa pero se escriben para guiarse de lo que se hace en cada línea del mismo; observe también, que no se usa la tilde del acento ortográfico en español ($'$) dentro de los comentarios de los programas, puesto que no es aceptada por SAS.

Programa 4.2.6.1 Con los datos del ancho de sépalo de las tres especies de *Iris spp* del Cuadro 3.2.1, en el Ejemplo 4.2.1, se probó la hipótesis de la Sección 4.2.2.1. Para calcular el estadístico de prueba correspondiente [expresión (4.2-1)] se usó este programa; en la segunda línea del programa (INFILE 'C:\HIPOTESIS\IRIS.DAT';) se llama el archivo IRIS.DAT que está en el subdirectorío HIPOTESIS, y después de la palabra INPUT de la tercera línea, están escritas las variables que conforman el archivo en cuestión.

```
DATA IRIS;
INFILE 'C:\HIPOTESIS\IRIS.DAT';
INPUT X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12;
/*X1,X5,X9: longitud de sepalo de I. setosa, I. versicolour e I. virginica, respectivamente*/
/*X2,X6,X10: ancho de sepalo de I. setosa, I. versicolour e I. virginica, respectivamente*/
/*X3,X7,X11: longitud de petalo de I. setosa, I. versicolour e I. virginica, respectivamente*/
/*X4,X8,X12: ancho de petalo de I. setosa, I. versicolour e I. virginica, respectivamente*/
PROC IML;
USE IRIS; START PRUEBA;                               /*comienza el modulo PRUEBA*/
READ ALL VAR {X2 x6 x10} INTO X; /*arregla los valores del ancho de sepalo de las tres
                                         especies de Iris en una matriz X*/
```



```

N=NROW(X); /*numero de individuos por especie*/
p=NCOL(X); /*numero de hileras de la matriz X*/
J=J(N,1,1); /*un vector de puros unos de orden n */
MED=(1/N)*X`*J; /*vector de medias del ancho de sepalo por especie*/
PRINT MED;
DIFE={0.428,-0.23,-0.026}; /*diferencia de MED menos mu cero*/
SIG={0.1 0 0,0 0.1 0,0 0 0.1}; /*matriz de covarianza conocida*/
INVE=INV(SIG); /*inversa de la matriz de covarianza conocida*/
EP=N*DIFE`*INVE*DIFE; /*calcula el estadistico de prueba del Ejemplo 2.1*/
PRINT DIFE SIG INVE EP; /*imprime resultados*/
FINISH; /*termina el modulo PRUEBA */
RUN PRUEBA /*correr el modulo PRUEBA */
QUIT;

```

Programa 4.2.6.2 Usando los datos del ancho de sépalo de las tres especies de *Iris spp* del Cuadro 3.3.1, se probó en el Ejemplo 4.2.2 la hipótesis de la Sección 4.2.2.2, que únicamente se diferencia con la del Ejemplo 4.2.1, porque la matriz de covarianzas es desconocida. El estadístico de prueba en este caso es el de la expresión (4.2-4) y se calculó con este programa.

```

DATA IRIS;
INFILE 'C:\HIPOTESIS\IRIS.DAT';
INPUT X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12;
/*X1,X5,X9: longitud de sepalo de I. setosa, I.versicolour e I. virginica, respectivamente*/
/*X2,X6,X10: ancho de sepalo de I. setosa, I.versicolour e I. virginica, respectivamente*/
/*X3,X7,X11: longitud de petalo de I. setosa, I.versicolour e I. virginica, respectivamente*/
/*X4,X8,X12: ancho de petalo de I. setosa, I.versicolour e I. virginica, respectivamente*/
PROC IML;
USE IRIS;
START PRUEBA; /*comienza el modulo PRUEBA */
READ ALL VAR {x2 x6 x10} INTO X; /*arreglan los valores del ancho de sepalo de las tres
especies de Iris en una matriz X*/

N=NROW(X); /*numero de hileras de la matriz X*/
p=NCOL(X); /*numero de variables por cada especie de Iris*/
J=J(N,1,1); /*un vector de puros unos de orden n=n1=n2=n3 */
MED=(1/N)*X`*J; /*vector de medias*/
H=I(N)-(1/N)*J*J`; /*matriz centrante*/
COV=(1/N)*X`*H*X; /*matriz de covarianzas muestral*/
PRINT MED COV; /*impresion del vector de medias y la matriz de covarianzas*/
DIFE={0.428,-0.23,-0.026}; /*diferencia de MED menos mu cero*/
ins4=inv(COV); /*inversa de la matriz de covarianzas*/
EP2=((n-p)/p)*dife`*ins4*dife; /*calcula el estadistico de prueba del Ejemplo 2.2*/
PRINT EP2;
FINISH; /*termina el modulo PRUEBA */

```



```
RUN PRUEBA /*correr el modulo PRUEBA */
QUIT;
```

Programa 4.2.6.3 Con los datos del largo de sépalo de las tres especies de *Iris spp*, se trató de ilustrar en el Ejemplo 4.2.3 la prueba de hipótesis de la Sección 4.2.2.3a. Es decir, se plantea si la matriz de covarianzas poblacional es igual a una matriz Σ_0 conocida. Puede observarse que en este ejemplo, la población está constituida por las longitudes del sépalo de las tres especies de *Iris*, y no, por las variables de alguna especie en particular; este comentario, es similar para los datos del ancho de sépalo usados en los Ejemplos 4.2.1 y 4.2.2.

```
DATA IRIS;
INFILE 'C:\HIPOTESIS\IRIS.DAT';
INPUT X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12;
/*X1,X5,X9: longitud de sepalo de I. setosa, I.versicolour e I. virginica, respectivamente*/
/*X2,X6,X10: ancho de sepalo de I. setosa, I.versicolour e I. virginica, respectivamente*/
/*X3,X7,X11: longitud de petalo de I. setosa, I.versicolour e I. virginica, respectivamente*/
/*X4,X8,X12: ancho de petalo de I. setosa, I.versicolour e I. virginica, respectivamente*/
PROC IML;
USE IRIS;
START PRUEBA; /*inicia el modulo PRUEBA */
READ ALL VAR {X1 x5 x9} INTO X; /*arregla los valores del largo de sepalo de las tres
especies de Iris en una matriz X*/
N=NROW(X); /*numero de individuos*/
p=NCOL(X); /*numero de variables por cada especie de Iris*/
J=J(N,1,1); /*un vector de puros unos de orden n=n1=n2=n3 */
MED=(1/N)*X`*J; /*vector de medias de largo de sépalo por especie*/
H=I(N)-(1/N)*J*J`;
COV=(1/N)*X`*H*X; /*matriz de covarianzas*/
CALL SVD(U,Q,V,COV);
A=SQRT(DIAG(Q));
COV12=U*A*U`;
SIG0={0.15 0 0,0 0.25 0,0 0 0.35}; /*matriz conocida*/
INVSIG0=INV(SIG0); /*inversa de la matriz conocida*/
WM1S=COV12*INVSIG0*COV12;
C=EIGVAL(WM1S);
a=(C`*J(P,1,1))/P; Media aritmética DA=TRACE(B)/P;
g=(DET(B))**(1/P);
EP=N*P*(a-LOG(g)-1); /*calcula el estadistico de prueba del Ejemplo 4.2.3*/
PRINT MED COV WM1S a g EP;
FINISH; /*termina el modulo PRUEBA*/
RUN PRUEBA /*correr el modulo PRUEBA */
QUIT;
```

Programa 4.2.6.4 Si se considera que los datos del Cuadro 3.3.4 están en un archivo nombrado CORCHOAR.DAT y están ordenados en 4 columnas que contienen las observaciones de las cuatro variables N,E,S,W y que cada hilera contiene la información de un árbol (son 28 árboles en total), entonces con este programa se puede obtener los resultados del Ejemplo 4.2.4.

```
DATA MIR;
INFILE 'C:\HIPOTESIS\CORCHOAR.DAT'; /*se abre el archivo CORCHOAR.DAT*/
INPUT N E S W; /*las variables del archivo CORCHOAR.DAT*/
PROC IML;
USE MIR;
START CHAANTE; /*inicia el modulo CHAANTE*/
READ ALL VAR {N E S W} INTO X; /*se arreglan los valores de los depositos de corcho en
las cuatro direcciones en una matriz X */
N=NROW(X); /*numero de arboles (de individuos)*/
J=J(N,1,1); /*un vector de puros unos de tamaño n */
K=NCOL(X); /*numero de columnas de X */
MED=(1/N)*X`*J; /*calcula el vector de medias*/
H=I(N)-(1/N)*J*J` /*matriz centrante*/
COV=(1/N)*X`*H*X; /*calcula la matriz de covarianzas*/
R={-1 1 -1 1, 0 -1 0 1, -1 0 1 0}; /*la matriz R para la prueba*/
EP=(N-1)*(R*MED)`*INV(R*COV*R`)*(R*MED); /*calcula el estadistico de prueba de la
expresion (2-16) */
PRINT MED COV R EP; /*imprime resultados*/
FINISH; /*termina el modulo CHAANTE*/
RUN CHAANTE; /*correr el modulo CHAANTE*/
QUIT;
```

Programa 4.2.6.5 Se escribe aquí el programa para obtener los resultados del Ejemplo 4.2.5; otra vez, se usan los datos del ancho de sépalo de las especies de *Iris spp*.

```
DATA IRIS;
INFILE 'C:\ARTURO\HIPOTESIS\IRIS.DAT';
INPUT X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12;
/*X1,X5,X9: longitud de sepalo de I. setosa, I.versicolour e I. virginica, respectivamente*/
/*X2,X6,X10: ancho de sepalo de I. setosa, I. versicolour e I. virginica, respectivamente*/
/*X3,X7,X11: longitud de petalo de I. setosa, I. versicolour e I. virginica, respectivamente*/
/*X4,X8,X12: ancho de petalo de I. setosa, I. versicolour e I. virginica, respectivamente*/
PROC IML;
USE IRIS;
START PRUEBA;
READ ALL VAR {X2 x6 x10} INTO X;
n=NROW(X); /*numero de hileras*/
```

```

p=NCOL(X);      /*numero de variables por cada especie de Iris*/
J=J(N,1,1);     /*un vector de puros unos de orden n*/
MED=(1/N)*X`*J;
H=I(N)-(1/N)*J*J;
COV=(1/N)*X`*H*X;
a0=(1/p)*trace(cov);
g0p=det(cov);
g0=g0p**(1/p);
EP=n*p*LOG(a0/g0); /*calcula el estadistico de prueba*/
PRINT MED COV a0 g0p g0 EP; /*imprime resultados*/
FINISH;         /*termina el modulo PRUEBA */
RUN PRUEBA      /*correr el modulo PRUEBA */
QUIT;

```

Programa 4.2.6.6 Con los datos del Cuadro 3.3.2, se cuenta con observaciones de tres variables en dos tipos de sitios de una especie de pájaro y se plantea en el Ejemplo 4.2.7 probar la siguiente hipótesis: “los vectores de medias poblacionales en ambas condiciones son iguales”. Los cálculos del Ejemplo 4.2.7 se efectuaron con este programa. Observe que en este caso, los datos no se leen de un archivo externo, sino que están incluidos en el programa. En el Ejemplo 4.2.8, se plantea la misma hipótesis pero se prueba usando un estadístico proporcional a la T^2 de Hotelling que se calcula en este mismo programa (ver comentarios del programa /* */).

```

DATA AVES;
INPUT x1a x2a x3a x4d x5d x6d;
/*x1a: altura del dosel en metros de los sitios de anidamiento*/
/*x2a: la distancia de la orilla del nido a la superficie externa mas cercana del grupo de arboles en que
esta situado el nido*/
/*x3a: el perimetro del grupo de arboles */
/*x4d: altura del dosel en metros de los sitios disponibles*/
/*x5d: distancia al contorno de los sitios disponibles */
/*x6d: perimetro del grupo de arboles */
CARDS;
1.20 14.0 8.90 0.47 13.5 3.17
0.58 25.0 4.34 0.62 8.0 3.23
0.74 14.0 2.30 0.75 19.0 2.44
0.70 12.0 5.16 0.52 5.0 1.56
1.36 14.5 2.92 0.73 8.0 2.28
0.78 17.0 3.30 0.62 16.0 3.16
0.45 15.0 3.17 0.60 17.0 2.78
0.78 15.0 4.81 0.26 4.5 3.07
0.63 16.0 2.40 0.46 15.0 3.84
0.75 12.0 3.74 0.28 12.0 3.33
0.55 12.0 4.86 0.53 11.0 2.80
0.45 20.0 2.88 0.42 17.0 2.92

```


1.56 16.0 4.90 0.47 20.0 4.40
 0.85 23.0 4.65 0.50 13.0 3.86
 0.58 12.0 4.02 0.54 16.0 3.48
 0.75 13.0 4.54 0.56 18.0 2.36
 0.55 18.0 3.22 0.32 7.0 3.08
 0.56 18.0 3.08 0.62 16.0 5.07
 0.57 19.5 4.43 0.39 15.0 2.02
 0.41 16.0 3.48 0.56 8.0 1.81
 0.65 18.0 4.50 0.27 9.0 2.05
 0.78 14.0 2.96 0.42 11.0 1.74
 0.64 18.0 5.25 0.70 13.0 2.85
 0.71 18.0 3.07 0.26 14.0 3.64

```
PROC IML;          /*se inicia el procedimiento IML*/
USE AVES;          /*se invoca al conjunto de datos AVES*/
START PRUEBA;      /*empieza el modulo PRUEBA */
READ ALL VAR {x1a x2a x3a} INTO X1; /*se arreglan los valores de los sitios de
                                     anidamiento en una matriz X1*/
READ ALL VAR {x4d x5d x6d} INTO X2; /*se arreglan los valores de los sitios disponibles en
                                     una matriz X2 */

q=NCOL(X1);        /*numero de columnas de X1 */
N1=NROW(X1);        /*numero de sitios de anidamiento */
N2=NROW(X2);        /*numero de sitios disponibles */
N=N1+N2;           /*numero total de sitios */
M1=J(N1,1,1);      /*un vector de puros unos de orden N1*/
M2=J(N2,1,1);      /*un vector de puros unos de orden N2*/
M=J(N,1,1);        /*un vector de puros unos de orden N */
H1=I(N1)-(1/N1)*M1*M1'; /*matriz centrante de los s. anidamiento*/
H2=I(N2)-(1/N2)*M2*M2'; /*matriz centrante de los s. disponibles */
H=I(N)-(1/N)*M*M';   /*matriz centrante tomando a todos los datos como si provinieran de una
                     sola muestra; podemos llamarle la matriz centrante
                     "conjunta"*/

MED1=(1/N1)*X1`*M1; /*vector de medias de los s. de anidamiento*/
MED2=(1/N2)*X2`*M2; /*vector de medias de los sitios disponibles*/
X=X1//X2;          /*se concatenan verticalmente las matrices X1 y X2*/
MED=(1/N)*X`*M;     /*calcula el vector de medias conjunto */
PRINT MED1 MED2 MED;
S1=(1/N1)*X1`*H1*X1; /*matriz de covarianzas de los s. anidamiento*/
S2=(1/N2)*X2`*H2*X2; /*matriz de covarianzas de los s. disponibles*/
W=N1*S1 + N2*S2;    /*matriz de suma de cuadrados y productos cruzados dentro grupos*/
SC=(1/N)*W;         /*estimador de maxima verosimilitud de la matriz de covarianzas comun; observe
                     que SC es una matriz de covarianzas "combinada", según la forma
                     como fue calculada W */

S=(1/N)*X`*H*X;     /*matriz de covarianzas tomando a todos los datos como si provinieran de un
                     solo grupo de sitios: matriz de covarianzas "conjunta" */
T=N*S;             /*matriz de las sumas de cuadrados y productos cruzados total */
PRINT S1 S2 SC S;
K=(N1*(MED1-MED)*(MED1-MED)'+(N2*(MED2-MED)*(MED2-MED)'));
/*K es la matriz de las SCPC "entre grupos"*/
```



```

INVW=INV(W);          /*inversa de la matriz W*/
W1K=INVW*K;
P=I(q)+W1K;
PRINT W K T;
LAMBDA=INV(DET(P));   /*calcula el estadistico de prueba del Ejemplo 4.2.6*/
PRINT LAMBDA;

d=med1-med2;
HOTELLING=1+(((n1*n2)/n)*d`*INVW*d); /*calcula el estadistico de prueba del Ejemplo
                                         4.2.7 que es proporcional a la T cuadrada de
                                         Hotelling para dos muestras*/

PRINT HOTELLING;
FINISH;               /*termina el modulo PRUEBA*/
RUN PRUEBA;           /*correr el modulo PRUEBA*/
QUIT;

```

Programa 4.2.6.7 Usando los datos del Cuadro 3.3.2, se prueba si las matrices de covarianzas de las dos condiciones son iguales. Ahora, a diferencia del Programa 4.2.6.6, los datos son llamados del archivo externo BIRDSDAT.SAS que está en el subdirectorio HIPOTESIS para hacer los cálculos que corresponden. Los resultados de ejecutar este programa se reportaron en el Ejemplo 4.2.9.

```

DATA AVES;
INFILE 'C:\HIPOTESIS\BIRDSDAT.SAS';
INPUT x1a x2a x3a x4d x5d x6d;
/*x1a: altura del dosel en metros de los sitios de anidamiento*/
/*x2a: la distancia de la orilla del nido a la superficie externa mas cercana del grupo de arboles en que
      esta situado el nido*/
/*x3a: el perimetro del grupo de arboles */
/*x4d: altura del dosel en metros de los sitios disponibles*/
/*x5d: distancia al contorno de los sitios disponibles */
/*x6d: perimetro del grupo de arboles */
PROC IML;          /*se inicia el procedimiento IML */
USE AVES;          /*se llama el conjunto de datos AVES */
START PRUEBCOV;    /*empieza el modulo PRUEBCOV */
READ ALL VAR {x1a x2a x3a} INTO X1; /*se arreglan los valores de los sitios de
                                         anidamiento en una matriz X1 */
READ ALL VAR {x4d x5d x6d} INTO X2; /*se arreglan los valores de los sitios disponibles
                                         en una matriz X2 */

q=NCOL(X1);        /*numero de columnas de X1 */
N1=NROW(X1);       /*numero de sitios de anidamiento */
N2=NROW(X2);       /*numero de sitios disponibles */
N=N1+N2;           /*numero total de sitios */
M1=J(N1,1,1);     /*un vector de puros unos de orden N1*/
M2=J(N2,1,1);     /*un vector de puros unos de orden N2*/

```

```

M=J(N,1,1);          /*un vector de puros unos de orden N */
H1=I(N1)-(1/N1)*M1*M1'; /*matriz centrante de los s. anidamiento*/
H2=I(N2)-(1/N2)*M2*M2'; /*matriz centrante de los s. disponibles */
H=I(N)-(1/N)*M*M';    /*matriz centrante conjunta*/
X=X1//X2;             /*se concatenan verticalmente las matrices X1 y X2*/
S1=(1/N1)*X1'*H1*X1; /*matriz de covarianzas de los s. anidamiento*/
Su1=(N1/(N1-1))*S1;
INVSu1=INV(Su1);
S2=(1/N2)*X2'*H2*X2; /*matriz de covarianzas de los s. disponibles*/
Su2=(N2/(N2-1))*S2;
INVSu2=INV(Su2);
S=(1/N)*(n1*S1+n2*S2); /*matriz de covarianzas combinada. Precaucion: en el programa
                        anterior, esta matriz corresponde a SC*/

Su=(N/(N-2))*S;
PRINT S1 Su1 INVSu1;
PRINT S2 Su2 INVSu2;
PRINT S;

p=q;                  /*p es el numero de variables por grupo*/
k=2;                  /*k es el numero de grupos*/
GAMA=1-(((2*P**2+3*P-1)/(6*(P+1)*(K-1)))*((1/(N1-1)-1/(N-K))+(1/(N2-1)-1/(N-
K)))));
MBOX=GAMA*((N1-1)*LOG(DET(INVSu1*Su))+(N2-1)
*LOG(DET(INVSu2*Su))); /*MBOX es el estadistico de prueba*/
PRINT GAMA MBOX;      /*impresion de resultados*/
TITLE1 'PRUEBA DE HIPOTESIS: COMPARACION DE DOS MATRICES DE COVARIANZAS';
FINISH;               /*termina el modulo PRUEBCOV*/
RUN PRUEBCOV;         /*correr el modulo PRUEBCOV*/
QUIT;

```

Programa 4.2.6.8 Con los datos de 24 tortugas hembras y 24 tortugas machos del Cuadro 3.3.4, se escribió en el Ejemplo 4.2.10 la prueba de igualdad de las matrices de covarianzas entre los dos sexos. Los cálculos necesarios para la prueba se hicieron con este programa y los resultados se reportan en dicho ejemplo.

```

DATA TORTUG;
INFILE 'C:\HIPOTESIS\TORTUG1.DAT';
INPUT x1h x2h x3h x4m x5m x6m;
/*x1h: largo de la coraza de hembras*/
/*x2h: ancho de la coraza de hembras*/
/*x3h: altura de la coraza de hembras */
/*x4m: largo de la coraza de machos*/
/*x5m: ancho de la coraza de machos*/
/*x6m: altura de la coraza de machos*/
PROC IML;             /*se inicia el procedimiento IML */
USE TORTUG;           /*se llama el conjunto de datos TORTUG*/

```

```

START PRUEBCOV;      /*empieza el modulo PRUEBCOV*/
READ ALL VAR {x1a x2a x3a} INTO X1;      /*se arreglan los valores de las tortugas
                                          hembras en una matriz X1 */
READ ALL VAR {x4d x5d x6d} INTO X2;      /*se arreglan los valores de las tortugas
                                          masculinas en una matriz X2 */

q=NCOL(X1);          /*numero de variables en X1*/
N1=NROW(X1);          /*numero de tortugas hembras */
N2=NROW(X2);          /*numero de tortugas machos */
N=N1+N2;              /*numero total de tortugas (machos + hembras)*/
M1=J(N1,1,1);         /*un vector de puros unos de orden N1*/
M2=J(N2,1,1);         /*un vector de puros unos de orden N2*/
M=J(N,1,1);           /*un vector de puros unos de orden N */
H1=I(N1)-(1/N1)*M1*M1'; /*matriz centrante de las tortugas hembras*/
H2=I(N2)-(1/N2)*M2*M2'; /*matriz centrante de las tortugas machos*/
H=I(N)-(1/N)*M*M';
MED1=(1/N1)*X1`*M1;    /*vector de medias de las hembras*/
MED2=(1/N2)*X2`*M2;    /*vector de medias de los machos*/
PRINT MED1 MED2;
S1=(1/N1)*X1`*H1*X1;   /*matriz de covarianzas de las hembras(bajo la hipotesis alternativa)*/
Su1=(N1/(N1-1))*S1;
INVSu1=INV(Su1);
S2=(1/N2)*X2`*H2*X2; /*matriz de covarianzas de los machos (bajo la hipotesis alternativa)*/
Su2=(N2/(N2-1))*S2;
INVSu2=INV(Su2);
S=(1/N)*(n1*S1+n2*S2); /*matriz de covarianzas comun (bajo la hipotesis nula)*/
Su=(N/(N-2))*S;
PRINT S1 Su1 INVSu1;
PRINT S2 Su2 INVSu2;
PRINT S;
p=q;                  /*numero de variables en cada grupo (sexo)*/
k=2;                  /*numero de grupos*/
A=DET(Su1);
B=DET(Su2);
C=DET(Su);
PRINT A B C; /*imprime los determinantes de las matrices de covarianzas insesgadas*/
GAMA=1-(((2*P**2+3*P-1)/(6*(P+1)*(K-1)))*((1/(N1-1)-1/(N-K))+1/(N2-1)-1/(N-K))));
MBOX=GAMA*((N1-1)*LOG(DET(INVSu1*Su))+(N2-1)*LOG(DET(INVSu2*Su))); /*MBOX es el estadistico de prueba*/
PRINT GAMA MBOX;      /*imprime los resultados*/
TITLE1 'PRUEBA DE HIPOTESIS: COMPARACION DE DOS MATRICES DE COVARIANZAS
      CON LOS DATOS DE TORTUGAS';
FINISH;                /*termina el modulo PRUEBCOV*/
RUN PRUEBCOV;          /*correr el modulo PRUEBCOV*/
QUIT;

```


Programa 4.2.6.9 Los cálculos para construir los intervalos de confianza simultáneos del Ejemplo 4.2.11, se hicieron con este programa.

```

DATA CLADI;
INFILE'C:\HIPOTESIS\CORCHOAR.DAT';    /*se invoca el archivo CORCHOAR.DAT*/
INPUT N E S W;                          /*las variables de CORCHOAR.DAT*/
PROC IML;
USE CLADI;
START INTERVAL;                          /*se inicia el modulo intervalo*/
READ ALL VAR {N E S W} INTO X;          /*se arreglan los valores de los depositos de
                                         corcho en las cuatro direcciones en una matriz
                                         X*/

R={-1 1 -1 1, 0 -1 0 1, -1 0 1 0};      /*la matriz R que se uso en el Ejemplo*/
Y=X*R;                                   /*genera las nuevas variables*/
N=NROW(Y);                               /*numero de arboles (individuos) que contiene
                                         el archivo que se usa*/

J=J(N,1,1);                             /*un vector de puros unos de tamaño n */
p=NCOL(Y);                               /*numero de columnas(variables) de la matriz Y*/
MED=(1/N)*Y`*J;                          /*calcula el vector de medias de las nuevas variables
                                         (i.e., las combinaciones lineales*/

H=I(N)-(1/N)*J*J;                        /*calcula la matriz centrante*/
COV=(1/N)*Y`*H*Y;                       /*calcula la matriz de covarianzas*/
PRINT MED COV;                           /*imprime el vector de medias y la matriz de
                                         covarianzas*/

a1={1,0,0};
a2={0,1,0};
a3={0,0,1};
A=a1||a2||a3;
m=n-p;
Ftab=FINV(0.99,p,m);                    /*valor de la F con p g.l. en el numerador y n-p g.l.
                                         en el denominador con un nivel alfa de 0.01*/

b1=(((P/(N-P))*a1`*cov*a1)*Ftab)**(0.5);
b2=(((P/(N-P))*a2`*cov*a2)*Ftab)**(0.5);
b3=(((P/(N-P))*a3`*cov*a3)*Ftab)**(0.5);
B=b1||b2||b3;

LI=A*MED-B;                             /*limites inferiores de los ICS, vease (4.2-41)*/
LS=A*MED+B;                             /*limites superiores de los ICS, vease (4.2-41)*/
PRINT LI LS;

FINISH;                                  /*termina el modulo INTERVAL*/
RUN INTERVAL;                            /*corre el modulo INTERVAL*/
QUIT;

```


Programa 4.2.6.10 Con este programa se hacen los cálculos del Ejemplo 4.2.6.

```

/*****/
/*PROGRAMA IML/SAS PARA PROBAR LA HIPOTESIS */
/*DE LA SECCION 4.2.3.2b */
/* aaas*/
/*****/

DATA ALKAB;
INFILE 'C:\CORRCAN\TREES.DAT';
INPUT NUM $ SP $ DN AT AFL RBC;
/*DN:diametro normal, AFT:altura de fuste limpio*/
/*RBC:radio de la base de la copa, AT:altura total*/
X1=DN;X2=AFL;X3=RBC;X4=AT;
PROC IML;USE ALKAB;
START PRUEB426;
READ ALL VAR {X1 X2 X3 X4} INTO X;
Q=NCOL(X);
N=NROW(X); /*numero de arboles*/
J=J(N,1,1); /*vector de unos de tamano n*/
H=I(N)-(1/N)*J*J;
MED=(1/N)*X`*J;
S=(1/N)*X`*H*X; /*calculo de la matriz de covarianzas muestral*/
SRC=S##0.5;
SREC=1/SRC;
DM12=DIAG(SREC);
R=DM12*S*DM12; /*calculo de la matriz de correlaciones*/

S11=S[1:2,1:2];
S22=S[3:4,3:4];
S12=S[1:2,3:4];
S21=S[3:4,1:2];
B1=INV(S11); /*inversa de S11*/
B2=INV(S22); /*inversa de S22*/
M=B2*S21*B1*S12;
EM=I(2)-M;
EP1=DET(EM);
PRINT MED S S11 S22 S12 S21;
PRINT B1 B2;
PRINT M;
PRINT EM EP1;
M11=N*S11;C1=INV(M11);
M22=N*S22; B=N*S21*B1*S12;
M221=N*(S22-S21*B1*S12);
AM1=INV(M221);
AM1B=AM1*N*S21*B1*S12;
PRINT M221 B AM1B;
FINISH; RUN PRUEB426; QUIT;

```

Programa 4.2.6.11 Programa para realizar los cálculos del Ejemplo 4.2.7

```

/*****
/*PROGRAMA IML/SAS PARA EL EJEMPLO 4.2.7      */
/*                                           aaas*/
*****/

DATA KIICHPAM;
INFILE 'C:\CORRCAN\TREES.DAT';
INPUT  NUM $ SP $ DN AT AFL RBC;
/*DN:diametro normal, AFT;altura de fuste limpio*/
/*RBC:radio de la base de la copa, AT:altura total*/
X1=DN;X2=AFL;X3=RBC;X4=AT;

PROC IML;USE KIICHPAM;
START PRUEB427;
READ ALL VAR {X1 X2 X3 X4} INTO X;

P=NCOL(X);
N=NROW(X);      /*numero de arboles*/
J=J(N,1,1);     /*vector de unos de tamano n*/
H=I(N)-(1/N)*J*J';
MED=(1/N)*X`*J;

S=(1/N)*X`*H*X; /*calculo de la matriz de covarianzas muestral*/
SRC=S##0.5;
SREC=1/SRC;
DM12=DIAG(SREC);
R=DM12*S*DM12; /*calculo de la matriz de correlaciones*/
DETR=DET(R);
NE=N-((1/6)*(2*P+11)); /*n estrella*/
EPM=-NE*LOG(DETR); /*EP modificado*/
ME=(1/2)*P*(P-1); /*m estrella*/
JIC1=CINV(0.95,ME); /*valor de jicuada con alfa igual a 0.05*/
JIC2=CINV(0.99,ME); /*valor de jicuada con alfa igual a 0.01*/
PRINT EPM jic1 jic2;
FINISH;
RUN PRUEB427;
QUIT;
```

4.3 REGRESIÓN LINEAL MULTIVARIADA

4.3.1 Introducción

Se hace, en primer lugar, una presentación muy somera de la regresión lineal múltiple en el caso univariado (una variable dependiente y más de una variable explicatoria o independiente), para tratar de extenderla después de manera natural, al contexto de la regresión multivariada (donde se incluye más de una variable dependiente).

4.3.2 Regresión Múltiple Univariada

4.3.2.1 El modelo

En una regresión, el modelo lineal puede escribirse como:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_q X_{iq} + \varepsilon_i, \quad (4.3-1)$$

donde Y_i es la i -ésima observación de la variable dependiente que está en función de las q variables independientes $X_{i1}, \dots, X_{iq}$; $\beta_0, \beta_1, \dots, \beta_q$ son los parámetros de regresión y ε_i es la variable aleatoria error con media cero y varianza σ^2 . Un caso particular de la regresión es cuando $q = 1$ y se le conoce como *regresión simple*.

El modelo de regresión múltiple (4.3-1) puede escribirse matricialmente como: (ver Johnson & Wichern, 1982, p. 293; Mardia *et al.*, 1989, p. 171; Searle, 1982; Steel & Torrie, 1988; p. 288)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (4.3-2)$$

con,

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1q} \\ 1 & X_{21} & X_{22} & \cdots & X_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nq} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

donde, $\mathbf{y}_{n \times 1}$ es un vector que contendrá las n observaciones de la variable dependiente a ser explicada; $\mathbf{X}_{n \times (q+1)}$ es la matriz de las q variables explicatorias más la primera columna de puros unos, si es que se considera un efecto promedio en el modelo (de otro modo no existe); $\boldsymbol{\beta}_{(q+1) \times 1}$ es el vector de parámetros¹ de regresión a estimarse en el análisis; $\boldsymbol{\varepsilon}_{n \times 1}$ es un vector de errores aleatorios² bajo los siguientes supuestos: $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ y $Cov(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \mathbf{I}$. Con el primer supuesto tenemos que la media de las ε_i es cero y con el segundo está implícito que $var(\varepsilon_i) = \sigma^2$ y $cov(\varepsilon_i, \varepsilon_k) = 0 \quad \forall i \neq k$; para mayor claridad, se puede ver que:

$$E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \begin{bmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \cdots & E(\varepsilon_2\varepsilon_n) \\ \vdots & \vdots & & \vdots \\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) & \cdots & E(\varepsilon_n^2) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}.$$

Comentario: Si en el modelo lineal de regresión no se considera la interceptada al origen (i.e $Y_i = \beta_1 X_{i1} + \cdots + \beta_q X_{iq} + \varepsilon_i$), entonces las dimensiones de $\mathbf{X}$ y $\boldsymbol{\beta}$ en (4.3-2) son $(n \times q)$ y $(q \times 1)$, respectivamente, ya que en la matriz $\mathbf{X}$ no se incluye la columna de unos y en el vector $\boldsymbol{\beta}$, se omite al elemento β_0 correspondiente a la interceptada.

¹ Los parámetros son valores desconocidos que se suponen fijos.

² Por esta razón, el vector $\mathbf{y}$ es considerado aleatorio cuando aún no se efectúan las mediciones de la variable dependiente.

El valor esperado del vector y en (4.3-2) es $E(y) = X\beta$, o bien, se puede escribir para cada elemento y_i de y , i.e., $E(y_i) = x_i'\beta$.

4.3.2.2 La estimación del vector de parámetros y las sumas de cuadrados

La estimación del vector de parámetros de regresión en (4.3-1) obtenida por el método de mínimos cuadrados es: (Searle, 1982, pp. 330-331 y 366; Mardia *et al.*, 1989, p. 172)

$$\hat{\beta} = (X'X)^{-1}X'y, \quad (4.3-3)$$

en donde se considera X de rango completo por columnas.

Anticipándose a las expresiones de las sumas de cuadrados, se dan las siguientes definiciones. Sea $X_{n \times m}$, con $r(X) = m$, entonces

$$P_X = X(X'X)^{-1}X'; \quad (4.3-4)$$

$$Q = I - P_X; \quad (4.3-5)$$

donde se cumple que: (a) P_X es simétrica e idempotente; (b) $P_X X = X$; (c) Q es simétrica e idempotente de rango $n - r(X) = n - m$; (d) $QX = X'Q = O$. Por otra parte, si $\hat{y} = X\hat{\beta} = P_X y$, entonces $\hat{\varepsilon} = y - \hat{y} = Qy$ y como resultado directo de (d), $X'\hat{\varepsilon} = 0$ y así también $\hat{y}'\hat{\varepsilon} = y'P_X(I - P_X)y = 0 = \hat{\varepsilon}'\hat{y}$. Las expresiones (4.3-4) y (4.3-5) se usarán en repetidas ocasiones en las secciones siguientes, así como también, con sus respectivas adecuaciones, en el tema de análisis de varianza multivariado del Subcapítulo 4.4.

La suma de cuadrados total sin corregir (SCT) $y'y = \sum_{i=1}^n y_i^2$ se puede particionar así:

$$\begin{aligned} y'y &= (\hat{y} + y - \hat{y})'(y - \hat{y}) = (\hat{y} + \hat{\varepsilon})'(\hat{y} + \hat{\varepsilon}) \\ &= y'P_X y + y'(I - P_X)y \\ &= SC\text{ Reg} + SC\text{Error}. \end{aligned} \quad (4.3-6)$$

El resultado $\mathbf{X}'\hat{\mathbf{e}} = \mathbf{0}$ implica que $\sum_{i=1}^n \hat{e}_i = 0$, i.e. $\mathbf{1}'\hat{\mathbf{e}} = 0$, y la descomposición en un modelo con interceptada de la suma de cuadrados total corregida por la media, denotada por $SCT(media)$, es:

$$\mathbf{y}'(\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{y} = \mathbf{y}'(\mathbf{P}_X - \frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{P}_X)\mathbf{y}, \quad (4.3-7)$$

donde,

$$SCT(media) = \mathbf{y}'[\mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}']\mathbf{y} \quad (4.3-8)$$

$$SC\text{ Reg}(media) = \mathbf{y}'[\mathbf{P}_X - \frac{1}{n}\mathbf{1}\mathbf{1}']\mathbf{y} \quad (4.3-9)$$

$$SCError = \hat{\mathbf{e}}'\hat{\mathbf{e}} = \mathbf{y}'[\mathbf{I} - \mathbf{P}_X]\mathbf{y} \quad (4.3-10)$$

Son importantes algunos comentarios aquí: (i) De acuerdo a la expresión (4.3-4), $\frac{1}{n}\mathbf{1}\mathbf{1}' = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' = \mathbf{P}_{\mathbf{1}_n}$ es el proyector del vector formado por la primera columna de puros 1's de la matriz $\mathbf{X}$. (ii) Se recuerda que en la partición (4.3-6) no se considera la corrección por la media; mientras que en la partición (4.3-7) el factor de corrección por la media está dado por $\mathbf{y}'\mathbf{P}_{\mathbf{1}_n}\mathbf{y}$. (iii) Los grados de libertad asociados a las sumas de cuadrados de las expresiones (4.3-8), (4.3-9) y (4.3-10), son $(n-1)$, q y $[n-(q+1)]$, respectivamente, que corresponden a los rangos de las matrices entre corchetes de esas expresiones. (iv) El lado derecho de (4.3-10) puede verse también como $(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, donde $\hat{\boldsymbol{\beta}}$ de la expresión (4.3-3) minimiza la $SCError$.

4.3.2.3 Propiedades muestrales de los estimadores mínimos cuadrados ordinarios

Algunas propiedades muestrales de los estimadores mínimos cuadrados ordinarios (MCO) son:

$$(a) E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \text{ y } Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1};$$

$$(b) E(\hat{\mathbf{e}}) = \mathbf{0}; Cov(\hat{\mathbf{e}}) = \sigma^2[\mathbf{I} - \mathbf{P}_X] = \sigma^2\mathbf{Q} \text{ y } E(\hat{\mathbf{e}}'\hat{\mathbf{e}}) = \sigma^2(n - r(\mathbf{X})); \quad (4.3-11)$$

(c) Si $\hat{\sigma}^2 = \frac{SCError}{n - r(\mathbf{X})} = \frac{\mathbf{y}'[\mathbf{I} - \mathbf{P}_X]\mathbf{y}}{n - r(\mathbf{X})}$ entonces $E(\hat{\sigma}^2) = \sigma^2$.

Se verifican a continuación estas propiedades:

(a.1) $E(\hat{\beta}) = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}] = E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \varepsilon)] = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon) = \beta$,

ya que $E(\varepsilon) = \mathbf{0}$. Por lo tanto, queda demostrado que $\hat{\beta}$ es un estimador insesgado de β .

(a.2)

$$\begin{aligned} Cov(\hat{\beta}) &= Cov((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Cov(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Cov(\varepsilon)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

(b.1) $E(\hat{\varepsilon}) = E[(\mathbf{I} - \mathbf{P}_X)\mathbf{y}] = (\mathbf{I} - \mathbf{P}_X)E(\mathbf{y}) = (\mathbf{I} - \mathbf{P}_X)\mathbf{X}\beta = (\mathbf{X} - \mathbf{X})\beta = \mathbf{0}$.

(b.2) $Cov(\hat{\varepsilon}) = E(\hat{\varepsilon}\hat{\varepsilon}')$

$$\begin{aligned} &= E[(\mathbf{I} - \mathbf{P}_X)\mathbf{y}\mathbf{y}'(\mathbf{I} - \mathbf{P}_X)] = (\mathbf{I} - \mathbf{P}_X)[E(\mathbf{y}\mathbf{y}')](\mathbf{I} - \mathbf{P}_X) \\ &= (\mathbf{I} - \mathbf{P}_X)[\sigma^2\mathbf{I} + \mathbf{X}\beta\beta'\mathbf{X}'](\mathbf{I} - \mathbf{P}_X) = (\mathbf{I} - \mathbf{P}_X)\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{P}_X) \\ &= \sigma^2[\mathbf{I} - \mathbf{P}_X]. \end{aligned}$$

(b.3) $E(\hat{\varepsilon}'\hat{\varepsilon}) = E[\mathbf{y}'(\mathbf{I} - \mathbf{P}_X)\mathbf{y}]$, es la esperanza de una forma cuadrática,

$$\begin{aligned} &= \text{tr}[(\mathbf{I} - \mathbf{P}_X)Var(\mathbf{y})] + E(\mathbf{y})'(\mathbf{I} - \mathbf{P}_X)E(\mathbf{y}) \\ &= \text{tr}[\sigma^2(\mathbf{I} - \mathbf{P}_X)], \text{ ya que } Var(\mathbf{y}) = \sigma^2\mathbf{I} \text{ y } E(\mathbf{y})'(\mathbf{I} - \mathbf{P}_X)E(\mathbf{y}) = \beta'\mathbf{X}'(\mathbf{X} - \mathbf{X})\beta = \mathbf{0}, \\ &= \sigma^2\text{tr}[\mathbf{I} - \mathbf{P}_X] \\ &= \sigma^2(n - r(\mathbf{X})). \end{aligned}$$

(c) $E(\hat{\sigma}^2) = \frac{1}{n - r(\mathbf{X})}E[\mathbf{y}'(\mathbf{I} - \mathbf{P}_X)\mathbf{y}]$, usando el resultado de la prueba (b.3) anterior,

$$\begin{aligned}
&= \frac{\sigma^2(n-r(\mathbf{X}))}{n-r(\mathbf{X})} \\
&= \sigma^2.
\end{aligned}$$

4.3.2.4 Intervalos de confianza

Para hacer pruebas de hipótesis y construir intervalos de confianza se puede suponer normalidad en la distribución del vector ε del modelo (4.3-2), i.e. $\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$; en esta situación el estimador de máxima verosimilitud para β coincide con el obtenido por mínimos cuadrados ordinarios (Johnson & Wichern, 1982, pp. 302-304 y Mardia *et al.*, 1989, p. 172) y se tiene que $\hat{\beta} \sim N_{q+1}(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$; así como que, $(n-r(\mathbf{X}))\hat{\sigma}^2 = \hat{\varepsilon}'\hat{\varepsilon} \sim \sigma^2 \chi^2_{n-r(\mathbf{X})}$. Una región de confianza al $100(1-\alpha)\%$ para β está dado por (Johnson & Wichern, 1982, p. 204):

$$(\beta - \hat{\beta})' \mathbf{X}'\mathbf{X}(\beta - \hat{\beta}) \leq (q+1)\hat{\sigma}^2 F_{q+1, n-q-1; \alpha} \quad (4.3-12)$$

donde $F_{q+1, n-q-1; \alpha}$ es el punto $100\alpha\%$ superior de una distribución F de Snedecor con $r(\mathbf{X}) = q+1$ grados de libertad (g.l.) en el numerador y $n-r(\mathbf{X}) = n-q-1$ g.l. en el denominador. Así también, para construir intervalos de confianza simultáneos al $100(1-\alpha)\%$ para β_i se usa:

$$\hat{\beta}_i \pm \sqrt{\hat{\text{var}}(\hat{\beta}_i)} \sqrt{(q+1)F_{q+1, n-q-1; \alpha}}, \quad i = 0, 1, \dots, q, \quad (4.3-13)$$

donde $\hat{\text{var}}(\hat{\beta}_i)$ es el i -ésimo elemento de la diagonal de $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$. Puede verse la prueba de estos intervalos simultáneos en Johnson & Wichern (1982; p. 304). Aunque son más recomendables los intervalos de confianza simultáneos de β_i , en la práctica frecuentemente se construyen intervalos para β_i individuales, cada una asociada a una variable independiente de interés, usando la expresión:

$$\hat{\beta}_i \pm t_{n-q-1; \alpha/2} \sqrt{\hat{\text{var}}(\hat{\beta}_i)}, \quad (4.3-14)$$

donde $t_{n-q-1; \alpha/2}$ es el punto $100(\alpha/2)\%$ superior de una t de Student con $n - q - 1$ g.l.

Los detalles sobre intervalos de confianza y presentación de pruebas de hipótesis en la regresión univariada pueden revisarse en Zamudio, en donde se desarrollan también ejemplos de tipo forestal.

EJEMPLO 4.3.1 Un ejemplo de regresión univariada en un escenario forestal, es cuando se tiene como variable dependiente al volumen (V), y como variable independiente, al cuadrado del diámetro multiplicado por la altura (D^2H), en un modelo de regresión simple como:

$$V_i = \beta_0 + \beta_1(D_i^2H_i) + \varepsilon_i, \quad i = 1, \dots, n, \quad (4.3-15)$$

donde V_i , D_i y H_i son el volumen, el diámetro normal y la altura del i -ésimo árbol de una muestra aleatoria de tamaño n . La variable D^2H es conocida como la variable combinada y ha sido la base de muchas ecuaciones del volumen de árboles, desde que Spurr lo presentó en 1952 (Burkhart & Gregoire, 1994).

Si hacemos $Y_i = V_i$ y $X_{i1} = (D_i^2H_i)$, puede verse a (4.3-15) como el modelo (4.3-1), quedando $Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$. Las mediciones de V_i se pueden arreglar en un vector y , y los datos de la variable independiente ($D_i^2H_i$), junto con el vector de unos asociado al parámetro β_0 , en una matriz X como la del modelo (4.3-2), para realizar el análisis de regresión. Con fines ilustrativos, se usan los datos del Cuadro 4.3.1 que contiene el diámetro normal en centímetros (D), la altura en metros (H), y el volumen en metros cúbicos de una muestra de 35 árboles, ejemplificando el uso de (4.3-15). En los siguientes resultados, la columna BETA contiene los estimadores de los parámetros del modelo; enseguida, se encuentra la columna EESTB que contiene los errores estándares de los estimadores de los parámetros; la columna T representa los valores del estadístico t que

tiene una distribución t de Student con el que se prueba si un parámetro en particular es igual a cero; la columna PROBT es el nivel de significancia estimado y la regla de decisión es rechazar la hipótesis nula si el nivel de significancia propuesto es mayor que el nivel de significancia estimado.

$$\begin{aligned} SCReg &= 0.534, & CMReg &= 0.534, & Fcal &= 273.250 \\ SCError &= 0.065, & CMErr &= 0.002, \\ SCTotal &= 0.599, \end{aligned}$$

$$R^2 = 0.8922$$

BETA	EESTB	T	PROBT
0.0260464	0.0130435	1.9968945	0.0541389
0.3592622	0.0217336	16.530258	0.0001

Covarianzas entre las β_i	
0.0001701	-0.000232
-0.000232	0.0004723

Correlaciones entre las β_i	
1	-0.819512
-0.819512	1

El valor de R^2 nos indica que el modelo explica casi 90 % de la varianza. Se observa también que β_1 es significativamente diferente de cero, siendo los valores estimados de los parámetros, $\hat{\beta}_0 = 0.0260464$ y $\hat{\beta}_1 = 0.3592622$. De esta forma, la estimación del volumen de un árbol, dado el diámetro normal y la altura, se establece por $\hat{V}_i = 0.026 + 0.359(D_i^2 H_i)$. Los cálculos de este ejemplo se hicieron con el Programa 4.3.1 del Anexo. ♦

EJEMPLO 4.3.2 Para construir intervalos de confianza simultáneos de los parámetros β_i , $i = 0, 1$, del modelo (4.3-15), se hace uso de la expresión (4.3-13). Así, los intervalos de confianza simultáneos al 95 % para β_0 y β_1 , haciendo uso de los resultados del Ejemplo 4.3.1, son:

$$\hat{\beta}_0 \pm \sqrt{\hat{var}(\hat{\beta}_0)} \sqrt{(2)F_{\alpha; 2, 33}} = 0.0260464 \pm 0.03361,$$

$$\hat{\beta}_1 \pm \sqrt{\hat{\text{var}}(\hat{\beta}_1)} \sqrt{(2)F_{\alpha; 2, 33}} = 0.3592622 \pm 0.056.$$

♦

EJEMPLO 4.3.3 Una generalización de (4.3-15) presentada por Burkhardt & Gregoire (1994) es:

$$V_i = \beta_0 + \beta_1 D_i^2 + \beta_2 H_i + \beta_3 D_i^2 H_i + \varepsilon_i, \quad (4.3-15b)$$

en donde se asume que D^2 y H y su interacción presentan “efectos” aditivos; en este modelo se tiene una regresión múltiple univariada: una variable dependiente (el volumen) y más de una variable independiente (el cuadrado del diámetro, la altura y la variable combinada). Usando los datos del Cuadro 4.3.1, se obtiene la siguiente salida en IML de SAS:

```

The SAS System      11:19 Monday, April 19, 1997
SCREG   SCE      SCTOTAL CMREG    CME    FCAL    R2
0.5585702 0.0403005 0.5988707 0.1861901 0.0013 143.22135 0.9327058

```

```

      BETA      EESTB    T        PROBT
0.0476444 0.051568 0.9239132 0.3626677
0.8346636 1.1644308 0.7167996 0.4788639
-0.009614 0.0058478 -1.643969 0.1102879
0.4284663 0.1149903 3.7261078 0.0007772

```

Covarianzas entre las β_i

```

0.0026593 -0.056688 -0.0002830 0.0055735
-0.056688 1.355899 0.0057785 -0.126059
-0.000283 0.0057785 0.0000342 -0.000644
0.0055735 -0.1260590 -0.000644 0.0132228

```

Correlaciones entre las β_i

```

1.000000 -0.944061 -0.937268 0.9399061
-0.944061 1.000000 0.8486069 -0.941451
-0.937268 0.8486069 1.000000 -0.95756
0.9399061 -0.941451 -0.95756 1.00000

```


Se puede observar de la salida anterior, que los parámetros asociados a las variables D^2 y H , no son significativamente diferentes de cero, mientras que el parámetro asociado a la variable combinada es diferente de cero con un nivel de significancia de 0.01. Los cálculos de este ejemplo se pueden hacer con el Programa 4.3.9.2. ♦

EJEMPLO 4.3.4 Otra ecuación de volumen que puede usarse para ilustrar una regresión múltiple univariada, es la que introdujeron Schumacher & Hall en 1933 (Burkhardt & Gregoire, 1994):

$$\ln V_i = \beta_0 + \beta_1 \ln D_i + \beta_2 \ln H_i + \varepsilon_i, \quad (4.3-16)$$

donde la variable dependiente es el logaritmo natural del volumen y los regresores son el logaritmo natural del diámetro normal y el logaritmo natural de la altura. Note que una vez que se tengan las estimaciones de los parámetros, para nuevos valores observados de D_i y H_i , el valor del volumen estimado se obtendría con:

$$\hat{V}_i = e^{\hat{\beta}_0} D_i^{\hat{\beta}_1} H_i^{\hat{\beta}_2}.$$

♦

4.3.3 Regresión Lineal Multivariada

4.3.3.1 El modelo de regresión lineal multivariada

Siguiendo a Mardia *et al.* (1989), veamos lo que sucede en los casos cuando se tiene que representar la relación entre $p > 1$ variables respuesta (variables dependientes) y un grupo de q variables explicatorias (variables independientes). Suponiendo que cada variable dependiente tiene su propia ecuación de regresión —univariada— en función de $\mathbf{X}$ como la del modelo (4.3-2), entonces se tendrían:

$$\mathbf{y}_{(1)} = \mathbf{X}\boldsymbol{\beta}_{(1)} + \boldsymbol{\varepsilon}_{(1)}$$

$$\mathbf{y}_{(2)} = \mathbf{X}\boldsymbol{\beta}_{(2)} + \boldsymbol{\varepsilon}_{(2)}$$

$$\vdots$$

$$\mathbf{y}_{(p)} = \mathbf{X}\boldsymbol{\beta}_{(p)} + \boldsymbol{\varepsilon}_{(p)}$$

donde $\boldsymbol{\beta}_{(j)} = [\beta_{0j} \ \beta_{1j} \ \cdots \ \beta_{qj}]'$ y $\boldsymbol{\varepsilon}_{(j)} = [\varepsilon_{1j} \ \varepsilon_{2j} \ \cdots \ \varepsilon_{nj}]'$, $j = 1, \dots, p$. Observe que cada variable dependiente (los vectores $\mathbf{y}_{(j)}$) es explicada por las mismas q variables explicatorias (las columnas de $\mathbf{X}$) pero con sus correspondientes vectores de parámetros, $\boldsymbol{\beta}_{(j)}$, y de errores, $\boldsymbol{\varepsilon}_{(j)}$.

Si $\mathbf{Y} = [\mathbf{y}_{(1)} \ \mathbf{y}_{(2)} \ \cdots \ \mathbf{y}_{(p)}]_{n \times p}$ es la matriz de variables dependientes cuyas columnas corresponden a cada una de ellas, $\mathbf{B} = [\boldsymbol{\beta}_{(1)} \ \boldsymbol{\beta}_{(2)} \ \cdots \ \boldsymbol{\beta}_{(p)}]_{(q+1) \times p}$ es la matriz de parámetros, $\mathbf{U} = [\boldsymbol{\varepsilon}_{(1)} \ \boldsymbol{\varepsilon}_{(2)} \ \cdots \ \boldsymbol{\varepsilon}_{(p)}]_{n \times p}$ es la matriz de errores aleatorios no observados y $\mathbf{X}_{n \times (q+1)}$ es una matriz de q variables independientes como la presentada en (4.3-2), entonces el modelo definido por:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}, \quad (4.3-17)$$

es el *modelo de regresión lineal multivariada*. Para una $\mathbf{X}$ dada, las hileras de $\mathbf{U}_{n \times p}$ no están correlacionadas, y cada una tiene media $\mathbf{0}$ y matriz de covarianzas común $\boldsymbol{\Sigma}$, i.e.

$\boldsymbol{\varepsilon}_i = [\varepsilon_{i1} \ \varepsilon_{i2} \ \cdots \ \varepsilon_{ip}]' \sim (\mathbf{0}, \boldsymbol{\Sigma})$; por lo tanto, las columnas $\boldsymbol{\varepsilon}_{(j)}$ pueden estar correlacionadas. En el caso de los diseños experimentales $\mathbf{X}$ en (4.3-17) es una matriz diseño, usualmente de 0's y 1's (*ibid.*, p. 157). Normalmente se supondrá, como en el caso univariado, que la primera columna de $\mathbf{X}$ será un vector de puros 1's, i.e.

$\mathbf{X}_{n \times (q+1)} = \begin{bmatrix} \mathbf{1} & \mathbf{X}_1 \\ n \times 1 & n \times q \end{bmatrix}$; en regresión $\mathbf{1}$ corresponde a la interceptada al origen mientras que

en diseños experimentales corresponde a un efecto general medio sobre la respuesta.

Como una extensión directa del caso univariado, se tiene que el valor esperado del j -ésimo vector columna de $\mathbf{Y}$, $\mathbf{y}_{(j)}$, del modelo (4.3-17) es:

$$E(\mathbf{y}_{(j)}) = \mathbf{X}\boldsymbol{\beta}_{(j)}, \quad (4.3-18)$$

por lo que se puede escribir que para cada elemento y_{ij} , $E(y_{ij}/\mathbf{X}) = \mathbf{x}_i'\boldsymbol{\beta}_{(j)}$. En palabras quiere decir, que el valor esperado de la j -ésima respuesta del i -ésimo individuo, dada $\mathbf{X}$, depende de la i -ésima hilera de $\mathbf{X}$ y la j -ésima columna de $\mathbf{B}$.

4.3.3.2 Estimadores de máxima verosimilitud de $\mathbf{B}$ y $\boldsymbol{\Sigma}$ y las sumas de cuadrados y productos cruzados

A continuación se escriben los estimadores de máxima verosimilitud de $\mathbf{B}$ y $\boldsymbol{\Sigma}$ (Teorema 6.2.1 de Mardia *et al.*, 1989) del modelo (4.3-17) bajo los supuestos de que las hileras de $\mathbf{U}$ son una muestra aleatoria de la población $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ y $\mathbf{U}$ es independiente de $\mathbf{X}$. Para $n \geq p + (q + 1)$,

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (4.3-19)^1$$

y

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\mathbf{Y}'\mathbf{Q}\mathbf{Y} \quad (4.3-20)$$

donde se recuerda que $\mathbf{Q} = \mathbf{I} - \mathbf{P}_\mathbf{X}$. Note que $\mathbf{X}$ debe ser de rango completo por columnas, para que $(\mathbf{X}'\mathbf{X})^{-1}$ exista, y en consecuencia, la estimación de $\mathbf{B}$ en (4.3-17) sea única. La demostración es desarrollada en Mardia *et al.* (1989, pp. 158-159).

El valor ajustado de $\mathbf{Y}$ es $\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{P}_\mathbf{X}\mathbf{Y}$ y la matriz estimada de errores es $\hat{\mathbf{U}} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{B}} = \mathbf{Y} - \mathbf{P}_\mathbf{X}\mathbf{Y} = (\mathbf{I} - \mathbf{P}_\mathbf{X})\mathbf{Y} = \mathbf{Q}\mathbf{Y}$. De ahí que (*op. cit.*, p. 159):

¹ También puede verse esta ecuación como la matriz que contiene en sus columnas los estimadores por mínimos cuadrados ordinarios de los vectores $\boldsymbol{\beta}_{(j)}$ univariados, $j = 1, \dots, p$.

$$\hat{U}'\hat{U} = Y'Y - \hat{Y}'\hat{Y} = Y'(I - P_X)Y = Y'QY, \quad (4.3-21)$$

[véase la analogía con la expresión (4.3-10)] lo cual se puede interpretar como: que la matriz de la suma de cuadrados y productos cruzados (SCPC) de errores, es igual a la diferencia entre la matriz de la SCPC total $Y'Y$ y la matriz de la SCPC estimada $\hat{Y}'\hat{Y}$. A partir de esto, podemos expresar la siguiente descomposición de las SCPC de una regresión multivariada:

$$\begin{aligned} Y'Y &= \hat{Y}'\hat{Y} + \hat{U}'\hat{U} \\ Y'(I)Y &= Y'P_X Y + Y'(I - P_X)Y \\ (SCPC(Total)) &= (SCPC(Estimada)) + (SCPC(Errores)) \end{aligned} \quad (4.3-22)$$

Si se hace $M_{11} = X'X$, $M_{22} = Y'Y$, $M_{12} = X'Y$ y $M_{21} = M'_{12}$, entonces: (*Ibid.*, p. 160)

$$n\hat{\Sigma} = \hat{U}'\hat{U} = M_{22} - M_{21}M_{11}^{-1}M_{12},$$

lo que significa que la matriz de la SCPC de errores puede verse como una matriz de errores Wishart $M_{22.1}$ condicionada en X .

Si en la partición (4.3-22) se incorpora el factor de corrección por la media $(1/n)Y'11'Y$ en las SCPC, la descomposición de las SCPC resultante es:

$$\begin{aligned} Y'(I - \frac{1}{n}11')Y &= Y'(P_X - \frac{1}{n}11')Y + Y'(I - P_X)Y \\ SCPC(Total\ corregida) &= SCPC[Estimada(corregida)] + SCPC(Errores) \end{aligned} \quad (4.3-23)$$

EJEMPLO 4.3.5 En los Cuadros 4.3.2 y 4.3.3, se cuenta con los datos del peso de leña y forraje, respectivamente, de ramas de *Gliricidia sepium* con diferentes diámetros (D) y longitudes (L); tomando el peso de leña y el peso de forraje como las variables dependientes y considerando los diámetros y las longitudes de las ramas como las variables explicatorias se pueden calcular las sumas de cuadrados y productos cruzados de la expresión (4.3-23), así como la estimación de la matriz de parámetros. En este

ejemplo se hace uso del modelo lineal $\beta_0 + \beta_1 L + \beta_2 D^2 L + \varepsilon$ para explicar las variables dependientes.

$$SCPC(Estimada\ corregida) = \begin{bmatrix} 89.4614 & 9.6678 \\ 9.6678 & 1.2469 \end{bmatrix}; \quad SCPC(Error) = \begin{bmatrix} 0.0042 & 0.011 \\ 0.0110 & 1.6578 \end{bmatrix};$$

$$SCPC(Total\ corregida) = \begin{bmatrix} 89.4656 & 9.6788 \\ 9.6788 & 2.9047 \end{bmatrix}; \quad \hat{\mathbf{B}} = \begin{bmatrix} 0.463 & 0.308 \\ 0.001 & 0.033 \\ 0.030 & 0.002 \end{bmatrix}.$$

Estos cálculos se obtuvieron con el Programa 4.3.9.3. Véase también el Ejemplo 4.3.6. ♦

4.3.3.3 Propiedades muestrales de los estimadores y la distribución de $\hat{\mathbf{B}}$ y $\hat{\Sigma}$

Se considera la distribución conjunta de los estimadores de máxima verosimilitud de las expresiones (4.3-19) y (4.3-20) del modelo lineal multivariado con errores normales (Mardia *et al.*, 1989; Teorema 6.2.2, p.160). Bajo el modelo (4.3-17):

- (a) $E(\hat{\mathbf{B}}) = \mathbf{B}$,
 - (b) $E(\hat{\mathbf{U}}) = \mathbf{O}$,
 - (c) $\hat{\mathbf{B}}$ y $\hat{\mathbf{U}}$ tienen distribución normal multivariada,
 - (d) $\hat{\mathbf{B}}$ es estadísticamente independiente de $\hat{\mathbf{U}}$, y entonces, también de $\hat{\Sigma}$.
- (4.3-24)

Las demostraciones de estas propiedades están en Mardia *et al.* (1989). Observe que (4.3-24a) y (4.3-24b) se pueden demostrar por analogía directa al caso univariado, desarrolladas en las pruebas (a.1) y (b.1) de las propiedades muestrales de los estimadores por MCO presentadas en la expresión (4.3-11). Aquí sólo escribiremos las demostraciones de las propiedades (4.3-24c) y (4.3-24d).

(4.3-24c) Para demostrar esta propiedad, se tiene que exhibir que $\hat{U}$ y $\hat{B}$ son funciones lineales de U , ya que se supone $U \sim N_p(0, \Sigma)$. Sabemos que $\hat{U} = Y - P_X Y = (I - P_X)Y = QY$, y que $QU = (I - P_X)U = (I - P_X)(Y - XB) = Y - P_X Y = QY$; esto es, $\hat{U} = QY = QU$. Además, substituyendo Y por $XB + U$ en el lado derecho de (4.3-19), obtenemos $\hat{B} = (X'X)^{-1}X'(XB + U) = B + (X'X)^{-1}X'U$. Entonces, tanto $\hat{U}$ como $\hat{B}$ son funciones lineales de U . Por lo tanto, como hemos supuesto que U tiene distribución normal multivariada, entonces también $\hat{U}$ y $\hat{B}$ son normales multivariadas. Observe que las hileras de estas matrices aleatorias no son independientes.

(4.3-24d) Para demostrar esta propiedad tenemos que usar el siguiente resultado (Teorema 3.3.3 de Mardia *et al.*, 1989; p. 65):

Resultado. Si X es una matriz de datos proveniente de una población que tiene distribución $N(\mu, \Sigma)$ y si $Y = AXB$ y $Z = CXD$, entonces los elementos de Y son independientes de los elementos de Z si y sólo si ocurre una de las siguientes dos posibilidades (i) $B'\Sigma D = O$ ó (ii) $AC' = O$.

Podemos escribir $\hat{U} = QUI$ y $\hat{B} - B = (X'X)^{-1}X'UI$; además $U \sim N_p(0, \Sigma)$. Hagamos $A = Q$, $B = I$, $C = (X'X)^{-1}X'$ y $D = I$. Como se cumple que $AC' = Q(X'X)^{-1}X' = O$ entonces podemos asegurar que $\hat{U}$ y $(\hat{B} - B)$ son estadísticamente independientes. Como B es una matriz constante, entonces $\hat{B}$ es estadísticamente independiente de $\hat{U}$. Así también, $\hat{B}$ es estadísticamente independiente de $\hat{\Sigma}$ ya que $\hat{\Sigma} = (1/n)\hat{U}'\hat{U}$.

Una forma alternativa de demostrar la propiedad (4.3-24d) es:

Sabemos que $\hat{B} = (X'X)^{-1}X'Y \Rightarrow \hat{B} = B + (X'X)^{-1}X'U \Rightarrow \hat{\beta}_{(j)} = \beta_{(j)} + (X'X)^{-1}X'\epsilon_{(j)}$.

También que $\hat{U} = QY = QU \Rightarrow \hat{\epsilon}_{(j)} = Q\epsilon_{(j)} = (I - P_X)\epsilon_{(j)}$. Entonces,

$$Cov(\hat{\beta}_{(j)}, \hat{\epsilon}_{(k)}) = E(\hat{\beta}_{(j)}\hat{\epsilon}'_{(k)})$$

$$\begin{aligned}
&= E \left[\left(\beta_{(j)} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\varepsilon}_{(j)} \right) \left((\mathbf{I} - \mathbf{P}_X) \boldsymbol{\varepsilon}_{(k)} \right)' \right], \\
&= E \left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \boldsymbol{\varepsilon}_{(j)} \boldsymbol{\varepsilon}_{(k)}' (\mathbf{I} - \mathbf{P}_X) \right] \quad \text{ya que } \beta_{(j)} \text{ es constante,} \\
&= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' E(\boldsymbol{\varepsilon}_{(j)} \boldsymbol{\varepsilon}_{(k)}') (\mathbf{I} - \mathbf{P}_X), \\
&= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \sigma_{jk} \mathbf{I} (\mathbf{I} - \mathbf{P}_X) \\
&= \sigma_{jk} \left[(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' - (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \right] = \mathbf{O}
\end{aligned}$$

Por lo tanto, queda demostrado que $\hat{\mathbf{B}}$ es estadísticamente independiente de $\hat{\mathbf{U}}$, ya que sus columnas lo son.

Se escriben a continuación dos resultados del Teorema 6.2.3 de Mardia *et al.* (1989) que se usarán en la Sección 4.3.5 sobre predicciones de la regresión multivariada.

$$(i) \text{ cov}(\hat{\beta}_{ij}, \hat{\beta}_{kl}) = \sigma_{jl} g_{ik} \text{ donde } \mathbf{G} = (\mathbf{X}'\mathbf{X})^{-1} \Rightarrow \text{Cov}(\hat{\beta}_{(j)}, \hat{\beta}_{(l)}) = \sigma_{jl} \mathbf{G}, \quad (4.3-25a)$$

$$(ii) n\hat{\Sigma} = \hat{\mathbf{U}}' \hat{\mathbf{U}} \sim W_p(\Sigma, n - q - 1). \quad (4.3-25b)$$

Pruebas:

(i) De $\hat{\mathbf{B}} = \mathbf{B} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{U}$, tenemos que $\hat{\mathbf{B}} - \mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{U}$ que implica $\hat{\beta}_{ij} - \beta_{ij} = \mathbf{a}'_i \boldsymbol{\varepsilon}_{(j)}$, donde $\mathbf{a}'_i$ es la i -ésima hilera de $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$ y $\boldsymbol{\varepsilon}_{(j)}$, como ya se ha dicho, es la j -ésima columna de $\mathbf{U}$. Por lo tanto,

$$\text{cov}(\hat{\beta}_{ij}, \hat{\beta}_{kl}) = \mathbf{a}'_i E[\boldsymbol{\varepsilon}_{(j)} \boldsymbol{\varepsilon}_{(l)}'] \mathbf{a}_k = \mathbf{a}'_i [\sigma_{jl} \mathbf{I}] \mathbf{a}_k = \sigma_{jl} (\mathbf{A} \mathbf{A}')_{ik} = \sigma_{jl} g_{ik}$$

ya que $\mathbf{A} \mathbf{A}' = \mathbf{G}$.

(ii) Usando $\hat{\mathbf{U}} = \mathbf{Q} \mathbf{U}$, $n\hat{\Sigma} = \hat{\mathbf{U}}' \hat{\mathbf{U}} = \mathbf{U}' \mathbf{Q} \mathbf{U}$. Como las hileras de $\mathbf{U}$ tienen vector media $\mathbf{0}$, usando el inciso (b) del Teorema 4.1.8.4 se concluye que $n\hat{\Sigma} \sim W_p(\Sigma, n - q - 1)$ ya que $(n - q - 1)$ es el rango de $\mathbf{Q} = (\mathbf{I} - \mathbf{P}_X)$.

Sería deseable que $\mathbf{X}'\mathbf{X} = \mathbf{I}$ o que sea muy cercano a, ya que elementos diferentes de una columna $\hat{\beta}_{(j)}$ tendrían correlación cero o cercana a cero.

4.3.4 Prueba de Razón de Verosimilitud (PRV) de los Parámetros de Regresión

4.3.4.1 Hipótesis de que las respuestas no dependen de las últimas $q-r$ variables de X

Esta prueba se presentará siguiendo a Johnson & Wichern (1982, p. 327). La matriz de variables independientes $X_{n \times (q+1)}$ del modelo (4.3-17) puede verse como

$$X = \left[\underbrace{x_{(0)} \quad x_{(1)} \quad \cdots \quad x_{(r)}}_{r+1 \text{ columnas}} \quad \underbrace{x_{(r+1)} \quad \cdots \quad x_{(q)}}_{q-r \text{ columnas}} \right], \quad r < q. \text{ De modo que se puede particionar}$$

$X = [X_1 \quad X_2]$ donde X_1 es de orden $n \times (r+1)$ y X_2 de $n \times (q-r)$; observe que $x_{(0)} = \mathbf{1}$ es de dimensión $(n \times 1)$. Así también, puede particiones la matriz de los

parámetros de regresión, B , en $B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$ donde la submatriz B_1 es de dimensión $(r+1) \times p$ y B_2 de $(q-r) \times p$. Entonces esta prueba de hipótesis se puede plantear como:

$$H_0: B_2 = 0 \quad \text{vs.} \quad H_1: B_2 \neq 0 \quad (4.3-26)$$

El modelo (4.3-17) puede escribirse como:

$$Y = [X_1 \quad X_2] \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} + U = X_1 B_1 + X_2 B_2 + U \quad (4.3-27)$$

y bajo la hipótesis nula H_0 , $Y = X_1 B_1 + U$. La prueba de razón de verosimilitud de H_0 se basa en:

$$Y'(I - P_{X_1})Y - Y'(I - P_X)Y = n(\hat{\Sigma}_1 - \hat{\Sigma}),$$

donde $\mathbf{P}_{X_1} = \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$ y $\hat{\Sigma}_1 = \frac{1}{n}\mathbf{Y}'(\mathbf{I} - \mathbf{P}_{X_1})\mathbf{Y}$ y se considera a $\hat{\mathbf{B}}$ como está definida en (4.3-19). Obviamente $\mathbf{X}_1\hat{\mathbf{B}}_1 = \mathbf{P}_{X_1}\mathbf{Y}$. La razón de verosimilitud, Λ , puede expresarse en términos de las varianzas generalizadas como:

$$\Lambda = \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|} \right)^{n/2} \quad (4.3-28)$$

Equivalentemente, se puede usar el estadístico lambda de Wilks, $\Lambda^{2/n}$.

Resultado. Considerar el modelo de regresión multivariada con $\mathbf{X}$ de rango completo por columnas igual a $q+1 < n$ y suponer que los errores $\mathbf{U}$ están distribuidos normalmente. Bajo H_0 en (4.3-26), $n\hat{\Sigma}$ se distribuye como una $W_p(\Sigma, n-q-1)$ independientemente de $n(\hat{\Sigma}_1 - \hat{\Sigma})$ los cuales, a su vez se distribuyen como una $W_p(\Sigma, q-r)$. La PRV para la prueba planteada en (4.3-26) indica rechazar H_0 para valores grandes de (Johnson & Wichern, 1982):

$$-2 \ln \Lambda = -n \ln \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|} \right) = -n \ln \left(\frac{|n\hat{\Sigma}|}{|n\hat{\Sigma} + n(\hat{\Sigma}_1 - \hat{\Sigma})|} \right). \quad (4.3-29)$$

Para n grande el estadístico de prueba modificado (*EPM*),

$$EPM = -\left(n - q - 1 - \frac{1}{2}(p - r)\right) \ln \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|} \right), \quad (4.3-30)$$

tiene una distribución muy aproximada a una χ_m^2 con $m = p(r+1)$ grados de libertad. Usando el *EPM*, la regla de decisión para una prueba de tamaño α es: rechazar H_0 si $EPM \geq \chi_{m, \alpha}^2$; no rechazar H_0 , de otro modo. Tanto $n-q$ como $n-p$ deben ser grandes también para que la aproximación de ji-cuadrada sea adecuada. La demostración de esta prueba de hipótesis, puede leerse en Johnson & Wichern (1982, pp. 328-329).

EJEMPLO 4.3.6 Considerando los datos y el modelo usados en el Ejemplo 4.3.5, se desea probar la hipótesis que plantea lo siguiente: *las respuestas no dependen de las dos variables independientes L y D^2L* . Particiónese por hileras a la matriz de parámetros del modelo de regresión como $\mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix}$ donde $\mathbf{B}_1 = [\beta_{01} \ \beta_{02}]$ y $\mathbf{B}_2 = \begin{bmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{bmatrix}$; de esta forma se puede plantear la hipótesis nula como $H_0: \mathbf{B}_2 = \mathbf{O}$. En este ejemplo $p = 2$, $r = 0$, $q = 2$, $n = 99$. Haciendo los cálculos correspondientes se obtienen:

$$\hat{\Sigma} = \begin{bmatrix} 0.0000425 & 0.0001106 \\ 0.0001106 & 0.016745 \end{bmatrix}; \hat{\Sigma}_1 = \begin{bmatrix} 0.9036931 & 0.0977654 \\ 0.0977654 & 0.0293400 \end{bmatrix};$$

usando ahora la expresión (4.3-30), se tiene que el estadístico de prueba es 959.14573 que se compara con $\chi^2_{2; 0.01} = 9.21$. La conclusión es que se debe rechazar la hipótesis nula con un nivel de significancia $\alpha = 0.01$; es decir, se rechaza la hipótesis de que los pesos de leña y de forraje no dependen de las variables L y D^2L , lo cual, intuitivamente, parece razonable. ♦

Por otro lado, si $\mathbf{X}$ es de rango incompleto, $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^- \mathbf{X}'\mathbf{Y}$, donde $(\mathbf{X}'\mathbf{X})^-$ es una inversa generalizada de $\mathbf{X}'\mathbf{X}$. Entonces los resultados sobre distribución y del EPM serán los mismos, sólo que substituyendo $(q+1)$ por $r(\mathbf{X})$ y $(r+1)$ por $r(\mathbf{X}_1)$.

4.3.5 Predicciones de la Regresión Multivariada

(Johnson & Wichern, 1982; pp. 331-332)

Cuando se hace una regresión con variables dasométricas, el manejador forestal puede buscar cualquiera de dos objetivos: (i) hacer estimaciones de la(s) variable(s) dependiente(s) en un tiempo fijo t_0 , i.e. de forma transversal, o (ii) hacer predicciones de la(s) variable(s) dependiente(s) en función del tiempo para los próximos años (i.e. longitudinalmente). Si el modelo (4.3-17), $\mathbf{Y} = \mathbf{XB} + \mathbf{U}$, con distribución normal de los

errores, se ha ajustado y resulta que “no presenta violaciones” a los supuestos del modelo lineal, entonces puede usarse para los propósitos del inciso (ii).

Un problema en regresión, es la predicción de la(s) respuesta(s) media(s) que corresponde(n) a valores fijos $\mathbf{x}'_0 = [1 \ x_{01} \ x_{02} \ \cdots \ x_{0q}]$, de las variables predictoras. En la regresión multivariada, se pueden hacer inferencias sobre las respuestas medias haciendo uso de los resultados en (4.3-24) y (4.3-25) donde se supone además, que $\mathbf{X}$ del modelo (4.3-17) es de rango completo por columnas. De esos resultados se tiene que:

$$\hat{\mathbf{B}}'\mathbf{x}_0 \sim N_p(\mathbf{B}'\mathbf{x}_0, \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\Sigma) \quad (4.3-31)$$

y $n\hat{\Sigma} \sim W_p(\Sigma, n - q - 1)$ independientemente.

El valor desconocido de la función de regresión en $\mathbf{x}_0$ es $\mathbf{B}'\mathbf{x}_0$, i.e. $E(\hat{\mathbf{Y}}_0|\mathbf{x}_0) = E(\hat{\mathbf{B}}'\mathbf{x}_0) = \mathbf{B}'\mathbf{x}_0$. Se usa también la T^2 de Hotelling, que puede verse como el análogo multivariado del cuadrado de $t = (\bar{X} - \mu_0)/(s/\sqrt{n})$, donde $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ y

$s^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X})^2$, utilizada para la prueba univariada $H_0: \mu = \mu_0$, es decir:

$$T^2 = (\bar{\mathbf{x}} - \mu_0)' \left(\frac{1}{n} \mathbf{S}_u \right)^{-1} (\bar{\mathbf{x}} - \mu_0) = n(\bar{\mathbf{x}} - \mu_0)' \mathbf{S}_u^{-1} (\bar{\mathbf{x}} - \mu_0),$$

donde $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X}'\mathbf{1}$ y $\mathbf{S}_u = \frac{1}{n-1}\mathbf{X}'\mathbf{H}\mathbf{X}$, con $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$ y μ_0 un vector $p \times 1$ de valores dados.

De lo anterior, se puede escribir el siguiente estadístico T^2 :

$$T^2 = \left(\frac{\hat{\mathbf{B}}'\mathbf{x}_0 - \mathbf{B}'\mathbf{x}_0}{\sqrt{\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \right)' \left(\frac{n\hat{\Sigma}}{n - q - 1} \right)^{-1} \left(\frac{\hat{\mathbf{B}}'\mathbf{x}_0 - \mathbf{B}'\mathbf{x}_0}{\sqrt{\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \right) \quad (4.3-32)$$

con el que el elipsoide de confianza al $100(1 - \alpha)\%$ para $\mathbf{B}'\mathbf{x}_0$ está dado por:

$$\left(\mathbf{B}'\mathbf{x}_0 - \hat{\mathbf{B}}'\mathbf{x}_0 \right)' \left(\frac{n\hat{\Sigma}}{n-q-1} \right)^{-1} \left(\mathbf{B}'\mathbf{x}_0 - \hat{\mathbf{B}}'\mathbf{x}_0 \right) \leq \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 \left[\left(\frac{p(n-q-1)}{n-q-p} \right) F_{p, n-q-p, \alpha} \right] \quad (4.3-33)$$

donde $F_{p, n-q-p, \alpha}$ es el punto superior al $100\alpha\%$ de una distribución F con p g.l. en el numerador y $(n-q-p)$ g.l. en el denominador.

Los intervalos de confianza simultáneos al $100(1-\alpha)\%$ para $E(\hat{Y}_{0j}) = \mathbf{x}_0'\beta_{(j)}$ son:

$$\mathbf{x}_0'\hat{\beta}_{(j)} \pm \sqrt{\left(\frac{p(n-q-1)}{n-q-p} \right) F_{p, n-q-p, \alpha}} \sqrt{\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 \left(\frac{n}{n-q-1} \hat{\sigma}_{ii} \right)}, \quad j=1, \dots, p \quad (4.3-34)$$

donde, como se ha convenido, $\hat{\beta}_{(j)}$ es la j -ésima columna de $\hat{\mathbf{B}}$ y $\hat{\sigma}_{ii}$ es el i -ésimo elemento de la diagonal principal de $\hat{\Sigma}$.

Otro problema de predicción, es el pronóstico de nuevas respuestas $\mathbf{Y}_0 = \mathbf{B}'\mathbf{x}_0 + \varepsilon_0$ en $\mathbf{x}_0$ no consideradas en las observaciones iniciales. Se supone que ε_0 es independiente de $\mathbf{U}$. Si sumamos $-\hat{\mathbf{B}}'\mathbf{x}_0$ en ambos lados de $\mathbf{Y}_0$,

$$\mathbf{Y}_0 - \hat{\mathbf{B}}'\mathbf{x}_0 = (\mathbf{B} - \hat{\mathbf{B}})'\mathbf{x}_0 + \varepsilon_0 \quad (4.3-35)$$

que se distribuye como $N_p\left(0, \left(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0\right)\Sigma\right)$ independientemente de $n\hat{\Sigma}$. El elipsoide de predicción al $100(1-\alpha)\%$ para $\mathbf{Y}_0$ está dado por la siguiente desigualdad:

$$\left(\mathbf{Y}_0 - \hat{\mathbf{B}}'\mathbf{x}_0 \right)' \left(\frac{n\hat{\Sigma}}{n-q-1} \right)^{-1} \left(\mathbf{Y}_0 - \hat{\mathbf{B}}'\mathbf{x}_0 \right) \leq \left(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 \right) \left[\frac{p(n-q-1)}{n-q-p} (F_{p, n-q-p, \alpha}) \right] \quad (4.3-36)$$

Los intervalos de predicción simultáneos al $100(1-\alpha)\%$ para el j -ésimo valor de $\mathbf{Y}_0$, denotado por Y_{0j} , son:

$$\mathbf{x}_0' \hat{\beta}_{(j)} \pm \sqrt{\left(\frac{p(n-q-1)}{n-q-p} \right) F_{p, n-q-p, \alpha}} \sqrt{\left(1 + \mathbf{x}_0' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_0 \right) \left(\frac{n}{n-q-1} \hat{\sigma}_{ii} \right)}, \quad (4.3-37)$$

con $j = 1, \dots, p$. Comparando los intervalos de confianza simultáneos para $E(\hat{Y}_{0j}) = \mathbf{x}_0' \beta_{(j)}$ y para la predicción individual Y_0 de nuevas respuestas, se puede observar que los intervalos de predicción para los valores actuales de la variable respuesta son más amplios que los intervalos correspondientes para los valores esperados. La amplitud extra viene del factor adicional $[n/(n-q-1)] \hat{\sigma}_{ii}$ y refleja la presencia del error aleatorio ε_{0j} en la observación futura.

4.3.6 Hipótesis Lineal General (Mardia *et al.*, 1989; pp. 161-167)

La hipótesis lineal general se puede plantear así:

$$H_0: \mathbf{C}_1 \mathbf{B} \mathbf{M}_1 = \mathbf{D} \text{ versus } H_1: \mathbf{C}_1 \mathbf{B} \mathbf{M}_1 \neq \mathbf{D}, \quad (4.3-38)$$

donde $\mathbf{C}_1$ tiene dimensión $(g \times m)$ con $m = q + 1$, $\mathbf{M}_1$ de orden $(p \times r)$ y $\mathbf{D}$ de $(g \times r)$; además $\mathbf{C}_1$ y $\mathbf{M}_1$ tienen rango g y r , respectivamente. Cuando $\mathbf{D} = \mathbf{O}$ y $\mathbf{M}_1 = \mathbf{I}$, la hipótesis toma la forma particular de:

$$H_0: \mathbf{C}_1 \mathbf{B} = \mathbf{O} \text{ versus } H_1: \mathbf{C}_1 \mathbf{B} \neq \mathbf{O} \quad (4.3-39)$$

Se considerará la hipótesis bajo el enfoque de la prueba de razón de verosimilitud (PRV) y la prueba unión intersección (PUI).

4.3.6.1 La prueba de razón de verosimilitud (PRV)

Considerar la hipótesis $H_{0a}: \mathbf{C}_1 \mathbf{B} = \mathbf{D}$ (i.e. $\mathbf{M}_1 = \mathbf{I}$). Sea $\mathbf{C}_2 ((m-g) \times m)$ tal que $\mathbf{C}' = (\mathbf{C}_1', \mathbf{C}_2')$ es una matriz no singular de orden $m \times m$; sea $\mathbf{B}_0 (m \times p)$ cualquier matriz que satisfaga $\mathbf{C}_1 \mathbf{B}_0 \mathbf{I} = \mathbf{D}_{g \times p}$.

Observe que si al modelo $Y = XB + U$ le sumamos $(-XB_0)$ en ambos lados, se tiene:

$$Y - XB_0 = X(B - B_0) + U,$$

con $X_{n \times m}$. Si definimos a la nueva matriz de respuestas como $Y_+ = Y - XB_0$, e igualamos $X = ZC$, tendremos:

$$Y_+ = ZC(B - B_0) + U$$

donde se ve que $Z = XC^{-1}$. Definiendo de forma particionada a $\Delta = (\Delta'_1, \Delta'_2)' = (C'_1, C'_2)'(B - B_0)$, el modelo queda finalmente como:

$$Y_+ = Z\Delta + U, \quad (4.3-40)$$

Probar $H_{0a}: C_1 B = D$ en el modelo original es equivalente a probar en el modelo (4.3-40) $\Delta_1 = 0$. Sea la partición de $C^{-1} = (C^{(1)}, C^{(2)})$, y sea

$$Q_1 = I - XC^{(2)}(C^{(2)'}X'XC^{(2)})^{-1}C^{(2)'}X'$$

El estadístico por RV está dado por (Mardia et al., 1989; p.162):

$$\lambda_a^{2/n} = \frac{|Y'_+ Q Y_+|}{|Y'_+ Q_1 Y_+|}. \quad (4.3-41)$$

en donde $Q = I - P_X$ como se definió en (4.3-5). Se hace también $Q_2 = Q_1 - Q$ y $QY = QY_+$ ya que $QX = 0$. (Observe que $QY_+ = Q(Y - XB_0) = QY - QXB_0 = QY$)

Resultado 4.3.6.1. La PRV de $H_0: C_1 B = D$ para el modelo $Y_+ = Z\Delta + U$, tiene como estadístico de prueba (EP):

$$\lambda_a^{2/n} = \frac{|Y'QY|}{|Y'QY + Y'_+ Q_2 Y_+|} \quad (4.3-42)$$

que bajo H_{0a} se distribuye como una $\Lambda(p, n - m, g)$. (Observe que (4.3-41) y (4.3-42) son equivalentes, puesto que $Y'_+ Q Y_+ = (Y' - B'_0 X')Q(Y - XB_0) = Y'Q(Y - XB_0) = Y'QY$ y $Y'QY + Y'_+ Q_2 Y_+ = Y'_+ Q Y_+ + Y'_+ Q_2 Y_+ = Y'_+(Q + Q_2)Y_+$)

$= \mathbf{Y}'_+ \mathbf{Q}_1 \mathbf{Y}_+$. La regla de decisión para una prueba de tamaño α es: rechazar H_{0a} si $\lambda_a^{2/n}$ es menor que $\Lambda(p, n-m, g; \alpha)$

Resultado 4.3.6.2. La PRV de $H_{0b}: \mathbf{C}_1 \mathbf{B} \mathbf{M}_1 = \mathbf{D}$ para el modelo $\mathbf{Y} = \mathbf{X} \mathbf{B} + \mathbf{U}$, con $\mathbf{X}_{n \times m}$, está dado por el EP:

$$\lambda_b^{2/n} = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|} = \prod_{i=1}^r (1 + \lambda_i)^{-1} \sim \Lambda(r, n-m, g) \quad (4.3-43)$$

donde $\mathbf{H} = \mathbf{M}'_1 \mathbf{Y}'_+ \mathbf{Q}_2 \mathbf{Y}_+ \mathbf{M}_1 = \mathbf{M}'_1 \mathbf{Y}'_+ \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{C}'_1 (\mathbf{C}_1(\mathbf{X}'\mathbf{X})^{-1} \mathbf{C}'_1)^{-1} \mathbf{C}_1(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \mathbf{Y}_+}_{\mathbf{Q}_2} \mathbf{M}_1$,

$\mathbf{E} = \mathbf{M}'_1 \mathbf{Y}'_+ \mathbf{Q} \mathbf{Y} \mathbf{M}_1 = \mathbf{M}'_1 \mathbf{Y}'_+ (\mathbf{I} - \mathbf{P}_X) \mathbf{Y} \mathbf{M}_1$ y $\lambda_1, \dots, \lambda_r$ son los valores característicos de $\mathbf{H} \mathbf{E}^{-1}$. La regla de decisión con un nivel de significancia α es rechazar $H_{0b}: \mathbf{C}_1 \mathbf{B} \mathbf{M}_1 = \mathbf{D}$ si $\lambda_b^{2/n} < \Lambda(r, n-m, g; \alpha)$; de otro modo, no rechazar H_{0b} .

Cuando $p=1$, se usa equivalentemente el siguiente estadístico de prueba, cuya distribución F está dado por la siguiente proporcionalidad:

$$\frac{\mathbf{H}}{\mathbf{E}} \sim \frac{g}{n-q} F_{g, n-m} \quad (4.3-44)$$

donde $\mathbf{H}$ y $\mathbf{E}$ son escalares. En este caso la regla de decisión con un nivel de significancia α es: rechazar la hipótesis nula si $\mathbf{H}/\mathbf{E} \geq [g/(n-m)] F_{g, n-m, \alpha}$; de otro modo no rechazar.

Los detalles de estos resultados se presentan en Mardia *et al.* (1989, pp. 162-163) en la demostración de los Teoremas 6.3.1 y 6.3.2.

4.3.6.2 La prueba de unión intersección (PUI)

Se parte del hecho de que la prueba de hipótesis multivariada $C_1 B M_1 = D$ es verdadera si y sólo si $b' C_1 B M_1 a = b' D a$ para cualesquiera dos vectores a y b ; de este argumento se deriva que el estadístico de prueba por el enfoque unión intersección, es λ_1 , el valor característico más grande de HE^{-1} ; sea $\theta = \lambda_1 / (1 + \lambda_1)$ la raíz más grande de $H(H + E)^{-1}$ que se distribuye como $\theta(r, n - m, g)$, entonces la regla de decisión es rechazar H_{0b} si $\theta > \theta(r, n - m, g)$. Además, se presentan los siguientes resultados (explicados en Mardia *et al.*, 1989; p. 163):

Resultado 4.3.6.3. Para la hipótesis $H_{0b}: C_1 B M_1 = D$, si $\text{rango}(M_1) = r = 1$, entonces la PUI es equivalente a la PRV.

Resultado 4.3.6.4. Para la hipótesis $H_{0b}: C_1 B M_1 = D$, si $r(C_1) = g = 1$, entonces la PUI y la PRV también coinciden; ya que en este caso $r(H) = r(HE^{-1}) = 1$.

Resultado 4.3.6.5. El único valor característico diferente de cero de HE^{-1} es su traza (tr), i.e. $\lambda_1 = \text{tr}(HE^{-1})$. Sea $d = X(X'X)^{-1}C_1'$. Bajo $H_{0b}: C_1 B M_1 = D$, $M_1'Y_+d$ se distribuye como una $N_p(0, (d'd)M_1'\Sigma M_1)$ y tanto bajo H_{0b} como bajo la hipótesis alternativa $H_{1b}: C_1 B M_1 \neq D$, $E \sim W_r(M_1'\Sigma M_1, n - m)$ independientemente. Por lo tanto, bajo la hipótesis nula λ_1 tiene la siguiente distribución:

$$(n - m)^{-1} T_{r, n - m}^2 = \left(r / (n - m - r + 1) \right) F_{r, n - m - r + 1} \quad (4.3-45)$$

La regla de decisión es rechazar $H_{0b}: C_1 B M_1 = D$ si $\lambda_1 > (r / (n - m - r + 1)) F_{r, n - m - r + 1; \alpha}$; no rechazar de otro modo.

4.3.6.3 Intervalos de confianza simultáneos (ICS)

Suponer que $\mathbf{B}$ es el valor verdadero de la matriz de parámetros de regresión y sea $\mathbf{Y}_+ = \mathbf{Y} - \mathbf{XB}$. Por el principio de unión intersección, los intervalos de confianza simultáneos (ICS) de $\mathbf{b}'\mathbf{C}_1\mathbf{B}\mathbf{M}_1\mathbf{a}$ al $100(1-\alpha)\%$ para cualesquier vectores $\mathbf{a}$, $\mathbf{b}$, están dados por la siguiente expresión:

$$\mathbf{b}'\mathbf{C}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{M}_1\mathbf{a} \pm \left\{ \frac{\theta(r, n-m, g; \alpha)}{1-\theta(r, n-m, g; \alpha)} (\mathbf{a}'\mathbf{E}\mathbf{a}) \left[\mathbf{b}'\mathbf{C}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}_1'\mathbf{b} \right] \right\}^{1/2}, \quad (4.3-46)$$

donde $\theta(r, n-m, g; \alpha)$ denota el valor crítico superior al nivel $100\alpha\%$ de la distribución θ ; para simplificar la notación se escribirá θ_α en lugar de $\theta(r, n-m, g; \alpha)$ en lo que sigue. Como $\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, $\mathbf{b}'\mathbf{C}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{M}_1\mathbf{a} = \mathbf{b}'\mathbf{C}_1\hat{\mathbf{B}}\mathbf{M}_1\mathbf{a}$ es un estimador insesgado de $\mathbf{b}'\mathbf{C}_1\mathbf{B}\mathbf{M}_1\mathbf{a}$.

Cuando los valores de $\mathbf{a}$ o $\mathbf{b}$ son establecidos *a priori*, los intervalos de confianza en (4.3-46) pueden ser restringidos. Si se fija $\mathbf{a}$, los ICS para $\mathbf{b}$ pueden obtenerse sustituyendo $\theta_\alpha/(1-\theta_\alpha)$ por $[g/(n-m)]F_{g, n-m; \alpha}$. Si se fija $\mathbf{b}$, los ICS para $\mathbf{a}$ se pueden construir reemplazando $\theta_\alpha/(1-\theta_\alpha)$ por $(n-m)^{-1}T_{r, n-m; \alpha}^2 = [r/n-m-r+1]F_{r, n-m-r+1; \alpha}$. Cuando tanto $\mathbf{a}$ como $\mathbf{b}$ son conocidos, se sustituye $\theta_\alpha/(1-\theta_\alpha)$ por $(n-m)^{-1}F_{1, n-m; \alpha}$, para obtener un intervalo de confianza individual.

4.3.7 Matrices Diseño de Rango Incompleto

(Mardia *et al.*, 1989; pp. 164-166)

En los diseños experimentales es común considerar matrices diseño $\mathbf{X}$ de rango incompleto. En este caso, la matriz $\mathbf{B}$ en el modelo (4.3-17) no está bien definida. Una forma de resolver la indeterminación, consiste en *restringir la regresión* a un subconjunto

de las columnas de la matriz X de orden $(n \times m)$; aquí, como en la Sección 4.3.7, $m = q + 1$.

Suponer $X_{n \times m}$ tiene $r(X) = k$, $k < m$. Suponer también que se ordenan las columnas de X y se particiona como $X = (X_1, X_2)$ de tal forma que $X_1(n \times k)$ es de rango completo por columnas; así que $X_2(n \times (m - k))$ puede escribirse como combinaciones lineales de las columnas de X_1 , $X_2 = X_1 A$ para alguna matriz $A_{k \times (m - k)}$. También se particiona la matriz de parámetros $B' = (B'_1, B'_2)$. Si el modelo de regresión (4.3-17) es válido, con estas nuevas condiciones puede ser reescrito sólo en términos de X_1 quedando como:

$$Y = X_1 B_1^* + U, \quad (4.3-47)$$

donde $B_1^* = B_1 + AB_2$ es unívocamente determinada.

Para la hipótesis $C_1 B M_1 = D$ y la partición $C_1 = (C_{11}, C_{12})$ conformable con $B' = (B'_1, B'_2)$, C_1 se considera verificable si, como función de B ,

$$XB = 0 \Rightarrow C_1 B = 0.$$

Si $\hat{B}_1^*$ es el estimador de máxima verosimilitud de B_1^* en el modelo (4.3-47) y si C_1 es verificable, entonces el único estimador de máxima verosimilitud de $C_1 B$ está dado por:

$$C_{11} \hat{B}_1^* = C_{11} (X_1' X_1)^{-1} X_1' Y$$

Para la estimación de combinaciones lineales estimables y la prueba de hipótesis verificables en matrices diseño de rango incompleto, es más conveniente usar el modelo (4.3-47) en lugar de (4.3-17).

Pueden verse algunos ejemplos relacionados a este tema en la Sección 4.4.

4.3.8 Correlación Múltiple (*ibíd.*, pp. 167-171)

4.3.8.1 El efecto de la media

En el modelo (4.3-17), la matriz $X_{n \times m}$ de rango completo contiene q columnas que corresponden a las variables independientes más la primera columna de puros unos, de tal manera que usando la partición de $X = [1, X_1]$ se puede escribir el modelo de regresión como:

$$Y = 1\mu' + X_1B + U \quad (4.3-48)$$

Suponer que todas las columnas de X_1 están centradas de modo que su media es cero, es decir,

$$X_1 = \begin{bmatrix} X_{11} - \bar{X}_1 & X_{12} - \bar{X}_2 & \cdots & X_{1q} - \bar{X}_q \\ X_{21} - \bar{X}_1 & X_{22} - \bar{X}_2 & \cdots & X_{2q} - \bar{X}_q \\ \vdots & \vdots & & \vdots \\ X_{n1} - \bar{X}_1 & X_{n2} - \bar{X}_2 & \cdots & X_{nq} - \bar{X}_q \end{bmatrix}_{n \times q}, \quad (4.3-49)$$

donde $\sum_{i=1}^n (X_{ij} - \bar{X}_j) = 0$ implica que $(x'_{(j)} - \bar{X}_j 1')1 = 0$, con $\bar{X}_j = n^{-1} \sum_{i=1}^n X_{ij}$. Entonces,

$$\left[\begin{pmatrix} 1' \\ X_1' \end{pmatrix} (1, X_1) \right]^{-1} = \begin{pmatrix} n^{-1} & 0' \\ 0 & (X_1' X_1)^{-1} \end{pmatrix} \quad (4.3-50)$$

Por el resultado (4.3-25a), $\hat{\mu} = \bar{y}$ es independiente de $\hat{B} = (X_1' X_1)^{-1} X_1' Y = (X_1' X_1)^{-1} X_1' (Y - 1\bar{y}')$. Si $\bar{y}$ (el vector columna que contiene las medias de las columnas de Y) es un estimador de la media de Y , entonces el modelo (4.3-48) puede replantearse como:

$$Y - 1\bar{y}' = X_1 B + W, \quad (4.3-51)$$

donde $W = U - 1\bar{u}'$ es la matriz de errores centrada. Si $W' = [w'_{(1)}, w'_{(2)}, \dots, w'_{(p)}]'$ es un vector que contiene las columnas de W una debajo de otra, de la primera a la última,

entonces $W^V \sim N_{np}(0, \Sigma \otimes H)$ con $H = I - n^{-1}11'$. Recuerde que $\Sigma \otimes H$ es el producto Kronecker de Σ y H .

4.3.8.2 Coeficiente de correlación múltiple

Suponer $p = 1$ de tal forma que el modelo (4.3-48) sea:

$$y - \bar{y}1 = X_1\beta + v, \quad (4.3-52)$$

donde el vector de errores centrado tiene distribución $N_n(0, \sigma^2 H)$ y X_1 tiene las columnas centradas para tener media cero, como en (4.3-49). La matriz de covarianzas muestral para y y las columnas de X_1 se define como:

$$S = \begin{pmatrix} s_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = n^{-1} \begin{pmatrix} y' - \bar{y}1' \\ X_1' \end{pmatrix} (y - \bar{y}1, X_1); \quad (4.3-53)$$

denótese por R a la correspondiente matriz de correlaciones. La correlación muestral entre y y $X_1\hat{\beta}$ se conoce como el *coeficiente de correlación múltiple de y con X_1* , y se denota por $R_{y,x}$. Por tanto:

$$\begin{aligned} R_{y,x} = \text{corr}(y, X_1\hat{\beta}) &= \frac{S_{12}S_{22}^{-1}S_{21}}{(s_{11}S_{12}S_{22}^{-1}S_{21})^{1/2}} \\ &= \left(\frac{S_{12}S_{22}^{-1}S_{21}}{s_{11}} \right)^{1/2}, \end{aligned} \quad (4.3-54)$$

usando el hecho de que $\hat{\beta} = (X_1'X_1)^{-1}X_1'y = (X_1'X_1)^{-1}X_1'(y - \bar{y}1) = S_{22}^{-1}S_{21}$. El cuadrado del coeficiente de correlación múltiple se puede presentar en términos de la matriz de correlaciones R como:

$$R_{y,x}^2 = R_{12}R_{22}^{-1}R_{21}. \quad (4.3-55)$$

Si $\hat{y} = X_1 \hat{\beta}$ es el valor estimado de y , y $\hat{u} = y - \bar{y}1 - X_1 \hat{\beta}$ es el vector de los residuales estimados, la suma de cuadrados total (corregida por la media) puede descomponerse como:

$$y'y - n\bar{y}^2 = \hat{y}'\hat{y} + \hat{u}'\hat{u},$$

que también puede escribirse como:

$$ns_y^2 = ns_y^2 R_{y,x}^2 + ns_y^2 (1 - R_{y,x}^2). \quad (4.3-56)$$

Observando en (4.3-56) que $R_{y,x}^2 + (1 - R_{y,x}^2) = 1$ se puede concluir que el cuadrado del coeficiente de correlación múltiple, expresa la proporción de la suma de cuadrados total explicada por la regresión con las q variables independientes arregladas en las columnas de X_1 .

4.3.8.3 Coeficiente de correlación parcial

Algunas veces es de interés estudiar la correlación entre dos variables después de descontar el efecto de otras variables. Considerar el modelo de regresión multivariado (4.3-51) con $p = 2$ variables dependientes y centrar las columnas de Y de tal forma que $\bar{y}_1 = \bar{y}_2 = 0$. Cada columna de Y puede escribirse como:

$$y_{(i)} = \hat{y}_i + \hat{u}_{(i)}, \quad i = 1, 2,$$

donde $\hat{y}_i = X_1(X_1'X_1)^{-1}X_1'y_{(i)}$ es el “valor ajustado” de $y_{(i)}$ y $\hat{u}_{(i)}$ es el “valor del residual” de $y_{(i)}$. Ya que tanto $\hat{u}_{(1)}$ como $\hat{u}_{(2)}$ son residuales después de ajustar X_1 , entonces una correlación entre $y_{(1)}$ y $y_{(2)}$ después de “eliminar” el efecto de X_1 puede definirse como:

$$r_{12 \cdot x} = \frac{\hat{u}_{(1)}' \hat{u}_{(2)}}{\|\hat{u}_{(1)}\| \|\hat{u}_{(2)}\|} \quad (4.3-57)$$

El coeficiente $r_{12 \cdot x}$ se conoce como el *coeficiente de correlación parcial (muestral) entre $y_{(1)}$ y $y_{(2)}$ dada X_1* .

Para calcular el coeficiente de correlación parcial, se particiona la matriz de covarianzas de Y y X_1 de la siguiente forma:

$$S = \begin{pmatrix} s_{11} & s_{12} & s'_1 \\ s_{21} & s_{22} & s'_2 \\ s_1 & s_2 & S \end{pmatrix} = n^{-1} \begin{pmatrix} y'_{(1)} \\ y'_{(2)} \\ X_1 \end{pmatrix} \begin{pmatrix} y_{(1)} & y_{(2)} & X_1 \end{pmatrix}.$$

Haciendo $s_{ij}^* = s_{ij} - s'_i S_{22}^{-1} s_j$, $i, j = 1, 2$, se tiene que:

$$r_{12 \cdot x} = \frac{s_{12}^*}{(s_{11}^* s_{22}^*)^{1/2}}. \quad (4.3-58)$$

Si se particiona R de forma similar que a S , y se hace $r_{ij}^* = r_{ij} - r_i R_{22}^{-1} r_j$, $i, j = 1, 2$, entonces puede escribirse $r_{12 \cdot x}$ como:

$$r_{12 \cdot x} = \frac{r_{12}^*}{(r_{11}^* r_{22}^*)^{1/2}}. \quad (4.3-59)$$

En general, si una matriz de datos $X_{l \times m}$ es particionada en dos grupos de variables como $X = (X_1 \ X_2)$ y tiene como matriz de covarianza a S , se puede definir la matriz de covarianzas "parcial" de X_1 después de descontar el efecto del grupo de variables X_2 como:

$$S_{11 \cdot 2} = S_{11} - S_{12} S_{22}^{-1} S_{21}, \quad (4.3-60)$$

donde $S_{11 \cdot 2}$ puede interpretarse como la matriz de covarianzas muestral de X_1 dado X_2 .

4.3.8.4 Medidas de correlación entre vectores

En el modelo de regresión univariada (4.3-52) donde los valores de y están centrados por la media de tal forma que $\bar{y} = 0$, se vio que el cuadrado del coeficiente de correlación

múltiple, R^2 , representa la “proporción de la varianza explicada” por el modelo, y de la expresión (4.3-56) se tiene que:

$$1 - R^2 = \frac{\hat{\mathbf{u}}' \hat{\mathbf{u}}}{\hat{\mathbf{y}}' \hat{\mathbf{y}}}.$$

Para buscar una medida análoga para la correlación multivariada entre las matrices $\mathbf{X}$ y $\mathbf{Y}$ en el modelo (4.3-17) se hace lo siguiente. Asumir que $\mathbf{Y}$ está centrada de forma que sus columnas tienen media cero, y escribir $\mathbf{D} = (\mathbf{Y}'\mathbf{Y})^{-1} \hat{\mathbf{U}}' \hat{\mathbf{U}}$, que es la generalización de $1 - R^2$ en el caso univariado. Los valores de la matriz de las SCPC de los errores $\hat{\mathbf{U}}' \hat{\mathbf{U}} = \mathbf{Y}'(\mathbf{I} - \mathbf{P}_X)\mathbf{Y}$ tienden a ceros cuando toda la variación en $\mathbf{Y}$ es “explicada” por el modelo, y viceversa, los valores de $\mathbf{Y}'\mathbf{Y}$ se acercan a ceros cuando casi nada de la variación en $\mathbf{Y}$ es explicada por el modelo. Por lo tanto $(\mathbf{I} - \mathbf{D})$ oscila entre la matriz identidad $\mathbf{I}$ y la matriz cero $\mathbf{O}$. Cualquier medida sensata de la correlación multivariada debe oscilar entre uno y cero y dos de los coeficientes frecuentemente usados que satisfacen esta propiedad son: (a) la correlación traza r_T y (b) la correlación determinante r_D :

$$r_T^2 = p^{-1} \text{tr}(\mathbf{I} - \mathbf{D}), \quad r_D^2 = \det(\mathbf{I} - \mathbf{D}). \quad (4.3-61)$$

Si $d_1, \dots, d_p$ son los valores característicos de $\mathbf{I} - \mathbf{D}$, entonces:

$$r_T^2 = p^{-1} \sum d_i, \quad r_D^2 = \prod d_i$$

Otra medida propuesta por Hotelling en 1936 conocido como “coeficiente de alienación de vector”, es equivalente al determinante de $\mathbf{D}$ que en términos de los valores característicos de $\mathbf{I} - \mathbf{D}$ es:

$$r_A = \prod (1 - d_i),$$

cuyos valores van de cero, cuando $\hat{\mathbf{U}} = \mathbf{O}$, a uno, cuando $\hat{\mathbf{U}} = \mathbf{Y}$.

4.3.9 Anexo de Datos y Programas

En el Cuadro 4.3.1, se reporta un conjunto de 35 árboles muestreados; cada observación contiene el diámetro normal en centímetros (D), la altura en metros (H), y el volumen en metros cúbicos (V).

Cuadro 4.3.1 Diámetro normal, altura y volumen de una muestra de 35 árboles

<i>i</i> -ésimo árbol	Diámetro en cm (D)	Altura en m (H)	Volumen en m ³ (V)	*Variable D^2H
1	0.13	4	0.041	0.068
2	0.14	5	0.040	0.098
3	0.14	7	0.058	0.137
4	0.15	8	0.092	0.180
5	0.16	8	0.095	0.205
6	0.17	5	0.076	0.145
7	0.17	7	0.069	0.202
8	0.17	7	0.092	0.202
9	0.17	9	0.104	0.260
10	0.18	6	0.102	0.194
11	0.18	8	0.101	0.259
12	0.18	8	0.113	0.259
13	0.19	6	0.099	0.217
14	0.19	7	0.117	0.253
15	0.19	8	0.136	0.289
16	0.20	9	0.145	0.360
17	0.20	10	0.148	0.400
18	0.21	7	0.146	0.309
19	0.21	10	0.186	0.441
20	0.21	14	0.232	0.617
21	0.22	8	0.147	0.387
22	0.22	27	0.347	1.307
23	0.23	12	0.374	0.635
24	0.24	8	0.181	0.461
25	0.24	14	0.331	0.806
26	0.25	6	0.219	0.375
27	0.25	9	0.290	0.563
28	0.25	14	0.256	0.875
29	0.26	11	0.327	0.744
30	0.26	16	0.421	1.082
31	0.27	9	0.256	0.656
32	0.27	15	0.478	1.094
33	0.28	12	0.383	0.941
34	0.28	13	0.379	1.019
35	0.28	15	0.515	1.176

*La variable D^2H fue generada a partir de los datos de la segunda y tercera columnas

Cuadro 4.3.2 Peso verde de la leña (en kg) de ramas de *Gliricidia sepium*

Largo	Diámetro (centímetros)								
(metros)	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0.4924	0.5267	0.5747	0.6365	0.7120	0.8012	0.9041	1.0208	1.1512
1.5	0.5061	0.5576	0.6296	0.7229	0.8355	0.9693	1.1237	1.2987	1.4943
2.0	0.5199	0.5885	0.6846	0.8081	0.9590	1.1374	1.3433	1.5766	1.8374
2.5	0.5336	0.6193	0.7394	0.8938	1.0825	1.3055	1.5629	1.8545	2.1805
3.0	0.5473	0.6502	0.7943	0.9796	1.2061	1.4738	1.7825	2.1324	2.5236
3.5	0.5610	0.6811	0.8492	1.0654	1.3296	1.6418	2.0020	2.4103	2.8667
4.0	0.5748	0.7120	0.9041	1.1512	1.4531	1.8099	2.2217	2.6882	3.2094
4.5	0.5885	0.7429	0.9590	1.2379	1.5766	1.9780	2.4412	2.9661	3.5529
5.0	0.6022	0.7737	1.0139	1.3227	1.7001	2.1471	2.6608	3.2441	3.8960
5.5	0.6159	0.8046	1.0688	1.4082	1.8236	2.3143	2.8804	3.5220	4.2391
6.0	0.6970	0.8355	1.1237	1.4943	1.9472	2.4824	3.1000	3.7998	4.5822

Fuente: Datos tomados de Montagnini *et al.* (1992)Cuadro 4.3.3 Peso verde del forraje (en kg) de ramas de *Gliricidia sepium*

Largo	Diámetro (en centímetros)								
(metros)	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
1.0	0.0496	0.0904	0.1475	0.2209	0.3107	0.4167	0.5391	0.6778	0.8328
1.5	0.1128	0.1491	0.1999	0.2652	0.3450	0.4393	0.5482	0.6715	0.8094
2.0	0.1761	0.2078	0.2523	0.3095	0.3794	0.4620	0.5573	0.6653	0.7860
2.5	0.2393	0.2620	0.3047	0.3538	0.4137	0.4846	0.5663	0.6590	0.7626
3.0	0.3026	0.3253	0.3572	0.4080	0.4481	0.5072	0.5755	0.6528	0.7392
3.5	0.3658	0.3840	0.4095	0.4423	0.4824	0.5298	0.5845	0.6465	0.7158
4.0	0.4291	0.4427	0.4620	0.4866	0.5169	0.5525	0.5937	0.6402	0.6924
4.5	0.4923	0.5015	0.5144	0.5309	0.5512	0.5751	0.6027	0.6340	0.6690
5.0	0.5556	0.5602	0.5668	0.5752	0.5856	0.5977	0.6119	0.6277	0.6456
5.5	0.6187	0.6189	0.6192	0.6195	0.6199	0.6204	0.6209	0.6215	0.6185
6.0	0.6820	0.6776	0.6716	0.6638	0.6543	0.6430	0.6300	0.6152	0.5988

Fuente: Datos tomados de Montagnini *et al.* (1992)

Programa 4.3.9.1¹ El análisis de regresión con los datos del Cuadro 4.3.1 usando el modelo 4.3-15, se puede realizar con un programa en IML (Interactive Matrix Language) de SAS como el siguiente, cuya salida se reporta en el Ejemplo 4.3.1.

```
DATA UNO;
INFILE 'C:\REGMUL\VOLTA1.DAT'; /*se llama el archivo VOLTA1.DAT
                                que esta en el subdirectorio REGMUL*/
INPUT D H V; /*el archivo contiene tres columnas que corresponden a D: diametro, H: altura,
              V: volumen*/
D2H=(D**2)*H; /*genera la variable combinada*/

PROC IML; /*se inicia el procedimiento IML */
USE UNO; /*se llama el conjunto de datos UNO */
START REGRE; /*empieza el modulo REGRE */
READ ALL VAR {V} INTO Y; /*se arreglan los valores del volumen
                          en un vector Y */
READ ALL VAR {D2H} INTO X; /*se arreglan los valores de la variable D2H en X; aqui se
                           incluirian las otras variables independientes */

N=NROW(X); /*numero de arboles (de individuos) */
M=J(N,1,1); /*un vector de puros unos de orden N*/
X=M||X; /*se anexa la primera columna de unos en la
        matriz de variables independientes; es la
        que corresponde a la interceptada */

K=NCOL(X); /*grados de libertad de X */
XX=X`*X; /*matriz de productos cruzados */
INVXX=INV(XX); /*inversa de la matriz de productos cruzados */
B=INVXX*X`*Y; /*el vector de parametros estimados */

PX=X*INVXX*X`; /*proyector de X: matriz simetrica e idempotente*/
Q=I(N)-PX;
FC=Y`*(1/N)*J(N,N,1)*Y; /*factor de correccion FC */
SCREG=(Y`*PX*Y)-FC; /*suma de cuadrados de regresion corregida*/
SCE=Y`*Q*Y; /*suma de cuadrados del error */
SCTOTAL=Y`*Y-FC; /*suma de cuadrados total corregida */

GLE=N-K; /*grados de libertad del error */
CMREG=SCREG/1; /*cuadrado medio de regresion */
CME=SCE/GLE; /*cuadrado medio del error=varianza estimada */
RAIZCME=SQRT(CME); /*raiz cuadrada del CME */
FCAL=CMREG/CME; /*estadistico de prueba del modelo */
R2=SCREG/SCTOTAL; /*coeficiente de determinacion */
PRINT B SCREG SCE SCTOTAL CMREG CME RAIZCME FCAL R2;

COVB=INVXX#CME; /*covarianza de los parametros estimados */
EESTB=SQRT(VECDIAG(COVB)); /*error estandar de los parametros estimados*/
T=B/EESTB; /*prueba t-Student para la hipotesis de que
            los estimadores son igual a cero */
```

¹ Para los programas de este apartado, véase la introducción de la Sección 4.2.6.

```

PROBT=1-PROBF(T#T,1,GLE);          /*nivel de significancia estimado */
PRINT B EESTB T PROBT;
S=DIAG(1/EESTB);
CORRB=S*COVB*S;                    /*correlacion de los parametros estimados */
PRINT COVB CORRB;

IF NROW(TVAL)=0 THEN RETURN;
PREDI=X*B;                          /*genera los volumen predichos*/
DIFER=Y-PREDI;                      /*error estimado*/
VRES=Q*CME;                         /*varianza de residuales*/
VPREDI=PX#CME;                      /*varianza de valores predichos*/
H=VECDIAG(PX);
LOWERM=PREDI-TVAL#SQRT(H*CME);
UPPERM=PREDI+TVAL#SQRT(H*CME);
LOWER=PREDI-TVAL#SQRT(H*CME+CME);
UPPER=PREDI+TVAL#SQRT(H*CME+CME);
PRINT Y PREDI DIFER H LOWERM UPPERM LOWER UPPER;

FINISH;                             /*termina el modulo REGRE*/
RUN REGRE;                          /*correr el modulo REGRE */
QUIT;

```

Programa 4.3.9.2 Para el Ejemplo 4.3.3, se hace uso también de los datos del Cuadro 4.3.1, pero haciendo el ajuste con el modelo (4.3-15b). A continuación se escribe el programa en SAS usado, que es muy parecido al del Ejemplo 4.3.1, con las adecuaciones a una regresión múltiple:

```

DATA DOS;
INFILE 'C:\REGMUL\VOLTA1.DAT';      /*se llama el archivo VOLTA1.DAT*/
INPUT D H V;
D2=D**2;                            /*cuadrado del diametro normal*/
D2H=(D**2)*H;                       /*variable combinada*/

PROC REG;
MODEL V=D2 H D2H;

PROC IML;                            /*se inicia el procedimiento IML */
USE DOS;                             /*se llama el conjunto de datos DOS */
START REGRE;                         /*empieza el modulo REGRE2 */
READ ALL VAR {V} INTO Y;             /*se arreglan los valores del volumen
                                     en un vector Y */
READ ALL VAR {D2 H D2H} INTO X;      /*se arreglan los valores de las variables independientes en
                                     X*/
N=NROW(X);                          /*numero de arboles (de individuos) */
M=J(N,1,1);                        /*un vector de puros unos de orden N */

```

```

X=M||X;          /*se anexa la primera columna de unos en la matriz de variables independientes;
                  es la que corresponde a la interceptada */
K=NCOL(X);       /*grados de libertad de X */
XX=X`*X;         /*matriz de productos cruzados */
INVXX=INV(XX);   /*inversa de la matriz de productos cruzados */
B=INVXX*X`*Y;    /*el vector de parametros estimados */

PX=X*INVXX*X`;   /*proyector de X: matriz simetrica e idempotente*/
Q=I(N)-PX;
FC=Y`*(1/N)*J(N,N,1)*Y; /*factor de correccion */
SCREG=(Y`*PX*Y)-FC;     /*suma de cuadrados de regresion corregida */
SCE= Y`*Q*Y;           /*suma de cuadrados del error */
SCTOTAL=Y`*Y-FC;       /*suma de cuadrados total corregida */

GLE=N-K;           /*grados de libertad del error */
CMREG=SCREG/(K-1); /*cuadrado medio de regresion */
CME=SCE/GLE;       /*cuadrado medio del error=varianza estimada */
RAIZCME=SQRT(CME); /*raiz cuadrada del CME */
FCAL=CMREG/CME;    /*estadistico de prueba del modelo */
R2=SCREG/SCTOTAL;  /*coeficiente de determinacion */
PRINT B SCREG SCE SCTOTAL CMREG CME RAIZCME FCAL R2;

COVB=INVXX#CME;    /*covarianza de los parametros estimados */
EESTB=SQRT(VECDIAG(COVB)); /*error estandar de los parametros estimados */
T=B/EESTB;         /*prueba t-Student para la hipotesis de que los estimadores son igual a cero */
PROBT=1-PROBF(T#T,1,GLE); /*nivel de significancia estimado */
PRINT B EESTB T PROBT;
S=DIAG(1/EESTB);
CORRB=S*COVB*S;    /*correlacion de los parametros estimados */
PRINT COVB CORRB;

FINISH;           /*termina el modulo REGRE2 */
RUN REGRE;        /*correr el modulo REGRE2 */
QUIT;

```

Programa 4.3.9.3 Programa para los Ejemplos 4.3.5 y 4.3.6.

```

DATA RMEBL;
INFILE 'C:\REGMUL\SEPIUM.DAT';
INPUT Y1 Y2 L D;
/*Y1:peso verde de lena*/
/*Y2:peso verde del forraje*/
/*L :longitud de rama*/
/*D :diametro de rama*/
x1=L;
X2=D**2*L;

PROC IML;          /*se inicia el procedimiento IML */

```



```

USE RMEBL; /*se llama el conjunto de datos RMEBL */
START REGREMUL; /*empieza el modulo REGREMUL */
READ ALL VAR {Y1 Y2} INTO Y; /*se arreglan los valores de las variables
dependientes en una matriz Y */
READ ALL VAR {X1 X2} INTO X; /*se arreglan los valores de las
variables independientes en X */
N=NROW(X); /*numero de ramas (de individuos) */
M=J(N,1,1); /*un vector de puros unos de orden N */
X=M||X; /*se anexa la primera columna de unos en la
matriz de variables independientes; es la
que corresponde a la interceptada */
K=NCOL(X); /*columnas de X */
XX=X*X; /*matriz de productos cruzados */
INVXX=INV(XX); /*inversa de la matriz de productos cruzados */
B=INVXX*X*Y; /*el vector de parametros estimados */

PX=X*INVXX*X; /*proyector de X:matriz simetrica e idempotente*/
PX1=M*INV(M*M)*M;
Q=I(N)-PX;
FC=Y*(1/N)*J(N,N,1)*Y; /*factor de correccion */
SCREG=(Y*PX*Y)-FC; /*suma de cuadrados de regresion corregida */
SCE=Y*Q*Y; /*suma de cuadrados del error */
SCTOTAL=Y*Y-FC; /*suma de cuadrados total corregida */
B=INVXX*X*Y;

R2=SCREG/SCTOTAL; /*coeficiente de determinacion */
PRINT B SCREG SCE SCTOTAL;
/*para hacer una prueba de hipotesis de la seccion 3.4 necesitamos los siguientes
calculos*/
S1=(1/N)*Y*(I(N)-PX1)*Y;
S=(1/N)*Y*(I(N)-PX)*Y;
DS=DET(S);
DS1=DET(S1);
RAZON=LOG(DS/DS1);
EPM=-(N-2-1-0.5*(2-0))*RAZON;
jic=CINV(0.99,2);

PRINT S S1 EPM JIC;

FINISH; /*termina el modulo REGREMUL */
RUN REGREMUL; /*correr el modulo REGREMUL */
QUIT;

```

Para correr el Programa 4.3.9.3 se organizaron los datos de los Cuadros 4.3.2 y 4.3.3 en el archivo externo SEPIUM.DAT de la siguiente manera.

0.4924 0.0496 1.0 1.0
 0.5267 0.0904 1.0 1.5
 0.5747 0.1475 1.0 2.0
 0.6365 0.2209 1.0 2.5
 0.7120 0.3107 1.0 3.0
 0.8012 0.4167 1.0 3.5
 0.9041 0.5391 1.0 4.0
 1.0208 0.6778 1.0 4.5
 1.1512 0.8328 1.0 5.0
 0.5061 0.1128 1.5 1.0
 0.5576 0.1491 1.5 1.5
 0.6296 0.1999 1.5 2.0
 0.7229 0.2652 1.5 2.5
 0.8355 0.3450 1.5 3.0
 0.9693 0.4393 1.5 3.5
 1.1237 0.5482 1.5 4.0
 1.2987 0.6715 1.5 4.5
 1.4943 0.8094 1.5 5.0
 0.5199 0.1761 2.0 1.0
 0.5885 0.2078 2.0 1.5
 0.6846 0.2523 2.0 2.0
 0.8081 0.3095 2.0 2.5
 0.9590 0.3794 2.0 3.0
 1.1374 0.4620 2.0 3.5
 1.3433 0.5573 2.0 4.0
 1.5766 0.6653 2.0 4.5
 1.8374 0.7860 2.0 5.0
 0.5336 0.2393 2.5 1.0
 0.6193 0.2620 2.5 1.5
 0.7394 0.3047 2.5 2.0
 0.8938 0.3538 2.5 2.5
 1.0825 0.4137 2.5 3.0
 1.3055 0.4846 2.5 3.5
 1.5629 0.5663 2.5 4.0

1.8545 0.6590 2.5 4.5
 2.1805 0.7626 2.5 5.0
 0.5473 0.3026 3.0 1.0
 0.6502 0.3253 3.0 1.5
 0.7943 0.3572 3.0 2.0
 0.9796 0.4080 3.0 2.5
 1.2061 0.4481 3.0 3.0
 1.4738 0.5072 3.0 3.5
 1.7825 0.5755 3.0 4.0
 2.1324 0.6528 3.0 4.5
 2.5236 0.7392 3.0 5.0
 0.5610 0.3658 3.5 1.0
 0.6811 0.3840 3.5 1.5
 0.8492 0.4095 3.5 2.0
 1.0654 0.4423 3.5 2.5
 1.3296 0.4824 3.5 3.0
 1.6418 0.5298 3.5 3.5
 2.0020 0.5845 3.5 4.0
 2.4103 0.6465 3.5 4.5
 2.8667 0.7158 3.5 5.0
 0.5748 0.4291 4.0 1.0
 0.7120 0.4427 4.0 1.5
 0.9041 0.4620 4.0 2.0
 1.1512 0.4866 4.0 2.5
 1.4531 0.5169 4.0 3.0
 1.8099 0.5525 4.0 3.5
 2.2217 0.5937 4.0 4.0
 2.6882 0.6402 4.0 4.5
 3.2094 0.6924 4.0 5.0
 0.5885 0.4923 4.5 1.0
 0.7429 0.5015 4.5 1.5
 0.9590 0.5144 4.5 2.0
 1.2379 0.5309 4.5 2.5
 1.5766 0.5512 4.5 3.0

1.9780 0.5751 4.5 3.5
 2.4412 0.6027 4.5 4.0
 2.9661 0.6340 4.5 4.5
 3.5529 0.6690 4.5 5.0
 0.6022 0.5556 5.0 1.0
 0.7737 0.5602 5.0 1.5
 1.0139 0.5668 5.0 2.0
 1.3227 0.5752 5.0 2.5
 1.7001 0.5856 5.0 3.0
 2.1471 0.5977 5.0 3.5
 2.6608 0.6119 5.0 4.0
 3.2441 0.6277 5.0 4.5
 3.8960 0.6456 5.0 5.0
 0.6159 0.6187 5.5 1.0
 0.8046 0.6189 5.5 1.5
 1.0688 0.6192 5.5 2.0
 1.4082 0.6195 5.5 2.5
 1.8236 0.6199 5.5 3.0
 2.3143 0.6204 5.5 3.5
 2.8804 0.6209 5.5 4.0
 3.5220 0.6215 5.5 4.5
 4.2391 0.6185 5.5 5.0
 0.6970 0.6820 6.0 1.0
 0.8355 0.6776 6.0 1.5
 1.1237 0.6716 6.0 2.0
 1.4943 0.6638 6.0 2.5
 1.9472 0.6543 6.0 3.0
 2.4824 0.6430 6.0 3.5
 3.1000 0.6300 6.0 4.0
 3.7998 0.6152 6.0 4.5
 4.5822 0.5988 6.0 5.0

4.4 ANÁLISIS DE VARIANZA MULTIVARIADO

4.4.1 Introducción

Al realizar un experimento, el investigador induce o manipula de forma sistemática las diferentes condiciones que desea comparar en un fenómeno de interés. Esto es, en un experimento se hace variar sistemáticamente una (más de una) variable, llamada(s) independiente(s), con la finalidad de evaluar su(s) efecto(s) en otra(s) variable(s), llamada(s) dependiente(s) o variable(s)-respuesta (Winer *et al.*, 1991). Para esto, se establece(n), desde la planeación del experimento, la(s) (diferentes) variable(s)-respuesta que se quiere(n) estudiar, misma(s) que se registrará(n) en las unidades experimentales (UE) —conocidas como parcelas en el ámbito agronómico.

La función del diseño experimental es describir cómo se va a ejecutar un experimento y con el análisis de varianza se hace el análisis de los datos a través de una descomposición de sumas de cuadrados (Winer *et al.*, 1991). Usando el análisis de varianza, puede analizarse también información que no provenga de un experimento, aunque para que sea posible, es necesario que los datos estén agrupados (Mardia *et al.*, 1989).

La variación total del fenómeno bajo estudio, es representada por la *suma de cuadrados total* (SCT) y ésta, se puede descomponer en dos partes: (i) la variación de las unidades experimentales representada por la *suma de cuadrados de las unidades experimentales* [$SC(UE)$] y (ii) la variación de mediciones de la respuesta dentro de unidades experimentales, representada por la *suma de cuadrados del error de muestreo*

(*SCEM*). Cuando en la toma de datos sólo se planea tomar una medición por parcela de cada variable-respuesta (i.e. cuando no se hace *submuestreo*), se está considerando que no existen diferencias significativas de las observaciones dentro de las unidades experimentales y entonces la $SCEM = 0$.

La $SC(UE)$ se puede descomponer, a su vez, en los siguientes dos grupos de sumas de cuadrados (SC): (i) las SC correspondientes exclusivamente a las fuentes de variación del diseño que se trabaja, más (ii) la *suma de cuadrados del error experimental* ($SCEE$). La $SCEE$ está constituida por, algo así como, los efectos en la respuesta que no se deben necesariamente a las condiciones que el investigador induce, sino que a factores externos —por así llamarlos— desconocidos, o bien, que deliberadamente no se contemplan en el modelo del diseño; sean lo que fueran, estos factores externos se tratan de mantener constantes. Las fuentes exclusivas correspondientes a un diseño son, por ejemplo, en el completamente al azar, los tratamientos; en el bloques al azar, los bloques y los tratamientos; en el cuadro latino, las hileras, las columnas y los tratamientos.

Por el número de variables-respuestas que se consideran en un experimento, los diseños experimentales pueden clasificarse en: univariados o multivariados. En los *diseños univariados*, se estudia una sola variable-respuesta del fenómeno de interés; cuando se consideran simultáneamente las mediciones de diferentes variables-respuesta (como el volumen, ancho de copa, incremento en altura, etc.), el análisis debe hacerse con las técnicas del *análisis de varianza multivariado* (MANOVA, por sus siglas en inglés) y para cada diseño univariado se puede tener de manera directa su extensión multivariada (Mardia, *et al.*, 1989). En este capítulo sólo se presenta la extensión multivariada de los diseños completamente al azar y bloques completos al azar.

El caso de submuestreo —también llamado “de mediciones repetidas” o “con remediciones”—, ya sea en los diseños univariados o en los multivariados, se presenta cuando de cada variable-respuesta se toma aleatoriamente más de una medición o muestra por UE. Estas clasificaciones pueden entenderse con el siguiente diagrama:

Diseño experimental	univariado (se estudia una sola variable-respuesta)	estándar (una sola medición de la variable-respuesta por unidad)
		con submuestreo (se toma más de una muestra de la variable-respuesta por unidad)
		estándar (una sola medición de cada variable-respuesta por unidad)
	multivariado (se estudian múltiples variables-respuesta)	con submuestreo (más de una medición de cada variable-respuesta por unidad)

En este capítulo, primero se exponen los resultados importantes del diseño completamente aleatorio estándar (o diseño de un solo criterio de clasificación) y posteriormente se hace la extensión a su formulación multivariada. Se procede de manera similar con el diseño de bloques completos al azar.

Para obtener información acerca de las condiciones que son recomendables para hacer uso de estos dos diseños, sus ventajas y desventajas, algunos ejemplos prácticos, el proceso de aleatorización adecuado, la presentación escalar de las sumas de cuadrados, pruebas de hipótesis sobre efectos de los tratamientos, casos particulares como el de “datos faltantes” en el bloques al azar, algunas interpretaciones geométricas, etc., se pueden revisar libros conocidos como el de Steel & Torrie (1988), Martínez (1988), Li (1969), Cochran & Cox (1983) y Snedecor & Cochran (1967). En las dos primeras referencias viene también el análisis de estos dos diseños con submuestreo, con la presentación escalar de las sumas de cuadrados.

4.4.2 Diseño Completamente al Azar (DCA)

4.4.2.1 Formulación univariada

Considere que a n unidades experimentales (UE) se les asignan aleatoriamente k tratamientos, $j = 1, 2, \dots, k$ y suponer que el j -ésimo tratamiento se aplica a n_j unidades experimentales de tal modo que $\sum_{j=1}^k n_j = n$

EJEMPLO 4.4.1 Como parte de una investigación para determinar la base química de la inmunidad del cedro australiano (*Toona ciliata* var *australis*) al barrenador de las Meliáceas (*Hypsipyla grandella*), Grijpma & Roberts (1976) condujeron un experimento de laboratorio para probar el efecto de seis tratamientos (incluyendo el testigo o control) sobre el peso en miligramos de pupas de *H. grandella*. Los tratamientos consistieron en mezclas de los ingredientes de la dieta artificial y de los extractos —obtenidos mediante una serie de solventes orgánicos— de hojas de *Toona* (T) y *Cedrela* (C), teniendo que la concentración de los extractos químicos fue equivalente a aproximadamente un gramo de hojas de *Toona* o *Cedrela* por 8 g de dieta artificial; el tratamiento testigo sólo contenía los ingredientes de la dieta artificial. Originalmente se tenían 120 pupas (que son las UE) a las que les asignaron aleatoriamente los $k = 6$ tratamientos; para motivos de este ejemplo, se puede suponer que las características de las pupas en laboratorio eran sensiblemente iguales, por lo que habían razones suficientes para considerar un DCA que exige homogeneidad del material experimental. Aunque inicialmente cada tratamiento se asignó a 20 pupas (i.e. a 20 UE), al final del experimento, y debido a las que murieron en el proceso, el número de UE por tratamiento quedó de la siguiente manera: $n_1 = 19$, $n_2 = 19$, $n_3 = 18$, $n_4 = 18$, $n_5 = 16$ y $n_6 = 15$; haciendo un total de $n = 105$ UE. Esto es, al término del experimento, quedaron 19 pupas del tratamiento 1 (testigo), 19 del tratamiento 2, y así enseguida. Los datos recabados que reportan Grijpma & Roberts (1976, p. 108) se presentan en el Cuadro 4.4.1. ♦

Cuadro 4.4.1 Pesos (en mg) de 105 pupas de *H. grandella*

UE <i>i</i>	<i>j</i> -ésimo tratamiento					
	Trat 1	Trat 2	Trat 3	Trat 4	Trat 5	Trat 6
1	140	155	157	154	95	92
2	145	160	148	178	115	58
3	161	159	90	172	120	45
4	174	164	155	180	108	46
5	150	152	180	104	170	80
6	149	141	185	128	112	75
7	162	158	163	160	65	105
8	175	128	155	155	140	85
9	173	165	106	161	90	65
10	165	161	166	155	115	100
11	153	143	105	125	95	95
12	162	163	153	135	125	85
13	146	164	171	108	124	70
14	170	156	185	167	116	120
15	171	139	97	136	98	80
16	116	154	155	133	120	.*
17	158	157	168	134	.*	.*
18	149	158	149	107	.*	.*
19	152	149	.*	.*	.*	.*
$\bar{Y}_{.j}$	156.4	154	149.3	144	113	80.1

El modelo lineal del diseño completamente aleatorio se puede plantear en forma escalar como:

$$Y_{ij} = \mu + \tau_j + \varepsilon_{ij}, \quad (4.4-1)$$

en donde Y_{ij} es la respuesta de la i -ésima unidad experimental, $i = 1, \dots, n_j$, que recibió el j -ésimo tratamiento, $j = 1, 2, \dots, k$; μ es un efecto general sobre la variable respuesta; τ_j es el efecto debido al j -ésimo tratamiento; y ε_{ij} es una variable aleatoria bajo el supuesto que $\varepsilon_{ij} \sim N(0, \sigma_e^2)$ independientemente.

En forma matricial, se puede reescribir el modelo (4.4-1) como sigue:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \boldsymbol{\varepsilon}, \quad (4.4-2)$$

donde $\mathbf{y}$ es el vector-respuesta que contiene a las n Y_{ij} del modelo (4.4-1); la matriz diseño $\mathbf{X}$ se puede particionar en el vector $\mathbf{1}$ formado por los unos de la primera

columna y la matriz de los tratamientos X_2 , i.e. $X = [1 \ X_2]$; $\tau = [\mu \ \tau_1 \ \tau_2 \ \dots \ \tau_k]'$ es el vector de parámetros de dimensión $(k+1) \times 1$. En este modelo, se supone que $\varepsilon \sim N_p(0, \sigma^2 I)$, es decir, se supone que los elementos de ε tienen media cero, varianza común y no están correlacionados.

En el DCA, se considera que las primeras n_1 UE's del vector y recibieron el tratamiento 1, las segundas n_2 UE's el tratamiento 2, así sucesivamente, hasta que las últimas n_k UE's recibieron el tratamiento k , de manera que, la matriz diseño del modelo (4.4-2), se puede escribir como:

$$X = \begin{bmatrix} 1_{n_1} & 1_{n_1} & 0 & \dots & 0 \\ 1_{n_2} & 0 & 1_{n_2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1_{n_k} & 0 & 0 & \dots & 1_{n_k} \end{bmatrix}_{n \times (k+1)} \quad (4.4.3)$$

donde, la primera columna de puros unos corresponde al efecto promedio de la respuesta que se representa con un vector $1'_n = [1'_{n_1}, 1'_{n_2}, \dots, 1'_{n_k}]'$ y las siguientes k columnas representan la presencia o ausencia del j -ésimo tratamiento sobre la i -ésima UE y son las que conforman a la submatriz X_2 .

4.4.2.1.1 Sumas de cuadrados y cuadrados medios

La suma de cuadrados de los tratamientos corregida por la media (SC_{trat}) es:

$$\begin{aligned} SC_{trat} &= y' P_{X_2} y - y' P_1 y \\ &= y' [P_{X_2} - P_1] y \end{aligned} \quad (4.4-4)$$

donde, $P_{X_2} = X_2(X_2'X_2)^{-1}X_2'$ es el proyector de la submatriz X_2 cuyo rango es $r(P_{X_2}) = r(X_2) = k$; recuerde también que $1_{n \times 1}$ es la primera columna de la matriz X y su proyector es $P_1 = 1(1'1)^{-1}1' = \frac{1}{n}11'$ con $r(P_1) = r(1_{n \times 1}) = 1$.

La suma de cuadrados del error experimental ($SCEE$) es:

$$\begin{aligned}
 SCEE &= y' [P_C - P_{X_2}] y \\
 &= y' [I_n - P_{X_2}] y
 \end{aligned}
 \tag{4.4-5}$$

donde $P_C = C(C'C)^{-1}C'$ y C es la siguiente matriz:

$$C = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n},$$

quedando que $C = I_{n \times n}$ y por lo tanto, $P_C = I$. Entonces $r(P_C) = r(I) = n$. En la matriz C , el 1 de la primera columna está asociado a la observación en la primera *UE*, el 1 de la segunda columna corresponde a la observación de la respuesta en la segunda *UE*, y así sucesivamente.

Adicionando la *SCtrat* y la *SCEE* se obtiene la suma de cuadrados total (*SCT*) expresada por:

$$SCT = y' [I_n - P_1] y
 \tag{4.4-6}$$

Usualmente, a la matriz centrante $(I_n - P_1)$ en (4.4-6) se le denota con H .

Los grados de libertad (g.l.) asociados a la *SCtrat* son $r(P_{X_2} - P_1) = (k - 1)$ y los g.l. asociados a la *SCEE* son $r(I_n - P_{X_2}) = (n - k)$. Los *cuadrados medios* [$CM(\cdot)$] se definen como el cociente de una *SC* y sus correspondientes g.l. Así,

$$CM(trat) = SCtrat / (k - 1);
 \tag{4.4-7}$$

$$CM(E) = SCEE / (n - k).
 \tag{4.4-8}$$

4.4.2.1.2 Prueba de hipótesis sobre los tratamientos

Para probar la hipótesis nula $H_0: \tau_1 = \tau_2 = \dots = \tau_k$ contra la hipótesis alternativa $H_1: \tau_i \neq \tau_j$, para alguna $i \neq j$, $i, j = 1, \dots, k$, se usa el estadístico de prueba:

$$F_{cal} = \frac{CM(trat)}{CM(E)} = \frac{(n-k)SC_{trat}}{(k-1)SCEE} \stackrel{H_0}{\sim} F_{k-1, n-k}. \quad (4.4-9)$$

Si $F_{k-1, n-k; \alpha}$ es el valor $100\alpha\%$ superior de una distribución F de Snedecor con $(k-1)$ g.l. en el numerador y $(n-k)$ g.l. en el denominador, la regla de decisión para una prueba de tamaño α es: rechazar la hipótesis nula $H_0: \tau_j = 0, \forall j = 1, \dots, k$ si $F_{cal} \geq F_{k-1, n-k; \alpha}$; no rechazar de otro modo.

EJEMPLO 4.4.2 Para los datos reportados en el Cuadro 4.4.1 del experimento descrito en el Ejemplo 4.4.1, se obtuvieron las siguientes sumas de cuadrados: $SC_{trat} = 73698.49$; $SCEE = 44877.35$; $SCT = 118575.85$. Para probar la hipótesis nula de que “no hay diferencias significativas entre los seis tratamientos”, se calcula el estadístico de prueba dado en la expresión (4.4.9), $F_{cal} = (14739.70/453.31) = 32.52$, que se compara con $F_{5,99;0.01} = 3.21$. Como $F_{cal} = 32.52 > F_{5,99;0.01}$ se concluye con un nivel de significancia $\alpha = 0.01$ que, al menos, existe un tratamiento cuyo efecto es diferente a los demás; esto era de esperarse por haber incluido el “testigo o control” como un tratamiento más. Excluyendo los datos del testigo en el análisis —esto es, probando si no existen diferencias entre los otros 5 tratamientos—, se obtiene que $F_{cal} = 31.17$ que se debe comparar con $F_{4,81;0.01} = 3.56$, lo cual, también permite concluir que hay una diferencia altamente significativa entre los tratamientos hechos a base de mezclas de la dieta artificial y los extractos de hojas. Para detectar qué tratamiento provoca el rechazo de la hipótesis nula se tendrían que hacer comparaciones de medias de tratamientos; para ello existen diferentes métodos que se reportan en la literatura (*vid.* Steel y Torrie, 1988; pp. 166-187). ♦

4.4.2.1.3 Esperanzas de las sumas de cuadrados

Se escriben a continuación, sin demostrar, las esperanzas $[E(\cdot)]$ de la SC_{trat} y la $SCEE$ en el diseño completamente aleatorio (DCA) univariado estándar (Winer *et al.*, 1991; p.267):

$$E(SC_{trat}) = \sigma_e^2(k-1) + \sum_{j=1}^k n_j(\tau_j - \bar{\tau})^2, \quad (4.4-10)$$

donde $\bar{\tau} = \frac{1}{k} \sum_{j=1}^k \tau_j$.

$$E(SCEE) = \sigma_e^2(n-k). \quad (4.4-11)$$

De forma directa se puede ver que las esperanzas de los cuadrados medios son: $E[CM(trat)] = \sigma_e^2 + \frac{1}{k-1} \sum_{j=1}^k n_j(\tau_j - \bar{\tau})^2$ y $E[CM(E)] = \sigma_e^2$. Véase que $CM(E)$ en (4.4-8) es un estimador insesgado de σ_e^2 .

4.4.2.2 Formulación multivariada estándar

Siguiendo a Mardia *et al.* (1989; pp. 333-335), suponer ahora que en lugar de sólo una respuesta como en el modelo (4.4-1), se incluyen $p > 1$ variables-respuesta contenidos en el vector aleatorio $\mathbf{y}'_{ij} = (Y_{ij1}, Y_{ij2}, \dots, Y_{ijp})$, con $i = 1, \dots, n_j$ y $j = 1, \dots, k$; $\boldsymbol{\tau}'_j = (\tau_{j1}, \tau_{j2}, \dots, \tau_{jp})$ es el vector de efectos debidos al j -ésimo tratamiento sobre las p respuestas en la i -ésima unidad, de tal forma que:

$$\mathbf{y}_{ij} = \boldsymbol{\mu} + \boldsymbol{\tau}_j + \boldsymbol{\varepsilon}_{ij}, \quad (4.4-12)$$

donde los $\boldsymbol{\varepsilon}_{ij}$ se distribuyen independientemente con una $N_p(\mathbf{0}, \Sigma)$ y $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)'$ es el efecto general sobre el vector-respuesta $\mathbf{y}_{ij}$. Por lo tanto, los n_j vectores que contienen las p variables-respuesta de las UE que recibieron el j -ésimo tratamiento (i.e. UE tratadas similarmente), $\mathbf{y}_{ij}$, se pueden considerar una muestra aleatoria de una población normal p -variada $N_p(\boldsymbol{\mu}_j, \Sigma)$, donde $\boldsymbol{\mu}_j = \boldsymbol{\mu} + \boldsymbol{\tau}_j$.

EJEMPLO 4.4.3 Ogbonnaya (1994) realizó un estudio para investigar los efectos de fertilizantes de nitrógeno (N) y potasio (K) y sus combinaciones, sobre el crecimiento, propiedades físicas e histoquímicas de plántulas de *Gmelina arborea* en un suelo oxisol. Se trasplantaron plántulas de dos semanas de edad con crecimiento uniforme en altura, poniéndose una sola plántula en cada maceta; las macetas fueron rellenas con suelo oxisol secado al sol durante una semana y con remoción de los materiales vegetales descompuestos. Se aplicaron cuatro niveles de N (0.00, 2.22, 4.44, y 6.66 g /maceta) y K (0.00, 5.70, 11.40 y 17.10 g /maceta) en todas sus combinaciones posibles, haciendo un total de 16 tratamientos; cada tratamiento tuvo cinco repeticiones. La dosis completa de cada nutriente se hizo en una sola aplicación, a una profundidad de 5 cm y alrededor de las plántulas. Después de realizar el experimento se dejaron las plántulas crecer durante 20 semanas, antes de hacer los registros de los valores que tomaron las siguientes variables-respuesta: crecimiento de altura total; diámetro del cuello de la raíz; producción de materia seca; área foliar; gravedad específica; mediciones de la longitud, el diámetro del lumen y el diámetro de la fibra; así como algunas propiedades químicas. El experimento se realizó con un factorial 4 x 4 —i.e., los 2 factores, N y K, tuvieron 4 niveles cada uno— en un diseño completamente al azar; es claro que es un diseño multivariado porque se consideró más de una variable-respuesta. En el Cuadro 4.4.2 se presentan los registros de las tres primeras variables-respuestas, generadas por simulación en la computadora, a partir de los valores de las medias aritméticas y su correspondientes desviaciones estándar que Ogbonnaya (1994) reporta. ♦

Cuadro 4.4.2 Efectos de fertilizantes de N y P y sus combinaciones, sobre tres variables-respuestas de plántulas de *Gmelina arborea*.

Trat	Variable-respuesta 1 Altura total de la plántula (cm)					Variable-respuesta 2 Diám. del cuello radical (mm)				
	UE 1	UE 2	UE 3	UE 4	UE 5	UE 1	UE 2	UE 3	UE 4	UE 5
I	19.56	18.95	22.71	13.41	23.33	9.88	9.74	8.79	9.13	8.91
II	25.12	22.00	22.60	25.26	24.20	4.97	6.29	7.89	7.18	7.38
III	22.41	27.79	25.21	22.58	27.74	8.54	10.07	8.39	7.85	10.88
IV	32.32	30.55	26.59	23.07	23.91	6.34	5.57	6.44	8.36	9.01
V	42.62	36.91	39.31	46.08	32.34	17.33	13.27	16.37	14.50	12.78

Trat	Variable-respuesta 1 Altura total de la plántula (cm)					Variable-respuesta 2 Diám. del cuello radical (mm)				
	UE 1	UE 2	UE 3	UE 4	UE 5	UE 1	UE 2	UE 3	UE 4	UE 5
VI	49.04	50.28	49.21	52.74	54.31	18.73	16.04	10.99	14.83	14.79
VII	51.94	55.54	56.90	47.85	47.23	21.32	12.81	12.13	14.55	10.81
VIII	43.51	32.44	41.86	47.95	45.09	12.06	12.15	13.22	14.94	10.57
IX	50.98	45.23	51.26	47.47	57.90	18.94	12.66	13.95	14.73	12.22
X	50.39	47.08	51.61	71.33	66.06	18.27	14.99	18.73	13.68	19.94
XI	46.41	46.32	42.86	41.99	43.80	14.67	12.33	13.63	12.43	12.93
XII	55.05	63.76	62.51	46.66	47.63	18.95	18.01	21.79	18.27	12.77
XIII	49.49	45.69	49.78	43.96	46.35	16.30	11.87	14.15	15.40	14.19
XIV	67.60	61.53	55.00	55.47	61.12	15.89	16.00	15.73	16.22	15.84
XV	60.91	55.77	68.80	60.48	61.26	15.36	15.21	15.36	15.30	15.18
XVI	56.93	59.63	53.33	59.57	61.19	21.32	13.56	22.36	16.54	17.53

Cuadro 4.4.2 Continuación.

Trat	Variable-respuesta 3 Producción de materia seca (g)				
	UE 1	UE 2	UE 3	UE 4	UE 5
I	6.71	8.08	10.03	8.82	9.70
II	4.59	5.52	4.26	3.65	3.98
III	1.45	3.99	3.74	2.97	1.65
IV	4.06	4.11	5.43	6.83	3.98
V	19.09	21.26	20.91	20.17	20.37
VI	23.03	30.64	18.98	22.63	25.78
VII	30.85	31.32	29.91	27.96	24.07
VIII	33.91	39.32	33.18	38.55	38.50
IX	32.61	31.07	32.66	34.47	32.75
X	37.13	34.16	41.55	30.80	28.54
XI	28.39	26.27	25.52	26.95	31.40
XII	29.98	34.09	28.04	27.28	30.16
XIII	40.59	39.30	30.63	40.55	37.04
XIV	28.86	31.76	28.39	33.98	29.16
XV	30.02	32.88	24.77	30.32	24.63
XVI	35.11	26.04	29.70	24.32	27.97

Normalmente se desea probar si los vectores de medias μ_j son iguales, lo cual se puede plantear como:

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k \text{ versus } H_1: \mu_i \neq \mu_j \text{ para alguna } i \neq j, \quad (4.4-13)$$

es decir, se plantea si los vectores de medias de las diferentes poblaciones son iguales y esto se puede escribir equivalentemente como:

$$H_0: \tau_1 = \tau_2 = \dots = \tau_k \text{ versus } H_1: \tau_i \neq \tau_j \text{ para alguna } i \neq j, \quad (4.4-14)$$

de forma análoga al caso univariado.

4.4.2.2.1 Sumas de cuadrados y productos cruzados (SCPC)

Con base al modelo (4.4-12) del diseño completamente al azar multivariado, se presentan los cálculos de las matrices de las siguientes SCPC (Mardia *et al.*, 1989; pp. 334-335).

La matriz de las SCPC “entre-muestras”, “entre-grupos” o “entre tratamientos” se calcula con:

$$\mathbf{K} = \sum_{j=1}^k n_j (\bar{\mathbf{y}}_j - \bar{\mathbf{y}})(\bar{\mathbf{y}}_j - \bar{\mathbf{y}})', \quad (4.4-15)$$

donde, $\bar{\mathbf{y}}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbf{y}_{ij}$, $\bar{\mathbf{y}} = \frac{1}{n} \sum_{j=1}^k \sum_{i=1}^{n_j} \mathbf{y}_{ij}$; recuerde que $n = \sum_{j=1}^k n_j$. En palabras, $\bar{\mathbf{y}}_j$ es el vector que contiene las medias de las p variables-respuesta en cada “muestra” de n_j UE donde se aplicó el j -ésimo tratamiento, $j = 1, \dots, k$ y $\bar{\mathbf{y}}$ es el vector general de medias.

La matriz de las SCPC dentro de muestras, de las UE tratadas análogamente o dentro de tratamientos, que es la contraparte de la *SCEE* en el caso univariado es:

$$\mathbf{W} = \sum_{j=1}^k \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_j)' \quad (4.4-16)$$

Finalmente, la matriz de las SCPC total (corregida por la media), la contraparte de la SCT (corregida por la media) univariada, se obtiene con:

$$\mathbf{T} = \sum_{j=1}^k \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - \bar{\mathbf{y}})(\mathbf{y}_{ij} - \bar{\mathbf{y}})' \quad (4.4-17)$$

La matriz de las SCPC(total), $\mathbf{T}$, se descompone con la identidad del análisis de varianza multivariado en $\mathbf{K}$ y $\mathbf{W}$:

$$\mathbf{T} = \mathbf{K} + \mathbf{W}.$$

Ahora suponer que en el modelo lineal:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{U}, \quad (4.4-18)$$

$$\mathbf{Y} = \begin{array}{c} \begin{array}{cccc} \text{var 1} & \text{var 2} & \dots & \text{var } p \end{array} \\ \left[\begin{array}{cccc} Y_{11} & Y_{12} & \dots & Y_{1p} \\ \vdots & \vdots & & \vdots \\ Y_{n_1,1} & Y_{n_1,2} & \dots & Y_{n_1,p} \\ Y_{n_1+1,1} & Y_{n_1+1,2} & \dots & Y_{n_1+1,p} \\ \vdots & \vdots & & \vdots \\ Y_{n_1+n_2,1} & Y_{n_1+n_2,2} & \dots & Y_{n_1+n_2,p} \\ \vdots & \vdots & & \vdots \\ Y_{n-n_k,1} & Y_{n-n_k,2} & \dots & Y_{n-n_k,p} \\ \vdots & \vdots & & \vdots \\ Y_{n1} & Y_{n2} & \dots & Y_{np} \end{array} \right] \end{array} \left\{ \begin{array}{l} n_1 \text{ UE del trat 1} \\ n_2 \text{ UE del trat 2} \\ \vdots \\ n_k \text{ UE del trat } k \end{array} \right.$$

es la matriz de las variables-respuesta; $\mathbf{X}_{n \times (k+1)}$ es la misma matriz diseño de 0's y 1's presentada en (4.4-3); la matriz de parámetros es:

$$\mathbf{B} = \begin{bmatrix} \mu' \\ \tau'_1 \\ \vdots \\ \tau'_k \end{bmatrix}_{(k+1) \times p} = \begin{bmatrix} \mu_1 & \mu_2 & \dots & \mu_p \\ \tau_{11} & \tau_{12} & \dots & \tau_{1p} \\ \vdots & \vdots & & \vdots \\ \tau_{k1} & \tau_{k2} & \dots & \tau_{kp} \end{bmatrix}; \text{ y } \mathbf{U} = [\epsilon_1, \epsilon_2, \dots, \epsilon_n]', \text{ con } \epsilon_i = \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{ip} \end{bmatrix}.$$

Así que (4.4-18) es el modelo lineal general del diseño completamente al azar. Se puede observar que la $\mathbf{y}'_{ij}$ del modelo (4.4-12) corresponde a la i -ésima hilera de $\mathbf{Y}$ en (4.4-18);

así también, el vector y del modelo (4.4-2) puede considerarse como una de las columnas de Y .

Con base en lo anterior y por analogía directa del caso univariado, pueden reescribirse las SCPC presentadas en (4.4-15), (4.4-16) y (4.4-17) como:

$$K = Y'(P_{X_2} - P_1)Y, \quad (4.4-19)$$

$$W = Y'(I_n - P_{X_2})Y, \quad (4.4-20)$$

$$T = Y'(I_n - P_1)Y, \quad (4.4-21)$$

respectivamente, donde P_{X_2} y P_1 son los que se definieron en la presentación univariada de este diseño, después de la estructura de la matriz diseño X en la expresión (4.4-3). Es conveniente reiterar que, las sumas de cuadrados en el caso univariado son un caso particular de las SCPC que se obtienen en el contexto multivariado.

4.4.2.2 Pruebas de hipótesis sobre los tratamientos (*ibíd.*, pp. 334-335)

Bajo H_0 definida en (4.4-14) se tiene que:

$$W \sim W_p(\Sigma, n-k) \text{ y } K \sim W_p(\Sigma, k-1), \quad (4.4-22)$$

independientemente. Además, si $n \geq p+k$,

$$\Lambda_{cal} = \frac{|W|}{|W+K|} \stackrel{H_0}{\sim} \Lambda(p, n-k, k-1), \quad (4.4-23)$$

donde Λ_{cal} es el estadístico lambda de Wilks.

La regla de decisión para una prueba de tamaño α es: rechazar H_0 si $\Lambda_{cal} \leq \Lambda(p, n-k, k-1; \alpha)$ donde $\Lambda(p, n-k, k-1; \alpha)$ es el punto $100\alpha\%$ inferior de una distribución lambda de Wilks con parámetros p , $n-k$ y $k-1$; no rechazar de otro modo. Esta prueba puede hacerse completando el cuadro de MANOVA del DCA como sigue.

Cuadro 4.4.3 Análisis de varianza multivariado del diseño completamente al azar

Fuente de variación	g.l. ¹	Matriz de SCPC	Criterio de Wilks
Entre tratamientos	$k - 1$	$\mathbf{K} = \mathbf{Y}'(\mathbf{P}_{\mathbf{X}_2} - \mathbf{P}_1)\mathbf{Y}$	$\frac{ \mathbf{W} }{ \mathbf{W} + \mathbf{K} } \sim \Lambda(p, n - k, k - 1)$
Dentro tratamientos (EE)	$n - k$	$\mathbf{W} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_{\mathbf{X}_2})\mathbf{Y}$	
Total	$n - 1$	$\mathbf{T} = \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_1)\mathbf{Y}$	

Para el cálculo de Λ se puede usar:

$$\Lambda = \prod_{j=1}^p (1 + \lambda_j)^{-1}, \quad (4.4-24)$$

donde $\lambda_1, \dots, \lambda_p$ son los valores característicos de $\mathbf{W}^{-1}\mathbf{K}$. En esta prueba, es útil la transformación de Λ para expresarla en términos de la distribución de F cuando $k = 3$ o $p = 2$ dada en (4.1-42).

EJEMPLO 4.4.4 Con los datos que se presentan en el Cuadro 4.4.2, del experimento multivariado descrito en el Ejemplo 4.4.3, se escriben a continuación las matrices de las SCPC de los tratamientos y del error (dentro tratamientos), para probar después la hipótesis planteada en (4.4-14) usando el criterio de Wilks.

$$\mathbf{K} = \begin{bmatrix} 14423.42 & 3426.63 & 10384.23 \\ 3426.63 & 976.07 & 2641.46 \\ 10384.23 & 2641.46 & 10675.52 \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} 1593.51 & 32.11 & -204.40 \\ 32.11 & 313.08 & 97.86 \\ -204.40 & 97.86 & 542.98 \end{bmatrix}$$

y

¹ Obsérvese que los g.l. asociados a las SCPC que se presentan en el Cuadro 4.4.3 son iguales a los que corresponden a su contraparte univariada, puesto que la matriz asociada a la forma cuadrática en ambas formulaciones, la univariada y la multivariada, son exactamente las mismas; de ahí que tengan los mismos rangos, la manera formal de definir a los g.l.

$$\mathbf{T} = \begin{bmatrix} 16016.93 & 3458.74 & 10179.83 \\ 3458.74 & 1289.14 & 2739.32 \\ 10179.83 & 2739.32 & 11218.51 \end{bmatrix}.$$

En el cálculo del estadístico Λ_{cal} para probar la hipótesis de que no hay diferencias entre los tratamientos (en este Ejemplo, son las 16 combinaciones de N y K), se requieren $|\mathbf{W}| = 240705821$ y $|\mathbf{T}| = 3.6553 \times 10^{10}$; así, $\Lambda_{cal} = |\mathbf{W}|/|\mathbf{W} + \mathbf{K}| = 0.0065851$. Aquí no se puede obtener la aproximación de F dada en (4.1-42), ya que $k-1 > 2$; sin embargo, para valores grandes de m , se puede usar la aproximación de Bartlett escrita en (4.1-44) que queda como:

$$\begin{aligned} EPM &= -\left\{64 - \frac{1}{2}(3 - 15 + 1)\right\} \ln \Lambda(3, 64, 15) \\ &= -\{69.5\} \ln (0.0065851) = 349.1, \end{aligned}$$

el cual se debe comparar con $\chi^2_{45,0.05} = 61.63$. Como $EPM = 349.1 > 61.63$, la conclusión para una prueba aproximada de tamaño $\alpha = 0.05$ es: rechazar la hipótesis nula, lo que es lo mismo a concluir que los tratamientos son diferentes entre sí. Usando un nivel de significancia de 0.01, $\chi^2_{45,0.01} = 69.92$, por lo que también se concluye que existen diferencias significativas entre los tratamientos. Mardia *et al.* (1989, p. 84) mencionan la limitante del EPM dado en (4.1-44). ♦

Por otra parte, Rao (1951, 1973, p.556) citado por Mardia *et al.* (1989) escribe que si $\Lambda \sim \Lambda(p, t - q, q)$, entonces una aproximación alternativa es:

$$R = \left(\frac{ms - 2\lambda}{pq} \right) \left(\frac{1 - \Lambda^{1/s}}{\Lambda^{1/s}} \right) \quad (4.4-25)$$

donde $m = t - \frac{1}{2}(p + q + 1)$, $\lambda = \frac{1}{4}(pq - 2)$, y $s^2 = (p^2 q^2 - 4)/(p^2 + q^2 - 5)$; R tiene una distribución F asintótica con pq y $(ms - 2\lambda)$ g.l. donde $(ms - 2\lambda)$ no tiene necesariamente que ser un número entero.

EJEMPLO 4.4.5 Del Ejemplo 4.4.3, el estadístico $\Lambda \sim \Lambda(3, 64, 15)$, así que: $p = 3$, $t = 79$, $q = 15$. De esta forma $m = 69.5$, $\lambda = 10.75$ y $s^2 = 8.8427$, por lo que de (4.4-25):

$$R = 18.166.$$

Como $R = 18.166$ es mayor que $F_{45, 185.17; 0.01} = 1.67$, entonces se concluye con un nivel de significancia $\alpha = 0.01$, de que al menos algún tratamiento es significativamente diferente de los otros. Para valores de p y $k - 1 > 2$, se recomienda usar la aproximación de Rao; además es esta aproximación en la que se basa la opción MANOVA de SAS. ♦

Los cálculos numéricos de las matrices **W** y **K** y Λ en el Ejemplo 4.4.4, y la aproximación de Rao obtenida en el Ejemplo 4.4.5, se pueden hacer con un sencillo programa en SAS como el que sigue:

```
DATA AILE; INFILE 'A:\MANOVA\MELINA.DAT';
INPUT ALTUP DIAMCR PESOMS TRAT $;
PROC ANOVA; CLASS TRAT;
MODEL ALTUP DIAMCR PESOMS=TRAT;
MANOVA H=TRAT/PRINTE PRINTH;
RUN;
```

Para ejecutar el programa anterior, se requiere que los datos del Cuadro 4.4.2 estén contenidos en el archivo externo llamado MELINA.DAT; las tres primeras columnas del archivo de datos contienen los valores que tomaron las variables-respuesta en el mismo orden en que se presentan en el Cuadro 4.4.2, y la cuarta columna corresponde al tratamiento (TRAT).

4.4.3 El Enfoque Unión Intersección

Considerar el modelo del diseño de un criterio de clasificación dado en (4.4-12) y la hipótesis nula planteada en (4.4-14). Esta hipótesis es verdadera si y sólo si todas y cada una de las hipótesis univariadas

$$H_{0\mathbf{a},\mathbf{c}}: \sum_{i=1}^k c_i \mathbf{a}' \boldsymbol{\tau}_i = 0$$

son verdaderas para todos los vectores $\mathbf{a}$ de dimensión p y todos los vectores de contrastes $\mathbf{c}$ de dimensión k . Un criterio de prueba adecuado es λ_1 , el valor característico más grande de $\mathbf{W}^{-1}\mathbf{K}$. La hipótesis nula H_0 se rechaza para valores grandes de λ_1 .

Con el enfoque unión intersección pueden construirse regiones de confianza simultáneas para hacer contrastes entre los vectores $\boldsymbol{\tau}_i$. Los intervalos de confianza simultáneos al $100(1-\alpha)\%$ para $\mathbf{a}'(\boldsymbol{\tau}_j - \boldsymbol{\tau}_k)$ están dados por:

$$\mathbf{a}'(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_k) \pm \left\{ \frac{\theta_\alpha}{1-\theta_\alpha} \mathbf{a}'\mathbf{W}\mathbf{a} \left(\frac{1}{n_j} + \frac{1}{n_k} \right) \right\}^{1/2}, \quad (4.4-26)$$

donde $\theta_\alpha \stackrel{\text{Notación}}{=} \theta(p, v_1, v_2; \alpha)$ es el punto $100\alpha\%$ superior de la distribución del valor característico más grande, donde v_1 son los g.l. del error y v_2 son los g.l. de la hipótesis; recuerde que $\theta_{cal} = \frac{\lambda_1}{1+\lambda_1} \sim \theta(p, v_1, v_2)$. Los valores de θ_α para α igual a 0.1, 0.05 y 0.01 con $p=2$, pueden encontrarse en la Tabla C.4 de Mardia *et al.* (1989; pp. 494-496). En general, a la variable aleatoria correspondiente a la distribución del valor característico más grande de $|\mathbf{K} - \theta(\mathbf{K} + \mathbf{W})| = 0$ se denota por $\theta(p, v_1, v_2)$ donde $\mathbf{K} \sim W_p(\Sigma, v_2)$ es independiente de $\mathbf{W} \sim W_p(\Sigma, v_1)$; v_1 son los grados de libertad del error y v_2 son los grados de libertad de la hipótesis.

EJEMPLO 4.4.6 (*ibid.*, p. 350) Usando los datos de Iris del Cuadro 3.2.1 se tiene que los vectores de medias para las tres especies (MED1, MED2, MED3) y el vector de medias general (MEDG) son:

	MED1	MED2	MED3	MEDG
Largo de sépalo	5.006	5.936	6.588	5.8433333
Ancho de sépalo	3.428	2.77	2.974	3.0573333
Largo de pétalo	1.462	4.26	5.552	3.758
Ancho de pétalo	0.246	1.326	2.026	1.1993333

Las matrices de las SCPC entre especies (K) y dentro especies (W) son:

K				W			
63.212133	-19.95267	165.2484	71.279333	38.9562	13.63	24.6246	5.645
-19.95267	11.344933	-57.2396	-22.93267	13.63	16.962	8.1208	4.8084
165.2484	-57.2396	437.1028	186.774	24.6246	8.1208	27.2226	6.2718
71.279333	-22.93267	186.774	80.413333	5.645	4.8084	6.2718	6.1566

respectivamente. Así también, la matriz $W^{-1}K$ es representada por F y sus valores característicos diferentes de cero están contenidos en la columna C. Estos valores característicos, se obtuvieron usando la propiedad de que los valores característicos diferentes de cero de $W^{-1}K$ son los mismos que los de $K^{1/2}W^{-1}K^{1/2}$. La matriz D contiene los vectores característicos asociados a los mencionados eigenvalores.

F				C	D			
-3.058369	1.0813826	-8.111923	-3.45865	32.191929	0.3143186	0.3430072	0.812331	0.351665
-5.561639	2.1782187	-14.96461	-6.30774	0.285391	-0.144393	0.8837743	-0.133497	-0.424585
8.0774388	-2.942719	21.511591	9.1420647	4.445E-16	0.8671595	-0.101771	-0.08472	-0.480104
10.497082	-3.419854	27.548525	11.84588	-1.09E-16	0.358314	0.3015487	-0.561355	0.6823196

Finalmente el término $\pm$ de (4.4-26) usando que $\theta(4,147,2) = \theta(2,145,4) = 0.1098$ tomado del Apéndice C.4 de Mardia *et al.* (1989) y escogiendo $\mathbf{a}' = (0,1,0,0)$ es:

$$MM = 0.289285$$

Todos estos cálculos se obtuvieron con el Programa 4.4.5.1.

Para probar la hipótesis de igualdad de vectores de medias poblacionales con las tres especies usando el enfoque unión-intersección se tiene el valor característico más grande de $W^{-1}K$ es $\lambda_1 = 32.192$ y $\theta_{cal} = \lambda_1 / (1 + \lambda_1) = 0.96987$. Como $\theta_{cal} = 0.97 > \theta_{0.01} = 0.1098$ se rechaza la hipótesis de que los vectores de medias de las tres especies son iguales con un nivel de significancia $\alpha = 0.01$.

Los intervalos de confianza (I.C.) al 99% para las diferencias del ancho de sépalo son los siguientes:

$$I. \text{ setosa e } I. \text{ versicolour} \quad \tau_{32} - \tau_{22} \in 3.43 - 2.77 \pm 0.29 = 0.66 \pm 0.29;$$

$$I. \text{ setosa e } I. \text{ virginica} \quad \tau_{32} - \tau_{12} \in 3.43 - 2.97 \pm 0.29 = 0.46 \pm 0.29;$$

$$I. \text{ virginica e } I. \text{ versicolour} \quad \tau_{12} - \tau_{22} \in 2.97 - 2.77 = 0.20 \pm 0.29.$$

Puede observarse que el único intervalo que contiene al cero es el que se construyó para *I. virginica* e *I. versicolour*, por lo que no se rechaza la hipótesis de igualdad de los anchos de sépalos entre estas dos especies con un nivel de $\alpha = 0.01$.

4.4.4 Diseño de Bloques Completos al Azar

4.4.4.1 Formulación univariada estándar

Suponer que Y_{ij} , la respuesta de la *UE* perteneciente al i -ésimo bloque que recibió el j -ésimo tratamiento, $i = 1, 2, \dots, b$, $j = 1, 2, \dots, k$, es generada por el modelo:

$$Y_{ij} = \mu + \alpha_i + \tau_j + \varepsilon_{ij} \quad (4.4-27)$$

donde, μ es el efecto de la media, α_i es el efecto del i -ésimo bloque, τ_j el efecto del j -ésimo tratamiento y ε_{ij} el error experimental (resultado de la interacción del bloque i y el tratamiento j) bajo el supuesto que se distribuye independientemente como una $N(0, \sigma_e^2)$. Existen $n = bk$ *UE* en total, ya que cada bloque contiene cada uno los k tratamientos y en cada *UE* hay una muestra de la única variable-respuesta que se estudia. Esto puede verse esquemáticamente en el Cuadro 4.4.4.

La forma matricial del modelo (4.4-27), puede verse como:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (4.4-28)$$

donde $\mathbf{y}$ es el vector-respuesta de dimensión $(n \times 1)$, $\mathbf{X}_{(n \times (1+b+k))}$ es la matriz diseño, $\boldsymbol{\beta}_{(1+b+k) \times 1}$ es el vector de parámetros y $\boldsymbol{\varepsilon}_{n \times 1} \sim N(\mathbf{0}, \Sigma)$ es el vector de errores aleatorios.

Cuadro 4.4.4 Observaciones de un
bloques al azar estándar

Fuente de variación	tratamientos		
bloques	1	...	k
1	y_{11}	...	y_{1k}
$\vdots$	$\vdots$		$\vdots$
b	y_{b1}	...	y_{bk}

Veamos de forma ampliada las estructuras de las matrices del modelo (4.4-28). El vector-respuesta $y = [Y_1 \ Y_2 \ \dots \ Y_n]'$; el vector de parámetros $\beta = [\mu \ \alpha' \ \tau']'$ con α de dimensión $(b \times 1)$ y τ de dimensión $(k \times 1)$; la estructura de la matriz diseño es:

$$X = \begin{bmatrix} 1 & 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \dots & 0 & 0 & 0 & \dots & 1 \\ 1 & 0 & 1 & \dots & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \dots & 0 & 0 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 \end{bmatrix}_{n \times (1+b+k)} \quad (4.4-29)$$

Para facilitar la presentación de lo que sigue, particionemos por columnas a la matriz diseño (4.4-29) en tres partes:

$$X = [\mu, \ X_1, \ X_2], \quad (4.4-30)$$

donde $\mu = 1$ es la primera columna del efecto promedio de la variable-respuesta, las siguientes b columnas de los bloques conforman a $X_1(n \times b)$ y a $X_2(n \times k)$ la forman las últimas k columnas correspondientes a igual número de tratamientos.

Para presentar las sumas de cuadrados (SC) se deben tener los proyectores del vector 1 , y de las submatrices X_1 y X_2 de la partición (4.4-30); $P_1 = n^{-1}11'$, $P_{X_1} = X_1(X_1'X_1)^{-1}X_1'$ y $P_{X_2} = X_2(X_2'X_2)^{-1}X_2'$.

En el Cuadro 4.4.5 se presentan las SC y las esperanzas de los cuadrados medios correspondientes a este diseño. El estadístico para probar la igualdad de tratamientos es:

$$F_{cal} = \frac{(n-b-t+1)SC_{trat}}{(k-1)SCEE} \sim F_{k-1, n-b-t+1} \quad (4.4-31)$$

La regla de decisión para una prueba de tamaño α es: rechazar que existe igualdad de los tratamientos si $F_{cal} \geq F_{k-1, n-b-t+1; \alpha}$ donde $F_{k-1, n-b-t+1; \alpha}$ es el punto $100\alpha\%$ superior de una distribución F de Snedecor; de otro modo, no rechazar la igualdad de los tratamientos.

Cuadro 4.4.5 Sumas de cuadrados y esperanzas de los cuadrados medios del diseño en bloques completos al azar estándar

f.v.	g.l.	SC	E(CM)
<i>bloq</i>	$b-1$	$y'[P_{X_1} - P_1]y$	$\sigma_e^2 + \frac{k}{b-1} \sum_{i=1}^b (\alpha_i - \bar{\alpha})^2$
<i>trat</i>	$k-1$	$y'[P_{X_2} - P_1]y$	$\sigma_e^2 + \frac{b}{k-1} \sum_{j=1}^k (\tau_j - \bar{\tau})^2$
<i>EE</i>	$(b-1)(k-1)$	$y'[I - P_{X_1} - P_{X_2} + P_1]y$	σ_e^2
<i>Total</i>	$kb-1$	$y'[I - P_1]y$	

donde, $\bar{\alpha} = \frac{1}{b} \sum_{i=1}^b \alpha_i$ y $\bar{\tau} = \frac{1}{k} \sum_{j=1}^k \tau_j$.

4.4.4.2 Formulación univariada con submuestreo

Si en cada una de las $n = bk$ unidades experimentales del diseño de bloques completos al azar (o diseño por dos criterios de clasificación) se hacen $m > 1$ mediciones aleatorias de la única variable-respuesta en estudio, se tiene la presentación univariada con submuestreo de este diseño. Por facilidad, considérese que existe el mismo número de

Cuadro 4.4.6 Observaciones de un bloques al azar con submuestreo

Fuente de variación	tratamientos		
bloques	1	...	k
1	$y_{111}, \dots, y_{11m}$	...	$y_{1k1}, \dots, y_{1km}$
$\vdots$	$\vdots$		$\vdots$
b	$y_{b11}, \dots, y_{b1m}$	...	$y_{bk1}, \dots, y_{bkm}$

submuestras por unidad experimental (m observaciones en cada una), por lo que, como puede observarse en el arreglo de las observaciones en el Cuadro 4.4.6, se tendrán $nm = kbm$ observaciones en total.

4.4.4.2.1 El modelo lineal

Si Y_{ijh} es la h -ésima submuestra tomada en la intersección del i -ésimo bloque y el j -ésimo tratamiento, $h = 1, \dots, m$, $i = 1, \dots, b$, $j = 1, \dots, k$, el modelo lineal puede expresarse como:

$$Y_{ijh} = \mu + \alpha_i + \tau_j + \eta_{ij} + \varepsilon_{ijh} \quad (4.4-32)$$

donde μ es el efecto de la media, α_i el efecto del i -ésimo bloque, τ_j el efecto del j -ésimo tratamiento, η_{ij} es el efecto de la interacción del i -ésimo bloque y el j -ésimo tratamiento (error experimental) y ε_{ijh} la variable de error (de muestreo, que representa

la variabilidad de submuestras dentro UE 's). Los supuestos del modelo son (a) $\eta_{ij} \sim (0, \sigma_e^2)$, (b) $\varepsilon_{ikh} \sim (0, \sigma_m^2)$ y (c) $cov(\eta_{ij}, \varepsilon_{ijh}) = 0$.

El modelo lineal en la expresión (4.4-32) se puede representar de forma matricial como:

$$y = X\beta + \varepsilon \quad (4.4-33)$$

donde y de dimensión $nm \times 1$ contiene las m submuestras de todas las $n = kb$ unidades experimentales; X es la matriz diseño del bloques al azar y es de dimensión $nm \times (1 + b + k + n)$; $\beta_{(1+b+k+n) \times 1}$ es el vector de parámetros; y ε es un vector aleatorio de dimensión $nm \times 1$. De forma ampliada, puede escribirse:

$$X = \begin{matrix} & \mu & \alpha_1 & \cdots & \alpha_b & \tau_1 & \cdots & \tau_k & \eta_{11} & \cdots & \eta_{1k} & \cdots & \eta_{b1} & \cdots & \eta_{bk} \\ \begin{matrix} \text{X} = \end{matrix} & \begin{bmatrix} 1 & 1 & \cdots & 0 & 1 & \cdots & 0 & 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ 1 & 1 & \cdots & 0 & 0 & \cdots & 1 & 0 & \cdots & 1 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ 1 & 0 & \cdots & 1 & 1 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ 1 & 0 & \cdots & 1 & 0 & \cdots & 1 & 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \end{matrix} \quad (4.4-34)$$

con $1 = [1, 1, \dots, 1]'$ y $0 = [0, 0, \dots, 0]$ de dimensión $m \times 1$. La matriz diseño (4.4-34) se puede particionar por columnas como sigue:

$$X = [1, X_1, X_2, X_3] \quad (4.4-35)$$

donde $1_{mn \times 1}$ es la primera columna de X correspondiente al efecto de la media; $X_1(mn \times b)$ es la matriz de los efectos de bloques; $X_2(mn \times k)$ es la matriz de los tratamientos; $X_3(mn \times n)$ es la matriz de las interacciones bloque-tratamiento. Hay dos cosas importantes que se deben hacer notar en la partición (4.4-35) de la matriz X : (a) las estructuras de $1, X_1$ y X_2 de esta partición son idénticas a las presentadas en (4.4-30), excepto que ahora, el número de observaciones por UE son $m > 1$; (b) si $m = 1$, la submatriz X_3 se reduciría a la matriz idéntica I de tamaño $n \times n$, que de hecho, es la

matriz de las interacciones en el diseño estándar (con sólo una observación por UE). Finalmente se escribe el vector de parámetros de forma ampliada como:

$$\beta = [\mu, \alpha_1, \alpha_2, \dots, \alpha_b, \tau_1, \tau_2, \dots, \tau_k, \eta_{11}, \dots, \eta_{1k}, \eta_{21}, \dots, \eta_{2k}, \dots, \eta_{b1}, \dots, \eta_{bk}]' \quad (4.4-36)$$

4.4.4.2.2 Sumas de cuadrados y esperanzas de los cuadrados medios

Con estos antecedentes, se pueden escribir las $SC(\cdot)$ y las $E(CM(\cdot))$ de este diseño resumidas en el Cuadro de análisis de varianza siguiente.

Cuadro 4.4.7 Sumas de cuadrados y esperanzas de los cuadrados medios del di-seño en bloques completos al azar con submuestreo

f.v.	g.l.	SC	E(CM)
<i>bloq</i>	$b-1$	$y' [P_{X_1} - P_1] y$	$\frac{\sigma^2}{m} + \frac{mk}{b-1} \sum_{i=1}^b (\alpha_i - \bar{\alpha})^2$
<i>trat</i>	$k-1$	$y' [P_{X_2} - P_1] y$	$\frac{\sigma^2}{m} + \frac{mb}{k-1} \sum_{j=1}^k (\tau_j - \bar{\tau})^2$
<i>EE</i>	$(b-1)(k-1)$	$y' [P_{X_3} - P_{X_1} - P_{X_2} + P_1] y$	$\frac{\sigma^2}{m}$
<i>EM</i>	$n(m-1)$	$y' [I_{mn} - P_{X_3}] y$	σ_m^2
<i>Total</i>	$mn-1$	$y' [I_{mn} - P_1] y$	

donde $\frac{\sigma^2}{m} = \sigma_m^2 + m \sigma_e^2$, $\bar{\alpha} = \frac{1}{b} \sum_{i=1}^b \alpha_i$, $\bar{\tau} = \frac{1}{k} \sum_{j=1}^k \tau_j$. En el Cuadro 4.4.7, $P_1 = \frac{1}{nm} \mathbf{1}\mathbf{1}'$ es de dimensión $nm \times nm$; P_{X_1} , P_{X_2} y P_{X_3} son los proyectores de las matrices correspondientes dadas en la partición (4.4-35) definidos de manera genérica para cualquier matriz $A_{p \times q}$ como $P_A = A(A'A)^{-1}A'$; si $m=1$ y $\sigma_m^2 = 0$, se sustituyen en el Cuadro 4.4.7 se obtienen los valores del Cuadro 4.4.5, que es el caso cuando no se hace submuestreo en el diseño de bloques al azar. Recuerde también que los grados de libertad asociados a las SC son los rangos de las correspondientes matrices entre corchetes.

Por último, si en el arreglo de las observaciones del Cuadro 4.4-6 sólo se considera si la observación corresponde o no a una *UE*, sin tomar en cuenta el bloque y el tratamiento al que pertenece, se puede escribir la matriz:

$$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{mn \times n}$$

donde las columnas representan las n unidades y los vectores $\mathbf{1}$ y $\mathbf{0}$ de tamaño m representan la presencia o no, respectivamente, de igual número de submuestras en cada *UE*; a esta matriz se le puede nombrar como la *matriz de las unidades experimentales* y es exactamente igual a la matriz $\mathbf{X}_3$ de la partición (4.4-35). Con esta notación se tiene que la suma de cuadrados de las unidades experimentales en este diseño es:

$$SC(UE) = \mathbf{y}'[\mathbf{P}_{\mathbf{X}_3} - \mathbf{P}_1]\mathbf{y} = SC_{bloq} + SC_{trat} + SC_{EE} \quad (4.4-37)$$

y se puede ver que:

$$SCTotal = SC(UE) + SC_{EM}, \quad (4.4-38)$$

identidad que es válida para cualquier diseño experimental, como se había mencionado en el tercer párrafo de la Sección 4.4.1.

4.4.4.2.3 Pruebas de hipótesis

Hay dos pruebas importantes a considerar: la prueba sobre el error experimental y la prueba sobre la igualdad de efectos de los tratamientos.

Prueba sobre el error experimental. Para probar la significancia del error experimental, se usa el estadístico de prueba siguiente:

$$F_{cal} = \frac{n(m-1)SC_{EE}}{(b-1)(k-1)SC_{EM}} \sim F_{n-b-k+1, nm-n}. \quad (4-4.39)$$

Si $F_{n-b-k+1, nm-n; \alpha}$ es el punto $100\alpha\%$ superior de una distribución F , la regla de decisión con un nivel de significancia α es: rechazar la hipótesis nula si $F_{cal} \geq F_{n-b-k+1, nm-n; \alpha}$ y concluir que el error experimental (formado por la interacciones bloque-tratamiento) sí es significativo; de otro modo, concluir que las interacciones bloque-tratamiento no son significativas. Recuerde que en total, hay $n = kb$ unidades experimentales.

Prueba sobre los tratamientos. Si se concluye que el error experimental es significativo, la hipótesis nula de que todos los tratamientos tienen el mismo efecto, se puede probar con el siguiente estadístico de prueba:

$$F_{cal} = \frac{CM_{trat}}{CM_{EE}} \sim F_{k-1, n-b-k+1} \quad (4.4-40)$$

La regla de decisión para una prueba de tamaño α es: rechazar $H_0: \tau_i = \tau_j$, $\forall i, j = 1, \dots, k$, $i \neq j$, si $F_{cal} \geq F_{k-1, n-b-k+1; \alpha}$; no rechazar de otro modo. Los cuadrados medios usados para calcular el estadístico de prueba anterior se derivan del Cuadro 4.4.7.

4.4.4.3 Formulación multivariada estándar

En el Cuadro 4.4.4 se tiene el arreglo de las observaciones de un bloques al azar, en donde se asume que en cada intersección bloque-tratamiento (que forman las n unidades) se toma una observación de la variable-respuesta que se estudia. Ahora, si en cada unidad experimental se hace una sola observación de cada una de $p > 1$ diferentes variables-respuesta, los datos colectados se pueden arreglar como en el Cuadro 4.4.8, donde y_{ijl} es la observación correspondiente a la l -ésima variable-respuesta tomada en la intersección del i -ésimo bloque y el j -ésimo tratamiento, $i = 1, \dots, b$, $j = 1, \dots, k$, $l = 1, \dots, p$.

Cuadro 4.4.8 Observaciones de un bloques al azar; el caso multivariado estándar

Fuente de variación	tratamientos		
bloques	1	...	k
1	$y_{111}, y_{112}, \dots, y_{11p}$	...	$y_{1k1}, y_{1k2}, \dots, y_{1kp}$
$\vdots$	$\vdots$		$\vdots$
b	$y_{b11}, y_{b12}, \dots, y_{b1p}$	...	$y_{bk1}, y_{bk2}, \dots, y_{bkp}$

4.4.4.3.1 El modelo lineal

De forma escalar, el modelo lineal que genera cada observación del Cuadro 4.4.8 es:

$$y_{ijl} = \mu_l + \alpha_{il} + \tau_{jl} + \eta_{ijl} \quad (4.4-41)$$

donde μ_l es el efecto promedio de la l -ésima respuesta, α_{il} es el efecto del i -ésimo bloque sobre la l -ésima respuesta, τ_{jl} es el efecto de la j -ésimo tratamiento sobre la l -ésima respuesta y η_{ijl} el efecto de la interacción del i -ésimo bloque con el j -ésimo tratamiento sobre la l -ésima respuesta.

Si en el modelo lineal general (4.4-18) —presentado primero en (4.3-17)— se considera la matriz diseño X como la presentada en (4.4-29) y la matriz de parámetros:

$$\mathbf{B} = \begin{bmatrix} \mu_1 & \cdots & \mu_p \\ \alpha_{11} & \cdots & \alpha_{1p} \\ \vdots & & \vdots \\ \alpha_{b1} & \cdots & \alpha_{bp} \\ \tau_{11} & \cdots & \tau_{1p} \\ \vdots & & \vdots \\ \tau_{k1} & \cdots & \tau_{kp} \end{bmatrix}, \quad (4.4-42)$$

entonces tenemos el modelo lineal general multivariado del diseño de bloques al azar en forma matricial.

4.4.4.3.2 Sumas de cuadrados y productos cruzados

Las matrices de las sumas de cuadrados y productos cruzados (SCPC) se presentan en el Cuadro 4.4.9, donde X_1 y X_2 son los de la partición (4.4-30).

Cuadro 4.4.9 Matrices de las SCPC del diseño en bloques completos al azar multivariado estandar

f.v.	g.l.	SCPC	Criterio de prueba
<i>bloq</i>	$b - 1$	$R = Y' [P_{X_1} - P_1] Y$	
<i>trat</i>	$k - 1$	$C = Y' [P_{X_2} - P_1] Y$	$\frac{ I }{ I + C } \sim \Lambda(p, v_1, v_2)$
<i>EE</i>	$(b - 1)(k - 1)$	$I = Y' [I - P_{X_1} - P_{X_2} - P_1] Y$	
<i>Total</i>	$tb - 1$	$T = Y' [I - P_1] Y$	

con $v_1 = kb - b - k + 1$ y $v_2 = k - 1$. En el Cuadro anterior se tiene el estadístico para probar la hipótesis de que todos los tratamientos son iguales. La regla de decisión con un nivel de significancia α es: rechazar H_0 si el estadístico de prueba es menor que $\Lambda(p, v_1, v_2; \alpha)$, concluyéndose que, al menos, el efecto de un tratamiento es diferente a los otros; de otro modo, no rechazar la hipótesis nula.

4.4.4.4 Formulación multivariada y con submuestreo

Si en cada unidad experimental ubicada en la intersección bloque-tratamiento (hilera-columna, en concordancia con Mardia *et al.*, 1989 y Morrison, 1981) se toman $m > 1$

submuestras de cada una de las $p > 1$ variables-respuesta en estudio, los datos se pueden arreglar como en el Cuadro 4.4.10.

Cuadro 4.4.10 Observaciones de un bloques al azar; el caso multivariado con submuestreo

Fuente de variación	tratamientos (columnas)		
bloques (hileras)	1	...	k
1	y'_{111}	...	y'_{1k1}
	y'_{112}		y'_{1k2}
	$\vdots$		$\vdots$
	y'_{11m}		y'_{1km}
$\vdots$	$\vdots$		$\vdots$
b	y'_{b11}	...	y'_{bk1}
	y'_{b12}		y'_{bk2}
	$\vdots$		$\vdots$
	y'_{b1m}		y'_{bkm}

donde,

$$y'_{ijh} = [y_{ijh1}, y_{ijh2}, \dots, y_{ijhp}],$$

con $i = 1, \dots, b$, $j = 1, \dots, k$, $h = 1, \dots, m$, $l = 1, \dots, p$.

4.4.4.4.1 El modelo lineal

Sea y_{ijhl} la h -ésima submuestra de la l -ésima variable-respuesta obtenida en la unidad experimental i - j , bajo el j -ésimo tratamiento en el i -ésimo bloque; de forma genérica, podemos llamarle *hileras* a los bloques y *columnas* a los tratamientos. El modelo lineal que genera la observación y_{ijhl} es (Morrison, 1981; p.186):

$$y_{ijhl} = \mu_l + \alpha_{il} + \tau_{jl} + \eta_{ijl} + \varepsilon_{ijhl}, \quad (4.4-43)$$

donde, μ_l es el efecto general de la l -ésima variable respuesta, α_{il} es el efecto del i -ésimo bloque sobre la l -ésima variable respuesta, τ_{jl} es el efecto del j -ésimo tratamiento

sobre la l -ésima variable respuesta, η_{ijl} es el efecto de interacción del i -ésimo bloque con el j -ésimo tratamiento sobre la l -ésima variable respuesta y ε_{ijhl} es el término de error aleatorio con distribución multinormal.

Para escribir el modelo (4.4-43) usando vectores (Mardia *et al.*, 1989; p. 350), suponer que en este diseño de dos criterios de clasificación se tienen mbk observaciones independientes generadas por el modelo:

$$y_{ijh} = \mu + \alpha_i + \tau_j + \eta_{ij} + \varepsilon_{ijh}, \quad (4.4-44)$$

$i = 1, \dots, b, j = 1, \dots, k, h = 1, \dots, m$; donde α_i es el efecto del i -ésimo bloque (hilera); τ_j es el efecto del j -ésimo tratamiento (columna); η_{ij} es el efecto de interacción entre la i -ésima hilera y la j -ésima columna; y ε_{ijk} es el término de error que se distribuye independientemente como una $N_p(0, \Sigma)$ para toda i, j, k . Se requiere que el número de observaciones en cada unidad experimental ($i-j$) sean iguales, para que la matriz de la SCPC total pueda descomponerse apropiadamente (*ibid.*)

4.4.4.4.2 Planteamiento de las pruebas de hipótesis

Bajo el supuesto de que los errores tienen distribución multinormal, con media cero y la misma matriz de covarianzas para todas las combinaciones de hileras y columnas, se pueden plantear las siguientes tres preguntas (Morrison, 1981; p. 185):

- (a) ¿Son iguales los vectores de medias de hileras? (esto es equivalente a plantear la igualdad de las α_i),
- (b) ¿Son iguales los vectores de medias de columnas? (que equivale a plantear la igualdad de las τ_j)
- (c) ¿Existe alguna interacción hilera-columna que produzca un efecto desfasado, ya sea hacia arriba o hacia abajo, del promedio de las respuestas? (equivale a plantear la igualdad de las η_{ij})

En el lenguaje del análisis de varianza, los incisos (a) y (b) se conocen como efectos principales de hileras y de columnas, respectivamente, y el inciso (c) tiene que ver con alguna posible interacción hilera-columna.

4.4.4.4.3 Sumas de cuadrados y productos cruzados

Siguiendo a Mardia *et al.* (1989), sea T la matriz de las SCPC total, que se puede descomponer en R , C y E , las matrices de la SCPC de hileras, de columnas y de errores, respectivamente; es decir,

$$T = R + C + E \quad (4.4-45)$$

donde,

$$\begin{aligned} T &= \sum_{i=1}^b \sum_{j=1}^k \sum_{h=1}^m (y_{ijh} - \bar{y}_{...})(y_{ijh} - \bar{y}_{...})', \\ R &= km \sum_{i=1}^b (\bar{y}_{i..} - \bar{y}_{...})(\bar{y}_{i..} - \bar{y}_{...})', \\ C &= bm \sum_{j=1}^k (\bar{y}_{.j.} - \bar{y}_{...})(\bar{y}_{.j.} - \bar{y}_{...})', \\ E &= \sum \sum \sum (y_{ijh} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})(y_{ijh} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})', \end{aligned}$$

con

$$\bar{y}_{i..} = \frac{1}{km} \sum_{j=1}^k \sum_{h=1}^m y_{ijh}, \quad \bar{y}_{.j.} = \frac{1}{bm} \sum_{i=1}^b \sum_{h=1}^m y_{ijh}, \quad \bar{y}_{...} = \frac{1}{bkm} \sum_{i=1}^b \sum_{j=1}^k \sum_{h=1}^m y_{ijh}.$$

Además, la matriz de la SCPC de los errores —que involucra tanto los experimentales como los de muestreo— se puede descomponer en la matriz de SCPC de las interacciones (error experimental) y la matriz de los residuales (error muestral), I y W , respectivamente, con la siguiente identidad:

$$E = I + W \quad (4.4-46)$$

donde,

$$I = m \sum_{i=1}^b \sum_{j=1}^k (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})'$$

y

$$\mathbf{W} = \sum_{i=1}^b \sum_{j=1}^k \sum_{h=1}^m (\mathbf{y}_{ijh} - \bar{\mathbf{y}}_{ij.})(\mathbf{y}_{ijh} - \bar{\mathbf{y}}_{ij.})',$$

con

$$\bar{\mathbf{y}}_{ij.} = \frac{1}{m} \sum_{h=1}^m \mathbf{y}_{ijh}$$

que es el vector de medias de las variables-respuesta por unidad experimental i - j .

4.4.4.4 Prueba para interacciones (*ibid.*, p. 352)

Haciendo uso de las identidades (4.4-45) y (4.4-46) se puede escribir que:

$$\mathbf{T} = \mathbf{R} + \mathbf{C} + \mathbf{I} + \mathbf{W}. \quad (4.4-47)$$

Bajo la hipótesis nula $\mathbf{B} = \mathbf{O}$, $\mathbf{T}$ se distribuye como una Wishart $W_p(\Sigma, kbm - 1)$; se va a usar también el resultado de que la matriz $\mathbf{W} \sim W_p(\Sigma, kb(m - 1))$. Bajo la hipótesis nula de que todos los vectores de interacciones $\boldsymbol{\eta}_{ij}$ son iguales,

$$\mathbf{I} \sim W_p(\Sigma, (b - 1)(k - 1)).$$

El estadístico de prueba para probar igualdad de las interacciones obtenido por el enfoque de razón de verosimilitud es:

$$\Lambda_{cal} = \frac{|\mathbf{W}|}{|\mathbf{W} + \mathbf{I}|} = \frac{|\mathbf{W}|}{|\mathbf{E}|}, \quad (4.4-48)$$

que se distribuye como una $\Lambda(p, v_1, v_2)$ donde $v_1 = kb(m - 1)$ y $v_2 = (b - 1)(k - 1)$. Si $\Lambda(p, v_1, v_2; \alpha)$ es el punto $100\alpha\%$ inferior de una distribución Λ de Wilks, la regla de decisión a un nivel de significancia α es: rechazar la hipótesis de que ninguna interacción es significativa si $\Lambda_{cal} \leq \Lambda(p, v_1, v_2; \alpha)$; de otro modo, no rechaza H_0 .

Si dentro cada unidad experimental i - j (intersección de la hilera i y la columna j) hay sólo una observación, $m = 1$, por cada variable respuesta, entonces $\mathbf{W}$ tendría cero grados de libertad, por lo que sería imposible probar la presencia de interacción; este es el

caso revisado en la Sección 4.4.4.3. Similarmente, en el caso univariado tiene que haber más de una observación por unidad, para que la prueba de interacción (error experimental) sea posible, como puede revisarse en la Sección 4.4.4.2.3.

4.4.4.4.5 Prueba para efectos principales

Si la hipótesis nula de la prueba de interacciones no fue rechazada, se concluye que no hay interacciones significativas y entonces se puede proceder a hacer la prueba de efectos principales. Según Mardia *et al.* (1989), el estadístico de razón de verosimilitud para probar la igualdad de las medias de columnas (que podrían ser las medias de tratamientos), indistintamente de las α_i y η_{ij} es:

$$\Lambda_{cal} = \frac{|W|}{|W + C|} \sim \Lambda(p, v_1, v_2) \quad (4.4-49)$$

donde $v_1 = kb(m-1)$, $v_2 = (k-1)$. La regla de decisión para una prueba de tamaño α es: rechazar H_0 si $\Lambda_{cal} \leq \Lambda(p, v_1, v_2; \alpha)$ y se concluye igualdad de columnas; no rechazar la hipótesis nula de otro modo. Si en las columnas, tenemos tratamientos, en caso de rechazar H_0 , se podrá concluir que los efectos de tratamientos son diferentes con un nivel de significancia α .

Para probar la igualdad de las b medias de hileras α_i indistintamente de los efectos de columnas τ_j , se reemplaza C por R y se intercambia b y k en las fórmulas ya escritas.

Si el efecto de interacción no es significativo, o sólo hay una observación por unidad experimental (Secciones 4.4.4.1 y 4.4.4.3), o por cualquier razón, el efecto de interacción no está considerado en el modelo lineal, entonces para hacer la prueba de la igualdad de medias de columnas, habrá que sustituir la matriz W por la matriz E quedando el estadístico de prueba como:

$$\Lambda_{cal} = \frac{|E|}{|E + C|} \sim \Lambda(p, v_1, v_2), \quad (4.4-50)$$

donde $v_1 = kbm - b - k$ y $v_2 = k - 1$. La regla de decisión para una prueba de tamaño α es rechazar H_0 si $\Lambda_{cal} \leq \Lambda(p, v_1, v_2; \alpha)$; no rechazar de otro modo.

EJEMPLO 4.4.7 Ali & Yap (1992) realizaron un estudio para comparar la efectividad de dos reguladores del crecimiento, conocidos como paclobutrazol y daminozide, en el control del crecimiento y las relaciones hídricas de plántulas de *Acacia mangium* con el objetivo de mejorar su establecimiento después de la plantación. Estos investigadores mencionan que cuando no se sincronizan los tiempos de la producción de planta y su plantación en campo, las plántulas sobrecrecen en tamaño y se tienen reportes de que la sobrevivencia y crecimiento de las mismas después del trasplante es pobre. En el experimento, se asperjaron plántulas de 14 semanas con ambos reguladores del crecimiento en concentraciones de 0, 25, 50, 75, 100, 150 y 200 $mg\ l^{-1}$. Se hicieron mediciones de incremento en altura y área foliar, transpiración y conducción estomática y otras, en intervalos semanales. Para fines de este ejemplo, suponer que las semillas que se usaron son de $b = 3$ procedencias y considérense las mediciones de $p = 3$ variables-respuesta: el incremento en altura (cm), el incremento en diámetro (cm) y el incremento en área foliar (cm^2). Incluyendo los datos del testigo o control, se cuenta con los datos de $k = 13$ tratamientos. Considérese también que en cada UE se midieron $m = 4$ plántulas (submuestras) para cada variable. Los datos se reportan en el Cuadro 4.4.11.

Cuadro 4.4.11 Datos de tres variables respuesta en un experimento de reguladores del crecimiento con plántulas de *Acacia mangium* de catorce semanas

f.v. bloq	Tratamientos (columnas)					
	trat 1	trat 2	trat 3	trat 4	trat 5	trat 6
1	30.94,0.35,388.60	24.88,0.24,380.06	23.69,0.19,465.48	11.99,0.24,459.11	18.54,0.23,322.05	17.64,0.27,320.70
	23.44,0.25,487.47	22.29,0.19,410.39	22.79,0.21,377.64	21.22,0.23,392.55	23.52,0.23,372.21	19.88,0.20,305.61
	35.55,0.24,442.93	27.46,0.24,400.75	25.78,0.25,312.73	27.24,0.23,364.68	23.57,0.30,461.69	21.07,0.21,355.84
	23.17,0.20, 414.57	27.95,0.24,423.30	30.75,0.28,438.46	25.22,0.22,405.15	20.04,0.25,327.93	11.05,0.19,159.35
2	26.66,0.29,470.01	22.34,0.28,348.50	18.69,0.20,451.09	15.41,0.20,408.62	22.52,0.24,318.98	21.91,0.18,289.05
	21.39,0.28,418.93	23.55,0.26,378.73	25.60,0.19,433.74	22.05,0.26,375.36	18.16,0.19,284.52	26.23,0.25,218.42
	26.75,0.25,395.26	23.54,0.19,362.60	18.66,0.28,376.78	28.10,0.24,250.06	15.17,0.22,296.56	20.07,0.19,332.23
	19.97,0.26,385.80	16.58,0.27,450.70	21.94,0.24,373.34	23.80,0.22,258.22	23.23,0.22,347.95	19.96,0.20,405.39
3	20.90,0.25,493.99	28.87,0.24,327.61	27.11,0.23,395.73	19.80,0.29,326.36	22.82,0.25,315.00	18.60,0.20,251.45
	22.26,0.29,429.05	30.89,0.28,401.01	22.23,0.29,421.23	22.71,0.29,403.64	23.50,0.21,359.47	17.56,0.22,243.46
	27.80,0.27,495.28	19.85,0.20,349.16	17.79,0.20,461.67	21.97,0.28,428.50	23.61,0.24,327.15	13.67,0.29,300.41
	17.85,0.23,366.93	23.20,0.24,433.83	19.08,0.26,434.52	13.50,0.30,420.52	20.12,0.27,314.39	20.77,0.23,194.53

Cuadro 4.4.11 Datos de tres variables respuestas en un experimento de reguladores del crecimiento con plántulas de *Acacia mangium* de catorce semanas (contin.)

f.v.	Tratamientos (columnas)					
bloq	trat 7	trat 8	trat 9	trat 10	trat 11	trat 12
1	26.87,0.26,265.09	18.28,0.27,422.33	18.49,0.22,307.18	24.49,0.19,405.89	17.83,0.16,411.96	14.93,0.14,325.88
	15.16,0.22,301.49	20.10,0.28,371.81	20.48,0.23,397.55	18.33,0.21,313.66	14.51,0.18,377.93	09.83,0.13,325.65
	18.58,0.23,279.93	19.86,0.23,432.98	26.03,0.22,522.92	21.78,0.21,479.74	22.22,0.23,376.09	10.54,0.14,333.55
	18.21,0.26,310.93	30.10,0.33,422.96	24.46,0.15,335.31	17.60,0.22,417.54	16.98,0.18,302.06	06.62,0.21,347.44
2	18.35,0.22,341.83	17.28,0.29,447.89	29.77,0.21,433.48	10.97,0.21,371.91	23.24,0.16,304.75	07.50,0.21,336.97
	17.63,0.21,295.65	19.75,0.26,415.82	18.68,0.22,405.78	22.50,0.17,362.95	15.98,0.18,320.34	12.90,0.20,338.60
	19.04,0.18,267.62	19.57,0.27,412.16	18.71,0.22,375.07	21.01,0.23,380.80	15.99,0.15,274.26	11.75,0.18,316.98
	08.27,0.21,286.33	27.36,0.31,365.02	19.89,0.21,390.75	19.84,0.18,452.62	18.37,0.17,269.16	14.59,0.21,357.91
3	15.19,0.22,286.03	20.70,0.28,474.45	17.49,0.17,479.70	23.98,0.25,393.86	25.42,0.19,356.49	11.66,0.14,373.61
	16.28,0.28,401.63	24.63,0.25,329.20	20.49,0.15,357.11	22.77,0.22,383.56	13.56,0.20,315.12	13.87,0.18,274.79
	10.01,0.23,243.79	15.78,0.32,421.16	20.68,0.18,418.21	22.96,0.24,383.11	11.06,0.18,320.65	09.29,0.17,245.96
	21.20,0.25,294.39	28.96,0.24,446.25	22.14,0.19,318.23	22.72,0.23,378.20	14.99,0.17,416.19	11.28,0.16,366.51

Cuadro 4.4.11 Continuación.

f.v.	
bloq	trat 13
1	26.87,0.26,265.09
	15.16,0.22,301.49
	18.58,0.23,279.93
	18.21,0.26,310.93
2	18.35,0.22,341.83
	17.63,0.21,295.65
	19.04,0.18,267.62
	08.27,0.21,286.33
3	15.19,0.22,286.03
	16.28,0.28,401.63
	10.01,0.23,243.79
	21.20,0.25,294.39

Para hacer las prueba de hipótesis de los bloque, los tratamientos y las interacciones es necesario tener las matrices de las sumas de cuadrados y productos cruzados siguientes.

Las SCPC para los bloques y los tratamientos son:

$$R = \begin{bmatrix} 51.172 & -0.047 & 486.891 \\ -0.047 & 0.002 & -0.440 \\ 486.891 & -0.440 & 4632.704 \end{bmatrix} \text{ y } C = \begin{bmatrix} 3088.720 & 18.594 & 25562.782 \\ 18.594 & 0.177 & 149.801 \\ 25562.782 & 149.801 & 377962.926 \end{bmatrix},$$

respectivamente.

Las SCPC para las interacciones bloque-tratamiento es:

$$\mathbf{I} = \begin{bmatrix} 361.872 & -0.196 & -1302.823 \\ -0.196 & 0.0324 & 12.109 \\ -1302.823 & 12.109 & 54048.762 \end{bmatrix},$$

y las SCPC de los residuales (variación entre submuestras) es:

$$\mathbf{W} = \begin{bmatrix} 2040.739 & 1.326 & 191.952 \\ 1.326 & 0.099 & -3.107 \\ 191.952 & -3.107 & 313255.594 \end{bmatrix}.$$

Para la prueba de hipótesis de las interacciones se calcula el estadístico de prueba siguiente:

$$A_{cal} = |\mathbf{W}|/|\mathbf{W} + \mathbf{I}| = 62808559 / 115243222 = 0.545$$

que se debe comparar con una $A(3,117,24; \alpha)$. Usando la aproximación de Bartlett el estadístico de prueba modificado

$$EPM = -\{117 - 0.5(3 - 24 + 1)\} \ln 0.545 = 77.09,$$

debe compararse con $\chi^2_{72;0.05} = 92.808$. Por lo tanto se concluye con una prueba aproximada de tamaño 0.05 que no hay diferencias entre las interacciones.

Como las interacciones no fueron significativamente diferentes, el estadístico de Wilks para probar si existen diferencias entre tratamientos es:

$$A_{cal} = |\mathbf{E}|/|\mathbf{E} + \mathbf{C}| = 115243222 / 803470056 = 0.1434319.$$

Usando la aproximación de Bartlett se tiene que:

$$EPM = -\{141 - 0.5(3 - 12 + 1)\} \ln 0.1434319 = 281.57$$

debe compararse con $\chi^2_{36;0.01} = 58.619$. Por lo tanto, con un nivel de significación de 0.01 se debe declarar diferencias altamente significativas entre los tratamientos.

La prueba de hipótesis entre bloques, para ver si hay efectos diferenciados entre las semillas de diferente procedencia se hace de manera similar. Véase el Programa 4.4.5.2.

4.4.5 Anexo de Programas

Programa 4.4.5.1

```
DATA IRIS;
INFILE 'C:\HIPOTESIS\IRIS.DAT';
INPUT Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8 Y9 Y10 Y11 Y12;
/*Y1,Y5,Y9 : longitud de sepalo de I. setosa, I.versicolour e I. virginica,respectivamente*/
/*Y2,Y6,Y10: ancho de sepalo de I. setosa, I.versicolour e I. virginica,respectivamente*/
/*Y3,Y7,Y11: longitud de petalo de I. setosa, I.versicolour e I. virginica,respectivamente*/
/*Y4,Y8,Y12: ancho de petalo de I. setosa, I.versicolour e I. virginica,respectivamente*/

PROC IML;      /*PARA EL EJEMPLO 4.4.6*/
USE IRIS;
START PRUEBA;

READ ALL VAR {Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8 Y9 Y10 Y11 Y12} INTO Y;

N=NROW(Y);  /*numero de individuos*/

p=NCOL(Y);  /*numero de variables*/

J=J(N,1,1);  /*un vector de puros unos de orden n=n1=n2=n3 */

MED=(1/N)*Y`*J;  /*vector de medias por var y por especie*/
MED1=MED[1:4,];
MED2=MED[5:8,];
MED3=MED[9:12,];
MEDG=(MED1+MED2+MED3)/3;

K=N*((MED1-MEDG)*(MED1-MEDG)`+(MED2-MEDG)*(MED2-
MEDG)`+(MED3-MEDG)*(MED3-MEDG)`);

/*K es la matriz de las SCPC entre especies*/

F1C1=J(4,4,0);F2C2=J(4,4,0);F3C3=J(4,4,0);

do h=1 to n;
F1C1=F1C1+((Y[h,{1,2,3,4}]-MED1)`*(Y[h,{1,2,3,4}]-MED1`));
F2C2=F2C2+((Y[h,{5,6,7,8}]-MED2)`*(Y[h,{5,6,7,8}]-MED2`));
F3C3=F3C3+((Y[h,{9,10,11,12}]-MED3)`*(Y[h,{9,10,11,12}]-MED3`));
END;

W=F1C1+F2C2+F3C3; /*W es la matriz de las SCPC dentro de especies*/
```

```

F=INV(W)*K;

CALL SVD(U,Q,V,K);
A=SQRT(DIAG(Q));
K12=U*A*U';
WM1PK=K12*INV(W)*K12;
C=EIGVAL(WM1PK);
D=EIGVEC(WM1PK);

PRINT MED1 MED2 MED3 MEDG;
PRINT K W F C D;

a={0,1,0,0};
TETA=0.1098; /*este valor de las tablas en el libro de Mardia*/
MM=((TETA/(1-TETA))*A`*W*A*(1/50+1/50))**0.5;
PRINT MM;

FINISH; /*termina el modulo PRUEBA */
RUN PRUEBA; /*correr el modulo PRUEBA */
QUIT;

```

Programa 4.4.5.2 Para la solución del Ejemplo 4.4.7.

```

DATA MELI;OPTIONS LS=80 NODATE;
INFILE 'C:MANOVA\ACACIA.DAT';
INPUT BLOQ $ TRAT $ Y1 Y2 Y3;
PROC ANOVA;CLASSES BLOQ TRAT;
MODEL Y1 Y2 Y3=BLOQ TRAT BLOQ*TRAT;
MANOVA H=BLOQ*TRAT/PRINTE PRINTH;
MANOVA H=TRAT/PRINTE PRINTH;
MANOVA H=BLOQ/PRINTE PRINTH;
RUN;

```


4.5 MÉTODO DE COMPONENTES PRINCIPALES

4.5.1 Introducción

Se cuenta con las calificaciones obtenidas en $p = 6$ materias de un grupo con $n = 55$ estudiantes (véase Cuadro 4.5.1), y existe el interés de generar con esos datos una “medida global” para cada individuo (Mardia *et al.*, 1989). De forma inmediata, una primera medida “global” que se podría considerar, es la *media aritmética* de sus calificaciones (i.e., los promedios), definida escalarmente para cada individuo como la siguiente combinación lineal:

$$\bar{x}_i = \frac{x_{i1} + x_{i2} + x_{i3} + x_{i4} + x_{i5} + x_{i6}}{6}, \quad i = 1, 2, \dots, 55,$$

lo que de manera genérica sería:

$$\bar{x}_i = \frac{1}{p} \sum_{j=1}^p x_{ij} = \frac{1}{p} x_i. \quad i = 1, 2, \dots, n. \quad (4.5-1)$$

Observe que x_i es la sumatoria de las 6 calificaciones del i -ésimo individuo y $\bar{x}_i$, es su promedio.

De manera equivalente, se puede presentar (4.5-1) como:

$$\bar{x}_i = \mathbf{x}'_i \mathbf{b} = \mathbf{b}' \mathbf{x}_i, \quad (4.5-2)$$

donde, $\mathbf{b} = \frac{1}{p} \mathbf{1} = \begin{bmatrix} \frac{1}{p} & \frac{1}{p} & \dots & \frac{1}{p} \end{bmatrix}_{p \times 1}$ es un vector de coeficientes de ponderaciones constantes. Recuerde que $\mathbf{x}'_i = \begin{bmatrix} x_{i1} & x_{i2} & \dots & x_{ip} \end{bmatrix}$.

Cuadro 4.5.1 Calificaciones de 55 estudiantes en 6 materias

Estud	Geom	Arit.	Hist. Méz.	Lect.y Redac	Comunic Educ.	Psic. Educ.	Prom.	MP1	MP2
1	8.0	7.7	8.0	9.7	8.3	9.0	8.45	8.30	8.57
2	5.7	5.0	7.0	8.7	7.3	7.3	6.83	6.46	7.13
3	5.7	5.0	6.7	8.3	7.0	7.3	6.67	6.34	6.93
4	8.7	7.7	8.0	9.7	8.0	8.0	8.35	8.31	8.38
5	6.3	6.0	8.7	10.0	8.7	9.0	8.12	7.63	8.51
6	6.3	6.7	7.3	8.7	8.3	8.7	7.67	7.38	7.90
7	5.7	5.0	6.3	8.3	6.3	7.3	6.48	6.20	6.71
8	7.3	5.0	6.7	8.7	7.0	8.0	7.12	6.88	7.31
9	8.7	6.3	8.0	9.3	8.0	8.0	8.05	7.91	8.16
10	5.0	5.0	6.7	6.0	6.3	6.0	5.83	5.63	6.00
11	7.0	6.7	7.7	9.3	8.3	7.7	7.78	7.55	7.97
12	6.0	7.7	7.0	9.3	7.3	7.7	7.50	7.34	7.63
13	6.7	7.3	8.0	9.7	9.3	9.0	8.33	8.00	8.60
14	7.0	6.3	7.0	8.7	8.0	8.3	7.55	7.33	7.73
15	5.7	5.0	6.7	8.3	7.7	6.7	6.68	6.35	6.95
16	8.3	6.3	8.7	9.7	7.3	8.0	8.05	7.86	8.20
17	6.3	7.0	6.3	8.3	7.7	6.7	7.05	6.95	7.13
18	8.0	6.0	6.7	8.7	8.7	8.7	7.80	7.60	7.96
19	6.7	5.0	7.3	9.0	8.3	8.3	7.43	7.04	7.75
20	6.0	5.3	6.7	8.3	6.3	7.3	6.65	6.40	6.85
21	5.0	5.0	7.3	8.0	6.3	7.7	6.55	6.16	6.86
22	5.3	5.0	7.3	8.3	8.0	7.3	6.87	6.44	7.21
23	6.0	7.3	7.0	8.7	6.3	7.0	7.05	6.95	7.13
24	6.7	7.3	6.3	8.3	8.0	6.7	7.22	7.16	7.26
25	7.3	7.7	6.7	7.0	7.3	7.0	7.17	7.25	7.10
26	5.7	5.7	7.0	9.0	8.0	7.3	7.12	6.76	7.4
27	6.0	5.0	7.0	9.0	8.0	7.3	7.05	6.66	7.36
28	7.0	7.3	8.0	9.7	8.7	9.7	8.4	8.09	8.65
29	7.0	8.0	8.7	9.0	9.0	8.7	8.40	8.18	8.58
30	7.3	6.3	6.0	9.7	7.0	7.3	7.27	7.15	7.36
31	6.0	6.0	7.7	8.3	7.0	8.3	7.22	6.91	7.46
32	8.0	6.0	7.3	8.7	9.0	8.7	7.95	7.71	8.14
33	7.7	5.7	8.0	9.3	9.0	8.3	8.00	7.68	8.26
34	6.0	6.3	8.0	8.7	8.3	9.3	7.77	7.36	8.09
35	7.7	7.3	8.3	9.7	9.3	9.0	8.55	8.29	8.76
36	7.7	6.0	8.7	9.3	8.3	8.7	8.12	7.80	8.37
37	5.7	5.3	6.7	8.3	7.7	7.7	6.90	6.55	7.18
38	7.7	6.0	7.3	9.3	8.7	9.3	8.05	7.75	8.29
39	7.0	7.0	8.7	9.7	8.7	8.7	8.30	7.98	8.56
40	5.0	5.3	6.7	8.7	7.0	7.0	6.62	6.25	6.91
41	6.7	7.0	7.7	9.7	8.3	8.3	7.95	7.68	8.17
42	7.0	7.3	6.7	9.0	7.0	7.3	7.38	7.33	7.43
43	6.7	5.7	7.0	9.3	8.3	9.0	7.67	7.3	7.96
44	9.0	8.7	8.3	10.0	9.3	10.0	9.21	9.13	9.29
45	5.0	5.7	7.0	8.7	7.0	7.0	6.73	6.39	7.01
46	7.7	8.8	7.3	10.0	8.3	10.0	8.68	8.58	8.77
47	7.7	6.0	6.7	9.7	8.7	9.3	8.02	7.73	8.25
48	6.3	5.7	6.7	8.7	6.3	8.0	6.95	6.71	7.14
49	5.0	5.0	7.0	8.0	5.7	7.3	6.33	6.00	6.60
50	8.0	8.3	6.7	9.7	7.3	9.7	8.28	8.25	8.31
51	6.7	6.7	7.7	9.3	7.7	8.3	7.73	7.48	7.94
52	5.3	6.7	8.0	8.3	7.7	6.3	7.05	6.79	7.26
53	5.7	5.0	5.3	6.0	6.3	6.7	5.83	5.71	5.93
54	5.0	5.0	6.3	7.0	5.7	6.7	5.95	5.71	6.14
55	5.0	5.0	5.3	6.7	5.3	6.3	5.6	5.45	5.72

Fuente: Calificaciones de un grupo de estudiantes del primer semestre (96-97), de la Escuela Normal Superior de Yucatán (módulo Oxkutzcab), usadas por cortesía del Profesor Daniel Suárez Burgos. [Est: Estudiante; Prom.: promedio; MP1: media ponderada 1; MP2: media ponderada 2.]

Para obtener, simultáneamente, los promedios de los 55 estudiantes, se define $\bar{x}$ como:

$$\bar{x} = Xb \quad (4.5-3)$$

donde X es la matriz de datos ($n \times p$). En este ejemplo, X es de orden 55×6 . Observe que el vector de medias definido en (4.1-4) es el vector que contiene las medias de cada una de las variables (materias) —y es el que se utilizará para el tema de este capítulo—, mientras que el vector de medias en (4.5-3) es un vector conformado por los promedios de cada uno de los individuos (estudiantes).

Otras medidas globales que se pueden construir para jerarquizar a los estudiantes, son las medias ponderadas, como las que aparecen en las dos últimas columnas del Cuadro 4.5.1. Con la media ponderada 1 (M.P.1) se usó el “criterio” de que en un estudiante de ciencias, deben tener mayor peso las materias con ese enfoque. Esta M.P.1 se obtuvo con la siguiente combinación lineal:

$$\bar{w}_i = \frac{x_1}{4} + \frac{x_2}{4} + \frac{x_3}{8} + \frac{x_4}{8} + \frac{x_5}{8} + \frac{x_6}{8}, \quad i = 1, \dots, 55, \quad (4.5-4)$$

dándole a Aritmética y Geometría, en forma conjunta, un 50 % de peso $\left(\frac{1}{4} + \frac{1}{4}\right)$ y un 50 % a las siguientes cuatro materias juntas $\left(\frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8}\right)$. En una media ponderada, los coeficientes de ponderación no siempre son constantes; de hecho, cuando los coeficientes son constantes se obtiene la media aritmética, un caso particular de la media ponderada. Puede obtenerse la M.P.1 simultáneamente para los 55 individuos, utilizando

$$w_{55 \times 1} = X_{55 \times 6} \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} & \frac{1}{8} \end{bmatrix}'.$$

De forma similar, se obtuvo la M.P.2 ubicada en la última columna del Cuadro 4.5.1, a través de la expresión:

$$w_{55 \times 1}^* = X_{55 \times 6} \begin{bmatrix} \frac{1}{10} & \frac{1}{10} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{bmatrix}' \quad (4.5-5)$$

bajo el “criterio” de que en la formación de un profesional social, las materias relacionadas con la expresión oral y escrita, la conducta y el entorno histórico, deben tener una ponderación más alta comparadas con las materias que promueven las estructuras formales que se usan en la ciencia.

Alguien podría opinar que no es apropiado dar el mismo peso a todas las materias como en la media aritmética, en tanto que otro podría decir que los “criterios” para el cálculo de las medias ponderadas son arbitrarios. Una de las medidas que no necesariamente asigna el mismo peso a las materias —sólo lo haría cuando las variables no estén correlacionadas completamente— y que tampoco las pondera arbitrariamente por la opinión de una persona, son las llamadas *componentes principales*. Este método fue introducido originalmente por Karl Pearson en 1901 y desarrollado por Hotelling en 1933 tal como se conoce ahora, según Pla (1986).

La fase esencial de este método, consiste en hacer combinaciones lineales de las p variables cuantitativas originales, correlacionadas en alguna medida, transformándolas en p nuevas variables cuantitativas no correlacionadas entre sí, conocidas como componentes principales. Entre otras características importantes, las componentes principales son índices que permiten ver con más claridad las diferencias entre individuos, jerarquizándolos por el criterio que se usa en la construcción de éstas sobre las variables originales, puesto que la varianza entre los individuos, de las combinaciones lineales resultantes, se hace máxima en la primera de ellas y decrece secuencialmente en las siguientes. Debido a que la primera componente principal tendrá la mayor varianza, es esta combinación lineal la que se usaría como medida global para diferenciar a cada individuo (véase el Cuadro 4.5.2).

Algunas veces, las componentes principales se usan sólo como paso previo a otro tipo de análisis, como la regresión múltiple o la regresión multivariada, para eliminar multicolinealidad en los datos.

En este capítulo se hace una breve descripción de las componentes principales poblacionales, las componentes principales muestrales, así como las propiedades más importantes de ambas, sin entrar en detalles de la teoría. A lo largo de este capítulo se desarrollan varios ejemplos con los datos del Cuadro 4.5.1 y se escribe el programa en IML (Interactive Matrix Language) de SAS para realizar la transformación componentes principales. Con datos de árboles y nutrientes del suelo, se ejemplifica al final el uso de las componentes principales en la regresión lineal.

4.5.2 Componentes Principales Poblacionales

Sea $\mathbf{x}$ un vector aleatorio con p variables, con vector de medias μ y matriz de varianzas y covarianzas Σ , i.e. $\mathbf{x} \sim (\mu, \Sigma)$. Por el teorema de la descomposición espectral, la matriz $\Sigma = \Sigma'$ se puede escribir como $\Sigma = \Gamma \Lambda \Gamma'$. La transformación componentes principales del vector $\mathbf{x}$ se define como:

$$\mathbf{y} = \Gamma'(\mathbf{x} - \mu) \quad (4.5-6)$$

donde Γ es una matriz ortogonal que contiene como columnas a los vectores característicos estandarizados de Σ , y $\Lambda = \Gamma' \Sigma \Gamma$ es una matriz diagonal ($p \times p$) que contiene los valores característicos de Σ , arreglados de tal manera que $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ (Mardia *et al.*, 1989; p.214). Es decir,

$$\Gamma = [\gamma_{(1)} \ \gamma_{(2)} \ \dots \ \gamma_{(i)} \ \dots \ \gamma_{(p)}] \quad (4.5-7)$$

y

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix} \quad (4.5-8)$$

De acuerdo a la definición de la transformación componentes principales se tiene que:

$$\begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_p \end{bmatrix} \rightarrow \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_p \end{bmatrix} = \begin{bmatrix} \gamma'_{(1)} \\ \vdots \\ \gamma'_{(i)} \\ \vdots \\ \gamma'_{(p)} \end{bmatrix} (\mathbf{x} - \boldsymbol{\mu}) = \begin{bmatrix} \gamma'_{(1)}(\mathbf{x} - \boldsymbol{\mu}) \\ \vdots \\ \gamma'_{(i)}(\mathbf{x} - \boldsymbol{\mu}) \\ \vdots \\ \gamma'_{(p)}(\mathbf{x} - \boldsymbol{\mu}) \end{bmatrix},$$

por lo que la i -ésima componente principal del vector $\mathbf{x}$ corresponde al i -ésimo elemento del vector $\mathbf{y} = (y_1, \dots, y_i, \dots, y_p)'$ dado por:

$$y_i = \gamma'_{(i)}(\mathbf{x} - \boldsymbol{\mu}) \quad (4.5-9)$$

donde $\gamma_{(i)}$ es la i -ésima columna de Γ y puede denominarse como el i -ésimo vector de ponderaciones de las variables originales.

La definición de las componentes principales sobre la matriz de covarianzas supone que las variables originales son conmensurables; cuando esto no es cierto, como frecuentemente ocurre, la matriz de covarianzas es sustituida por la matriz de correlaciones para estandarizar las varianzas.

4.5.2.1 Propiedades de las componentes principales poblacionales (*ibid.*, 214-217)

Si $\mathbf{x} \sim (\boldsymbol{\mu}, \Sigma)$ y $\mathbf{y}$ es el vector de la transformación componentes principales de $\mathbf{x}$ como se definió en (4.5-6), entonces se cumple que:

$$a) E(y_i) = 0 \quad (4.5-10a)$$

$$b) \text{var}(y_i) = \lambda_i \quad (4.5-10b)$$

$$c) \text{corr}(y_i, y_j) = 0, \quad \forall i \neq j \quad (4.5-10c)$$

$$d) \text{var}(y_1) \geq \text{var}(y_2) \geq \dots \geq \text{var}(y_p) \quad (4.5-10d)$$

$$e) \sum_{i=1}^p \text{var}(y_i) = \text{tr}(\Sigma) \quad (4.5-10e)$$

$$f) \prod_{i=1}^p \text{var}(y_i) = |\Sigma|$$

(4.5-10f)

Además se tiene que:

- La varianza de la primera componente principal, λ_1 , es mayor o igual que cualquier otra combinación lineal estandarizada (Teorema 8.2.2, *ibid.*)
- La última componente principal de $\mathbf{x}$, y_p , tiene la varianza más pequeña que cualquier otra combinación lineal estandarizada (Teorema 8.2.2, *ibid.*)
- Si $\alpha = \mathbf{a}'_i \mathbf{x}$ es una combinación lineal estandarizada que no está correlacionada con las primeras k componentes principales de $\mathbf{x}$, entonces la varianza de α se maximiza cuando α es la $(k+1)$ -ésima componente principal (Teorema 8.2.3, *ibid.*).
- La suma de los primeros k valores característicos dividida por la suma de todos los valores característicos, $(\lambda_1 + \lambda_2 + \dots + \lambda_k) / (\lambda_1 + \lambda_2 + \dots + \lambda_k + \lambda_{k+1} + \dots + \lambda_p)$, representa la "proporción de la variación total" que explican las primeras k componentes principales. En este contexto, la variación total es la $\text{tr } \Sigma = \sum_{i=1}^p \lambda_i$. La proporción de la variación total, cuantifica la información retenida al reducir la dimensión de p a k (*ibid.*, p. 218).
- Una desventaja de las componentes principales de un vector aleatorio es que no son invariantes a cambios de escala (*ibid.*, p. 219). Esta propiedad se puede interpretar como sigue.

Suponer que se toman $p = 3$ variables en n árboles, viz., la altura (h) en pies, el área basal (AB) en pulg², y la edad (E) en lustros, y se quieren generar componentes

principales expresadas en unidades diferentes a las de las variables originales, digamos, en metros, centímetros cuadrados, y años. Los dos caminos alternativos son: (a) primero transformar la escala de las variables originales (i.e. convertir a metros, centímetros cuadrados y años, respectivamente) y luego generar las componentes principales, o (b) primero generar las componentes principales sobre los datos en sus unidades originales y después multiplicar a las componentes relevantes por 0.3048, 6.45 y 0.2 para expresarlas en *m*, *cm*² y *años*. La teoría señala que, en general, las alternativas (a) y (b) no conducen a los mismos resultados. Esto se puede expresar algebraicamente así.

Sean μ y Σ , el vector de medias y la matriz de covarianzas, del vector aleatorio $\mathbf{x}$. Si la i -ésima variable se transforma multiplicándola por d_i , entonces la variable transformada $D\mathbf{x}$ tendrá como vector de medias a $D\mu$ y como matriz de covarianzas a $D\Sigma D$ donde D es la matriz diagonal que contiene a las d_i . Ahora bien, si $\gamma_{(i)}$ es un vector característico de Σ , se esperaría que $D^{-1}\gamma_{(i)}$ sea un vector característico de $D\Sigma D$ para que sea indiferente seguir (a) o (b); pero esto no ocurre ya que para ello, siguiendo las alternativas (a) $y_i^* = \gamma_{(i)}^*(D\mathbf{x} - D\mu)$ y (b) $y_i = \gamma_{(i)}(\mathbf{x} - \mu)$ tendrían que producirse resultados proporcionales. Y esto sólo puede cumplirse si $D\Sigma D'\gamma_{(i)}^* = a_i\gamma_{(i)}^*$. Al multiplicar en ambos lados de la igualdad por D^{-1} , nos queda $\Sigma D'\gamma_{(i)}^* = a_i D^{-1}\gamma_{(i)}^*$; pero $D'\gamma_{(i)}^* \neq D^{-1}\gamma_{(i)}^*$. Por lo tanto, $D^{-1}\gamma_{(i)}$ no es eigenvector de $D\Sigma D$ y en consecuencia, las componentes principales de un vector aleatorio no son invariantes a cambios de escala.

- Si la matriz de covarianzas de $\mathbf{x}$ tiene rango $r < p$, entonces la variación total de $\mathbf{x}$ puede ser explicada en su totalidad por las primeras r componentes principales (*ibid.*, p. 220).

4.5.3 Componentes Principales Muestrales

Siguiendo a Mardia *et al.* (1989), la transformación componentes principales muestral es análoga a la transformación componentes principales poblacional, si en lugar de usar los parámetros poblacionales μ y Σ , se emplean sus estimaciones de máxima verosimilitud $\bar{x}$ y S , respectivamente. Sea $X = \begin{bmatrix} x_{(1)} & x_{(2)} & \cdots & x_{(p)} \end{bmatrix}$ una matriz de datos muestrales de orden $(n \times p)$ y a un vector estandarizado de dimensión $(p \times 1)$, i.e. $a'a = 1$. Se puede transformar linealmente a las variables originales generando un vector $y = Xa$ de dimensión $(n \times 1)$ que represente la suma ponderada de las columnas de X . La varianza muestral de y se denota como S_y y es igual a $a'Sa$ donde S es la matriz de varianzas y covarianzas estimada por máxima verosimilitud de la matriz de datos X .

La transformación componentes principales de una muestra aleatoria multivariada se realiza bajo los siguientes criterios:

- Seleccionar un vector a en particular, digamos a_1 , para hacer la combinación lineal de las variables en X mediante Xa_1 , de modo que la varianza de Xa_1 , igual a $a_1'Sa_1$, sea mayor o igual que la varianza de cualquier otra combinación lineal estandarizada, esto es,

$$var(Xa_1) \geq var(Xb) \quad \forall \quad b_{p \times 1} \cdot \exists \cdot b'b = 1 \quad (4.5-11)$$

- Seleccionar otro vector $a \neq a_1$, digamos a_2 , de manera que se cumplan las siguientes características:

$$(i) \quad cov(Xa_1, Xa_2) = 0$$

$$(ii) \quad var(Xa_1) \geq var(Xa_2) \geq var(Xc) \quad \forall \quad c_{p \times 1} \cdot \exists \cdot c'c = 1 \text{ y que satisfaga } cov(Xa_1, Xc) = 0. \quad (4.5-12)$$

- Así sucesivamente hasta obtener el vector a_p , que haga que la combinación lineal Xa_p tenga la varianza más pequeña que cualquier otra combinación lineal

estandarizada y que, al mismo tiempo, no esté correlacionada con ninguna de las anteriores $p - 1$ combinaciones lineales ya escogidas.

Se cumple por el teorema de la descomposición espectral, la relación $S = GLG'$, donde L es una matriz diagonal que tiene ordenados de mayor a menor en la diagonal principal, los valores característicos de S ; G es la matriz ortogonal que contiene en sus columnas los vectores característicos asociados a los valores característicos de S . El vector a_1 que proporciona la varianza más grande en el anterior proceso es $g_{(1)}$, que es el vector característico asociado al valor característico más grande de la matriz de covarianzas muestral S . De lo anterior se tiene que la primera componente principal muestral, por analogía con la contraparte poblacional en la expresión (4.5-6), se obtiene usando:

$$y_{(1)} = (X - 1\bar{x}')g_{(1)}, \quad (4.5-13)$$

donde $\bar{x} = \frac{1}{n}X'1$ [como en la expresión (4.1-4)]. Obsérvese que en la expresión (4.5-6) se dan las p componentes principales *de un vector aleatorio* (el resultado es un vector aleatorio), mientras que en (4.5-13) se obtiene el vector de valores de la primera componente principal correspondiente a todos los individuos estudiados, siendo de dimensión $n \times 1$. En general, el vector de valores de los n individuos que corresponden a la i -ésima componente principal es:

$$y_{(i)} = (X - 1\bar{x}')g_{(i)} \quad (4.5-14)$$

El r -ésimo elemento del vector $y_{(i)}$, denotado como y_{ri} , representa el valor de la i -ésima componente principal del r -ésimo individuo, $r = 1, \dots, n$, y obviamente, $i = 1, \dots, p$.

Para agrupar todas las componentes principales que se obtienen individualmente por (4.5-14), se tiene:

$$Y = (X - 1\bar{x}')G \quad (4.5-15)$$

de modo que $Y = [y_{(1)} \ y_{(2)} \ \cdots \ y_{(p)}]$ es la matriz de valores de las p componentes principales.

La matriz de varianzas y covarianzas muestral de Y se obtiene con la siguiente expresión:

$$S_Y = \frac{1}{n} Y' H Y = L \quad (4.5-16)$$

donde $H = I - \frac{1}{n} 11'$. Es decir, la matriz de varianzas y covarianzas de Y está dada por la matriz L que contiene en su diagonal (ordenados descendientemente) los valores característicos de S :

$$L = \begin{bmatrix} l_1 & 0 & \cdots & 0 \\ 0 & l_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & l_p \end{bmatrix},$$

lo cual significa que las columnas de Y no están correlacionadas y la varianza de la i -ésima componente principal, $y_{(i)}$, es l_i ; esto es, las nuevas variables generadas a partir de las variables originales como se define en (4.5-6) tienen la propiedad de no estar correlacionadas y tener una varianza que disminuye conforme se toma una componente menor.

4.5.3.1 Propiedades de las componentes principales muestrales

Por analogía con la contraparte poblacional, se tienen las siguientes propiedades de las componentes principales muestrales:

$$a) \bar{y}_{(i)} = 0 \quad (4.5-17a)$$

$$b) \text{var}(y_{(i)}) = l_i \quad (4.5-17b)$$

$$c) \text{corr}(y_{(i)}, y_{(j)}) = 0 \quad i \neq j \quad (4.5-17c)$$

$$d) \text{var}(y_{(1)}) \geq \text{var}(y_{(2)}) \geq \dots \geq \text{var}(y_{(p)}) \geq 0 \quad (4.5-17d)$$

$$e) \sum_{i=1}^p \text{var}(y_{(i)}) = \sum_{i=1}^p l_i = \text{tr}(\mathbf{S}) \quad (4.5-17e)$$

$$f) \prod_{i=1}^p \text{var}(y_{(i)}) = \prod_{i=1}^p l_i = |\mathbf{S}| \quad (4.5-17f)$$

EJEMPLO 4.5.1 Con las calificaciones del Cuadro 4.5.1, se puede hacer la transformación componentes principales, para ejemplificar la transformación propuesta en la expresión (4.5-15). En primer lugar, se requieren el vector de medias y la matriz de varianzas y covarianzas de las variables originales. Usando las expresiones (4.1-4) y (4.1-7) se obtienen:

$$\bar{\mathbf{x}} = \begin{bmatrix} 6.6127 \\ 6.2564 \\ 7.2345 \\ 8.7909 \\ 7.6836 \\ 7.9673 \end{bmatrix} \quad \text{y} \quad \mathbf{S} = \begin{bmatrix} 1.1735 & 0.6989 & 0.3857 & 0.6258 & 0.6673 & 0.7153 \\ 0.6989 & 1.1606 & 0.3555 & 0.5203 & 0.5038 & 0.5411 \\ 0.3857 & 0.3555 & 0.6419 & 0.4632 & 0.5153 & 0.4551 \\ 0.6258 & 0.5203 & 0.4632 & 0.8343 & 0.6117 & 0.6621 \\ 0.6673 & 0.5038 & 0.5153 & 0.6117 & 0.9876 & 0.6875 \\ 0.7153 & 0.5411 & 0.4551 & 0.6621 & 0.6875 & 0.9957 \end{bmatrix}.$$

Enseguida se tienen que obtener los valores característicos de $\mathbf{S}$ y sus vectores característicos asociados:

$l_1 = 3.8319$	$l_2 = 0.6989$	$l_3 = 0.4564$	$l_4 = 0.3202$	$l_5 = 0.2701$	$l_6 = 0.2160$
$\mathbf{g}_{(1)}$	$\mathbf{g}_{(2)}$	$\mathbf{g}_{(3)}$	$\mathbf{g}_{(4)}$	$\mathbf{g}_{(5)}$	$\mathbf{g}_{(6)}$
0.4675	-0.2391	-0.7233	0.2800	0.3215	0.1390
0.4074	-0.7831	0.4417	-0.0243	-0.1583	-0.0040
0.2909	0.3156	0.4802	0.1899	0.4842	0.5609
0.3969	0.1934	0.1088	-0.4029	0.4465	-0.6569
0.4258	0.3645	0.1110	0.5907	-0.4806	-0.3060
0.4382	0.2443	-0.1642	-0.6113	-0.4547	0.3753

Se sabe que las columnas $\mathbf{g}_{(i)}$ conforman la matriz $\mathbf{G}_{6 \times 6}$ y también que cada columna de la matriz de datos originales, $\mathbf{X}_{55 \times 6}$, puede denotarse como $\mathbf{x}_{(j)}$. A partir de la expresión (4.5-14), se pueden escribir las 6 componentes principales resultantes como:

$$\begin{aligned} y_{(1)} &= 0.468x_{(1)} + 0.407x_{(2)} + 0.291x_{(3)} + 0.397x_{(4)} + 0.426x_{(5)} + 0.438x_{(6)} - \bar{x}g_{(1)}, \\ y_{(2)} &= -0.239x_{(1)} - 0.783x_{(2)} + 0.316x_{(3)} + 0.193x_{(4)} + 0.365x_{(5)} + 0.244x_{(6)} - \bar{x}g_{(2)}, \\ y_{(3)} &= -0.723x_{(1)} + 0.442x_{(2)} + 0.480x_{(3)} + 0.109x_{(4)} + 0.111x_{(5)} - 0.164x_{(6)} - \bar{x}g_{(3)}, \\ y_{(4)} &= 0.280x_{(1)} - 0.024x_{(2)} + 0.190x_{(3)} - 0.403x_{(4)} + 0.591x_{(5)} - 0.611x_{(6)} - \bar{x}g_{(4)}, \\ y_{(5)} &= 0.321x_{(1)} - 0.158x_{(2)} + 0.484x_{(3)} + 0.447x_{(4)} - 0.481x_{(5)} - 0.455x_{(6)} - \bar{x}g_{(5)}, \\ y_{(6)} &= 0.139x_{(1)} - 0.004x_{(2)} + 0.561x_{(3)} - 0.657x_{(4)} - 0.306x_{(5)} + 0.375x_{(6)} - \bar{x}g_{(6)}, \end{aligned}$$

que resultan de una descomposición espectral de $\mathbf{S}$. Para los datos del Cuadro 4.5.1, se genera una nueva matriz $\mathbf{Y}$ como la definida en (4.5-15) la cual puede verse en el Cuadro 4.5.2. La variación total está dada por la sumatoria de todos los l_i , que en este caso es igual a 5.7935; la primera componente principal explica alrededor de 66 % de la variación total; las primeras dos componentes principales explican 78.2 %; y las últimas dos componentes, la quinta y la sexta, explican aproximadamente 8.38 % de la variación total de los datos. Los cálculos de este ejemplo se obtuvieron con el Programa 4.5.9.1.

4.5.4 Correlación entre las Variables Originales y las Componentes Principales

Se trata de presentar la correlación poblacional entre el vector aleatorio $\mathbf{x}$ y su correspondiente vector de componentes principales $\mathbf{y}$. Se tiene el supuesto de que el vector aleatorio $\mathbf{x} \sim (\mu, \Sigma)$; para mayor sencillez, se asume que $\mu = 0$ y siempre se tiene que $E(\mathbf{y}) = 0$. En este contexto, la covarianza entre $\mathbf{x}$ y $\mathbf{y}$ es: (Mardia *et al.*, 1989; p. 223)

$$\text{Cov}(\mathbf{x}, \mathbf{y}) = \Gamma \Lambda, \quad (4.5-18)$$

Cuadro 4.5.2 Las seis componentes principales generadas a partir de las calificaciones del Cuadro 4.5.1

Est.	Identific.	C.P. I	C.P. II	C.P. III	C.P. IV	C.P. V	C.P. VI
1	A	2.535	-0.568	-0.000	-0.135	0.228	0.218
2	B	-1.499	0.808	0.050	-0.052	0.239	-0.327
3	#	-1.872	0.526	-0.171	-0.125	0.059	-0.141
4	D	2.297	-1.089	-0.376	0.495	1.052	0.032
5	E	1.541	1.595	0.891	-0.321	0.231	0.062
6	F	0.601	0.134	0.392	-0.133	-0.809	0.138
7	G	-2.287	0.145	-0.441	-0.614	0.202	-0.151
8	H	-0.659	0.392	-1.400	-0.266	0.434	0.082
9	\$	1.567	-0.070	-1.038	0.691	1.095	0.300
10	J	-3.980	-0.324	0.221	0.987	-0.265	0.999
11	K	0.845	-0.035	0.307	0.508	0.332	-0.310
12	L	0.155	-1.165	1.025	-0.519	-0.006	-0.540
13	M	2.190	0.421	0.874	0.102	-0.607	-0.267
14	N	0.375	-0.022	-0.403	0.083	-0.340	0.010
15	O	-1.837	0.635	0.005	0.656	-0.004	-0.580
16	P	1.445	0.069	-0.446	0.137	1.821	0.589
17	Q	-0.858	-1.201	0.262	0.699	-0.322	-0.729
18	R	1.106	0.232	-1.391	0.482	-0.635	-0.082
19	S	0.039	1.330	-0.550	0.144	-0.096	-0.147
20	T	-1.908	-0.036	-0.333	-0.461	0.445	0.114
21	U	-2.267	0.668	0.447	-0.744	0.145	0.660
22	V	-1.459	1.176	0.517	0.468	-0.259	-0.166
23	W	-0.979	-1.503	0.787	-0.431	0.588	-0.101
24	X	-0.421	-1.422	0.139	0.981	-0.385	-0.766
25	Y	-0.544	-2.186	-0.195	1.142	-0.442	0.721
26	Z	-0.796	0.573	0.469	0.224	-0.074	-0.741
27	a	-0.941	1.049	-0.057	0.325	0.133	-0.696
28	b	2.382	0.301	0.476	-0.597	-0.540	0.221
29	c	2.282	-0.296	1.242	0.590	-0.314	0.604
30	d	-0.243	-0.825	-0.938	-0.405	0.654	-1.236
31	e	-0.596	0.231	0.370	-0.486	0.027	0.833
32	f	1.409	0.531	-1.069	0.773	-0.488	0.163
33	g	1.413	1.077	-0.518	0.832	0.252	-0.030
34	h	0.764	0.887	0.670	-0.441	-0.776	0.715
35	i	2.745	0.276	0.295	0.439	-0.140	0.041
36	j	1.616	0.905	-0.192	0.300	0.698	0.726
37	k	-1.277	0.644	-0.027	0.037	-0.506	-0.206
38	l	1.642	0.756	-0.919	-0.096	-0.445	0.045
39	m	2.025	0.513	0.844	0.155	0.301	0.240
40	n	-2.050	0.463	0.560	-0.306	0.102	-0.614
41	o	1.248	0.026	0.602	-0.111	0.094	-0.391
42	p	-0.050	-1.450	-0.019	-0.099	0.426	-0.429
43	q	0.663	0.916	-0.467	-0.479	-0.536	-0.253
44	r	4.480	-0.828	-0.158	0.037	-0.265	0.394
45	s	-1.800	0.244	0.881	-0.259	0.184	-0.448
46	t	3.197	-1.276	0.235	-1.111	-0.702	-0.042
47	u	1.626	0.644	-1.163	-0.371	-0.557	-0.556
48	v	-1.139	-0.172	-0.445	-0.976	0.338	0.154
49	w	-2.785	0.256	0.302	-0.911	0.470	0.525
50	x	2.282	-1.641	-0.585	-1.415	-0.334	0.056
51	y	0.712	-0.036	0.359	-0.297	0.251	0.057
52	z	-1.129	-0.288	1.735	0.994	0.409	-0.063
53	+	-3.753	-0.762	-1.073	0.489	-1.036	0.574
54	*	-3.648	-0.305	-0.044	-0.274	-0.042	0.564
55	@	-4.404	-0.922	-0.536	-0.335	-0.286	0.173

(Est.: Estudiante; C.P.: Componente Principal).

Prueba.

$$Cov(\mathbf{x}, \mathbf{y}) = E[(\mathbf{x} - \mathbf{0})(\mathbf{y} - \mathbf{0})'] = E(\mathbf{x}\mathbf{y}'),$$

donde $\mathbf{y} = \Gamma' \mathbf{x}$; usando el supuesto que $\mu = \mathbf{0}$; i.e., $\mathbf{y}' = \mathbf{x}'\Gamma$, entonces,

$$Cov(\mathbf{x}, \mathbf{y}) = E(\mathbf{x}\mathbf{x}'\Gamma) = E(\mathbf{x}\mathbf{x}')\Gamma;$$

como $Var(\mathbf{x}) = E[(\mathbf{x} - \mu)(\mathbf{x} - \mu)'] = \Sigma = E(\mathbf{x}\mathbf{x}')$, cuando $\mu = \mathbf{0}$, se tiene:

$$Cov(\mathbf{x}, \mathbf{y}) = \Sigma\Gamma.$$

Además se sabe que $\Sigma = \Gamma\Lambda\Gamma'$ y que Γ es ortogonal; por lo tanto:

$$Cov(\mathbf{x}, \mathbf{y}) = \Gamma\Lambda\Gamma'\Gamma = \Gamma\Lambda\mathbf{I} = \Gamma\Lambda$$

como se presentó en (4.5-18). $\square$

Si $var(x_i) = \sigma_{ii}$ y $var(y_j) = \lambda_j$, la covarianza entre x_i y y_j es $\gamma_{ij}\lambda_j$. Entonces la correlación simple entre x_i y y_j denotada por ρ_{ij} es:

$$\rho_{ij} = \frac{cov(x_i, y_j)}{\sqrt{var(x_i)var(y_j)}} = \frac{\gamma_{ij}\lambda_j}{\sqrt{\sigma_{ii}\lambda_j}} = \frac{\gamma_{ij}\sqrt{\lambda_j}}{\sqrt{\sigma_{ii}}} = \gamma_{ij}\left(\frac{\lambda_j}{\sigma_{ii}}\right)^{1/2}. \quad (4.5-19)$$

Si en vez de Σ se usa una matriz de correlaciones poblacional $\mathbf{P}$ (que tiene puros unos en la diagonal principal), $var(x_i) = \sigma_{ii} = 1$, entonces $\rho_{ij} = \gamma_{ij} \sqrt{\lambda_j}$.

Se puede asumir que el cuadrado de la correlación simple entre x_i y y_j , es la proporción de la variación de x_i explicada por y_j (Mardia *et al.*, 1989; p. 223). Es decir, la proporción de la variación de x_i explicada por y_j se puede expresar así:

$$\rho_{ij}^2 = \frac{\gamma_{ij}^2 \lambda_j}{\sigma_{ii}} = \frac{1}{\sigma_{ii}} (\gamma_{ij}^2 \lambda_j); \quad (4.5-20)$$

como los elementos de $\mathbf{y}$ no están correlacionados, cualquier conjunto G de componentes explican una proporción (*loc. cit.*, p. 223):

$$\rho_{iG}^2 = \sum_{j \in G} \rho_{ij}^2 = \frac{1}{\sigma_{ii}} \sum_{j \in G} \lambda_j \gamma_{ij}^2 \quad (4.5-21)$$

de la variación en x_i . En esta expresión σ_{ii} representa la variación en x_i a ser explicada y $\sum_{j \in G} \lambda_j \gamma_{ij}^2$ es la cantidad de la variación en x_i que explica el conjunto G de componentes. Observe que $\sum_{j=1}^p \rho_{ij}^2 = \frac{1}{\sigma_{ii}} \gamma_i' \Lambda \gamma_i = \frac{1}{\sigma_{ii}} \sigma_{ii} = 1$; es decir, cuando G incluye a todas las componentes, el numerador de ρ_{iG}^2 es el i -ésimo elemento de $\Sigma = \Gamma \Lambda \Gamma'$ que es σ_{ii} .

La proporción de la variación total atribuida a las componentes en G puede expresarse como la suma sobre todas las p variables de la proporción en cada variable explicada por las componentes en G , donde cada variable es ponderada por su varianza, esto es (*Ibid.*, p. 223)

$$\sum_{j \in G} \lambda_j = \sum_{i=1}^p \sigma_{ii} \rho_{iG}^2 \quad (4.5-22)$$

4.5.5 Estructura de Correlación Muestral

Por analogía con la parte poblacional, la proporción de la variación en x_i que explica y_j está dada por:

$$r_{ij}^2 = \frac{1}{s_{ii}} l_j g_{ij}^2 \quad (4.5-23)$$

y la raíz cuadrada del lado derecho de esta expresión, es la correlación simple muestral entre x_i y y_j .

EJEMPLO 4.5.2 Usando los datos del Ejemplo 4.5.1, se tienen los siguientes cálculos hechos con el Programa 4.5.9.1. del Anexo 4.5.9.

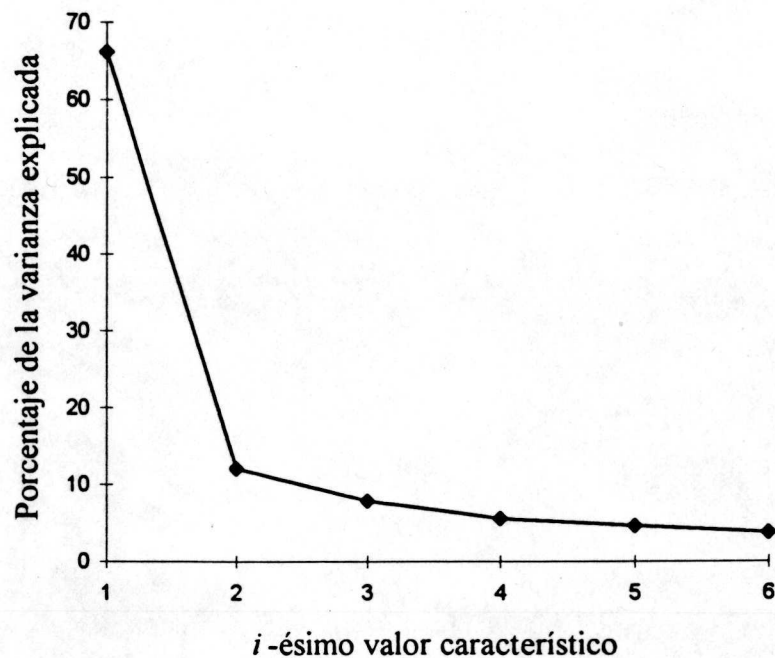
		<i>j</i> -ésima Componente Principal					
<i>r</i> _{<i>ij</i>}		1	2	3	4	5	6
<i>i</i> -ésima Variable	1	0.8448	-0.1846	-0.4511	0.1463	0.1542	0.0597
	2	0.7403	-0.6077	0.2770	-0.0128	-0.0764	-0.0017
	3	0.7107	0.3293	0.4049	0.1341	0.3141	0.3254
	4	0.8507	0.1770	0.0805	-0.2496	0.2541	-0.3342
	5	0.8387	0.3067	0.0754	0.3363	-0.2513	-0.1431
	6	0.8596	0.2047	-0.1112	-0.3467	-0.2368	0.1748

		<i>j</i> -ésima Componente Principal					
r_{ij}^2		1	2	3	4	5	6
<i>i</i> -ésima Variable	1	0.7137	0.0341	0.2034	0.0214	0.0238	0.0036
	2	0.5480	0.3693	0.0767	0.0002	0.0058	0.0000
	3	0.5051	0.1084	0.1640	0.0180	0.0987	0.1059
	4	0.7236	0.0313	0.0065	0.0623	0.0646	0.1117
	5	0.7035	0.0940	0.0057	0.1131	0.0632	0.0205
	6	0.7389	0.0419	0.0124	0.1202	0.0561	0.0306

Los valores de r_{ij}^2 se pueden interpretar de la siguiente manera: la primera componente principal explica 71.37 % de la variación de x_1 , 54.8 % de x_2 , 50.51 % de x_3 , 72.36 % de x_4 , 70.35 % de x_5 y 73.89 % de x_6 ; así enseguida, hasta la sexta componente principal que explica 0.36 % de la variación de x_1 , 0 % de x_2 , 10.59 % de x_3 , 11.17 % de x_4 , 2.05 % de x_5 y 3.06 % de x_6 . Se puede observar también que las dos primeras componentes juntas explican alrededor de 75 % de la variación de x_1 , aproximadamente 92 % de la de x_2 , y así sucesivamente. Observe que las tres primeras

componentes explican conjuntamente arriba del 95 % de la variación de x_1 y x_2 , mientras que las últimas tres componentes explican alrededor del 20 % de x_3 , x_4 , x_5 y x_6 . Si se ignorara la última componente principal se perdería más información sobre las variables x_3 y x_4 que sobre las otras; ignorando las dos últimas componentes principales se pierde información sustancial sobre las últimas 4 variables, siendo más drástico en x_3 y x_4 . Como puede notarse, examinando la estructura de correlación muestral, se tiene una idea de lo que pasaría si una componente se elimina.

EJEMPLO 4.5.3 Para visualizar las contribuciones de las 6 diferentes componentes principales generadas en el Ejemplo 4.5.1 se usará una gráfica de puntos (debido a Cattell, 1966 y citado por Mardia *et al.* 1989), donde aparecen los porcentajes de la varianza explicada por los valores característicos muestrales. En este tipo de gráfica frecuentemente se puede indicar dónde terminan los valores característicos grandes y comienzan los pequeños.



Otras “reglas de dedo” para la exclusión de componentes principales que mencionan Mardia *et al.* (1989) son: (a) incluir sólo las componentes suficientes que expliquen el 90 % de la variación total; (b) excluir aquellas componentes principales cuyos valores característicos son menores que el promedio, i.e. menos que 1 si se está usando la matriz de correlaciones (citando a Kaizer). Se ha sugerido que cuando $p \leq 20$, el criterio de Kaiser tiende a incluir muy pocas componentes. Similarmente, se dice que la prueba de puntos de Cattell es muy conservadora, i.e. desecha muy pocas componentes. Por lo tanto, en la práctica frecuentemente se usa una combinación de ellas.

La prueba de isotropía que se revisará en este mismo capítulo puede dar también alguna indicación del número de componentes a incluir.

4.5.6 Graficación de las Componentes Principales

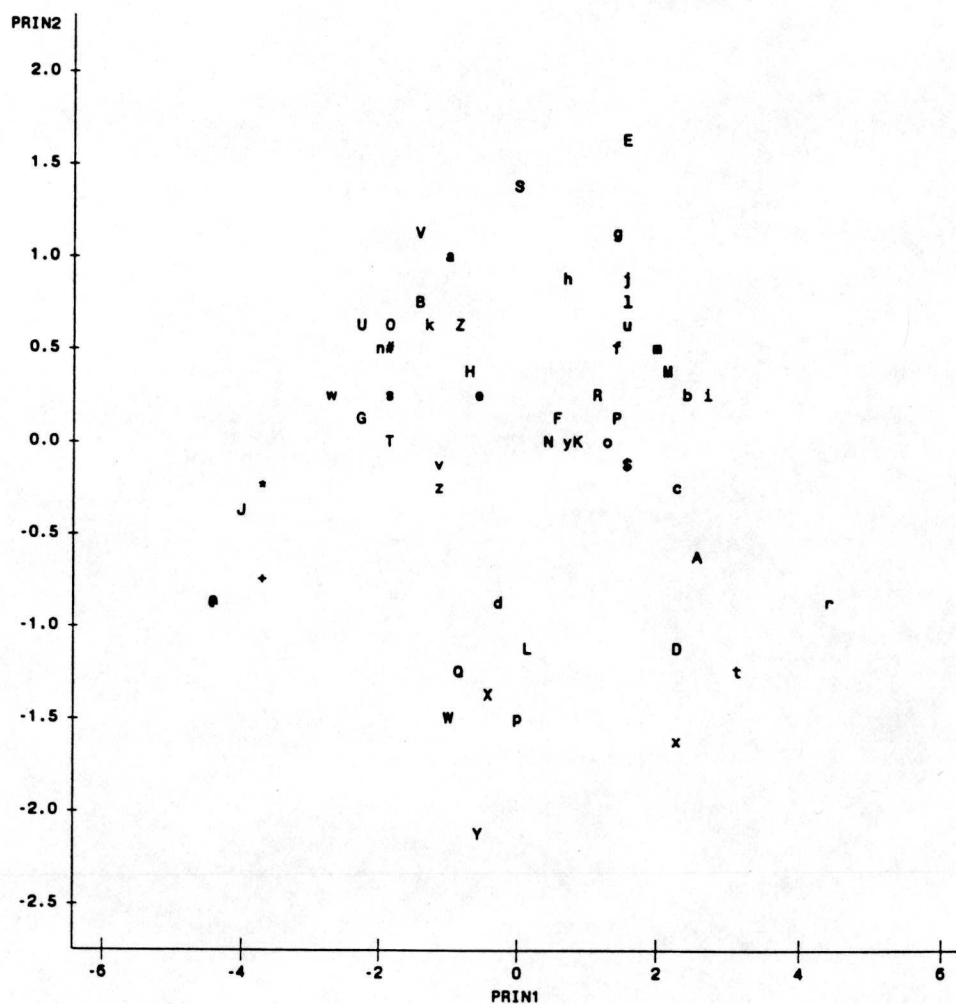
Pla (1986) remarca que independientemente del grado de complejidad de un método estadístico, al concluir el análisis de los datos será necesario interpretar adecuadamente los resultados; cuando es posible presentar gráficamente un resultado, a menudo, se facilita su interpretación.

El método de componentes principales genera nuevas variables no correlacionadas entre sí, a partir de combinaciones lineales estandarizadas de las p variables iniciales. Usualmente, el número de componentes que explican la mayor parte de la información contenida en las variables iniciales es menor que p ; si pasa que $k < p$ componentes principales explican la mayor parte de la variabilidad de las p variables originales, se dice que la dimensionalidad del problema inicial se reduce de p a k . Por ejemplo, si de un grupo de árboles, inicialmente se estudia la relación entre ellos sobre mediciones de edad, altura, diámetro normal, diámetro de copa, altura de copa, área basal y volumen, y después de generar las componentes principales se puede estudiar la mayor parte de esa información inicial con sólo las primeras tres componentes, se dirá que el problema

redujo su dimensionalidad de $p = 7$ a $k = 3$. Esto de entrada es una ventaja, debido a que nos es imposible hacer interpretaciones geométricas en más de tres dimensiones.

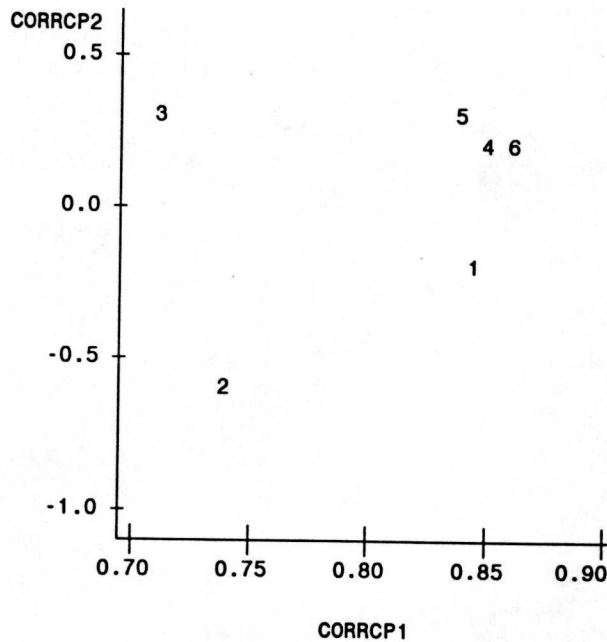
Mardia *et al.* (1989) describen cómo puede usarse gráficamente la reducción de la dimensionalidad que proveen las componentes principales. Si las dos primeras componentes explican la mayor parte de la varianza, entonces una gráfica de dispersión que muestre la distribución de los individuos en estas dos dimensiones dará frecuentemente una indicación adecuada de la distribución completa de los datos. Tales gráficas pueden aumentar o mejorar los métodos gráficos sobre las variables originales.

EJEMPLO 4.5.4 En la figura que sigue se puede ver la dispersión de los 55 estudiantes de la Tabla 4.5.1 sobre los ejes de las dos primeras componentes que se generaron.



Otra forma que ayuda a la interpretación de resultados, es usar como ejes coordenados a las correlaciones de las variables originales con la dos primeras componentes, para graficar las 6 variables del ejemplo.

The SAS System 01:45 Saturday, August 3, 1997 4
Plot of CORRCP2*CORRCP1=VAR.



En las abscisas se tienen la correlaciones de las variables originales con la primera componente principal y en la ordenada, las correlaciones de las variables originales con la segunda componente principal. Mientras más cerca esté un punto de la circunferencia unitaria, estará mejor explicado por esas dos componentes.

4.5.7 Pruebas de Hipótesis sobre las Componentes Principales

Se presentarán dos pruebas de hipótesis basadas en la matriz de covarianzas: (a) una plantea si las primeras k componentes principales explican una proporción de la variación total igual a un valor que se fija anticipadamente y posteriormente, (b) se escribirá la prueba de isotropía en la que se indaga si las últimas $(p - k)$ componentes principales

son iguales. Estas pruebas pueden usarse como criterio para decidir las componentes principales posibles de incluir o excluir.

En todas estas pruebas se asumirá que se tiene una muestra aleatoria de datos normales p -dimensional de tamaño n .

4.5.7.1 La hipótesis que $\frac{(\lambda_1 + \dots + \lambda_k)}{(\lambda_1 + \dots + \lambda_p)} = \psi$ (*ibid.*, pp. 233-234)

Si se decide incluir las primeras k componentes en un análisis, podrían ignorarse las restantes $(p - k)$ si puede asegurarse que todos sus correspondientes valores característicos poblacionales son cero, pero esta situación se detecta fácilmente en la práctica, ya que sólo ocurre cuando Σ tiene rango k , en cuyo caso con probabilidad uno también S tendría rango k (y entonces, los l_i , $i = k + 1, \dots, p$, valdrían cero). Frecuentemente es conveniente contar con un procedimiento para decidir si las primeras k componentes principales incluyen toda la variación importante en $\mathbf{x}$, o viendo el problema desde otra perspectiva, se puede probar si la proporción de la varianza explicada por las últimas $(p - k)$ componentes es menor que un valor determinado.

En esta prueba de hipótesis se plantea si la proporción de la variación total explicada por las primeras k componentes principales es igual a un valor fijo; en otras palabras, se propone un valor de referencia y se prueba si la proporción de la variación total que explican las primeras k componentes con respecto al total que explican las p que existen, no es diferente al valor fijo.

Sean $l_1, l_2, \dots, l_p$ los valores característicos de S_u y denótese la contraparte muestral de ψ por:

$$\hat{\psi} = \frac{(l_1 + \dots + l_k)}{(l_1 + \dots + l_p)} \quad (4.5-24)$$

Se puede demostrar que $\hat{\psi}$ se distribuye asintóticamente normal con (ψ, τ^2) [ibid., pp. 233-234] donde,

$$\tau^2 = \frac{2 \operatorname{tr} \Sigma^2}{(n-1)(\operatorname{tr} \Sigma)^2} (\psi^2 - 2\alpha\psi + \alpha^2) \quad (4.5-25)$$

$$\text{con } \alpha = \frac{\lambda_1^2 + \lambda_2^2 + \dots + \lambda_k^2}{\lambda_1^2 + \lambda_2^2 + \dots + \lambda_p^2}.$$

El valor de α puede interpretarse como la proporción de la varianza explicada por las primeras k componentes de un vector que tiene como matriz de covarianzas a Σ^2 . La prueba resulta la misma si se toman los valores característicos de S o los de S_u , debido a que se trabaja con cocientes que involucran tanto en el numerador como en el denominador a estos valores característicos, por lo que en los cálculos resultan las mismas proporciones.

Con la ecuación de la varianza de $\hat{\psi}$ se puede construir un intervalos de confianza (I.C.) aproximado para ψ que estaría dado por:

$$\text{I.C.} = \hat{\psi} \pm Z_{0.05/2} \hat{\tau} \quad (4.5-26)$$

donde $100(1 - 0.05)\%$ es el nivel de confianza del intervalo que se propone y $\hat{\tau}$ es la desviación estándar de $\hat{\psi}$. En la estimación de los valores de α y τ^2 para construir este I.C. aproximado, los valores característicos poblacionales se sustituyen directamente por los muestrales.

El I.C. va a sugerir, que el valor verdadero de la proporción de la variación total, que explican las primeras k componentes principales, se encuentra entre $\hat{\psi} - Z_{0.05/2} \hat{\tau}$ y $\hat{\psi} + Z_{0.05/2} \hat{\tau}$ con un 95% de confiabilidad. Si se desea expresar el intervalo en porcentajes, sólo habrá que multiplicar las proporciones por 100; el nivel de confiabilidad del intervalo se controla con el nivel de significancia α en los valores tabulados de Z .

EJEMPLO 4.5.5 Usando las calificaciones del Cuadro 4.5.1 podemos construir un I.C. para la proporción de la varianza explicada por las dos primeras componentes principales; en este caso $k = 2$, $p = 6$, $n = 55$. Así:

$$\begin{aligned}\text{tr } \mathbf{S} &= l_1 + l_2 + \dots + l_6 = 5.794, & \text{tr } \mathbf{S}^2 &= l_1^2 + l_2^2 + \dots + l_6^2 = 15.602, \\ \hat{\psi} &= \frac{l_1 + l_2}{\text{tr } \mathbf{S}} = \frac{3.8319 + 0.6989}{5.794} = 0.782, & \hat{\alpha} &= \frac{l_1^2 + l_2^2}{\text{tr } \mathbf{S}^2} = \frac{14.683 + 0.488}{15.602} = 0.972, \\ \hat{\tau}^2 &= \frac{2 \text{tr } \mathbf{S}^2}{(n-1)(\text{tr } \mathbf{S})^2} (\hat{\psi}^2 - 2\hat{\alpha}\hat{\psi} + \hat{\alpha}^2) = 0.001, & \hat{\tau} &= \sqrt{0.001} = 0.025.\end{aligned}$$

Entonces el I.C. aproximado al 95 % está dado por,

$$\text{I.C.} = \hat{\psi} \pm Z_{0.05/2} \hat{\tau} = 0.782 \pm (1.96)(0.025) = (0.733, 0.831)$$

lo que puede interpretarse como sigue: la varianza explicada por las primeras dos componentes principales se encuentra entre 73 % y 83 % del total con un 95 % de confianza.

4.5.7.2 La hipótesis sobre variación isotrópica (*ibid.*, p. 235)

Una hipótesis más conveniente es preguntarse si los últimos $(p - k)$ valores característicos de Σ son iguales, conocida como la hipótesis de variación isotrópica. De no rechazarse H_0 , querría decir que la variación es igual en todas las direcciones del espacio $(p - k)$ -dimensional generado por los últimos $(p - k)$ vectores característicos, e implicaría que si alguna de las $(p - k)$ componentes se elimina entonces deberían eliminarse todas las $(p - k)$ componentes (es bueno recordar que por construcción, las componentes principales están ordenadas de mayor a menor).

Debe tenerse en cuenta que cualquier valor característico que no es cero, por muy pequeño que sea, es “significativamente” diferente de cero, ya que si $\lambda_p = 0$, entonces $l_p = 0$ con probabilidad uno; en cuyo caso no tendría sentido probar si $\lambda_p = 0$, puesto

que se observaría de forma directa en los datos. Sin embargo, sí tiene sentido probar si $\lambda_p = \lambda_{p-1}$; sería probar la hipótesis de que la dispersión es parcialmente isotrópica en un subespacio bidimensional.

De forma más general, se puede probar si $\lambda_p = \lambda_{p-1} = \dots = \lambda_{k+1}$. La prueba de variación isotrópica que se plantea como:

$$H_0: \lambda_{k+1} = \dots = \lambda_p \text{ vs. } H_1: \text{Al menos una } \lambda_i \text{ es diferente } \forall i = k+1, \dots, p, \quad (4.5-27)$$

es útil para determinar cuántas componentes principales se pueden usar para describir los datos.

Una vez que ya se ha decidido incluir al menos k componentes, suponga que se plantea si se incluye o no alguna otra. Si no se rechaza H_0 en (4.5-27), implica que si se incluyen más de k componentes principales, se deberán incluir todas las p que existen, puesto que cada una de las restantes $(p - k)$ componentes contienen, estadísticamente, la misma cantidad de información.

Esta prueba de hipótesis puede comenzarse con $k = 0$, es decir, probar si todas las λ_i desde $i = 1, 2, \dots, p$ son iguales; e ir aumentando k hasta que la hipótesis nula no se rechace. El estadístico de prueba (EP) por el criterio de razón de verosimilitud es:

$$-2 \ln \lambda = np(a - 1 - \log g) \quad (4.5-28)$$

donde a y g , son la media aritmética y geométrica, respectivamente, de los valores característicos de $\tilde{\Sigma}^{-1}S$ y $\tilde{\Sigma}$ es el estimador por máxima verosimilitud de Σ bajo la hipótesis nula. Del desarrollo que tiene Mardia (1989; p. 235) de esta prueba, se tiene $a = 1$ y $g = (g_0/a_0)^{(p-k)/p}$, donde $g_0 = (l_{k+1} \times \dots \times l_p)^{1/(p-k)}$ es la media geométrica de los estimadores muestrales de los valores característicos repetidos (de los que se prueban si son iguales) y $a_0 = (l_{k+1} + \dots + l_p)/(p - k)$ es la media aritmética respectiva.

Observe que los valores característicos de $\tilde{\Sigma}^{-1}S$ son $(\underbrace{1,1,\dots,1}_k, \underbrace{\frac{l_{k+1}}{a_0}, \dots, \frac{l_p}{a_0}}_{p-k})$, de ahí que

$$a = \frac{k + \frac{(p-k)a_0}{a_0}}{p} = 1.$$

Finalmente, sustituyendo los valores de a y g en (4.5-28) queda el EP como:

$$-2 \ln \lambda = n(p-k) \ln \left(\frac{a_0}{g_0} \right). \quad (4.5-29)$$

En esta hipótesis el EP se distribuye aproximadamente como una χ^2_h donde $h = \frac{1}{2}(p-k+2)(p-k-1)$, son los grados de libertad de H_1 menos los grados de libertad de H_0 .

Una mejor aproximación a la ji-cuadrada es la aproximación de Bartlett que se obtiene sustituyendo n por $n' = n - \frac{2p+11}{6}$ en (4.5-29):

$$n' (p-k) \ln \left(\frac{a_0}{g_0} \right) \sim \chi^2_{(p-k+2)(p-k-1)/2} \quad (4.5-30)$$

La relación a_0/g_0 resulta igual ya sea si se obtiene con los valores característicos de S o con los de S_u . El criterio de decisión para una prueba aproximada de tamaño α es el siguiente: rechazar H_0 si EP en (4.5-30) es mayor o igual que $\chi^2_{(p-k+2)(p-k-1)/2; \alpha}$ y no rechazar H_0 en caso contrario.

EJEMPLO 4.5.6 Con los datos del ejemplo que se sigue, se hacen los cálculos para probar si los 6 últimos valores característicos de Σ son iguales. Se han obtenido que los valores característicos de S son $l_1 = 3.832$, $l_2 = 0.699$, $l_3 = 0.456$, $l_4 = 0.320$, $l_5 = 0.270$, $l_6 = 0.216$. Los valores de a_0 y g_0 se obtienen con:

$$a_0 = \frac{(l_{k+1} + \dots + l_p)}{p-k} = \frac{3.832 + 0.699 + 0.456 + 0.320 + 0.270 + 0.216}{6-0} = 0.966,$$

$$g_0 = (l_{k+1} \times \dots \times l_p)^{1/p-k} = (3.832 \times 0.699 \times 0.456 \times 0.320 \times 0.270 \times 0.216)^{1/6} = 0.532;$$

para usar la aproximación de Bartlett dada en (4.5-30), se requiere también

$$n' = n - \frac{2p+11}{6} = 55 - \frac{2(6)+11}{6} = 51.167. \text{ Entonces, el estadístico de prueba (EP) es:}$$

$$EP = n'(p-k) \ln\left(\frac{a_0}{g_0}\right) = 51.167(6) \ln\left(\frac{0.966}{0.532}\right) = 183.133$$

que se debe comparar con una χ_h^2 donde $h = \frac{1}{2}(p-k+2)(p-k-1) = 20$, teniendo que a un nivel de significancia de 0.05, $\chi_{20; 0.05}^2 = 31.41$ y a 0.01, $\chi_{20; 0.01}^2 = 35.57$. Como $EP = 183.33 > \chi_{20; 0.01}^2 = 35.57$ entonces se rechaza H_0 con un nivel de significancia de $\alpha = 0.01$ y se puede concluir que, al menos, alguno de los 6 valores característicos de Σ es diferente a los demás.

EJEMPLO 4.5.7 Como se rechazó la prueba de isotropía para todos los valores característicos de Σ , convendría continuar indagando si los últimos cinco valores característicos son iguales (i.e. ahora $k=1$); haciendo todos los pasos como en el Ejemplo 4.5.6, se obtiene que $EP = 22.498 < \chi_{14,0.05}^2 = 23.68$ por lo que no se rechaza la hipótesis nula y se concluye que los últimos cinco valores característicos de Σ son iguales. ¿Cómo se interpreta este resultado? El no rechazo de la hipótesis nula sugiere que si se incluye la segunda componente habría que incluir a todas las restantes, puesto que contienen la misma cantidad de información. Sin embargo, en el caso de este ejemplo hay otros aspectos importantes que considerar: la primera componente principal explica sólo el 66.14 % de la variación total de los datos originales (lo cual es muy poco), en tanto que juntando las primeras dos componentes se explica el 78.21 % (véase también el I.C. que se construyó para las primeras dos componentes en el Ejemplo 4.5.5); comparando los valores del EP y χ_h^2 , se puede admitir que el criterio para no rechazar la hipótesis nula no es tan “contundente” (una diferencia del 5 %), ya que si se hace la prueba de isotropía para los últimos 4 valores característicos poblacionales, el valor de la χ_h^2 es más que el doble que su correspondiente EP ; así también, si se observa la gráfica

de los porcentajes de la varianza total que explican los valores característicos (la gráfica de puntos) y la estructura de correlación muestral, parece razonable no desechar la segunda componente principal. Por estas razones, se pueden tomar las primeras dos componentes para estudiar la variabilidad de la información contenida en los datos originales.

4.5.7.3 Hipótesis que involucran a la matriz de correlación

En ocasiones las variables originales sobre las que se desea generar las componentes principales son inconmensurables, es decir, están en unidades de medición diferentes y no son comparables. En estos casos se sugiere el uso de la matriz de correlaciones en lugar de la matriz de covarianzas en la construcción de las componentes principales, con la finalidad de estandarizar las variables y hacerlas conmensurables. Las pruebas de hipótesis sobre la matriz de correlaciones poblacional son las que se presentan enseguida.

4.5.7.3.1 La hipótesis de que todos los eigenvalores de la matriz poblacional P de correlaciones son iguales

De acuerdo a Mardia *et al.* (1989) esta prueba fue propuesta por Box en 1949. Cuando las variables están estandarizadas, la matriz de correlación muestral R puede considerarse como un estimador de la matriz de correlación poblacional P . Esta hipótesis es equivalente a:

$$H_0: P = I \text{ vs. } H_1: P \neq I. \quad (4.5-31)$$

Para probarla se emplea el siguiente estadístico de prueba:

$$EP = -n' \ln |R|, \quad (4.5-32)$$

donde $n' = n - (2p + 11)/6$, que tiene una distribución asintótica ji-cuadrada con $\frac{1}{2}p(p+1) - p = \frac{1}{2}p(p-1) = h$ grados de libertad. La regla de decisión para una prueba aproximada de tamaño α es: rechazar H_0 si EP en (4.5-32) es mayor o igual que $\chi^2_{h;\alpha}$; no rechazar H_0 cuando EP sea menor que $\chi^2_{h;\alpha}$.

4.5.7.3.2 La hipótesis que los últimos $(p - k)$ valores característicos de P son iguales, donde $0 < k < p - 1$

Mardia *et al.* (1989) presentan la prueba sugerida por Bartlett en 1951 en donde se usa:

$$EP = (n - 1)(p - k) \ln \left(\frac{a_0}{g_0} \right), \quad (4.5-33)$$

donde a_0 y g_0 son las medias aritmética y geométrica de los $(p - k)$ eigenvalores más pequeños de R . Pero se tiene el inconveniente de que este EP en (4.5-33) no tiene una distribución asintótica ji-cuadrada; sin embargo, si las primeras k componentes principales reúnen claramente una alta proporción de la varianza, o si se está consciente de que se va a realizar una prueba conservadora (es decir, que no rechaza H_0 en varios de los casos en los que debería hacerlo), se puede considerar que el EP se distribuye como un ji-cuadrada con $\frac{1}{2}(p - k + 2)(p - k - 1)$ grados de libertad. Por supuesto que la regla de decisión para una prueba (muy) aproximada de tamaño α es: rechazar H_0 si el EP en (4.5-33) es mayor que ó igual a $\chi^2_{0.5(p-k+2)(p-k-1)}$; no rechazar H_0 de otro modo.

4.5.8 El Uso de las Componentes Principales en la Regresión Múltiple
(Mardia *et al.*, 1989; 244-245)

Se dice que cuando las variables independientes en un modelo de regresión múltiple son altamente dependientes —esto es, que se explican en gran medida entre sí— se tiene el problema de que los estimadores de los coeficientes de regresión serán muy imprecisos; en este caso, recomiendan desechar alguna de las variables independientes con la finalidad de incrementar la estabilidad de los coeficientes de regresión.

También se puede reducir la dimensionalidad usando componentes principales (Massy, 1966 citado por Mardia *et al.* 1989). Las componentes principales se usan para eliminar multicolinealidad en las variables independientes. Hay que tener en cuenta que el criterio de selección de las componentes principales en un contexto de regresión es, en cierta

medida, diferente a como se hizo en el ejemplo de las calificaciones. En ese caso, convenía seleccionar a las componentes principales con las varianzas más grandes que explicaran al máximo, la variación total de la matriz de los datos originales. En cambio, en el escenario de la regresión múltiple, como el objetivo es que un grupo de variables independientes expliquen el comportamiento de la variable dependiente, es más adecuado seleccionar todas aquellas componentes principales que tengan las mayores correlaciones con la variable dependiente. En la regresión multivariada se examinan las correlaciones de las componentes con cada una de las variables dependientes. Muchas veces, las componentes con varianzas mayores coinciden en explicar mejor a la variable dependiente, lo que facilita la selección de componentes para el análisis de regresión.

Las componentes principales pueden o no tener un significado intuitivo claro; cuando sí lo tienen, posiblemente sea mejor obtener la ecuación de regresión en términos de la componentes, mientras que cuando no resulta fácil interpretar a las componentes, se sugiere volver a transformar las componentes a las variables originales.

Considere que se toman n observaciones, sobre una variable dependiente y q variables independientes, las cuales se centran de forma que $\bar{y} = \bar{x}_i = 0, i = 1, \dots, q$; entonces la ecuación de regresión es:

$$y = X\beta + \varepsilon \quad \text{donde } \varepsilon \sim N_n(0, \sigma^2 H) \quad \text{y} \quad H = I - \frac{1}{n} 11'. \quad (4.5-34)$$

Note que y es de orden $n \times 1$ y X de dimensión $n \times q$. Si se denota la transformación componentes principales como $W = XG$ (ya que las observaciones están centradas), entonces la ecuación de regresión puede escribirse como:

$$y = W\alpha + \varepsilon, \quad (4.5-35)$$

donde $\alpha = G'\beta$. El hecho de que las columnas de W son ortogonales, es lo que hace que los estimadores de mínimos cuadrados $\hat{\alpha}_i$ y su distribución no se alteren si alguna de las columnas de W no se considera en la regresión. El estimador de α_i obtenido por mínimos cuadrados es:

$$\hat{\alpha}_i = \frac{1}{nl_i} \mathbf{w}_{(i)}' \mathbf{y}, \quad i = 1, 2, \dots, p, \quad (4.5-36)$$

donde l_i es el i -ésimo valor característico de la matriz de covarianzas, $\frac{1}{n} \mathbf{X}'\mathbf{X}$ (por la estandarización convenida). Todas las $\hat{\alpha}_i$ se pueden disponer en un vector $\hat{\alpha}$ usando:

$$\hat{\alpha} = (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'\mathbf{y}$$

En la ecuación de regresión escrita en términos de las componentes, $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{W}\hat{\alpha} = (\mathbf{I} - \mathbf{P}_\mathbf{W})\mathbf{y}$, donde $\mathbf{P}_\mathbf{W} = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'$. Otros resultados importantes son: $E(\hat{\alpha}_i) = \alpha_i$, $\text{var}(\hat{\alpha}_i) = (nl_i)^{-1} \sigma^2$; y la covarianza entre la componente $\mathbf{w}_i$ y $\mathbf{y}$ es $\text{cov}(\mathbf{w}_i, \mathbf{y}) = n^{-1} \mathbf{w}_{(i)}' \mathbf{y} = n^{-1} \mathbf{g}_{(i)}' \mathbf{X}'\mathbf{y}$.

Trabajando con componentes, se puede probar si la contribución de $\mathbf{w}_i$ es o no significativa en la regresión, plantéandolo como:

$$H_0: \alpha_i = 0 \quad \text{versus} \quad H_a: \alpha_i \neq 0, \quad i = 1, \dots, q, \quad (4.5-37)$$

La hipótesis se prueba usando el estadístico de prueba siguiente:

$$t_{cal} = \frac{\hat{\alpha}_i (nl_i)^{1/2}}{(\hat{\mathbf{e}}'\hat{\mathbf{e}}/(n-q-1))^{1/2}} = \frac{\mathbf{w}_{(i)}' \mathbf{y}}{(nl_i \hat{\mathbf{e}}'\hat{\mathbf{e}}/(n-q-1))^{1/2}} \sim t_{n-q-1; \alpha/2}. \quad (4.5-38)$$

La regla de decisión con un nivel de significancia α es: rechazar H_0 si $|t_{cal}| > t_{n-q-1; \alpha/2}$; no rechazar de otro modo.

El uso de componentes principales no es sugerible cuando las variables consideradas en el modelo sean inconmensurables, por la sensibilidad de este método a los cambios de escala. Como solución a este problema, Weisberg (1980, citado por Pla, 1988) sugiere particionar el grupo de variables en grupos más homogéneos y realizar el análisis de componentes principales dentro de cada grupo, y después, ajustar el modelo de regresión en función de los grupos de componentes principales. Haciendo este tipo de ajuste no se eliminará la multicolinealidad por completo, pero no habrá correlación dentro de grupos.

EJEMPLO 4.5.8 Ejemplo del uso de componentes principales en la regresión lineal. En el Cuadro 4.5.3 se tienen las observaciones de 20 árboles de eucalipto con las variables altura y diámetro (en m y cm, respectivamente). Para cada árbol, que corresponden a diferentes sitios, se cuenta con las cantidades de nueve nutrientes (N, P, K, Ca, Mg, Fe, Cu, Zn, Mn) en partes por millón de una muestra de suelo.

Cuadro 4.5.3 Variables dasométricas de 20 eucaliptos y nutrientes de las muestras de suelo correspondientes

árbol	alt.	diám	N	P	K	Ca	Mg	Fe	Cu	Zn	Mn
1	6.90	6.7	4.9	0.75	342	2598	875.6	16.27	0.230	1.53	17.69
2	8.50	9.7	3.5	3.55	394	2075	969.5	12.65	0.008	0.67	5.66
3	10.40	14.7	3.5	14.86	338	2468	934.3	0.49	0.410	0.49	2.07
4	10.80	17.0	4.9	0.15	330	2862	651.7	7.32	0.160	1.15	2.27
5	13.00	15.8	4.2	0.78	254	1974	679.4	1.60	0.270	0.36	17.7
6	13.20	19.1	4.2	0.08	410	1832	111.7	1.15	0.410	0.66	3.69
7	14.10	22.3	4.9	0.06	598	1951	790.1	3.57	0.270	0.63	4.82
8	9.90	9.2	3.2	3.75	661	2126	759.3	4.67	0.170	0.58	6.89
9	13.02	29.2	3.5	3.22	562	2544	946.2	9.24	0.100	0.82	14.31
10	13.27	13.1	3.0	5.78	539	2949	808.0	1.85	0.360	0.15	10.24
11	12.01	19.8	4.5	6.54	540	2434	921.9	4.36	0.120	0.91	14.64
12	9.19	8.1	3.2	3.90	245	2298	838.2	14.20	0.430	0.83	3.72
13	14.11	10.9	4.6	5.05	293	2243	711.2	1.45	0.050	0.80	15.24
14	9.46	15.9	3.6	13.50	497	1934	1085	11.75	0.320	0.43	1.53
15	12.39	13.7	4.3	4.77	508	1947	864.0	5.06	0.150	0.55	16.23
16	16.11	19.6	4.1	6.46	427	2255	407.7	0.36	0.380	0.65	11.27
17	8.33	13.6	3.2	6.36	441	2054	957.7	9.00	0.340	0.76	10.77
18	16.70	14.2	3.9	6.63	376	2154	591.1	10.91	0.080	0.18	10.14
19	11.41	21.4	2.8	4.50	198	1960	672.3	7.71	0.500	0.63	0.26
20	18.45	20.1	3.1	2.32	372	2686	359.7	11.54	0.260	0.55	8.60

Fuente: Datos adaptados y usados por cortesía del Biól. F. Vargas López.

Con los datos de los nueve nutrientes del Cuadro 4.5.3 se construyeron las componentes principales que se reportan en el Cuadro 4.5.4.

Cuadro 4.5.4 Componentes principales generadas a partir de los nueve nutrientes de los 20 sitios del Cuadro 4.5.3

obser	Componente principal								
	1	2	3	4	5	6	7	8	9
1	339.53	87.67	-92.87	8.63	8.31	0.98	0.86	0.27	0.12
2	-173.71	231.02	-59.42	-2.48	5.10	-1.67	-0.27	-0.14	-0.22
3	214.61	155.28	-106.96	-8.39	-11.50	2.14	0.27	0.08	0.00
4	584.36	-155.97	-66.76	-6.11	1.33	-5.23	1.04	-0.07	-0.08
5	-299.09	-68.95	-150.27	11.42	-4.15	-2.96	-0.50	-0.27	0.08

obser	Componente principal								
	1	2	3	4	5	6	7	8	9
6	-485.26	-590.00	95.89	-3.21	0.11	-0.97	0.27	0.17	0.02
7	-310.06	98.05	170.73	-2.77	1.11	-5.98	0.70	-0.20	0.13
8	-137.58	63.89	238.72	-2.68	0.95	-1.16	-0.65	0.18	-0.12
9	293.33	197.47	112.38	4.16	2.81	-0.33	-0.51	0.11	-0.07
10	685.56	24.86	114.29	-0.76	-5.10	-1.07	-1.04	-0.25	0.13
11	181.52	178.92	94.20	4.42	-2.96	0.79	0.54	0.15	-0.05
12	36.61	59.41	-183.87	-4.36	5.26	-0.62	-0.39	0.07	0.10
13	-28.07	-53.16	-115.78	7.23	-6.14	-0.31	0.41	0.02	-0.19
14	-304.00	372.96	22.57	-9.07	0.94	4.81	0.57	-0.12	0.08
15	-308.82	156.36	69.85	7.50	-0.99	0.69	0.15	-0.09	0.02
16	-39.53	-330.57	66.17	1.89	-5.17	3.11	0.27	0.21	0.05
17	-195.18	228.83	-11.17	1.75	0.96	1.00	-0.59	0.27	0.04
18	-125.77	-150.34	-14.78	0.64	3.31	5.26	0.22	-0.49	-0.06
19	-314.13	-83.93	-204.57	-6.53	0.04	-1.98	-0.81	0.18	0.02
20	385.68	-421.79	21.66	-1.27	5.85	3.49	-0.55	-0.08	-0.02

Cada componente principal explica los siguientes porcentajes de la variación total contenida en los datos originales.

Comp. princ.	%	% acumulado
1	60.03	60.03
2	31.75	91.78
3	8.19	99.97
4	0.02	99.99
5	0.01	100.00
6	0.00	100.00
7	0.00	100.00
8	0.00	100.00
9	0.00	100.00

Con las primeras tres componentes principales se hace una regresión lineal (múltiple) usando como variable dependiente a la altura. La intención de esto, es explorar la posible relación que existe entre estas componentes principales, construidas a partir de los datos originales de las muestras de suelo, y la respuesta en altura de los árboles. Los resultados obtenidos con la computadora son los siguientes:

Model: MODEL1. Dependent Variable: ALT

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	78.39428	26.13143	4.592	0.0168
Error	16	91.05449	5.69091		
C Total	19	169.44877			

Root MSE	2.38556	R-square	0.4626
Dep Mean	12.06250	Adj R-sq	0.3619
C.V.	19.77668		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	12.062500	0.53342787	22.613	0.0001
PRIN1	1	0.000431	0.00166421	0.259	0.7990
PRIN2	1	-0.007848	0.00228837	-3.430	0.0034
PRIN3	1	0.006288	0.00450694	1.395	0.1820

Puede observarse que el segundo coeficiente de regresión $\hat{\alpha}_2 = -0.00785$ resulta significativamente diferente de cero (nivel de significancia de 0.01), mientras que el primero y el tercero no son significativamente diferentes de cero. Es notorio también que el tercer coeficiente de regresión $\hat{\alpha}_3 = 0.00629$ estuvo "más cerca" de considerarse diferente de cero con respecto al primer coeficiente $\hat{\alpha}_1 = 0.000431$. Los cálculos de este ejemplo se hicieron con el Programa 4.5.9.2.

4.5.9 Anexo de Programas

Programa 4.5.9.1 Si los datos del Cuadro 4.5.1 están en el archivo externo MATEDAT.TXT, los cálculos para los Ejemplos 4.5.1 y 4.5.2 se pueden hacer con este programa.

```
DATA CALIFMAT;
INFILE 'C:\COMPRIN\MATEDAT.TXT';
INPUT AL SEX GEO ARIT HIS LECT COM PSI;
/* AL:no.alumno SEX GEO:geometria ARIT:aritmética HIS:historia LECT:lectura y redaccion
COM:comunicación educativa PSI:psicología educativa */
PROC IML WORKSIZE=80; USE CALIFMAT;
READ ALL VAR {GEO ARIT HIS LECT COM PSI} INTO X;
START XIMBAL (X,prom,mp1,mp2,VM,S,DIAGS,G,L,VARITOT,PROP, Y, GG, rijCUA,
rij);
n=NROW(X);p1={ .25,.25,.125,.125,.125,.125 };p2={ .1,.1,.2,.2,.2,.2 };
PROM=X*(1/6)*J(6,1,1); /*promedios de los estudiantes */
MP1=X*p1; /*media ponderada 1*/
MP2=X*p2; /*media ponderada 2*/
VM=(1/n)*X'*J(n,1,1); /*VM es el vector que contiene las medias de las variables */
H=I(n)-(1/n)*J(n,n,1); /* H es la matriz centrante */
S=(1/n)*X'*H*X; /* S es la matriz de varianzas y covarianzas muestral */
/*Cuando la capacidad de la computadora no sea la requerida para el
paso anterior, se puede usar S=(1/n)*(X'*X-n*VM*VM'); */
```

```

G=EIGVEC(S);      /*G es la matriz que contiene los eigenvectores de la matriz S*/
L=G`*S*G; L=DIAG(L);
                  /*L es una matriz diagonal que contiene los eigenvalores de la matriz S*/
                  /* cada valor de la diagonal de L expresa la varianza de la componente
                     correspondiente, L=G`SG */
VARITOT=TRACE(L);  /*Contiene la variacion total en S, =tr(L)=tr(S) */
PROP=(1/VARITOT)*L; /*Calcula el peso de cada componente principal en la variacion total */
Y=(X-J(N,1,1)*VM`)*G; /* Y es la matriz que contiene la transformacion comprin,
                       de las variables originales */
GG=G#G; DIAGS=DIAG(1/S); /* DIAGS es una matriz diagonal que
                           contiene los reciprocos de los elementos
                           de la diagonal de S */
rij=DIAGS##0.5*G*L##0.5; /*es una matriz en donde cada rij representa la correlacion simple
                           entre la variable original xi y la componente principal yj */
rijCUA=DIAGS*GG*L; /* es una matriz en donde cada rijCUA representa la proporcion de la
                     variacion de xi que explica la componente yj */
FINISH;
RUN XIMBAL(X,prom,mp1,mp2,VM,S,DIAGS,G,L,VARITOT, PROP,Y, GG, rijCUA, rij);
PRINT X prom mp1 mp2 VM S DIAGS G GG L VARITOT PROP Y rijCUA rij;
QUIT;

```

Programa 4.5.9.2 Programa que hace los cálculos del Ejemplo 4.5.8.

```

DATA MINERV;
INFILE 'C:/COMPRIN/PANCHO1.DAT';
INPUT ALT DIAM N P K CA MG FE CU ZN MN;
DATA B;SET MINERV;
PROC DATASETS;DELETE MINERV;
PROC PRINCOMP DATA=B COV OUT=D;
VAR N P K CA MG FE CU ZN MN;
PROC DATASETS;DELETE B;
PROC REG;
MODEL ALT=PRIN1 PRIN2 PRIN3;
OUTPUT OUT=A P=PHAT R=RES;
PROC DATASETS;DETETE D;
DATA C;SET A;
PROC DATASETS;DELETE A;
PROC PRINT;
RUN;

```

4.6 MÉTODO DE CORRELACIONES CANÓNICAS

4.6.1 Introducción

Este método tiene por objeto evaluar el grado de relación entre dos conjuntos de variables. En el análisis de correlaciones canónicas, una colección de variables es particionada en dos grupos: un conjunto $\mathbf{x}$ y un conjunto $\mathbf{y}$, presumiendo que hay algún tipo de relación entre las variables $\mathbf{x}$ y las variables $\mathbf{y}$; lo que se trata es de encontrar dos combinaciones lineales, $\eta = \mathbf{a}'\mathbf{x}$ y $\phi = \mathbf{b}'\mathbf{y}$, tal que η y ϕ tengan la mayor correlación posible. Observe que η es la combinación lineal sobre las variables $X_1, \dots, X_q$ contenidas en el vector $\mathbf{x}$ y el vector que contiene los coeficientes de la combinación lineal es $\mathbf{a}$; similarmente, ϕ es la combinación lineal sobre las variables $Y_1, \dots, Y_p$ arregladas en el vector $\mathbf{y}$, y $\mathbf{b}$ es el vector que contiene sus respectivos coeficientes (Harris, 1985; Mardia *et al.*, 1989). Esta técnica fue introducida por Harold Hotelling en 1936 para analizar la estructura de la relación entre dos conjuntos de variables continuas y en los años sesenta Kendall & Stuart presentaron una adaptación del análisis de correlaciones canónicas para tablas de contingencia de doble entrada (Srikantan, 1970; Morrison, 1981).

Existe cierta relación entre el método de componentes principales y el de correlaciones canónicas; sin embargo, el primero estudia las relaciones *dentro* un conjunto de variables y las correlaciones canónicas centran su atención en las posibles relaciones *entre* dos grupos de variables, las $\mathbf{x}$ y las $\mathbf{y}$ (Mardia *et al.*, 1989).

El análisis de correlaciones canónicas puede visualizarse como una generalización de la regresión múltiple. En la regresión múltiple hay un grupo de q variables $\mathbf{x}$ y un grupo

de $p = 1$ variable y . En la solución de la regresión se encuentra la combinación lineal $\mathbf{a}'\mathbf{x}$ que tenga la máxima correlación con la variable dependiente y (*ibid.*).

En el análisis de correlaciones canónicas, el conjunto $\mathbf{y}$ contiene $p \geq 1$ variables y se buscan dos vectores $\mathbf{a}$ y $\mathbf{b}$ tal que $\text{corr}(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y})$ sea máxima. En este análisis no existe el supuesto de asimetría causal —como sucede en la regresión donde, explícitamente, cierta(s) variable(s), la(s) dependiente(s), está(n) en función de otras, llamadas predictoras— sino que $\mathbf{x}$ y $\mathbf{y}$ se toman simétricamente en el análisis estadístico; sin embargo, si se considerara que las variables $\mathbf{x}$ son las que “explican” a las variables $\mathbf{y}$, entonces puede decirse que la combinación lineal $\eta = \mathbf{a}'\mathbf{x}$ es el “mejor predictor” y que la combinación lineal $\phi = \mathbf{b}'\mathbf{y}$ es “el criterio más predecible”, bajo el entendido de que ambas combinaciones lineales tienen la mayor correlación posible. Puede observarse que la cuestión es enteramente análoga, si se supone que las variables $\mathbf{x}$ están en función de las variables $\mathbf{y}$ (*ibid.*).

EJEMPLO 4.6.1 (Stiteler, 1979). Es usual en la dasonomía describir la *calidad de sitio*¹ para algún propósito, tal como la producción maderable, usando un número conocido como *índice de sitio*². Un estudio de la relación entre varias características ambientales y la calidad de sitio puede involucrar el coeficiente de correlación múltiple entre este índice de sitio y las variables ambientales. Podría, sin embargo, ser más razonable describir la calidad de sitio a través de varias variables (clima, geología, tipo de suelo, topografía y vegetación actual) que necesitan una técnica para correlacionar dos grupos de variables. ♦

¹ En el contexto del manejo forestal, la calidad de sitio (o calidad de estación) puede definirse como la producción maderable potencial de un sitio para una especie o tipo de bosque particular. La calidad de sitio sólo tiene sentido con respecto a una o más especies que podrían considerarse para el manejo en una ubicación particular (Clutter *et al.*, 1983; p. 31).

² Es un criterio tradicional a través del cual se ha clasificado al suelo forestal, bajo el supuesto que para una especie dada, los sitios más fértiles producirán, a una determinada edad base, árboles de mayores dimensiones que los sitios con menor fertilidad (Mendoza, 1993; p. 37).

EJEMPLO 4.6.2 Con los datos del Cuadro 4.6.1 suponer que X_1 = diámetro normal, X_2 = altura a la base de la copa (altura de fuste limpio), Y_1 = radio de la base de la copa y Y_2 = altura total. Se puede explorar la relación que existe entre el grupo de variables $\mathbf{x} = (X_1, X_2)'$ y las variables $\mathbf{y} = (Y_1, Y_2)'$ con un análisis de correlaciones canónicas. Véase el Ejemplo 4.6.3. ♦

Cuadro 4.6.1 Mediciones de 4 dimensiones en 31 coníferas

	Diám. en <i>cm</i>	Altura total en <i>m</i>	altura a la ba- se de la copa (hbc) en <i>m</i>	Radio de la base de la copa en <i>cm</i>
1	86.1	32.7	9	279.9
2	97.3	33.8	9.6	682.5
3	47.8	27.7	2.8	261.6
4	40.9	19.5	9.1	139.5
5	80.3	35.1	12.4	279.9
6	49.5	24.9	7.3	351.8
7	75.2	33.7	11.7	340.9
8	63.5	30.8	7.2	239.8
9	35.6	12.0	1.6	193.3
10	27.4	10.7	1.1	166.6
11	64.8	33.7	13.2	239.0
12	76.5	38.6	13.7	299.5
13	68.3	31.1	11.2	400.3
14	53.3	22.1	3.1	181.6
15	46.2	16.2	4.1	117.8
16	30.5	11.2	0.0	211.3
17	67.6	26.8	12.3	147.9
18	69.9	4.4	2.4	244.1
19	51.3	21.1	8.2	332.5
20	38.6	20.8	6.1	336.1
21	71.6	34.1	9.8	151.4
22	56.1	28.7	11.0	263.6
23	55.9	29.1	6.4	308.1
24	34.0	13.0	0.4	136.2
25	28.2	13.4	0.6	168.4
26	29.5	14.4	0.5	207.8
27	79.0	32.1	6.2	358.7
28	36.1	21.3	3.1	258.6
29	45.7	34.7	5.1	110.0
30	70.4	34.8	10.0	287.3
31	75.4	36.8	9.3	251.9

Fuente: Tomado de Biging & Gill (1997, p. 27)

4.6.2 Correlaciones Canónicas Poblacionales

(Mardia *et al.*, 1989; pp. 282-287)

Suponga que $\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$ es un vector aleatorio de orden $(q + p) \times 1$ que se puede particionar en un vector aleatorio $\mathbf{x}$ de orden $q \times 1$ y un vector aleatorio $\mathbf{y}$ de dimensión $p \times 1$. Suponer que $\mathbf{x}$ tiene vector de medias μ_1 y matriz de covarianzas Σ_{11} ; que $\mathbf{y}$ tiene vector de medias μ_2 y matriz de covarianzas Σ_{22} . Esto es:

$$\mathbf{x} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_q \end{bmatrix} \sim (\mu_1, \Sigma_{11}) \quad \text{y} \quad \mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{bmatrix} \sim (\mu_2, \Sigma_{22}).$$

Suponer también que $\text{Cov}(\mathbf{x}, \mathbf{y}) = E\{(\mathbf{x} - \mu_1)(\mathbf{y} - \mu_2)'\} = \Sigma_{12} = \Sigma_{21}'$. Considerando las combinaciones lineales $\eta = \mathbf{a}'\mathbf{x}$ y $\phi = \mathbf{b}'\mathbf{y}$, la correlación entre η y ϕ es:

$$\rho(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y}) = \frac{\mathbf{a}'\Sigma_{12}\mathbf{b}}{(\mathbf{a}'\Sigma_{11}\mathbf{a} \mathbf{b}'\Sigma_{22}\mathbf{b})^{1/2}}, \quad (4.6-1)$$

Ahora el objetivo es encontrar los valores de $\mathbf{a}$ y $\mathbf{b}$, tal que la correlación $\rho(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y})$ sea máxima. Los valores que tome la $\rho(\mathbf{a}'\mathbf{x}, \mathbf{b}'\mathbf{y})$ dependen de los vectores $\mathbf{a}$ y $\mathbf{b}$.

La presentación de los resultados, se facilita conviniendo en la notación que viene enseguida. Suponer que Σ_{11} y Σ_{22} son matrices no-singulares (es decir, que tienen determinante diferente de cero y entonces inversa única) y sea:

$$\mathbf{K} = \Sigma_{11}^{-1/2} \Sigma_{12} \Sigma_{22}^{-1/2}. \quad (4.6-2)$$

Hacer también:

$$\mathbf{N}_1 = \mathbf{K}\mathbf{K}', \quad \mathbf{N}_2 = \mathbf{K}'\mathbf{K}, \quad (4.6-3)$$

y

$$\begin{aligned} \mathbf{M}_1 &= \Sigma_{11}^{-1/2} \mathbf{N}_1 \Sigma_{11}^{1/2} = \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}, \\ \mathbf{M}_2 &= \Sigma_{22}^{-1/2} \mathbf{N}_2 \Sigma_{22}^{1/2} = \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}. \end{aligned} \quad (4.6-4)$$

Las matrices $\mathbf{M}_1$, $\mathbf{N}_1$, $\mathbf{M}_2$ y $\mathbf{N}_2$ tienen los mismos valores característicos diferentes de cero y como $\mathbf{N}_1 = \mathbf{K}\mathbf{K}'$ es positiva semi-definida, todos sus valores característicos diferentes de cero, son positivos. Si k es el número de valores característicos diferentes de cero, entonces $k = r(\mathbf{K}) = r(\Sigma_{12})$ donde $r(\cdot)$ denota el rango de $(\cdot)$. Por facilidad, suponer que todos los k valores característicos son distintos entre sí y que están ordenados de la siguiente manera $\lambda_1 > \lambda_2 > \dots > \lambda_k > 0$.

Por el teorema de la descomposición del valor singular (ver Sección 2.4.6), puede escribirse la matriz $\mathbf{K}$ de la expresión (4.6-2) como:

$$\mathbf{K} = (\alpha_1, \dots, \alpha_k) \mathbf{D} (\beta_1, \dots, \beta_k)' \quad (4.6-5)$$

donde α_i y β_i es el i -ésimo vector característico estandarizado de $\mathbf{N}_1$ y $\mathbf{N}_2$, respectivamente, asociados a λ_i (note que los valores característicos diferentes de cero de $\mathbf{N}_1$ y $\mathbf{N}_2$ son los mismos) y

$$\mathbf{D} = \begin{bmatrix} \sqrt{\lambda_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sqrt{\lambda_k} \end{bmatrix}.$$

Como se está suponiendo que $\lambda_i \neq \lambda_j$, $\forall i \neq j$, $i, j = 1, \dots, k$, los vectores característicos son únicos excepto por el signo y éste se escogerá de tal forma que las raíces cuadradas $\sqrt{\lambda_i}$ en $\mathbf{D}$ sean positivas. También se tiene que como $\mathbf{N}_1 = \mathbf{N}_1'$ y $\mathbf{N}_2 = \mathbf{N}_2'$ (ambas matrices son simétricas), sus vectores característicos son ortogonales, por lo tanto:

$$\alpha_i' \alpha_j = \delta_{ij}, \quad \beta_i' \beta_j = \delta_{ij}, \quad (4.6-6)$$

donde δ_{ij} es la delta de Kronecker, que es igual a 1 para toda $i = j$, $i, j = 1, \dots, k$ y vale cero, de otro modo.

Definición Usando la notación anterior, se establecen:

$$\mathbf{a}_i = \Sigma_{11}^{-1/2} \alpha_i, \quad \mathbf{b}_i = \Sigma_{22}^{-1/2} \beta_i, \quad i = 1, \dots, k. \quad (4.6-7)$$

Donde,

- (a) los vectores $\mathbf{a}_i$ y $\mathbf{b}_i$ se conocen como los i -ésimos vectores de correlación canónica para $\mathbf{x}$ y $\mathbf{y}$, respectivamente;
- (b) las variables aleatorias $\eta_i = \mathbf{a}_i' \mathbf{x}$ y $\phi_i = \mathbf{b}_i' \mathbf{y}$ son denominadas las i -ésimas variables de correlación canónica;
- (c) $\rho_i = \sqrt{\lambda_i}$ se denomina como el i -ésimo coeficiente de correlación canónica.

Se tiene entonces que:

$$\begin{aligned} \text{corr}(\eta_i, \eta_j) &= \mathbf{a}_i' \Sigma_{11} \mathbf{a}_j = \alpha_i' \alpha_j = \delta_{ij} \\ \text{corr}(\phi_i, \phi_j) &= \mathbf{b}_i' \Sigma_{22} \mathbf{b}_j = \beta_i' \beta_j = \delta_{ij}. \end{aligned} \quad (4.6-8)$$

Por lo tanto, las i -ésimas variables de correlación canónica para $\mathbf{x}$ no están correlacionadas y están estandarizadas con varianza 1; lo mismo pasa con las correspondientes para $\mathbf{y}$.

Algunas propiedades de las correlaciones canónicas

Resultado 1. (Teorema 10.2.3, *ibid.*, p. 285) Sean η_i y ϕ_i las i -ésimas variables de correlación canónica, $i = 1, \dots, k$, y sean $\boldsymbol{\eta} = (\eta_1, \dots, \eta_k)'$ y $\boldsymbol{\phi} = (\phi_1, \dots, \phi_k)'$. Entonces:

$$\text{Var} \begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\phi} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \Lambda^{1/2} \\ \Lambda^{1/2} & \mathbf{I} \end{pmatrix}. \quad (4.6-9)$$

Prueba. (*ibid.*).

Resultado 2. (Teorema 10.2.4, *ibid.*, p. 285) Si $\mathbf{x}^* = \mathbf{U}'\mathbf{x} + \mathbf{u}$ y $\mathbf{y}^* = \mathbf{V}'\mathbf{y} + \mathbf{v}$, donde $\mathbf{U}_{q \times q}$ y $\mathbf{V}_{p \times p}$ son matrices no-singulares y $\mathbf{u}_{q \times 1}$ y $\mathbf{v}_{p \times 1}$ son vectores con valores fijos, entonces:

- (a) las correlaciones canónicas entre $\mathbf{x}^*$ y $\mathbf{y}^*$ son las mismas que las de $\mathbf{x}$ y $\mathbf{y}$;
- (b) los vectores de correlación canónica para $\mathbf{x}^*$ y $\mathbf{y}^*$ están dados por $\mathbf{a}_i^* = \mathbf{U}^{-1}\mathbf{a}_i$ y $\mathbf{b}_i^* = \mathbf{V}^{-1}\mathbf{b}_i$, $i = 1, \dots, k$, donde $\mathbf{a}_i$ y $\mathbf{b}_i$ son los vectores de correlaciones canónicas para $\mathbf{x}$ y $\mathbf{y}$, respectivamente.

La prueba de este resultado puede verse en la referencia citada.

El análisis de correlaciones canónicas usando la matriz de correlaciones $\mathbf{P}$ conduce esencialmente a los mismos resultados que usando la matriz Σ , propiedad de invarianza que no se cumple en las componentes principales (*ibid.*).

4.6.3 Correlaciones Canónicas Muestrales (*ibid.*, pp 287-288)

El desarrollo poblacional puede extenderse de forma directa en el análisis de datos muestrales. Lo que se tiene que hacer es substituir todas las Σ_{ij} por $\mathbf{S}_{ij}$, $i, j = 1, 2$; y usar l_i en lugar de λ_i , r_i en vez de ρ_i , etc., es decir, reemplazar todos los valores poblacionales por sus análogos muestrales. Para datos con distribución normal, $\mathbf{S}$ es el estimador de máxima verosimilitud de Σ , así que los valores de las correlaciones canónicas muestrales, en este caso, serán los estimadores de máxima verosimilitud de los valores poblacionales correspondientes, excepto cuando hayan valores característicos poblacionales repetidos.

Si $(\mathbf{X}, \mathbf{Y})$ es la matriz de datos y $\mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \mathbf{S}_{21} & \mathbf{S}_{22} \end{bmatrix}$ es el estimador de $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$,

se tiene que la correlación máxima entre las i -ésimas variables (de correlaciones)

canónicas de ambos grupos de variables (originales), escrita por directa analogía de (4.6-1), es:

$$r(\mathbf{Xa}_i, \mathbf{Yb}_i) = \frac{\mathbf{a}_i' \mathbf{S}_{12} \mathbf{b}_i}{(\mathbf{a}_i' \mathbf{S}_{11} \mathbf{a}_i \mathbf{b}_i' \mathbf{S}_{22} \mathbf{b}_i)^{1/2}}.$$

EJEMPLO 4.6.3 Haciendo el análisis descrito en el Ejemplo 4.6.2 para los datos del Cuadro 4.6.1, se obtuvieron los siguientes resultados corriendo el Programa 4.6.7.1 escrito en IML de SAS; A1 y A2 son el primer y segundo vectores de correlaciones canónicas para las variables $\mathbf{x}$; B1 y B2 son los dos vectores de correlaciones canónicas para el grupo de variables $\mathbf{y}$.

A1	A2	B1	B2
0.5067733	1.413429	0.2440686	1.0559984
0.5632894	-1.391871	0.8801917	-0.632428

La primeras variables de correlación canónica (muestral) para el h -ésimo árbol, $h = 1, \dots, 31$ se obtuvieron con $\eta_{h1} = 0.507x_{h1} + 0.563x_{h2}$, y $\phi_{h1} = 0.244y_{h1} + 0.880y_{h2}$; las segundas variables de correlación canónica para el h -ésimo árbol fueron calculadas usando $\eta_{h2} = 1.413x_{h1} - 1.392x_{h2}$ y $\phi_{h2} = 1.056y_{h1} - 0.632y_{h2}$. Los valores de η_{h1} se escribieron en la columna VCCA1 de la salida de SAS; los de η_{h2} , en VCCA2; los de ϕ_{h1} , en VCCB1; y los de ϕ_{h2} , en VCCB2.

VCCA1	VCCA2	VCCB1	VCCB2
48.702787	109.16939	97.097057	274.89358
54.716622	124.16468	196.32727	699.34288
25.800975	63.664666	88.229644	258.73095
25.852962	45.143217	51.211302	134.97944
47.678686	96.239144	99.209517	273.37575
29.197292	59.804074	107.78009	355.7528
44.699839	90.004966	112.86543	338.67706
36.235789	79.731267	85.637544	233.74965
18.942393	48.091078	57.740752	196.53537
14.505207	37.196896	50.079872	169.16236
40.274331	73.217497	87.994845	231.07082
46.485224	89.058681	107.07393	291.85983
40.921459	80.948241	125.0746	403.04767
28.757215	71.020964	63.775086	177.79267
25.722414	59.593747	43.010381	114.15129
15.456586	43.109584	61.429832	216.04928
41.186336	78.427783	59.686877	139.23311
36.77535	95.458195	63.449977	254.98654
30.616444	61.095562	99.724839	337.77526
22.997515	46.067944	100.33943	341.76658
41.805206	87.561177	66.966517	138.31238
34.626166	63.982782	89.597973	260.21051

31.933681	70.102704	100.8111	306.94947
17.455609	47.499837	44.684629	135.60543
14.628981	39.023575	52.895713	169.35561
15.231458	41.00022	63.392206	210.32952
43.527487	103.03129	115.80154	358.48571
20.040714	46.709986	81.864211	259.61049
26.032317	57.495161	57.390194	94.21459
41.309736	85.586688	100.75157	281.37987
43.4493	93.628143	93.871924	242.73267

Puede verse que las primeras variables de correlaciones canónicas, η_{h1} y ϕ_{h1} , representan la suma del ancho y la altura correspondientes; en un caso, el diámetro normal y la altura de fuste limpio y para el grupo de variables y, el radio de la copa y la altura total. Las segundas variables de correlaciones canónicas, η_{h2} y ϕ_{h2} , miden las diferencias entre diámetros y alturas; esta medida puede relacionarse de algún modo con los conceptos forestales del *coeficiente de esbeltez* y el *coeficiente de robustez* para diferenciar entre árboles “esbeltos” y árboles “robustos”. Por propiedades se tiene que

$Cov(\eta_{h1}, \eta_{h2}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ y esta situación se cumple también para las variables de

correlación canónica de y. Finalmente, los coeficientes de correlaciones canónicas indican la correlación más alta entre un grupo de variables y el otro.

	R1	R2
	0.8460664	0.3595589

Así $corr(\eta_{h1}, \phi_{h1}) = 0.846$ y $corr(\eta_{h2}, \phi_{h2}) = 0.360$. Estos valores muestrales pueden denotarse por r_1 y r_2 , respectivamente. ♦

4.6.4 La Hipótesis $\Sigma_{12} = \mathbf{O}$ (*ibid.*, pp. 288)

Considerar la hipótesis $H_0: \Sigma_{12} = \mathbf{O}$, para probar si los dos conjuntos de variables, $\mathbf{x}$ y $\mathbf{y}$, no están correlacionadas entre sí. Bajo el supuesto de normalidad, el estadístico de máxima verosimilitud para probar la hipótesis nula es:

$$\lambda^{2/n} = |\mathbf{I} - \mathbf{S}_{22}^{-1} \mathbf{S}_{21} \mathbf{S}_{11}^{-1} \mathbf{S}_{12}| = \prod_{i=1}^k (1 - r_i^2),$$

que tiene una distribución $\Lambda(p, n-1-q, q)$ de Wilks donde $r_i, i=1, \dots, k$, son los coeficientes de correlaciones canónicas muestrales con $k = \min(p, q)$. La regla de decisión para una prueba de tamaño α es: rechazar H_0 si $\lambda^{2/n} < \Lambda(p, n-1-q, q; \alpha)$ donde $\Lambda(p, n-1-q, q; \alpha)$ es el punto $100\alpha\%$ inferior de una distribución lambda de Wilks con parámetros $p, n-1-q$ y q .

Con la aproximación de Bartlett, se tiene el siguiente estadístico:

$$EP = -\left\{n - \frac{1}{2}(p+q+3)\right\} \ln \prod_{i=1}^k (1-r_i^2)$$

que para valores grandes de n se distribuye asintóticamente como una χ^2_{pq} . La regla de decisión para una prueba aproximada de tamaño α es: rechazar H_0 si $EP \geq \chi^2_{pq}$; no rechazar H_0 , de otro modo.

Para probar la hipótesis de que “sólo s de los coeficientes de correlación poblacionales son diferentes de cero”, se usa el estadístico propuesto por Bartlett en 1939 (*ibid.*, 1989):

$$-\left\{n - \frac{1}{2}(p+q+3)\right\} \ln \prod_{i=1}^k (1-r_i^2) \sim \chi^2_{(p-s)(q-s)}, \quad (4.6-10)$$

asintóticamente. Similarmente, se rechaza la hipótesis nula si $EP \geq \chi^2_{(p-s)(q-s); \alpha}$ con un nivel de significancia aproximado de α ; no rechazar H_0 , de otro modo.

4.6.5 Puntaje¹ y Predicción (*ibid.*, p. 289)

Si en la matriz $\mathbf{X}$ se arreglan las mediciones de las q variables $\mathbf{x}$ tomadas sobre n individuos y $\mathbf{Y}$ es la matriz de datos que contiene las mediciones de las p variables $\mathbf{y}$ sobre los mismos n individuos, entonces la matriz de datos $(\mathbf{X}, \mathbf{Y})$ será de orden

¹ Esta palabra se usa en el sentido de darle una “calificación” a un individuo (*score*, en inglés).

$n \times (q + p)$ y sean los vectores $\mathbf{a}_i$ y $\mathbf{b}_i$ los i -ésimos vectores de correlaciones canónicas. Entonces los vectores $\mathbf{Xa}_i$ y $\mathbf{Yb}_i$ de tamaño n denotan los *puntajes* de los n individuos sobre las i -ésimas variables de correlaciones canónicas para $\mathbf{x}$ y $\mathbf{y}$, respectivamente. En términos de los valores de las variables para un individuo en particular, digamos el h -ésimo individuo, estos puntajes pueden tomar la forma:

$$\eta_{hi} = \mathbf{a}_i' \mathbf{x} \quad \text{y} \quad \phi_{hi} = \mathbf{b}_i' \mathbf{y}. \quad (4.6-11)$$

Para el caso de los 31 árboles del Cuadro 4.6.1, estos puntajes ya se han calculado en el Ejemplo 4.6.3.

Como las correlaciones canónicas son invariantes a cambios de escala, a veces es conveniente reemplazar estos puntajes por otros nuevos, como:

$$\eta_{hi}^* = c_1 \mathbf{a}_i' \mathbf{x} + d_1 \quad \text{y} \quad \phi_{hi}^* = c_2 \mathbf{b}_i' \mathbf{y} + d_2, \quad (4.6-12)$$

donde $c_1, c_2 > 0$ y d_1, d_2 son números reales. Estos puntajes son más importantes para los primeros vectores de correlaciones canónicas y pueden calcularse para cada uno de los n individuos.

Suponer que r_i denota al i -ésimo coeficiente de correlación canónica. Si las variables $\mathbf{x}$ y $\mathbf{y}$ son interpretadas como las variables “predictoras” y “dependientes”, respectivamente, entonces el puntaje η_i puede usarse para predecir un valor del puntaje ϕ_i usando regresión por mínimos cuadrados. Como los vectores $\mathbf{Xa}_i$ y $\mathbf{Yb}_i$ ambos tienen varianza muestral igual a 1, el valor predicho de ϕ_i dado η_i es:

$$\hat{\phi}_i = r_i (\eta_i - \mathbf{a}_i' \bar{\mathbf{x}}) + \mathbf{b}_i' \bar{\mathbf{y}}, \quad (4.6-13)$$

o equivalentemente

$$\hat{\phi}_i^* = \frac{c_2 r_i}{c_1} (\eta_i^* - d_1 - c_1 \mathbf{a}_i' \bar{\mathbf{x}}) + d_2 + c_2 \mathbf{b}_i' \bar{\mathbf{y}} \quad (4.6-14)$$

Obsérvese que r_i^2 representa la proporción de la varianza de ϕ_i que es “explicada” por la regresión en X . La predicción más importante es para la primera variable de correlación canónica, ϕ_1^* .

4.6.6 Datos Cualitativos y el Uso de Variables Ficticias (*ibid.*, p. 290)

El análisis de correlaciones canónicas puede aplicarse a datos *cualitativos*. Cualquier tabla de contingencia de dos entradas puede analizarse por correlaciones canónicas. Considerar una tabla $N_{r \times c}$ en donde los individuos se ordenan de acuerdo a dos características, cuyos atributos en cada una están representados por las categorías de las r hileras y las categorías de las c columnas, y n_{ij} denota el número de individuos que reúnen los atributos de la i -ésima hilera y la j -ésima columna. El objetivo es estudiar la relación entre esas dos características.

EJEMPLO 4.6.5 Considerar los datos del Cuadro 3.3.3 donde el atributo de las hileras son las etapas sucesionales de una selva mediana subperennifolia (son cuatro etapas sucesionales, de la perturbación más antigua a la perturbación más reciente) y el atributo de las columnas son los estratos de altura medidos en centímetros; los valores dentro del Cuadro indican el número de árboles por hectárea de la regeneración natural que pertenecen al estrato de la j -ésima altura y están en la i -ésima etapa sucesional.

Etapas sucesional	Estratos de alturas (en cm)			
	0-30	31-100	101-300	>300
I	29,050	7,475	9,450	11,350
II	16,025	5,575	5,025	5,250
III	43,250	8,600	4,225	2,225
IV	83,175	9,275	2,075	2,000

Una tabla de contingencia N no es una matriz de datos puesto que sus columnas no representan variables ni sus hileras representan individuos. Sin embargo, se pueden representar los datos en una matriz de datos $Z = (X, Y)$ de dimensión $[n \times (r + c)]$ donde

tanto las columnas de X como las de Y son variables ficticias (también llamadas “de engaño” o “pseudovariables”) constituidas de ceros y unos para las categorías de hileras y columnas, respectivamente, usando la siguiente convención:

$$x_{ki} \begin{cases} 1 & \text{si el } k\text{-ésimo individuo pertenece a la } i\text{-ésima categoría de hilera,} \\ 0 & \text{de otro modo,} \end{cases}$$

y

$$y_{kj} \begin{cases} 1 & \text{si el individuo } k \text{ pertenece a la categoría de columna } j, \\ 0 & \text{de otro modo,} \end{cases}$$

para $k = 1, \dots, n$ individuos, $i = 1, \dots, r$ hileras, $j = 1, \dots, c$ columnas. Es importante observar que $\mathbf{x}'_{(i)}\mathbf{y}_{(j)} = n_{ij}$ y que tanto las columnas de X como las de Y suman $\mathbf{1}_n$.

El propósito del análisis de correlaciones canónicas es encontrar vectores $\mathbf{a}_{r \times 1}$ y $\mathbf{b}_{c \times 1}$ tal que las combinaciones lineales $\eta = \mathbf{a}'\mathbf{x}$ y $\phi = \mathbf{b}'\mathbf{y}$ tengan la máxima correlación. Como $\mathbf{x}$ tiene un solo elemento diferente de cero, entonces η puede tomar los valores ya sea $a_1, a_2, \dots$, ó a_r ; de igual forma, como $\mathbf{y}$ tiene un solo elemento diferente de cero, $\phi = b_1, b_2, \dots$, ó b_c . Por lo tanto, un individuo en la i -ésima categoría de hilera y la j -ésima categoría de columna se le puede asociar con su “puntaje” (a_i, b_j) . Estos puntajes pueden graficarse, observándose los puntos n_{ij} en (a_i, b_j) , y las correlaciones representadas por estos puntos podrán evaluarse. La correlación depende de los valores de $\mathbf{a}$ y $\mathbf{b}$, y si estos vectores maximizan esta correlación, entonces ellos darán el vector de las primeras ponderaciones de correlaciones canónicas. Tales puntajes son conocidos como “puntajes canónicos”.

Habiendo calculado el primer conjunto de puntajes canónicos, debe uno indagar un segundo grupo que maximice la correlación entre todos los conjuntos que no están correlacionados con la primera; lo anterior correspondería a la segunda correlación canónica.

Sea $\mathbf{Z} = (\mathbf{X}, \mathbf{Y})$ la matriz de datos que corresponde a la tabla de contingencia $\mathbf{N}$. Sea $f_i = \sum_j n_{ij}$ y $g_j = \sum_i n_{ij}$, de tal forma que $\mathbf{f}$ y $\mathbf{g}$ son las sumas marginales de hileras y columnas, respectivamente, de la tabla. Por facilidad, suponer que $f_i > 0$, $g_j > 0$, $\forall i, j$.

Se puede mostrar que:

$$n\mathbf{S} = \mathbf{Z}'\mathbf{H}\mathbf{Z} = \mathbf{Z}'\mathbf{Z} - n\bar{\mathbf{z}}\bar{\mathbf{z}}'$$

$$= \begin{bmatrix} nS_{11} & nS_{12} \\ nS_{21} & nS_{22} \end{bmatrix} = \begin{bmatrix} \text{diag}(\mathbf{f}) - n^{-1}\mathbf{f}\mathbf{f}' & \mathbf{N} - \hat{\mathbf{N}} \\ \mathbf{N}' - \hat{\mathbf{N}}' & \text{diag}(\mathbf{g}) - n^{-1}\mathbf{g}\mathbf{g}' \end{bmatrix}, \quad (4.6-15)$$

donde $\hat{\mathbf{N}} = n^{-1}\mathbf{f}\mathbf{g}'$ es el valor estimado de $\mathbf{N}$ bajo el supuesto de que las categorías de hileras y columnas de la tabla son independientes.

Como tanto las columnas de $\mathbf{X}$ como las columnas de $\mathbf{Y}$ suman $\mathbf{1}$ (vector), entonces $\mathbf{S}_{11}^{-1}$ y $\mathbf{S}_{22}^{-1}$ no existen. Una forma de encontrar una salida es eliminando una columna tanto en $\mathbf{X}$ como en $\mathbf{Y}$, digamos la primera columna en ambas. Denotar con $\mathbf{S}_{ij}^*$ la componente de la matriz de covarianzas obtenida por eliminar la primera hilera y la primera columna de $\mathbf{S}_{ij}$ para $i, j = 1, 2$. Análogamente, sean $\mathbf{f}^*$ y $\mathbf{g}^*$ los vectores obtenidos por eliminar las primeras componentes de $\mathbf{f}$ y $\mathbf{g}$, respectivamente. Entonces se tiene que:

$$\left(n\mathbf{S}_{11}^*\right)^{-1} = [\text{diag}(\mathbf{f}^*)]^{-1} + f_1^{-1}\mathbf{1}\mathbf{1}', \quad \left(n\mathbf{S}_{22}^*\right)^{-1} = [\text{diag}(\mathbf{g}^*)]^{-1} + g_1^{-1}\mathbf{1}\mathbf{1}' \quad (4.6-16)$$

así que $\mathbf{M}_1^* = \mathbf{S}_{11}^{*-1}\mathbf{S}_{12}^*\mathbf{S}_{22}^{*-1}\mathbf{S}_{21}^*$ y $\mathbf{M}_2^*$ se calculan directamente para el conjunto de variables reducidas. Para los individuos que están en la primera categoría de hilera de $\mathbf{N}$ así como los de la primera categoría de columna se le asocia un puntaje 0.

EJEMPLO 4.6.6 Con los datos de número de individuos por etapa sucesional y estrato de altura del Cuadro 3.3.3 (escrito otra vez en esta sección) se puede hacer un análisis de

correlaciones canónicas para estudiar la relación entre etapas sucesionales y estratos de alturas. Los resultados obtenidos con el Programa 4.6.7.2 de SAS son los siguientes.

R1	R2	R3
0.3875822	0.0764427	0.0117915

R1 indica la primera correlación canónica r_1 y así enseguida. También se tiene que A1 y

A1	A2	A3
0	0	0
0.184749	1.9845821	2.8177751
1.7288354	2.32307	-0.238789
2.3042959	0.2399547	1.1906353

B1	B2	B3
0	0	0
0.9433311	2.4607113	1.5390834
2.436569	1.1612866	-2.425597
2.8850967	-1.518767	1.5844049

B1 representan los primeros vectores de correlaciones canónicas; A2 y A2, los segundos y A3 y B3 los terceros. Tomando en cuenta sólo A1 y B1, puede observarse que sus últimos dos coeficientes se aproximan más entre ellos que los primeros dos; esto podría interpretarse como la estabilidad que adquiere el bosque en etapas sucesionales superiores en su estratificación de alturas.

4.6.7. Anexo de Programas

Programa 4.6.7.1

```

/*****
/*PROGRAMA IML/SAS PARA EL ANALISIS DE CORRE */
/*LACIONES CANONICAS MUESTRAL USANDO R */
/*
aaas*/
/*****
DATA XIMBAL;
INFILE 'C:\CORRCAN\TREES.DAT';
INPUT NUM $ SP $ DN AT AFL RBC;
/*DN:diametro normal, AFT;altura de fuste limpio*/
/*RBC:radio de la base de la copa, Y2:altura total*/
X1=DN;X2=AFL;Y1=RBC;Y2=AT;

PROC IML;USE XIMBAL;
START CORRCA;

```

```

READ ALL VAR {X1 X2 Y1 Y2} INTO X;
READ ALL VAR {X1 X2} INTO X1;
READ ALL VAR {Y1 Y2} INTO Y1;

Q=NCOL(X);
N=NROW(X);           /*numero de individuos*/
J=J(N,1,1);           /*vector de unos de tamaño n*/
H=I(N)-(1/N)*J*J';
MED=(1/N)*X'*J;

S=(1/N)*X'*H*X;       /*calculo de la matriz de covarianzas muestral*/
SRC=S##0.5;
SREC=1/SRC;
DM12=DIAG(SREC);
R=DM12*S*DM12;        /*calculo de la matriz de correlaciones*/
R11=R[1:2,1:2];
R22=R[3:4,3:4];
R12=R[1:2,3:4];
R21=R[3:4,1:2];
B1=INV(R11);           /*inversa de R11*/
B2=INV(R22);           /*inversa de R22*/

PRINT MED S R;
PRINT R11 R22 R12 R21; /*imprime las particiones de R */
PRINT B1 B2;

CALL EIGEN(Q,U,B1);    /*Q contiene eigenvalores de B1*/
A=SQRT(DIAG(Q));       /*U sus corresp. eigenvectores*/
RC1=U*A*U';            /*R11 a la menos un medio*/

CALL EIGEN(Q,U,B2);    /*Q contiene eigenvalores de B2*/
A=SQRT(DIAG(Q));       /*U sus corresp. eigenvectores*/
RC2=U*A*U';            /*R22 a la menos un medio*/

K=RC1*R12*RC2;         /*K de la expresion */

CALL SVD(ALFA,D,BETA,K); /*ALFA contiene eigenvectores de N1=K*K'*/
PRINT ALFA D BETA;      /*BETA contiene eigenvectores de N2=K'*K*/
                        /*D contiene sqrt(eigenvalores) de N1*/

K1=ALFA*DIAG(D)*BETA'; /*descomposicion del valor singular de K*/

Ai=RC1*ALFA;
Bi=RC2*BETA;
R1=D[1,];               /*primer coeficiente de correlacion canonica*/
R2=D[2,];               /*segundo coeficiente de correlacion canonica*/
A1=Ai[,1];A2=Ai[,2];    /*vectores de correlaciones canonicas para primer grupo de var*/
B1=Bi[,1];B2=Bi[,2];    /*vectores de correlaciones canonicas para segundo grupo de var*/
VCCA1=X1*A1; VCCB1=Y1*B1; /*primeras variables de correlacion canonica*/

```

VCCA2=X1*A2; VCCB2=Y1*B2; /*segundas variables de correlacion canonica*/

PRINT A1 A2 B1 B2;
PRINT VCCA1 VCCA2 VCCB1 VCCB2;
PRINT R1 R2;

FINISH;
RUN CORRCA;
QUIT;

Programa 4.6.7.2

```
/******  
/*PROGRAMA IML/SAS PARA UN ANALISIS DE CORRELACIONES CANO-*/  
/*NICAS CON DATOS CUALITATIVOS (TABLAS DE DOBLE ENTRADA)1 */  
/*  
/******  
/*cualitva.sas)*/
```

```
PROC IML;  
START CUALITA;  
N={29050 7475 9450 11350,  
    16025 5575 5025 5250,  
    43250 8600 4225 2225,  
    83175 9275 2075 2000};  
nt=N[+,+]; /*suma todos los valores del Cuadro de contingencia N*/  
PRINT N NT; /*imprime N y NT*/  
  
F=N[+,+]; /*un vector que contiene las sumas de columnas de N*/  
F1=F[1,]; /*el primer valor de F*/  
FE=F[2:4,]; /*los ultimos 3 valores de F = F estrella*/  
  
G=N[+,+]; /*un vector que contiene las sumas de hileras de N*/  
G1=G[1,]; /*el primer valor de G*/  
GE=G[2:4,]; /*los ultimos 3 valores de G = G estrella*/  
  
NES=(F*G')/nt; /*N estimada (gorro) bajo el supuesto de independencia entre  
categorias de hileras y de columnas*/  
PRINT F G NES;  
  
nS11=DIAG(F)-((F*F')/nt); S11=nS11/nt; S11E=S11[2:4,2:4]; /* S11 estrella */  
nS22=DIAG(G)-((G*G')/nt); S22=nS22/nt; S22E=S22[2:4,2:4]; /* S22 estrella */  
nS12=N-NES; S12=nS12/nt; S12E=S12[2:4,2:4]; /* S12 estrella */  
nS21=N'-NES'; S21=nS21/nt; S21E=S21[2:4,2:4]; /* S21 estrella */  
PRINT S11E S22E S12E S21E;
```

L1=EIGVAL(S11E);

¹ Fuentes Romero, A. Margarita (Comunicación Personal)


```

G1=EIGVEC(S11E);
LI1=INV(SQRT(DIAG(L1)));
NS1=G1*LI1*G1';
M1=NS1*S12E*INV(S22E)*S21E*NS1;    /* M1 estrella */
L2=EIGVAL(S22E);
G2=EIGVEC(S22E);
LI2=INV(SQRT(DIAG(L2)));
NS2=G2*LI2*G2';
M2=NS2*S21E*INV(S11E)*S12E*NS2;    /* M2 estrella */
LM1=EIGVAL(M1);
CORRCAN=SQRT(LM1);                  /* correlaciones canonicas */
GM1=EIGVEC(M1);GM2=EIGVEC(M2);
A=NS1*GM1;                          /* vector canonico para el estatus del padre */
B=NS2*GM2;                          /* vector canonico para el estatus del hijo */
R1=CORRCAN[1,];                      /* correlacion canonica */
R2=CORRCAN[2,];                      /* correlacion canonica */
R3=CORRCAN[3,];

A1=A[,1];                           /* vector canonico */
A2=A[,2];                           /* vector canonico */
A3=A[,3];                           /* vector canonico */

B1=B[,1];                           /* vector canonico */
B2=B[,2];                           /* vector canonico */
B3=B[,3];                           /* vector canonico */

PRINT R1 R2 R3;
PRINT A1 A2 A3;
PRINT B1 B2 B3;
FINISH;
RUN CUALITA;
QUIT;

```

V CONCLUSIONES

Los métodos multivariados permiten el análisis de datos del bosque que contienen más de una variable. Los ejemplos desarrollados en este documento fueron con la intención de hacer hincapié en este hecho. Los métodos multivariados pueden significar una ventaja para los investigadores en recursos naturales que trabajan con datos sobre varias dimensiones, ya que permiten un análisis que incluye diferentes aspectos del fenómeno que se estudia. Es decir, los métodos multivariados permiten que un problema se explore, simultáneamente, desde distintos ángulos; esto es menos lejano de como ocurre en la práctica. Sin embargo, hay que admitir que no siempre son necesarios.

En ocasiones, el tipo de estudio que se realiza o los alcances que se le quieren dar a éste, no exige más que el uso de estadísticos univariados o, como ocurre en no muy pocas situaciones, ¡sólo basta con la inspección visual! Lo anterior puede ser explicado con un fenómeno análogo. El ser humano tiene la capacidad de observar en tres dimensiones; esta característica le permite distinguir si el objeto A está más cerca de él en comparación al objeto B, o viceversa; y esta distinción lo puede hacer, razonablemente, con varios objetos. Sin embargo, aún cuando para la mayoría de los fenómenos le resulta cómodo mirar tridimensionalmente, habrán situaciones en las que le baste mirar sólo dos dimensiones del entorno. De forma similar, los métodos multivariados permiten visualizar un problema desde varias dimensiones, pero algunos diagnósticos no lo requieren. Entonces, los métodos multivariados son útiles para estudios amplios que intenten ser detallados y más completos.

Muchas de estas técnicas que se discuten y que se ejemplifican requieren algunas condiciones para ser aplicadas. Es importante considerarlas desde la planeación del

estudio ya que de ahí dependerá obtener resultados buenos o mediocres. Normalmente las técnicas estadísticas pueden ser aplicadas a una variedad de problemas, pero debe ser el fenómeno y las características, así como los alcances del estudio lo que determine finalmente qué método es conveniente. No sería correcto, y ni siquiera ético, usar alguna de estas técnicas sólo porque suenen bonito. Hay que entender qué es lo que hacen, cómo lo hacen y cuáles son sus limitaciones, para poder emplearlas ventajosamente.

VI LITERATURA CITADA

- Ali A., S. & S. W. Yap. 1992. Effects of two different growth regulators on the growth and water relations of *Acacia mangium* seedlings. J. of Tropical Forest Science, 6(4):489-501.
- Biging G., S & S. J. Gill. 1997. Stochastic models for conifer tree crown profiles. For. Sci. 43(1):25-34.
- Burkhart, H. E. & T. G. Gregoire. 1994. Forest Biometrics. In: Handbook of Statistics, Vol. 12; Environmental Statistics. Edited by G.P. Patil & C.R. Rao, Elsevier Science B.V., The Netherlands. pp. 377-407.
- Caballero D., M. 1967. Estudio comparativo de dos especies de pinos mexicanos (*Pinus pseudostrobus* Lindl. y *Pinus montezumae* Lamb.) con base en características de plántula y semilla. Boletín técnico No. 20. Instituto Nacional de Investigaciones Forestales.
- Clark, D. C. & K. F. Haynes. 1992. Sublethal effects of cypermethrin on chemical communication, courtship, and oviposition in the cabbage looper (Lepidoptera: Noctuidae). Journal of Economic Entomology, 85(5): 1171-1178.
- Cochran, W. G. & G. M. Cox. 1983. Diseños Experimentales. Traducido del inglés por el Centro de Estadística y Cálculo del Colegio de Postgraduados. Octava reimpresión. Trillas, México. 654 p.
- Cortez J., G. 1994. Respuesta de *Pinus cooperi* Martínez a aclareos en el S.P.E.F. Cielito Azul en San Miguel de Cruces, Durango. Tesis de maestría. División de Ciencias Forestales. Universidad Autónoma Chapingo. Chapingo, Edo. de Méx., México. 110 p.
- Daniel, W. 1983. Bioestadística: Base para el Análisis de las Ciencias de la Salud. Traducido del inglés por J. H. Pérez Castellanos. Cuarta reimpresión. Limusa, México. 485 p.
- Freese, F. 1969. Muestreo Forestal Elemental. Traducido del inglés por el Centro Regional de Ayuda Técnica, Agencia para el Desarrollo Internacional (A.I.D.) de los EE.UU. A.I.D., México/Buenos Aires. 96 p.
- Grijpma, P. & S. C. Roberts. 1976. Biological and chemical screening for the basis of resistance of *Toona ciliata* M. J. Roem. Var. Australis. In: Studies on the Shootborer *Hypsipyla grandella* (Zeller); Lep. Pyralidae. Vol 2. CATIE, Turrialba, Costa Rica. pp. 102-109.

- Hocker, H. W. 1984. Introducción a la Biología Forestal. Traducido del inglés por F. Bellomo López. AGT, México. 446 p.
- Johnson, R. A. & D. W. Wichern. 1982. Applied Multivariate Statistical Analysis. Prentice-Hall, New Jersey. 594 p.
- López T., J. L. 1995. Alometría de variables dasométricas en árboles de vegetación secundaria derivada de una selva mediana subperennifolia. Tesis de maestría. División de Ciencias Forestales. Universidad Autónoma Chapingo. Chapingo, Edo. de Méx., México. 83 p.
- Manly, B. F. J. 1994. Ecological Statistics. *In*: Handbook of Statistics, Vol. 12; Environmental Statistics. Edited by G.P. Patil & C.R. Rao, Elsevier Science B.V., The Netherlands. pp. 337-376.
- Mardia, K. V; J. T. Kent & J. M. Bibby. 1989. Multivariate Analysis. Seventh printing. Academic Press, London. 521 p.
- Mendoza B., M. A. 1993. Conceptos Básicos de Manejo Forestal. UTEHA, México, D.F. 161 p.
- Montagnini, F. *et al.* 1992. Sistemas Agroforestales: principios y aplicaciones en los trópicos. 2da. edición. OET-IICA. San José, Costa Rica. 622 p.
- Morrison, D. F. 1981. Multivariate Statistical Methods. Second ed., third printing. International Student Edition. McGraw-Hill, Tokio. 415 p.
- Ogbonnaya, C. Y. 1994. Growth and histo-chemical response of *Gmelina arborea* seedlings to applications of N and K fertilizers and their combinations on oxisolic soil. *Journal of Tropical Forest Science*, 6(4): 363-378.
- Olf, H.; D. M. Pegtel; J.M. Van Groenendael; J. P. Bakker. 1994. Germination strategies during grassland succession. *Journal of Ecology*, 82: 69-77.
- Ostle, B. 1973. Estadística Aplicada. Traducido de inglés por D. de la Serna Valdivia. Tercera reimpresión. Limusa, México. 629 p.
- Pla, L. E. 1986. Análisis Multivariado: Método de Componentes Principales. Organización de los Estados Americanos. 90 p.
- Repasky, R. R & D. Schluter. 1994. Habitat distributions of wintering sparrows along an elevational gradient: tests of the food, predation and microhabitat structure hypotheses. *Journal of Animal Ecology*, 63: 569-582.
- Sánchez M., V. 1993. Estratificación vertical de una selva mediana subperennifolia en San Felipe Bacalar, Quintana Roo. Tesis de maestría. División de Ciencias Forestales. Universidad Autónoma Chapingo. Chapingo, Edo. de México, México. 130 p.
- Santillán P., J. 1986. Elementos de Dasonomía. Universidad Autónoma Chapingo. Chapingo, Edo. de Méx., México. 346 p.
- SAS Institute Inc. 1987. SAS/STAT Guide for Personal Computers, Version 6 Edition. Cary, NC: SAS Institute Inc. 1028 p.

- SAS Institute Inc. 1988. SAS/IML User's Guide, Release 6.03 Edition. Cary, NC: SAS Institute Inc. 357 p.
- Searle, S. R. 1982. Matrix Algebra Useful for Statistics. John Wiley & Sons, New York. 438 p.
- Srikantan, K. S. 1970. Canonical association between nominal measurements. *Journal of the American Statistical Association*, 65(329):284-292.
- Steel, R. & J. Torrie. 1988. Bioestadística: Principios y Procedimientos. Traducido por Ricardo Martínez. McGraw-Hill, México, D.F. 621 p.
- Stiteler, W. M. 1979. Multivariate Statistics with Applications in Statistical Ecology. *In*: Statistical Ecology Series, Vol 7; Multivariate Methods in Ecological Work. Edited by L. Orloci, C. R. Rao and W. M. Stiteler. International Co-operative Publishing House, Fairland, Maryland. pp. 279-300.
- Winer, B. J.; D. R. Brown; K. M. Michels. 1991. Principles of Experimental Design. Third edition. McGraw-Hill, U.S.A.
- Young, R. A. 1991. Introducción a las Ciencias Forestales. Traducido del inglés por José Hurtado Vega. Noriega-Limusa, México.
- Zamudio S., F. J. (En prensa, a). Regresión: Geometría y Pruebas de Hipótesis. Un nivel intermedio. Universidad Autónoma Chapingo. Chapingo, Edo. de Méx., México. 119 p.
- Zamudio S., F. J. y A. A. Alvarado S. (En prensa) Análisis de diseños experimentales con igual número de submuestras. Universidad Autónoma Chapingo. Chapingo, Edo. de Méx., México.
- Zimmerman J. K.; E. M. Everham III; R. B. Waide; D. J. Lodge; C. M. Taylor; N. V. L. Brokaw. 1994. Responses of tree species to hurricane winds in subtropical wet forest in Puerto Rico: implications for tropical tree life histories. *Journal of Ecology*, (82): 911-922.

